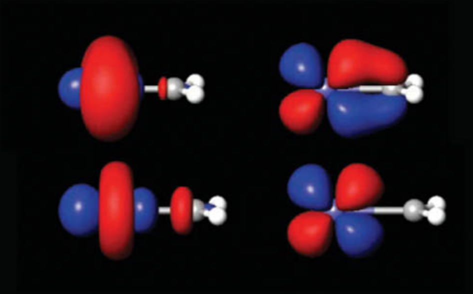
Introduction to
Relativistic Quantum Chemistry



KENNETH G. DYALL

KNUT FÆGRI, JR.

Introduction to Relativistic Quantum Chemistry

This page intentionally left blank

Introduction to RELATIVISTIC QUANTUM CHEMISTRY

Kenneth G. Dyall
Knut Fægri, Jr.

OXFORD
UNIVERSITY PRESS

2007

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Library of Congress Cataloging-in-Publication Data
Dyall, Kenneth G., 1955–
Introduction to relativistic quantum chemistry / Kenneth G. Dyall,
Knut Faegri, Jr.
p. cm.
Includes bibliographical references.
ISBN: 978-0-19-514086-6
1. Quantum chemistry. 2. Quantum field theory. I. Faegri, Knut.
II. Title.
QD462.D93 2006
541'.28–dc22 2006014117

9 8 7 6 5 4 3 2 1
Printed in the United States of America
on acid-free paper

Preface

The emergence of relativistic quantum chemistry has been one of the more remarkable developments within computational chemistry over the past decades. Since the early work of Dirac, relativity has always been a part of the overall quantum chemical picture, but it has mostly been neglected on the grounds that the effects were considered small and the methods to treat them were poorly developed and expensive to use. However, as nonrelativistic quantum chemistry became more powerful and accurate, the lower rows of the periodic system came within reach of computational studies, and it became clear that relativistic effects had a significant influence on a number of physical and chemical properties. The start of the "modern" era of relativistic quantum chemistry may be traced back to a review article by Pyykkö (1978) and to articles by Pitzer (1979) and by Pyykkö and Desclaux (1979).

The developments in the field have been well documented through articles in scientific journals, conferences and symposia, and review volumes. Unfortunately, this specialist literature is not easily accessible to newcomers to the field. For many years the book by Moss, *Advanced Molecular Quantum Mechanics* (1973), served as an introduction to the field, but today this book suffers from two major drawbacks—it is out of print and it does not cover the developments of the past three decades. We therefore decided that there was a need for a book to fill the gap between the standard texts on quantum mechanics, which have little if anything on relativity, the advanced texts, which treat relativity in detail but have little connection with quantum chemistry, and the literature, where there is a large amount of both theory and applications.

Our ambition is to provide a modern introduction to the field of relativistic quantum chemistry, aimed at the advanced student and the practicing nonspecialist researcher. The material has been divided into five parts. Parts I and II provide the necessary background from classical physics, relativistic quantum mechanics, and group theory. Part III covers the application of these principles to fully relativistic methods for quantum chemistry within a four-component framework. Part IV deals with the main approximate methods that have been developed, and Part V treats the consequences of relativity for chemical bonding.

The book is introductory in the sense that it introduces many of the concepts needed for a firm background in relativistic theory and quantum chemistry but without going into all the details. References are given to fuller treatments of the introductory material. Our intention is not to give all the details of proofs that can be found in other standard works on relativistic quantum mechanics, but rather to present the relevant parts for the purpose of constructing a relativistic quantum chemistry. It is also introductory in the sense that the details of many of the methods that are found in the literature are reproduced or elaborated in this book, so that the educated quantum chemist does not have to search through the literature for them. Finally, it is introductory in the sense that it contains descriptive material to do with relativistic effects on bonding, structure, and energetics of molecules. We have no ambitions of providing an extensive review of what has become a large and quite heterogeneous field, nor of providing a historical overview of the development of relativistic methods. The specialist will probably have no difficulty identifying one or more "pet subjects" or "key references" that are missing. We do provide a guide to the review literature in the field, and also in this respect the work is introductory.

The book started out as a set of lecture notes by KGD for a 1995 graduate course given at (then) Odense University. These notes developed through further presentations, and the process of turning them into a book was begun in 1996 during a research visit by KGD to the University of Oslo. Since then the writing has been a gradual process, hampered by the academic work uncertainties of one of us and by the extensive administrative load carried by the other. The work would not have been possible had it not been for the support of our employers through this period, for KGD: Eloret and Schrödinger; for KF: the University of Oslo. Support by the Research Council of Norway, NASA, and DOE is gratefully acknowledged, as is the hospitality both of us enjoyed during research visits to the IRSAMC of Université Paul Sabatier in Toulouse.

Most of all, this book could never have been written without support and criticism from and frequent discussions with our friends in the field. These include in particular Harry M. Quiney, Luuk Visscher, Trond Saue, Hans Jørgen Aa. Jensen, and Trygve Helgaker, as well as most of the people involved in the DIRAC and MOLFDIR collaborations. Peter Schwerdfeger did a wonderful job of commenting on a late version of the manuscript. Trond Saue and Luuk Visscher also provided useful comments on substantial parts of a final draft. Others who have very kindly read and commented on parts of the manuscript in various stages of completion are Joost v. Stralen, Raimo v. Handel, Leif Veseth, Werner Kutzelnigg, Timo Fleig, and Hans Jørgen Aa. Jensen. In addition to the direct contributions to the book, there are many who have in one way or another influenced our thinking or contributed to the work on which this book was based. Among those not already mentioned we would like to acknowledge Peter Taylor, Jeppe Olsen, Wim Nieuwpoort, and Gustavo Aucar. Finally, we would like to express our gratitude to Pekka Pyykkö and Ian Grant, both pioneers and leaders in this field, and to Bernd A. Hess, Jaap Snijders, Jan Almlöf, and Odd Gropen, who are no longer with us, but who all four in various ways have helped us in our efforts in this field.

In our families the book has taken on almost mythical stature. It is with some relief that we are now able to present a final product. We do this in grateful acknowledgement of their almost infinite tolerance and support.

Notation Conventions

We have adopted a number of conventions in this book in order to maintain a consistent, clear, and identifiable notation. As far as possible we have kept to common conventions for symbols and quantities in quantum chemistry. We have also tried to avoid the duplication of symbols where possible. These goals conflict to some extent, so some quantities are given unconventional symbols. The following list identifies symbols and typography used throughout the book.

r, ϑ, φ for spherical coordinates
$\mathit{\Psi}$ for a general one-particle or multideterminant many-particle wave function
$\mathit{\Phi}$ for a one-determinant or CSF many-particle wave function
Ψ for time-dependent 4-spinors
$\theta(t)$ for the time part of a spinor or a general time-dependent function
$\uppsi$ for time-independent 4-spinors, that is $\Psi = \uppsi\theta$
Ψ^X for time-dependent 2-spinors, with $X = L,\ S$
$\uppsi^X$ for time-independent 2-spinors, with $X = L,\ S$
$\upxi(\vartheta, \varphi, \tau)$ for 2-spinor angular-momentum functions (including spin)
$\upphi, \upphi^X$ for 2-spinors in a modified representation
$\upchi_\mu^X$ for 2-spinor basis functions
$\psi, \psi^{X\tau}$ for scalar functions
$\upeta, \upalpha, \upbeta$ for spin functions
τ, α, β for spin labels; α and β also denote the spin functions, and β one of the Dirac matrices;
$\phi, \phi^{X\tau}$ for scalar space functions; the superscripts X and τ are for component and spin.
$\chi_\mu^X, \chi_\mu^{\bar{X}}$ for scalar basis function
$\mathbf{A}, \boldsymbol{\Omega}$ boldface for matrices
$\mathbf{a}, \boldsymbol{\sigma}$ boldface for vectors

a Roman type for four-vectors
$\hat{F}$, $\hat{\Omega}$ hats for operators
h_{pq} for one-electron Hamiltonian matrix elements
H_{PQ} for N-electron Hamiltonian matrix elements

In addition to these font conventions, we have adopted some conventions for the indices of functions. In general, Roman letters are used for Fock space functions (i.e. orbitals or spinors), while Greek letters are used for basis functions. Specific ranges of letters are used as follows:

$p, q, r, s, \ldots$ general orbital or spinor indices
$i, j, k, l, \ldots$ occupied or inactive orbital or spinor indices
$a, b, c, d, \ldots$ virtual orbital or spinor indices
$t, u, v, w, \ldots$ active orbital or spinor indices
$\kappa, \lambda, \mu, \nu, \ldots$ basis function indices

In this book we have used two systems of units. The first is the SI system, which we use in the early chapters of the book and in particular for electromagnetic quantities. Factors of c therefore always represent the speed of light and never a conversion factor for magnetic units. The second is the Hartree atomic units system, defined by $\hbar = e = m_e = 1$. In these units, $c \sim 137$. Many physics texts use the system $\hbar = e = c = 1$, since they are dealing with particles of different masses. Our concern is principally with the electron and chemistry, and the size of relativistic effects, which are measured by c, so Hartree atomic units are more appropriate. However, to keep the connection with SI units and to track quantities that involve the charge, the mass, or spin, the symbols $\hbar$, e, and $m \equiv m_e$ are retained in much of the development, whereas $1/4\pi\epsilon_0$ is usually omitted for clarity.

There are some situations where, in the interests of clarity, we have allowed some inconsistency or sacrificed some rigor of expression. We do not usually multiply scalars by the unit matrix in expressions where the context would demand it—such as where an operator is a combination of scalar and spin-dependent operators or in a matrix expression—and we do not always indicate the rank of the unit matrix or the zero matrix by a subscript. In some places the notation would be overloaded if $\mathbf{I}_2$ were used instead of 1, but the matrix notation is to be inferred from the context.

We have also used both $\hat{\mathcal{H}}$ and $\hat{h}$ for the one-electron Hamiltonian operator. The latter is used for the free-particle Dirac Hamiltonian where a distinction between it and the full one-electron Hamiltonian is necessary, and is also used in a sum over one-electron Hamiltonians for a single electron. The former is usually used in formal developments, and to represent the total Hamiltonian. In many of the formal developments, the total Hamiltonian is simply the one-electron Hamiltonian, so $\hat{\mathcal{H}}$ is used. However, for the one-electron Hamiltonian matrix elements, lower case is always used, and for the N-electron Hamiltonian matrix elements, upper case is always used.

Contents

I: Foundations

II: The Dirac Equation: Solutions and Properties

III: Four-Component Methodology

IV: Approximations to the Dirac Equation

Appendices

Part I

Foundations

This page intentionally left blank

1

Introduction

> The general theory of quantum mechanics is now almost complete, the imperfections that still remain being in connection with the exact fitting of the theory with relativity ideas. These give rise to difficulties only when high-speed particles are involved, and are therefore of no importance in the consideration of atomic and molecular structure and ordinary chemical reactions, in which it is indeed usually sufficiently accurate if one neglects relativity variation of mass with velocity and assumes only Coulomb forces between the various electrons and atomic nuclei. The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations that are too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.

The quote above from Paul Adrian Maurice Dirac (1929) has been somewhat of an article of faith for modern quantum chemistry. Intensive efforts on the development of theory, algorithms, and techniques have made computational quantum chemistry a very successful representative of the “third way” in modern science—computer modeling has come into its own alongside experiment and theory. Fifty years ago this was a branch of science where predictions were at best qualitative, founded on rather approximate models. Many of these models were quite sophisticated, and much of the insight gained is still valid and valuable, but the developments in both methods and computer hardware up to the present have very much transformed this field. Today standard quantum chemical methods are capable of predicting results with chemical accuracy: reaction energies may be determined within a few kilojoules per mole, and spectral data within a few reciprocal centimeters. At least, this sort of reliability can be expected for “normal” areas of application.

Dirac’s main contribution to science was a merging of the two great developments of 20th century physics—quantum mechanics and the (special) theory of relativity.

Most of the successful development in quantum chemistry has been based on nonrelativistic quantum mechanics. This may be justified by considering that special relativity is needed primarily to describe objects moving at velocities approaching the speed of light, and that this is mostly not the case for chemical systems. After all, most chemical reactions and phenomena occur at energies below the relativistic domain.

Or could relativistic effects nevertheless be important? Even without the recent advances in computational chemistry, it became clear fairly early that nonrelativistic theory was unable to explain certain trends in observed properties. A few examples will suffice to illustrate the anomalies.

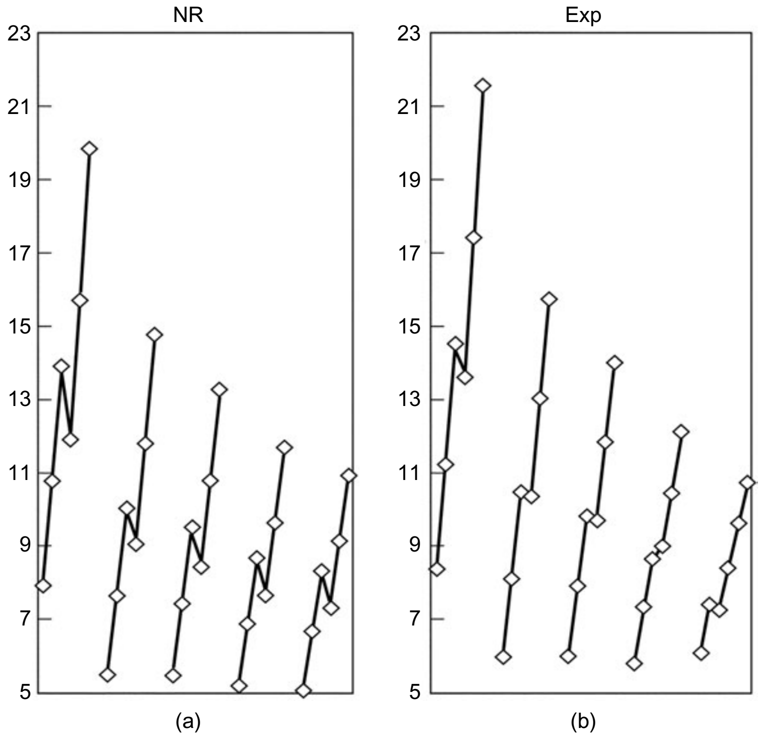
Experimental determination of the metal–carbon bond length in the group 12 dimethyl compounds showed an increase in bond length from Zn to Cd, but a decrease in bond length from Cd to Hg (Rao et al. 1960). The expected trend was an increase from Zn to Cd and again from Cd to Hg. The conclusion was "this anomalous magnitude of bond lengths for the Cd and Hg compounds is not understood." The decrease in bond length is due to relativistic effects. One very well studied series is the coinage metal hydrides, CuH, AgH, and AuH, which show the same trend in the bond lengths. Nonrelativistic Hartree–Fock (NRHF) calculations on AuH predict a bond length of 183 pm, which is longer than that of AgH, whereas fully relativistic Dirac–Hartree–Fock (DHF) calculations give 157 pm, which is shorter than that of AgH and in line with the observed trend. The experimental bond length for AuH is 152 pm, and the error in the DHF result is largely due to neglect of electron correlation.

A second anomaly is evident in the trends in the ionization potentials of the p-block elements. From nonrelativistic theory we would expect an increase in ionization potential as the occupation of the valence p shell goes from p^1 to p^3. This is due to an increase in nuclear charge that is only partly compensated by the screening from the other valence electrons. From p^3 to p^4 we expect a decrease in ionization potential due to the loss of exchange energy with electron pairing. Finally, from p^4 to p^6 we expect a similar increase as from p^1 to p^3 for the same reason.

This pattern is indeed followed quite faithfully when the principal quantum number n is 2, 3, and 4. For $n = 5$ there is a marked reduction in the increase from p^2 to p^3 compared to the previous series, but the pattern still holds. For $n = 6$, however, we observe a decrease of ionization potential from p^2 to p^3 and an increase from p^3 to p^4. These trends are depicted in figure 1.1, where the experimental values are shown with predictions from nonrelativistic Hartree–Fock calculations. It is clear that the nonrelativistic calculations do not produce the correct trend. The explanation lies in the spin–orbit splitting of the p shell.

There are a number of other effects, in particular for systems containing heavy elements, which nonrelativistic theory fails to predict or adequately explain, but which are explained by relativistic theory. Among these are the stabilization of lower valence states for the $6p$ block (the "inert pair" effect) and the stabilization of the higher valence states for the $5d$ block.

This book is devoted to developing the necessary tools for extending the realm of reliable quantum chemical calculations to cover cases like those listed above. In order to do that we will have to review some basic subjects in physics. Most of these are the



(a) (b)

Figure 1.1 Ionization potentials in eV of the p block elements as a function of occupation number for each principal quantum number from $n = 2$ (left) to $n = 6$ (right) : (a) nonrelativistic Koopman's theorem; (b) experiment.

subject of entire textbooks by themselves, and we will only briefly recapitulate those themes that are of interest in developing a relativistic quantum chemistry. With the necessary tools from physics in place, we will show how to establish a fully relativistic quantum chemistry, discuss the development of various approximations to this model, and end with a qualitative overview of the effects of relativity on bonding.

2

Basic Special Relativity

Special relativity is a fascinating and challenging branch of physics. It describes the physics of the high velocity/high energy regime, frequently turning up phenomena that appear paradoxical in view of our everyday experience. In this book we will be quite selective in our presentation of the theory of special relativity: we will concentrate on those features that we consider necessary for the later applications to relativistic quantum chemistry. We do this in good conscience, knowing that there is a vast literature on the subject, catering to a wide range of audiences—from the quite elementary to the very sophisticated. A few examples are listed in the reference list, but a visit to any nearby physics library will provide an ample selection of reading material for those wishing to delve deeper into the matter.

In the present chapter we adopt a minimalist approach. We develop some of the basic concepts and formulas of special relativity, building on a rather elementary level of basic physics. The aim is to provide a sufficient foundation for those who want to proceed as quickly as possible to the more quantum chemical parts of the text. In later chapters we will introduce more advanced tools of physics and revisit some of the subjects treated here.

2.1 Inertial Frames and Newtonian Mechanics

The theory of special relativity deals with the description of physical phenomena in frames that move at constant velocity relative to each other. The classroom is one such frame, the car passing at constant speed outside the classroom is another. The trajectory of a ball being thrown up vertically in the car will look quite different whether we describe it relative to the interior of the car or relative to the interior of the classroom. In particular we will be concerned with inertial frames. We define an inertial frame as a frame where spatial relations are Euclidean and where there is a universal time such that free particles move with constant velocities.

In classical Newtonian mechanics, relations between the spatial parameters and time in two inertial frames S and S' are expressed in terms of the Galilean transformations. Assume that S' is moving with constant speed v in the direction of the positive x axis of S. If the coordinate axes of S' are parallel to those of S, the Galilean coordinate transformations are

$$\begin{aligned} x' &= x - vt; \quad x = x' + vt; \\ y' &= y; \\ z' &= z; \\ t' &= t. \end{aligned} \tag{2.1}$$

A basic tenet of classical physics is that a physical law should have the same form in all inertial frames. This certainly holds for ordinary classical mechanics at “normal” speeds and energies using the Galilean transformations. However, when the same transformations are applied to the fundamental equations of the electromagnetic theory of radiation, the Maxwell equations, they change form. This, together with several other clues, indicated that the Galilean transformations might be inadequate, at least for electromagnetic phenomena.

We will not follow the historical development, but proceed straight to the resolution of these difficulties. The solution was provided by Albert Einstein in 1905, and may be cast in the form of the two postulates of special relativity:

> *Postulate 1. The laws of physics are identical in all inertial frames.*
> *Postulate 2. In empty space light signals propagate in straight lines with speed c in all inertial frames.*[1]

Postulate 1 is basically a reassertion of the old principle of transferability of physical laws between inertial frames. The consequence of these postulates was that the transformation equations, and in some cases the laws themselves, had to be modified. The following sections will be devoted to these modified laws and transformations.

2.2 Relativistic Coordinate Transformations

If the Galilean transformations fail, what should they be replaced by? Let us start from the postulates and consider transformations between the two simplest possible inertial frames: two one-dimensional frames S and S', where S' is moving along the only dimension (x axis) with constant speed v. (The approach adopted in this section follows that of French (1968).) The relations

$$x' = x - vt; \quad x = x' + vt; \quad t' = t \tag{2.2}$$

---

[1] This postulate is usually presented as an independent postulate in most texts, but in fact postulate 2 follows from postulate 1, as shown by Frank and Rothe (1911).

must be replaced by something different, but something that should reduce to the expressions above for situations where we know the Galilean transformations give the correct results. It appears reasonable to try a set of linear transformations of the form

$$x' = ax - bt; \quad x = ax' + bt'. \tag{2.3}$$

The negative sign of b in the first relation is motivated by the need for correspondence with (2.1). We note that we no longer assume time to be the same in both frames.

We calibrate time for the two frames by setting $t = t' = 0$ for the moment when the origins of the two frames coincide. If a light signal is emitted from the common origin at this moment, it will travel a distance

$$x = ct \text{ and } x' = ct' \tag{2.4}$$

in the frames S and S' respectively, according to postulate 2. From our assumed form of the transformations we get by substitution

$$ct = act' + bt'; \quad ct' = act - bt. \tag{2.5}$$

Isolating the ratio of t and t' we get

$$\frac{t}{t'} = a + \frac{b}{c}; \quad \frac{t'}{t} = a - \frac{b}{c} \tag{2.6}$$

and eliminating the time ratios we find that

$$1 = a^2 - \frac{b^2}{c^2}. \tag{2.7}$$

The origin of S' has the coordinate $x' = 0$, and at time t we can write

$$0 = ax - bt \Rightarrow \frac{b}{a} = \frac{x}{t} \tag{2.8}$$

But x/t is just the speed of the origin of S' in S, that is, the relative speed of the frames, or v. Thus

$$\frac{b}{a} = v \Rightarrow b = va \tag{2.9}$$

and substituting this expression into the relation between a and b in (2.7), we get

$$1 = a^2 - \frac{a^2 v^2}{c^2} \Rightarrow a = \frac{1}{\sqrt{1 - v^2/c^2}} \equiv \gamma, \tag{2.10}$$

where we have introduced the *Lorentz factor* γ, a factor that appears frequently in the relativistic transformation equations. We may now easily derive the transformations between our two inertial frames, and get

$$\begin{gathered} x = \gamma(x' + vt'); \quad x' = \gamma(x - vt); \\ t = \gamma\left(t' + \frac{vx'}{c^2}\right); \quad t' = \gamma\left(t - \frac{vx}{c^2}\right). \end{gathered} \tag{2.11}$$

2.3 Transformation of Lengths and Relativistic Invariants

Let us examine what effect this transformation has on the measurement of distances in one frame from the other. As before we assume a one-dimensional case where an inertial frame S' is moving with velocity v relative to a stationary frame S. The distance between two points x'_1 and x'_2 in S' is

$$\ell' = x'_2 - x'_1. \tag{2.12}$$

The same two points are moving with speed v relative to S, where we measure their distance as

$$\ell = x_2 - x_1. \tag{2.13}$$

The two measurements are related by the transformation

$$\ell' = x'_2 - x'_1 = \gamma(x_2 - vt) - \gamma(x_1 - vt) = \gamma(x_2 - x_1) = \gamma\ell. \tag{2.14}$$

Thus, a rod of length ℓ' in the moving frame would be measured as only ℓ'/γ from the stationary frame. Because $\gamma > 1$, we have the apparent paradox that moving objects contract along the direction of motion (the *Lorentz–Fitzgerald contraction*). For "normal" speeds this effect would be small and unobservable for all practical purposes. However, for speeds close to that of light, c, the effect would be noticeable. Thus for a rod of length ℓ' moving at speed $v = c\sqrt{3}/2$, the observer in the stationary frame would measure a rod of length only $\ell'/2$.

If lengths are not conserved between frames, we might want to look for other quantities that are invariant under the transformations in (2.11). We have seen that these transformations mix time and position, and so we would expect that any invariant quantity should involve both variables. We can take a clue from the constancy of the speed of light. From the expression for the distance traveled by a light signal in (2.4), we may write

$$x^2 - c^2t^2 = 0; \quad x'^2 - c^2t'^2 = 0. \tag{2.15}$$

By postulate 2 this quantity is conserved in a transformation between frames for a light signal, or equivalently a photon. What about a particle traveling at a slower speed?

For this we have

$$\begin{aligned} x'^2 - c^2t'^2 &= \gamma^2(x - vt)^2 - \gamma^2c^2\left(t - \frac{vx}{c^2}\right)^2 \\ &= \gamma^2(x^2 - c^2t^2)\left(1 - \frac{v^2}{c^2}\right) \\ &= (x^2 - c^2t^2). \end{aligned} \tag{2.16}$$

Thus, for the one-dimensional frames considered by us, the quantity $x^2 - c^2t^2$ is certainly conserved under the transformations of (2.11). The reader can easily verify by insertion that this is not the case if the Galilean transformations of (2.1) are used instead.

At this point we may easily generalize to the full three-dimensional spatial coordinate space. The complete set of transformation equations, known as the *Lorentz transformation*, is

$$\begin{aligned} x &= \gamma(x' + vt'); \quad x' = \gamma(x - vt) \\ y &= y' \\ z &= z' \\ t &= \gamma\left(t' + \frac{vx'}{c^2}\right); \quad t' = \gamma\left(t - \frac{vx}{c^2}\right). \end{aligned} \tag{2.17}$$

It is easily seen that the inclusion of the two spatial dimensions perpendicular to the direction of relative motion produces no great changes in the invariant above. We may define an *interval* s by

$$s^2 = x'^2 + y'^2 + z'^2 - c^2t'^2 = x^2 + y^2 + z^2 - c^2t^2 \tag{2.18}$$

where s is clearly invariant under the Lorentz transformations. This is quite similar in appearance to a scalar product, and we may indeed introduce a four-vector

$$\mathrm{w} = (x, y, z, ict) \tag{2.19}$$

with length $\mathrm{w} \cdot \mathrm{w} = s^2$. In the future we will encounter other four-vectors where the components transform as x, y, z and ict under these transformations. A list of common four-vectors is given in Appendix A. It may be shown that the scalar product of two four-vectors $\mathrm{a} = (a_1, a_2, a_3, a_4)$ and $\mathrm{b} = (b_1, b_2, b_3, b_4)$,

$$\mathrm{a} \cdot \mathrm{b} = a_1b_1 + a_2b_2 + a_3b_3 - a_4b_4, \tag{2.20}$$

is invariant under Lorentz transformations.[2]

---

[2] Many texts use a more complicated notation involving a metric and covariant and contravariant vectors, and the sign of the scalar product is the opposite of the one given here. The more complicated notation is not necessary for the developments in this book.

For arbitrary relative motion of the two particles, we will later need a more general form of the Lorentz transformation for the position vector $\mathbf{r}$. According to this form the position vector $\mathbf{r}'$ in the moving frame is expressed as

$$\mathbf{r}' = \gamma(\mathbf{r} - \mathbf{v}t) + (\gamma - 1)\frac{\mathbf{v} \times (\mathbf{v} \times \mathbf{r})}{v^2} \tag{2.21}$$

where $\mathbf{v}$ is the relative velocity of the two frames. The first term is immediately recognizable from the expressions for the Lorentz transformation in (2.17) and describes the translational motion. The second term arises from nontranslational motion, as indicated by the fact that any component of $\mathbf{v}$ along $\mathbf{r}$ vanishes in the cross product $\mathbf{v} \times \mathbf{r}$.

2.4 Transformation of Velocities

In the previous section we have shown that the Lorentz transformations also affect time. We will therefore have to abandon our simple one-dimensional model when we turn to velocity transformations. Even velocities perpendicular to the relative motion of the two frames are affected by the time transformation. This is in contrast to the Galilean transformations for velocities, which take the simple form

$$\begin{aligned} u_x' &= \frac{\mathrm{d}x'}{\mathrm{d}t} = u_x - v; \quad u_x = u_x' + v; \\ u_y' &= u_y; \\ u_z' &= u_z. \end{aligned} \tag{2.22}$$

If we retain the x axis as the direction of relative motion of the frames S and S', the velocity components in this direction are

$$u_x = \frac{\mathrm{d}x}{\mathrm{d}t}; \quad u_x' = \frac{\mathrm{d}x'}{\mathrm{d}t'}. \tag{2.23}$$

We use the Lorentz transformations to find the differentials:

$$\begin{aligned} x' &= \gamma(x - vt); \quad \mathrm{d}x' = \gamma(\mathrm{d}x - v\mathrm{d}t), \\ t' &= \gamma\left(t - \frac{vx}{c^2}\right); \quad \mathrm{d}t' = \gamma\left(\mathrm{d}t - \frac{v\mathrm{d}x}{c^2}\right). \end{aligned} \tag{2.24}$$

Substituting these into the velocity expressions, we get

$$u_x' = \frac{\gamma(\mathrm{d}x - v\mathrm{d}t)}{\gamma(\mathrm{d}t - v\mathrm{d}x/c^2)} = \frac{u_x - v}{1 - vu_x/c^2}. \tag{2.25}$$

Likewise for frame S, we get

$$u_x = \frac{u'_x + v}{1 + vu'_x/c^2} \tag{2.26}$$

For the transverse velocities

$$u_y = \frac{\mathrm{d}y}{\mathrm{d}t}; \quad u'_y = \frac{\mathrm{d}y'}{\mathrm{d}t'} \tag{2.27}$$

and from the Lorentz transformations we find

$$u'_y = \frac{u_y}{\gamma(1 - vu_x/c^2)} \tag{2.28}$$

and

$$u_y = \frac{u'_y}{\gamma(1 + vu'_x/c^2)} \tag{2.29}$$

These velocity transformations have a number of interesting implications. One of these is that the usual rules for the addition of velocities do not hold. From the Galilean transformations it is easy to verify that

$$u_x = u'_x + v \tag{2.30}$$

If we express velocities as fractions of c, that is,

$$u_x = pc, \ v = qc; \quad p, q < 1, \tag{2.31}$$

then the Galilean transformations would yield

$$u_x = (p + q)c, \tag{2.32}$$

which would lead to objects or signals moving faster than the speed of light if $p + q > 1$. Setting $p = 0.8$ and $q = 0.5$, S' would move with speed $0.5c$ relative to S, and a particle with speed $0.8c$ in frame S' would then according to the Galilean transformations move with speed $1.3c$ relative to S. The Lorentz transformations, however, yield

$$u_x = \frac{u'_x + v}{(1 + vu'_x/c^2)} = \frac{(p + q)c}{1 + pq} < c. \tag{2.33}$$

Thus objects moving slower than the speed of light in one inertial frame also move slower than c in all other inertial frames.

2.5 Transformation of Mass

We have now discussed how the description of coordinates and velocities is affected by the postulates of special relativity. To develop a relativistic dynamics, we also need to account for any relativistic effects on the mass of the particles involved. We imagine a simple collision experiment in the frame S': A particle with mass m' moves along the y' direction, that is, perpendicular to the direction of motion of S' relative to S. The particle undergoes a totally elastic collision with a wall in the $x'z'$ plane and rebounds in the $-y'$ direction. If the speed along the y' axis before the collision was u'_y, the total change of momentum is

$$\Delta p' = 2m'u'_y. \tag{2.34}$$

We assume that the speed u_y can be made arbitrarily small, such that ordinary nonrelativistic mechanics applies to the collision. (This assumption may appear questionable, but more refined arguments produce the same result.)

Now, consider the process viewed from S. Before the collision, the y-component of the particle velocity is

$$u_y = \frac{u'_y}{\gamma(1 + vu'_x/c^2)} = \frac{u'_y}{\gamma} \tag{2.35}$$

because the only nonzero component of velocity in S' is along the y' axis, and therefore $u'_x = 0$. Observed from the S frame the total change in momentum is therefore

$$\Delta p = \frac{2mu'_y}{\gamma}. \tag{2.36}$$

For the change of momentum to be the same in the two frames, we must have

$$\Delta p = \Delta p' \Rightarrow 2m'u'_y = \frac{2mu'_y}{\gamma} \tag{2.37}$$

and therefore

$$m = \gamma m'. \tag{2.38}$$

Here m' is the mass of the particle measured in the S' frame where conditions are nonrelativistic. Thus, normal Newtonian mechanics apply, and the mass m' would be the same as we would measure for the particle at rest. We call this the rest mass, and write it conventionally as m_0. In order to describe the same process from the frame S, we must use a mass

$$m = \gamma m_0 = \frac{m_0}{\sqrt{1 - v^2/c^2}} \tag{2.39}$$

indicating that mass increases with velocity and in fact would become infinite as the particle approaches the speed of light.

The alert reader will have noticed that we based this derivation on the assumption that the change of momentum should be the same in the two systems. We have made no justification of this, and indeed with lengths, velocities, and masses all being affected by special relativity there is every reason to question this assumption. However, the same result could have been derived from a similar, but slightly more complicated experiment which ensures symmetry between the situations in S and S', including the equality of change in momentum in the two systems. We will not do this in detail here, but descriptions of such experiments may be found in various textbooks (e.g., French 1968).

2.6 Relativistic Energy

One of the main pillars of classical Newtonian dynamics is the relation between force and acceleration in Newton's second law:

$$F = \frac{\mathrm{d}}{\mathrm{d}t}(mu) = ma. \tag{2.40}$$

We know that mass and velocities, and therefore also acceleration, are affected by special relativity, and may suspect that this law also must be modified. Nevertheless, the work done on a system by a constant force acting over a small distance may still be expressed as

$$F\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}t}(mu)\mathrm{d}x. \tag{2.41}$$

Assume that this force is applied to a particle at rest with rest mass m_0. If this work leads to acceleration of the particle to a final speed of u_f, the work done on the system should be

$$T = \int_0^{u_f} F\mathrm{d}x. \tag{2.42}$$

The appropriate indefinite integral is

$$\int F\mathrm{d}x = \int \frac{\mathrm{d}}{\mathrm{d}t}(mu)\mathrm{d}x = \int \frac{\mathrm{d}x}{\mathrm{d}t}\mathrm{d}(mu) = \int mu\mathrm{d}u + u^2\mathrm{d}m. \tag{2.43}$$

The mass and the velocity of the particle are related as shown in the previous section

$$m = \frac{m_0}{\sqrt{1 - u^2/c^2}}. \tag{2.44}$$

Squaring this expression and applying it to the particle moving at speed u, we find

$$m^2 - \frac{m^2u^2}{c^2} = m_0^2 \tag{2.45}$$

and therefore

$$u^2 = c^2\left(1 - \frac{m_0^2}{m^2}\right). \tag{2.46}$$

From this expression we can derive the required differentials for the integration above:

$$2u\mathrm{d}u = c^2\left(2\frac{m_0^2}{m^3}\mathrm{d}m\right) \Rightarrow u\mathrm{d}u = \frac{c^2 m_0^2}{m^3}\mathrm{d}m. \tag{2.47}$$

This enables us to eliminate u from the indefinite integral:

$$\int mu\mathrm{d}u + u^2\mathrm{d}m = \int \frac{c^2 m_0^2}{m^2}\mathrm{d}m + c^2\mathrm{d}m - \frac{c^2 m_0^2}{m^2}\mathrm{d}m = \int c^2\mathrm{d}m. \tag{2.48}$$

Initially, the particle is at rest with mass m_0, and it is accelerated to a velocity u_f with a corresponding mass m_f. The work done on the particle is therefore

$$T = \int_{m_0}^{m_f} c^2\mathrm{d}m = m_f c^2 - m_0 c^2. \tag{2.49}$$

If we assume that energy conservation is valid, then this work must appear as increased energy of the system. And if the work results in frictionless motion of an undeformable particle, this work must appear as kinetic energy. If we interpret $m_0 c^2$ as a *rest energy*, that is, an energy that the particle possesses due to its rest mass, then the total energy of the particle is

$$E = mc^2 = T + m_0 c^2 \tag{2.50}$$

with $m_f \equiv m$.

2.7 Relativistic Momentum

The usual linear momentum may be written

$$\mathbf{p} = m\mathbf{u}. \tag{2.51}$$

Squaring this and using the expression for u^2 from (2.46) we get

$$p^2 = m^2u^2 = m^2c^2 - m_0^2c^2. \tag{2.52}$$

If we multiply both sides of the equation by c^2, we get the following relation between momentum and energy:

$$p^2c^2 = m^2c^4 - m_0^2c^4 = E^2 - m_0^2c^4. \tag{2.53}$$

For the particular case of particles with no rest mass, such as photons, we get

$$E^2 = p^2c^2. \tag{2.54}$$

From our experience this far with vector lengths and velocities, we do not expect the magnitude of ordinary linear momentum to be invariant under the Lorentz transformations. By analogy with our previous derivation of the four-vector, we can take a cue from the relations for light signals. For photons we know that the relation

$$p^2 - \frac{E^2}{c^2} = 0 \tag{2.55}$$

must hold for all frames. For a particle of nonzero mass this quantity is

$$p^2 - \frac{E^2}{c^2} = p^2 - m^2c^2 = -m_0^2c^2 \tag{2.56}$$

which also must be invariant regardless of which frame it is referred to. We can therefore define a momentum four-vector of the form

$$\mathrm{p} = \left(p_x, p_y, p_z, \frac{iE}{c}\right) = (p_x, p_y, p_z, imc). \tag{2.57}$$

The length of this vector, defined by the scalar product

$$\mathrm{p} \cdot \mathrm{p} = -m_0^2c^2, \tag{2.58}$$

is invariant under the Lorentz transformations, as the reader can easily verify—for example, by inserting (2.39), (2.46), and (2.51) into (2.57) . A similar exercise using the Galilean velocity transformations of (2.22) and a constant mass of m_0 yields a quantity that is dependent on the velocity of the moving frame, and is therefore not invariant.

We may use the definition of the momentum to derive an expression for the energy, which we will use later. With the help of (2.46) we can write the scalar product of **u** and **p** as

$$\mathbf{u} \cdot \mathbf{p} = mu^2 = mc^2 - m_0c^2\left(\frac{m_0}{m}\right). \tag{2.59}$$

With $E = mc^2$ and (2.39) we can rearrange this expression into an equation for the energy,

$$E = \mathbf{u} \cdot \mathbf{p} + m_0c^2/\gamma. \tag{2.60}$$

3

Relativistic Electromagnetic Interactions

Chemical concepts are conveniently formulated in terms of molecules—aggregates of atoms linked by electromagnetic interactions. The proper relativistic description of these interactions is a prerequisite for the development of a theory of relativistic quantum chemistry. As a simple starting point we will consider classical systems made up of point charges, postponing the transition to a quantum mechanical description until later. From the previous chapter we know something about how an electron's particle properties might be affected by relativity. In this chapter we describe the effects of relativity on the interaction with the electromagnetic field.

Again, we adopt a minimalist approach. Electromagnetism and electrodynamics are subjects covered in numerous textbooks for a wide variety of target audiences. To develop the necessary theory from first principles is far beyond the scope of this book. We will only highlight those parts necessary for the later development and understanding of a theory of relativistic quantum chemistry. This means that some of the fundamental equations must be presented without derivation, requiring that the reader either knows these from before or that they must be taken on faith. In particular, in this chapter we make use of the Maxwell equations, the Lorentz force equation, and the generalized potential. The reader will be able to find descriptions or derivations of these in Jackson (1975), for example. We will also need to use a number of relations from vector calculus, and these will normally be introduced in the general form when required.

In dealing with fields that vary over time and space, we will need various differential operators. In the nonrelativistic theory of electrodynamics the gradient operator, ∇, and the time derivative, d/dt, are used. From our experience in the previous chapter with mixing of space and time coordinates under Lorentz transformations, we might expect these to combine in a four-space differential operator also. Indeed, in our notation, the appropriate gradient operator to use for four-vectors is $\Box$, defined by

$$\Box = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}, \frac{-i}{c}\frac{\partial}{\partial t}\right) = \left(\nabla, \frac{-i}{c}\frac{\partial}{\partial t}\right) \tag{3.1}$$

and termed *quabla*, in analogy to nabla.[1] This equation is rather straightforwardly derived by requiring Lorentz invariance and applying the transformation equations (2.17). The interested reader may make things simpler by working out the case of one spatial dimension as done in section 2.3.

The four-vector analogue of the Laplacian ∇^2 is the *d'Alembertian* $\Box^2$ and is correspondingly defined as

$$\Box^2 = \left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right). \tag{3.2}$$

This operator is the square of a Lorentz-invariant four-vector, and is therefore itself Lorentz invariant.

3.1 The Maxwell Equations

The basic equations of electrodynamics are the Maxwell equations. For point charges in a vacuum, which is what we are mainly interested in, these equations take the form (in SI units):

$$\begin{aligned} \nabla \cdot \mathbf{B} &= 0 \\ \nabla \times \mathbf{E} &= -\left(\frac{\partial \mathbf{B}}{\partial t}\right) \\ \nabla \cdot \mathbf{E} &= \rho/\epsilon_0 \\ \nabla \times \mathbf{B} &= \frac{1}{c^2}\left[\frac{\mathbf{j}}{\epsilon_0} + \frac{\partial \mathbf{E}}{\partial t}\right]. \end{aligned} \tag{3.3}$$

Here, $\mathbf{B}$ is the magnetic field, $\mathbf{E}$ is the electric field, ρ is the charge density, $\mathbf{j}$ is the current density, and ϵ_0 is the permittivity of the vacuum. The Maxwell equations are invariant under Lorentz transformations. In fact, it was the search for a transformation that would leave the Maxwell equations formally invariant that originally led to the Lorentz transformations.

Current and charge densities are related through the *continuity equation*

$$\nabla \cdot \mathbf{j} + \frac{\partial \rho}{\partial t} = 0 \tag{3.4}$$

---

[1] This is not the conventional definition; $\Box$ is often used for the d'Alembertian. The current definition is appropriate for our purposes, in which we need a notation for both the four-vector and its square.

which follows from the Maxwell equations (3.3) by taking the time derivative of the third and the divergence of the fourth and combining them in a suitable manner. The magnetic field disappears during this operation thanks to the relation

$$\nabla \cdot (\nabla \times \mathbf{B}) = 0 \tag{3.5}$$

which holds for any vector,[2] not just **B**.

A closer examination of the continuity equation reveals that this is just the scalar product of the gradient for the four-vector space from (3.1) and a vector

$$\mathrm{j} = (\mathbf{j}, ic\rho) = (j_x, j_y, j_z, ic\rho) \tag{3.6}$$

and the continuity equation can be recast in the form

$$\Box \cdot \mathrm{j} = 0. \tag{3.7}$$

This equation must be Lorentz invariant because it was derived from the invariant Maxwell equations. Thus the product of the four-vector gradient and the j-vector is invariant, and therefore j is another Lorentz-invariant four-vector, the four-current.

3.2 Potentials and Gauge Transformations

For a number of purposes it is more convenient to work in terms of field-related potentials rather than the fields **B** and **E** themselves. To find these potentials, we start with the two homogeneous Maxwell equations. The simplest of these is the equation in **B**:

$$\nabla \cdot \mathbf{B} = 0. \tag{3.8}$$

Using the relation in (3.5), we see that the magnetic field may be expressed as the curl of another vector **A**:

$$\mathbf{B} = \nabla \times \mathbf{A}. \tag{3.9}$$

With this choice for **B**, (3.8) is automatically satisfied, and the corresponding properties of the field are built into any further theoretical developments.

We now turn to the homogeneous equation in **E**:

$$\nabla \times \mathbf{E} + \left(\frac{\partial \mathbf{B}}{\partial t}\right) = 0. \tag{3.10}$$

---

[2] A list of vector relations can be found in appendix B.

If we substitute (3.9) into this equation, we get

$$\nabla \times \mathbf{E} + \frac{\partial}{\partial t} (\nabla \times \mathbf{A}) = 0 \tag{3.11}$$

or

$$\nabla \times \left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} \right) = 0. \tag{3.12}$$

This time we can use another orthogonality relation involving differential operators. For any scalar s we have

$$\nabla \times (\nabla s) = 0. \tag{3.13}$$

Using this relation, we can introduce a scalar $-\phi$ defined by

$$\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} = -\nabla \phi \qquad \Rightarrow \qquad \mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} - \nabla \phi, \tag{3.14}$$

so that the homogeneous equation in **E**, (3.12), is satisfied. The choice of a negative scalar $-\phi$ turns out to be convenient in later work.

The quantities **A** and ϕ are referred to as the *vector potential* and *scalar potential*, respectively. The definition of these quantities contains some ambiguity. Making use of (3.13), we see that we can always add any gradient of a scalar quantity to **A** without changing **B** . Thus, if Λ is any scalar,

$$\mathbf{A}' = \mathbf{A} + \nabla \Lambda \qquad \Rightarrow \qquad \nabla \times \mathbf{A}' = \nabla \times \mathbf{A} + \nabla \times \nabla \Lambda = \nabla \times \mathbf{A} = \mathbf{B}. \tag{3.15}$$

If we modify **A** we will, of course, also have to change ϕ. The new scalar potential ϕ' is given by

$$-\nabla \phi' = \mathbf{E} + \frac{\partial \mathbf{A}'}{\partial t} = \mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} + \frac{\partial \nabla \Lambda}{\partial t} = -\nabla \phi + \nabla \frac{\partial \Lambda}{\partial t} \tag{3.16}$$

and therefore

$$\phi' = \phi - \frac{\partial \Lambda}{\partial t}. \tag{3.17}$$

The particular choice of **A** is referred to as a *gauge*, and this freedom in choice of gauge can be exploited in transforming expressions to forms that are convenient for further manipulations, as we will see below. It should be emphasized that the choice of gauge only influences the form of the mathematical expressions. The physics must remain the same, regardless of gauge.

We can also add a constant to ϕ without changing the electric field **E**. This choice of scalar defines an *electric* gauge, and is used to fix the zero of energy at infinite charge separation, for example. The choice of Λ could be said to constitute a *magnetic* gauge. In both cases, the physics must be independent of the gauge.

We can now use the two remaining (inhomogeneous) Maxwell equations to derive equations for **A** and ϕ. We start with the equation for **E**:

$$\nabla \cdot \mathbf{E} = \rho/\epsilon_0. \tag{3.18}$$

If we insert the expression for **E** derived above, we get

$$\begin{aligned} &\nabla \cdot \left(-\frac{\partial \mathbf{A}}{\partial t} - \nabla\phi \right) = \rho/\epsilon_0 \\ \Rightarrow \quad &\nabla^2\phi + \frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}) = -\rho/\epsilon_0. \end{aligned} \tag{3.19}$$

The equation in **B** is

$$\nabla \times \mathbf{B} - \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t} = \frac{1}{c^2}\frac{\mathbf{j}}{\epsilon_0}. \tag{3.20}$$

To manipulate this equation, we need the vector operator expression

$$\nabla \times (\nabla \times \mathbf{a}) = \nabla(\nabla \cdot \mathbf{a}) - \nabla^2\mathbf{a} \tag{3.21}$$

where **a** is any vector. Substituting for **B** and **E** in (3.20), we get

$$\begin{aligned} &\nabla \times (\nabla \times \mathbf{A}) + \frac{1}{c^2}\frac{\partial}{\partial t}\left(\frac{\partial \mathbf{A}}{\partial t} + \nabla\phi \right) = \frac{1}{c^2}\frac{\mathbf{j}}{\epsilon_0} \\ \Rightarrow \quad &\nabla^2\mathbf{A} - \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} - \nabla\left(\nabla \cdot \mathbf{A} + \frac{1}{c^2}\frac{\partial \phi}{\partial t} \right) = -\frac{1}{c^2}\frac{\mathbf{j}}{\epsilon_0}. \end{aligned} \tag{3.22}$$

This equation together with (3.19) provides us with two equations for **A** and ϕ. Unfortunately, the equations are coupled; that is, both equations contain both variables. It would be simpler if we could have just one variable for each equation. This is where the freedom of gauge comes in handy. We can use this freedom to choose a gauge that decouples the equations. We can eliminate ϕ from (3.22) by setting the last parenthesis on the left side to 0,

$$\nabla \cdot \mathbf{A} + \frac{1}{c^2}\frac{\partial \phi}{\partial t} = 0. \tag{3.23}$$

This condition is called the *Lorentz condition* and defines the *Lorentz gauge*. Using this condition in (3.19) provides us finally with a set of decoupled equations:

$$\begin{aligned} \nabla^2\phi - \frac{1}{c^2}\frac{\partial^2 \phi}{\partial t^2} &= -\rho/\epsilon_0, \\ \nabla^2\mathbf{A} - \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} &= -\frac{1}{c^2}\frac{\mathbf{j}}{\epsilon_0}. \end{aligned} \tag{3.24}$$

A closer examination of these equations reveals that they may be given a particularly simple form using the d'Alembertian $\Box^2$:

$$\begin{aligned}\Box^2\phi &= -\rho/\epsilon_0,\\ \Box^2\mathbf{A} &= -\frac{1}{c^2}\frac{\mathbf{j}}{\epsilon_0}.\end{aligned} \tag{3.25}$$

In this form, we see that if we multiply the top equation by i/c and add the two equations, we recover the four-vector current density j from (3.6) on the right-hand side. If we define a four-vector potential A accordingly, we get

$$\Box^2\mathrm{A} = \Box^2\left(\mathbf{A}, \frac{i\phi}{c}\right) = -\frac{1}{c^2\epsilon_0}\mathrm{j}. \tag{3.26}$$

Because $\Box^2$ and j are invariant, it follows that the four-vector potential A is also Lorentz invariant. Note that this is really implied in the Lorentz condition (3.23), which is a product of the Lorentz-invariant four-gradient $\Box$ and the four-vector A and can be written

$$\Box\cdot\mathrm{A} = 0. \tag{3.27}$$

The theory of gauge transformations is quite extensive, and we will return to some of it later. However, we close this section with a few remarks about another frequently used gauge. Using the first of the coupled equations for **A** and ϕ, (3.19), we can define another gauge that yields an equation for ϕ only, by setting

$$\nabla\cdot\mathbf{A} = 0. \tag{3.28}$$

This is called the *Coulomb gauge*, and the resulting equation (Poisson's equation),

$$\nabla^2\phi = -\rho/\epsilon_0, \tag{3.29}$$

has the ordinary Coulomb potential from a charge distribution ρ as its solution. For most of our purposes, we will use the Coulomb gauge, and we note that for time-independent scalar potentials ϕ we have

$$\frac{\partial\phi}{\partial t} = 0, \tag{3.30}$$

and the Lorentz gauge becomes identical to the Coulomb gauge for such potentials.

3.3 The Relativistic Potential from a Moving Charge

In developing a relativistic quantum chemistry we will need to describe a system of interacting particles, all of which have charge, and most of which are assumed to be pointlike. In order to set up a Hamiltonian, we need to know the potential energy of the system or any relativistic equivalent of this. In particular, we will need expressions for the interaction of two moving pointlike particles 1 and 2 with charges q_1 and q_2, respectively. To arrive at this interaction we proceed in three steps: first we derive the potential set up by a moving charge; then we find the expression for the potential experienced by a moving charge in an electromagnetic field; and finally we build on these results to obtain the potential for the relativistic interaction of two moving charged particles.

We start with the potential set up by a moving charge. Having established that $\mathrm{A} = \left(\mathbf{A}, \frac{i}{c}\phi\right)$ is a four-vector, we expect it to transform in analogy with the position four-vector, and the Lorentz transformation of (2.17) should apply if we replace $\mathbf{r}$ with $\mathbf{A}$ and t with ϕ/c^2. More specifically—if S is the stationary frame and S' is moving along the x axis with velocity v relative to S—we have the transformation equations

$$\begin{aligned} &A_x = \gamma\left(A'_x + \frac{v\phi'}{c^2}\right); \qquad A'_x = \gamma\left(A_x - \frac{v\phi}{c^2}\right) \\ &A_y = A'_y; \qquad A_z = A'_z \\ &\phi = \gamma(\phi' + vA'_x); \qquad \phi' = \gamma(\phi - vA_x). \end{aligned} \tag{3.31}$$

From this we can deduce the relativistic potential set up by a charge q moving at velocity $\mathbf{u}$. We let S be the observer frame, and place the x axis in the direction of $\mathbf{u}$. In a charge-centered frame S', the potential is the same as from a stationary charge, that is, the ordinary Coulomb potential

$$\phi' = \frac{q}{4\pi\epsilon_0 r'}, \tag{3.32}$$

and in the absence of external magnetic fields the vector potential vanishes,

$$\mathbf{A}' = 0. \tag{3.33}$$

From the Lorentz transformations we get the potential in the stationary frame S:

$$\begin{aligned} &A_x = \frac{\gamma u\phi'}{c^2}; \qquad \phi = \gamma\phi' \\ &A_y = 0; \qquad A_z = 0. \end{aligned} \tag{3.34}$$

The problem here is that while we now have expressions for the potential in a stationary frame, these are given in terms of variables referring to the moving frame. In addition we know that the "signal" mediating the interaction from the charge to the observer travels with a definite speed c, and this has to be taken into account. Let us fix the time scales such that the two frames coincide at $t = t' = 0$ and assume that at this time we measure the potential a distance r' away from the charge in the moving frame.

This signal has traveled some distance before reaching our measuring apparatus, and was therefore emitted at

$$t' = -r'/c. \tag{3.35}$$

Expressed in terms of the variables of the stationary frame, we must have

$$t' = \gamma\left(t - \mathbf{u}\cdot\mathbf{r}/c^2\right). \tag{3.36}$$

Here we use a slightly more general form of the transformation with $\mathbf{u}\cdot\mathbf{r}$ to allow for arbitrary $\mathbf{r}$ directions, not only on the x axis. We can now calculate r':

$$r' = -ct' = -\gamma\,(tc - \mathbf{u}\cdot\mathbf{r}/c). \tag{3.37}$$

In the stationary frame the signal traveled a distance r and was therefore emitted at

$$t = -r/c. \tag{3.38}$$

Thus

$$r' = \gamma\,(r + \mathbf{u}\cdot\mathbf{r}/c), \tag{3.39}$$

and the potentials in the stationary frame are

$$\mathbf{A} = \mathbf{u}\phi/c^2; \qquad \phi = \frac{q}{4\pi\epsilon_0}\,(r + \mathbf{u}\cdot\mathbf{r}/c)^{-1}. \tag{3.40}$$

This potential is called the *Lienard–Weichart* potential, and refers to a signal emitted at a time previous to $t = t' = 0$. Thus, to evaluate it we must use quantities $\mathbf{u}$ and $\mathbf{r}$ determined at the time of emission. Both the velocity and the position may have changed by the time the observation is made, and we characterize the quantities to be used in (3.40) as *retarded* values, that is, relating to a time preceding the time of observation. The finite speed of propagation of the potential, or equivalently the signal from the charge, is a fundamental feature of relativistic electrodynamics. We cannot transmit information faster than the speed of light.

3.4 The Potential Experienced by a Moving Charge

We now move on to the potential experienced by a moving charge in an electromagnetic field. Potentials are related to the force exerted on a particle. For a particle of charge q and velocity $\mathbf{u}$ in an electromagnetic field with electric field strength $\mathbf{E}$ and magnetic field **B**, the force acting on the particle (in SI units) is given by the *Lorentz force equation* as

$$\mathbf{F} = q\,(\mathbf{E} + \mathbf{u} \times \mathbf{B}). \tag{3.41}$$

For a conservative force we have

$$\mathbf{F} = -\nabla V \tag{3.42}$$

where V is the potential energy giving rise to the force. However, the presence of the velocity-dependent term $\mathbf{u} \times \mathbf{B}$ in the Lorentz force equation necessitates the use of generalized potentials fulfilling the equation

$$\mathbf{F} = -\nabla V + \sum_i \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial V}{\partial u_i}\mathbf{e}_i; \qquad i = x, y, z \tag{3.43}$$

where $\mathbf{e}_i$ is a unit vector.

It is possible to derive the potential corresponding to **F** directly from (3.41) (see Goldstein 1950). Here we take a somewhat simpler approach and start with a suggested form of the potential and show that the application of (3.43) to this satisfies the Lorentz force equation. We choose the potential

$$V = q\phi - q\mathbf{u} \cdot \mathbf{A}. \tag{3.44}$$

The expression for the generalized potential, (3.43), has two parts. We first evaluate the part involving the gradient:

$$\frac{1}{q}\nabla V = \nabla\phi - \nabla\,(\mathbf{u} \cdot \mathbf{A}). \tag{3.45}$$

The gradient of a scalar product of two vectors **a** and **b** has the general form

$$\nabla\,(\mathbf{a} \cdot \mathbf{b}) = (\mathbf{a} \cdot \nabla)\,\mathbf{b} + (\mathbf{b} \cdot \nabla)\,\mathbf{a} + \mathbf{a} \times (\nabla \times \mathbf{b}) + \mathbf{b} \times (\nabla \times \mathbf{a}). \tag{3.46}$$

We also note that x_i and u_i are independent canonical variables, and therefore

$$\frac{\partial}{\partial x_i}u_j = 0 \tag{3.47}$$

for any i and j. Using these expressions, we get the gradient of the scalar product between **u** and **A**

$$\nabla\,(\mathbf{u} \cdot \mathbf{A}) = (\mathbf{u} \cdot \nabla)\,\mathbf{A} + \mathbf{u} \times (\nabla \times \mathbf{A}). \tag{3.48}$$

The curl in the last term above is easily recognized from (3.9) as the magnetic field **B**, and the gradient part of the general potential, (3.43), thus yields

$$\frac{1}{q}\nabla V = \nabla\phi - (\mathbf{u}\cdot\nabla)\,\mathbf{A} - \mathbf{u}\times\mathbf{B}. \tag{3.49}$$

To evaluate the time derivative part of the generalized potential, we note that the scalar potential ϕ has no explicit velocity dependence, and that the vector potential **A** cannot depend explicitly on particle velocity. Thus, the only contributions from the suggested potential, (3.44), will be terms of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial\mathbf{A}\cdot\mathbf{u}}{\partial u_i}\right)\mathbf{e}_i = \frac{\mathrm{d}A_i}{\mathrm{d}t} = \left(\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}t}\right)_i. \tag{3.50}$$

With this, the force due to the generalized potential, (3.44), becomes

$$\frac{1}{q}\mathbf{F} = -\nabla\phi + (\mathbf{u}\cdot\nabla)\,\mathbf{A} + \mathbf{u}\times\mathbf{B} - \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}t}. \tag{3.51}$$

But in general, the time derivative of any vector **a** may be written

$$\frac{\mathrm{d}\mathbf{a}}{\mathrm{d}t} = \frac{\partial\mathbf{a}}{\partial t} + (\mathbf{u}\cdot\nabla)\,\mathbf{a}. \tag{3.52}$$

Using this in the expression for the force above, we get

$$\frac{1}{q}\mathbf{F} = -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t} + \mathbf{u}\times\mathbf{B}. \tag{3.53}$$

The first two terms here may be expressed in terms of the electric field **E** using (3.14), and we then finally obtain the force as

$$\mathbf{F} = q(\mathbf{E} + \mathbf{u}\times\mathbf{B}), \tag{3.54}$$

which is indeed the Lorentz force equation. We therefore conclude that (3.44) describes the potential experienced by a particle of charge q and velocity **u** in an electromagnetic field given by **E** and **B**.

3.5 The Interaction of Two Charged Particles

According to the preceding section the potential energy of a particle with charge q_1 and velocity $\mathbf{u}_1$ in an electromagnetic field is

$$V = q_1\phi - q_1\mathbf{u}_1\cdot\mathbf{A} = q_1(\phi - \mathbf{u}_1\cdot\mathbf{A}). \tag{3.55}$$

We now consider the situation where the field is due to another particle with charge q_2 and velocity $\mathbf{u}_2$. According to section 3.3, the potential set up by this particle is

$$\mathbf{A}_2 = \mathbf{u}_2\phi_2/c^2; \qquad \phi_2 = \frac{q_2}{4\pi\epsilon_0}\,(r + \mathbf{u}_2 \cdot \mathbf{r}/c)^{-1} \tag{3.56}$$

where $\mathbf{r}$ is $\mathbf{r}_1 - \mathbf{r}_2$ (remember that the potential from section 3.3 was expressed in terms of distance from the moving charge). This gives an interaction energy of

$$V_{12} = q_1\phi_2\left(1 - \frac{\mathbf{u}_1 \cdot \mathbf{u}_2}{c^2}\right) = \frac{q_1 q_2}{4\pi\epsilon_0}(r + \mathbf{u}_2 \cdot \mathbf{r}/c)^{-1}\left(1 - \frac{\mathbf{u}_1 \cdot \mathbf{u}_2}{c^2}\right). \tag{3.57}$$

The obvious problem with this expression is that the variables relating to particle 2 are retarded quantities, and we really need to know something about the motion of particle 2 before we can use it. Another less obvious problem appears if we carry out the same exercise starting with particle 2 instead of particle 1. Then we get the interaction energy

$$V_{21} = \frac{q_1 q_2}{4\pi\epsilon_0}(r - \mathbf{u}_1 \cdot \mathbf{r}/c)^{-1}\left(1 - \frac{\mathbf{u}_1 \cdot \mathbf{u}_2}{c^2}\right). \tag{3.58}$$

The first parenthesis in this expression, arising from the use of the retarded potential from particle 1, is clearly not in general equal to the corresponding term in the expression for V_{12}. In fact, the forces on the particles will in general be neither opposite in direction nor equal in magnitude in a complete relativistic treatment. However, we foresee possible difficulties in constructing a Hamiltonian from interaction energies that are not unambiguously defined. In particular, we should expect the expressions to be symmetric to the relabeling of particles.

To escape this dilemma we seek refuge in the fact that although we are developing a relativistic theory, the electrons in molecules *mostly* belong to the low-energy regime, moving at speeds considerably lower than the speed of light. In this time scale, the signal transmitting the interaction between two particles will be instantaneous to a very good approximation. Thus, we do not have to use retarded quantities, and the transformation of potentials from one particle frame to another may be done by using the Lorentz transformation directly at the time of observation. While we are mainly concerned with the potentials acting on particle 1 from particle 2, the following derivation is quite general. We will therefore suppress the indices until we return to the interaction term, which we derive by inserting the new expressions for the potentials into (3.55).

For the case at hand the relative motion of the two particles may be arbitrary, and we need the general form of the Lorentz transformation for the position vector $\mathbf{r}$ given in the previous chapter:

$$\mathbf{r}' = \gamma(\mathbf{r} - \mathbf{u}t) + (\gamma - 1)\frac{\mathbf{u} \times (\mathbf{u} \times \mathbf{r})}{u^2}. \tag{3.59}$$

We may cast this in a more convenient form by expanding the triple cross product. For that purpose we use the general expression

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}\,(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}\,(\mathbf{a} \cdot \mathbf{b}) \tag{3.60}$$

(also known as the *back cab* rule), where **a**, **b**, and **c** are any vectors. Applying this to the transformation expression above, we get

$$\mathbf{u} \times (\mathbf{u} \times \mathbf{r}) = \mathbf{u}\,(\mathbf{u} \cdot \mathbf{r}) - \mathbf{r}\,(\mathbf{u} \cdot \mathbf{u}). \tag{3.61}$$

Inserting this expression in the general Lorentz transformation, (3.59), yields

$$\mathbf{r}' = \mathbf{r} + \mathbf{u}\left[(\gamma - 1)\frac{\mathbf{u} \cdot \mathbf{r}}{u^2} - \gamma t\right]. \tag{3.62}$$

If we decide to disregard retardation effects, we are free to fix the time scale, and in particular we can decide that we observe the interaction between the particles at the time $t' = t = 0$. Then, the expression for $\mathbf{r}'$ further simplifies to

$$\mathbf{r}' = \mathbf{r} + \mathbf{u}\,(\gamma - 1)\frac{\mathbf{u} \cdot \mathbf{r}}{u^2}. \tag{3.63}$$

To derive an expression for the potential in the observer frame, we start from (3.34) and

$$\phi = \gamma\phi' = \frac{\gamma q}{4\pi\epsilon_0 r'}. \tag{3.64}$$

Here r' is the length of $\mathbf{r}'$. We have

$$(\mathbf{r}')^2 = \mathbf{r}' \cdot \mathbf{r}' = r^2\left[1 + \left(\frac{\gamma \mathbf{u} \cdot \mathbf{r}}{cr}\right)^2\right]. \tag{3.65}$$

The scalar potential takes the form

$$\phi = \frac{\gamma q}{4\pi\epsilon_0 r}\left[1 + \left(\frac{\gamma \mathbf{u} \cdot \mathbf{r}}{cr}\right)^2\right]^{-1/2}. \tag{3.66}$$

We have already made the assumption that our particles move at speeds well below that of light, that is

$$u/c \ll 1. \tag{3.67}$$

We can use this assumption to develop the expression for the potential as a power series in u/c, retaining only terms of order u^2/c^2. In particular

$$\begin{aligned}\gamma&\left[1+\left(\frac{\gamma\mathbf{u}\cdot\mathbf{r}}{cr}\right)^2\right]^{-1/2}\\&=\left(1-\frac{u^2}{c^2}\right)^{-1/2}\left[1+\left(\frac{\gamma\mathbf{u}\cdot\mathbf{r}}{cr}\right)^2\right]^{-1/2}\\&=1+\frac{1}{2}\frac{u^2}{c^2}-\frac{1}{2}\left(\frac{\mathbf{u}\cdot\mathbf{r}}{cr}\right)^2+\mathcal{O}\left(\frac{u^4}{c^4}\right).\end{aligned}\tag{3.68}$$

Inserting this in (3.66), we get the scalar potential to the desired order as

$$\phi=\frac{q}{4\pi\epsilon_0 r}\left[1+\frac{1}{2}\frac{u^2}{c^2}-\frac{1}{2}\left(\frac{\mathbf{u}\cdot\mathbf{r}}{cr}\right)^2+\mathcal{O}\left(\frac{u^4}{c^4}\right)\right]\tag{3.69}$$

and the vector potential to the same order is

$$\mathbf{A}=\mathbf{u}\phi/c^2=\frac{q}{4\pi\epsilon_0 rc}\left[\frac{\mathbf{u}}{c}+\mathcal{O}\left(\frac{u^3}{c^3}\right)\right].\tag{3.70}$$

If we use these potentials in the expression for the interaction energy, we get

$$\begin{aligned}V_{12}&=q_1\phi_2-q_1\mathbf{u}_1\cdot\mathbf{A}_2\\&=\frac{q_1q_2}{4\pi\epsilon_0 r}\left[1+\frac{1}{2}\frac{u_2^2}{c^2}-\frac{1}{2}\left(\frac{\mathbf{u}_2\cdot\mathbf{r}}{cr}\right)^2+\mathcal{O}\left(\frac{u_2^4}{c^4}\right)\right]-\frac{q_1q_2\mathbf{u}_1\cdot\mathbf{u}_2}{4\pi\epsilon_0 rc^2}\end{aligned}\tag{3.71}$$

where we have temporarily restored the particle indices. As we see, this is still not symmetric in the particle labels. However, these potentials are expressed in the Lorentz gauge, (3.23), which may be verified by showing that the potentials satisfy the gauge condition. From nonrelativistic theory we know that the Coulomb interaction provides us with an interaction energy that is symmetric in the particle labels. If we transform to the Coulomb gauge, the scalar potential will just give us the symmetric Coulomb term, which should be dominant, and the vector potential may in the end turn out to either be symmetric or require further manipulation.

To transform the Lorentz gauge scalar potential ϕ^{L} to a Coulomb potential ϕ^{C}, we must add a gauge function Λ as indicated in (3.15) such that

$$\phi^{\mathrm{C}}=\phi^{\mathrm{L}}-\frac{\partial\Lambda}{\partial t}.\tag{3.72}$$

We could find the gauge transformation by applying the Coulomb condition (3.28). But in this case we know that we want to retain just the Coulomb part of the potential, that is

$$\phi^{\mathrm{C}} = \frac{q}{4\pi\epsilon_0 r}. \tag{3.73}$$

For simplicity, we set

$$\Lambda = \frac{q}{4\pi\epsilon_0}\lambda. \tag{3.74}$$

Then, λ must satisfy the equation

$$\frac{q}{4\pi\epsilon_0 r} = \frac{q}{4\pi\epsilon_0 r}\left[1 + \frac{1}{2}\frac{u^2}{c^2} - \frac{1}{2}\left(\frac{\mathbf{u}\cdot\mathbf{r}}{cr}\right)^2 + \mathcal{O}\left(\frac{u^4}{c^4}\right)\right] - \frac{q}{4\pi\epsilon_0}\frac{\partial\lambda}{\partial t}, \tag{3.75}$$

which (dropping higher order terms in c^{-1}) readily reduces to

$$\frac{\partial\lambda}{\partial t} = \frac{1}{2c^2}\left[\frac{u^2}{r} - \frac{(\mathbf{u}\cdot\mathbf{r})^2}{r^3}\right]. \tag{3.76}$$

To solve this equation, we make the assumption that changes in u with t are small, in other words u is approximately constant. This may appear somewhat unmotivated, but we should recall that our theories will be applied to time-independent stationary states of systems, which moreover are reasonably well described in a mean field model, and so this may after all not be a very drastic assumption. Anyway, it allows us to find a simple solution for λ:

$$\lambda = -\frac{\mathbf{u}\cdot\mathbf{r}}{2c^2 r} = -\frac{\mathbf{u}\cdot\mathbf{n}}{2c^2} \tag{3.77}$$

where $\mathbf{n}$ is $\mathbf{r}/r$, the unit vector along $\mathbf{r}$. The minus sign appears because $\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$ and we are dealing with the potential from particle 2, therefore $\mathrm{d}\mathbf{r}/\mathrm{d}t = -\mathbf{u}_2 \equiv -\mathbf{u}$. The reader may verify that this solution satisfies (3.76), assuming $\mathrm{d}\mathbf{u}/\mathrm{d}t = 0$, with $\mathrm{d}(1/r)/\mathrm{d}t = \mathbf{u}\cdot\mathbf{r}/r^3$. The vector potential is transformed according to

$$\mathbf{A}^{\mathrm{C}} = \mathbf{A}^{\mathrm{L}} + \nabla\Lambda = \mathbf{A}^{\mathrm{L}} + \frac{q}{4\pi\epsilon_0}\nabla\lambda. \tag{3.78}$$

To evaluate the gradient term above, we use the general vector relation, (3.46), applied to $\mathbf{u}\cdot\mathbf{n}$,

$$\nabla\,(\mathbf{u}\cdot\mathbf{n}) = (\mathbf{u}\cdot\nabla)\,\mathbf{n} + (\mathbf{n}\cdot\nabla)\,\mathbf{u} + \mathbf{u}\times(\nabla\times\mathbf{n}) + \mathbf{n}\times(\nabla\times\mathbf{u}). \tag{3.79}$$

Any gradient operating on $\mathbf{u}$ gives a null result according to (3.47). Also, for the vector $\mathbf{n}$ we have in general

$$\nabla\times\mathbf{n} = 0; \qquad (\mathbf{a}\cdot\nabla)\,\mathbf{n} = \frac{1}{r}[\mathbf{a} - \mathbf{n}\,(\mathbf{a}\cdot\mathbf{n})] \tag{3.80}$$

for any vector **a**. Thus

$$\nabla(\mathbf{u}\cdot\mathbf{n}) = \frac{\mathbf{u}}{r} - \frac{\mathbf{r}(\mathbf{u}\cdot\mathbf{r})}{r^3}. \tag{3.81}$$

This we can insert into the expression for $\mathbf{A}^{\mathrm{C}}$ to obtain

$$\mathbf{A}^{\mathrm{C}} = \frac{q}{8\pi\epsilon_0 cr}\left[\frac{\mathbf{u}}{c} + \frac{\mathbf{r}(\mathbf{u}\cdot\mathbf{r})}{cr^2} + \mathcal{O}\left(\frac{u^3}{c^3}\right)\right]. \tag{3.82}$$

We have included one c in the outside denominator here in anticipation of the **u** that multiplies the vector potential in the expression for the interaction energy.

We can now omit the superscript C, and restore the particle labels. The potentials from particle 2 are

$$\begin{aligned}\phi_2 &= \frac{q_2}{4\pi\epsilon_0 r}\\ \mathbf{A}_2 &= \frac{q_2}{8\pi\epsilon_0 r}\left[\frac{\mathbf{u}_2}{c^2} + \frac{\mathbf{r}\,(\mathbf{u}_2\cdot\mathbf{r})}{c^2 r^2} + \mathcal{O}\left(\frac{u_2^3}{c^3}\right)\right].\end{aligned} \tag{3.83}$$

Finally, we get the interaction energy by substituting these potentials into (3.55):

$$V_{12} = \frac{q_1 q_2}{4\pi\epsilon_0 r}\left[1 - \frac{1}{2}\frac{\mathbf{u}_1\cdot\mathbf{u}_2}{c^2} - \frac{1}{2}\frac{(\mathbf{u}_1\cdot\mathbf{r})\,(\mathbf{u}_2\cdot\mathbf{r})}{c^2 r^2}\right] + \mathcal{O}\left(\frac{u^4}{c^4}\right). \tag{3.84}$$

The only remaining possible cause for particle asymmetry lies in the definition of **r** as $\mathbf{r}_1 - \mathbf{r}_2$, but from the expression above it is clear that this is of no consequence. In addition to the Coulomb term, the expression for the interaction between two charged particles contains two relativistic terms. These will be discussed in greater detail when we later introduce similar expressions in the relativistic Hamiltonian.

While the expression above for the potential between two charged particles is indeed symmetric in the particle labels, it is not Lorentz invariant. This can be demonstrated by carrying out the appropriate transformation. To make things simple, we consider only the case of two one-dimensional inertial frames, where S' moves with velocity v relative to S. In the moving (one-dimensional) system, $\mathbf{u}'$ and $\mathbf{r}'$ are parallel, and if we discard terms of order 4 and higher in c, the interaction takes the form

$$V'_{12} = \frac{q_1 q_2}{4\pi\epsilon_0 r'}\left[1 - \frac{u'_1 u'_2}{c^2}\right]. \tag{3.85}$$

Applying the Lorentz transformation to r' and u', we get

$$V'_{12} = \frac{q_1 q_2}{4\pi\epsilon_0 \gamma r}\left[1 - \frac{1}{c^2}\frac{(u_1 - v)}{(1 - u_1 v/c^2)}\frac{(u_2 - v)}{(1 - u_2 v/c^2)}\right], \tag{3.86}$$

which expanded to order c^2 yields

$$V'_{12} = \frac{q_1 q_2}{4\pi\epsilon_0 r}\left[1 - \frac{u_1 u_2}{c^2} + \frac{(u_1 + u_2)v}{c^2} - \frac{3v^2}{2c^2}\right]. \tag{3.87}$$

This is clearly different from the expression

$$V_{12} = \frac{q_1 q_2}{4\pi\epsilon_0 r}\left[1 - \frac{u_1 u_2}{c^2}\right] \tag{3.88}$$

obtained by applying (3.84) in the inertial frame S.

Part II

The Dirac Equation: Solutions and Properties

4

The Dirac Equation

The purpose of this chapter is to introduce the Dirac equation, which will provide us with a basis for developing the relativistic quantum mechanics of electronic systems. Thus far we have reviewed some basic features of the classical relativistic theory, which is the foundation of relativistic quantum theory. As in the nonrelativistic case, quantum mechanical equations may be obtained from the classical relativistic particle equations by use of the correspondence principle, where we replace classical variables by operators. Of particular interest are the substitutions

$$\mathcal{H} \to i\hbar\frac{\partial}{\partial t}; \quad \mathbf{p} \to -i\hbar\nabla. \tag{4.1}$$

In terms of the momentum four-vector introduced earlier, this yields

$$\mathrm{p} = \left(p_x, p_y, p_z, \frac{i}{c}\mathcal{H}\right) \to -i\hbar\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}, -\frac{i}{c}\frac{\partial}{\partial t}\right) = -i\hbar\Box. \tag{4.2}$$

In going from a classical relativistic description to relativistic quantum mechanics, we require that the equations obtained are invariant under Lorentz transformations. Other basic requirements, such as gauge invariance, must also apply to the equations of relativistic quantum mechanics.

We start this chapter by reexamining the quantization of the nonrelativistic Hamiltonian and draw out some features that will be useful in the quantization of the relativistic Hamiltonian. We then turn to the Dirac equation and sketch its derivation. We discuss some properties of the equation and its solutions, and show how going to the nonrelativistic limit reduces it to a Schrödinger-type equation containing spin.

4.1 Quantization of the Nonrelativistic Hamiltonian

The nonrelativistic classical Hamiltonian for a free particle is

$$\mathcal{H} = T = \mathbf{p}^2/2m. \tag{4.3}$$

Using the correspondence principle, (4.1), we get the time-dependent Schrödinger equation:

$$-\frac{\hbar^2}{2m}\left[\frac{\partial^2\psi(\mathbf{r},t)}{\partial x^2}+\frac{\partial^2\psi(\mathbf{r},t)}{\partial y^2}+\frac{\partial^2\psi(\mathbf{r},t)}{\partial z^2}\right] = i\hbar\frac{\partial\psi(\mathbf{r},t)}{\partial t}. \tag{4.4}$$

This equation is obviously not Lorentz invariant—it has x, y, and z appearing quadratically but t appearing linearly, which violates the relativistic principle of equivalence of spatial and temporal variables. Since we know that the nonrelativistic classical Hamiltonian is not Lorentz invariant, it is no surprise that neither is the nonrelativistic Schrödinger equation.

In most quantum chemical applications the electrons are not free, but are subject to electrostatic potentials in the form of Coulomb interactions with nuclei and other electrons. These are accounted for in the Hamiltonian by adding a term involving the scalar potential and the charge of the electron ($-e$):

$$\mathcal{H} = T + V = \mathbf{p}^2/2m - e\phi. \tag{4.5}$$

However, if we are concerned about Lorentz invariance, we should at this point remember that the scalar potential is only one component of a four-vector $\mathrm{A} = (\mathbf{A}, i\phi/c)$. If the scalar potential modifies $\mathcal{H}$, or equivalently E, then we would expect the vector potential to modify the momentum, which accounts for the remaining components of the four-vector.

From classical mechanics (e.g., Goldstein 1950), we can show that the presence of a vector potential requires that the Hamiltonian function must be constructed using the *kinetic momentum* (or *mechanical momentum*), which is the momentum that is given in nonrelativistic theory by $m\mathbf{v}$. We must express this momentum in terms of the *canonical momentum* of Lagrangian mechanics, because it is the canonical momentum to which the quantization rule $\mathbf{p} \rightarrow -i\hbar\nabla$ applies. Here (and hereafter) we will use $\mathbf{p}$ for the canonical momentum and $\boldsymbol{\pi}$ for the kinetic momentum. The relation between the two is

$$\boldsymbol{\pi} = \mathbf{p} - q\mathbf{A} \tag{4.6}$$

for a particle of charge q in a vector potential $\mathbf{A}$, and for the electron[1] we must therefore use $\boldsymbol{\pi} = \mathbf{p} + e\mathbf{A}$. This relation is known as *minimal coupling*.

---

[1] Some texts use e for the charge rather than q. However, we will always use e for the elementary charge and develop equations for the electron, whose charge is $-e$.

With this prescription for including potentials, the nonrelativistic classical Hamiltonian becomes

$$\mathcal{H} + e\phi = \boldsymbol{\pi}^2/2m = (\mathbf{p} + e\mathbf{A})^2/2m. \tag{4.7}$$

If we apply the normal quantization procedure to this Hamiltonian, we get (expanding the square)

$$\left[-\frac{\hbar^2}{2m}\nabla^2 - \frac{ie\hbar}{2m}\left[\nabla \cdot \mathbf{A} + \mathbf{A} \cdot \nabla\right] + \frac{e^2}{2m}\mathbf{A}^2 - e\phi\right]\psi(\mathbf{r}, t) = i\hbar\frac{\partial\psi(\mathbf{r}, t)}{\partial t}. \tag{4.8}$$

The term $\nabla \cdot \mathbf{A}$ tells us that here we have a choice of gauge, and we usually choose the Coulomb gauge, $\nabla \cdot \mathbf{A} = 0$, giving

$$\left[-\frac{\hbar^2}{2m}\nabla^2 - \frac{ie\hbar}{m}\mathbf{A} \cdot \nabla + \frac{e^2}{2m}\mathbf{A}^2 - e\phi\right]\psi(\mathbf{r}, t) = i\hbar\frac{\partial\psi(\mathbf{r}, t)}{\partial t}. \tag{4.9}$$

For molecules and atoms the scalar potential ϕ includes the nuclear and electronic Coulomb potentials and is normally incorporated into the variational procedure. In many applications, the terms involving the vector potential are treated as a perturbation, for example to obtain radiative transition probabilities. In this case the term involving $\mathbf{A} \cdot \nabla$ gives the usual one-photon transition probability, and the term involving $\mathbf{A}^2$ gives two-photon probabilities.

4.2 Spin in the Nonrelativistic Hamiltonian

One of the missing features in the usual form of the Schrödinger equation is spin, and it is often stated that this is because it is a relativistic phenomenon. This is at best a tenuous statement, and we will show that there is a method for introducing spin into the nonrelativistic equation.[2] This demonstration also serves to introduce some quantities that will turn up later in a fully relativistic equation that accounts for spin. Among these are the 2 × 2 *Pauli spin matrices*:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{4.10}$$

often simply called the Pauli matrices, which we can collect into a vector, $\boldsymbol{\sigma}$. These matrices are unitary, Hermitian matrices that do not commute:

$$\sigma_x\sigma_y = i\sigma_z; \qquad \sigma_y\sigma_x = -i\sigma_z. \tag{4.11}$$

---

[2] This demonstration was first made by Lévy-Leblond (1967, 1970).

The commutator of σ_x and σ_y is therefore

$$\left[\sigma_x\,,\,\sigma_y\right] = 2i\sigma_z. \tag{4.12}$$

Similar relations for the other products of Pauli matrices may be derived by cyclic permutation of the indices. In fact, the Pauli matrices anticommute:

$$\sigma_i\sigma_j + \sigma_j\sigma_i = 2\delta_{ij}\mathbf{I}_2 \tag{4.13}$$

where $\mathbf{I}_2$ is the 2×2 unit matrix.[3]

We will often be dealing with the scalar product of $\boldsymbol{\sigma}$ with various vectors $\mathbf{u}$; that is, with $(\boldsymbol{\sigma} \cdot \mathbf{u})$. Using the commutation properties of the Pauli matrices, (4.13), it may be shown that

$$(\boldsymbol{\sigma} \cdot \mathbf{u})(\boldsymbol{\sigma} \cdot \mathbf{v}) = (\mathbf{u} \cdot \mathbf{v})\mathbf{I}_2 + i\boldsymbol{\sigma} \cdot \mathbf{u} \times \mathbf{v}, \tag{4.14}$$

where we have assumed that $\mathbf{u}$ commutes with $\boldsymbol{\sigma}$. This relation is called the *Dirac relation* or the *Dirac identity*, and will be applied numerous times in this book. If $\mathbf{u} = \mathbf{v}$, the vector product vanishes, so that

$$(\boldsymbol{\sigma} \cdot \mathbf{u})(\boldsymbol{\sigma} \cdot \mathbf{u}) = u^2\mathbf{I}_2. \tag{4.15}$$

These 2×2 matrices can be regarded as operators on a space of two-component vector functions,

$$\psi^P = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}. \tag{4.16}$$

The two-component wave function ψ^P is called a *Pauli spinor*. We will show in a later chapter that for this basis the Pauli matrices form a representation of the spin operators such that $\hbar\boldsymbol{\sigma} = 2\mathbf{s}$. With the conventional choice of basis, that is, the eigenfunctions of σ_z, the upper component represents the part of the wave function with spin projection $m_s = \frac{1}{2}$, or α spin, and the lower component the part with $m_s = -\frac{1}{2}$, or β spin. The primitive α and β spin functions are represented by the vectors

$$\eta(\tfrac{1}{2}) = \eta(\alpha) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \qquad \eta(-\tfrac{1}{2}) = \eta(\beta) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{4.17}$$

To incorporate spin into the nonrelativistic Hamiltonian we now replace $\boldsymbol{\pi}$ by $\boldsymbol{\sigma} \cdot \boldsymbol{\pi}$ in (4.7). First, we consider the square of $\boldsymbol{\sigma} \cdot \boldsymbol{\pi}$, expanded in terms of the momentum

---

[3] We will indicate the rank of the unit matrix $\mathbf{I}$ and the zero matrix $\mathbf{0}$ by a subscript where we feel it is necessary or useful. However, we will omit it when the context makes it clear what the rank is, or the rank is unspecified.

and the vector potential,

$$(\boldsymbol{\sigma} \cdot \boldsymbol{\pi})^2 = (\boldsymbol{\sigma} \cdot \mathbf{p})^2 + e(\boldsymbol{\sigma} \cdot \mathbf{p})(\boldsymbol{\sigma} \cdot \mathbf{A}) + e(\boldsymbol{\sigma} \cdot \mathbf{A})(\boldsymbol{\sigma} \cdot \mathbf{p}) + e^2(\boldsymbol{\sigma} \cdot \mathbf{A})^2. \tag{4.18}$$

Making use of the relations (4.14), (4.15), and the correspondence principle gives

$$(\boldsymbol{\sigma} \cdot \boldsymbol{\pi})^2 = -\hbar^2\nabla^2 - ie\hbar\big[\nabla \cdot \mathbf{A} + \mathbf{A} \cdot \nabla\big] + e^2\mathbf{A}^2 + e\hbar\boldsymbol{\sigma} \cdot \big[\nabla \times \mathbf{A} + \mathbf{A} \times \nabla\big]. \tag{4.19}$$

The term in the first square bracket is the same as in the equation without spin. The effect of the first term in the second square bracket can be demonstrated by letting it operate on a function ϕ. For the case of an arbitrary vector $\mathbf{a}$, we have the general vector relation

$$\nabla \times (\phi\mathbf{a}) = (\nabla\phi) \times \mathbf{a} + \phi(\nabla \times \mathbf{a}) = -\mathbf{a} \times (\nabla\phi) + \phi(\nabla \times \mathbf{a}). \tag{4.20}$$

If we apply this and insert the definition of the vector potential, $(\nabla \times \mathbf{A}) = \mathbf{B}$, we get

$$\nabla \times \mathbf{A} + \mathbf{A} \times \nabla = (\nabla \times \mathbf{A}) - \mathbf{A} \times \nabla + \mathbf{A} \times \nabla = \mathbf{B}. \tag{4.21}$$

Thus we end up with an extra term in the nonrelativistic Schrödinger equation that accounts for the interaction of the spin of the electron with a magnetic field:

$$\left[-\frac{\hbar^2}{2m}\nabla^2 - \frac{ie\hbar}{m}\mathbf{A} \cdot \nabla + \frac{e^2}{2m}\mathbf{A}^2 + \frac{e\hbar}{2m}\boldsymbol{\sigma} \cdot \mathbf{B} - e\phi\right]\psi^P = i\hbar\frac{\partial\psi^P}{\partial t}. \tag{4.22}$$

The fourth term is the *spin Zeeman* term and represents the interaction of the electron spin with the external magnetic field, and under the substitution $\hbar\boldsymbol{\sigma} = 2\mathbf{s}$ it takes the form $(e/m)\mathbf{s} \cdot \mathbf{B}$.

We have demonstrated that by replacing $\boldsymbol{\pi}$ with $\boldsymbol{\sigma} \cdot \boldsymbol{\pi}$ we can indeed introduce spin into the nonrelativistic Schrödinger equation. In this form, spin appears explicitly in the wave function through the Pauli spinors (or products of these for many-electron systems), and its interaction with magnetic fields appears naturally in the Hamiltonian and need not be grafted on ad hoc when required. However, apart from the fact that it yields a convenient form of the Schrödinger equation, it is not immediately evident why the operator $\boldsymbol{\sigma} \cdot \boldsymbol{\pi}$ should be used. And we still have the problem that the free-electron part of the Hamiltonian is not Lorentz invariant. So we must look for an alternative route to a relativistic quantum theory for the electron, one which preferably also accounts for spin. Our experiences from the derivations in this section show us that this route may lead to multicomponent wave functions.

4.3 The Dirac Equation

We now turn to the quantization of the classical relativistic Hamiltonian, and in particular to its representation in the Dirac equation. Quantization of the classical relativistic Hamiltonian has been treated in detail in many texts, such as Rose (1961), Bjorken and Drell (1964), Moss (1973), and Sakurai (1967), to which the reader is referred for greater detail than is presented here.

In the previous section, we noted that the nonrelativistic Schrödinger equation for the free particle contains second derivatives with respect to the space variables, but only the first derivative with respect to the time, and therefore it does not treat time and space on an equal footing, precluding Lorentz invariance. The second derivatives come from the form of the nonrelativistic Hamiltonian, but the first derivative of the time is fundamental. We therefore seek an equation that is homogeneous and first order in space and time.

For a classical expression of this kind, we turn to the relation between energy and momentum derived in section 2.7. Omitting the subscript for the rest mass of the electron[4] and introducing the mechanical momentum, we may write this relation as

$$(E + e\phi)^2 = m^2c^4 + \boldsymbol{\pi}^2c^2. \tag{4.23}$$

We could go ahead and quantize this expression according to our usual prescription, but this results in four-electron operators, which would be impractical for relativistic quantum chemistry. The squared form does have its place in physics, but it can be shown that it does not describe the electron.

We can reduce the rank of the operators by taking the square root of the expression above, to obtain the Hamiltonian

$$\mathcal{H} = c\sqrt{m^2c^2 + \boldsymbol{\pi}^2} - e\phi. \tag{4.24}$$

There are some obvious difficulties to be encountered in an attempt to quantize this Hamiltonian. If we simply use the correspondence principle, we have the problem of interpreting the square root of the Laplacian operator. We could circumvent this problem by extracting a factor of mc from the square root and then expanding the square root expression as a series in powers of $(\boldsymbol{\pi}/mc)^2$. This yields

$$\mathcal{H} + e\phi = mc^2\sqrt{1 + (\boldsymbol{\pi}/mc)^2} = mc^2 + \frac{\boldsymbol{\pi}^2}{2m} - \frac{\boldsymbol{\pi}^4}{8m^3c^2} + \mathcal{O}\left(\frac{\boldsymbol{\pi}^6}{m^5c^4}\right) \tag{4.25}$$

where the first term is the rest mass energy and the second term is related to the kinetic energy. Unfortunately, this approach fails for several reasons. First, the series is infinite and would have to be truncated. Second, the expansion is only valid if $\pi < mc$, but π can take values up to ∞. Third, the operators become more and more pathological the higher one goes in the series. Finally, the Lorentz invariance is not obvious, and would certainly be lost if the series were truncated, as would be necessary in any practical method.

Dirac's approach to quantization of the Hamiltonian above was to assume that the argument of the square root operator could be written as a perfect square,

$$\pi^2 + m^2c^2 = (\boldsymbol{\alpha} \cdot \boldsymbol{\pi} + \beta mc)^2 \tag{4.26}$$

---

[4] The mass m appearing hereafter is the rest mass, the subscript of m_0 being omitted.

where $\boldsymbol{\alpha}$ and β remain to be specified. Thus, the relativistic Hamiltonian for an electron can be written

$$\mathcal{H} + e\phi = c\boldsymbol{\alpha} \cdot (\mathbf{p} + e\mathbf{A}) + \beta mc^2. \tag{4.27}$$

Using the correspondence principle to quantize this Hamiltonian, we get the Dirac equation for an electron in an electromagnetic field:

$$\left(i\hbar\frac{\partial}{\partial t} + e\phi\right)\Psi = c\boldsymbol{\alpha} \cdot (-i\hbar\nabla + e\mathbf{A})\,\Psi + \beta mc^2\Psi. \tag{4.28}$$

Transferring the term involving $\boldsymbol{\alpha}$ to the left-hand side, this equation may be written in four-vector form as

$$c\mathrm{\alpha} \cdot (i\hbar\Box - e\mathsf{A})\Psi = \beta mc^2\Psi \tag{4.29}$$

where we have introduced the four-vector $\mathrm{\alpha}$

$$\mathrm{\alpha} = (\boldsymbol{\alpha}, i). \tag{4.30}$$

We know that the four-vector momentum and the four-vector potential are conserved quantities under a Lorentz transformation, so we expect that $\mathrm{\alpha}$ is also conserved, as may be shown with the conditions below.

It remains to determine the quantities $\boldsymbol{\alpha}$ and β. In order for the perfect square to be obtained, $\boldsymbol{\alpha}$ and β must obey the following conditions:

$$\alpha_i^2 = \beta^2 = 1, \tag{4.31a}$$

$$\alpha_i\alpha_j = -\alpha_j\alpha_i, \; i \neq j, \tag{4.31b}$$

$$\alpha_i\beta = -\beta\alpha_i, \tag{4.31c}$$

or, with $\beta \equiv \alpha_0$

$$\alpha_i\alpha_j + \alpha_j\alpha_i = 2\delta_{ij}. \tag{4.32}$$

(Note that α_0 is not part of the four-vector $\mathrm{\alpha}$.) The second and third conditions of equation (4.31) mean that the algebra of the α_i is noncommutative—in fact, it is anticommutative. Scalars cannot satisfy these relations, and therefore an appropriate noncommutative algebra must be found to satisfy these conditions. The most obvious choice is matrix algebra.

The first condition, (4.31a), means that the matrices must be unitary. For matrices of rank 2, any unitary matrix can be expressed as a linear combination of the 2 × 2 unit matrix and the Pauli matrices. These four will not do, because even though the Pauli matrices anticommute, the unit matrix commutes with the Pauli matrices, and therefore the anticommutation requirement cannot be met. Thus, the rank of the $\boldsymbol{\alpha}$ and β matrices must be greater than 2. Through some elementary matrix algebra, it may also be shown that the rank of the $\boldsymbol{\alpha}$ and β matrices must be even. From this we conclude that the minimum matrix rank required is 4, and we can indeed find a rank 4 matrix representation of $\boldsymbol{\alpha}$ and β that satisfies the requirements of (4.31). One such representation is

$$\alpha_k = \begin{pmatrix} \mathbf{0}_2 & \sigma_k \\ \sigma_k & \mathbf{0}_2 \end{pmatrix}, \; k = 1, 2, 3; \quad \beta = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & -\mathbf{I}_2 \end{pmatrix}, \tag{4.33}$$

where $\mathbf{I}_2$ and $\mathbf{0}_2$ are the 2×2 unit and null matrices, and σ_k are the Pauli matrices defined above. This representation of the α matrices is called the *standard representation*, and is defined only up to a similarity transformation,

$$\alpha'_\mu = \mathbf{Q}^{-1} \alpha_\mu \mathbf{Q}, \tag{4.34}$$

which for invariance requires that

$$\Psi' = \mathbf{Q}^{-1} \Psi. \tag{4.35}$$

As can be expected, the α_k matrices do not commute. Introducing the vector $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \left[\begin{pmatrix} \sigma_x & \mathbf{0}_2 \\ \mathbf{0}_2 & \sigma_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \mathbf{0}_2 \\ \mathbf{0}_2 & \sigma_y \end{pmatrix}, \begin{pmatrix} \sigma_z & \mathbf{0}_2 \\ \mathbf{0}_2 & \sigma_z \end{pmatrix} \right] \tag{4.36}$$

we can write

$$\alpha_x \alpha_y = i \Sigma_z; \qquad \alpha_y \alpha_x = -i \Sigma_z. \tag{4.37}$$

The commutation relations for the α matrices can be summed up as

$$\boldsymbol{\alpha} \times \boldsymbol{\alpha} = 2i \boldsymbol{\Sigma}. \tag{4.38}$$

While this provides us with an equation for the relativistic electron, the $\boldsymbol{\alpha}$ and β matrices arose from the mathematical treatment, and only indirectly from the physics. It would be nice if we could also give these quantities a physical interpretation. In order to find some classical operator or quantity corresponding to $\boldsymbol{\alpha}$ and β, we compare the Dirac Hamiltonian with the classical relativistic Hamiltonian. For this purpose we use the classical relativistic expression for the energy from (2.60) for a field-free system,

$$E = \mathbf{u} \cdot \mathbf{p} + mc^2/\gamma. \tag{4.39}$$

Comparing with (4.27), we see that we may identify $c\boldsymbol{\alpha}$ with the velocity operator $\mathbf{u}$. With this identification, we see that the velocity four-vector $\mathrm{\alpha}$ can be identified with the classical velocity four-vector, which is the time derivative of the position four-vector w. We can also identify β with the inverse of the γ factor arising in the Lorentz transformations.

We will not prove the Lorentz invariance of the Dirac equation. However, we will point to a few features that at least suggest Lorentz invariance. The first thing to note is that the rest mass is obviously a relativistic invariant. If we multiply (4.29) by β and define $\gamma = \beta\alpha$, we have

$$c\gamma \cdot (\hbar\Box - e\mathrm{A})\Psi = mc^2\Psi. \tag{4.40}$$

It is fairly obvious that if the operator on the left has as its eigenvalue a Lorentz invariant quantity, the operator itself must be Lorentz invariant. This, of course, does not constitute a proof. Such proofs may be found in most standard textbooks in physics that cover the Dirac equation.

What is of further concern is whether the probability density $\Psi^\dagger\Psi$ is time-independent, which we expect for a bound state, and whether it is conserved under a Lorentz transformation, since this has implications for the normalization of the wave function. If $c\boldsymbol{\alpha}$ is the velocity operator, we may write the current density for the Dirac wave function as

$$\mathbf{j} = \Psi^\dagger c\boldsymbol{\alpha}\Psi; \tag{4.41}$$

with the charge density defined by

$$\rho = \Psi^\dagger\Psi. \tag{4.42}$$

Charge is invariant under a Lorentz transformation, and the law of conservation is expressed in the continuity equation

$$\nabla \cdot \mathbf{j} + \frac{\partial\rho}{\partial t} = 0. \tag{4.43}$$

Inserting $\mathbf{j}$ and ρ from the above definitions into the charge conservation equation, we get

$$c\nabla \cdot (\Psi^\dagger\boldsymbol{\alpha}\Psi) + \frac{\partial}{\partial t}\Psi^\dagger\Psi = \left[c\boldsymbol{\alpha} \cdot \nabla\Psi + \frac{\partial\Psi}{\partial t}\right]^\dagger\Psi + \Psi^\dagger\left[c\boldsymbol{\alpha} \cdot \nabla\Psi + \frac{\partial\Psi}{\partial t}\right]. \tag{4.44}$$

From the free-particle Dirac equation,

$$c\boldsymbol{\alpha} \cdot \nabla\Psi + \frac{\partial\Psi}{\partial t} = \frac{mc^2}{i\hbar}\beta\Psi. \tag{4.45}$$

On substitution of this and its conjugate into (4.44) it is clear that the continuity equation is satisfied:

$$\nabla \cdot \mathbf{j} + \frac{\partial\rho}{\partial t} = -\frac{mc^2}{i\hbar}\Psi^\dagger\beta\Psi + \frac{mc^2}{i\hbar}\Psi^\dagger\beta\Psi = 0. \tag{4.46}$$

In order to show that the probability density is time independent, we now take its space integral and find by Gauss's theorem that

$$\frac{\partial}{\partial t}\int \rho \mathrm{d}\mathbf{r} = -\int \nabla \cdot \mathbf{j} \mathrm{d}\mathbf{r} = -\int j_n \mathrm{d}S \tag{4.47}$$

where j_n is the component of the current along the outward normal to the surface S. The surface is of course at infinity, so that if the current density vanishes sufficiently strongly at infinity—which we expect to be the case for a bound, normalizable state—the time derivative of the integral of the probability density is zero, and the integral has a constant value. That this integral is invariant under a Lorentz transformation is shown, for example, in Rose (1961).

4.4 The Time-Independent Dirac Equation

Now that we have a valid equation, we need to manipulate it to obtain a workable form. It is almost invariably the case that we are looking for stationary states, and most considerations of time-dependent phenomena proceed from a set of states that are time-independent.

A stationary state is the eigenfunction of a Hamiltonian that is independent of the time variable and has the energy as its eigenvalue. If we have such a Hamiltonian, we can write the wave function as a product of a spatial and a temporal part,

$$\Psi(\mathbf{r}, t) = \psi(\mathbf{r})\theta(t), \tag{4.48}$$

which on substitution into the time-dependent Dirac equation

$$\hat{\mathcal{H}}\Psi = i\hbar\frac{\partial \Psi}{\partial t} \tag{4.49}$$

yields

$$\hat{\mathcal{H}}\psi(\mathbf{r})/\psi(\mathbf{r}) = i\hbar\frac{\partial \theta}{\partial t}/\theta. \tag{4.50}$$

If $\hat{\mathcal{H}}$ is time-independent, the left side is a function of $\mathbf{r}$ only and the right side a function of t only. For the two to be equal, they must both be equal to a constant, which is the total energy E. The solution of the temporal part is easily obtained as a simple exponential:

$$\theta(t) = e^{Et/i\hbar}. \tag{4.51}$$

The spatial part is the time-independent Dirac equation

$$\hat{\mathcal{H}}\psi = c\boldsymbol{\alpha} \cdot (-i\hbar\nabla + e\mathbf{A})\,\psi + \beta mc^2\psi - e\phi\psi = E\psi, \tag{4.52}$$

which we may now proceed to solve.

In the light of the chapter on special relativity (chapter 2), it is apparent that there is a possible problem in performing this separation of the space and time variables, because the Lorentz transformation mixes them. The separation would have to be performed in a particular frame of reference, and *only be valid in this frame of reference.* If we want results in another frame of reference, we must perform a Lorentz transformation to that frame, and there is no guarantee that we will still have a stationary state. However, if our Hamiltonian is Lorentz invariant, the choice of the frame of reference is arbitrary, and, as we saw above, the probability density is independent of time and of the frame of reference. We may therefore choose the frame that is most convenient. In molecules (and in atoms) the Born–Oppenheimer frame is the most convenient frame of reference for electronic structure calculations because the nuclear potential is then simply the static Coulomb potential. Regardless of whether the Hamiltonian is Lorentz invariant or not, it is this frame that we work in from here on.

4.5 The Dirac Wave Function

Because of the appearance of 4 × 4 matrices in the Dirac equation, the wave function must be a four-component vector:

$$\Psi(\mathbf{r},t) = \begin{pmatrix} \psi_1(\mathbf{r},t) \\ \psi_2(\mathbf{r},t) \\ \psi_3(\mathbf{r},t) \\ \psi_4(\mathbf{r},t) \end{pmatrix}. \tag{4.53}$$

We have therefore moved from a scalar function in the nonrelativistic regime to a four-component vector function in the relativistic regime, and we might expect that the work to describe this wave function would increase in proportion.

What is the significance of the four components? We noted above that in nonrelativistic quantum mechanics we can introduce the spin by replacing the mechanical momentum $\boldsymbol{\pi}$ by $\boldsymbol{\sigma} \cdot \boldsymbol{\pi}$ and making the wave function a two-component vector, or 2-spinor, where the upper component corresponds to spin $\frac{1}{2}$ and the lower component corresponds to spin $-\frac{1}{2}$. The same concept applies to the Dirac wave function—components 1 and 3 correspond to spin $\frac{1}{2}$ and components 2 and 4 correspond to spin $-\frac{1}{2}$, and the wave function is called a 4-spinor.

It is useful to classify the upper two and lower two components of the Dirac wave function as two-component spinors. From the discussion of spin in the nonrelativistic equation, it might be guessed that these would be Pauli spinors. We write

$$\Psi(\mathbf{r},t) = \begin{pmatrix} \Psi^L(\mathbf{r},t) \\ \Psi^S(\mathbf{r},t) \end{pmatrix}; \tag{4.54}$$

Ψ^L and Ψ^S are termed the *large* and *small* components of the wave function. The reason for this nomenclature will become clear later. For the time-independent case, we have correspondingly:[5]

$$\psi(\mathbf{r}) = \begin{pmatrix} \psi^L(\mathbf{r}) \\ \psi^S(\mathbf{r}) \end{pmatrix}. \tag{4.55}$$

For various applications and derivations it is easier to work with the two-component form in (4.54) rather than with the full four-component form of (4.53), and in later chapters we will use either form according to what is convenient for the case at hand.

To obtain further information on the nature of the Dirac wave function, we can solve the equation for a simple model system. The simplest case is the time-independent equation for a free particle. In the nonrelativistic case the Schrödinger equation for a free particle moving along the x axis is

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = E\psi. \tag{4.56}$$

This has the solution

$$\psi = ae^{ik_x x}; \quad k_x = \sqrt{\frac{2mE}{\hbar^2}} \tag{4.57}$$

where a is a normalization constant. k_x can be interpreted by noting that ψ is also an eigenfunction of the momentum operator

$$\hat{p}_x\psi = -i\hbar\frac{\mathrm{d}\psi}{\mathrm{d}x} = -i\hbar\frac{\mathrm{d}ae^{ik_x x}}{\mathrm{d}x} = \hbar k_x\psi. \tag{4.58}$$

This type of exponential wave function is known as a *plane wave*, and for the general case of a particle moving along the direction $\mathbf{k}$ it has the form

$$\psi = ae^{i\mathbf{k}\cdot\mathbf{r}}. \tag{4.59}$$

The time-independent Dirac equation for a free particle follows easily from (4.52) as

$$i\hbar c\boldsymbol{\alpha}\cdot\nabla\psi + mc^2\beta\psi = E\psi. \tag{4.60}$$

We now draw on our experience from the nonrelativistic case and assume that the 4-spinor ψ has the form

$$\psi = e^{i\mathbf{k}\cdot\mathbf{r}}\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix}. \tag{4.61}$$

---

[5] The notation convention is that upper case is used for the time-dependent spinors; lower case for the time-independent spinors.

We insert this expression into the Dirac equation, expand the 4 × 4 matrices, and carry out the differentiations. The result is the following set of equations:

$$\begin{aligned}(E-mc^2)a_1 - K_z a_3 - K_- a_4 &= 0,\\ (E-mc^2)a_2 - K_+ a_3 + K_z a_4 &= 0,\\ -K_z a_1 - K_- a_2 + (E+mc^2)a_3 &= 0,\\ -K_+ a_1 + K_z a_2 + (E+mc^2)a_4 &= 0,\end{aligned} \tag{4.62}$$

where we have set

$$\mathbf{K} = c\hbar\mathbf{k} \qquad \text{and} \qquad K_\pm = K_x \pm iK_y. \tag{4.63}$$

This set of four equations for four variables has solutions only if there is linear dependence; that is, the determinant of coefficients must equal zero:

$$\begin{vmatrix} (E-mc^2) & 0 & -K_z & -K_- \\ 0 & (E-mc^2) & -K_+ & K_z \\ -K_z & -K_- & (E+mc^2) & 0 \\ -K_+ & K_z & 0 & (E+mc^2) \end{vmatrix} = 0. \tag{4.64}$$

This reduces to a fourth degree equation for E,

$$(E^2 - m^2c^4 + K^2)^2 = 0, \tag{4.65}$$

which has the doubly degenerate solutions

$$E_+ = +\sqrt{m^2c^4 + K^2}, \qquad E_- = -\sqrt{m^2c^4 + K^2}. \tag{4.66}$$

The equations for the coefficients turn out to be pairwise linearly dependent for these values of E. For E_+ we choose the equations

$$\begin{aligned}-K_z a_1 - K_- a_2 + (E_+ + mc^2)a_3 &= 0,\\ -K_+ a_1 + K_z a_2 + (E_+ + mc^2)a_4 &= 0.\end{aligned} \tag{4.67}$$

With four variables and only two equations, we can choose two of the variables freely, but one of these will be determined by the normalization conditions (which we do not consider in this example). If we set $a_2 = 0$, we get

$$a_3 = \frac{a_1 K_z}{E_+ + mc^2}; \quad a_4 = \frac{a_1 K_+}{E_+ + mc^2}. \tag{4.68}$$

The choice $a_1 = 0$ gives

$$a_3 = \frac{a_2 K_-}{E_+ + mc^2}; \qquad a_4 = -\frac{a_2 K_z}{E_+ + mc^2}. \tag{4.69}$$

Correspondingly, for E_- we get the solutions

$$a_1 = \frac{a_3 K_z}{E_- - mc^2}; \qquad a_2 = \frac{a_3 K_+}{E_- - mc^2}; \qquad a_4 = 0 \tag{4.70}$$

and

$$a_1 = \frac{a_4 K_-}{E_- - mc^2}; \qquad a_2 = -\frac{a_4 K_z}{E_- - mc^2}; \qquad a_3 = 0. \tag{4.71}$$

To sum up, for the free-particle Dirac equation we get the solutions:

$$\begin{aligned}
E_1 &= +\sqrt{m^2c^4 + K^2}; & \psi_1 &= ae^{i\mathbf{k}\cdot\mathbf{r}}\begin{pmatrix} 1 \\ 0 \\ K_z/(E+mc^2) \\ K_+/(E+mc^2) \end{pmatrix}, \\
E_2 &= +\sqrt{m^2c^4 + K^2}; & \psi_2 &= ae^{i\mathbf{k}\cdot\mathbf{r}}\begin{pmatrix} 0 \\ 1 \\ K_-/(E+mc^2) \\ -K_z/(E+mc^2) \end{pmatrix}, \\
E_3 &= -\sqrt{m^2c^4 + K^2}; & \psi_3 &= ae^{i\mathbf{k}\cdot\mathbf{r}}\begin{pmatrix} K_z/(E-mc^2) \\ K_+/(E-mc^2) \\ 1 \\ 0 \end{pmatrix}, \\
E_4 &= -\sqrt{m^2c^4 + K^2}; & \psi_4 &= ae^{i\mathbf{k}\cdot\mathbf{r}}\begin{pmatrix} K_-/(E-mc^2) \\ -K_z/(E-mc^2) \\ 0 \\ 1 \end{pmatrix}.
\end{aligned} \tag{4.72}$$

Comparing coefficients from the lower two components with those of the upper two, we see that there is a ratio of the order $K/(E + mc^2)$. Remembering that $K = cp$, this may be expanded as

$$\frac{K}{E + mc^2} = \frac{cp}{mc^2 + \sqrt{m^2c^4 + p^2c^2}} < \frac{p}{2mc}. \tag{4.73}$$

For nonrelativistic velocities $p \ll mc$, and when the upper two components are large, the lower two are small, and vice versa. While we normally do not have situations where the upper or lower components are pure Pauli spinors, as in our present example, the magnitude difference between upper and lower components always holds one way or the other.

Which way the difference goes depends, as we have seen, on the sign of E. We have not commented on the negative-energy solutions, which would appear to be unphysical. Admittedly, we regard the bound states of the hydrogen atom as negative, for example, but that is relative to the ionization limit and not on an apparently absolute scale. We will postpone the discussion of these negative-energy states until the next chapter. For now we merely register the fact that solutions with positive energies have large values for the upper components of the 4-spinor, while solutions with negative energies have large values for the lower components.

The free-particle Dirac equation provides a nice demonstration of some of the properties of the 4-spinor solutions, but quantum chemistry is mainly concerned with electrons bound in molecules by electromagnetic forces. In a static potential V, such as that provided by the nuclei in the Born–Oppenheimer frame, where the vector potential $\mathbf{A}$ is zero, the time-independent electronic Dirac equation is

$$c(\boldsymbol{\alpha} \cdot \mathbf{p})\psi + (\beta mc^2 + V)\psi = E\psi. \tag{4.74}$$

Writing ψ in terms of large and small components, we have

$$\begin{pmatrix} V - E + mc^2 & c(\boldsymbol{\sigma} \cdot \mathbf{p}) \\ c(\boldsymbol{\sigma} \cdot \mathbf{p}) & V - E - mc^2 \end{pmatrix} \begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} = 0 \tag{4.75}$$

or as a pair of coupled equations,

$$\begin{aligned} (V - E + mc^2)\psi^L + c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^S &= 0 \\ c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L + (V - E - mc^2)\psi^S &= 0. \end{aligned} \tag{4.76}$$

Solution of these equations for various forms of V is the subject of most of the following chapters. But we see from comparing these equations with the corresponding free-particle equations that the overall structure is similar, and that we would expect the solutions to have much the same form; that is, two large components and two small components.

4.6 Nonrelativistic Limit of the Dirac Equation

While we cannot provide general solutions to the Dirac equation of an electron in a molecular field, we can gain further insight by investigating what happens when we go to the nonrelativistic limit; that is, when we let the speed of light go to infinity. Just as in classical mechanics we get the Newtonian nonrelativistic limit of the relativistic Hamiltonian function by subtracting the rest energy and taking the $c \to \infty$ limit, so we expect that in quantum mechanics the limit of the Dirac equation as we allow c to go to infinity after the rest energy has been subtracted should be the nonrelativistic Schrödinger equation.

If we partition the energy into the rest term plus a remainder,

$$E = mc^2 + E' \tag{4.77}$$

we may define a new Hamiltonian

$$\hat{\mathcal{H}}' = \hat{\mathcal{H}} - \mathbf{I}_4 mc^2 \tag{4.78}$$

which amounts to defining

$$\beta' = \beta - \mathbf{I}_4. \tag{4.79}$$

Using this Hamiltonian and E' as the separation variable in (4.50) to obtain a time-independent equation, the only difference is that the new temporal wave function is related to the old by a phase factor,

$$\theta'(t) = e^{-(mc^2/i\hbar)t}\theta(t). \tag{4.80}$$

Henceforth, we will feel free to make use of this new definition of the Hamiltonian and of the energy when convenient, and drop the primes everywhere. We will make it clear when the rest mass term is included.

The time-independent Dirac equation in two-component form with the rest mass subtracted becomes

$$(V - E)\psi^L + c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^S = 0 \tag{4.81a}$$

$$c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L + (V - E - 2mc^2)\psi^S = 0. \tag{4.81b}$$

Before taking the limit $c \to \infty$, this equation must be rearranged for two reasons: first, because we need to change it to a form where c occurs in some form of denominator—this will provide us with terms that vanish and hopefully other terms that remain finite—and, second, because the nonrelativistic wave function is a scalar function, whereas the Dirac wave function is a four-component vector function. If we use the two-component nonrelativistic Schrödinger equation that we derived in section 4.2, we can write the nonrelativistic wave function in terms of spin–orbitals, which can be transformed to two-component spinors. Then it is only necessary to reduce the Dirac equation from four-component to two-component form.

If we set $c \to \infty$ in (4.81b) after dividing it by c^2, we find that $\psi^S \to 0$, which implies that we must eliminate the small component before allowing c to go to infinity in order to obtain the nonrelativistic limit. Examining (4.81b), we see that it can be rearranged to make ψ^S the subject, provided that $V - E - 2mc^2$ is nonzero. Now V is always zero or negative for the nuclear potential, so we must have $E > -2mc^2$ in order that $V - E - 2mc^2$ is never zero. (Compared with the free-particle case discussed in the previous section, this corresponds to choosing the positive energy solution.) With this condition,

$$\psi^S = \left[2mc^2 + E - V\right]^{-1} c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L \tag{4.82}$$

which on substitution into (4.81a) gives

$$(V - E)\psi^L + c^2(\boldsymbol{\sigma} \cdot \mathbf{p})\left[2mc^2 + E - V\right]^{-1}(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L = 0. \tag{4.83}$$

The last term can be rearranged by noting that

$$(A + B)^{-1} = A^{-1} - A^{-1}B(A + B)^{-1} \tag{4.84}$$

for any quantities A and B for which $(A + B)^{-1}$ and A^{-1} exist. Letting $A = 2mc^2$ and $B = (E - V)$, the result is

$$(V - E)\psi^L + \frac{1}{2m}(\boldsymbol{\sigma} \cdot \mathbf{p})(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L - \frac{1}{2m}(\boldsymbol{\sigma} \cdot \mathbf{p})\frac{E - V}{2mc^2 + E - V}(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L = 0. \tag{4.85}$$

Using (4.15) to reduce the second term, we find that it is the nonrelativistic kinetic energy operator $\hat{T}$. The above equation is therefore

$$(\hat{T} + V - E)\psi^L + \frac{1}{2m}(\boldsymbol{\sigma} \cdot \mathbf{p})\frac{V - E}{2mc^2 + E - V}(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L = 0 \tag{4.86}$$

which is the nonrelativistic Schrödinger equation with a relativistic correction. Since c is in the denominator of the last term, if we allow c to go to infinity we get the nonrelativistic Schrödinger equation with ψ^L the nonrelativistic wave function.

We may also partition the second term using (4.14) to arrive at

$$\begin{aligned}\frac{1}{2m}(\boldsymbol{\sigma} \cdot \mathbf{p})\frac{V - E}{2mc^2 + E - V}(\boldsymbol{\sigma} \cdot \mathbf{p}) = \frac{V - E}{2mc^2 + E - V}T - \frac{c^2}{(2mc^2 + E - V)^2}\\ \times\left[(\mathbf{p}V) \cdot \mathbf{p} + i\boldsymbol{\sigma} \cdot (\mathbf{p}V) \times \mathbf{p}\right].\end{aligned} \tag{4.87}$$

Here we see that the relativistic correction includes a correction to the kinetic energy, a scalar potential term, and a term involving the spin that can be classified as a spin–orbit interaction.

In the previous section, we showed that the ratio of the large component to the small component was approximately $p/2mc$ for a free-electron state. Given that the large component becomes the nonrelativistic wave function in two-component form in the limit $c \to \infty$, we may use the nonrelativistic wave function as an approximation to the large component to gain some idea of how the small component behaves for a bound state. Taking, for example, the nonrelativistic hydrogenic atom 1s wave function as an approximation to ψ^L, we find that

$$\begin{aligned} &\psi^L = N_{1s} e^{-Zr} \Rightarrow \\ &\psi^S \approx \frac{(\boldsymbol{\sigma} \cdot \mathbf{p})}{2mc} \psi^L = N_{1s} \frac{(\boldsymbol{\sigma} \cdot \mathbf{p})}{2mc} e^{-Zr} \approx \frac{Z}{2mc} \psi^L. \end{aligned} \tag{4.88}$$

The magnitude of c is approximately 137 in Hartree atomic units, which we have used here. We will obtain a more accurate value for this ratio in chapter 7.

For light elements, the small component will be one to two orders of magnitude smaller than the large component for the 1s orbital, and any expectation value involving it will be two to four orders of magnitude smaller than the expectation value of the same operator over the large component. The names *large* and *small* components are in common usage, but, as we have seen previously, the large component is large only for the positive energy solutions. More appropriate names might be *upper* and *lower* components, and indeed some textbooks use this nomenclature. Unfortunately, this may also lead to problems, because for some purposes it is convenient to interchange the order of the components of the 4-spinor (for example, in some computational applications). In this case the upper components may not both be upper anymore, but as long as we are dealing with positive energies, the large components will at least always be large. For this book, we will adhere to the conventional large/small terminology.

One slightly subtle note of caution needs to be added in connection with the limiting process carried out above. If the potential V is a Coulomb potential, $V = -Z/r$, the last term in (4.86) does *not* vanish for infinite c at $r = 0$. Employing the correspondence principle for $\mathbf{p}$, the operator in the last term is

$$\begin{aligned} &\frac{1}{2m} (\boldsymbol{\sigma} \cdot \mathbf{p}) \frac{V - E}{2mc^2 + E - V} (\boldsymbol{\sigma} \cdot \mathbf{p}) \\ &\quad = \frac{i\hbar c^2}{(2mc^2 + E + Z/r)^2} \frac{Z}{r^2} (\boldsymbol{\sigma} \cdot \mathbf{e}_r)(\boldsymbol{\sigma} \cdot \mathbf{p}) - \frac{Z/r + E}{2mc^2 + E + Z/r} T \\ &\quad = \frac{i\hbar Z}{(2mcr + Er/c + Z/c)^2} (\boldsymbol{\sigma} \cdot \mathbf{e}_r)(\boldsymbol{\sigma} \cdot \mathbf{p}) - \frac{Z/c + Er/c}{2mcr + Er/c + Z/c} T \end{aligned} \tag{4.89}$$

where $\mathbf{e}_r$ is a unit radial vector. Taking both limits, but keeping terms in the product of r and c, we find

$$\lim_{r \to 0, c \to \infty} \frac{1}{2m} (\boldsymbol{\sigma} \cdot \mathbf{p}) \frac{V - E}{2mc^2 + E - V} (\boldsymbol{\sigma} \cdot \mathbf{p}) = \frac{i\hbar Z}{(2mcr)^2} (\boldsymbol{\sigma} \cdot \mathbf{e}_r)(\boldsymbol{\sigma} \cdot \mathbf{p}) \tag{4.90}$$

In order for the limit to vanish, cr must approach infinity. We have no reason to suppose that c would approach infinity faster than r approaches zero, so we cannot assume that this limit vanishes. Thus, for a point nuclear model, the nonrelativistic limit ($c \to \infty$) is only obtained for $r > 0$. This points to the need for a more realistic model of the nuclear charge distribution, which we consider in chapter 7.

In later chapters we will consider the nonrelativistic limit of other relativistic operators. Here we will only look at one, the velocity operator, to show that it has the correct nonrelativistic limit. Extracting $2mc^2$ from the square bracket in (4.82) we get

$$\psi^S = \frac{1}{2mc}\left[1 + (E - V)/2mc^2\right]^{-1}(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L. \tag{4.91}$$

In the nonrelativistic limit the term in square brackets may be replaced with unity, again provided that V is finite. Using this and the definition of $\boldsymbol{\alpha}$, (4.33), we write the current density as

$$\begin{aligned}\lim_{c\to\infty} \psi^\dagger c\boldsymbol{\alpha}\psi &= \frac{1}{2m}\psi^{L\dagger}\left[\boldsymbol{\sigma}(\boldsymbol{\sigma} \cdot \mathbf{p}) + (\boldsymbol{\sigma} \cdot \mathbf{p})\boldsymbol{\sigma}\right]\psi^L \\ &= \frac{1}{m}\psi^{L\dagger}\mathbf{p}\psi^L,\end{aligned} \tag{4.92}$$

from the anticommuting properties of the Pauli matrices. Classically, $\mathbf{p} = m\mathbf{v}$, so the velocity operator does indeed have the correct nonrelativistic form.

Finally, at the interface between relativistic and nonrelativistic representations, we should consider the nuclear potential, which we have introduced only by way of an unspecified potential V above. So far we have only discussed the Dirac equation for an electron, but in quantum chemistry we are often concerned with the nuclear motion as well. From the discussion of the potentials due to a static and a moving charge in chapter 3, it is obvious that the nuclear motion creates a magnetic field, which should affect the electronic motion. The same arguments as are used for separating nuclear and electronic motion in nonrelativistic quantum chemistry may be applied here also: the mass of the nuclei is at least three orders of magnitude larger than the electron mass, and the velocities of the nuclei are correspondingly smaller. Thus we conclude that as far as electromagnetic electron–nucleus interactions go, we may well regard the nuclei as stationary. Relativity also affects the nucleus–nucleus interaction in a molecule, but again the effect is small. We therefore conclude that relativity is a very small perturbation on the nuclear motion, and for most chemical purposes it can be neglected entirely. Thus we will treat the electronic motion relativistically but the nuclear motion nonrelativistically. We will, however, return to these questions in chapter 13 when we discuss the calculation of molecular properties.

5

Negative-Energy States and Quantum Electrodynamics

We now return to the problem of the negative-energy solutions that appeared when we solved the free particle Dirac equation in the previous chapter. The energy eigenvalues obtained there were either $E_+ = +\sqrt{m^2c^4 + p^2c^2}$ or $E_- = -\sqrt{m^2c^4 + p^2c^2}$. The minimum absolute values we can have are therefore $|E| = mc^2$ for $p = 0$ (that is, the particle at rest). As the momentum of the particle increases, we generate a continuum of solutions, either below $-mc^2$ or above $+mc^2$. This is a general feature of all solutions we will obtain from the Dirac equation—we will have continuum solutions on both sides of an energy gap stretching from $-mc^2$ to $+mc^2$, in addition to any discrete solutions.

Classically and nonrelativistically we would expect a free particle to have a positive energy. The addition of a rest mass term mc^2, which is definitely also positive, should not change this. However, the fact that we now have negative-energy states means that a single particle in a positive-energy state could spontaneously fall to a negative-energy state with the emission of a photon. The interaction with the radiation field occurs via the operator $\boldsymbol{\alpha} \cdot \mathbf{A}$. The radiative transition moment therefore connects the large component of the positive-energy solution with the small component of the negative-energy solution. As we have seen, both of these should be large in magnitude, giving a large transition moment. Calculations (Bjorken and Drell 1964) show that the transition rate into the highest mc^2 section of the negative continuum is approximately 10^8 s^{-1}, and the rate of decay into the whole continuum is infinite. Any bound state would therefore immediately dissolve into the negative continuum with the emission of photons. This is clearly an unphysical situation.

To resolve this dilemma, Dirac postulated in 1930 that the negative-energy states are fully occupied. The implications of this postulate are significant and wide-ranging. First and foremost, it prevents the collapse of bound states. Second, if we excite a particle out of a filled negative-energy state, it must leave behind a "hole" in this state that behaves as a particle with complementary properties to the one we excite. This means that, for example, total charge must be conserved. Thus, if the negative-energy states arose from the solution of the Dirac equation for a free electron, a hole in the negative-energy continuum must behave as a particle with the same mass as the electron but with opposite charge. The negative-energy states are therefore connected with *positrons*. This explains the phenomenon of *pair creation* and the energy required to create an electron–positron pair with zero kinetic energy is $2mc^2$, as expected if this involves exciting a particle across the gap between negative- and positive-energy states. The discovery of the positron in 1933 was seen as a triumph for the Dirac theory and its reinterpretation.

There is, however, a problem with this approach. The filling of the negative continuum implies that the "vacuum" state is infinitely charged, and as a consequence we would expect an infinite interaction between any state and the vacuum. We therefore need to take the reinterpretation one step further. What we actually measure is not the absolute properties of a state but the *differences* in properties between a state and the vacuum. The Dirac equation is therefore only the starting point for a theory that is an infinitely-many-body theory, even for a system with only one electron—or indeed for the vacuum itself!

Before we go any further, we should reflect on how we arrived at this position. We solved the Dirac equation for a free particle and found that the spectrum contained two disjoint continua. If the particle is an electron, we interpret the positive-energy states as describing the electron with an infinite rate of decay into the negative continuum. But we could equally well have solved the Dirac equation for a particle of positive charge—a positron—in the same field, and found exactly the same situation, with two continua, and an infinite rate of decay of the positive-energy states, which now describe positrons. So if the negative continuum for the electron should be filled, so should the negative continuum for a positron (even if we have no particles), and the vacuum, which should be a reference point for all particles, would then have no charge.

In fact, there is a more direct connection between the charge of the particle and the mass-energy, so that we only need to solve *one* Dirac equation to obtain solutions for *both* signs of the charge. We write the Dirac equation for an electron of charge $-e$ in four-vector form as

$$c\mathrm{\alpha} \cdot (i\hbar\Box - e\mathrm{A})\,\Psi = \beta mc^2\Psi. \tag{5.1}$$

We can take the complex conjugate of this equation, but we have to exercise some care with regard to the four-vector quantities, which represent a mixture of real and complex entities. Showing the space and the time components of the four-vectors explicitly, the equation above takes the form

$$c(\boldsymbol{\alpha}, i) \cdot \left[i\hbar \left(\nabla, \frac{-i}{c}\frac{\partial}{\partial t} \right) - e\left(\mathbf{A}, \frac{i\phi}{c} \right) \right] \Psi = \beta mc^2\Psi. \tag{5.2}$$

Complex conjugation yields

$$c(\boldsymbol{\alpha}^*, -i) \cdot \left[-i\hbar \left(\nabla, \frac{i}{c}\frac{\partial}{\partial t} \right) - e\left(\mathbf{A}, \frac{-i\phi}{c} \right) \right] \Psi^* = \beta mc^2\Psi^*. \tag{5.3}$$

Due to the presence of a scalar product between the vectors, we can recover the four-vector quantities $\Box$ and A by changing the sign of the last component in each four-vector and writing this expression as

$$c(\boldsymbol{\alpha}^*, i) \cdot [-i\hbar\Box - e\mathrm{A}]\,\Psi^* = \beta mc^2\Psi^*. \tag{5.4}$$

This looks very similar to the Dirac equation, but although $\boldsymbol{\alpha}$ is Hermitian, the complex conjugate is not the same as the original because $\alpha_y^* = -\alpha_y$. In order to recover the four-vector $\mathrm{\alpha}$, we introduce the operator

$$\hat{C}' = i\beta\alpha_y. \tag{5.5}$$

Operating on the Dirac matrices we find that

$$\hat{C}'\beta = -\beta\hat{C}'; \qquad \hat{C}'(\boldsymbol{\alpha})^* = \boldsymbol{\alpha}\hat{C}'. \tag{5.6}$$

We apply this operator to the conjugated Dirac equation and arrive at the result

$$c\mathrm{\alpha} \cdot \left(i\hbar\Box + e\mathrm{A}_\mu\right)\hat{C}'\Psi^* = \beta mc^2\hat{C}'\Psi^*. \tag{5.7}$$

which, by comparison with (5.1), is the Dirac equation for a particle of the same mass but opposite charge. Thus, if Ψ is the wave function for the electron, $\hat{C}'\Psi^*$ must be the wave function for a positron. Introducing the complex conjugation operator, $\hat{\mathcal{K}}_0$, the positron wave function can be written $\hat{C}'\hat{\mathcal{K}}_0\Psi$. $\hat{C} = \hat{C}'\hat{\mathcal{K}}_0$ is called the *charge conjugation* operator, which in matrix form is

$$\hat{C} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}\hat{\mathcal{K}}_0. \tag{5.8}$$

Applying $\hat{C}$ to the fourth of the wave functions defined in (4.72), which is for a negative-energy electron, we find

$$\hat{C}\psi_4 = \hat{C}ae^{i\mathbf{k}\cdot\mathbf{r}} \begin{pmatrix} K_-/(E_- - mc^2) \\ -K_z/(E_- - mc^2) \\ 0 \\ 1 \end{pmatrix} = a^*e^{-i\mathbf{k}\cdot\mathbf{r}} \begin{pmatrix} 1 \\ 0 \\ K_z/(E_- - mc^2) \\ K_+/(E_- - mc^2) \end{pmatrix} \tag{5.9}$$

Noting that $E_- = -E_+$, we may write this as

$$\hat{C}\psi_4 = a^*e^{-i\mathbf{k}\cdot\mathbf{r}} \begin{pmatrix} 1 \\ 0 \\ -K_z/(E_+ + mc^2) \\ -K_+/(E_+ + mc^2) \end{pmatrix}. \tag{5.10}$$

This is equal to ψ_1 if a is real and we change the sign of the momentum. Thus, we have the wave function for a positive-energy particle of opposite momentum. If we had included the time factor we would have seen a change in sign there too. So we may interpret the negative-energy electrons as positive-energy positrons moving backwards in time. We will examine the issue of time reversal in more detail later when we address the symmetry of the Dirac equation. We note also that there has been a change in spin, which we would also expect from time reversal.

The same Dirac equation therefore provides solutions that describe electrons *and* positrons. Whether they appear as negative-energy states or positive-energy states will depend on which sign we choose for the charge. The physics of the particles ought to be independent of this choice, and we therefore need to have a mathematical model that permits the description of both positrons and electrons on the same footing. The necessity for an alternative model is further obvious if we follow the line of reasoning about the filling of the negative-energy states a little further. We argued that the negative-energy states for both the electron and positron Dirac equations must be filled, producing a "vacuum" with no net charge but an infinite number of particles. However, we now have an infinite density of particle-antiparticle pairs, which must spontaneously annihilate each other at an infinite rate, leaving the vacuum empty again! This absurdity is only resolved if we do not insist that the Dirac equation on its own is a complete description of the physics. When we developed the Dirac equation, we chose the positive sign for the square root, implying that only the solutions with positive (classical) mass were meaningful. We must therefore view the Dirac equation with its negative-energy solutions as a mathematical means to obtain solutions for particles of both signs of the charge at the same time; that is, we do not need to solve the Dirac equation for both signs separately. The existence of negative-energy states is the mathematical but not the physical embodiment of the existence of particles and antiparticles.

For an alternative model we must turn to quantum field theory and quantum electrodynamics. This is suggested by one final point from the heuristic model of the filling of the negative-energy states: the annihilation of the particles produces a radiation field. We therefore require a theory that treats particles, antiparticles and their interaction with the electromagnetic field in a consistent way. Quantum electrodynamics (QED) has been hailed as the most accurate theory ever devised, and it is indeed remarkable both in its predictive power and its logically coherent treatment of the entire range of electromagnetic interactions within the relativistic quantum regime. We do not expect most practicing chemists to have a background in QED, and a complete introduction to the subject certainly falls outside the scope of this text. We will therefore follow our usual approach—beg, borrow, or steal what we need from the advanced physics theories and adapt it to our immediate needs. It is, however, possible to develop much of the theory presented here within the context of QED. The interested reader should consult the articles by Quiney and coworkers (Quiney 2002, Quiney et al. 2001). For more detailed derivations, see Berestetskii et al. (1982) or Bjorken and Drell (1964).

5.1 Second Quantization

It is convenient to use the language of *second quantization* when treating a general many-body problem. Most quantum chemists are familiar with this formalism, and here we will only provide some brief reminders of the basic features. More thorough discussions may be found in standard texts such as Merzbacher (1970), Schiff (1968), or Jørgensen and Simons (1981).

In second quantization we assume that a complete set of one-particle operators with corresponding eigenvalues generates a complete set of one-particle states for the system under consideration. The general *N*-particle state is then labeled by the eigenvalues of each of the *N* particles in this representation. The entire space of such many-particle states is called a *Fock space*. The basis for this space is a distribution of particles with the specified eigenvalues. If the eigenvalues of the one-particle operators are (collectively) $\omega_1, \omega_2, \ldots$, and n_i is the number of particles with eigenvalue ω_i, then the general basis vector in Fock space is

$$|n_1, n_2, \ldots\rangle. \tag{5.11}$$

The simplest state is the state with no particles, the *vacuum state*, which we denote $|vac\rangle$. This state is assumed to be normalized:

$$|0, 0, \ldots\rangle \equiv |vac\rangle; \qquad \langle vac|vac\rangle = 1. \tag{5.12}$$

We can generate the other states in the Fock space by use of *creation operators*. The creation operator $a_i{}^\dagger$ increases the number of particles in state i by 1:

$$a_i{}^\dagger|n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots\rangle = C_i|n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots\rangle \tag{5.13}$$

where C_i is a normalization constant. Correspondingly, the Hermitian conjugate of the creation operator, a_i, removes a particle from state i, and is called an *annihilation operator*:

$$a_i|n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots\rangle = C_i{}'|n_1, \ldots, n_{i-1}, n_i - 1, n_{i+1}, \ldots\rangle. \tag{5.14}$$

The general state in Fock space may be constructed by applying a succession of creation operators to the vacuum state:

$$(a_1{}^\dagger)^{n_1}(a_2{}^\dagger)^{n_2}\ldots|vac\rangle = C_x|n_1, n_2, \ldots\rangle. \tag{5.15}$$

Application of an annihilation operator to the vacuum state yields 0, and by Hermitian conjugation we must have

$$a_i|vac\rangle = 0 \quad \Rightarrow \quad \langle vac|a_i{}^\dagger = 0. \tag{5.16}$$

For electrons and positrons, which will be our main concern, the creation and annihilation operators fulfill the anticommutation relations:

$$\begin{aligned} a_q{}^\dagger a_p{}^\dagger + a_p{}^\dagger a_q{}^\dagger &= 0 \\ a_q a_p + a_p a_q &= 0 \\ a_q{}^\dagger a_p + a_p a_q{}^\dagger &= \delta_{pq}. \end{aligned} \tag{5.17}$$

From these relations it is straightforward to show that the *number operator*

$$\hat{N} \equiv \sum_{p} a_p{}^{\dagger} a_p \tag{5.18}$$

yields the number of particles in the system. Here and in the following, the sum is over the complete set of one-particle states unless otherwise indicated. We can also show from the commutation rules that trying to create two electrons in the same state yields a null result:

$$2\left(a_i{}^{\dagger} a_i{}^{\dagger}\right)|vac\rangle = \left(a_i{}^{\dagger} a_i{}^{\dagger} + a_i{}^{\dagger} a_i{}^{\dagger}\right)|vac\rangle = 0. \tag{5.19}$$

This means that there can only be one electron or positron (or in general, fermion) in any state.

The second-quantized representation of a one-particle operator $\hat{\Omega}$ is given as

$$\hat{\Omega} = \sum_{p,q} \Omega_{pq} a_p{}^{\dagger} a_q. \tag{5.20}$$

Here Ω_{pq} is the usual first-quantized matrix element between the one-particle states with eigenvalues ω_p and ω_q

$$\Omega_{pq} = \langle \psi_p | \Omega | \psi_q \rangle. \tag{5.21}$$

5.2 Relativistic Second-Quantized Hamiltonians

In the second-quantized representation of the Dirac Hamiltonian for the electron, the complete set of states must be used to obtain a proper representation of the first-quantized Dirac Hamiltonian. We partition this set into two sets, containing the positive-energy and negative-energy states. We will use the index p^+ for the positive-energy states and p^- for the negative-energy states. A positive-energy state is generated by

$$a_{p^+}{}^{\dagger}|vac\rangle, \tag{5.22}$$

while the corresponding negative-energy state is

$$a_{p^-}{}^{\dagger}|vac\rangle, \tag{5.23}$$

and the Dirac Hamiltonian may be written[1]

$$\hat{\mathcal{H}} = \sum_{pq} \left[h_{p^+q^+}\ a_{p^+}{}^{\dagger} a_{q^+} \ + h_{p^+q^-}\ a_{p^+}{}^{\dagger} a_{q^-} \ + h_{p^-q^+}\ a_{p^-}{}^{\dagger} a_{q^+} \right.$$

$$\left. + h_{p^-q^-}\ a_{p^-}{}^{\dagger} a_{q^-} \ \right]. \tag{5.24}$$

There is a possible inconsistency in the expression for the negative-energy state: while (5.23) follows from the second quantization formalism, the existence of a filled vacuum as postulated by Dirac would prevent the use of a creation operator. We can get around this difficulty by using our previous interpretation of negative-energy states as related to positrons. The excitation of an electron from a (filled) negative-energy state corresponds to the creation of a positron hole. Thus, if we create an electron in one of these states, it is equivalent to annihilating a positron. So we can write

$$a_{p^-}{}^{\dagger} |vac\rangle \equiv \tilde{a}_p\ |vac\rangle \tag{5.25}$$

where we use the tilde to refer to the positron quantities. Note that we have placed the tilde over the operator rather than over the index (and we also drop the + superscript on the electron indices). This is to emphasize that positron and electron solutions really form two distinct spaces. However, in matrix elements we have no choice but to place the tilde over the index.

With this formalism, the Hamiltonian operator is

$$\hat{\mathcal{H}} = \sum_{pq} \left[h_{pq}\ a_p{}^{\dagger} a_q \ + h_{p\tilde{q}}\ a_p{}^{\dagger} \tilde{a}_q{}^{\dagger} + h_{\tilde{p}q}\ \tilde{a}_p\ a_q \ + h_{\tilde{p}\tilde{q}}\ \tilde{a}_p\ \tilde{a}_q{}^{\dagger} \right]. \tag{5.26}$$

There are several points to be noted about this operator. First, the second term creates an electron–positron pair, and the third term annihilates an electron–positron pair. This means that the Hamiltonian connects states with different particle numbers, that is, particle number is *not* conserved, though charge is. The existence of these terms embodies the idea of an infinitely-many-body problem that arose from the filling of the negative-energy states in Dirac's interpretation. Second, the order of the operators in the fourth term means that the vacuum expectation value of this operator is *not* zero, but

$$\langle vac|\hat{\mathcal{H}}|vac\rangle = \sum_{pq} h_{\tilde{p}\tilde{q}} \langle vac|\tilde{a}_p\ \tilde{a}_q{}^{\dagger}|vac\rangle = \sum_{pq} h_{\tilde{p}\tilde{q}} \delta_{pq} = \sum_{p} h_{\tilde{p}\tilde{p}}. \tag{5.27}$$

Thus the vacuum has an energy that is equal to the sum of the energies of the negative solutions of the Dirac equation, as is expected from Dirac's interpretation. Note that the matrix elements are the same as in the Dirac equation, so the sum is negative and infinite. This Hamiltonian operator therefore represents the first stage of the Dirac reinterpretation with the negative-energy states all filled.

---

[1] We write the matrix elements of the one-particle Hamiltonian with a lower case h, regardless of the case of the operator. Upper case H is reversed for many-particle Hamiltonian matrix elements.

To avoid the negative infinite vacuum energy, the vacuum expectation value is subtracted from the Hamiltonian to define a new, QED Hamiltonian[2]:

$$\begin{aligned}\hat{\mathcal{H}}^{\text{QED}} = \sum_{pq} \big[& h_{pq}\, a_p{}^{\dagger} a_q \ + h_{p\tilde{q}}\, a_p{}^{\dagger} \tilde{a}_q{}^{\dagger} + h_{\tilde{p}q}\, \tilde{a}_p\, a_q \\ & + h_{\tilde{p}\tilde{q}}\, (\tilde{a}_p\, \tilde{a}_q{}^{\dagger} - \langle vac|\tilde{a}_p\, \tilde{a}_q{}^{\dagger}|vac\rangle)\big].\end{aligned} \tag{5.28}$$

The last term can be rewritten using the anticommutation relations as

$$\tilde{a}_p\, \tilde{a}_q{}^{\dagger} - \langle vac|\tilde{a}_p\, \tilde{a}_q{}^{\dagger}|vac\rangle = \tilde{a}_p\, \tilde{a}_q{}^{\dagger} - \delta_{\tilde{p}\tilde{q}} = -\tilde{a}_q{}^{\dagger}\tilde{a}_p \tag{5.29}$$

and thus the Hamiltonian is

$$\hat{\mathcal{H}}^{\text{QED}} = \sum_{pq} \big[h_{pq}\, a_p{}^{\dagger} a_q \ + h_{p\tilde{q}}\, a_p{}^{\dagger} \tilde{a}_q{}^{\dagger} + h_{\tilde{p}q}\, \tilde{a}_p\, a_q \ - h_{\tilde{p}\tilde{q}}\, \tilde{a}_q{}^{\dagger}\tilde{a}_p \ \big]. \tag{5.30}$$

The subtraction of the vacuum expectation value is therefore equivalent to permuting all annihilation operators to the right as if all anticommutators had vanished. This is called *normal ordering*, and operators that are normal-ordered are enclosed in colons:

$$\hat{\mathcal{H}}^{\text{QED}} \ = :\hat{\mathcal{H}}: . \tag{5.31}$$

The QED Hamiltonian has the features we were looking for above. Its vacuum expectation value is zero (by construction), that is, the vacuum has zero energy. The positronic term now has a negative sign, and the energy of a positron state $|\tilde{p}\rangle = \tilde{a}_p{}^{\dagger}|vac\rangle$ is

$$\langle\tilde{p}|\ :\hat{\mathcal{H}}:\ |\tilde{p}\rangle = -h_{\tilde{p}\tilde{p}} > 0. \tag{5.32}$$

Thus, the negative-energy electron solutions have now become positive-energy positron solutions, and because of this there is no more danger of radiative decay of bound electrons into the negative continuum.

To gain some more insight into the transition to the QED picture, we calculate the total charge Q^{vac} of the vacuum. We do this by using the number operator $\hat{N}$ of the previous section and multiply the number of electrons found by the electronic charge:

$$Q^{\text{vac}} = -e\langle vac|\hat{N}|vac\rangle = -e\sum_{p}\langle vac|a_p{}^{\dagger}a_p \ + \tilde{a}_p\, \tilde{a}_p{}^{\dagger}|vac\rangle. \tag{5.33}$$

Only the last term contributes, but this yields an infinite charge, consistent with all negative-energy states being filled. However, if we use the normal-ordered

---

[2] We use the term "QED Hamiltonian" rather loosely to mean a Hamiltonian that is more or less consistent with the ideas of QED, recognizing that, strictly speaking, it is difficult if not impossible to define a QED Hamiltonian.

operator $:\hat{N}:$ we get

$$Q^{\text{vac}} = -e\langle vac| :\hat{N}: |vac\rangle = -e\sum_p \langle vac|a_p{}^\dagger a_p \;- \tilde{a}_p{}^\dagger \tilde{a}_p \;|vac\rangle = 0, \tag{5.34}$$

an uncharged vacuum. The normal-ordered number operator also gives the correct charge of a state with a given number of electrons or positrons.

5.3 Definition of the Vacuum

The subtraction of the vacuum expectation value from the Hamiltonian makes the operator vacuum-dependent. The argument for this proposition may be developed as follows.

If we have chosen a basis where the Hamiltonian matrix is diagonal, the terms that create and annihilate pairs vanish, because the matrix elements for these terms contain one electron and one positron index. We are left with an operator that conserves particle number. It is in *this* basis that the interpretation of particles is made. A basis change in Fock space may be effected by the unitary exponential operator (see for example Helgaker et al. 2000)

$$\hat{U} = e^{i\hat{\lambda}} \tag{5.35}$$

where $\hat{\lambda}$ is a Hermitian operator,

$$\hat{\lambda} = \sum_{rs} \lambda_{rs} a_r{}^\dagger a_s \;= \hat{\lambda}^\dagger. \tag{5.36}$$

It is easy to show that $\hat{U}$ is indeed Hermitian:

$$\hat{U}^\dagger \hat{U} = e^{-i\hat{\lambda}^\dagger} e^{i\hat{\lambda}} = e^{-i\hat{\lambda}+i\hat{\lambda}} = 1. \tag{5.37}$$

The reinterpreted form of $\hat{\lambda}$ is

$$\hat{\lambda} = \sum_{pq} \left[\lambda_{pq}\; a_p{}^\dagger a_q \;+ \lambda_{p\tilde{q}}\; a_p{}^\dagger \tilde{a}_q{}^\dagger + \lambda_{\tilde{p}q}\; \tilde{a}_p\; a_q \;+ \lambda_{\tilde{p}\tilde{q}}\; \tilde{a}_q\; \tilde{a}_p{}^\dagger\right] \tag{5.38}$$

which contains pair creation and destruction terms. Exponentiating this operator gives an operator containing terms that create one, two, ... pairs, and normal-ordering does not affect these terms. The vacuum state in the new basis, $|vac'\rangle$, is derived by operating on the vacuum in the old basis with this exponential operator,

$$|vac'\rangle = \hat{U}|vac\rangle. \tag{5.39}$$

Therefore, the new vacuum contains a whole series of pair states from the old basis. The structure of the vacuum has changed. This may be seen by evaluating the number of "old" electrons in the new vacuum,

$$\langle vac'|\hat{N}^e|vac'\rangle = \langle vac|\hat{U}^\dagger \sum_p a_p{}^\dagger a_p \; \hat{U}|vac\rangle. \tag{5.40}$$

By expanding $\hat{U}$ to first order, for example, it is fairly simple to verify that there will be terms that have a nonzero vacuum expectation value, such as

$$\begin{aligned}
&\langle vac| \sum_{rs} \lambda^*_{\tilde{r}s}\tilde{a}_r \; a_s \; \sum_p a_p{}^\dagger a_p \; \sum_{vw} \lambda_{v\tilde{w}} a_v{}^\dagger \tilde{a}_w{}^\dagger |vac\rangle \\
&= \sum_p \sum_{rs} \sum_{vw} \lambda^*_{\tilde{r}s}\lambda_{v\tilde{w}}\delta_{ps}\delta_{pv}\delta_{rw} \\
&= \sum_{pr} \lambda^*_{\tilde{r}p}\lambda_{p\tilde{r}}.
\end{aligned} \tag{5.41}$$

The number of electrons is, however, equal to the number of positrons, so that charge is conserved.

With respect to the old vacuum, the new vacuum is called a *dressed* vacuum and the states in it are called *dressed* states. The new particles are called *quasiparticles* because they are a composite of an "old" particle and a series of "old" pairs: they are dressed with a series of pairs rather than being bare particles. The new vacuum is also called a *polarized vacuum* because with respect to the old vacuum there is a charge polarization expressed by the presence of undressed pairs. The concept of vacuum polarization will be discussed more later. Finally, another way of looking at the change in the vacuum is that because the transformation mixes positron creation and electron annihilation operators, a normal-ordered operator in the new basis will definitely not be normal-ordered in the old basis.

So far, no prescription has been given for choosing a representation of the vacuum. If the structure of the vacuum changes between representations, the energy of the vacuum may also change. It is of course zero in the chosen representation, but the energy of the new vacuum with respect to the old Hamiltonian operator will not necessarily be zero. For example, the vacuum in the presence of two nuclei will depend on the distance between the nuclei.

In the calculations of relativistic atomic physics, the *bare* or free-particle vacuum is chosen as a reference. The energy of any other vacuum with respect to this vacuum can be evaluated in the chosen basis set, removing the potential problems of a "floating" vacuum. This brings with it some complications, because a simple-minded evaluation of the energy of a new, polarized vacuum state is actually infinite, and it is necessary to introduce the concept of renormalization in order to remove the singular terms. Such considerations are beyond the scope of this book. Instead, we will work with the "floating" vacuum, in which the vacuum is defined by the current set of one-particle states (Mittleman 1981).

5.4 The Electron–Electron Interaction

From the discussion of classical relativistic electromagnetic interactions in chapter 3, we know that the Coulomb interaction between charged particles is not Lorentz invariant, and that we need to take into account the finite transmission speed of the signal between particles (the retardation). If we make the replacement $\mathbf{u} = c\boldsymbol{\alpha}$ in (3.84), we get an interaction for low velocities that might well be suitable for relativistic quantum chemistry:

$$V_{12} = \frac{q_1 q_2}{4\pi\epsilon_0}\left[\frac{1}{r_{12}} - \frac{\boldsymbol{\alpha}_1 \cdot \boldsymbol{\alpha}_2}{2r_{12}} - \frac{1}{2}\frac{(\boldsymbol{\alpha}_1 \cdot \mathbf{r}_{12})(\boldsymbol{\alpha}_2 \cdot \mathbf{r}_{12})}{r_{12}^3}\right] + \mathcal{O}(c^{-4}). \tag{5.42}$$

However, this interaction is not retarded and is definitely not Lorentz invariant. It might be suitable, but we must at least start with an interaction that is Lorentz invariant.

As foreshadowed earlier, the interaction of charged particles in a proper relativistic theory must propagate at the speed of light. In QED, the procedure by which this occurs is the emission and absorption of "virtual" photons. The derivation of the Hamiltonian for the interaction is obtained from perturbation theory. This is usually expressed in terms of Feynman diagrams, which are a pictorial way of representing the interaction as a function of time. Here we will use these diagrams mainly for illustrating the processes, but they are also a powerful tool in their own right for the development of many-body theories.

Feynman diagrams (figure 5.1) are written with time on the vertical axis. Straight lines represent electrons. The direction of propagation is given by an arrow: an electron that is propagating backwards in time is interpreted as a positron. Squiggly lines represent photons. The number of particles at any given time can be found by drawing a horizontal line across the diagram. In some cases it is useful to distinguish between bare particles (i.e., free particles) and dressed particles. The dressed particles appear with a double line. Similar diagrams may be familiar to our readers from their use in many-body perturbation theory, where they are applied to the calculation and analysis of correlation energy between electrons (see for example Jørgensen and Simons 1981).

There are three possible diagrams for the exchange of one photon, as shown in figure 5.1(a)–(c). The first is the "normal" interaction, where a particle interacts with another particle by exchange of a photon. The second incorporates the creation of an electron–positron pair (the closed loop). This represents the interaction of the electron with the vacuum, and is known as *vacuum polarization*. In the third diagram, the electron is interacting with itself by emitting a photon and then reabsorbing it at a later time. This is called the *self-energy* term.

The diagrams for two-photon exchange are more numerous and complicated. A few of these are also given in figure 5.1(d)–(f). The first is a "ladder" diagram: two electrons exchange a photon at one time, and then exchange another a bit later. These ladder diagrams can be summed to infinite order by incorporating the one-photon interaction into a variational procedure. Diagrams of this type are called *reducible*, because they can be split into two lower-order diagrams by drawing a horizontal line across the graph between the two parts. The second is a diagram in which a pair is created and annihilated. One electron emits a photon, which creates a pair. The positron from this

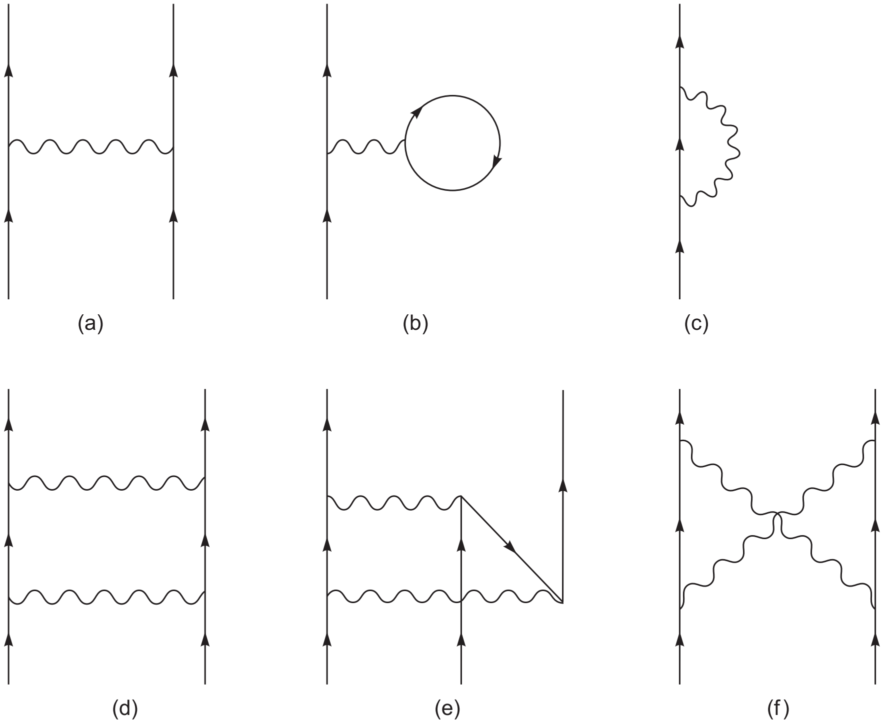


Figure 5.1 Feynman diagrams for (a)–(c) one-photon and (d)–(f) some two-photon exchanges

pair annihilates the second electron, creating a photon that is reabsorbed by the first electron. In the third diagram, two electrons simultaneously emit a photon each, and later each absorbs the photon from the other electron. Both of these are irreducible.

The development of a Lorentz-invariant interaction is a lengthy procedure, involving time-dependent perturbation theory. It is given, for example, in Berestetskii et al. (1982). Here we cite results that correspond to the first diagram, that of a single photon exchange.

The potential for the interaction between electron i and electron j is derived in the Feynman (or Lorentz) gauge. It is given in Hartree atomic units by

$$V^{\mathrm{F}}(\omega, r_{ij}) = \frac{1 - \boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j}{r_{ij}} e^{i|\omega|r_{ij}/c} \tag{5.43}$$

where ω is the frequency of the exchanged photon. Although the results should be independent of the gauge, the Coulomb gauge is usually preferred in calculations. The Coulomb gauge form of the interaction is

$$V^{\mathrm{C}}(\omega, r_{ij}) = V^{\mathrm{F}}(\omega, r_{ij}) + \boldsymbol{\alpha}_i \cdot \left\{\nabla_i \boldsymbol{\alpha}_j \cdot \left[\nabla_j f(|\omega|/c, r_{ij})\right]\right\} \tag{5.44}$$

where the function $f(\eta, r)$ is

$$f(\eta, r) = \frac{1 + i\eta r - e^{i\eta r}}{\eta^2 r}. \tag{5.45}$$

The additional term in the Coulomb-gauge interaction is properly called the *gauge term* because it arises from the gauge transformation of the four-potential from the Feynman gauge to the Coulomb gauge. The interaction in either gauge has a real and an imaginary part. The imaginary part gives the energy levels a width, and is often neglected to leave a real interaction operator.

The Lorentz-invariant electron–electron interaction operator just presented contains a photon frequency, which is given by the conservation of energy for each of the particles in the interaction. For an interaction given by the integral

$$\int \mathrm{d}\mathbf{r}_i \int \mathrm{d}\mathbf{r}_j \psi_p{}^\dagger(\mathbf{r}_i)\psi_q(\mathbf{r}_i)V(\omega, r_{ij})\psi_r{}^\dagger(\mathbf{r}_j)\psi_s(\mathbf{r}_j) \tag{5.46}$$

we have

$$\omega = \epsilon_s - \epsilon_r = \epsilon_q - \epsilon_p \tag{5.47}$$

where the ϵ are the one-particle eigenvalues. This expression is of course only appropriate for contributions to the diagonal of the Hamiltonian matrix where there is energy conservation, but it is sufficient for the present purpose. In general, each matrix element will have different photon frequencies. The one-particle functions here are the *molecular* functions, so each molecular integral will have a different photon frequency, requiring N^4 integrals over basis functions to be evaluated. The calculation of integrals becomes an enormous task, scaling as N^8. For molecules this would be impractical, and we would therefore like to have some justifiable approximation that did not require so much work. We must therefore abandon strict Lorentz invariance, but with care the lack of invariance will not significantly affect our calculations.

The oscillatory factor in (5.43), which arises from the retardation of the interaction due to its finite speed of transmission, contributes only if $\omega r_{ij}/c$ is large. That is, the factor will only be significantly different from one when the eigenvalue differences are large or the distances are large, or both. For the direct interactions, that is, $p = q$ and $r = s$, the frequency is zero, and therefore the retardation is exclusively an exchange/correlation effect. The eigenvalue differences are large only for core–valence (and core–core) interactions. The densities only have an appreciable magnitude for one-center terms, for which the distance between the electrons is small, and therefore the retardation correction will be small. In the valence shell, where most of the significant molecular effects originate, eigenvalue differences rarely exceed a few hartrees, so that ω/c will be of order 10^{-2}. The average distance between the electrons must then be $10a_0$ for a 0.5% reduction in the magnitude of the interaction, and $100a_0$ for a 50% reduction. For these distances, the magnitude of the interaction will be small because of the inverse power of the distance, and the retardation correction will again be small.

There is one circumstance in which the retardation could be significant, and that is in dispersion interactions, which are generally fairly weak. Here it is possible to have one-center densities in the exchange integrals, and the distances can be large. For the He dimer, for example, the dependence of the dispersion energy on the internuclear separation R changes from $1/R^6$ to $1/R^7$ due to retardation (Jamieson et al. 1995).

Given the small magnitude of the retardation correction in situations of chemical interest, it is a reasonable approximation to let $\omega/c \to 0$ in (5.43) and (5.44), to give

$$V^{\mathrm{F}}(0, r_{ij}) = \frac{1 - \boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j}{r_{ij}} \tag{5.48}$$

$$V^{\mathrm{C}}(0, r_{ij}) = \frac{1 - \boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j}{r_{ij}} + \frac{(\boldsymbol{\alpha}_i \times \mathbf{r}_{ij}) \cdot (\boldsymbol{\alpha}_j \times \mathbf{r}_{ij})}{2r_{ij}^3}. \tag{5.49}$$

The last term in V^{C} is the gauge term. The addition to the Coulomb interaction in the Feynman gauge is called the *Gaunt interaction*, and in the Coulomb gauge it is the *Breit interaction*.

Note that if one is interested in pair-creation effects, such as in the second of the two-photon diagrams, the frequency will always be of order $2c^2$, and therefore the use of the full frequency-dependent interaction is mandatory. In quantum chemistry we are not concerned with such effects.

An alternative form of the Breit interaction is obtained by using relations for the scalar product of two vector products,

$$V^{\mathrm{C}}(0, r_{ij}) = \frac{1}{r_{ij}} - \frac{1}{2r_{ij}} \left[\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j + \frac{(\boldsymbol{\alpha}_i \cdot \mathbf{r}_{ij})(\boldsymbol{\alpha}_j \cdot \mathbf{r}_{ij})}{r_{ij}^2} \right]. \tag{5.50}$$

This is precisely the interaction we obtained in (5.42) by substitution of $c\boldsymbol{\alpha}$ for $\mathbf{u}$ in the classical interaction expanded in powers of $1/c$. In some places, the last term in square brackets is referred to as the retardation correction, but this is not correct because the whole term in square brackets can be derived from an unretarded classical interaction. Moreover, the retardation is expressed here by the finite photon frequency, which does not contribute to this interaction. It is better to describe the Coulomb gauge interaction as a sum of the Coulomb interaction, a current–current interaction, and a gauge term.

One further point should be made about the Feynman and Coulomb gauge interactions. It has been shown that the Coulomb–Breit interaction is correct to $\mathcal{O}(c^{-2})$. The Coulomb–Gaunt interaction is not correct to this order, and does not become correct by summing ladder diagrams. The missing term comes from the crossed two-photon exchange diagram shown in figure 5.1. Thus, in a situation where one is interested in an accurate calculation of low-order effects, the Coulomb gauge interaction is preferable.

The operator presented above is essentially the first-quantized operator. The second-quantized form is

$$\hat{g} = \sum_{pqrs} a_p{}^{\dagger} a_r{}^{\dagger} a_s \, a_q \, (pq|rs) \tag{5.51}$$

where the integral is defined as

$$(pq|rs) = \int d^3r_i \int d^3r_j \psi_p{}^\dagger(i)\psi_r{}^\dagger(j)V(i,j)\psi_s(j)\psi_q(i). \tag{5.52}$$

Finally, it should be emphasized that we must use the QED reinterpretation of the Dirac one-particle states when we introduce the electron–electron interaction, or we will again be running the risk of having a bound state decay into the negative continuum. Consider two noninteracting bound one-particle states, each containing one electron. These states are degenerate with an infinity of states in which one particle is excited into the positive continuum while the other de-excites into the negative continuum. Turn on the interaction, and the system dissolves into the two continua. This problem is called *continuum dissolution*, or the *Brown–Ravenhall disease* (Brown and Ravenhall 1951). The possibility of continuum dissolution has been cited as a reason that the Breit interaction should only be used in perturbation theory. This argument is inconsistent, for several reasons. First, we only need the Coulomb interaction for continuum dissolution to occur: to single out the Breit term as the cause of the trouble is erroneous. Second, the Coulomb and the Breit terms both arise from the same fully retarded interaction in the zero frequency limit, and should therefore be treated on the same footing. Third, the interaction between the particles is mediated by the emission and absorption of photons, and, as we have seen already, this causes problems even for one particle. In order to avoid problems of collapse, it is necessary to employ the reinterpretation of QED, even for a single particle.

5.5 The Lamb Shift

The retarded electron–electron interaction presented above arises from the first of the one-photon Feynman diagrams in figure 5.1. In terms of an expansion of the relativistic interactions in powers of $1/c$, this interaction contains the lowest-order terms. As pointed out above, the Breit interaction contains all terms of order c^{-2}. After the Breit interaction, the lowest-order interactions come from the other two one-photon diagrams, the vacuum polarization and self-energy terms, which are $\mathcal{O}(c^{-3})$. The energy contribution from these two terms is called the *Lamb shift*, after its discoverer W. E. Lamb Jr. (1952), and its calculation has been an important testing ground for QED theories.

Since the classical self-energy of an electron is infinite, and the vacuum is infinitely charged in the Dirac picture, it is not surprising that these terms are difficult to handle. The operators derived for the self-energy and the vacuum polarization have divergences in the photon frequency, either at large or at small momenta. The divergences are removed by redefining the mass and charge of the electron, a process called *renormalization*. As the mass and charge measured experimentally incorporate these terms, but in the Dirac equation they do not, there is an extra term to be included in the Hamiltonian that arises from the difference between the measured and the bare charge and mass.

The evaluation of the vacuum polarization term is the most straightforward, and can be put in the form of a series of potentials. The lowest-order potential is the

Uehling potential, which provides the contribution of $\mathcal{O}(c^{-3})$. For a point nucleus, the Uehling potential consists of a multiplicative correction to the nuclear Coulomb potential that is independent of the nuclear charge. The evaluation of the self-energy is a complicated procedure and involves summations over partial wave expansions for atoms; only recently has it become possible to obtain values for one-electron functions of general form. Extensive tabulations now exist for one-electron atoms for a range of principal and angular quantum numbers. In quantum chemistry we will generally not need to be concerned with the Lamb shift, unless we are looking for high accuracy in the energies. Then, we may be able to get an estimate that is accurate enough for chemical purposes, for example by a perturbative treatment of the total Lamb shift to lowest order (Pyykkö et al. 2001).

6

Relativistic Symmetry

Symmetry is one of the great unifying principles of physics and chemistry. Symmetry governs the geometry of molecules, as well as their spectra and the shapes of their crystals. The symmetry of a system provides fingerprints that show up under various probes of the system. Knowledge of the symmetry of a system provides us with valuable information about its properties and behavior, and symmetry may also be used to simplify quantum chemical calculations on the system. All this—provided we have the theory and tools to extract this information. The purpose of this chapter is to provide some of the tools to deal with the symmetries of relativistic quantum chemistry and examine the symmetries of the Dirac equation.

Most chemists have a good working knowledge of the finite point group symmetry of molecules. Crystallographers and solid state chemists are also familiar with the symmetries of extended and infinite systems. This is usually accompanied by some knowledge of the mathematical discipline of group theory, which provides a powerful tool for the theoretical treatment of symmetry. Group theory in itself is a vast subject, and only a small fraction of it is actively used in physics, and even less in chemistry. For this presentation our selection will be aimed at those aspects of group theory immediately connected to relativistic quantum chemistry either as a foundation for the theory or in its application and interpretation. We will assume that the reader has a solid background in the finite point group symmetry of nonrelativistic quantum chemistry, as well as a familiarity with group theory at the level of a good advanced undergraduate course (e.g., Atkins and Friedman 1997). For completeness and ease of reference, some of the basic features of group theory commonly encountered by chemists are reviewed in appendix C. Group theoretical concepts beyond this level will be introduced in the text as they are needed.

In this chapter, we start with a qualitative introduction of double groups and relativistic molecular symmetries, connecting to the more or less phenomenological introduction this subject is frequently accorded in quantum chemistry. The aim is to provide the necessary insight for those who only need an operational familiarity with double groups. We then turn to a more formal discussion of the symmetry invariance of the Dirac equation, and in particular the role of the two infinite continuous Lie groups $SU(2)$ and $SO(3)$—the groups that describe rotations of spin and spatial coordinates. While our aim is to introduce the central features of this theory in a coherent, consistent, and comprehensive manner, our target audience is still the practicing quantum chemist, and thus a completely rigorous mathematical treatment would carry us far beyond the scope of this book. Fortunately, there exists a large number of books dealing with group and representation theory at any desired level of rigor, and those readers who need to go beyond our treatment are referred to this extensive literature. The references at the end of the book represent only a small fraction of this material, consisting primarily of texts that we ourselves have found useful.

6.1 The Symmetry of the Relativistic One-Electron Atom

The area where chemists normally first meet relativistic effects is in the discussion of the spectra of atoms and molecules. This is an example of how knowledge of the symmetry of a system enables us to make a priori predictions about its properties. The simplest case here is atomic spectra, in particular for one-electron atoms, so let us use that as a starting point.

We will discuss the Dirac equation for the one-electron atom in more detail in chapter 7. Here we are only interested in the symmetry properties of the Hamiltonian for such systems. We know that for the corresponding nonrelativistic case, angular momentum and spin are *normal constants of motion*, represented by operators that commute with the Hamiltonian. In particular

$$[\hat{\mathcal{H}}^{\mathrm{NR}}, \boldsymbol{\ell}] = 0; \qquad [\hat{\mathcal{H}}^{\mathrm{NR}}, \mathbf{s}] = 0 \tag{6.1}$$

and we know that the angular momentum is a normal constant of motion due to the spherical potential. The Dirac Hamiltonian for a one-electron atom has the form

$$\hat{\mathcal{H}} = c\boldsymbol{\alpha} \cdot \mathbf{p} + \beta mc^2 - e\phi \tag{6.2}$$

where $-e\phi$ is the same spherical potential as in the nonrelativistic case. Any change in normal constants of motion must therefore be due to the change in the kinetic energy part of the operator. For two-component quantities it is easy to show that

$$[\boldsymbol{\sigma} \cdot \mathbf{p}, \ell_z] = -i\hbar(\sigma_x p_y - \sigma_y p_x); \qquad [\boldsymbol{\sigma} \cdot \mathbf{p}, \sigma_z] = 2i(\sigma_x p_y - \sigma_y p_x) \tag{6.3}$$

and while neither of these commutators vanish, we see that a new operator $\boldsymbol{j}$ defined as

$$\boldsymbol{j} = \boldsymbol{\ell} + \frac{\hbar}{2}\boldsymbol{\sigma} = \boldsymbol{\ell} + \mathbf{s} \tag{6.4}$$

has the desired commutation property

$$[\boldsymbol{\sigma} \cdot \mathbf{p}, j_z] = 0. \tag{6.5}$$

The extension to the case of the four-component Dirac Hamiltonian above follows readily by noting that the spin operator and the orbital angular momentum operator for this case are

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\sigma} & \mathbf{0}_2 \\ \mathbf{0}_2 & \boldsymbol{\sigma} \end{pmatrix}, \qquad \mathbf{L} = \boldsymbol{\ell}\mathbf{I}_4, \tag{6.6}$$

which may be shown to commute with the β matrix.

The operator $\boldsymbol{j}$ may already be familiar from the treatment of spin–orbit coupling in nonrelativistic systems. For the one-electron atom with nuclear charge Z we would write the Hamiltonian as

$$\begin{aligned} \hat{\mathcal{H}} &= -\frac{\hbar^2}{2m}\nabla^2 + \frac{Z}{r} + \xi(r)\mathbf{s}\cdot\boldsymbol{\ell} \\ &= \hat{\mathcal{H}}_0^{\mathrm{NR}} + \hat{\mathcal{H}}^{\mathrm{so}} \end{aligned} \tag{6.7}$$

where $\hat{\mathcal{H}}^{\mathrm{so}}$ is a phenomenological effective spin–orbit interaction Hamiltonian. Again it may be shown that while neither $\mathbf{s}$ nor $\boldsymbol{\ell}$ are normal constants of motion for this Hamiltonian, $\boldsymbol{j}$ is. In the many-electron case, this leads to the *jj*-coupling scheme, which is important for describing cases where spin–orbit coupling dominates over electron–electron interactions.

We can show that $\boldsymbol{j}$ has all the properties of angular momentum operators. For orbital angular momentum, $\boldsymbol{\ell}$, the effect of a rotation by an angle ϕ about the z axis is (Brink and Satchler 1993)

$$\hat{R}_z(\phi)|\ell, m_\ell\rangle = e^{im_\ell\phi}|\ell, m_\ell\rangle \tag{6.8}$$

where $|\ell, m_\ell\rangle$ is an eigenfunction of $\boldsymbol{\ell}^2$ and ℓ_z. Similarly, we expect the effect of a rotation on an eigenfunction $|j, m_j\rangle$ of $\boldsymbol{j}$ to be

$$\hat{R}_z(\phi)|j, m_j\rangle = e^{im_j\phi}|j, m_j\rangle. \tag{6.9}$$

In particular, if we rotate through an angle of 2π, we get

$$\hat{R}_z(2\pi)|j, m_j\rangle = e^{im_j 2\pi}|j, m_j\rangle = (-1)^{2m_j}|j, m_j\rangle. \tag{6.10}$$

For orbital angular momentum, this result presents no problems because m_ℓ is always an integer, and thus a rotation by 2π amounts to an identity operation. However, m_j may assume half-integer values and a rotation by 2π changes the sign of the function. For example, applying this operation to the simplest half-integer function $|\frac{1}{2}, \frac{1}{2}\rangle$ yields

$$\hat{R}_z(2\pi)|\tfrac{1}{2}, \tfrac{1}{2}\rangle = -|\tfrac{1}{2}, \tfrac{1}{2}\rangle. \tag{6.11}$$

This is what we would get for the rotation of a $1s\alpha$ hydrogenic function, for instance. This function has a symmetric spatial part, and so the antisymmetry must be associated with the spin part.

In general, for functions describing systems with half-integer spin, the function must change sign under a rotation by 2π. In particular this will be the case for one-particle functions of particles with half-integer spin, *fermions*. This operation is given a special symbol, $\bar{E}$, and is interpreted as rotation by an angle 2π around an arbitrary axis.

6.2 Double Groups

In order to find a normal constant of motion for a one-electron atom we have combined the spin and orbital angular momentum of an electron into a total angular momentum $\hat{\boldsymbol{j}}$. This has profound consequences for the description of the symmetry of atomic systems.

For molecules described in a nonrelativistic formulation, the symmetry is described by finite point groups, none of which contain the $\bar{E}$ operation. Any symmetry operation of a molecular point group may be composed of rotations and inversion. Thus, we would also expect these groups to require the addition of $\bar{E}$ in order to describe the relativistic symmetry of molecules. To be precise, the relativistic group must be the direct product of the nonrelativistic group and the group $\{E, \bar{E}\}$. This has the effect that the nonrelativistic group of order n, $\{g_i;\ i = 1, n\}$ is expanded to order $2n$ by adding the operations $\{g_{n+i} = g_i\bar{E};\ i = 1, n\}$, and the new group is called the *double group*.

The great orthogonality theorem must still hold for the double group, as must the various orthogonalities between rows and columns of the character table. Also, the sum of squares of the dimensions of the irreducible representations (irreps) must equal the order of the group. If the nonrelativistic group has k_1 irreps with dimensions $d_i, i = 1, \ldots, k_1$, then we must have

$$\sum_{i=1}^{k_1} d_i^2 = n. \tag{6.12}$$

The double group retains these irreps to describe basis functions that do not change sign under $\bar{E}$, and these will appear as the first k_1 irreps of the double group. Thus, expansion of the group just adds k_2 new irreps such that

$$\sum_{i=1}^{k_2} d_{k_1+i}^2 = n. \tag{6.13}$$

We will consider three examples of double groups and their character tables. The simplest of these is for the molecule with no spatial symmetry, which belongs to the point group C_1. Including spin symmetry and the $\bar{E}$ operation[1], we get the C_1^* double

---

[1] We consistently try to use "hats", for example $\hat{\omega}$, for operators throughout the text. However, hats are conventionally not used with symmetry operations in the context of character tables.

group, which has the character table

C_1^* character table

Symbol	E	$\bar{E}$
A	1	1
$B_{1/2}$	1	-1

While it is conventional to use E to denote a rotation of 0 or 2π around an arbitrary axis, and this use is appropriate for the single groups, here we must restrict it to rotations of 0 or 4π, and 2π rotations are covered by $\bar{E}$. This usage applies to all subsequent discussion of symmetry operations.

Our second example is for a system that has an axis of twofold rotation, denoted C_2, as its only spatial symmetry. This system belongs to the point group C_2, which has the elements E and C_2. The character table for C_2 is:

C_2 character table

Symbol	E	C_2
A	1	1
B	1	-1

The double group C_2^* has the additional elements $\bar{E}$ and $C_2\bar{E} = \bar{C}_2$. The character table for the double group is easily found to be[2]

C_2^* character table

Symbol	E	C_2	$\bar{E}$	$\bar{C}_2$
A	1	1	1	1
B	1	-1	1	-1
$E_{1/2}$	1	i	-1	$-i$
$E_{-1/2}$	1	$-i$	-1	i

The C_2^* group only has one-dimensional irreps. Given that there are already two one-dimensional irreps in the group C_2, the extra irreps added in C_2^* must also be one-dimensional in order to have the sum of the squares of the irrep dimensions equal the order of the group. In this case the number of classes equals the order of the group, and the group is Abelian. The imaginary characters for C_2 in the additional irreps follow from the fact that $C_2C_2 = \bar{E}$, which has the character -1 in these representations.

As our final example of the expansion of a regular finite point group to a double group, we consider D_2, which has the elements E, C_{2x}, C_{2y}, and C_{2z}, which are

---

[2] The notation from Altmann and Herzig (1994) is ${}^1E_{1/2}$ and ${}^2E_{1/2}$ for the fermion irreps. However, we prefer the labels $E_{1/2}$ and $E_{-1/2}$, which display the connection with the m_j quantum number (see appendix D).

the identity and twofold rotations about the x, y, and z axes. D_2 has the character table

D_2 character table

Symbol	E	C_{2z}	C_{2y}	C_{2x}
A_1	1	1	1	1
B_1	1	1	−1	−1
B_2	1	−1	1	−1
B_3	1	−1	−1	1

The double group has the additional operations $\bar{E}$, $\bar{C}_{2x} = C_{2x}\bar{E}$, $\bar{C}_{2y} = C_{2y}\bar{E}$, $\bar{C}_{2z} = C_{2z}\bar{E}$, and has order 8. Construction of the character table can be done from general group theoretical considerations. The additional irreps must be either four irreps of dimension 1 or one irrep of dimension 2. In the first case the total number of irreps is 8, there must be 8 classes, every element is in a class by itself, the group is Abelian, and the elements must commute. For any of these twofold rotations we can show that

$$C_2C_2 = \bar{E}, \quad C_2\bar{C}_2 = E. \tag{6.14}$$

The product of two rotations about different axes must be one of the rotations about the third axis. In particular

$$C_{2x}C_{2y} = \tilde{C}_{2z} \tag{6.15}$$

where $\tilde{C}_{2z}$ may be C_{2z} or $\bar{C}_{2z}$. We must also have

$$C_{2y}C_{2x} = \tilde{C}'_{2z} \tag{6.16}$$

where $\tilde{C}'_{2z}$ may or may not be equal to $\tilde{C}_{2z}$. If the group is Abelian, we must have equality, because elements commute in Abelian groups. Multiplication of $\tilde{C}_{2z}$ by $\tilde{C}'_{2z}$ yields

$$\tilde{C}_{2z}\tilde{C}'_{2z} = C_{2x}C_{2y}C_{2y}C_{2x} = C_{2x}\bar{E}C_{2x} = \bar{C}_{2x}C_{2x} = E. \tag{6.17}$$

But the square of any rotation is $\bar{E}$, and

$$C_{2z}C_{2z} = \bar{C}_{2z}\bar{C}_{2z} = \bar{E}. \tag{6.18}$$

Therefore, we conclude that $\tilde{C}'_{2z}$ is not equal to $\tilde{C}_{2z}$, C_{2x} and C_{2y} do not commute, the double group is not Abelian, and the expansion of the group adds one two-dimensional irrep. From the orthogonality relations, the character table is easily deduced as

D_2^* character table

Symbol	E	C_{2z}	C_{2y}	C_{2x}	$\bar{E}$	$\bar{C}_{2z}$	$\bar{C}_{2y}$	$\bar{C}_{2x}$
A_1	1	1	1	1	1	1	1	1
B_1	1	1	−1	−1	1	1	−1	−1
B_2	1	−1	1	−1	1	−1	1	−1
B_3	1	−1	−1	1	1	−1	−1	1
$E_{1/2}$	2	0	0	0	−2	0	0	0

This table does not reflect the theorem that the number of irreps must equal the number of classes. Indeed, inspection of the table shows that except for E and $\bar{E}$, the barred and unbarred operations are in the same class. A more appropriate character table is therefore

D_2^* character table

Symbol	E	$2C_{2z}$	$2C_{2y}$	$2C_{2x}$	$\bar{E}$
A_1	1	1	1	1	1
B_1	1	1	−1	−1	1
B_2	1	−1	1	−1	1
B_3	1	−1	−1	1	1
$E_{1/2}$	2	0	0	0	−2

We see here the structure of a table containing the “old” irreps from D_2, that is, those that have a positive character for the $\bar{E}$ operation, and one new irrep, $E_{1/2}$, with a negative character for $\bar{E}$. The first group of irreps is spanned by functions having integer j values, and the irreps are termed *boson* irreps, whereas irreps with a negative sign for $\bar{E}$ are spanned by functions with half-integer j values and are called *fermion* irreps. We justify this by noting that only the spin can take half-integer values and thus a half-integer j value really reflects the fermion character of the basis functions. In standard tabulations sometimes only the left half of the character table is listed (that is, for the usual nonrelativistic operations), and the other half is then given by the symmetry properties of the boson and fermion irreps. A comprehensive collection of character tables for point groups has been provided by Altmann and Herzig (1994).

The construction of these character tables illustrates how symmetries that are hidden when only the nonrelativistic case is considered become apparent when the spin space is also considered, providing new degrees of freedom. It is similar to a “flatlander” living on one side of a cube discovering the additional symmetries provided by a third dimension, expanding the world from a D_{4h} object to one of O_h symmetry. We will follow convention and adhere to standard usage of the term “double groups”, meaning groups of order $2n$, where n is the order of the nonrelativistic or “single” group. We may, however, argue that the name “double group” is really a misnomer. Rather, the “ordinary” point groups might be called “half groups” as they only tell half the story. On the other hand, there is the possibility of further hidden symmetries that might give rise to further doublings of the group. While these are not likely to affect quantum chemistry, such further symmetries play a major role in high-energy physics.

The purpose of these two sections has been to provide an easy introduction to the subject of double groups. We now turn to a more formal treatment of the symmetry of the Dirac Hamiltonian. However, the material covered above should provide sufficient background to follow our subsequent discussion of the use of symmetry in applications and implementations of the methods, and the reader mainly oriented toward these aspects may get by with skimming the remainder of this chapter at a first reading.

6.3 Spin and the *SU*(2) Group

The most obvious new feature of the Dirac equation as compared with the standard nonrelativistic Schrödinger equation is the explicit appearance of spin through the term $\boldsymbol{\alpha} \cdot \mathbf{p}$. Any spin operator trivially commutes with a spin-free Hamiltonian, but the introduction of spin-dependent terms may change this property, as demonstrated in the case of *jj*-coupling. A further scrutiny of spin symmetry is therefore a natural first step in discussing the symmetry of the Dirac Hamiltonian. This requires a basis of spin functions on which to carry out the various operations, and a convenient choice is the familiar eigenfunctions of the s_z operator, $|\tfrac{1}{2}, \tfrac{1}{2}\rangle$ and $|\tfrac{1}{2}, -\tfrac{1}{2}\rangle$, also called the α and β spin functions.

With this choice any spin function can be expressed as a linear combination of the two basis functions. Representing these as column vectors, we have

$$|\tfrac{1}{2}, \tfrac{1}{2}\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \qquad |\tfrac{1}{2}, -\tfrac{1}{2}\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{6.19}$$

An arbitrary spin function may then be written:

$$\eta = a|\tfrac{1}{2}, \tfrac{1}{2}\rangle + b|\tfrac{1}{2}, -\tfrac{1}{2}\rangle = \begin{pmatrix} a \\ b \end{pmatrix}. \tag{6.20}$$

From our choice of eigenfunctions of s_z as a basis, it follows that

$$s_z\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \tfrac{1}{2}\begin{pmatrix} 1 \\ 0 \end{pmatrix}; \qquad s_z\begin{pmatrix} 0 \\ 1 \end{pmatrix} = -\tfrac{1}{2}\begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{6.21}$$

from which we deduce that the operator s_z in this basis corresponds to the transformation matrix

$$s_z = \tfrac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \tfrac{1}{2}\sigma_z \tag{6.22}$$

which is recognized as the σ_z Pauli matrix multiplied by $\tfrac{1}{2}$.

The familiar relations involving step operators s_+ and s_- take the form

$$\begin{aligned} s_+\begin{pmatrix}1\\0\end{pmatrix} &= \begin{pmatrix}0\\0\end{pmatrix}; \qquad & s_+\begin{pmatrix}0\\1\end{pmatrix} &= \begin{pmatrix}1\\0\end{pmatrix} \\ s_-\begin{pmatrix}1\\0\end{pmatrix} &= \begin{pmatrix}0\\1\end{pmatrix}; \qquad & s_-\begin{pmatrix}0\\1\end{pmatrix} &= \begin{pmatrix}0\\0\end{pmatrix} \end{aligned} \tag{6.23}$$

which yields the following matrix representations for the operators.

$$s_+ = \begin{pmatrix}0 & 1\\0 & 0\end{pmatrix}; \qquad s_- = \begin{pmatrix}0 & 0\\1 & 0\end{pmatrix}. \tag{6.24}$$

We may now use the relations

$$s_x = \frac{s_+ + s_-}{2}; \qquad s_y = -i\frac{s_+ + s_-}{2} \tag{6.25}$$

to find the matrix representations for s_x and s_y

$$s_x = \tfrac{1}{2}\begin{pmatrix}0 & 1\\1 & 0\end{pmatrix} = \tfrac{1}{2}\sigma_x; \qquad s_y = \tfrac{1}{2}\begin{pmatrix}0 & -i\\i & 0\end{pmatrix} = \tfrac{1}{2}\sigma_y. \tag{6.26}$$

Thus, with this standard basis, the spin operations are represented by the Pauli matrices. Any other choice of basis might have led to a different representation. We note that the basis is irreducible, a manifestation of the impossibility of determining several components of the spin simultaneously. A general spin operation $\hat{\omega}(\mathbf{s})$ may be written as

$$\hat{\omega}(\mathbf{s}) = \sigma_x a_x + \sigma_y a_y + \sigma_z a_z = \boldsymbol{\sigma} \cdot \mathbf{a} \tag{6.27}$$

or products of such linear combinations.

For the systems we are concerned with, any physically acceptable transformation in spin space must preserve scalar products and expectation values. For an operation $\hat{\Omega}$ on the scalar product of two spin functions η and ξ we must have

$$\hat{\Omega}\langle\eta|\xi\rangle = \langle\hat{\Omega}\eta|\hat{\Omega}\xi\rangle = \langle\eta|\hat{\Omega}^\dagger\hat{\Omega}|\xi\rangle = \langle\eta|\xi\rangle. \tag{6.28}$$

This is satisfied if $\hat{\Omega}^\dagger\hat{\Omega} = 1$, that is, $\hat{\Omega}$ is unitary. A general unitary operation in spin space may be written

$$\hat{u} = e^{-i\hat{\omega}(\mathbf{s})} \tag{6.29}$$

where the Hermiticity of the spin operators ensures the unitarity,

$$\hat{u}^\dagger\hat{u} = e^{i\hat{\omega}(\mathbf{s})^\dagger}e^{-i\hat{\omega}(\mathbf{s})} = 1. \tag{6.30}$$

In our standard basis a general unitary transformation will take the form

$$\mathbf{U} = e^{-\frac{i}{2}\boldsymbol{\sigma}\cdot\mathbf{a}} \tag{6.31}$$

where **a** is a real coefficient vector.

The use of matrices in an exponential may require some explanation. Basically, the matrix exponential is defined through the series expansion

$$e^{\mathbf{A}} = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{A}^k \tag{6.32}$$

and for most purposes behaves analogously to ordinary exponentials. Note however that the familiar multiplication rule

$$e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}} \tag{6.33}$$

only holds if **A** and **B** commute, $[\mathbf{A}, \mathbf{B}] = 0$. Also, the determinant of an exponential of this type is given by

$$|e^{\mathbf{A}}| = e^{Tr(\mathbf{A})} \tag{6.34}$$

where $Tr(\mathbf{A})$ is the trace of the matrix **A**. To show this, we note that if **B** is a unitary matrix, we have

$$|e^{\mathbf{A}}| = |\mathbf{B}e^{\mathbf{A}}\mathbf{B}^{-1}|. \tag{6.35}$$

By expansion of the exponential and insertion of $\mathbf{BB}^{-1}$ it is easy to show that

$$\mathbf{B}e^{\mathbf{A}}\mathbf{B}^{-1} = e^{\mathbf{BAB}^{-1}}. \tag{6.36}$$

We can choose **B** such that it diagonalizes **A**, and (6.36) certainly holds when **A** is diagonal. But the trace of a product of matrices is invariant to cyclic permutation of the matrices, and thus (6.36) also holds for general **A**.

The product of two unitary transformations is a new unitary transformation, and it is easy to show that the set of unitary transformations on the spin space defined above forms a group. Because the transformations all have determinant 1, due to the Pauli matrices being traceless, this is called the special unitary group of dimension 2, or $SU(2)$.

A general rotation by an angle ϕ around an axis along the unit vector **n** takes the form

$$\mathbf{U}_{\mathbf{n}}(\phi) = e^{-\frac{i}{2}\boldsymbol{\sigma}\cdot\mathbf{n}\phi}. \tag{6.37}$$

This may be expressed in trigonometric functions in a manner similar to the corresponding nonmatrix exponential. Using (4.15) we can show that

$$(\boldsymbol{\sigma} \cdot \mathbf{n})^2 = \mathbf{I}_2 \tag{6.38}$$

(remember **n** is a real unit vector), and therefore a term by term expansion of the exponential yields

$$e^{-\frac{i}{2}\boldsymbol{\sigma}\cdot\mathbf{n}\phi} = \mathbf{I}_2 \cos\frac{\phi}{2} - i(\boldsymbol{\sigma}\cdot\mathbf{n}) \sin\frac{\phi}{2}. \tag{6.39}$$

This is possible only due to the properties of the Pauli matrices, and will not hold for arbitrary exponents.

A major goal of the present chapter is to demonstrate how the Dirac equation transforms under operations affecting spin and spatial coordinates. Assuming that the two types of coordinates do not mix, such operations must appear as products of separate operations in the two spaces. In this section, we have built on our knowledge of spin from nonrelativistic quantum mechanics, and derived some of the characteristics of the operations in spin space. We have shown that the only physically acceptable operations on spin are carried by unitary transformations that belong to the group $SU(2)$. We have also shown that our standard basis, the α and β spin functions, spans a representation of matrix exponentials that effect the acceptable transformations in spin space. We note that we could equally well have started our discussion from the $\boldsymbol{\sigma}$ matrices of the Dirac equation and their domain and arrived at the same result, avoiding the use of the nonrelativistic ad hoc concept of spin. Nothing much would be gained by this, as we eventually would have to provide some interpretation of the domain of the $\boldsymbol{\sigma}$ matrices, leading us back to a spin concept.

6.4 Spatial Rotations and the *SO*(3) Group

We now turn from transformations and operations in spin space to the more familiar three-dimensional real Euclidean space, which describes the structure of atoms and molecules. This space is usually referred to as R_3. Molecules and atoms are finite systems, unlike solids and surfaces. Their Hamiltonians are most conveniently expressed in a body-fixed coordinate system where translational motion has been separated out, and so we will not consider translations in this treatment. We will also restrict our attention to operations that do not deform the molecule, and we will therefore be dealing with transformations that preserve length.

For a position vector **r**, length is given by the scalar product $\langle \mathbf{r}|\mathbf{r}\rangle$. The operation $\hat{\omega}$ represented by the linear transformation $\boldsymbol{\Omega}$ must satisfy

$$\langle \boldsymbol{\Omega}\mathbf{r}|\boldsymbol{\Omega}\mathbf{r}\rangle = \langle \mathbf{r}|\mathbf{r}\rangle. \tag{6.40}$$

Using the identity

$$4\langle \mathbf{r}|\mathbf{r}'\rangle = \langle \mathbf{r}+\mathbf{r}'|\mathbf{r}+\mathbf{r}'\rangle - \langle \mathbf{r}-\mathbf{r}'|\mathbf{r}-\mathbf{r}'\rangle \tag{6.41}$$

we can show, by applying $\boldsymbol{\Omega}$ to each coordinate in this expression and using (6.40) on the right-hand side, that these transformations preserve scalar products in general

$$\langle \boldsymbol{\Omega}\mathbf{r}|\boldsymbol{\Omega}\mathbf{r}'\rangle = \langle \mathbf{r}|\mathbf{r}'\rangle \tag{6.42}$$

and therefore also preserve angles.

The operations are performed on objects in R_3, and therefore the transformation $\boldsymbol{\Omega}$ must be a real 3×3 matrix. The requirement of length preservation leads to

$$\langle \boldsymbol{\Omega}\mathbf{r}|\boldsymbol{\Omega}\mathbf{r}\rangle = \langle \mathbf{r}|\boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{\Omega}|\mathbf{r}\rangle = \langle \mathbf{r}|\mathbf{r}\rangle. \tag{6.43}$$

This is satisfied if $\boldsymbol{\Omega}$ is an orthogonal matrix, that is

$$\boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{\Omega} = \mathbf{I}_3. \tag{6.44}$$

Thus, the transformations we consider here are all carried out by real, orthogonal 3×3 matrices. It is easy to show that these matrices form a group, as does the set of length-preserving operations on R_3. The two groups are isomorphic, and are both called $O(3)$—the real orthogonal group of dimension 3.

The operations used for describing molecular symmetry are: E, the identity operation; C_n, rotation of angle $2\pi/n$ about some axis; I, inversion; σ, reflection; and S_n, rotation–reflection. The σ and S_n operations may be expressed as products of rotations and the inversion, and they need not be treated separately in detail. Inversion turns out to require special consideration in connection with the Dirac equation, so for the next few sections we will only consider spatial rotations, and we will return to inversion as a separate theme later.

Rotation of a system by an angle ϕ around the z axis, which we denote $\hat{R}_z(\phi)$, is effected by the transformation matrix $\mathbf{R}_z(\phi)$. We know that the position vector $\mathbf{r}$ is transformed to $\mathbf{r}'$ according to

$$\mathbf{r}' = (x', y', z') = \hat{R}_z(\phi)\mathbf{r} = (x\ \cos\phi - y\ \sin\phi, x\ \sin\phi + y\ \cos\phi, z). \tag{6.45}$$

If this corresponds to matrix multiplication of the column vector by $\mathbf{R}_z(\phi)$

$$\mathbf{R}_z(\phi)\mathbf{r} = (x\ \cos\phi - y\ \sin\phi, x\ \sin\phi + y\ \cos\phi, z) \tag{6.46}$$

then $\mathbf{R}_z(\phi)$ must have the form

$$\mathbf{R}_z(\phi) = \begin{pmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{6.47}$$

The matrices $\mathbf{R}_x(\phi)$ and $\mathbf{R}_y(\phi)$ for rotations about the other axes are found by cyclic permutation. An infinitesimal rotation of this type may be written as

$$\mathbf{R}_z(\phi) = \mathbf{I}_3 - i\phi\mathbf{X}_z + \mathcal{O}\left(\phi^2\right) \tag{6.48}$$

with the $\mathbf{X}_z$ matrix given by

$$-i\mathbf{X}_z = \left.\frac{d\mathbf{R}_z(\phi)}{d\phi}\right|_{\phi=0} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{6.49}$$

Any finite rotation can in principle be seen as the limit of an infinite sequence of infinitesimal rotations. We may write this rotation as

$$\mathbf{R}_z(\phi) = \lim_{n\to\infty} \mathbf{R}_z\left(\frac{\phi}{n}\right)^n = \lim_{n\to\infty}\left(1 - i\mathbf{X}_z\frac{\phi}{n}\right)^n = e^{-i\mathbf{X}_z\phi}. \tag{6.50}$$

Rotations about the x and y axes may similarly be generated by the matrices

$$-i\mathbf{X}_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}; \qquad -i\mathbf{X}_y = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}. \tag{6.51}$$

The $\mathbf{X}_q$ matrices satisfy the usual cyclic commutation relations, for example

$$\left[\mathbf{X}_x, \mathbf{X}_y\right] = i\mathbf{X}_z. \tag{6.52}$$

Rotation by an angle ϕ about an axis in the direction of unit vector $\mathbf{n}$ is given by

$$\mathbf{R}_{\mathbf{n}}(\phi) = e^{-i\mathbf{X}\cdot\mathbf{n}\phi}. \tag{6.53}$$

The transformation matrices are orthogonal

$$\mathbf{R}_{\mathbf{n}}(\phi)^{\mathrm{T}}\mathbf{R}_{\mathbf{n}}(\phi) = \mathbf{I}_3 \tag{6.54}$$

and because the $\mathbf{X}_q$ matrices are traceless, the transformation matrices have unit determinants,

$$|\mathbf{R}_{\mathbf{n}}(\phi)| = 1. \tag{6.55}$$

These transformations form a subgroup of $O(3)$ called the special orthogonal group in three dimensions, $SO(3)$.

There is a connection between the groups $SU(2)$ and $SO(3)$. Recall that the operations of $SU(2)$ were represented by the unitary matrices

$$\mathbf{U}_{\mathbf{n}}(\phi) = e^{-\frac{i}{2}\sigma\cdot\mathbf{n}\phi}. \tag{6.56}$$

Comparing this with the rotations in R_3

$$\mathbf{R}_{\mathbf{n}}(\phi) = e^{-i\mathbf{X}\cdot\mathbf{n}\phi} \tag{6.57}$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_x \\ \mathbf{X}_y \\ \mathbf{X}_z \end{pmatrix} \tag{6.58}$$

we see that we can map $SU(2)$ onto $SO(3)$ by making the connection

$$\mathbf{U_n}(\phi) \rightarrow \mathbf{R_n}(\phi). \tag{6.59}$$

This mapping preserves the group properties, that is

$$\mathbf{U}_{\mathbf{n}_3}(\phi_3) = \mathbf{U}_{\mathbf{n}_1}(\phi_1)\mathbf{U}_{\mathbf{n}_2}(\phi_2) \rightarrow \mathbf{R}_{\mathbf{n}_3}(\phi_3) = \mathbf{R}_{\mathbf{n}_1}(\phi_1)\mathbf{R}_{\mathbf{n}_2}(\phi_2) \tag{6.60}$$

and is therefore a homomorphism. However, the mapping is 2 to 1. We see this by considering the spin transformations $U_\mathbf{n}(0)$ and $U_\mathbf{n}(2\pi)$, which map according to

$$\begin{aligned} \mathbf{U_n}(0) &\rightarrow \mathbf{R_n}(0) \\ \mathbf{U_n}(2\pi) &\rightarrow \mathbf{R_n}(2\pi). \end{aligned} \tag{6.61}$$

The rotation matrices for 2π and 0 are identical

$$\mathbf{R_n}(2\pi) = \mathbf{R_n}(0) \tag{6.62}$$

but the trigonometric expression in (6.39)

$$\mathbf{U_n}(\phi) = \mathbf{I}_2 \cos\frac{\phi}{2} - i(\boldsymbol{\sigma}\cdot\mathbf{n})\sin\frac{\phi}{2} \tag{6.63}$$

gives different values for $\mathbf{U_n}(0)$ and $\mathbf{U_n}(2\pi)$,

$$\mathbf{U_n}(0) = \mathbf{I}_2; \qquad \mathbf{U_n}(2\pi) = -\mathbf{I}_2. \tag{6.64}$$

Thus, both these two distinct elements of $SU(2)$ map onto the same element of $SO(3)$.

Finally, we list the trigonometric expression for $R_\mathbf{n}(\phi)$, analogous to (6.39)

$$e^{-i\mathbf{X}\cdot\mathbf{n}\phi} = \mathbf{n}\mathbf{n}^\mathrm{T} + i(\mathbf{X}\cdot\mathbf{n})\sin\phi + (\mathbf{X}\cdot\mathbf{n})^2\cos\phi \tag{6.65}$$

where $\mathbf{n}$ is a column vector and the 3 × 3 matrix $\mathbf{n}\mathbf{n}^\mathrm{T}$ has elements

$$\left(\mathbf{n}\mathbf{n}^\mathrm{T}\right)_{ij} = n_i n_j. \tag{6.66}$$

6.5 Transformation of Operators

We have thus far only discussed the effect of spin and space operations as carried out on the basis of a particular representation. Our aim is to discover the transformation properties of the Dirac Hamiltonian under these operations, and we must therefore examine the effect of the operations on operators. We first apply some operation $\hat{\omega}$ to a function ψ. If the transformed function is ψ', then we have

$$\hat{\omega}\psi = \psi'. \tag{6.67}$$

Applying the same operation $\hat{\omega}$ to the product of an operator $\hat{Q}$ and ψ we get

$$\begin{aligned} \hat{\omega}(\hat{Q}\psi) &= \hat{Q}'\psi' = \hat{Q}'\hat{\omega}\psi \\ \hat{\omega}^{-1}\hat{Q}'\hat{\omega}\psi &= \hat{Q}\psi \end{aligned} \tag{6.68}$$

and thus we conclude that the transformed operator $\hat{Q}'$ is

$$\hat{Q}' = \hat{\omega}\hat{Q}\hat{\omega}^{-1}. \tag{6.69}$$

For the case that both the operator and the function are vector quantities expressed in our standard basis, the operator $\hat{\omega}$ may be replaced by the corresponding matrix, $\boldsymbol{\Omega}$. For spin operations, the function is always expressed in our standard basis, the α and β spin functions. If the operator is also a simple vector function of spin, $\hat{Q}(\mathbf{s})$, then we have

$$\mathbf{Q}' = \boldsymbol{\Omega}\mathbf{Q}\boldsymbol{\Omega}^{-1} \tag{6.70}$$

where $\mathbf{Q}$ is the matrix representation of $\hat{Q}$.

Let us consider in particular the transformation of the $\boldsymbol{\sigma}$ vector in the role of spin operator. For a rotation by an angle ϕ around the z axis, we have

$$\begin{aligned} \boldsymbol{\sigma}' &= \mathbf{U}_z(\phi)\boldsymbol{\sigma}\mathbf{U}_z(\phi)^{-1} \\ &= e^{-\frac{i}{2}\sigma_z\phi}\boldsymbol{\sigma}e^{\frac{i}{2}\sigma_z\phi}. \end{aligned} \tag{6.71}$$

We can carry out this transformation component by component and use the trigonometric form of the transformation matrices from (6.39)

$$e^{-\frac{i}{2}\sigma_z\phi} = \mathbf{I}_2\cos\frac{\phi}{2} - i\sigma_z\sin\frac{\phi}{2}. \tag{6.72}$$

For the σ_x component we get

$$\begin{aligned}\sigma_x' &= e^{-\frac{i}{2}\sigma_z\phi}\sigma_x e^{\frac{i}{2}\sigma_z\phi}\\ &= \left(\mathbf{I}_2\cos\frac{\phi}{2} - i\sigma_z\sin\frac{\phi}{2}\right)\sigma_x\left(\mathbf{I}_2\cos\frac{\phi}{2} + i\sigma_z\sin\frac{\phi}{2}\right)\\ &= \sigma_x\cos^2\frac{\phi}{2} - i\sigma_z\sigma_x\cos\frac{\phi}{2}\sin\frac{\phi}{2} + i\sigma_x\sigma_z\cos\frac{\phi}{2}\sin\frac{\phi}{2} + \sigma_z\sigma_x\sigma_z\sin^2\frac{\phi}{2}.\end{aligned} \tag{6.73}$$

But $\sigma_z\sigma_x\sigma_z = -\sigma_x$, and

$$\begin{aligned}\sigma_x' &= \sigma_x\left(\cos^2\frac{\phi}{2} - \sin^2\frac{\phi}{2}\right) - i\left(\sigma_z\sigma_x - \sigma_x\sigma_z\right)\cos\frac{\phi}{2}\sin\frac{\phi}{2}\\ &= \sigma_x\cos\phi + \sigma_y\sin\phi\end{aligned} \tag{6.74}$$

where we have used the commutation relations for the Pauli matrices and the usual trigonometric expressions for the sine and cosine of the double angle.

Similarly

$$\sigma_y' = -\sigma_x\sin\phi + \sigma_y\cos\phi \tag{6.75}$$

while σ_z is unchanged. Thus

$$\boldsymbol{\sigma}' = (\sigma_x\cos\phi + \sigma_y\sin\phi, -\sigma_x\sin\phi + \sigma_y\cos\phi, \sigma_z) \tag{6.76}$$

which is the same as the matrix transformation

$$\boldsymbol{\sigma}' = \boldsymbol{\sigma}\begin{pmatrix}\cos\phi & -\sin\phi & 0\\ \sin\phi & \cos\phi & 0\\ 0 & 0 & 1\end{pmatrix} = \boldsymbol{\sigma}\mathbf{R}_z(\phi). \tag{6.77}$$

In fact, with considerably more algebra we can show that for the general operator $\hat{u}_{\mathbf{n}}(\phi)$ effecting a rotation by an angle ϕ around an axis $\mathbf{n}$ we have

$$\hat{u}_{\mathbf{n}}(\phi)\boldsymbol{\sigma} = \boldsymbol{\sigma}\mathbf{R}_{\mathbf{n}}(\phi). \tag{6.78}$$

We have regarded $\boldsymbol{\sigma}$ as a row vector throughout this derivation. This is purely for convenience in later application of the results.

The spin functions are expressed explicitly in terms of our standard basis in spin space. The standard basis in R_3—the coordinates x, y, z—is usually implicit in the functions and operators in some functional form. Thus we cannot insert the transformation matrices directly as we did above for the $\boldsymbol{\sigma}$ vector in spin space. There is, however, one class of operators for which we can readily derive the transformation properties. These are operators that behave like the position vector, in the sense that the components transform like the x, y, z components of the position vector $\mathbf{r}$. These operators are called *vector operators* (Tinkham 1964). We demonstrate this on the position column vector using a rotation about the **n** axis,

$$\mathbf{r}'\psi(\mathbf{r}) = \hat{R}_{\mathbf{n}}(\phi)\ \mathbf{r}\ \hat{R}_{\mathbf{n}}(\phi)^{-1}\psi(\mathbf{r}). \tag{6.79}$$

Using the "passive" interpretation, where a symmetry operation on a function is defined by carrying out the inverse operation on the coordinate system, the effect of $\hat{R}$ on a function $f(\mathbf{r})$ is given by

$$\hat{R}f(\mathbf{r}) = f(\mathbf{R}^{-1}\mathbf{r}) \tag{6.80}$$

where $\mathbf{R}$ is the matrix representation of $\hat{R}$. Using this in (6.79), we get

$$\begin{aligned}\hat{R}_{\mathbf{n}}(\phi)\ \mathbf{r}\ \hat{R}_{\mathbf{n}}(\phi)^{-1}\psi(\mathbf{r}) &= \hat{R}_{\mathbf{n}}(\phi)\ \mathbf{r}\ \psi(\mathbf{R}_{\mathbf{n}}(\phi)\mathbf{r}) \\ &= \mathbf{R}_{\mathbf{n}}(\phi)^{-1}\mathbf{r}\ \psi(\mathbf{R}_{\mathbf{n}}(\phi)^{-1}\mathbf{R}_{\mathbf{n}}(\phi)\mathbf{r}) = \mathbf{R}_{\mathbf{n}}(\phi)^{-1}\mathbf{r}\ \psi(\mathbf{r})\end{aligned} \tag{6.81}$$

and we conclude that for a vector operator

$$\hat{R}_{\mathbf{n}}(\phi)\ \mathbf{r}\ \hat{R}_{\mathbf{n}}(\phi)^{-1} = \mathbf{R}_{\mathbf{n}}(\phi)^{-1}\mathbf{r}. \tag{6.82}$$

We have here chosen to treat the vectors as column vectors, which allows us to premultiply by the transformation matrices, making a visually easier connection to the operator expressions. Corresponding expressions are, of course, obtained if we use row vectors instead.

6.6 Transformation of the Dirac Equation under *SU*(2) and *SO*(3)

We now have the necessary tools to discuss the transformation of the Dirac Hamiltonian. We consider the time-independent equation with a static scalar potential, (4.74)

$$(\hat{h}^{\mathrm{D}} + V)\psi = c(\boldsymbol{\alpha}\cdot\mathbf{p})\psi + (\beta mc^2 + V)\psi = E\psi, \tag{6.83}$$

where $\hat{h}^{\mathrm{D}} = c(\boldsymbol{\alpha}\cdot\mathbf{p}) + \beta mc^2$ is the free-particle Dirac Hamiltonian. As a further initial restriction we assume that the potential V is spherically symmetric. The general symmetry operation composed of spin rotations in $SU(2)$ and space rotations in $SO(3)$ takes the form $\hat{u}_{\mathbf{n}_1}(\phi_1)\hat{R}_{\mathbf{n}_2}(\phi_2)$ and the transformed operator is

$$(\hat{h}^{\mathrm{D}} + V)' = \hat{u}_{\mathbf{n}_1}(\phi_1)\hat{R}_{\mathbf{n}_2}(\phi_2)(\hat{h}^{\mathrm{D}} + V)\hat{R}_{\mathbf{n}_2}(\phi_2)^{-1}\hat{u}_{\mathbf{n}_1}(\phi_1)^{-1}. \tag{6.84}$$

The potential is spin-independent, and with the assumption of spherical symmetry this part of the operator obviously must be invariant under the transformation above.

The term involving β is independent of spatial coordinates. The effect of the spin transformation on β is

$$\hat{u}_{\mathbf{n}_1}(\phi_1)\beta\hat{u}_{\mathbf{n}_1}(\phi_1)^{-1} = \begin{pmatrix} \hat{u}_{\mathbf{n}_1}(\phi_1)(\mathbf{I}_2)\hat{u}_{\mathbf{n}_1}(\phi_1)^{-1} & \mathbf{0}_2 \\ \mathbf{0}_2 & \hat{u}_{\mathbf{n}_1}(\phi_1)(-\mathbf{I}_2)\hat{u}_{\mathbf{n}_1}(\phi_1)^{-1} \end{pmatrix} = \beta \tag{6.85}$$

and therefore this term is also invariant under $SU(2) \otimes SO(3)$ transformations.

It remains to consider the term $\boldsymbol{\alpha} \cdot \mathbf{p}$. We have

$$\begin{aligned} \boldsymbol{\alpha}' \cdot \mathbf{p}' &= \hat{u}_{\mathbf{n}_1}(\phi_1)\hat{R}_{\mathbf{n}_2}(\phi_2)\ \boldsymbol{\alpha} \cdot \mathbf{p}\ \hat{R}_{\mathbf{n}_2}(\phi_2)^{-1}\hat{u}_{\mathbf{n}_1}(\phi_1)^{-1} \\ &= \left[\hat{u}_{\mathbf{n}_1}(\phi_1)\ \boldsymbol{\alpha}\ \hat{u}_{\mathbf{n}_1}(\phi_1)^{-1}\right] \cdot \left[\hat{R}_{\mathbf{n}_2}(\phi_2)\ \mathbf{p}\ \hat{R}_{\mathbf{n}_2}(\phi_2)^{-1}\right]. \end{aligned} \tag{6.86}$$

The spin transformation here is the same as in the case considered in the previous section with the operator working on the four-component spinor expressed in our standard basis. Expanding $\boldsymbol{\alpha}$ in terms of $\boldsymbol{\sigma}$ yields

$$\hat{u}_{\mathbf{n}_1}(\phi_1)\alpha_k\hat{u}_{\mathbf{n}_1}(\phi_1)^{-1} = \begin{pmatrix} \mathbf{0}_2 & \hat{u}_{\mathbf{n}_1}(\phi_1)\sigma_k\hat{u}_{\mathbf{n}_1}(\phi_1)^{-1} \\ \hat{u}_{\mathbf{n}_1}(\phi_1)\sigma_k\hat{u}_{\mathbf{n}_1}(\phi_1)^{-1} & \mathbf{0}_2 \end{pmatrix} \tag{6.87}$$

and we find that

$$\hat{u}_{\mathbf{n}_1}(\phi_1)\boldsymbol{\alpha}\hat{u}_{\mathbf{n}_1}(\phi_1)^{-1} = \boldsymbol{\alpha}\mathbf{R}_{\mathbf{n}_1}(\phi_1). \tag{6.88}$$

The momentum operator $\mathbf{p}$ has the form

$$\mathbf{p} = -i\hbar\nabla. \tag{6.89}$$

This is a vector operator of the type considered in the previous section, that is, an operator that transforms as

$$\mathbf{v}' = \hat{R}_{\mathbf{n}}(\phi)\mathbf{v}\hat{R}_{\mathbf{n}}(\phi)^{-1} = \mathbf{R}_{\mathbf{n}}^{-1}(\phi)\mathbf{v}. \tag{6.90}$$

To demonstrate this, we assume that the transformation of the gradient may be written as

$$\nabla' = \hat{R}_{\mathbf{n}}(\phi)\nabla\hat{R}_{\mathbf{n}}(\phi)^{-1} = \mathbf{W}\nabla \tag{6.91}$$

where $\mathbf{W}$ is some matrix to be determined. We now recall the divergence of the position vector $\mathbf{r}$, which is

$$\nabla \cdot \mathbf{r} = \frac{\partial}{\partial x}x + \frac{\partial}{\partial y}y + \frac{\partial}{\partial z}z = 3. \tag{6.92}$$

This transforms as

$$\begin{aligned}\hat{R}_{\mathbf{n}}(\phi)\nabla\cdot\mathbf{r}\;\hat{R}_{\mathbf{n}}(\phi)^{-1} &= \left(\hat{R}_{\mathbf{n}}(\phi)\nabla\hat{R}_{\mathbf{n}}(\phi)^{-1}\right)^{\mathrm{T}}\left(\hat{R}_{\mathbf{n}}(\phi)\mathbf{r}\hat{R}_{\mathbf{n}}(\phi)^{-1}\right)\\ &= (\mathbf{W}\nabla)^{\mathrm{T}}\mathbf{R}_{\mathbf{n}}(\phi)^{-1}\mathbf{r} = \nabla^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{R}_{\mathbf{n}}(\phi)^{-1}\mathbf{r}.\end{aligned} \tag{6.93}$$

But the divergence here is just a real number that must be invariant under the transformations, thus

$$\hat{R}_{\mathbf{n}}(\phi)\nabla\cdot\mathbf{r}\hat{R}_{\mathbf{n}}(\phi)^{-1} = \nabla\cdot\mathbf{r} \quad\Rightarrow\quad \mathbf{W}^{\mathrm{T}}\mathbf{R}_{\mathbf{n}}(\phi)^{-1} = \mathbf{I}_3 \tag{6.94}$$

which is satisfied if $\mathbf{W}^{\mathrm{T}} = \mathbf{R}_{\mathbf{n}}(\phi)$, or $\mathbf{W} = \mathbf{R}_{\mathbf{n}}(\phi)^{-1}$. Then

$$\hat{R}_{\mathbf{n}}(\phi)\nabla\hat{R}_{\mathbf{n}}(\phi)^{-1} = \mathbf{R}_{\mathbf{n}}(\phi)^{-1}\nabla \tag{6.95}$$

and $\mathbf{p}$ transforms as a vector operator.

Putting all this together in the expression for the transformed $\boldsymbol{\alpha}\cdot\mathbf{p}$ term, (6.86), we get

$$\boldsymbol{\alpha}'\cdot\mathbf{p}' = \left[\boldsymbol{\alpha}\mathbf{R}_{\mathbf{n}_1}(\phi_1)\right]\left[\mathbf{R}_{\mathbf{n}_2}(\phi_2)^{-1}\mathbf{p}\right]. \tag{6.96}$$

This expression leaves the $\boldsymbol{\alpha}\cdot\mathbf{p}$ term invariant if $\mathbf{n}_1 = \mathbf{n}_2$ and $\phi_1 = \phi_2$, and we conclude that spin and space transformations of the type

$$\hat{\omega}_{\mathbf{n}}(\phi) = \hat{u}_{\mathbf{n}}(\phi)\hat{R}_{\mathbf{n}}(\phi) \tag{6.97}$$

leave $\hat{h}^{\mathrm{D}} + V$ invariant.

The extension of this result to the usual molecular Dirac Hamiltonian is easily made. In the Born–Oppenheimer approximation the Hamiltonian takes the form

$$\hat{\mathcal{H}} = \sum_i \hat{h}_i^{\mathrm{D}} + \sum_{i,N} V_{iN} + \sum_{i,j} V_{ij} \tag{6.98}$$

in obvious notation. V_{ij} is the electron–electron interaction potential, which is spherically symmetric. The pure Coulomb potential is spin independent. The Breit terms contain $\boldsymbol{\alpha}$, but these appear either in the form $\boldsymbol{\alpha}\cdot\boldsymbol{\alpha}$ or $\boldsymbol{\alpha}\cdot\mathbf{r}$, both of which are invariant under $SU(2)\otimes SO(3)$ transformations. The nuclear potential restricts the $\hat{R}_{\mathbf{n}}(\phi)$ to those $\mathbf{n}$ and ϕ that leave V_{iN} invariant. These form a subgroup, say G of $SO(3)$, and within these restrictions the molecular Dirac Hamiltonian is invariant under $SU(2)\otimes G$ operations, with the restriction that the operations in $SU(2)$ and G are "the same", for example a rotation ϕ about some axis $\mathbf{n}$ in both Euclidean space and spin space. However, $SO(3)$ does not cover inversion and neither therefore can G, which is a subgroup of $SO(3)$. So in order to have a complete discussion of all molecular symmetry operations we must now examine the effects of inversion.

6.7 Space Inversion

The effect of space inversion $\hat{I}$ is to change the sign of all spatial coordinates

$$\hat{I}f(x,y,z) = f(-x,-y,-z). \tag{6.99}$$

This corresponds to multiplication of the position vector by the matrix $-\mathbf{I}_3$

$$\hat{I}\mathbf{r} = -\mathbf{I}_3\mathbf{r} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \mathbf{r} = -\mathbf{r}. \tag{6.100}$$

$-\mathbf{I}_3$ is a linear 3×3 transformation with $(-\mathbf{I}_3)^{\mathrm{T}}(-\mathbf{I}_3) = \mathbf{I}_3$, and it belongs to the group $O(3)$. But the determinant $|-\mathbf{I}_3|$ is -1, and $-\mathbf{I}_3$ cannot therefore be a member of $SO(3)$. The inversion is also different from the rotations of $SO(3)$ in another respect: all the rotations may be conceived as derived by a continuous series of transformations starting from the identity. The inversion cannot be generated in this manner because no combination of rotations can alter the signs of an odd number of coordinates. Getting from the identity to the inversion operations therefore involves a discontinuous jump.

Our experience with inversion is related to the nonrelativistic case and operations in R_3. We can always apply the operator in the nonrelativistic form, but our aim is to find operations that leave the Dirac Hamiltonian invariant, and there is no guarantee that the nonrelativistic inversion operator will do this. However, the operator we are looking for should have the same effect as the familiar inversion when applied only to R_3. We may therefore assume that the inversion operator we are looking for may be written as the product of one part that acts only in R_3 in the manner described in (6.99) and one part that does not act in R_3, but which may affect other coordinates. We write the relativistic inversion operator as

$$\hat{I} = \hat{a}\hat{I}_R \tag{6.101}$$

where $\hat{I}_R$ is the part acting in R_3, and $\hat{a}$ contains other effects not noticed in applications on R_3 only. For the present purpose it is sufficient to consider only the free-particle Dirac Hamiltonian,

$$\hat{h}^{\mathrm{D}} = c(\boldsymbol{\alpha} \cdot \mathbf{p}) + \beta mc^2 \tag{6.102}$$

because it contains the spin-dependent terms. This transforms as

$$\hat{h}^{\mathrm{D}'} = \hat{I}\hat{h}^{\mathrm{D}}\hat{I}^{-1} = c\left(\hat{I}\boldsymbol{\alpha}\hat{I}^{-1}\right)\cdot\left(\hat{I}\mathbf{p}\hat{I}^{-1}\right) + mc^2\left(\hat{I}\beta\hat{I}^{-1}\right). \tag{6.103}$$

Here, $\mathbf{p}$ is only affected by the $\hat{I}_R$ part of the operator, and thus

$$\hat{I}\mathbf{p}\hat{I}^{-1} = \hat{I}_R\mathbf{p}\hat{I}_R^{-1} = -\mathbf{p}. \tag{6.104}$$

The other terms are unaffected by operations on the spatial coordinates, but may be affected by the operator $\hat{a}$. Therefore

$$\hat{h}^{\mathrm{D}'} = -c\left(\hat{a}\boldsymbol{\alpha}\hat{a}^{-1}\right)\cdot\mathbf{p} + mc^2\left(\hat{a}\beta\hat{a}^{-1}\right). \tag{6.105}$$

If this is to be equal to $\hat{h}^{\mathrm{D}}$, then we must have

$$\begin{aligned} \hat{a}\boldsymbol{\alpha}\hat{a}^{-1} &= -\boldsymbol{\alpha} \\ \hat{a}\beta\hat{a}^{-1} &= \beta. \end{aligned} \tag{6.106}$$

It is easy to see that setting $\hat{a} = \beta$ satisfies these equations, and we conclude that the operator representing inversion in the relativistic case is (to within an arbitrary complex phase, which we set to 1)

$$\hat{I} = \beta\hat{I}_R. \tag{6.107}$$

For the molecular Dirac Hamiltonian the symmetry of the nuclear potential will determine whether inversion leaves the operator invariant. In the absence of spin-dependent terms, this reduces to normal symmetry consideration of objects in R_3, and depends on how the nuclear potential transforms under $\hat{I}_R$.

For a relativistic Hamiltonian with inversion symmetry, it can be shown that the large and small components of the spinor have different parity. We do this by writing the spinor, ψ, as a column vector of large and small components

$$\psi = \begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix}. \tag{6.108}$$

Partitioning the expression for the inversion operator, we get

$$\hat{I} = \begin{pmatrix} \hat{I}_R & 0 \\ 0 & -\hat{I}_R \end{pmatrix}. \tag{6.109}$$

We apply this to the spinor ψ

$$\hat{I}\psi = \begin{pmatrix} \hat{I}_R & 0 \\ 0 & -\hat{I}_R \end{pmatrix}\begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} = \begin{pmatrix} \hat{I}_R\psi^L \\ -\hat{I}_R\psi^S \end{pmatrix}. \tag{6.110}$$

This demonstrates that if ψ is an eigenfunction of the inversion operator $\hat{I}$, the large and small components of the spinor must have different parity.

Our discussion of spatial inversion symmetry concludes our project of demonstrating the symmetry properties of the Dirac equation under transformations of spin and spatial coordinates. With a spherical potential, the symmetry group is $SU(2) \otimes O(3)$. For potentials of lower (e.g., molecular) symmetries the appropriate group is $SU(2)\otimes G$, where G is the nonrelativistic (spatial) point group of the potential—that is, the single group. These direct product forms show why we get the double groups, and we could have gone on from here and derived character tables for double groups from our general knowledge of $SU(2)$ and $O(3)$. As mentioned previously, this has been done by others[3], and appropriate tables may be found in Altmann and Herzig (1994).

6.8 Reflections and Rotation–Inversions

Now that we have derived the operation for space inversion, we can proceed to define the remaining operations in the finite point groups. The reflections in the single group have the form $\sigma_q = \hat{I}C_{2q}$, where q defines a twofold axis. The simplest of these are the reflections in the xy, xz, and yz planes, which change the sign of the z, y, and x coordinates. Their effect on the position vector is

$$\begin{aligned}
\sigma_{xy}\mathbf{r} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ -z \end{pmatrix}, \\
\sigma_{xz}\mathbf{r} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ -y \\ z \end{pmatrix}, \\
\sigma_{yz}\mathbf{r} &= \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -x \\ y \\ z \end{pmatrix}.
\end{aligned} \tag{6.111}$$

Like the inversion operation, the reflections are linear 3 × 3 transformation that belong to the group $O(3)$, but not to $SO(3)$ because their determinants are -1.

The double-group reflection operators are similarly formed by taking the product of the double-group inversion operator with each of the possible C_2 operations, to give two reflections,

$$\sigma_q = \beta\hat{I}C_{2q}; \qquad \bar{\sigma}_q = \beta\hat{I}\bar{C}_{2q}. \tag{6.112}$$

From the relation between C_{2q} and $\bar{C}_{2q}$ it is easy to show that $\bar{\sigma}_q = \sigma_q\bar{E}$. The presence of β indicates that the large and small components have the opposite parity under reflection.

The rotation-inversions are similarly formed by the product of an inversion and a rotation about an axis of order $2n$, $S_{2n} = \hat{I}C_{2n}$, where $n > 1$. The reflections are in fact simply a special case of rotation–inversion with $n = 1$. As for the reflections, the large and small components have the opposite parity under the rotation–inversion operations.

It is important to note that 2-spinors and 4-spinors both transform as the fermion irreps of the double group. For 2-spinors, however, there is no factor of β in the

---

[3] See the bibliographies by Pyykkö (1986, 1993, 2000) for references.

inversion-derived operations. Hence, the large and small components may transform as different irreps of the double group, when considered as 2-spinors, but the 4-spinor as a whole transforms as only one irrep. The simplest example is the group C_i, which contains only the identity and the inversion operation in the single group. The extra irreps in the double group are $B_{1/2g}$ and $B_{1/2u}$. If the large component transforms as the first, the small component must transform as the second, and vice versa. However, because of the factor of β in the inversion operation, the 4-spinor that has a $B_{1/2g}$ large component and a $B_{1/2u}$ small component transforms as $B_{1/2g}$ under the four-component operations. Thus, the 4-spinors can be labeled with the symmetry of the large component.

6.9 Time Reversal

We have previously commented on the Lorentz invariance of the Dirac equation. Considering that this places time and space coordinates on an equal footing, it may seem inconsistent to discuss transformations in spin space and R_3 only. We therefore now turn our attention to time transformations. With only one coordinate, there are only two possible transformations: translation and reversal. Translation will be treated in connection with a discussion of the Lorentz transformations in the next section. Here, we will consider the symmetry of the Dirac equation under time reversal.

Before looking at the consequences of time reversal, we need to introduce some operator concepts not normally encountered in introductory quantum chemistry texts, which deal almost exclusively with linear operators ($\hat{\mathcal{A}}$) defined by the relation

$$\hat{\mathcal{A}}\,(\lambda_1\psi_1 + \lambda_2\psi_2) = \lambda_1\hat{\mathcal{A}}\psi_1 + \lambda_2\hat{\mathcal{A}}\psi_2. \tag{6.113}$$

However, there is also a class of antilinear operators ($\hat{\mathcal{B}}$) that have the property

$$\hat{\mathcal{B}}\,(\lambda_1\psi_1 + \lambda_2\psi_2) = \lambda_1^*\hat{\mathcal{B}}\psi_1 + \lambda_2^*\hat{\mathcal{B}}\psi_2. \tag{6.114}$$

If an antilinear operator is unitary, it is called antiunitary and

$$\hat{\mathcal{B}}\hat{\mathcal{B}}^\dagger = \hat{\mathcal{B}}^\dagger\hat{\mathcal{B}} = 1. \tag{6.115}$$

As before, any linear operator transforms under an antiunitary transformation as

$$\hat{\mathcal{A}}' = \hat{\mathcal{B}}\hat{\mathcal{A}}\hat{\mathcal{B}}^\dagger, \tag{6.116}$$

and by using the defining equations for linear and antilinear operators, it is easily seen that $\hat{\mathcal{A}}'$ is linear.

Antiunitary transformations affect commutation relations between operators. For

$$\mathbf{r}' = \hat{\mathcal{B}}\mathbf{r}\hat{\mathcal{B}}^\dagger; \qquad \mathbf{p}' = \hat{\mathcal{B}}\mathbf{p}\hat{\mathcal{B}}^\dagger \tag{6.117}$$

it is easy to see that the antiunitary transformation leads to

$$[x, p_x] = i\hbar \qquad \Rightarrow \qquad \left[x', p'_x\right] = -i\hbar. \tag{6.118}$$

Similarly, for the spin

$$\left[s_x, s_y\right] = i\hbar s_z \qquad \Rightarrow \qquad \left[s'_x, s'_y\right] = -i\hbar s'_z. \tag{6.119}$$

Now let us turn to time reversal. In nonrelativistic physics, time reversal leaves positions invariant, but changes the sign of all velocities. If $\hat{\mathcal{K}}$ is the time-reversal operator, we must have

$$\hat{\mathcal{K}}\mathbf{r}\hat{\mathcal{K}}^\dagger = \mathbf{r}; \qquad \hat{\mathcal{K}}\mathbf{p}\hat{\mathcal{K}}^\dagger = -\mathbf{p}. \tag{6.120}$$

Comparing with the commutation relations above, we see that for $\mathbf{r}$ and $\mathbf{p}$ at least, $\hat{\mathcal{K}}$ has the effect of an antiunitary operator. Expressing orbital angular momentum as $\boldsymbol{\ell} = \mathbf{r} \times \mathbf{p}$, we see that $\hat{\mathcal{K}}\boldsymbol{\ell}\hat{\mathcal{K}}^\dagger = -\boldsymbol{\ell}$. For spin we can draw on the analogies between the transformation of commutation relations for spin and orbital angular momentum. From these we see that the transformed commutation relations are consistent with $\hat{\mathcal{K}}\mathbf{s}\hat{\mathcal{K}}^\dagger = -\mathbf{s}$, and the spin has the same transformation properties as orbital angular momentum.

In the nonrelativistic Hamiltonian we need not worry about spin. With $\mathbf{r}$ being real and $\mathbf{p}$ purely imaginary in our standard representation, the complex conjugation operator, $\hat{\mathcal{K}}_0$, can be used to effect time reversal. For b and d real numbers, the effect of $\hat{\mathcal{K}}_0$ is

$$\begin{aligned} \hat{\mathcal{K}}_0(b + id) &= b - id \\ \hat{\mathcal{K}}_0\mathbf{r}\hat{\mathcal{K}}_0 &= \mathbf{r} \\ \hat{\mathcal{K}}_0\mathbf{p}\hat{\mathcal{K}}_0 &= \hat{\mathcal{K}}_0\left(-i\hbar\nabla\right)\hat{\mathcal{K}}_0 = i\hbar\nabla = -\mathbf{p}. \end{aligned} \tag{6.121}$$

Extending our treatment to the two-component relativistic case, we have a situation analogous to that previously encountered for inversion: there may be parts of a total time-reversal operator that do not give explicit effects when applied in the nonrelativistic realm, but which are essential for the relativistic treatment. Taking the same approach as for inversion, we write the total time-reversal operator as

$$\hat{\mathcal{K}} = \hat{a}\hat{\mathcal{K}}_0 \tag{6.122}$$

where $\hat{a}$ is an operator to be determined. If spin changes sign under time reversal, we must have

$$\hat{\mathcal{K}}\boldsymbol{\sigma}\hat{\mathcal{K}}^\dagger = -\boldsymbol{\sigma}. \tag{6.123}$$

Expanding the spin transformation in separate components, we get

$$\begin{aligned} \hat{a}\hat{\mathcal{K}}_0\sigma_x\hat{\mathcal{K}}_0^{-1}\hat{a}^{-1} &= \hat{a}\sigma_x\hat{a}^{-1} = -\sigma_x \\ \hat{a}\hat{\mathcal{K}}_0\sigma_y\hat{\mathcal{K}}_0^{-1}\hat{a}^{-1} &= -\hat{a}\sigma_y\hat{a}^{-1} = -\sigma_y \\ \hat{a}\hat{\mathcal{K}}_0\sigma_z\hat{\mathcal{K}}_0^{-1}\hat{a}^{-1} &= \hat{a}\sigma_z\hat{a}^{-1} = -\sigma_z. \end{aligned} \tag{6.124}$$

These relations are satisfied if we take

$$\hat{a} = e^{i\eta}\sigma_y \tag{6.125}$$

where $e^{i\eta}$ is an arbitrary phase factor. Conventionally we choose this factor such that the total time-reversal operator becomes

$$\hat{\mathcal{K}} = -i\sigma_y\hat{\mathcal{K}}_0. \tag{6.126}$$

This operator is unitary

$$\hat{\mathcal{K}}^\dagger\hat{\mathcal{K}} = (\hat{\mathcal{K}}_0{}^\dagger\sigma_y{}^\dagger)(\sigma_y\hat{\mathcal{K}}_0) = \mathbf{I}_2, \tag{6.127}$$

but we note that

$$\hat{\mathcal{K}}^2 = (-i\sigma_y\hat{\mathcal{K}}_0)(-i\sigma_y\hat{\mathcal{K}}_0) = (-i\sigma_y)(i\sigma_y^*)\hat{\mathcal{K}}_0\hat{\mathcal{K}}_0 = \sigma_y\sigma_y^* = -\mathbf{I}_2, \tag{6.128}$$

that is, a double time reversal changes the sign of the wave function. Note, however, that this derivation, and indeed this form of the time-reversal operator, is derived for a one-electron system. For a spin-independent system, $\hat{\mathcal{K}} = \hat{\mathcal{K}}_0$, and clearly $\hat{\mathcal{K}}^2 = \mathbf{I}_2$. For more than one electron, we must have a σ_y for each spin of $\frac{1}{2}$, and thus $\hat{\mathcal{K}}^2$ is $\mathbf{I}_2$ or $-\mathbf{I}_2$ depending on whether the system has an even or odd number of electrons, or equivalently whether it has boson or fermion symmetry.

In spin space the active part of the time-reversal operator is the σ_y part. Applying time reversal to our standard basis in spin space we get

$$\begin{aligned} \hat{\mathcal{K}}\begin{pmatrix}1\\0\end{pmatrix} &= -i\sigma_y\hat{\mathcal{K}}_0\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}0 & -1\\1 & 0\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}0\\1\end{pmatrix} \\ \hat{\mathcal{K}}\begin{pmatrix}0\\1\end{pmatrix} &= -i\sigma_y\hat{\mathcal{K}}_0\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}0 & -1\\1 & 0\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix} = -\begin{pmatrix}1\\0\end{pmatrix}. \end{aligned} \tag{6.129}$$

Thus time reversal in spin space causes a spin flip.

For four-component wave functions it is the spin operator $\boldsymbol{\Sigma}$ that changes sign under time reversal, and by simple extension of the algebra above we can write the four-component operator as

$$\hat{\mathcal{K}} = -i\Sigma_y\hat{\mathcal{K}}_0. \tag{6.130}$$

Let us first demonstrate the effect of time reversal on the time-dependent Schrödinger equation:

$$i\hbar\frac{\partial}{\partial t}\psi(t) = \hat{\mathcal{H}}^{\mathrm{NR}}\psi(t). \tag{6.131}$$

In the absence of external vector potentials, the only variable in $\hat{\mathcal{H}}^{\mathrm{NR}}$ affected by time reversal is $\mathbf{p}$ in the kinetic energy expression. But this occurs only in the form $\mathbf{p}\cdot\mathbf{p}$, and therefore $\hat{\mathcal{H}}^{\mathrm{NR}}$ must be invariant to time reversal. Transforming the time-dependent equation above, we have

$$-i\hbar\frac{\partial}{\partial t}\hat{\mathcal{K}}\psi(t) = \hat{\mathcal{H}}^{\mathrm{NR}}\hat{\mathcal{K}}\psi(t) \tag{6.132}$$

which after a change of variables becomes

$$i\hbar\frac{\partial}{\partial t}\hat{\mathcal{K}}\psi(-t) = \hat{\mathcal{H}}^{\mathrm{NR}}\hat{\mathcal{K}}\psi(-t). \tag{6.133}$$

This is the equation for the time-reversed function of $\psi(t)$, which we denote $\bar{\psi}(t)$. We have

$$\hat{\mathcal{K}}\psi(-t) = \bar{\psi}(t) \tag{6.134}$$

where $\hat{\mathcal{K}}$ both changes the functional form of ψ to the time-reversed form and reverses the time coordinate itself. The time-reversed function $\bar{\psi}(t)$ satisfies the equation

$$i\hbar\frac{\partial}{\partial t}\bar{\psi}(t) = \hat{\mathcal{H}}^{\mathrm{NR}}\bar{\psi}(t). \tag{6.135}$$

Under the assumption of no external vector potentials, the time dependence may be separated out in the usual manner to show that $\psi(t)$ and $\bar{\psi}(t)$ have the same eigenvalue:

$$\hat{\mathcal{H}}^{\mathrm{NR}}\psi(t) = E\psi(t); \qquad \hat{\mathcal{H}}^{\mathrm{NR}}\bar{\psi}(t) = E\bar{\psi}(t). \tag{6.136}$$

For the relativistic case of the time-independent Dirac equation for an electron subjected only to time-independent electrostatic fields, we have

$$(\hat{h}^{\mathrm{D}} + V)\psi = c(\boldsymbol{\alpha}\cdot\mathbf{p})\psi + (\beta mc^2 + V)\psi = E\psi. \tag{6.137}$$

Time-reversing this operator yields

$$\begin{aligned}\hat{\mathcal{K}}(\hat{h}^{\mathrm{D}} + V)\hat{\mathcal{K}}^{-1} &= c(\hat{\mathcal{K}}\boldsymbol{\alpha}\hat{\mathcal{K}}^{-1})\cdot(\hat{\mathcal{K}}\mathbf{p}\hat{\mathcal{K}}^{-1}) + mc^2(\hat{\mathcal{K}}\beta\hat{\mathcal{K}}^{-1}) + \hat{\mathcal{K}}V\hat{\mathcal{K}}^{-1}\\ &= c(\boldsymbol{\alpha}\cdot\mathbf{p}) + mc^2(\hat{\mathcal{K}}\beta\hat{\mathcal{K}}^{-1}) + \hat{\mathcal{K}}V\hat{\mathcal{K}}^{-1}\end{aligned} \tag{6.138}$$

where we have used (6.130) for time reversal in the four-component space. It follows from (6.123) that $\boldsymbol{\alpha}$ must change sign under time reversal, and hence the $c\boldsymbol{\alpha} \cdot \mathbf{p}$ term is invariant. The change in sign of $\boldsymbol{\alpha}$ is consistent with the role of $c\boldsymbol{\alpha}$ as the velocity operator. The term involving β is easily shown to be invariant. The one-electron potential is time-independent, and for the static nuclear potential it is also $\boldsymbol{\alpha}$-independent. Therefore, the Hamiltonian is invariant under time reversal, and in analogy with the nonrelativistic case we get

$$\begin{aligned}(\hat{h}^{\mathrm{D}} + V)\psi &= E\psi \\ (\hat{h}^{\mathrm{D}} + V)\bar{\psi} &= E\bar{\psi}.\end{aligned} \tag{6.139}$$

The extension to the many-electron case is straightforward, and the considerations involving space inversion of the Coulomb potential and the Breit terms also hold for time reversal.

We have now shown that a wave function and its time reverse are energy degenerate in both the relativistic and the nonrelativistic case. One way this can happen is if time reversal just produces the same function—that is, the function is invariant under time reversal, possibly with the exception of a phase factor. In this case we have

$$\hat{\mathcal{K}}\psi(t) = a\psi(t) \tag{6.140}$$

where a is a phase factor. Double time reversal then yields

$$\hat{\mathcal{K}}^2\psi(t) = \hat{\mathcal{K}}a\psi(t) = a^*\hat{\mathcal{K}}\psi(t) = |a|^2\psi(t). \tag{6.141}$$

We have previously shown that $\hat{\mathcal{K}}^2$ can only take the values 1 or -1, depending on whether the system has boson or fermion symmetry. This leaves us with two possibilities:

1. $\hat{\mathcal{K}}^2 = 1$. In this case, the equation above yields $|a|^2 = 1$, and the function may indeed be invariant under time reversal.
2. $\hat{\mathcal{K}}^2 = -1$. In this case, the equation above implies $|a|^2 = -1$, which clearly is impossible.

For a system with fermion symmetry—that is, half-integer spin—the wave function cannot therefore be invariant under time reversal. Further, the time-reversed wave function cannot be a simple multiple of the original.

We can explore the fermion case further. Because $\hat{\mathcal{K}}$ is an antiunitary operator, we must have

$$\langle\hat{\mathcal{K}}\psi|\hat{\mathcal{K}}\psi\rangle = 1 \tag{6.142}$$

for arbitrary, normalized ψ, and the time-reversed wave function is thus normalized. We can also show that $\hat{\mathcal{K}}\psi$ is orthogonal to ψ, as follows. If we treat the overlap of the wave function and the time-reversed wave function as a number, we can apply the time-reversal operator to get

$$\hat{\mathcal{K}}\langle\hat{\mathcal{K}}\psi\,|\,\psi\rangle = \langle\hat{\mathcal{K}}\psi\,|\,\psi\rangle^* = \langle\psi\,|\,\hat{\mathcal{K}}\psi\rangle. \tag{6.143}$$

If we apply the time-reversal operator to the functions in the overlap integral, we get

$$\hat{\mathcal{K}}\langle\hat{\mathcal{K}}\psi\,|\,\psi\rangle = \langle\hat{\mathcal{K}}^2\psi\,|\,\hat{\mathcal{K}}\psi\rangle = -\langle\psi\,|\,\hat{\mathcal{K}}\psi\rangle \tag{6.144}$$

because $\hat{\mathcal{K}}^2 = -1$. Therefore

$$\langle\hat{\mathcal{K}}\psi\,|\,\psi\rangle = 0 \tag{6.145}$$

and $\hat{\mathcal{K}}\psi$ and ψ are linearly independent. These conclusions may, of course, be verified by using the explicit expression for the time-reversal operator.

Thus, for the fermion case, ψ and $\hat{\mathcal{K}}\psi$ form a degenerate pair. It is easy to see that any third function that is degenerate in energy with ψ and $\hat{\mathcal{K}}\psi$ will also have its own time-reversed partner. This illustrates Kramers' theorem (Kramers 1930): for a system with half-integer spin, the energy levels are at least doubly degenerate, and any degeneracy is even-fold. This theorem holds in the absence of external vector potentials. We will use the term *Kramers pair* for a pair of functions related by time reversal. The $\psi(t)$ and $\bar{\psi}(t)$ above form a Kramers pair, and the hydrogenic $1s\alpha$ and $1s\beta$ spin–orbitals form another example of a Kramers pair. Note, however, that the Kramers partner of the $2p_1\alpha$ spin–orbital is $2p_{-1}\beta$, not $2p_1\beta$, because time-reversal includes complex conjugation. As can be seen from the above demonstration, the theorem is equally valid for relativistic and nonrelativistic systems, provided any potentials involved are invariant with respect to time reversal.

As a last comment, we note that although the time-reversal operator, $\hat{\mathcal{K}}$, commutes with the Hamiltonian, it is antilinear and therefore cannot correspond to an observable.

6.10 Lorentz Transformations and the Lorentz Group

While we still have no intention of proving the invariance of the Dirac equation under Lorentz transformations, we do want to conclude this discussion of relativistic symmetry by demonstrating the relationship between the Lorentz transformations and the spatial transformations already discussed. To make the symmetry between space and time more explicit we introduce the variable τ, defined as

$$\tau = ict. \tag{6.146}$$

The general four-vector introduced in chapter 2 is then

$$\mathrm{w} = (x, y, z, \tau), \tag{6.147}$$

and for motion along the x axis, the Lorentz transformations take the form

$$\begin{aligned} x' &= \gamma \left(x + i\tau \frac{v}{c} \right) \\ y' &= y \\ z' &= z \\ \tau' &= \gamma \left(\tau - ix \frac{v}{c} \right). \end{aligned} \tag{6.148}$$

If we write this transformation as the result of applying an operator $\hat{\lambda}_x(v)$ to the four-vector w, we get

$$\mathrm{w}' = \hat{\lambda}_x(v)\mathrm{w} = \hat{\lambda}_x(v)(x, y, z, \tau) = \left(\gamma \left(x + i\tau \frac{v}{c} \right), y, z, \gamma \left(\tau - ix \frac{v}{c} \right) \right). \tag{6.149}$$

This is equal to a linear transformation $\Lambda_x(v)$, such that

$$\mathrm{w}' = \Lambda_x(v)\mathrm{w} = \begin{pmatrix} \gamma & 0 & 0 & \gamma i \frac{v}{c} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\gamma i \frac{v}{c} & 0 & 0 & \gamma \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ \tau \end{pmatrix}. \tag{6.150}$$

Here we recognize a situation that is quite similar to that encountered for the $SO(3)$ rotations in section 6.4. We have transformations that conserve the length of the four-vector, and which may be expressed as matrices operating on the set of basis vectors. We may in fact proceed in a completely analogous manner to the SO(3) case. Thus, an infinitesimal Lorentz transformation—that is, $v/c \ll 1$—may be written as

$$\Lambda_x(v) = \mathbf{I}_4 - i\frac{v}{c}\mathbf{Y}_x + \mathcal{O}\left(\frac{v^2}{c^2} \right). \tag{6.151}$$

The matrix $\mathbf{Y}_x$ is the equivalent of the $\mathbf{X}_x$ matrix for spatial rotations, and it may be found to be

$$-i\mathbf{Y}_x = \left. \frac{d\boldsymbol{\Lambda}_x(v)}{d\left(\frac{v}{c}\right)} \right|_{\frac{v}{c}=0} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}. \tag{6.152}$$

Any finite Lorentz transformation may now be expressed in exponential form, and for motion along the x axis,

$$\Lambda_x(v) = e^{-i(v/c)\mathbf{Y}_x}. \tag{6.153}$$

Corresponding generators for motion along the other two axes are easily derived.

The **X** matrices were defined only on R_3, but they may easily be cast into four-space form by adding a row and a column of zeros. The six matrices

$$\mathbf{X}_x = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -i & 0 \\ 0 & i & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \ \mathbf{X}_y = \begin{pmatrix} 0 & 0 & -i & 0 \\ 0 & 0 & 0 & 0 \\ i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \ \mathbf{X}_z = \begin{pmatrix} 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \tag{6.154}$$

and

$$\mathbf{Y}_x = \begin{pmatrix} 0 & 0 & 0 & i \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -i & 0 & 0 & 0 \end{pmatrix}; \ \mathbf{Y}_y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & i \\ 0 & 0 & 0 & 0 \\ 0 & -i & 0 & 0 \end{pmatrix}; \ \mathbf{Y}_z = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & i \\ 0 & 0 & -i & 0 \end{pmatrix} \tag{6.155}$$

generate a group that is known as the *proper Lorentz group*. If the space inversion operator, represented by

$$\mathbf{I} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6.156}$$

is added to the set, the resulting group is the *orthochronous Lorentz group*. Finally, we may add the time-inversion operation

$$\mathbf{K}_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{6.157}$$

This defines the *full Lorentz group*. The symmetry of the Dirac equation may therefore be expressed as the direct product of $SU(2)$ and the full Lorentz group.

This concludes our introduction to relativistic symmetry. Our aim has been to relate closely to features that should be familiar to the practicing quantum chemist. In particular, we have put some emphasis on the double groups, which represent a rather straightforward extension of the methods and concepts of nonrelativistic symmetry. We have also provided a more general discussion that shows how the double group symmetry arises as the direct product of the underlying symmetries in the two separate physical spaces considered—spin space and the four-space spanned by the Lorentz transformations. In the chapters to follow, we will repeatedly exploit both $SU(2) \otimes G$, $G \subseteq O(3)$ symmetry and Kramers symmetry to develop and simplify methods for quantum chemical calculations on relativistic systems.

7

One-Electron Atoms

The development of quantum chemistry, that is, the solution of the Schrödinger equation for molecules, is almost exclusively founded on the expansion of the molecular electronic wave function as a linear combination of atom-centered functions, or atomic orbitals—the LCAO approximation. These orbitals are usually built up out of some set of *basis functions*. The properties of the atomic functions at large and small distances from the nucleus determines to a large extent what characteristics the basis functions must have, and for this purpose it is sufficient to examine the properties of the hydrogenic solutions to the Schrödinger equation. If we are to do the same for relativistic quantum chemistry, we should first examine the properties of the atomic solutions to determine what kind of basis functions would be appropriate.

However, the atomic solutions of the Dirac equation provide more than merely a guide to the choice of basis functions. The atoms in a molecule retain their atomic identities to a very large extent, and the modifications caused by the molecular field are quite small for most properties. In order to arrive at a satisfactory description of the relativistic effects in molecules, we must first of all be able to treat these effects at the atomic level. The insight gained into the effects of relativity on atomic structure is therefore a necessary and useful starting point for relativistic quantum chemistry.

7.1 Separation of Variables in the Dirac Equation

As in the nonrelativistic case, most of the salient features of the atomic systems are exposed in the treatment of the simplest of these, the hydrogen-like one-electron atoms. In Hartree atomic units the time-independent Dirac equation yields the coupled equations

$$\begin{aligned} (V - E)\psi^L + c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^S &= 0 \\ c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L + (V - E - 2mc^2)\psi^S &= 0 \end{aligned} \tag{7.1}$$

where we have shifted the energy by $-mc^2$ (with $m = 1$), as discussed in section 4.6. We will use this shifted energy scale for the rest of the book unless otherwise explicitly indicated. V is here a scalar, central potential. For a point nucleus this is $-Z/r$, but we use V here in order to account for other possible models of the nucleus. In solving the Schrödinger equation for the hydrogen atom, we can separate angular and radial variables in spherical polar coordinates and show that the solutions take the form

$$\psi^{\mathrm{NR}}(r, \vartheta, \varphi) = \frac{1}{r}\, R_{n\ell}(r) Y_{\ell}^{m_\ell}(\vartheta, \varphi) \tag{7.2}$$

where the $Y_{\ell}^{m_\ell}(\vartheta, \varphi)$ are the usual spherical harmonics, defined with the Condon and Shortley phase conventions,

$$Y_{\ell}^{m_\ell}(\vartheta, \varphi) = (-1)^{(m_\ell+|m_\ell|)/2} \left[\frac{(2\ell+1)}{4\pi} \frac{(\ell-|m_\ell|)!}{(\ell+|m_\ell|)!}\right]^{1/2} P_{\ell}^{m_\ell}(\cos\vartheta)\, e^{im_\ell\varphi}. \tag{7.3}$$

For the Dirac equation, we know already that the large and small components have different radial functions, so we must seek a separation of the form

$$\psi(r, \vartheta, \varphi, \tau) = \frac{1}{r} \begin{pmatrix} P(r)\ \xi_L(\vartheta, \varphi, \tau) \\ iQ(r)\ \xi_S(\vartheta, \varphi, \tau) \end{pmatrix}. \tag{7.4}$$

Here, P and Q are the *radial* large and small components of the wave function. The factor of i has been introduced to make the radial components real. The angular functions ξ are two-component spinors, that is, a product of angular and spin functions; the spin variable τ has been explicitly shown.

To achieve this separation of variables in the Dirac equation, we must be able to factorize the operator $\boldsymbol{\sigma} \cdot \mathbf{p}$. The nuclear potential V is already a function of r only. We follow the procedure in Schiff (1968) and introduce the radial operators for momentum and velocity

$$\hat{p}_r = r^{-1}(\mathbf{r} \cdot \mathbf{p} - i); \qquad \sigma_r = r^{-1}(\boldsymbol{\sigma} \cdot \mathbf{r}). \tag{7.5}$$

We now apply the Dirac relation, (4.14), in the form

$$(\boldsymbol{\sigma} \cdot \mathbf{r})(\boldsymbol{\sigma} \cdot \boldsymbol{\ell}) = \mathbf{r} \cdot \boldsymbol{\ell} + i\boldsymbol{\sigma} \cdot \mathbf{r} \times \boldsymbol{\ell} = i\boldsymbol{\sigma} \cdot \mathbf{r} \times \boldsymbol{\ell}. \tag{7.6}$$

The scalar product $\mathbf{r} \cdot \boldsymbol{\ell}$ vanishes because $\boldsymbol{\ell} = \mathbf{r} \times \mathbf{p}$. The vector product $\mathbf{r} \times \boldsymbol{\ell}$ may be evaluated using standard vector relations:

$$\mathbf{r} \times \boldsymbol{\ell} = \mathbf{r} \times (\mathbf{r} \times \mathbf{p}) = \mathbf{r}(\mathbf{r} \cdot \mathbf{p}) - r^2\mathbf{p} = \mathbf{r}(r\hat{p}_r + i) - r^2\mathbf{p}. \tag{7.7}$$

Substituting this expression for $\mathbf{r} \times \boldsymbol{\ell}$ in (7.6) yields

$$i\boldsymbol{\sigma} \cdot \mathbf{r} \times \boldsymbol{\ell} = i\boldsymbol{\sigma} \cdot \mathbf{r}(r\hat{p}_r + i) - ir^2\boldsymbol{\sigma} \cdot \mathbf{p} = (\boldsymbol{\sigma} \cdot \mathbf{r})(\boldsymbol{\sigma} \cdot \boldsymbol{\ell}). \tag{7.8}$$

Making use of (7.5) and rearranging, we get

$$\boldsymbol{\sigma} \cdot \mathbf{p} = \sigma_r \hat{p}_r + i\sigma_r r^{-1}(\boldsymbol{\sigma} \cdot \boldsymbol{\ell} + 1) \equiv \sigma_r \hat{p}_r - i\sigma_r r^{-1} \hat{K} \tag{7.9}$$

where $\hat{K}$ is a pure angular momentum operator,

$$\hat{K} = -1 - \boldsymbol{\sigma} \cdot \boldsymbol{\ell} = \hat{\ell}^2 + \hat{s}^2 - \hat{j}^2 - 1 \tag{7.10}$$

where we have used $\boldsymbol{\sigma} = 2\mathbf{s}$ and (6.4) in the second equality. We can now use this expression for $(\boldsymbol{\sigma} \cdot \mathbf{p})$ in the Dirac equation. The first of the coupled equations becomes

$$(V - E)\,\psi^L + c\sigma_r \left[\hat{p}_r - ir^{-1}\hat{K}\right] \psi^S = 0. \tag{7.11}$$

The radial momentum operator $\hat{p}_r$ can be written

$$\hat{p}_r = -i\left(\frac{\mathrm{d}}{\mathrm{d}r} + \frac{1}{r}\right). \tag{7.12}$$

This operator obviously acts only on the radial large or small component. The “radial” velocity operator σ_r is in fact a pure angular operator, because the components of $\hat{\mathbf{r}} = \mathbf{r}/r$ are pure angular functions:

$$\sigma_r = \sigma_x \sin\vartheta \cos\varphi + \sigma_y \sin\vartheta \sin\varphi + \sigma_z \cos\vartheta. \tag{7.13}$$

This operator therefore acts only on the angular large or small component. The operator $\hat{K}$ also operates only on the angular function, and in order to achieve the separation the angular function must be an eigenfunction of $\hat{K}$. Substituting for the radial momentum operator, using the factorized large and small components, with κ^L and κ^S the eigenvalues of the angular large and small components, the first of the coupled equations can be written

$$(V - E)\,\frac{P}{r}\xi^L = -c\left(\frac{\mathrm{d}}{\mathrm{d}r} + \frac{\kappa^S + 1}{r}\right)\frac{Q}{r}\sigma_r \xi^S \tag{7.14}$$

which is obviously factorized. Applying the same procedure to the second of the coupled equations yields

$$\left(V - E - 2mc^2\right)\frac{Q}{r}\xi^S = c\left(\frac{\mathrm{d}}{\mathrm{d}r} + \frac{\kappa^L + 1}{r}\right)\frac{P}{r}\sigma_r \xi^L. \tag{7.15}$$

In the next section we will examine the properties of the angular functions. The radial functions will be the subject of the subsequent sections, where we also make comparisons with nonrelativistic radial functions and expectation values to gain some insight into the effects of relativity on electronic structure.

7.2 Angular Wave Functions

As discussed in the chapter on symmetry (chapter 6), neither orbital nor spin angular momentum provide good quantum numbers for the Dirac equation in a central field, and we must instead turn to eigenfunctions of the operators $\hat{j}^2$ and $\hat{j}_z$ with eigenvalues $j(j+1)$ and m_j. For a one-electron wave function the angular momentum part can be expressed in a basis of coupled products of a spherical harmonic and a Pauli spinor $\eta(m_s)$

$$\xi_{j,m_j}(\vartheta,\varphi) = \sum_{m_\ell,m_s} \langle \ell m_\ell s m_s | j m_j \rangle \; Y_\ell^{m_\ell}(\vartheta,\varphi)\; \eta(m_s) \tag{7.16}$$

where $\langle \ell m_\ell s m_s | j m_j \rangle$ is a Clebsch–Gordan coefficient.

There is a problem with the labeling of the function ξ: the total angular momentum label j is not unique. We can couple $s = \frac{1}{2}$ to either ℓ or $\ell + 1$ to form $j = \ell + \frac{1}{2}$. The ambiguity may be resolved by resorting to the operator $\hat{K}$ derived in the previous section during the separation of variables in the Dirac equation. If we apply this operator to one of the angular momentum basis functions, we find that it is indeed an eigenfunction of $\hat{K}$ with eigenvalue κ, given by

$$\begin{aligned}\hat{K} Y_\ell^{m_\ell}(\vartheta,\varphi)\;\eta(m_s) &= \kappa Y_\ell^{m_\ell}(\vartheta,\varphi)\;\eta(m_s)\\ &= \left[\ell(\ell+1) - j(j+1) - \frac{1}{4}\right] Y_\ell^{m_\ell}(\vartheta,\varphi)\;\eta(m_s).\end{aligned} \tag{7.17}$$

There are two possibilities:

CASE 1 $j = \ell + \dfrac{1}{2}$

Then

$$\kappa = \ell\left(j+\tfrac{1}{2}\right) - j(j+1) - \frac{1}{4} = (\ell - j)\left(j+\tfrac{1}{2}\right) - \tfrac{1}{2}\left(j+\tfrac{1}{2}\right) = 2(l-j)\left(j+\tfrac{1}{2}\right), \tag{7.18}$$

where we use the fact that $\ell - j = -\frac{1}{2}$ in this case.

CASE 2 $j = \ell - \dfrac{1}{2}$

Then

$$\begin{aligned}\kappa &= \left(j+\tfrac{1}{2}\right)\left(j+\frac{3}{2}\right) - j(j+1) - \frac{1}{4} = \left(j+\frac{3}{2}-j\right)\left(j+\tfrac{1}{2}\right) - \tfrac{1}{2}\left(j+\tfrac{1}{2}\right)\\ &= 2(l-j)\left(j+\tfrac{1}{2}\right),\end{aligned} \tag{7.19}$$

where we use the fact that $\ell - j = +\frac{1}{2}$ in this case.

We conclude that the operator $\hat{K}$ has the eigenvalues

$$\kappa = 2(l-j)\left(j+\tfrac{1}{2}\right) = -a\left(j+\tfrac{1}{2}\right), \tag{7.20}$$

where $a = 2(j - \ell) = \pm 1$ is a sign factor, and $j = \ell + \frac{1}{2}a$. The general angular function can therefore be written with a κ label as

$$\xi_{\kappa,m}(\vartheta, \varphi) = \sum_{m_s=\pm 1/2} \langle \ell m_\ell s m_s | jm \rangle \, Y_\ell^{m_\ell}(\vartheta, \varphi)\eta(m_s), \tag{7.21}$$

where $m_\ell + m_s = m$.

These functions are suitable functions for the angular large and small components. We must now determine which of these is to be used for the large and the small components, for there is no a priori reason to expect the angular parts of the large and small components to be the same. From the separation of variables in the previous section, we have the relations

$$\sigma_r \xi^L = \xi^S; \qquad \sigma_r \xi^S = \xi_L \tag{7.22}$$

which implies that $\sigma_r^2 = 1$, a fact that can be proved from the expression given above.[1]

Now $\hat{\mathbf{r}}$ may be expanded in terms of spherical harmonics with $\ell = 1$, so the angular large and small components must differ in the orbital angular momentum by a single unit. Since the total angular momentum commutes with the Dirac Hamiltonian, j is a good quantum number, and therefore both angular components must have the same j value. To have ℓ values that differ by one, the large and small components must have values of κ that are opposite in sign. Thus, we can set $\kappa^L = \kappa$ and $\kappa^S = -\kappa$, and the factorized 4-spinor can be written

$$\psi(r, \vartheta, \varphi, \tau) = \frac{1}{r} \begin{pmatrix} P_{n\kappa}(r) \; \xi_{\kappa,m}(\vartheta, \varphi) \\ i\, Q_{n\kappa}(r) \; \xi_{-\kappa,m}(\vartheta, \varphi) \end{pmatrix}. \tag{7.23}$$

As well as being eigenfunctions of the operators $\hat{j}^2$, $\hat{j}_z$, and $\hat{K}$, the two-spinor angular functions are eigenfunctions of the inversion operator $\hat{I}$ with eigenvalue $(-1)^\ell$. This follows directly from the inversion properties of the spherical harmonics. Because the ℓ value of the spherical harmonics in the angular function for the small component differs from that of the large component by 1, the small component has the opposite parity under inversion. This fact was demonstrated in chapter 6.

A summary of the quantum numbers for the angular functions along with the labels that are often used is given in table 7.1.

We can insert the expressions for the Clebsch–Gordan coefficients into the angular 2-spinors (and suppress the angular variables in the spherical harmonics) to get

$$\xi_{\kappa,m} = \frac{1}{\sqrt{2\ell + 1}} \begin{pmatrix} a\sqrt{\ell + \frac{1}{2} + am} \; Y_\ell^{m-1/2} \\ \sqrt{\ell + \frac{1}{2} - am} \; Y_\ell^{m+1/2} \end{pmatrix}. \tag{7.24}$$

---

[1] The relations in (7.22) are only defined up to a phase, which we are free to choose. We have already chosen a phase factor of i to multiply the small component, so we choose a phase factor of 1 in this case.

Table 7.1 Quantum numbers and labels for atomic spinors

l	0	1	1	2	2	3	3
j	1/2	1/2	3/2	3/2	5/2	5/2	7/2
κ	-1	1	-2	2	-3	3	-4
Labels	s	$\bar{p}$	p	$\bar{d}$	d	$\bar{f}$	f
	s	p_-	p_+	d_-	d_+	f_-	f_+
	$s_{1/2}$	$p_{1/2}$	$p_{3/2}$	$d_{3/2}$	$d_{5/2}$	$f_{5/2}$	$f_{7/2}$

For $a = +1$, that is, $j = \ell + \frac{1}{2}$, and with $m_\ell = m - \frac{1}{2}$

$$\xi_{\kappa,m} = \frac{1}{\sqrt{2\ell+1}} \begin{pmatrix} \sqrt{\ell+m_\ell+1}\; Y_\ell^{m_\ell} \\ \sqrt{\ell-m_\ell}\; Y_\ell^{m_\ell+1} \end{pmatrix} \tag{7.25}$$

and for $a = -1$, that is, $j = \ell - \frac{1}{2}$, with $m_\ell = m - \frac{1}{2}$

$$\xi_{\kappa,m} = \frac{1}{\sqrt{2\ell+1}} \begin{pmatrix} -\sqrt{\ell-m_\ell}\; Y_\ell^{m_\ell} \\ \sqrt{\ell+m_\ell+1}\; Y_\ell^{m_\ell+1} \end{pmatrix}. \tag{7.26}$$

Explicit forms are given in table 7.2 for a few values of κ.

The coupling of the spin and orbital angular momentum has implications for the electron density of the one-particle functions. We can compare the Dirac angular functions with the Schrödinger angular functions. This is done in figure 7.1 using polar plots of the angular density, integrated over spin. Notice that the relativistic densities

Table 7.2 Angular atomic 2-spinors for *s*, *p* and *d* subshells

$s_{1/2}$	$\begin{pmatrix} s \\ 0 \end{pmatrix}$		
$\sqrt{3}p_{1/2}$	$\begin{pmatrix} -p_0 \\ \sqrt{2}p_1 \end{pmatrix}$		
$\sqrt{3}p_{3/2}$	$\begin{pmatrix} \sqrt{2}p_0 \\ p_1 \end{pmatrix}$	$\begin{pmatrix} p_1 \\ 0 \end{pmatrix}$	
$\sqrt{5}d_{3/2}$	$\begin{pmatrix} -\sqrt{2}d_0 \\ \sqrt{3}d_1 \end{pmatrix}$	$\begin{pmatrix} -d_1 \\ 2d_2 \end{pmatrix}$	
$\sqrt{5}d_{5/2}$	$\begin{pmatrix} \sqrt{3}d_0 \\ \sqrt{2}d_1 \end{pmatrix}$	$\begin{pmatrix} 2d_1 \\ d_2 \end{pmatrix}$	$\begin{pmatrix} d_2 \\ 0 \end{pmatrix}$

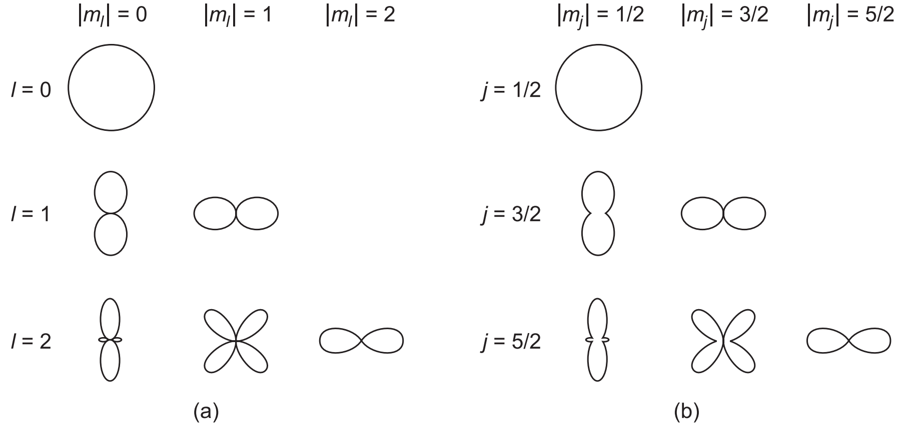


Figure 7.1 Polar density plots of (a) nonrelativistic orbitals and (b) relativistic 2-spinors

depend on j and not on ℓ. Thus $s_{1/2}$ and $p_{1/2}$ 2-spinors are *both* spherical. A relativistic filled shell also has a spherical density; for example, the ground state of the Pb atom has the configuration $6p^2$, which in a nonrelativistic scheme would have ^3P symmetry and components whose density looks like that of a p function, but in a relativistic scheme the configuration is $6p_{1/2}^2$, which has $J = 0$ and a spherical density.

Another feature that emerges from these plots is the loss of nodal structure. Because the spin-up and spin-down components of each spinor have nodes in different places, the directional properties of the angular functions are smeared out compared with the properties of the nonrelativistic angular functions. Only for the highest m value does the spinor retain the nodal structure of the nonrelativistic angular function, and that is because it is a simple product of a spin function and a spherical harmonic. The admixture of m_ℓ and $m_\ell + 1$ character approaches equality as ℓ increases and as m_ℓ approaches zero, resulting in a loss of spatial directionality. The implications of this loss of directionality for molecular structure could be significant, particularly where the structure is not determined simply from the molecular symmetry or from electrostatics.

7.3 Solutions of the Radial Dirac Equation

We now turn to the radial functions and the radial Dirac equation. Using (7.22) in (7.14) and (7.15) we arrive at the pair of coupled radial equations[2]

$$(V - E)P + c\left[\frac{\mathrm{d}Q}{\mathrm{d}r} - \frac{\kappa Q}{r}\right] = 0 \tag{7.27a}$$

$$-c\left[\frac{\mathrm{d}P}{\mathrm{d}r} + \frac{\kappa P}{r}\right] + (V - E - 2c^2)\,Q = 0. \tag{7.27b}$$

---

[2] We omit the mass in this section, and work strictly in Hartree atomic units.

or in matrix form

$$\begin{pmatrix} V - E & c\left[\dfrac{\mathrm{d}}{\mathrm{d}r} - \dfrac{\kappa}{r}\right] \\ \\ -c\left[\dfrac{\mathrm{d}}{\mathrm{d}r} + \dfrac{\kappa}{r}\right] & V - E - 2c^2 \end{pmatrix} \begin{pmatrix} P(r) \\ \\ Q(r) \end{pmatrix} = 0. \tag{7.28}$$

For the purposes of comparison with the nonrelativistic limit, we eliminate the small component from these equations, using (7.27b) to express Q in terms of P and substituting into (7.27a), with the result

$$\begin{aligned} \Bigg[\frac{\mathrm{d}^2}{\mathrm{d}r^2} - \frac{\kappa(\kappa+1)}{r^2} + \frac{(V - E - 2c^2)}{c^2}(V - E) \\ - \frac{1}{(V - E - 2c^2)}\frac{\mathrm{d}V}{\mathrm{d}r}\left(\frac{\mathrm{d}}{\mathrm{d}r} + \frac{\kappa}{r}\right)\Bigg] P(r) = 0. \end{aligned} \tag{7.29}$$

Comparing this with the nonrelativistic radial equation,

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}r^2} - \frac{\ell(\ell+1)}{r^2} + 2(E - V)\right] R(r) = 0, \tag{7.30}$$

we see that the former reduces to the latter in the limit $c \to \infty$, provided that $\kappa(\kappa+1) = \ell(\ell+1)$, which indeed holds, as may be verified from the definition of κ above.

The solutions of these equations depend on the choice of the nuclear potential. For a point nucleus $V = -Z/r$, analytic solutions have been derived (Darwin 1928, Gordon 1928). We will not go into the details of the derivation, but we reproduce the essential features of the solutions for the purposes of comparison between nonrelativistic and relativistic solutions and the examination of the effects of relativity.

The solutions of the point nucleus radial Dirac equation can be written as a product of an exponential function, a power of r, and a polynomial:

$$\begin{pmatrix} P(r) \\ Q(r) \end{pmatrix} = \mathcal{N} e^{-\lambda r}\, r^{\gamma} \begin{pmatrix} \mathcal{N}_P[F_1(r) + F_2(r)] \\ \mathcal{N}_Q[F_1(r) - F_2(r)] \end{pmatrix}. \tag{7.31}$$

The exponent λ is defined by

$$\lambda = \sqrt{-E(2 + E/c^2)} \tag{7.32}$$

and γ by

$$\gamma = +\sqrt{\kappa^2 - Z^2/c^2} < |\kappa|. \tag{7.33}$$

The nonrelativistic solution in a point nuclear potential can be similarly written

$$R(r) = \mathcal{N}' e^{-\sqrt{-2E}r}\ r^{\ell+1}\ F'(r). \tag{7.34}$$

There are some obvious similarities between the functions, and some obvious differences. Further discussion of the exponents is deferred to the next two sections, where the implications of the form of the exponent are drawn out. Here, we first want to make some remarks on the overall shape of the radial functions.

The apparent angular quantum number γ is clearly not an integer.[3] For negative κ, that is, for $j = \ell + \frac{1}{2}$, $|\kappa| = \ell + 1$ and the exponent of r is less than the nonrelativistic exponent. This means that for s functions, instead of a cusp at the nucleus, there is a singularity, which will inevitably place much greater demands on the basis set in an LCAO expansion of the wave function than in the nonrelativistic case. It also means that the value of the large and small components at the origin $r = 0$ are *discontinuous* at the nonrelativistic limit $c = \infty$,

$$\begin{aligned} \psi^L(0, c), \psi^S(0, c) &= \infty, \quad c < \infty \\ &= \mathcal{N}', \quad c = \infty. \end{aligned} \tag{7.35}$$

This discontinuity causes some difficulties in the interpretation of the nonrelativistic limit of the wave function, to which we return later.

For positive κ, that is, for $j = \ell - \frac{1}{2}$, $|\kappa| = \ell$, and the exponent of r in the relativistic wave function is smaller than the nonrelativistic exponent of r by at least one unit. This means that the $2p_{1/2}$ radial function should behave in the same way as the $1s_{1/2}$ radial function, and have a singularity! To further understand what is going on here, and how the nonrelativistic limit is attained, we need to examine the polynomial functions F_1 and F_2.

F_1 and F_2 can be expressed in terms of the confluent hypergeometric function $F(a, b; x)$, which is defined by

$$F(a, b; x) = 1 + \frac{a}{b}x + \frac{a(a+1)}{b(b+1)}x^2 + \cdots\ . \tag{7.36}$$

This sum clearly truncates after a finite number of terms if a is a negative integer. The two functions F_1 and F_2 appearing in the radial solution of the point nucleus Dirac equation are

$$\begin{aligned} F_1 &= (\mathrm{N}_\kappa - \kappa) F(-n_r, 2\gamma + 1; 2\lambda r) \\ F_2 &= -n_r F(1 - n_r, 2\gamma + 1; 2\lambda r) \end{aligned} \tag{7.37}$$

---

[3] As a matter of fact, for hydrogenic atoms with nuclear charge $Z > 137$, it could even be imaginary. In practice, other effects, such as finite nuclear size, become important at that stage, and the imaginary γ remains mainly a formal problem.

where we have defined the variable N_κ by

$$\begin{aligned}\mathrm{N}_\kappa &= \sqrt{(n_r+\gamma)^2 + Z^2/c^2} \\ &= \sqrt{n^2 - 2n_r(|\kappa| - \gamma)} \\ &\leq n\end{aligned} \tag{7.38}$$

and the radial quantum number n_r by

$$n_r = n - |\kappa| \tag{7.39}$$

where n is the nonrelativistic principal quantum number. The nonrelativistic equivalent of the radial quantum number is $n - \ell - 1$, which counts the number of radial nodes. As n_r is non-negative, F_1 will always truncate. F_2 truncates if $n_r > 0$, but for $n_r = 0$ it vanishes. Hence both functions are indeed polynomial functions of the radial variable. If F_2 vanishes, the large and small components have exactly the same radial form, and the ratio between them is that of the relative normalization factors,

$$\mathcal{N}_P = \sqrt{2 + E/c^2}, \quad \mathcal{N}_Q = \sqrt{-E/c^2}. \tag{7.40}$$

This case applies whenever $n = |\kappa|$, that is, to the $1s_{1/2}$, $2p_{3/2}$, ... spinors.

The $2p_{1/2}$ case has $n_r = 1$ and $\kappa = 1$. Then

$$F_2 = -n_r F(0, 2\gamma + 1; 2\lambda r) = -n_r \tag{7.41}$$

and

$$F_1 = (\mathrm{N}_\kappa - \kappa)F(-1, 2\gamma + 1; 2\lambda r) = (\mathrm{N}_\kappa - \kappa)(1 - 2\lambda r/(2\gamma + 1)) \tag{7.42}$$

and the wave function is

$$\begin{pmatrix} P(r) \\ Q(r) \end{pmatrix} = \mathcal{N} e^{-\lambda r}\, r^\gamma \begin{pmatrix} \mathcal{N}_P \left[(\mathrm{N}_\kappa - \kappa)\left(1 - \frac{2\lambda r}{2\gamma+1}\right) - n_r \right] \\ \mathcal{N}_Q \left[(\mathrm{N}_\kappa - \kappa)\left(1 - \frac{2\lambda r}{2\gamma+1}\right) + n_r \right] \end{pmatrix}. \tag{7.43}$$

Given that $\gamma \simeq \ell$, both P and Q behave approximately as r^ℓ at the origin, not $r^{\ell+1}$ as the nonrelativistic radial function does. There is in fact a singularity at the origin of the same order as the $1s_{1/2}$. However,

$$\mathrm{N}_\kappa = \sqrt{n^2 - 2n_r(|\kappa| - \gamma)}; \qquad |\kappa| - \gamma = \mathcal{O}\left(\frac{Z^2}{c^2}\right) \tag{7.44}$$

and therefore $\mathrm{N}_\kappa - \kappa$ is quite close to n_r, and in fact equality occurs in the nonrelativistic limit. Thus, for $2p_{1/2}$

$$F_1 \approx n_r \left(1 - \frac{2\lambda r}{2\gamma + 1}\right). \tag{7.45}$$

Inserting this approximate value for F_1 and the exact value for F_2 into the expression for the radial solutions, we get for $2p_{1/2}$

$$\begin{aligned} \begin{pmatrix} P(r) \\ Q(r) \end{pmatrix} &\simeq \mathcal{N} e^{-\lambda r}\, r^\gamma \begin{pmatrix} \mathcal{N}_P \left[n_r \left(1 - \frac{2\lambda r}{2\gamma+1}\right) - n_r \right] \\ \mathcal{N}_Q \left[n_r \left(1 - \frac{2\lambda r}{2\gamma+1}\right) + n_r \right] \end{pmatrix} \\ &\simeq \mathcal{N} e^{-\lambda r}\, r^\gamma \begin{pmatrix} \mathcal{N}_P n_r \left[\left(- \frac{2\lambda r}{2\gamma+1}\right) \right] \\ \mathcal{N}_Q n_r \left[\left(2 - \frac{2\lambda r}{2\gamma+1}\right) \right] \end{pmatrix}. \end{aligned} \tag{7.46}$$

In this form the $2p_{1/2}$ does appear better behaved. The large component adds in another power of r in the lowest order, and yields the correct power dependence for a nonrelativistic $2p$ function. The problem of the singularity persists in the lowest order for the small component, but here the normalization factor $\mathcal{N}_Q$ vanishes in the nonrelativistic limit, and so the small component also tends towards the correct behavior. As discussed in chapter 4, this is not true at $r = 0$, where the singularity persists.

We have compared the relativistic and nonrelativistic radial functions, and we now compare some of the properties of the solutions. We select two of these, the eigenvalue and the mean radius. The orbital eigenvalue $\epsilon_{n\kappa}$ is given by

$$\begin{aligned} \epsilon_{n\kappa} &= -c^2 \left[1 - \sqrt{1 - \left(\frac{Z}{c\mathrm{N}_\kappa}\right)^2} \,\right] \\ &= -\frac{Z^2}{\mathrm{N}_\kappa^2} \left[1 + \sqrt{1 - \left(\frac{Z}{c\mathrm{N}_\kappa}\right)^2} \,\right]^{-1} \\ &< -\frac{Z^2}{2n^2} \end{aligned} \tag{7.47}$$

since $\mathrm{N}_\kappa \leq n$ and the factor in square brackets is less than 2. We see from this that N_κ has the role of an apparent principal quantum number. The direct effect of relativity is therefore to provide an energetic stabilization of the orbitals. One further point of interest here is that the eigenvalue depends only on the magnitude of κ (via N_κ) and not on its sign; thus the $2s_{1/2}$ and the $2p_{1/2}$ orbitals are degenerate, but the $2p_{3/2}$ is higher in energy. The difference in energy between the $2p_{1/2}$ and the $2p_{3/2}$ is the spin–orbit splitting. The degeneracy between the $2s_{1/2}$ and $2p_{1/2}$ levels is removed by the Lamb shift. In contrast, the nonrelativistic eigenvalues have no dependence on the angular momentum at all, and the $2s$ and all $2p$ orbitals are degenerate.

The mean radius is given by

$$\begin{aligned}\langle r\rangle_{n\kappa} &= \frac{1}{2Z}\left[\frac{\gamma+n_r}{\mathrm{N}_\kappa}(3\mathrm{N}_\kappa^2-\kappa^2)-\kappa\right]\\ &< \frac{1}{2Z}\left[3\mathrm{N}_\kappa^2-\kappa(\kappa+1)\right]\\ &< \frac{1}{2Z}\left[3n^2-\ell(\ell+1)\right]\end{aligned}\tag{7.48}$$

since $\ell(\ell+1)=\kappa(\kappa+1)$. The last line is the nonrelativistic mean radius. The direct effect of relativity is to decrease the mean radius, which is consistent with the lowering of the energy.

7.4 Behavior at Large *r*

At large distances from the nucleus, the potential energy term as well as the angular momentum term of the radial Dirac equation becomes small, and the equation approaches the limiting form:

$$EP - c\left[\frac{\mathrm{d}Q}{\mathrm{d}r}\right] = 0 \tag{7.49a}$$

$$c\left[\frac{\mathrm{d}P}{\mathrm{d}r}\right] + (2c^2+E)\,Q = 0. \tag{7.49b}$$

This pair of equations is easily seen to be satisfied by exponential solutions, and in fact at large r the behavior of the solution is dominated by the exponential term. The behavior of the solutions—bound or continuum—depends on λ from (7.32):

$$\lambda = \sqrt{-E(2+E/c^2)}. \tag{7.50}$$

For positive energies, λ is imaginary in both relativistic and nonrelativistic cases, and the wave function is oscillatory, representing a free electron in a Coulomb field. For small energies, $E < mc^2$, the wavelengths will be very similar, but as E becomes large they will diverge from each other, the relativistic wavelength becoming shorter than that of the nonrelativistic function for the same energy.

For negative energies, λ is real for all energies in the nonrelativistic case, and hence the spectrum in this region is composed of bound states. In the relativistic case, λ is real in the range $0 > E > -2c^2$, providing a bound spectrum, but for $E < -2c^2$, λ is again imaginary and the solutions are oscillatory. The spectrum for both cases is shown in figure 7.2.

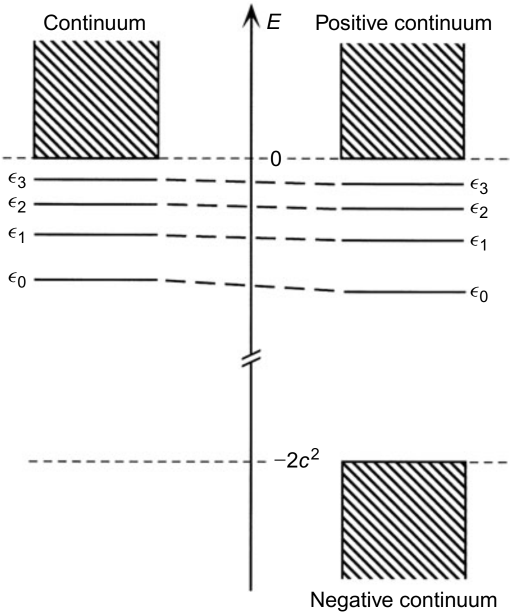


Figure 7.2 Spectrum of the Schrödinger Hamiltonian (left) and the Dirac Hamiltonian (right)

It is interesting to note that λ is related to the product of the large and small component relative normalization factors,

$$\lambda^2 = c^2 \mathcal{N}_P^2 \mathcal{N}_Q^2. \tag{7.51}$$

The two points at which $\lambda = 0$ correspond to $E = 0$, where the small component vanishes, and $E = -2mc^2$, where the large component vanishes.

7.5 Behavior at Small *r*

At small distances from the nucleus, the solutions of the Dirac equation are determined largely by the nuclear potential. In section 7.3, the solutions for a point nuclear potential were presented. Here, we adopt a more general approach, to determine the behavior of the solutions for an arbitrary (but realistic) nuclear potential. The radial functions are expanded in a power series,

$$\begin{aligned} P(r) &= r^{\gamma_p}(p_0 + p_1 r + p_2 r^2 + \cdots) \\ Q(r) &= r^{\gamma_q}(q_0 + q_1 r + q_2 r^2 + \cdots). \end{aligned} \tag{7.52}$$

Note that P and Q do not necessarily have the same lowest-order term. In this expansion, we require p_0 and q_0 to be nonzero. Other coefficients may however be zero. The potential V is also expanded in a power series,

$$-rV(r) = v_0 + v_1 r + v_2 r^2 + \cdots . \tag{7.53}$$

For a point nucleus, $v_0 = Z,\ v_i = 0,\ i > 0$. The form of the expansion for a finite nucleus can be derived by considering the potential due to a distributed charge, but in general we have $v_0 = v_2 = 0$ for finite nuclei. Inserting these expansions into the radial Dirac equation gives

$$\begin{aligned} &r^{\gamma_p - 1}[v_0 p_0] + r^{\gamma_p}[(v_1 + E)p_0 - v_0 p_1] + \ldots \\ &\quad - r^{\gamma_q - 1}[c(\gamma_q - \kappa)]q_0 - r^{\gamma_q}[c(\gamma_q - \kappa + 1)]q_1 - \cdots = 0 \end{aligned} \tag{7.54a}$$

$$\begin{aligned} &r^{\gamma_q - 1}[v_0 q_0] + r^{\gamma_q}[(2c^2 + v_1 + E)q_0 - v_0 q_1] + \ldots \\ &\quad + r^{\gamma_p - 1}[c(\gamma_p + \kappa)]p_0 + r^{\gamma_p}[c(\gamma_p + \kappa + 1)]p_1 + \cdots = 0. \end{aligned} \tag{7.54b}$$

Equating powers of r, there are three cases to consider: $\gamma_p = \gamma_q$, $\gamma_p < \gamma_q$, and $\gamma_p > \gamma_q$.

CASE 1 $\gamma_p = \gamma_q = \gamma$

The lowest power of r in the Dirac equation gives the two equations

$$\begin{aligned} v_0 p_0 - c(\gamma - \kappa)q_0 &= 0 \\ v_0 q_0 + c(\gamma + \kappa)p_0 &= 0. \end{aligned} \tag{7.55}$$

If $v_0 = 0$, we have $\gamma = -\kappa$ *and* $\gamma = \kappa$ at the same time. This is an impossibility because κ is never zero, so this case applies exclusively to the point nucleus. Expressing q_0 in terms of p_0 and taking ratios,

$$v_0^2 = c^2(\kappa^2 - \gamma^2) \tag{7.56}$$

Since $v_0 = Z$, we get the definition of γ given earlier. The two equations above determine q_0 in terms of p_0, and the value of p_0 is determined by normalization. Equating higher powers of r yields expressions for the higher coefficients.

CASE 2 $\gamma_p < \gamma_q$

From (7.54a), this choice implies that $v_0 p_0 = 0$, from which it follows that $v_0 = 0$, and this case applies to a finite, or distributed, nuclear charge. The lowest term in (7.54b) then yields $c(\gamma_p + \kappa)p_0 = 0$, from which we deduce that $\gamma_p = -\kappa$, and thus $\kappa < 0$ in order for the solution to be normalizable.

We now need to determine γ_q. The term in $r^{\gamma_q - 1}$ must match the term in r^{γ_p}. If $\gamma_q \neq \gamma_p + 1$ then from (7.54a) we have $c(\gamma_q - \kappa)q_0 = 0$ and hence $\gamma_q = \kappa$. This contradicts the assumption that $\gamma_p < \gamma_q$. If the term in $r^{\gamma_q - 1}$ matches the term in $r^{\gamma_p + 1}$, then the term in r^{γ_p} yields $v_1 = -E$, which cannot be the case since v_1 must be linear in Z and E is quadratic in Z. We conclude that $\gamma_q = \gamma_p + 1$.

As for the point nucleus case, p_0 is determined by the normalization, and the other coefficients are determined relative to p_0. Because $v_0 = 0$, the lowest-order nonzero term in (7.54a) relates q_0 to p_0:

$$(v_1 + E)p_0 - c(\gamma_q - \kappa)q_0 = 0. \tag{7.57}$$

The r^{γ_p} term in (7.54b) has no matching term in $r^{\gamma_q - 1}$, and consequently $p_1 = 0$. The next term from the same equation relates p_2 to q_0. The r^{γ_p+1} term in (7.54a) (not listed) turns out to be zero because $v_0 = v_2 = 0$ and $p_1 = 0$. This leads to $q_1 = 0$. It can be shown that if the potential is an even function of r, then $p_n = q_n = 0$ for n odd.

CASE 3 $\gamma_p > \gamma_q$

Then $v_0 q_0 = 0$, and again $v_0 = 0$. The lowest term in (7.54a) gives $c(\gamma_q - \kappa)q_0 = 0$ from which $\gamma_q = \kappa$, and thus $\kappa > 0$. Similar arguments can be marshaled as for case 2 to show that $\gamma_p = \gamma_q + 1$ and to derive the coefficients. This case is the finite nucleus case for $\kappa > 0$.

At this point it is useful to summarize some of the features of the solutions for point and finite nuclear charge distributions.

POINT NUCLEUS CASE

(1) P and Q have the same behavior at the origin, and both are series in r (not r^2).

(2) The solutions for $|\kappa| = 1$ have a singularity at the origin. This could be a serious problem in quantum chemistry where Gaussian basis functions are commonly used to expand the wave function. The cusp that has to be represented in the Schrödinger solutions is now replaced by a singularity, which will inevitably make greater demands on the flexibility of the basis.

(3) For $\gamma > 1/2$, there is only one normalizable solution, that is, the regular solution with leading power r^γ. This is called the *limit point* case. For $\gamma < 1/2$ both the regular (r^γ) and irregular ($r^{-\gamma}$) solutions are normalizable, and the general solution is a linear combination of the two. This is called the *limit circle* case. The extra condition that the integral over the potential must also be finite eliminates the irregular solution. The point at which $\gamma < 1/2$ occurs for $Z/c = \sqrt{\kappa^2 - \frac{1}{4}}$, which for $\kappa = 1$ gives a Z value of approximately 119.

FINITE NUCLEUS CASE

(1) P and Q have the same lowest-order term as the corresponding nonrelativistic function of the same angular momentum at the origin. This means that the large component behaves like the nonrelativistic function, $P/r \sim r^\ell$, but the small component behavior depends on κ. For $\kappa < 0$, $Q/r \sim r^{\ell+1}$, and for $\kappa > 0$, $Q/r \sim r^{\ell-1}$. The spin–orbit splitting can then be understood in terms of the expectation of the potential over the small component, which is larger in magnitude for positive κ than for negative κ.

(2) For at least the first few terms, the solutions are either odd or even functions of r. If the nuclear potential is an even function for all r (e.g. for a Gaussian nuclear charge distribution), P and Q are odd or even for all r.

(3) At least inside the nucleus, P and Q are essentially *Gaussian* in shape. This means that in a method using a Gaussian basis set a nuclear charge distribution with a finite radius is preferred to a point nucleus: the basis then has the right behavior at the origin, and the demands on the basis are smaller because of the cutoff in the potential (Visser et al. 1987, Ishikawa et al. 1985).

We conclude from this discussion that for relativistic quantum chemistry it is preferable to adopt a model of the nucleus that has a charge distribution with a finite, rather than a zero, radius.

7.6 Nuclear Models

If a nuclear charge distribution with a finite radius is to be used, the question of the functional form of the distribution must be raised. The point nuclear model was simple: now we have to consider some form of the nuclear charge distribution that bears some relation to experimentally determined distributions. The Coulomb potential at a point $\mathbf{r}$ from a charge distribution $\rho(\mathbf{r})$ is

$$V(\mathbf{r}) = \int_{v'} \frac{\rho(\mathbf{r}')}{|r - r'|} \, \mathrm{d}\mathbf{r}'. \tag{7.58}$$

In relativistic atomic structure calculations, the two most common models are the uniformly charged sphere and the Fermi two-parameter distribution. The radial density for the uniformly charged sphere is given by

$$\begin{aligned} \rho^{\mathrm{nuc}}(r) &= \rho_0, \quad r \leq r^{\mathrm{nuc}} \\ &= 0, \quad r > r^{\mathrm{nuc}}. \end{aligned} \tag{7.59}$$

This is the simplest model, and gives rise to a quadratic potential inside the nucleus and a Coulombic potential outside:

$$\begin{aligned} V^{\mathrm{nuc}}(r) &= -\frac{Z}{2r^{\mathrm{nuc}}}\left[3 - \left(\frac{r}{r^{\mathrm{nuc}}}\right)^2\right], \quad r \leq r^{\mathrm{nuc}} \\ &= -\frac{Z}{r}, \quad r > r^{\mathrm{nuc}}. \end{aligned} \tag{7.60}$$

With a quadratic potential inside the nucleus, the wave function should be Gaussian in shape.

The Fermi two-parameter distribution has the form

$$\rho^{\mathrm{nuc}}(r) = \frac{\rho_0}{1 + \exp[(r - r^{\mathrm{nuc}})/s]}, \tag{7.61}$$

which introduces a smooth decay of the nuclear radial density rather than a sudden drop to zero. The potential is difficult to represent in closed form. It reduces to the uniform distribution in the limit $s \to 0$. A variant of this distribution, the three-parameter distribution, multiplies it by a quadratic function of r. This model has been used to fit a range of nuclear scattering data.

Both of these distributions present some problems for implementation in quantum chemical programs. The uniform sphere potential can be represented as the sum of the point nuclear potential and the difference between the point and uniform sphere nuclear potentials inside the nuclear radius. Then one has to perform integrals over a small finite region of space, which may have to be done numerically. A method for their evaluation has been presented by Matsuoka (1987). The Fermi distribution is difficult to represent in closed form and therefore the integrals would have to be performed numerically. An attractive alternative is to choose a Gaussian distribution,

$$\rho^{\mathrm{nuc}}(r) = Z\rho_0 \exp\left(-\zeta^{\mathrm{nuc}}\, r^2\right), \tag{7.62}$$

whose potential is simply

$$V^{\mathrm{nuc}}(r) = -\frac{Z}{r}\, \mathrm{erf}\left(\sqrt{\zeta^{\mathrm{nuc}}}\, r\right). \tag{7.63}$$

This is a convenient choice for quantum chemical calculations because the integrals are readily evaluated. For this choice, the nuclear charge distribution falls off too slowly for heavy atoms, but the effect is not large.

The principal parameter used in these distributions is the nuclear radius. For many nuclei the rms radius has been determined. These data could be used directly in calculations, but it is probably more useful in quantum chemistry, where the accuracy of the nuclear model is not very important, to use a fitted value. Two common fits to the cube root of the mass number A are used:

$$r^{\mathrm{rms}} = 1.2A^{1/3} \tag{7.64}$$

$$r^{\mathrm{rms}} = 0.836A^{1/3} + 0.57 \tag{7.65}$$

where r^{rms} is given in fm.[4] This value may be used directly in the uniform and Fermi distributions; for the Gaussian distribution it may be used to determine the exponent by matching the rms radii, so that

$$\zeta^{\mathrm{nuc}} = \frac{3}{2(r^{\mathrm{rms}})^2}. \tag{7.66}$$

The Fermi distribution has in addition the parameter s, which can be taken to be constant, determined from fits to scattering data.

---

[4] The latter has become a standard for the Gaussian nucleus in relativistic quantum chemistry (Visscher and Dyall 1997).

Table 7.3 Nuclear size and shape effects on the 1*s* orbital of Hg^{79+}

Nuclear model	*Eigenvalue*	*Difference*
Point	−3532.191 849	
		2.017 574
Uniform	−3530.174 275	
		0.007 881
Fermi	−3530.182 156	
		0.011 843
Gaussian	−3530.193 999	

The difference between the models is not large, as shown by the data in table 7.3 for hydrogen-like mercury. The effect of the finite radius is approximately 2 E_{h}, but the difference between the models is only 0.01 E_{h}. For the valence spinors of many-electron atoms, the difference between the models is well below the limits of chemical interest[5].

---

[5] Finite nuclear models have been discussed in detail by Andrae (2002).

8

Properties of Relativistic Mean-Field Theory

We have previously seen how the Dirac equation for one particle requires some rather special consideration and interpretation in order to arrive at a form that is able to treat electrons and positrons on an equal footing. These problems persist also when we go to systems with more than one electron. One might think that the extension to several electrons should not introduce dramatic changes. After all, we noted that even the one-electron problem must be viewed as a many-electron (and -positron) system in order to arrive at a consistent description. The problem with introducing more electrons is that electron–electron interactions that were previously small—for the one-electron case typically arising from vacuum polarization and self-interaction—now occur to the same order as the kinetic energy and the interaction with the potential. So while a perturbative approach such as QED can use the solutions of the one-electron Dirac equations as a very good starting approximation to a more accurate description of the full system, the same would not work for a system with more electrons because it would mean neglecting interactions of the same magnitude as the zeroth-order energy. For applications to quantum chemistry, the treatment of the entire electron–electron interaction as a perturbation would be hopelessly impractical, as it is even in many-electron relativistic atomic structure calculations.

The technique for dealing with this problem is well known from nonrelativistic calculations on many-electron systems. One-particle basis sets are developed by considering the behavior of the single electron in the mean field of all the other electrons, and while this neglects a smaller part of the interaction energy, the electron correlation, it provides a suitable starting point for further variational or perturbational treatments to recover more of the electron–electron interaction. It is only natural to pursue the same approach for the relativistic case. Thus one may proceed to construct a mean-field method that can be used as a basis for the perturbation theory of QED. In particular, the inclusion of the Breit interaction in the mean-field calculations ensures that the terms of $\mathcal{O}(\alpha^2)$ are included to infinite order in QED.

Deriving a mean-field or self-consistent field (SCF) method is not an entirely straightforward matter, however. In the QED reinterpretation, we take a particular approach to the negative-energy states. Now we come to develop a mean-field theory, do we regard these states as occupied or not? It is this question and the implications of the answers to it that will occupy the attention of the present chapter. As a first step, we analyze the traditional mean-field approach in a second-quantized form.

8.1 Mean-Field Formalism in Second Quantization

In deriving a set of relativistic one-particle functions, that is, spinors, from a mean-field approach, we typically start by making a simple guess at these functions (or the electron density), and then try to refine them iteratively. The refinement can be done by diagonalizing a suitable Hamiltonian (or Fock) matrix, which defines a rotation of the spinors in the entire function space available. Normally, this iterative process reaches a stage where further rotations do not change the spinors, that is, they are *self-consistent*. Provided we have chosen our sequence of rotations carefully, this should correspond to the optimal set of spinors from the mean field. For the present chapter our main concern is the rotation of the set of one-particle functions, and how this can be cast in a consistent theoretical framework that also accounts for the positron contributions.

The second-quantized Hamiltonian that defines our many-electron system is

$$\hat{\mathcal{H}} = \sum_{pq} h_{pq}\, a_p{}^{\dagger} a_q \;+ \tfrac{1}{2} \sum_{pqrs} (pq|rs)\, a_p{}^{\dagger} a_r{}^{\dagger} a_s\, a_q . \tag{8.1}$$

Most of the theory can be developed using only the one-electron part of this operator, with rather straightforward extensions to include the two-electron part. To effect the rotations in the function space we employ the exponential rotation operator $\hat{U} = \exp(i\hat{\lambda})$, introduced in (5.35), but parametrized in terms of the operator $\hat{\kappa} = i\hat{\lambda}$. We want the rotations to preserve orthonormality in the set of one-particle functions, and therefore require that $\hat{U}$ be a unitary operator, that is

$$\hat{U}\hat{U}^{\dagger} = e^{\hat{\kappa}} e^{\hat{\kappa}^{\dagger}} = 1. \tag{8.2}$$

The operator $\hat{\kappa}$ may be expressed as

$$\hat{\kappa} = \sum_{pq} \kappa_{pq}\, a_p{}^{\dagger} a_q . \tag{8.3}$$

The Hermitian conjugate of this operator is

$$\hat{\kappa}^{\dagger} = \sum_{pq} \kappa_{pq}{}^{\dagger}\, a_q{}^{\dagger} a_p \;= \sum_{pq} \kappa_{qp}{}^{\dagger}\, a_p{}^{\dagger} a_q . \tag{8.4}$$

From this it is easy to see that the unitarity condition is fulfilled if

$$\kappa_{pq} = -\kappa_{qp}^{*}, \tag{8.5}$$

that is, the matrix $\boldsymbol{\kappa}$ is anti-Hermitian.

The first step is to choose a reference state, or initial guess, which we will denote $|0\rangle$. To this reference state we apply the general rotation to obtain a new reference state,

$$|0'\rangle = \hat{U}|0\rangle = e^{\hat{\kappa}}|0\rangle. \tag{8.6}$$

The energy of the reference state in the rotated basis is the expectation of the Hamiltonian in the rotated basis:

$$E' = \langle 0'|\hat{\mathcal{H}}'|0'\rangle. \tag{8.7}$$

Since the Hamiltonian involves a complete sum over one-particle states, we can just as well express it in terms of the old basis. Using (8.6) to express the new reference in terms of the old, we get an expression for the energy of the rotated reference in terms of the original basis,

$$E' = \langle 0'|\hat{\mathcal{H}}|0'\rangle = \langle 0|e^{-\hat{\kappa}}\hat{\mathcal{H}}e^{\hat{\kappa}}|0\rangle. \tag{8.8}$$

The operator can be expanded in a commutator series

$$e^{-\hat{\kappa}}\hat{\mathcal{H}}e^{\hat{\kappa}} = \hat{\mathcal{H}} + \left[\hat{\mathcal{H}}, \hat{\kappa}\right] + \frac{1}{2!}\left[\left[\hat{\mathcal{H}}, \hat{\kappa}\right], \hat{\kappa}\right] + \cdots \tag{8.9}$$

so that the energy may be written

$$E' = \langle 0|\hat{\mathcal{H}}|0\rangle + \langle 0|\left[\hat{\mathcal{H}}, \hat{\kappa}\right]|0\rangle + \frac{1}{2!}\langle 0|\left[\left[\hat{\mathcal{H}}, \hat{\kappa}\right], \hat{\kappa}\right]|0\rangle + \cdots . \tag{8.10}$$

The energy at the new point depends on the Hamiltonian matrix elements and the parameters of the rotation. Not all of these parameters will necessarily cause a change in the energy, and we need to determine which parameters are the nonredundant ones.

Substituting $\hat{\kappa}$ into the first term in the energy expansion, (8.10),

$$\begin{aligned}\langle 0|\left[\hat{\mathcal{H}}, \hat{\kappa}\right]|0\rangle &= \sum_{pq}\kappa_{pq}\big(\langle 0|\hat{\mathcal{H}}a_p{}^{\dagger}a_q\ |0\rangle - \langle 0|a_p{}^{\dagger}a_q\ \hat{\mathcal{H}}|0\rangle\big) \\ &= \sum_{pq}\kappa_{pq}\big(\langle 0|\hat{\mathcal{H}}a_p{}^{\dagger}a_q\ |0\rangle - \langle 0|\hat{\mathcal{H}}a_q{}^{\dagger}a_p\ |0\rangle^{\dagger}\big)\end{aligned} \tag{8.11}$$

we may make the following observations:

- If q is the index of an unoccupied spinor, the action of a_q on the reference produces a zero result, and the first term in (8.11) is zero.
- If p is the index of an unoccupied spinor, the action of a_p on the reference produces a zero result, and the second term in (8.11) is zero.
- Therefore, if both p and q are unoccupied spinor indices, the right-hand side of (8.11) vanishes.

- If p and q are the indices of occupied spinors, we may permute $a_p{}^\dagger$ and a_q to get

$$\begin{aligned}\langle 0|\left[\hat{\mathcal{H}},\hat{\kappa}\right]|0\rangle &= \sum_{pq}\kappa_{pq}\big(\langle 0|\hat{\mathcal{H}}(\delta_{pq}-a_q\,a_p{}^\dagger)|0\rangle - \langle 0|\hat{\mathcal{H}}(\delta_{pq}-a_p\,a_q{}^\dagger)|0\rangle^\dagger\big)\\ &= \sum_{pq}\kappa_{pq}\big(\langle 0|\hat{\mathcal{H}}a_p\,a_q{}^\dagger|0\rangle^\dagger - \langle 0|\hat{\mathcal{H}}a_q\,a_p{}^\dagger|0\rangle\big).\end{aligned} \tag{8.12}$$

  Now the action of $a_p{}^\dagger$ and $a_q{}^\dagger$ on the reference gives a zero result because they are already occupied.

Similar arguments to these can be marshaled for the higher-order terms in (8.10). The conclusion is that the energy is invariant to rotations within the occupied spinor space and within the unoccupied spinor space. The matrix elements of $\hat{\kappa}$ within these spaces may therefore be set to zero, and the only matrix elements to be considered are those that connect occupied and unoccupied spinors.

To make the energy stationary, we differentiate with respect to the nonredundant parameters of $\boldsymbol{\kappa}$ and set the derivative to zero. For this purpose, we write the energy expansion explicitly in terms of the parameters as

$$\begin{aligned}E' &= \langle 0|\hat{\mathcal{H}}|0\rangle + \sum_{pq}\kappa_{pq}\langle 0|\left[\hat{\mathcal{H}},\,a_p{}^\dagger a_q\right]|0\rangle\\ &\quad + \tfrac{1}{2}\sum_{pqrs}\kappa_{pq}\kappa_{rs}\langle 0|\left[\left[\hat{\mathcal{H}},\,a_p{}^\dagger a_q\right],\,a_r{}^\dagger a_s\right]|0\rangle + \cdots\\ &= E + \sum_{pq}\kappa_{pq}g_{pq} + \tfrac{1}{2}\sum_{pqrs}\kappa_{pq}\kappa_{rs}\tilde{G}_{pq,rs} + \cdots .\end{aligned} \tag{8.13}$$

The matrix $\mathbf{g}$ is the *gradient* of the energy with respect to the parameters of the rotation, and the matrix $\tilde{\mathbf{G}}$ is related to the matrix of second derivatives. Differentiating the energy with respect to κ_{pq} we find

$$\frac{\partial E'}{\partial \kappa_{pq}} = g_{pq} + \tfrac{1}{2}\sum_{rs}\kappa_{rs}(\tilde{G}_{pq,rs}+\tilde{G}_{rs,pq}) + \cdots . \tag{8.14}$$

Truncating the expression at the dots gives a set of linear equations for $\boldsymbol{\kappa}$, which can be solved by standard methods. The rotation operator $\hat{U}$ may then be constructed and the spinors rotated to give a new initial guess. This will not be the exact solution because we have truncated the energy expression to second order to obtain the rotation parameters. However, it should give a good—or at least a better—approximation to the exact solution. The procedure is repeated until the spinor gradient is zero. At this point we have arrived at a stationary point of the energy as a function of the rotation parameters.

We could equally well have differentiated the energy expression with respect to the complex conjugate of κ_{pq} since this may be regarded as an independent parameter. This is equivalent to differentiating with respect to $-\kappa_{qp}$, and in fact we end up with the same result as (8.14) after reindexing, if the gradient is an anti-Hermitian matrix. This makes sense, because the gradient appears in the energy expression contracted with $\boldsymbol{\kappa}$, which is an anti-Hermitian matrix.

The second derivative of the energy with respect to the rotation parameters,

$$\left.\frac{\partial^2 E}{\partial \kappa^*_{pq}\kappa_{rs}}\right|_{\kappa_{pq}=\kappa_{rs}=0} = -\tfrac{1}{2}(\tilde{G}_{qp,rs} + \tilde{G}_{rs,qp}) = G_{pq,rs}, \tag{8.15}$$

is called the *Hessian*[1] matrix. Since the energy expression actually contains the product of the *conjugate* of the rotation operator with the Hamiltonian times the rotation operator, we have differentiated with respect to $\kappa^*_{pq} = -\kappa_{qp}$. The Hessian gives us information about the nature of the stationary point. If we have reached the ground state, all the eigenvalues of the Hessian should be positive. If we have reached an excited state, on the other hand, there should be as many negative eigenvalues of the Hessian as there are states below the state reached. Other stationary points are characterized by other conditions, but for the present these two will suffice.

Up to now, our discussion has been quite general and independent of the form of the Hamiltonian. To obtain explicit expressions for the gradient and the Hessian, we must use the Hamiltonian from (8.1), but as previously remarked it is initially sufficient to consider only the one-electron part. We assume that the many-electron wave function is described by a single determinant. The gradient is

$$g^{(1)}_{pq} = \langle 0| \sum_{rs} \left[a_r{}^{\dagger} a_s \; h_{rs} \, , \, a_p{}^{\dagger} a_q \right] |0\rangle. \tag{8.16}$$

By using the commutation relations for creation and annihilation operators, we can get two rank 4 expressions that cancel, and we are left with

$$g^{(1)}_{pq} = \sum_r \Big(\langle 0| a_r{}^{\dagger} a_q \; |0\rangle h_{rp} - \langle 0| a_p{}^{\dagger} a_r \; |0\rangle h_{qr} \Big). \tag{8.17}$$

The Fock space matrix element is called the *first-order density matrix* or one-particle density matrix,

$$D_{pq} = \langle 0| a_p{}^{\dagger} a_q \; |0\rangle. \tag{8.18}$$

Using this definition, we may write the gradient as

$$g^{(1)}_{pq} = \sum_r \big(D_{rq} h_{rp} - D_{pr} h_{qr} \big). \tag{8.19}$$

If p and q are both occupied spinor indices, the first-order density matrix element is a delta function, $D_{pq} = \delta_{pq}$, and the gradient matrix element is zero. This is in accord

---

[1] We give the Hessian the symbol **G** here, where in many other places it is given the symbol **H**; we do this to avoid confusion with the Hamiltonian.

with the findings above, in which the energy was not dependent on the parameters for two occupied spinors. Similarly, if both p and q are unoccupied spinor indices, the first-order density matrix elements are zero, and again the energy does not depend on the rotation parameters. However, if p is occupied and q is unoccupied, $a_q\,|0\rangle$ vanishes, and $g^{(1)}_{pq} = -h_{qp}$. Similarly, if p is unoccupied and q is occupied, $g^{(1)}_{pq} = h_{qp}$. Thus for any pair of indices t, u, where we have

$$g^{(1)}_{tu} = h_{ut}, \tag{8.20}$$

we get

$$g^{(1)}_{ut} = -h_{tu} = -g^{(1)*}_{tu} \tag{8.21}$$

by virtue of the Hermiticity of **h**. The gradient is therefore an anti-Hermitian matrix, as we had supposed. The condition for a stationary state is that the matrix elements h_{pq} are zero for $p \neq q$, as we should expect.

The Hessian matrix is

$$G^{(1)}_{pq,rs} = 2(D_{qs}h_{pr} + D_{rp}h_{sq}) - \sum_t \Big(\delta_{rp}(D_{ts}h_{tq} + D_{qt}h_{st}) + \delta_{qs}(D_{tp}h_{tr} + D_{rt}h_{pt})\Big). \tag{8.22}$$

The matrix elements are nonzero only for q and s unoccupied and p and r occupied (or vice versa), and the Hessian is then

$$G^{(1)}_{pq,rs} = 2(\delta_{pr}h_{sq} - \delta_{sq}h_{pr}). \tag{8.23}$$

If we ensure that **h** is diagonal, including the occupied–occupied and unoccupied–unoccupied blocks, which, since they are arbitrary, we are always free to do, the Hessian reduces to

$$G^{(1)}_{pq,pq} = 2(h_{qq} - h_{pp}). \tag{8.24}$$

If p is occupied and q is unoccupied in the ground state, the Hessian eigenvalues (which in this case are just the diagonal elements given above) are all positive. The state is therefore a minimum with respect to rotations among the spinors. The first excited state will have one $h_{qq} < h_{pp}$, and therefore one of the eigenvalues of the Hessian will be negative. This state represents a saddle point of the energy surface spanned by rotations in the spinor space—it is a maximum with respect to rotations involving the spinor q for which $h_{qq} < h_{pp}$, but a minimum with respect to all other rotations. If we wish to optimize this state, we must employ a *minimax* technique. There are several ways of doing this, including counting eigenvalues of the Hessian, making use of maximum overlap between iterations to follow the desired root, inverting the negative eigenvalues of the Hessian in the equations used to obtain the rotation parameters, and so on.

The two-electron contribution to the spinor gradient is given by

$$g_{pq}^{(2)} = \langle 0|\tfrac{1}{2}\sum_{rstu}\left[a_r^\dagger a_t^\dagger a_u\, a_s\, (rs|tu),\, a_p^\dagger a_q\,\right]|0\rangle. \tag{8.25}$$

Similarly to our treatment of the one-electron contribution, we can commute operators to achieve a rank reduction through cancellation of two terms, and we are left with

$$\begin{aligned} g_{pq}^{(2)} = \tfrac{1}{2}\sum_{rstu}(rs|tu)\, \langle 0|\,&\delta_{ps}\, a_r^\dagger a_t^\dagger a_u\, a_q \;+\, \delta_{pu}\, a_r^\dagger a_t^\dagger a_q\, a_s \\ &-\,\delta_{tq}\, a_r^\dagger a_p^\dagger a_u\, a_s \;-\, \delta_{rq}\, a_p^\dagger a_t^\dagger a_u\, a_s\;|0\rangle. \end{aligned} \tag{8.26}$$

Reindexing the integrals enables a reduction to two terms,

$$g_{pq}^{(2)} = \sum_{rst}(rs|tp)\,\langle 0|a_r^\dagger a_t^\dagger a_q\, a_s\;|0\rangle - \sum_{stu}(qs|tu)\,\langle 0|a_p^\dagger a_t^\dagger a_u\, a_s\;|0\rangle. \tag{8.27}$$

The Fock space matrix element is called the *second-order density matrix* or two-particle density matrix,

$$P_{pqrs} = \langle 0|a_p^\dagger a_r^\dagger a_s\, a_q\;|0\rangle, \tag{8.28}$$

which is zero unless all four indices are of occupied spinors. For a single determinant, the value of the matrix element is then

$$P_{pqrs} = \delta_{pq}\delta_{rs} - \delta_{ps}\delta_{rq}. \tag{8.29}$$

The second-order density matrix is invariant to the interchange of the two index pairs pq and rs.

Combining the one- and two-electron contributions to the gradient, we get

$$g_{pq} = \sum_r\Big(D_{rq}h_{rp} + \sum_{st}P_{strq}\,(st|rp)\Big) - \sum_r\Big(D_{pr}h_{qr} + \sum_{st}P_{prst}\,(qr|st)\Big). \tag{8.30}$$

The terms in brackets may appear familiar, and are indeed nothing but the expression for the usual spinor or spin–orbital Fock matrix. If we insert the values of the density matrices for a single determinant, we get

$$f_{qp} = h_{qp} + \sum_r\Big((rr|qp) - (qr|rp)\Big) \tag{8.31}$$

where q is an occupied spinor index. The gradient defines the generalized Fock matrix

$$g_{pq} = f_{qp} - f_{pq}^*. \tag{8.32}$$

The Fock matrix defined in terms of the densities in (8.30) is only nonzero if the second index is an occupied spinor index. A set of spinors that transforms the Fock matrix into a diagonal matrix corresponds to a gradient where all elements are zero, and thus corresponds to a stationary point of the energy surface. The form of the Fock matrix in (8.31) is used to define the matrix for all indices p and q, but, as already noted, the choice of the occupied–occupied and unoccupied–unoccupied blocks has no influence on the energy or on the parameters of the optimization. It can of course influence the rate of convergence of a self-consistent field procedure.

We may similarly derive the two-electron terms for the Hessian, which are presented here for completeness:

$$\begin{aligned} G^{(2)}_{pqrs} = 2\sum_{vw} \Big(& P_{rpvw}\big\{(sq|vw) - (sw|vq)\big\} + P_{qsvw}\big\{(pr|vw) - (pw|vr)\big\} \\ & - P_{vpws}(vq|wr) - P_{qvrw}(pv|sw)\Big) \\ - \sum_{uvw} \Big(& \delta_{pr}\big\{P_{quvw}(su|vw) + P_{wvus}(wv|uq)\big\} \\ & + \delta_{qs}\big\{P_{ruvw}(pu|vw) + P_{wvup}(wv|ur)\big\}\Big). \end{aligned} \tag{8.33}$$

8.2 Structure of the Spinor Rotation Operator

Having carried out this analysis of the variational process, we are now equipped to consider the question raised at the start of this chapter—what role should the negative-energy spinors play in this process? As we saw, the various contributions to the gradient and the Hessian above depend on whether the spinors involved are occupied or not in the many-electron state being optimized. The variational process is therefore directly influenced by our choice of model for the negative-energy states: are they filled or not? Here, we define three possible approaches, in which the definition of the occupied and unoccupied spinors depends on the interpretation used. In the first, which we will call the *empty Dirac* approach, the negative-energy states are empty, and they are not reinterpreted as positron states. The representation of the orbital rotation matrix $\hat{\kappa}$ in this approach is

$$\begin{aligned} \hat{\kappa} = \sum_{pq} \big[& \kappa_{p^+q^+}\, a_{p^+}{}^{\dagger} a_{q^+} \;+\; \kappa_{p^+q^-}\, a_{p^+}{}^{\dagger} a_{q^-} \\ & + \kappa_{p^-q^+}\, a_{p^-}{}^{\dagger} a_{q^+} \;+\; \kappa_{p^-q^-}\, a_{p^-}{}^{\dagger} a_{q^-} \;\big] \end{aligned} \tag{8.34}$$

where the $+$ and $-$ superscripts label the positive-energy and negative-energy states, respectively. The unoccupied spinors consist of the unoccupied electron spinors and the negative-energy spinors. In the second approach, the *filled Dirac* approach, the negative-energy states are filled, so that the occupied spinors include these states, and the only empty spinors are the unoccupied electron spinors. The representation of $\hat{\kappa}$ is the same as for the empty Dirac approach. In the third approach, the negative-energy states are reinterpreted according to the QED approach as positive-energy positrons, and the operators are presented in normal-ordered form,

$$:\hat{\kappa}: = \sum_{pq} \left[\kappa_{pq}\, a_p{}^{\dagger} a_q \; + \; \kappa_{p\tilde{q}}\, a_p{}^{\dagger} \tilde{a}_q{}^{\dagger} \; + \; \kappa_{\tilde{p}q}\, \tilde{a}_p\, a_q \; - \; \kappa_{\tilde{p}\tilde{q}}\, \tilde{a}_q{}^{\dagger} \tilde{a}_p \right] \tag{8.35}$$

where the tildes label positrons. The unoccupied spinors consist of the unoccupied electron spinors and the positron spinors, just as in the empty Dirac approach. Thus the structure of $\hat{\kappa}$ will depend on whether the negative-energy states are filled or not.

If no reinterpretation is employed, that is, if the negative-energy states are kept as negative-energy states, the structure of $\hat{\kappa}$ is the same as that used in the energy expansion, (8.13), and the conclusions from the general considerations above are valid. The nonredundant parameters of $\hat{\kappa}$ are those that connect occupied and unoccupied spinors. In the empty Dirac approach the unoccupied spinors are the empty electron spinors of positive and negative energies. In the filled Dirac picture the only unoccupied spinors are the empty electron spinors, and the occupied spinors consist of the occupied electron spinors plus the negative-energy spinors. For a system where we have n spinors of both positive and negative energy and n_o occupied electron spinors, the parameter space in the empty Dirac picture is of the size $n_o(2n - n_o)$, whereas in the filled Dirac picture it is of the size $(n + n_o)(n - n_o)$, which will generally be much larger.

In the QED approach $:\hat{\kappa}:$ also contains terms of the form $a_p{}^{\dagger}\tilde{a}_q{}^{\dagger}, \tilde{a}_p\, a_q$ and $\tilde{a}_q{}^{\dagger}\tilde{a}_p$, which were not covered by the discussion of the previous section. We therefore need to investigate how these terms affect the first-order correction to the energy, $\langle 0 | \left[\hat{\mathcal{H}}, \hat{\kappa} \right] | 0 \rangle$. However, we must apply the normal ordering *after* we have performed the commutation, because the product of two normal-ordered operators is not itself a normal-ordered operator but must be placed in normal order.

To illustrate this approach, it is again sufficient to consider the one-electron part of the Hamiltonian, and the relevant commutator in the Dirac approach (filled or empty) becomes

$$\begin{aligned} \left[\hat{\mathcal{H}}, \hat{\kappa} \right] = \Bigg[& \sum_{rs} \left(h_{r^+s^+}\, a_{r^+}{}^{\dagger} a_{s^+} + h_{r^+s^-}\, a_{r^+}{}^{\dagger} a_{s^-} + h_{r^-s^+}\, a_{r^-}{}^{\dagger} a_{s^+} + h_{r^-s^-}\, a_{r^-}{}^{\dagger} a_{s^-} \right), \\ & \sum_{pq} \Big(\kappa_{p^+q^+}\, a_{p^+}{}^{\dagger} a_{q^+} \; + \; \kappa_{p^+q^-}\, a_{p^+}{}^{\dagger} a_{q^-} \\ & \qquad + \kappa_{p^-q^+}\, a_{p^-}{}^{\dagger} a_{q^+} \; + \; \kappa_{p^-q^-}\, a_{p^-}{}^{\dagger} a_{q^-} \; \Big) \Bigg] \end{aligned} \tag{8.36}$$

where we have used the same indexing conventions for $\hat{h}$ and $\hat{\kappa}$. The evaluation of the various terms in this expression is straightforward, but somewhat lengthy. Thus, after commutation of operators and reindexing, the first term gives

$$\begin{aligned} & \sum_{rs} \sum_{pq} \left[h_{r^+s^+}\; a_{r^+}{}^{\dagger} a_{s^+}, \kappa_{p^+q^+}\; a_{p^+}{}^{\dagger} a_{q^+} \right] \\ & \quad = \sum_{rq} a_{r^+}{}^{\dagger} a_{q^+} \sum_{s} \left(\kappa_{s^+q^+} h_{r^+s^+} - \kappa_{r^+s^+} h_{s^+q^+} \right). \end{aligned} \tag{8.37}$$

Collecting all the terms arising from evaluation of the commutator, we get from the Dirac approach (empty or filled)

$$\hat{\mathcal{H}}' = \left[\hat{\mathcal{H}}, \hat{\kappa}\right] = \sum_{rq} \left(a_{r^+}{}^{\dagger} a_{q^+} \ h'_{r^+q^+} + a_{r^+}{}^{\dagger} a_{q^-} \ h'_{r^+q^-} \right.$$
$$\left. + a_{r^-}{}^{\dagger} a_{q^+} \ h'_{r^-q^+} + a_{r^-}{}^{\dagger} a_{q^-} \ h'_{r^-q^-} \right) \qquad (8.38)$$

where the matrix elements are defined by

$$h'_{ab} = \sum_{r} h_{ar^+}\kappa_{r^+b} \ + \ h_{ar^-}\kappa_{r^-b} \ - \ \kappa_{ar^+}h_{r^+b} \ - \ \kappa_{ar^-}h_{r^-b} \qquad (8.39)$$

with a and b any index, positive or negative. The QED approach gives a normal-ordered one-particle operator

$$:\hat{\mathcal{H}}':=:\left[\mathcal{H}, \hat{\kappa}\right]:= \sum_{rq} \left(a_r{}^{\dagger} a_q \ h'_{rq} \ + \ a_r{}^{\dagger} \tilde{a}_q{}^{\dagger} h'_{r\tilde{q}} \ + \ \tilde{a}_r \ a_q \ h'_{\tilde{r}q} \ - \ \tilde{a}_q{}^{\dagger} \tilde{a}_r \ h'_{\tilde{r}\tilde{q}} \right) \qquad (8.40)$$

where the matrix elements are defined by

$$h'_{ab} = \sum_{r} h_{ar}\kappa_{rb} + h_{a\tilde{r}}\kappa_{\tilde{r}b} - \kappa_{ar}h_{rb} - \kappa_{a\tilde{r}}h_{\tilde{r}b} \qquad (8.41)$$

with a and b any index, electron or positron. Note that this matrix is the same as in (8.39); all that differs is the convention for labeling the indices.

If we now take the reference expectation of this operator in the QED approach,

$$\langle 0| :\hat{\mathcal{H}}': |0\rangle = \sum_{pq} \Big(\langle 0| a_p{}^{\dagger} a_q \ |0\rangle h'_{pq} + \langle 0| a_p{}^{\dagger} \tilde{a}_q{}^{\dagger} \ |0\rangle h'_{p\tilde{q}}$$
$$+ \langle 0| \tilde{a}_p \ a_q \ |0\rangle h'_{\tilde{p}q} - \langle 0| \tilde{a}_q{}^{\dagger} \tilde{a}_p \ |0\rangle h'_{\tilde{p}\tilde{q}} \Big), \qquad (8.42)$$

we find that the second and third terms vanish because the operators create and destroy pairs and hence connect states of different particle number, and the fourth term vanishes because there are no positrons in the reference, leaving only the first term. Taking the expectation of the Dirac approach operator, (8.38),

$$\langle 0| \hat{\mathcal{H}}' |0\rangle = \sum_{pq} \Big(\langle 0| a_{p^+}{}^{\dagger} a_{q^+} \ |0\rangle h'_{p^+q^+} + \langle 0| a_{p^+}{}^{\dagger} a_{q^-} \ |0\rangle h'_{p^+q^-}$$
$$+ \langle 0| a_{p^-}{}^{\dagger} a_{q^+} \ |0\rangle h'_{p^-q^+} + \langle 0| a_{p^-}{}^{\dagger} a_{q^-} \ |0\rangle h'_{p^-q^-} \Big) \qquad (8.43)$$

the last three terms vanish in the empty Dirac approach because there are no negative-energy states occupied in the reference, and again only the first term survives.

Thus we find that in both the QED approach and the empty Dirac approach all terms except the first vanish, because the reference contains only electrons in positive-energy states. But in the filled Dirac approach all terms in principle survive. Thus the properly normal-ordered QED approach gives the same result for this term in the energy expansion as the empty Dirac approach, and not the same as the filled Dirac approach. The parameter space for the QED approach is therefore *identical* with that of the empty Dirac approach.

This is not quite the end of the matter. What we have assumed in the QED approach is that the reference vacuum is that of the current guess. If we were to take an absolute reference, such as the free-particle vacuum, the normal ordering should take place with respect to this fixed vacuum, and then the QED approach would give the same results as the filled Dirac approach, in which rotations between the negative-energy states and the unoccupied electron states affect the energy. By this means a vacuum polarization term has been introduced into the procedure, but without the renormalization term[2]. In atomic structure calculations in which QED effects are introduced, the many-particle states employed are usually the Dirac–Fock states (Mittleman 1981), that is, those that result from the "empty Dirac" picture. We will therefore take as our reference the QED approach with the "floating vacuum"—a vacuum defined with respect to the current set of spinors.

The conclusion of this analysis is that the normal-ordered QED approach as presented here, with a floating vacuum, is equivalent to the empty Dirac approach. It appears that the reinterpretation of the negative-energy states as positron states has *no* influence on the combination of matrix elements that results from the commutator. Normal ordering then only affects the terms involving the positron operators, and at least for the one-electron Hamiltonian this means that the reference energy will be identical in both the empty Dirac and the QED approaches, since they only have occupied electron states, and the terms that survive in the reference expectation value are identical in the two approaches.

8.3 Relativistic Stationarity Conditions

Having determined which of the parameters of κ are nonredundant, we can proceed with the development of the mean-field theory using the methods presented in section 8.1. Again it is sufficient to treat only the one-electron part of the Hamiltonian for purposes of detailed derivation. The gradient must now be expanded to include the negative-energy states, and it can easily be derived from equations (8.42) and (8.43). Where both indices in the gradient refer to states with positive energies, the gradient is the same as that derived above,

$$g^{(1)}_{p^+q^+} = \sum_r \left(h_{rp^+} D_{rq^+} - h_{q^+r} D_{p^+r}\right), \tag{8.44}$$

---

[2] See Chaix and Iracene (1989) for a discussion of this kind of approach.

and has the same properties. The new term comes from one positive-energy and one negative-energy index,

$$g^{(1)}_{p^+q^-} = \sum_r \left(h_{rp^+}D_{rq^-} - h_{q^-r}D_{p^+r}\right). \tag{8.45}$$

The first density matrix is zero for all r, and the second reduces to a delta function in which p must be an occupied spinor index, giving

$$g^{(1)}_{p^+q^-} = -h_{q^-p^+}. \tag{8.46}$$

A similar expression arises if the first index is a negative-energy spinor index, while the element with two negative-energy indices obviously vanishes. The gradient is then simply related to the one-electron matrix, which on diagonalization will make the gradient zero.

Of more interest is the Hessian, because this determines what sort of state we have. When all four indices are electron indices, the result is the same for all approaches, and is given by (8.22) and the following equations. The important—and interesting—case is where two indices are negative-energy indices. As we noted at the end of the preceding section, the empty Dirac and the QED pictures give the same result. In the empty Dirac approach, the expression is derived quite simply by inserting the appropriate indices in (8.22), giving

$$\begin{aligned} G^{(1)}_{p^+q^-,r^+s^-} &= 2\left(D_{q^-s^-}h_{p^+r^+} + D_{r^+p^+}h_{s^-q^-}\right) \\ &\quad - \sum_t \left[\delta_{rp}\left(D_{ts^-}h_{tq^+} + D_{q^-t}h_{s^-t}\right) + \delta_{qs}\left(D_{tp^+}h_{tr^+} + D_{r^+t}h_{p^+t}\right)\right]. \end{aligned} \tag{8.47}$$

Of course, q^- and s^- are unoccupied, so that if p and r are occupied spinor indices and $\mathbf{h}$ is diagonal,

$$G^{(1)}_{p^+q^-,r^+s^-} = \delta_{pr}\delta_{qs}\, 2\left(h_{q^-q^-} - h_{p^+p^+}\right) \tag{8.48}$$

which is the same as the general result but with negative-energy spinor indices for q. Now, however, $h_{q^-q^-} < -2mc^2$, so that the Hessian eigenvalue is large and negative. This means that the eigenstate we have labeled as the ground state is not in fact the "ground state" of the Dirac one-electron Hamiltonian matrix $\mathbf{h}$, but an *excited* state: in fact it is the $(N + 1)$th excited state, where there are N negative-energy solutions. The ground state of the Dirac matrix is the lowest of the negative-energy states, and does indeed collapse towards minus infinity as we increase the size of the basis set. The phenomenon of variational collapse is therefore a real one, but it is irrelevant to the issue of bound states. Provided that we can label the eigenvalues, which we can always do in a finite basis set, we can always identify the state that corresponds to the "real" ground state, the one that has the minimum energy in the reinterpreted QED approach.

An important point in this discussion, and one that appears to have been missed all too often, is that the QED approach is *always* based on some set of one-particle functions, for which there are both positive- and negative-energy states. It is only *after* the reinterpretation and the normal ordering that we have a Hamiltonian in which the ground state is the electron state of lowest energy, and that the positron states appear as positive-energy states of particles with a positive charge. Therefore, the spinor optimization process will always have to contend with a set of negative-energy states, whose existence is the mathematical embodiment in a single formalism of the occurrence of particles and antiparticles.

The implication of the discussion above is that the mean-field theory we are looking for is not one of *minimization* of the energy with respect to the rotation parameters, but one in which we are seeking an excited state in the space of one-particle functions. It is therefore a *minimax* problem (Talman 1986). As noted above, the procedures for attacking this problem are well-established in quantum chemistry. In fact, there is likely to be no problem with convergence to the correct solution, even in a first-order SCF method, because the separation between the positive- and negative-energy states is very large and the coupling between them is small. This has been found to be the case in practice in a large number of four-component SCF calculations on atoms and molecules.

8.4 Projection and Bounds

If we are seeking an excited state, it is worth asking whether there are any bounds to the solution that would give us confidence (apart from experience) that our calculations will not collapse. In nonrelativistic variational theory, the ground-state energy is bounded from below by the exact energy, and the iterative solution of the self-consistent field equations converges to the exact energy from above. For an excited state there is no such variational lower bound, but we may obtain some kind of bound from a consideration of the properties of matrices and eigenvalues.

Suppose we have a matrix **h** for which we wish to find the first excited state, that is, the second-lowest eigenstate. From the discussion above, we know that it is a minimum with respect to rotations with all the higher-lying states in the initial guess. Therefore, if we project out the parts of the matrix involving the ground state, we end up with a matrix in which the state of interest is now the ground state, and will converge from above to the desired solution. The excited state is bounded from below by the projection.

Now if we reintroduce the ground state and diagonalize the matrix, the excited state solution *must* go *up*, provided that the projected eigenvalue is higher than the eigenvalue of the ground state. (If it does not, the initial guess must be very poor.) This is a consequence of theorems involving the diagonalization of a matrix and its submatrices and the relations between the eigenvalues.

Applying this to the Dirac one-electron Hamiltonian, the lowest electron state is the state of interest. Projecting out the negative-energy states leaves only the positive-energy block of the matrix. The lowest electron state in the projected matrix is therefore bounded from below, and we may minimize its energy to obtain a projected eigenvalue. Now, if we reintroduce the negative-energy states, there is a large gap between the projected eigenvalue and the highest negative-energy eigenvalue, which will be at about $-2mc^2$. The projected eigenvalue must therefore go up to the eigenvalue of the full matrix. The true eigenvalue is therefore bounded from below by the projected eigenvalue. Second-order perturbation theory, combined with the information that the coupling between the positive and negative energy states is $\mathcal{O}(c^{-1})$, then shows that the energy can go up by at most $\mathcal{O}(c^{-4})$.

What this means for mean-field theory is that the lowest electron eigenvalue of the one-particle matrix that we are diagonalizing can *never* fall below the lowest eigenvalue of the positive–positive block of the matrix in any one iteration, and therefore there is no problem with variational collapse in a self-consistent field procedure, provided that the set of states in which we have formed the matrix represent the solutions of some one-particle Dirac equation. Failure to ensure a proper representation in a finite basis has been the occasion of problems that appear to exhibit variational collapse. Further discussion of this issue will be postponed to chapter 11, which covers finite basis methods.

It has been argued that in order to prevent variational collapse, it is *necessary* to introduce a projection onto some set of positive-energy states (Sucher 1980, 1984). This, as we have just seen, is not the case. Even if it were the case, there is no prescription for choosing the set of positive-energy states onto which a projection must be made, and in the absence of any such prescription, one might as well choose the self-consistent field states (Mittleman 1981).

8.5 Many-Electron Theory

In all of the preceding discussion, we have simply assumed that we have some kind of one-particle potential V, which is used to set up the one-particle Hamiltonian matrix. We should now consider the effect of the electron–electron interaction in the discussion. Intuition tells us that since the solutions are mostly dominated by the nuclear Coulomb attraction, things will not change much, if at all.

The second-quantized Hamiltonian in the general case is

$$\hat{\mathcal{H}} = \sum_{pq} h_{pq}\, a_p{}^{\dagger} a_q \; + \tfrac{1}{2} \sum_{pqrs} (pq|rs)\, a_p{}^{\dagger} a_r{}^{\dagger} a_s\, a_q. \tag{8.49}$$

Rewriting this in terms of positron and electron indices is more involved, because the two-electron part produces 16 terms. Permutational symmetry and reindexing permits the combining of six of these, and the result in the empty Dirac picture is

$$\begin{aligned}
\hat{\mathcal{H}} = \sum_{pq} \Big(& h_{p^+q^+}\, a_{p^+}{}^{\dagger} a_{q^+} + h_{p^+q^-}\, a_{p^+}{}^{\dagger} a_{q^-} \\
& + h_{p^-q^+}\, a_{p^-}{}^{\dagger} a_{q^+} + h_{p^-q^-}\, a_{p^-}{}^{\dagger} a_{q^-} \Big) \\
+ \tfrac{1}{2} \sum_{pqrs} \Big(& (p^+q^+|r^+s^+) a_{p^+}{}^{\dagger} a_{r^+}{}^{\dagger} a_{s^+}\, a_{q^+} + 2(p^+q^+|r^+s^-) a_{p^+}{}^{\dagger} a_{r^+}{}^{\dagger} a_{s^-}\, a_{q^+} \\
& + 2(p^+q^+|r^-s^+) a_{p^+}{}^{\dagger} a_{r^-}{}^{\dagger} a_{s^+}\, a_{q^+} + 2(p^+q^+|r^-s^-) a_{p^+}{}^{\dagger} a_{r^-}{}^{\dagger} a_{s^-}\, a_{q^+} \\
& + 2(p^+q^-|r^-s^+) a_{p^+}{}^{\dagger} a_{r^-}{}^{\dagger} a_{s^+}\, a_{q^-} + (p^+q^-|r^+s^-) a_{p^+}{}^{\dagger} a_{r^+}{}^{\dagger} a_{s^-}\, a_{q^-} \\
& + (p^-q^+|r^-s^+) a_{p^-}{}^{\dagger} a_{r^-}{}^{\dagger} a_{s^+}\, a_{q^+} + 2(p^-q^-|r^-s^+) a_{p^-}{}^{\dagger} a_{r^-}{}^{\dagger} a_{s^+}\, a_{q^-} \\
& + 2(p^-q^-|r^+s^-) a_{p^-}{}^{\dagger} a_{r^+}{}^{\dagger} a_{s^-}\, a_{q^-} + (p^-q^-|r^-s^-) a_{p^-}{}^{\dagger} a_{r^-}{}^{\dagger} a_{s^-}\, a_{q^-} \Big).
\end{aligned} \tag{8.50}$$

Applying the reinterpretation of the negative-energy states and normal-ordering the result, we get the QED Hamiltonian

$$\begin{aligned}
:\hat{\mathcal{H}}:=\sum_{pq}\Big(& h_{pq}\, a_p^\dagger a_q + h_{p\tilde{q}}\, a_p^\dagger \tilde{a}_q^\dagger + h_{\tilde{p}q}\, \tilde{a}_p\, a_q - h_{\tilde{p}\tilde{q}}\, \tilde{a}_q^\dagger \tilde{a}_p \Big) \\
+\tfrac{1}{2}\sum_{pqrs}\Big(& (pq|rs)\, a_p^\dagger a_r^\dagger a_s\, a_q + 2(pq|r\tilde{s})\, a_p^\dagger a_r^\dagger \tilde{a}_s^\dagger a_q + 2(pq|\tilde{r}s)\, a_p^\dagger \tilde{a}_r\, a_s\, a_q \\
& -2(pq|\tilde{r}\tilde{s})\, a_p^\dagger \tilde{a}_s^\dagger \tilde{a}_r\, a_q + 2(p\tilde{q}|\tilde{r}s)\, a_p^\dagger \tilde{a}_q^\dagger \tilde{a}_r\, a_s \\
& +(p\tilde{q}|r\tilde{s})\, a_p^\dagger a_r^\dagger \tilde{a}_s^\dagger \tilde{a}_q^\dagger + (\tilde{p}q|\tilde{r}s)\, \tilde{a}_p\, \tilde{a}_r\, a_s\, a_q - 2(\tilde{p}\tilde{q}|\tilde{r}s)\, \tilde{a}_q^\dagger \tilde{a}_p\, \tilde{a}_r\, a_s \\
& -2(\tilde{p}\tilde{q}|r\tilde{s})\, a_r^\dagger \tilde{a}_s^\dagger \tilde{a}_q^\dagger \tilde{a}_p + (\tilde{p}\tilde{q}|\tilde{r}\tilde{s})\, \tilde{a}_s^\dagger \tilde{a}_q^\dagger \tilde{a}_p\, \tilde{a}_r \Big).
\end{aligned} \tag{8.51}$$

The first, fourth, fifth, and tenth of the two-electron terms all preserve the numbers of particles, and it is these terms that will contribute to the energy. The second and third and the eighth and ninth terms are terms that create or annihilate one pair, and the sixth and seventh terms create and annihilate two pairs, respectively.

With all these terms, the derivation of the gradient and Hessian will be a very tedious exercise. However, we may make use of results from the previous sections to analyze the terms that will be required and the terms that may be ignored. First, it was noted that the commutators in the QED and Dirac approaches gave the same matrix elements, regardless of whether the negative-energy states were reinterpreted as the conjugates of positron states or not. Second, the density matrices are only nonzero when all indices are those of occupied spinors, which in the present case means that they are electron indices. The normal ordering of the QED operators will therefore have no influence on the density matrices. It follows that the QED and the empty Dirac approaches must give the same results, so that it is only necessary to consider the empty Dirac approach. Third, the contributions from the positron spinors can be derived by inserting a tilde on the appropriate indices in the expression for the electron contributions. The only terms that survive are those in which none of the creation operators and annihilation operators—or alternatively none of the density matrices—have tilded indices.

With these principles, the gradients for positive–negative rotations are determined to be

$$\begin{aligned}
g_{\tilde{p}q} &= \sum_r \Big(D_{rq} h_{r\tilde{p}} + \sum_{st} P_{strq}\ (st|r\tilde{p}) \Big) = f_{q\tilde{p}}, \\
g_{p\tilde{q}} &= -\sum_r \Big(D_{pr} h_{\tilde{q}r} + \sum_{st} P_{prst}\ (\tilde{q}r|st) \Big) = -f^*_{p\tilde{q}}.
\end{aligned} \tag{8.52}$$

The Fock matrix is identical in form to the electron Fock matrix given above, and the gradient is the same in form as the spinor gradient involving one unoccupied electron spinor. Thus, for the purposes of self-consistent field methods, we may treat the negative-energy spinors as simply an extension of the set of unoccupied spinors. The Hessian may similarly be defined using the principles and formulas set out above.

Part III

Four-Component Methodology

9

Operators, Matrix Elements, and Wave Functions under Time-Reversal Symmetry

We now take on the task of developing the theory and methods for a relativistic quantum chemistry. The aim is to arrive at a qualitative as well as a quantitative understanding of the relativistic effects in molecules. We must be able to predict the effects of relativity on the wave functions and electron densities of molecules, and on the molecular properties arising from these. And we must develop methods and algorithms that enable us to calculate the properties and interactions of molecules with an accuracy comparable to that achieved for lighter systems in a nonrelativistic framework. Parts of this development follow fairly straightforwardly from our considerations of the atomic case in part II, but molecular systems represent challenges of their own. This is particularly true for the computational techniques. From the nonrelativistic experience we know that present-day quantum chemistry owes much of its success to the enormous effort that has gone into developing efficient methods and algorithms. This effort has yielded powerful tools, such as the use of basis-set expansions of wave functions, the exploitation of molecular symmetry, the description of correlation effects by calculations beyond the mean-field approximation, and so on. In developing a relativistic quantum chemistry, we must be able to reformulate these techniques in the new framework, or replace them by more suitable and efficient methods.

In nonrelativistic theory, spin symmetry provides one of the biggest reductions in computational effort, such as in the powerful and elegant Graphical Unitary Group Approach (GUGA) for configuration interaction (CI) calculations (Shavitt 1988). For relativistic applications, time-reversal symmetry takes the place of spin symmetry, and this chapter is devoted to developing a formalism for efficient incorporation of this symmetry in our theory and methods. Time-reversal symmetry includes the spin symmetry of nonrelativistic systems, but there are significant differences from spin symmetry for systems with a Hamiltonian that is spin-dependent.

The development of techniques that incorporate time-reversal symmetry presented here are primarily aimed at four-component calculations, but they are equally applicable to two-component calculations in which the spin-dependent operators are included at the self-consistent field (SCF) stage of a calculation.

9.1 Time Reversal and Kramers-Restricted Representation of Operators

In the previous discussion of the symmetry of the Dirac equation (chapter 6), it was shown that the Dirac equation was symmetric under time reversal, and that the fermion functions occur in Kramers pairs where the two members are related by time reversal. We will have to deal with a variety of operators, and in most cases the methodologies will be developed in the absence of an external magnetic field, or with the magnetic field considered as a perturbation. Consequently, we can make the developments in terms of a basis of Kramers pairs, which are the natural representation of the wave function in a system that is symmetric under time reversal. The development here is primarily the development of a second-quantized formalism. We will use the term *Kramers-restricted* to cover techniques and methods based on spinors that in some well-defined way appear as Kramers pairs. The analogous nonrelativistic situation is the spin-restricted formalism, which requires that orbitals appear as pairs with the same spatial part but with α and β spins respectively. Spin restriction thus appears as a special case of Kramers restriction, because of the time-reversal connection between α and β spin functions.

The first step in the implementation of time-reversal symmetry is to classify the basic operators according to their behavior under time reversal effected by the operator $\hat{\mathcal{K}} = -i\Sigma_y\hat{\mathcal{K}}_0$:

- The position operator is symmetric: $\hat{\mathcal{K}}\mathbf{r} = \mathbf{r}\hat{\mathcal{K}}$.
- The momentum operator is antisymmetric: $\hat{\mathcal{K}}\mathbf{p} = -\mathbf{p}\hat{\mathcal{K}}$.
- The Dirac velocity operator $\boldsymbol{\alpha}$ is antisymmetric: $\hat{\mathcal{K}}\boldsymbol{\alpha} = -\boldsymbol{\alpha}\hat{\mathcal{K}}$.
- The Dirac kinetic energy operator $c\boldsymbol{\alpha}\cdot\mathbf{p}$ is therefore symmetric, as it should be.
- The β matrix is symmetric: $\hat{\mathcal{K}}\beta = \beta\hat{\mathcal{K}}$.
- The scalar potential V is symmetric: $\hat{\mathcal{K}}V = V\hat{\mathcal{K}}$.
- For a radiation field in the Coulomb gauge, the vector potential is real and hence symmetric under time reversal: $\hat{\mathcal{K}}\mathbf{A} = \mathbf{A}\hat{\mathcal{K}}$.
- The interaction of a current with the vector radiation field $c\boldsymbol{\alpha}\cdot\mathbf{A}$ is therefore antisymmetric.

Our basis of Kramers pairs is $\{\psi_p, \psi_{\bar{p}}\}$. We adopt the convention that general spinors are labeled p, q, r, s, occupied spinors are labeled i, j, k, l, unoccupied spinors are labeled a, b, c, d, and active spinors (where necessary) are labeled t, u, v, w. The time-reversed conjugate of a function or operator is denoted by a bar. We place the bar over the index of a function rather than the function, that is, we use $\psi_{\bar{p}}$ rather than $\bar{\psi}_p$. However, the two are equivalent, and we place the bar over the function when there is no index.

To develop relations under time reversal, we use the creation operator $a_p^\dagger$ corresponding to ψ_p, that is

$$\psi_p = a_p^\dagger|vac\rangle. \tag{9.1}$$

Time-reversing this expression yields

$$\hat{\mathcal{K}}a_p^\dagger|vac\rangle = \hat{\mathcal{K}}\psi_p = \psi_{\bar{p}} = a_{\bar{p}}^\dagger|vac\rangle, \tag{9.2}$$

while another application of $\hat{\mathcal{K}}$ provides the expected sign change

$$\hat{\mathcal{K}} a_{\bar{p}}^{\dagger} |vac\rangle = \hat{\mathcal{K}} \psi_{\bar{p}} = -\psi_{p} = -a_{p}^{\dagger} |vac\rangle. \tag{9.3}$$

We therefore have the following result for time reversal of creation operators

$$\hat{\mathcal{K}} a_{p}^{\dagger} = a_{\bar{p}}^{\dagger} \hat{\mathcal{K}}; \qquad \hat{\mathcal{K}} a_{\bar{p}}^{\dagger} = -a_{p}^{\dagger} \hat{\mathcal{K}} \tag{9.4}$$

and by conjugation we get the corresponding relations for annihilation operators,

$$\hat{\mathcal{K}} a_{p} = a_{\bar{p}} \hat{\mathcal{K}}; \qquad \hat{\mathcal{K}} a_{\bar{p}} = -a_{p} \hat{\mathcal{K}}. \tag{9.5}$$

In this basis, the second-quantized representation of a general one-particle operator $\hat{\Omega}$ is

$$\hat{\Omega} = \sum_{pq} \left[\Omega_{pq} a_{p}^{\dagger} a_{q} + \Omega_{p\bar{q}} a_{p}^{\dagger} a_{\bar{q}} + \Omega_{\bar{p}q} a_{\bar{p}}^{\dagger} a_{q} + \Omega_{\bar{p}\bar{q}} a_{\bar{p}}^{\dagger} a_{\bar{q}} \right]. \tag{9.6}$$

Let us for a moment consider the nonrelativistic case of the second-quantized representation of operators. Here, time reversal flips the spin function between α and β, and in a spin-restricted formalism the basis of Kramers pairs is $\{\phi_{p}\alpha, \phi_{p}\beta\}$, where ϕ is the spatial part of the orbital. Normally the operators used in nonrelativistic calculations are real and spin free, and we can then use the representation

$$\begin{aligned} \hat{\Omega} &= \sum_{pq} \left[\Omega_{pq} a_{p}^{\dagger} a_{q} + \Omega_{\bar{p}\bar{q}} a_{\bar{p}}^{\dagger} a_{\bar{q}} \right] \\ &= \sum_{pq} \Omega_{pq} \left(a_{p}^{\dagger} a_{q} + a_{\bar{p}}^{\dagger} a_{\bar{q}} \right). \end{aligned} \tag{9.7}$$

We see that we can define a new basis generated by the operators

$$\hat{E}_{pq} = a_{p}^{\dagger} a_{q} + a_{\bar{p}}^{\dagger} a_{\bar{q}} \tag{9.8}$$

and then the operator in this basis is

$$\hat{\Omega} = \sum_{pq} \Omega_{pq} \hat{E}_{pq}. \tag{9.9}$$

The $\hat{E}_{pq}$ operators are known as the generators of the unitary group of order N, and are extensively used in nonrelativistic theory. They are also known as one-particle excitation operators. The interested reader can find a more extensive discussion of these operators in the books listed in the bibliography, such as Jørgensen and Simons (1981) and Helgaker et al. (2000). Here we will explore the possibilities of an analogous development for the relativistic case.

The relativistic basis is no longer the set of products of orbital functions with α and β spin functions, but general four-component spinors grouped as Kramers pairs. Likewise, the operators $\hat{\Omega}$ are no longer necessarily spin free. If we apply the time-reversal operator to matrix elements of $\hat{\Omega}$, we can derive some relations between matrix elements

$$\begin{aligned} \hat{\mathcal{K}}\Omega_{pq} &= \hat{\mathcal{K}}\langle p|\hat{\Omega}|q\rangle = \langle \bar{p}|\hat{\bar{\Omega}}|\bar{q}\rangle = \langle p|\hat{\Omega}|q\rangle^* = \Omega_{pq}^* \\ \hat{\mathcal{K}}\Omega_{\bar{p}q} &= \hat{\mathcal{K}}\langle \bar{p}|\hat{\Omega}|q\rangle = -\langle p|\hat{\bar{\Omega}}|\bar{q}\rangle = \langle \bar{p}|\hat{\Omega}|q\rangle^* = \Omega_{\bar{p}q}^*. \end{aligned} \tag{9.10}$$

The last two equalities follow from the fact that matrix elements are simple numbers, and application of the time-reversal operator generates the complex conjugate of the number.

We should be prepared to handle operators that are antisymmetric under time reversal as well as those that are symmetric. We introduce a sign factor t to keep track of this behavior, that is

$$\hat{\bar{\Omega}} = \hat{\mathcal{K}}\hat{\Omega}\hat{\mathcal{K}}^{-1} = \mathrm{t}\hat{\Omega}; \quad \mathrm{t} = \pm 1. \tag{9.11}$$

Using these relations, we can represent $\hat{\Omega}$ as

$$\hat{\Omega} = \sum_{pq} \left[\Omega_{pq} a_p^{\dagger} a_q + \mathrm{t}\Omega_{pq}^* a_{\bar{p}}^{\dagger} a_{\bar{q}} + \Omega_{\bar{p}q} a_{\bar{p}}^{\dagger} a_q - \mathrm{t}\Omega_{\bar{p}q}^* a_p^{\dagger} a_{\bar{q}} \right]. \tag{9.12}$$

Here, although we have not reduced the number of terms in the expression, we have reduced the number of unique matrix elements by half.

The operators encountered within the relativistic framework can be complex as well as spin-dependent. To recover something similar to the $\hat{E}$ operators above, we must split the matrix elements of $\hat{\Omega}$ into real and imaginary parts:

$$\begin{aligned} \hat{\Omega} = \sum_{pq} \big[&\mathrm{Re}(\Omega_{pq})(a_p^{\dagger} a_q + \mathrm{t} a_{\bar{p}}^{\dagger} a_{\bar{q}}) + i\,\mathrm{Im}(\Omega_{pq})(a_p^{\dagger} a_q - \mathrm{t} a_{\bar{p}}^{\dagger} a_{\bar{q}}) \\ &+ \mathrm{Re}(\Omega_{\bar{p}q})(a_{\bar{p}}^{\dagger} a_q - \mathrm{t} a_p^{\dagger} a_{\bar{q}}) + i\,\mathrm{Im}(\Omega_{\bar{p}q})(a_{\bar{p}}^{\dagger} a_q + \mathrm{t} a_p^{\dagger} a_{\bar{q}}) \big]. \end{aligned} \tag{9.13}$$

The first term contains an operator reminiscent of the nonrelativistic $\hat{E}_{pq}$, and so we are led to the relativistic analogs

$$\hat{E}_{pq}^{s} = (a_p^{\dagger} a_q + s a_{\bar{p}}^{\dagger} a_{\bar{q}}); \qquad \hat{E}_{\bar{p}q}^{s} = (a_{\bar{p}}^{\dagger} a_q - s a_p^{\dagger} a_{\bar{q}}), \tag{9.14}$$

where s is a sign factor that takes the values ± 1. Using these operators, we can cast $\hat{\Omega}$ in the form

$$\hat{\Omega} = \sum_{pq} \left[\mathrm{Re}(\Omega_{pq})\,\hat{E}_{pq}^{\mathrm{t}} + i\mathrm{Im}(\Omega_{pq})\,\hat{E}_{pq}^{-\mathrm{t}} + \mathrm{Re}(\Omega_{\bar{p}q})\,\hat{E}_{\bar{p}q}^{\mathrm{t}} + i\mathrm{Im}(\Omega_{\bar{p}q})\,\hat{E}_{\bar{p}q}^{-\mathrm{t}} \right]. \tag{9.15}$$

These $\hat{E}^{s}_{pq}$ operators are the equivalent of the singlet and triplet excitation operators of nonrelativistic theory: $\hat{E}^{+}_{pq}$ is the singlet operator and the remaining three are the triplet operators. The presence of triplet operators should not be surprising in a formalism in which spin is incorporated from the start. Although we are dealing with general operators and matrix elements, and there is no explicit mention of spin, the basis for the matrix elements is a general, complex, mixed-spin basis, which is appropriate for operators that incorporate spin.

Two features of this representation of $\hat{\Omega}$ are noteworthy. First, we still have the same number of terms as in the original primitive representation. Second, we have a mixture of $\hat{E}^{+}_{pq}$ and $\hat{E}^{-}_{pq}$ operators, and thus we have not achieved a basis set reduction, while the nonrelativistic $\hat{E}_{pq}$ allowed us to formally halve the size of the basis. The conclusion is that even if it is possible to construct operators analogous to $\hat{E}_{pq}$ for the relativistic case, there is less, if anything, to be gained compared with the nonrelativistic case because we must evaluate expressions involving both triplet and singlet operators. This is to be expected because the operators include spin-dependent terms. These excitation operators have their uses in correlated methods, and there are reductions that occur due to point group symmetry, which will be considered later.

To complete the development of this operator formalism, we can define the two-particle excitation operators,

$$\hat{e}^{s_1 s_2}_{pq,rs} = \hat{E}^{s_1}_{pq}\hat{E}^{s_2}_{rs} - \delta_{rq}\hat{E}^{s_1 s_2}_{ps}. \tag{9.16}$$

The operator with $s_1 = s_2 = 1$ is identical in form to the nonrelativistic two-particle excitation operator. To arrive at expressions for the time-reversed operators, any index in this expression can be replaced with the barred index. The final expression for the Dirac–Coulomb Hamiltonian is

$$\begin{aligned}
\hat{\mathcal{H}} = \sum_{pq} \Big[& \mathrm{Re}\, h_{pq}\hat{E}^{+}_{pq} + i\,\mathrm{Im}\, h_{pq}\hat{E}^{-}_{pq} + \mathrm{Re}\, h_{\bar{p}q}\hat{E}^{+}_{\bar{p}q} + i\,\mathrm{Im}\, h_{\bar{p}q}\hat{E}^{-}_{\bar{p}q}\Big] \\
+ \frac{1}{2}\sum_{pqrs} \Big[& (\mathrm{Re}\, pq|\,\mathrm{Re}\, rs)\hat{e}^{++}_{pq,rs} + i(\mathrm{Re}\, pq|\,\mathrm{Im}\, rs)\hat{e}^{+-}_{pq,rs} \\
& + i(\mathrm{Im}\, pq|\,\mathrm{Re}\, rs)\hat{e}^{-+}_{pq,rs} - (\mathrm{Im}\, pq|\,\mathrm{Im}\, rs)\hat{e}^{--}_{pq,rs} \\
& + (\mathrm{Re}\, p\bar{q}|\,\mathrm{Re}\, \bar{r}s)\hat{e}^{++}_{p\bar{q},\bar{r}s} + i(\mathrm{Re}\, p\bar{q}|\,\mathrm{Im}\, \bar{r}s)\hat{e}^{+-}_{p\bar{q},\bar{r}s} \\
& + i(\mathrm{Im}\, p\bar{q}|\,\mathrm{Re}\, \bar{r}s)\hat{e}^{-+}_{p\bar{q},\bar{r}s} - (\mathrm{Im}\, p\bar{q}|\,\mathrm{Im}\, \bar{r}s)\hat{e}^{--}_{p\bar{q},\bar{r}s} \\
& + (\mathrm{Re}\, pq|\,\mathrm{Re}\, r\bar{s})\hat{e}^{++}_{pq,r\bar{s}} + i(\mathrm{Re}\, pq|\,\mathrm{Im}\, r\bar{s})\hat{e}^{+-}_{pq,r\bar{s}} \\
& + i(\mathrm{Im}\, pq|\,\mathrm{Re}\, r\bar{s})\hat{e}^{-+}_{pq,r\bar{s}} - (\mathrm{Im}\, pq|\,\mathrm{Im}\, r\bar{s})\hat{e}^{--}_{pq,r\bar{s}} \\
& + (\mathrm{Re}\, \bar{p}q|\,\mathrm{Re}\, rs)\hat{e}^{++}_{\bar{p}q,rs} + i(\mathrm{Re}\, \bar{p}q|\,\mathrm{Im}\, rs)\hat{e}^{+-}_{\bar{p}q,rs} \\
& + i(\mathrm{Im}\, \bar{p}q|\,\mathrm{Re}\, rs)\hat{e}^{-+}_{\bar{p}q,rs} - (\mathrm{Im}\, \bar{p}q|\,\mathrm{Im}\, rs)\hat{e}^{--}_{\bar{p}q,rs}\Big].
\end{aligned} \tag{9.17}$$

Here, to obtain an expression in terms of the $\hat{e}$ operators, we had to partition each two-electron integral into four parts, by the real and imaginary parts of each density. The expressions in (9.10) were used to combine integrals with common density components.

In this Hamiltonian we separate the matrix elements into real and imaginary parts. This is not always convenient, and the formalism is no more compact than the original. There is, however, an alternative. We can invoke the properties of $\hat{\Omega}$ under Hermitian conjugation,

$$\Omega^*_{pq} = \mathrm{h}\Omega_{qp}; \qquad \Omega^*_{\bar{p}q} = \mathrm{h}\Omega_{q\bar{p}}; \qquad \mathrm{h} = \pm 1. \tag{9.18}$$

Using these expressions with some regrouping and reindexing, we can cast the expression for $\hat{\Omega}$ in the form

$$\hat{\Omega} = \sum_{pq} \big[\Omega_{pq}(a_p{}^\dagger a_q \; + \mathrm{th}\; a_{\bar{q}}{}^\dagger a_{\bar{p}}\,) + \tfrac{1}{2}\Omega_{p\bar{q}}(a_p{}^\dagger a_{\bar{q}} \; - \mathrm{th}\; a_q{}^\dagger a_{\bar{p}}\,) \\ + \tfrac{1}{2}\Omega_{\bar{p}q}(a_{\bar{p}}{}^\dagger a_q \; - \mathrm{th}\; a_{\bar{q}}{}^\dagger a_p)\big]. \tag{9.19}$$

The quantities in parentheses suggest the introduction of a new set of operators defined by (Aucar et al. 1995)

$$\begin{aligned} \hat{X}^{\mathrm{s}}_{pq} &= a_p{}^\dagger a_q \; + \mathrm{s}a_{\bar{q}}{}^\dagger a_{\bar{p}} \\ \hat{X}^{\mathrm{s}}_{p\bar{q}} &= a_p{}^\dagger a_{\bar{q}} \; - \mathrm{s}a_q{}^\dagger a_{\bar{p}} \\ \hat{X}^{\mathrm{s}}_{\bar{p}q} &= a_{\bar{p}}{}^\dagger a_q \; - \mathrm{s}a_{\bar{q}}{}^\dagger a_p \end{aligned} \tag{9.20}$$

where s is a sign factor. These operators can now be used to write the one-electron operator as

$$\hat{\Omega} = \sum_{pq} [\Omega_{pq}\hat{X}^{\mathrm{th}}_{pq} + \frac{1}{2}\Omega_{p\bar{q}}\hat{X}^{\mathrm{th}}_{p\bar{q}} + \frac{1}{2}\Omega_{\bar{p}q}\hat{X}^{\mathrm{th}}_{\bar{p}q}]. \tag{9.21}$$

This expression represents a simplification of the original expansion in the primitive Kramers pairs basis. Although there is not a reduction in the size of the one-particle basis, we need only consider half the matrix elements, and there is therefore a 50% reduction in the amount of work. The lack of reduction might be expected because the matrix elements are potentially spin-dependent. The $\hat{X}^{\pm}_{pq}$ operators are called *Kramers single-replacement operators*, and they define what we will call a *Kramers basis*.

The $\hat{X}^{\pm}_{pq}$ operators have the following features:

1. They give the most compact representation of an operator.
2. There is no division of matrix elements into real and imaginary parts.
3. The behavior under time reversal is

$$\hat{\mathcal{K}}\hat{X}^{\mathrm{s}}_{pq}\hat{\mathcal{K}}^{-1} = \mathrm{s}\;\hat{X}^{\mathrm{s}}_{qp}; \quad \hat{\mathcal{K}}\hat{X}^{\mathrm{s}}_{p\bar{q}}\hat{\mathcal{K}}^{-1} = \mathrm{s}\;\hat{X}^{\mathrm{s}}_{\bar{p}q} \tag{9.22}$$

that is, they transform into their conjugates.

4. The symmetry under time reversal and Hermitian conjugation is transparently displayed in the sign index of $\hat{X}^{\pm}_{pq}$.

This is in contrast to the $\hat{E}^{\pm}_{pq}$ operators, which have the following properties:

1. They give a representation of an operator in terms of real matrices.
2. They are genuine excitation operators, which can be used in the same way as in nonrelativistic theory, and must be used for coupled-cluster theory.
3. The behavior under time reversal[1] is:

$$\hat{\mathcal{K}}\hat{E}^{\mathrm{s}}_{pq}\hat{\mathcal{K}}^{-1} = s\hat{E}^{\mathrm{s}}_{pq}; \qquad \hat{\mathcal{K}}\hat{E}^{\mathrm{s}}_{p\bar{q}}\hat{\mathcal{K}}^{-1} = s\hat{E}^{\mathrm{s}}_{p\bar{q}}. \tag{9.23}$$

A further difference between the $\hat{X}^{\pm}_{pq}$ and $\hat{E}^{\pm}_{pq}$ is in their relation to point group symmetry, to which we will return later.

There is an even more compact way of defining the Kramers single-replacement operators, $\hat{X}^{\pm}_{pq}$, which also has the advantage of displaying the permutational symmetry. To do this, we introduce two auxiliary operators. One is the *bar-reversal operator*, $\hat{K}_p$, which is the time-reversal operator for a spinor with index p. The effect of this operator is

$$\hat{K}_p\psi_p = \psi_{\bar{p}}; \qquad \hat{K}_p\psi_{\bar{p}} = -\psi_p; \qquad \hat{K}_p\psi_q = \psi_q, \tag{9.24}$$

where $p \neq q$. In terms of annihilation and creation operators, the effect of $\hat{K}_p$ is

$$\hat{K}_p a_p = a_{\bar{p}}; \qquad \hat{K}_p a_{\bar{p}} = -a_p; \qquad \hat{K}_p a_q = a_q. \tag{9.25}$$

The other auxiliary operator is the *Kramers permutation operator*, $\hat{T}_{pq}$, whose action is given by

$$\hat{T}_{pq} a_p^{\dagger} a_q = a_{\bar{q}}^{\dagger} a_{\bar{p}}, \qquad \hat{T}_{pq} a_{\bar{p}}^{\dagger} a_q = -a_{\bar{q}}^{\dagger} a_p \tag{9.26}$$

and which may be considered as the application of the time-reversal operator followed by Hermitian conjugation. These two operators commute because

$$\begin{aligned} \hat{K}_p\hat{T}_{pq} a_p^{\dagger} a_q &= \hat{K}_p a_{\bar{q}}^{\dagger} a_{\bar{p}} &= -a_{\bar{q}}^{\dagger} a_p \\ \hat{T}_{pq}\hat{K}_p a_p^{\dagger} a_q &= \hat{T}_{pq} a_{\bar{p}}^{\dagger} a_q &= -a_{\bar{q}}^{\dagger} a_p. \end{aligned} \tag{9.27}$$

The Kramers single-replacement operators, $\hat{X}^{\mathrm{s}}_{pq}$, can now be expressed in terms of these auxiliary operators,

$$\hat{X}^{\mathrm{s}}_{pq} = \left(1 + \mathrm{s}\hat{T}_{pq}\right) a_p^{\dagger} a_q, \tag{9.28}$$

---

[1] These operators can be made symmetric under time reversal by multiplying the $\hat{E}^{-}$ operators by i; however, this makes it more difficult to define the two-particle excitation operators e.

and the operation of $\hat{K}_p$ and $\hat{K}_q$ on this expression defines the other members of the set:

$$\hat{X}^{\mathrm{s}}_{p\bar{q}} = \hat{K}_q \hat{X}^{\mathrm{s}}_{pq} = \left(1 + \mathrm{s}\hat{T}_{pq}\right) a_p{}^{\dagger} a_{\bar{q}} \; = a_p{}^{\dagger} a_{\bar{q}} \; - \mathrm{s} a_q{}^{\dagger} a_{\bar{p}}. \tag{9.29}$$

The reduction of products of second-quantized operators makes use of commutators of the replacement operators. The various commutators of Kramers single-replacement operators are given by application of bar reversal to the basic commutator relation

$$\left[\hat{X}^{\mathrm{s}_1}_{pq}\,,\, \hat{X}^{\mathrm{s}_2}_{rs}\right] = (1 + \mathrm{s}_2\hat{T}_{pq})(1 - \mathrm{s}_1\mathrm{s}_2\hat{T}_{ps}\hat{T}_{rq})(\delta_{rq} a_p{}^{\dagger} a_s \; - \delta_{ps} a_r{}^{\dagger} a_q). \tag{9.30}$$

The Kramers permutation operator must be taken to apply to the Kronecker delta as well as the operators. Thus, for $\mathrm{s}_1 = \mathrm{s}_2 = +1$ we have

$$\left[\hat{X}^{+}_{pq}\,,\, \hat{X}^{+}_{rs}\right] = \delta_{rq}\hat{X}^{-}_{ps} - \delta_{ps}\hat{X}^{-}_{rq}. \tag{9.31}$$

Using these auxiliary operators, it is easy to define Kramers double-replacement operators in an analogous manner:

$$\hat{x}^{\mathrm{s}_1\mathrm{s}_2}_{pq,rs} = \left(1 + \mathrm{s}_1\hat{T}_{pq}\right)\left(1 + \mathrm{s}_2\hat{T}_{rs}\right) a_p{}^{\dagger} a_r{}^{\dagger} a_s \; a_q \; = \hat{x}^{\mathrm{s}_2\mathrm{s}_1}_{rs,pq}, \tag{9.32}$$

and the $\hat{x}$ operators with bars follow from the application of the bar-reversal operators. The double-replacement operators can be expressed in terms of the single-replacement operators, but not in the same simple manner as in nonrelativistic theory:

$$\hat{x}^{\mathrm{s}_1\mathrm{s}_2}_{pq,rs} = \hat{X}^{\mathrm{s}_1}_{pq}\hat{X}^{\mathrm{s}_2}_{rs} - \delta_{rq} a_p{}^{\dagger} a_s \; - \mathrm{s}_1\delta_{r\bar{p}} a_{\bar{q}}{}^{\dagger} a_s \; - \mathrm{s}_2\delta_{\bar{s}q} a_p{}^{\dagger} a_{\bar{r}} \; - \mathrm{s}_1\mathrm{s}_2\delta_{\bar{p}\bar{s}} a_{\bar{q}}{}^{\dagger} a_{\bar{r}}. \tag{9.33}$$

The Kronecker delta terms with an odd number of bars have been retained both to show the structure and to enable the expressions for the other $\hat{x}$ operators to be derived using the bar-reversal operator.

Finally, the commutator of a single- and a double-replacement operator is given by

$$\begin{aligned}\left[\hat{X}^{\mathrm{s}_1}_{tu}\, \hat{x}^{\mathrm{s}_2\mathrm{s}_3}_{pq,rs}\right] = {} & (1 + \mathrm{s}_2 T_{pq})(1 + \mathrm{s}_3 T_{rs}) \\ & \times \left[(1 + \mathrm{s}_1\mathrm{s}_2 T_{pu} T_{tq})(\delta_{pu} a_t{}^{\dagger} a_r{}^{\dagger} a_s \; a_q \; - \delta_{tq} a_p{}^{\dagger} a_r{}^{\dagger} a_s \; a_u \,)\right. \\ & \quad \left. + (1 + \mathrm{s}_1\mathrm{s}_3 T_{ru} T_{ts})(\delta_{ru} a_t{}^{\dagger} a_p{}^{\dagger} a_q \; a_s \; - \delta_{ts} a_r{}^{\dagger} a_p{}^{\dagger} a_q \; a_u \,)\right].\end{aligned} \tag{9.34}$$

9.2 Matrix Elements under Time Reversal

We have already examined the matrix elements of a one-particle operator. The integrals of the one-particle Hermitian operator $\hat{f}$ have the following relations:

$$f_{\bar{p}\bar{q}} = \mathrm{th} f_{qp} = \mathrm{t} f^{*}_{pq}; \qquad f_{\bar{p}q} = -\mathrm{th} f_{\bar{q}p} = -\mathrm{t} f^{*}_{p\bar{q}}. \tag{9.35}$$

Thus, the application of time reversal reduces the number of unique matrix elements by a factor of 2.

Matrix elements of two-particle operators arise primarily either from the Coulomb or the Breit (or Gaunt) interaction,

$$(pq|rs) = \int \mathrm{d}^3 r_i \int \mathrm{d}^3 r_j \, \psi_p{}^{\dagger}(i) \psi_r{}^{\dagger}(j) V^{\mathrm{C}}(i, j) \psi_s(j) \psi_q(i) \tag{9.36}$$

and

$$(p\boldsymbol{\alpha}q|r\boldsymbol{\alpha}s) = \int \mathrm{d}^3 r_i \int \mathrm{d}^3 r_j \, \psi_p{}^{\dagger}(i) \psi_r{}^{\dagger}(j) V^{\mathrm{B}}(i, j) \psi_s(j) \psi_q(i), \tag{9.37}$$

where V^{C} and V^{B} are the Coulomb and Breit interactions, respectively.

The Coulomb integral $(pq|rs)$ is essentially a matrix element over two charge densities. In the relativistic case the permutational symmetry of the integrals is reduced relative to the nonrelativistic case because the functions are in general complex. Thus, we have

$$(pq|rs) = (qp|sr)^* \neq (qp|rs). \tag{9.38}$$

Particle interchange symmetry is of course retained. The Kramers permutation operator is effectively a unit operator when applied to a charge density, for

$$\begin{aligned} \hat{T}_{pq} \psi_p{}^{\dagger} \psi_q &= (\psi_{\bar{p}}{}^{\dagger} \psi_{\bar{q}})^{\dagger} = \psi_{\bar{q}}{}^{\dagger} \psi_{\bar{p}} \\ &= \hat{\mathcal{K}}(\psi_{\bar{p}}{}^{\dagger} \psi_{\bar{q}}) = \hat{\mathcal{K}}(\hat{\mathcal{K}} \psi_p{}^{\dagger} \psi_q) = \psi_p{}^{\dagger} \psi_q \end{aligned} \tag{9.39}$$

where we have used the antiunitarity of the time-reversal operator. The charge density is a product of two fermion functions and behaves as a boson under time reversal, that is, $\hat{\mathcal{K}}^2 = 1$. The relations between the 16 types of two-electron integrals arising from all possible combinations of barred and unbarred spinors can then be determined with the use of this operator and the bar-reversal operator. Thus, for the Coulomb interaction,

$$\begin{aligned} (pq|rs) &= \hat{T}_{rs}(pq|rs) &&= \hat{T}_{pq}(pq|rs) &&= \hat{T}_{pq}\hat{T}_{rs}(pq|rs) \\ &= (pq|\bar{s}\bar{r}) &&= (\bar{q}\bar{p}|rs) &&= (\bar{q}\bar{p}|\bar{s}\bar{r}) \end{aligned} \tag{9.40}$$

giving a reduction of a factor of 4. Application of the bar-reversal operator to this expression yields expressions for all the other integral types, and we see that time reversal reduces the number of unique integrals by a factor of 4.

The Breit interaction involves current densities of the type $\psi_p \boldsymbol{\alpha} \psi_q$. When applied to a current density, the Kramers permutation operator gives

$$\begin{aligned} \hat{T}_{pq} \psi_p{}^{\dagger} \boldsymbol{\alpha} \psi_q &= (\psi_{\bar{p}}{}^{\dagger} \boldsymbol{\alpha} \psi_{\bar{q}})^{\dagger} = \psi_{\bar{q}}{}^{\dagger} \boldsymbol{\alpha} \psi_{\bar{p}} \\ &= \hat{\mathcal{K}}(\psi_{\bar{p}}{}^{\dagger} \boldsymbol{\alpha} \psi_{\bar{q}}) = -\hat{\mathcal{K}}(\hat{\mathcal{K}} \psi_p{}^{\dagger} \boldsymbol{\alpha} \psi_q) = -\psi_p{}^{\dagger} \boldsymbol{\alpha} \psi_q. \end{aligned} \tag{9.41}$$

Thus, for the Gaunt (or Breit) interaction

$$(p\boldsymbol{\alpha}q|r\boldsymbol{\alpha}s) = -(p\boldsymbol{\alpha}q|\bar{s}\boldsymbol{\alpha}\bar{r}) = -(\bar{q}\boldsymbol{\alpha}\bar{p}|r\boldsymbol{\alpha}s) = (\bar{q}\boldsymbol{\alpha}\bar{p}|\bar{s}\boldsymbol{\alpha}\bar{r}). \tag{9.42}$$

Application of the bar-reversal operator to this expression yields expressions for all the other integral types, and time reversal again reduces the number of unique integrals by a factor of 4.

If the Coulomb and Gaunt interactions are combined into a single interaction, there is of course a loss of one of these relations, and then the only relation between the integrals involves permutation of both pair indices. However, the factor of 2 lost by this procedure is regained by the use of a single interaction instead of two separate interactions.

With the use of these relations, the Kramers-restricted Dirac–Coulomb–Breit Hamiltonian may be written as

$$\begin{aligned}\hat{\mathcal{H}} = &\sum_{pq}\left[h_{pq}\hat{X}^{+}_{pq} + \frac{1}{2}\left(h_{\bar{p}q}\hat{X}^{+}_{\bar{p}q} + h_{p\bar{q}}\hat{X}^{+}_{p\bar{q}}\right)\right] \\ &+ \frac{1}{2}\sum_{pqrs}\Big[(pq|rs)\hat{x}^{++}_{pq,rs} - (p\boldsymbol{\alpha}q|r\boldsymbol{\alpha}s)\hat{x}^{--}_{pq,rs} + (\bar{p}q|rs)\hat{x}^{++}_{\bar{p}q,rs} \\ &\qquad - (\bar{p}\boldsymbol{\alpha}q|r\boldsymbol{\alpha}s)\hat{x}^{--}_{\bar{p}q,rs} + (p\bar{q}|rs)\hat{x}^{++}_{p\bar{q},rs} - (p\boldsymbol{\alpha}\bar{q}|r\boldsymbol{\alpha}s)\hat{x}^{--}_{p\bar{q},rs}\Big] \\ &+ \frac{1}{4}\sum_{pqrs}\left[(\bar{p}q|r\bar{s})\hat{x}^{++}_{\bar{p}q,r\bar{s}} - (\bar{p}\boldsymbol{\alpha}q|r\boldsymbol{\alpha}\bar{s})\hat{x}^{--}_{\bar{p}q,r\bar{s}}\right] \\ &+ \frac{1}{8}\sum_{pqrs}\left[(\bar{p}q|\bar{r}s)\hat{x}^{++}_{\bar{p}q,\bar{r}s} - (\bar{p}\boldsymbol{\alpha}q|\bar{r}\boldsymbol{\alpha}s)\hat{x}^{--}_{\bar{p}q,\bar{r}s} + (p\bar{q}|r\bar{s})\hat{x}^{++}_{p\bar{q},r\bar{s}} - (p\boldsymbol{\alpha}\bar{q}|r\boldsymbol{\alpha}\bar{s})\hat{x}^{--}_{p\bar{q},r\bar{s}}\right].\end{aligned} \tag{9.43}$$

This expression is more compact than the equivalent expression in terms of the one- and two-particle excitation operators $\hat{E}$ and $\hat{e}$.

9.3 Many-Particle States and Time Reversal

A many-particle (Fock space) single-determinant state in Kramers-unrestricted form can be written

$$|L\rangle = \prod_{k}^{N} a_k^{\dagger}|vac\rangle. \tag{9.44}$$

To represent the state in Kramers-restricted form we may partition the creation operators into "barred" and "unbarred" sets, and anticommute the barred operators to the right, so that, up to a sign,

$$|L\rangle = \prod_p a_p^\dagger \prod_q a_{\bar{q}}^\dagger |vac\rangle. \tag{9.45}$$

The two products of creation operators are in nonrelativistic theory termed α and β strings. To avoid confusion with these operators, but also to retain the analogy between Kramers pairs and spin–orbital pairs, we will term these A and B strings:

$$\begin{aligned} \hat{A}_I(N_A) &= \prod_p^{N_A} a_p^\dagger \\ \hat{B}_J(N_B) &= \prod_p^{N_B} a_{\bar{p}}^\dagger. \end{aligned} \tag{9.46}$$

The subscripts I and J on the strings are indices of the particular set of Kramers pairs from which the creation operators that make up the string are taken. The spinors that make up the Kramers pairs can, in this notation, be labeled A and B spinors. The many-particle state can now be written

$$|L\rangle = \hat{A}_I(N_A)\hat{B}_J(N_B)|vac\rangle. \tag{9.47}$$

The behavior of these strings under time reversal is given by

$$\hat{\mathcal{K}}\hat{A}_I(N_A) = \hat{B}_I(N_A); \qquad \hat{\mathcal{K}}\hat{B}_J(N_B) = (-1)^{N_B}\hat{A}_J(N_B). \tag{9.48}$$

For a closed-shell system, it is often convenient to reorder the creation operators so that the operators for each Kramers pair are together. Thus, we define a Kramers-pair creation operator

$$\hat{O}_p = {a_p}^\dagger {a_{\bar{p}}}^\dagger \tag{9.49}$$

and a closed-shell state is then given by

$$|L\rangle = \hat{A}_I(N_A)\hat{B}_I(N_A)|vac\rangle = (-1)^{(N_A-1)(N_A-2)/2}\prod_p^{N_A}\hat{O}_p|vac\rangle. \tag{9.50}$$

It is clear that $\hat{O}_p$ is symmetric under time reversal, for

$$\hat{\mathcal{K}}\hat{O}_p\hat{\mathcal{K}}^{-1} = \hat{\mathcal{K}}{a_p}^\dagger {a_{\bar{p}}}^\dagger\hat{\mathcal{K}}^{-1} = -{a_{\bar{p}}}^\dagger {a_p}^\dagger = {a_p}^\dagger {a_{\bar{p}}}^\dagger = \hat{O}_p. \tag{9.51}$$

Thus, all closed-shell states are symmetric under time reversal.

We now consider the effect of time reversal on a general many-particle state (determinant), which we write in terms of A and B strings:

$$|I, N_A; J, N_B\rangle = \hat{A}_I(N_A)\hat{B}_J(N_B)|vac\rangle. \tag{9.52}$$

Operating with $\hat{\mathcal{K}}$ on this determinant produces the result

$$\begin{aligned}\hat{\mathcal{K}}\ |I, N_A; J, N_B\rangle &= (-1)^{N_B}\hat{B}_I(N_A)\ \hat{A}_J(N_B)\ |vac\rangle \\ &= (-1)^{N_B(N_A+1)}\hat{A}_J(N_B)\ \hat{B}_I(N_A)\ |vac\rangle \\ &= (-1)^{N_B(N_A+1)}|J, N_B; I, N_A\rangle.\end{aligned} \tag{9.53}$$

In this notation the first two indices in the ket are always for the A string, and the second two are for the B string.

The sign in this expression is determined by whether N is even or odd. If N is even, then $N_A = N_B + 2k$ for integer k, and therefore N_A can be replaced by N_B in the phase factor. $N_B(N_B + 1)$ must be even and the phase is $+1$. If N is odd, then $N_A = N_B + 2k - 1$ for integer k, and we can replace $N_A + 1$ by N_B in the phase factor. Then we have N_B^2, which is even or odd, depending on whether N_B is even or odd, and it can be replaced by N_B. We can therefore write

$$\begin{aligned}\hat{\mathcal{K}}\ |I, N_A; J, N_B\rangle &= |J, N_B; I, N_A\rangle, && N \text{ even} \\ &= (-1)^{N_B}|J, N_B; I, N_A\rangle, && N \text{ odd.}\end{aligned} \tag{9.54}$$

Operating again on the determinant with $\hat{\mathcal{K}}$ produces the result

$$\begin{aligned}\hat{\mathcal{K}}^2|I, N_A; J, N_B\rangle &= \hat{\mathcal{K}}(-1)^{N_B}\hat{B}_I(N_A)\hat{A}_J(N_B)|vac\rangle \\ &= (-1)^{N_A+N_B}\hat{A}_I(N_A)\hat{B}_J(N_B)|vac\rangle \\ &= (-1)^{N}|I, N_A; J, N_B\rangle.\end{aligned} \tag{9.55}$$

We see that, for many-electron functions, a double time reversal produces a phase that depends on the number of electrons. If N is odd, there is a change of phase; if N is even, there is no change of phase. This is a manifestation of Kramers' theorem: a system with an odd number of fermions behaves like a fermion, but a system with an even number of fermions behaves like a boson.

We can summarize the effect of time reversal on many-electron states as follows:

$$\begin{aligned}\hat{\mathcal{K}}|M\rangle &= |\bar{M}\rangle \\ \hat{\mathcal{K}}|\bar{M}\rangle &= (-1)^{N}|M\rangle.\end{aligned} \tag{9.56}$$

We finally consider the effect of time reversal on matrix elements of the Hamiltonian over many-electron states. According to what we have shown above, we must have

$$\begin{aligned}\hat{\mathcal{K}}\langle L|\hat{\mathcal{H}}|M\rangle &= \langle \bar{L}|\hat{\mathcal{H}}|\bar{M}\rangle = \langle L|\hat{\mathcal{H}}|M\rangle^* \\ \hat{\mathcal{K}}\langle L|\hat{\mathcal{H}}|\bar{M}\rangle &= (-1)^N \langle \bar{L}|\hat{\mathcal{H}}|M\rangle = \langle L|\hat{\mathcal{H}}|\bar{M}\rangle^*\end{aligned} \tag{9.57}$$

where $|L\rangle$ and $|M\rangle$ are N-particle states. Thus the phase relations between the matrix elements of $\hat{H}$ are also determined by the parity of N.

These relations between the Hamiltonian matrix elements reduce the number of unique elements by a factor of 2. If the states are arranged so that the time-reversed partner of a state is the same distance from the end of the list as the state is from the beginning, time-reversal symmetry provides a symmetry about the diagonal running from upper right to lower left. The Hermitian character of the Hamiltonian corresponds to symmetry about the leading diagonal (upper left to lower right). Therefore, only a quarter of the Hamiltonian matrix is unique.

For the case of N even, it is possible to construct a real basis, that is, one in which the Hamiltonian matrix is real,

$$|M^s\rangle = \frac{i^{(1-s)/2}}{\sqrt{2(1+\delta_{M,\bar{M}})}}\big(|M\rangle + s|\bar{M}\rangle\big) \tag{9.58}$$

where the sign s can be ± 1. Both of these functions, $|M^+\rangle$ and $|M^-\rangle$, are symmetric under time reversal, by construction. We can write these states in terms of the A and B strings as

$$|M^s\rangle = \frac{i^{(1-s)/2}}{\sqrt{2(1+\delta_{M,\bar{M}})}}\big(\hat{A}_I(N_A)\hat{B}_J(N_B) + s(-1)^{N_B(N_A+1)}\hat{A}_J(N_B)\hat{B}_I(N_A)\big). \tag{9.59}$$

Transforming the Hamiltonian matrix elements into this basis and making use of the behavior of the functions under time reversal, we have

$$\begin{aligned}H_{L^+M^+} &= \mathrm{Re}(H_{LM} + H_{L\bar{M}})/G_{LM} \\ H_{L^+M^-} &= -\mathrm{Im}(H_{LM} - H_{L\bar{M}})/G_{LM} \\ H_{L^-M^+} &= \mathrm{Im}(H_{LM} + H_{L\bar{M}})/G_{LM} \\ H_{L^-M^-} &= \mathrm{Re}(H_{LM} - H_{L\bar{M}})/G_{LM}\end{aligned} \tag{9.60}$$

where $G_{LM} = [(1+\delta_{L,\bar{L}})(1+\delta_{M,\bar{M}})]^{1/2}$ is a normalization factor required when one or both of the determinants is a closed shell. We will return to these expressions later when we consider the effect of symmetry on the structure of the many-electron Dirac Hamiltonian matrix.

10

Matrices and Wave Functions under Double-Group Symmetry

Symmetry is one of the most versatile theoretical tools of physics and chemistry. It provides qualitative insight into the wave functions and properties of systems, and it has also been used successfully to obtain great savings in computational efforts. In the preceding chapter we examined time-reversal symmetry, and now we turn to the more familiar point-group symmetry. We show how relativity requires special consideration and extensions of the concepts developed for the nonrelativistic case, and how time-reversal symmetry and double-group symmetry are connected.

Although the techniques that incorporate double-group symmetry presented here are primarily aimed at four-component calculations, they are equally applicable to two-component calculations in which the spin-dependent operators are included at the SCF stage of a calculation.

10.1 Time-Reversal and Point-Group Symmetry

In the preceding chapter, we have shown how the use of time-reversal symmetry can lead to considerable reduction in the number of unique matrix elements that appear in the operator expressions. However, we are also interested in the overall structure of the matrices of the operators. In particular, we are interested in possible block structures, where classes of matrix elements may be set to zero a priori. If the matrices can be cast in block diagonal form, we may save on storage as well as computational effort in solving eigenvalue problems, for example.

Matrix blocking will already be effected by the point-group symmetry of the molecule. We do not expect time reversal to reduce the symmetry because the time-reversal operator $\hat{\mathcal{K}}$ commutes with all operations of the point group. If time reversal introduces new symmetries, we will gain further blocking. We must therefore establish the relation between time-reversal symmetry and point-group symmetry.

One possibility would be to extend the point group $G = \{\hat{g}_i; i = 1, h\}$ in the same manner as already done for the double groups in chapter 6. There we constructed a new set of operations $\bar{E}\hat{g}_i$ and obtained a new group $G^* = \{\hat{g}_i, \bar{E}\hat{g}_i; i = 1, h\}$. The same recipe for time reversal would give us a group $G_t^* = \{\hat{g}_i, \hat{\mathcal{K}}\hat{g}_i; i = 1, h\}$ of order $2h$. This is unfortunately only of limited use, due to the antiunitary character of $\hat{\mathcal{K}}$. Wigner (1959) has shown that with antiunitary elements in the group it is not possible to find a homomorphism of the group onto a set of linear transformations.

We can illustrate this problem by considering the associative conditions that the group has to fulfil (Tinkham 1964). The action of $\hat{\mathcal{K}}$ on a product of two group elements yields

$$(\hat{\mathcal{K}}(\hat{g}_i\hat{g}_j)) = (\hat{\mathcal{K}}\hat{g}_i)(\hat{\mathcal{K}}\hat{g}_j). \tag{10.1}$$

If the time-reversal operator is considered an element of the group, then the associative law must hold, such that

$$\hat{\mathcal{K}}(\hat{g}_i\hat{g}_j) = (\hat{\mathcal{K}}\hat{g}_i)\hat{g}_j \tag{10.2}$$

which implies that $(\hat{\mathcal{K}}\hat{g}_j) = \hat{g}_j$. If this is true, the group elements must be real, something they are not in the general case. Not only that, the time-reversal operator must be the identity operator E, which is also not the case. We can, of course, still construct transformation matrices corresponding to the symmetry operations of the group, using a given basis, and these matrices may even be decomposed to irreducible members. Wigner (1959) has introduced the term *co-representation* for such sets of matrices.

We must therefore go back to the Kramers pairs and see how they behave under the symmetry operations of the group G. Consider the Kramers pair $(\psi_p^\gamma, \psi_{\bar{p}}^{\gamma'})$, transforming under the irreps γ and γ', respectively. The function ψ_p^γ transforms under the group operation $\hat{g}$ according to

$$\hat{g}\psi_p^\gamma = \sum_q \psi_q^\gamma D_{pq}^\gamma(\hat{g}) \tag{10.3}$$

The transformation of $\psi_{\bar{p}}^{\gamma'}$ under the operation $\hat{g}$ is

$$\begin{aligned}\hat{g}\bar{\psi}_{\bar{p}}^{\gamma'} &= \hat{g}\hat{\mathcal{K}}\psi_p^\gamma = \hat{\mathcal{K}}\hat{g}\psi_p^\gamma = \hat{\mathcal{K}}\sum_q \psi_q^\gamma D_{qp}^\gamma(\hat{g}) = \sum_q (\hat{\mathcal{K}}\psi_q^\gamma) D_{qp}^\gamma(\hat{g})^* \\ &= \sum_q \psi_{\bar{q}}^{\gamma'} D_{qp}^\gamma(\hat{g})^* = \sum_q \psi_{\bar{q}}^{\gamma'} D_{qp}^{\gamma'}(\hat{g})\end{aligned} \tag{10.4}$$

We have made use of the fact that $\hat{\mathcal{K}}$ commutes with all elements of the group.[1] Thus, the representation matrices of the time-reversed spinor $\psi_{\bar{p}}^{\gamma'}$ are the complex conjugates of those for ψ_p^γ,

$$\mathbf{D}^{\gamma'}(\hat{g}) = \mathbf{D}^\gamma(\hat{g})^*. \tag{10.5}$$

---

[1] This fact may be established by using angular momentum eigenfunctions for the Kramers pair basis and applying $\hat{\mathcal{K}}\hat{g}$ and $\hat{g}\hat{\mathcal{K}}$ to them using a rotation and an inversion for $\hat{g}$ and using the explicit expressions for the rotation matrices and the angular momentum functions.

What consequences does this have for the relation of the Kramers pairs to the representations of the (double) point group? Let us initially restrict our discussion to groups with irreps of dimension 1 or 2, which covers all groups except the cubic and icosahedral groups. There are three cases to consider:

1. The irrep γ is one-dimensional, and all the (1×1) representation matrices $\mathbf{D}^{\gamma}(\hat{g})$ are real. Then ψ_p^{γ} and $\psi_{\bar{p}}^{\gamma'}$ obviously belong to the same irrep and $\gamma = \gamma'$.
2. The representation matrices $\mathbf{D}^{\gamma}(\hat{g})$ are complex, and the operations of the point group do not mix ψ_p^{γ} and $\psi_{\bar{p}}^{\gamma'}$. In this case ψ_p^{γ} and $\psi_{\bar{p}}^{\gamma'}$ belong to two different one-dimensional irreps and $\gamma \neq \gamma'$.
3. The representation matrices $\mathbf{D}^{\gamma}(\hat{g})$ are not all real, and the operations of the point group mix ψ_p^{γ} and $\psi_{\bar{p}}^{\gamma'}$. In this case ψ_p^{γ} and $\psi_{\bar{p}}^{\gamma'}$ are a basis for a two-dimensional irrep and $\gamma = \gamma'$.

Let us consider a few simple examples to see how this works. We only show the half of the character tables that covers the "single group" operations. The other half follows by symmetry.

EXAMPLE 1

C_1^* character table

Symbol	E
A	1
$B_{1/2}$	1

Here, the fermion irrep is $B_{1/2}$, the representation matrices are all real, and we have case 1.

EXAMPLE 2

C_2^* character table

Symbol	E	C_2
A	1	1
B	1	-1
$E_{1/2}$	1	i
$E_{-1/2}$	1	$-i$

The fermion irreps[2] are $E_{1/2}$ and $E_{-1/2}$, some of the representation matrices are complex, and we have case 2. Note that although the irreps are one-dimensional (in accordance with Lagrange's theorem), they are labeled E in recognition of the equivalence between them: together they form an $E_{1/2}$ rep.

---

[2] See appendix D for an explanation of the notation for the irrep labels, which differs a little from the conventional notation.

EXAMPLE 3

C^*_{2v} character table

Symbol	E	C_2	σ_x	σ_y
A_1	1	1	1	1
A_2	1	1	−1	−1
B_1	1	−1	1	−1
B_2	1	−1	−1	1
$E_{1/2}$	2	0	0	0

The fermion functions span the two-dimensional $E_{1/2}$ irrep, and we have case 3.

EXAMPLE 4

C^*_{3v} character table

Symbol	E	$2C_3$	$3\sigma_v$
A_1	1	1	1
A_2	1	1	−1
E	2	−1	0
$E_{1/2}$	2	1	0
$E_{3/2}$	1	−1	i
$E_{-3/2}$	1	1	$-i$

The fermion functions are either in the two-dimensional $E_{1/2}$ irrep, or they are in the two one-dimensional irreps $E_{3/2}$ and $E_{-3/2}$. Thus we have both case 2 and case 3 in this group.

For case 1 and case 2, time reversal introduces extra degeneracies in the group that are not formally evident from the character table. In case 1, fermion functions are pairwise degenerate; in case, 2 time-reversal makes two formally inequivalent one-dimensional irreps degenerate. This phenomenon is not confined to the relativistic case. In the group C_3, for example, the boson irreps consist of an A irrep and two E irreps that are related by complex conjugation, and are degenerate, forming an E rep. We can regard this as the application of time reversal in a nonrelativistic system: the time-reversal operator for scalar functions is simply the complex conjugation operator.

For the groups of higher symmetry, initially excluded from our considerations, we encounter irreps that display higher degeneracies. These fall into two categories:

1. Genuine $2m$-fold degenerate irreps, where the members of Kramers pairs can be assigned to different rows. An example is T_d, which has two real irreps of degeneracy 2 and one of degeneracy 4.
2. m-fold degenerate irreps that occur in conjugate pairs, so that the pair together forms a $2m$-fold degenerate rep. The members of a Kramers pair belong to the different irreps. An example is T, which has 6 fermion irreps: 2 doubly degenerate irreps that together form a rep, and 4 singly degenerate irreps that form 2 reps.

There is a simple test, the *Frobenius–Schur test*, that can be applied to determine which case we have for a given irrep γ. We calculate the quantity τ defined by

$$\tau = \frac{1}{h}\sum_{i=1}^{h} \chi^{\gamma}(\hat{g}_i^2), \tag{10.6}$$

where $\chi^{\gamma}(\hat{g}_i^2)$ is the character of the operation $\hat{g}_i\hat{g}_i$. The test gives the following results:

$$\begin{aligned} \tau = 1 \quad &\Rightarrow \quad \text{case } 1, \\ \tau = 0 \quad &\Rightarrow \quad \text{case } 2, \\ \tau = -1 \quad &\Rightarrow \quad \text{case } 3. \end{aligned} \tag{10.7}$$

We demonstrate this test for C_2^*, remembering that $C_2C_2 = \bar{E}$. Then for $E_{1/2}$

$$\tau = \frac{1}{2}\left[\chi(E^2) + \chi(C_2^2)\right] = \frac{1}{2}\left[\chi(E) + \chi(\bar{E})\right] = 0. \tag{10.8}$$

Thus this is indeed a case 2 irrep, as already found above. We have only used the single group operations for this test: the result would have been the same if we had included all operations of the double group and doubled the order of the group.

In order to classify the irreps of the higher groups—and also to gain insight into the properties of the lower groups—it is useful to make use of a *group chain*, as follows.

1. The higher group is decomposed into a product of two groups, one of which is of order 2.
2. A correspondence is set up between the fermion irreps in the higher group and the lower group.
3. This correspondence is used to define the rows of the irreps in the higher group.

It is particularly useful to make this definition so that real irreps reduce to complex irreps, where the rows and the multiplication table are well-defined.

A simple example is the decomposition $C_{2v} = C_2 \otimes C_s$. The reduction $C_{2v} \to C_2$ gives the correspondence $A_1, A_2 \to A$, $B_1, B_2 \to B$ for the boson irreps, and $(E_{1/2}, E_{-1/2}) \to (E_{1/2}, E_{-1/2})$ for the fermion irreps. The notation used for the rows of the $E_{1/2}$ irrep make the correspondence perfectly clear. A more extensive example is the reduction $T_d \to T$, which follows the same pattern.

10.2 Time-Reversal Symmetry and Matrix Block Structure

We now return to the question of how time-reversal symmetry relates to double-group symmetry and the block structure of operator matrices. For case 2 above, with the two components of the Kramers pair belonging to different irreps, there are no matrix elements of totally symmetric operators between a fermion function and its time-reversed partner. For case 3, we can always carry out transformations between the Kramers pairs such that they correspond to different rows of the irreducible matrix. Again, the great orthogonality theorem assures us that the matrix is blocked.

It remains to discuss case 1, where both functions in a Kramers pair belong to the same irrep. For this case, ordinary point group symmetry does not promise any a priori blocking. We will therefore look for a unitary transformation on the Kramers pair basis that can block-diagonalize a matrix constructed over functions belonging to a case 1 irrep. The problem can be reduced to the diagonalization of a general 2 × 2 matrix spanned by the components of a Kramers pair. For the operator $\hat{\Omega}$ this is

$$\boldsymbol{\Omega} = \begin{pmatrix} \Omega_{pq} & \Omega_{p\bar{q}} \\ \Omega_{\bar{p}q} & \Omega_{\bar{p}\bar{q}} \end{pmatrix} \tag{10.9}$$

We have previously shown that in a Kramers basis we have the following relations:

$$\Omega_{\bar{p}\bar{q}} = \mathrm{t}\Omega^*_{pq}; \qquad \Omega_{\bar{p}q} = -\mathrm{t}\Omega^*_{p\bar{q}}. \tag{10.10}$$

Using these relations, we see that the matrix takes the form

$$\boldsymbol{\Omega} = \begin{pmatrix} a & b \\ -\mathrm{t}b^* & \mathrm{t}a^* \end{pmatrix}. \tag{10.11}$$

For simplicity we will assume that t = 1, that is, $\hat{\Omega}$ is symmetric under time reversal. We will comment briefly on the antisymmetric t = −1 case later.

The general unitary 2 × 2 matrix has the form

$$\mathbf{u} = \begin{pmatrix} p\cos\vartheta & q\sin\vartheta \\ -q^*\sin\vartheta & p^*\cos\vartheta \end{pmatrix} \tag{10.12}$$

where p and q are phase factors such that $pp^* = qq^* = 1$. Without loss of generality, we can assume that $p = 1$, and that

$$\mathbf{u} = \begin{pmatrix} \cos\vartheta & q\sin\vartheta \\ -q^*\sin\vartheta & \cos\vartheta \end{pmatrix}. \tag{10.13}$$

Applying this transformation to $\boldsymbol{\Omega}$, we get

$$\mathbf{u}^\dagger\boldsymbol{\Omega}\mathbf{u} = \boldsymbol{\Omega}' = \begin{pmatrix} \Omega'_{11} & \Omega'_{12} \\ \Omega'_{21} & \Omega'_{22} \end{pmatrix}. \tag{10.14}$$

By insertion we get the following expressions for the elements of $\boldsymbol{\Omega}'$:

$$\begin{aligned}
\Omega'_{11} &= a\cos^2\vartheta + qa^*q^*\sin^2\vartheta + (qb^* - bq^*)\sin\vartheta\cos\vartheta \\
\Omega'_{12} &= b\cos^2\vartheta + qb^*q\sin^2\vartheta + (aq - qa^*)\sin\vartheta\cos\vartheta \\
\Omega'_{21} &= -b^*\cos^2\vartheta - q^*b^*q^*\sin^2\vartheta + (q^*a - a^*q)\sin\vartheta\cos\vartheta \\
\Omega'_{22} &= a^*\cos^2\vartheta + q^*aq\sin^2\vartheta + (q^*b - b^*q)\sin\vartheta\cos\vartheta.
\end{aligned} \tag{10.15}$$

For $\boldsymbol{\Omega}'$ to be diagonal, the elements Ω'_{12} and Ω'_{21} must vanish. A sufficient condition for this is that the terms in a and b vanish independently of each other. Thus, from Ω'_{12} we get

$$aq - qa^* = \mathrm{Re}(a)q + i\mathrm{Im}(a)q - q\mathrm{Re}(a) + qi\mathrm{Im}(a) = 0. \tag{10.16}$$

$\mathrm{Re}(a)$ and $\mathrm{Im}(a)$ are real numbers and commute with all other quantities. We cannot expect a to be real in general, and thus we are left with the equation

$$iq + qi = 0, \tag{10.17}$$

that is, q anticommutes with i. Let us accept this somewhat unexpected result for the moment and see where it takes us. For the b-dependent terms to vanish, we must have

$$qb^*q\sin^2\vartheta + b\cos^2\vartheta = 0. \tag{10.18}$$

Accepting the premise of anticommutation between q and i, this yields

$$q^2 b\sin^2\vartheta + b\cos^2\vartheta = 0. \tag{10.19}$$

Our starting assumption was that the matrix $\boldsymbol{\Omega}$ was not in general diagonal, and thus b will not in general be 0. Then

$$q^2\sin^2\vartheta + \cos^2\vartheta = 0. \tag{10.20}$$

Here $\sin^2\vartheta$ and $\cos^2\vartheta$ are real positive numbers, and the equation can only be satisfied if q^2 is a negative real number. As q is assumed to have unit modulus, this means that $q^2 = -1$. For this value of q, the equation above is satisfied if

$$\sin\vartheta = \cos\vartheta = \frac{1}{\sqrt{2}}. \tag{10.21}$$

The diagonal elements of $\boldsymbol{\Omega}$ are now

$$\begin{aligned}
\Omega'_{11} &= \frac{1}{2}(a + qa^*q^*) + \frac{1}{2}(qb^* - bq^*) = a + bq \\
\Omega'_{22} &= \frac{1}{2}(a^* + q^*aq) + \frac{1}{2}(q^*b - b^*q) = a^* - b^*q
\end{aligned} \tag{10.22}$$

and the $\boldsymbol{\Omega}'$ matrix takes the form

$$\boldsymbol{\Omega}' = \begin{pmatrix} a+bq & 0 \\ 0 & a^* - b^*q \end{pmatrix}. \tag{10.23}$$

The question remains, what is q? We have the following circumstantial evidence in the case:

$$q^*q = 1; \qquad q^2 = -1; \qquad qi = -iq. \tag{10.24}$$

We see that the imaginary unit i itself satisfies the first two criteria, but obviously not the last. The solution is that q is one of the *quaternion* units. Quaternions were developed by Hamilton as a further extension of the complex numbers in a progression:

real numbers → complex numbers → quaternions.

A general quaternion number is expressed in terms of the units $(1, i, j, k)$, where 1 and i are the usual units for complex numbers. Thus the quaternion number Q is given as

$$Q = r + si + tj + uk \tag{10.25}$$

with r, s, t, u real numbers. These quaternion units fulfill the relations

$$\begin{aligned} &i^2 = j^2 = k^2 = -1 \\ &ij = k; \quad jk = i; \quad ki = j \\ &i^* = -i; \quad j^* = -j; \quad k^* = -k. \end{aligned} \tag{10.26}$$

As we can see from these expressions, only two of the three units are truly independent: the last one may be expressed in terms of the two others. Conventionally, we choose j as the independent unit, which allows us to write Q as a combination of two complex numbers:

$$Q = v + wj; \qquad v = r + is, w = t + iu. \tag{10.27}$$

Quaternions as well as complex numbers are members of a more general class of structures called *Clifford algebras*. The Clifford algebra $C(p, q)$ is generated by starting with the real numbers and then adding p quantities that squared yield -1 and q quantities that squared give 1. The complex numbers then correspond to C(1,0), quaternions to C(2,0), and the series may be extended to octonions and so on.

An important realization of quaternions is to be found in the Pauli matrices. The set of matrices $\{\mathbf{I}_2, i\sigma_z, i\sigma_y, i\sigma_x\}$ is isomorphic to the set of quaternion units $\{1, i, j, k\}$. This isomorphism has been exploited in computational schemes for the construction of symmetry spinors (Saue and Jensen 1999).

Returning to the diagonalization problem at hand, we see that the transformation matrix $\mathbf{u}$ that diagonalizes the 2×2 matrix $\boldsymbol{\Omega}$ is

$$\mathbf{u} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & j \\ -j^* & 1 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix}. \tag{10.28}$$

Because we need a quaternion transformation to diagonalize a matrix over a basis that follows case 1 above, and the resulting matrix is a quaternion matrix, the corresponding irrep can be called a *quaternion irrep*. In the same vein, a case 2 irrep is called a *complex irrep*, while case 3 yields a *real irrep*.

The case of an operator that is antisymmetric under time reversal, that is, $\mathrm{t} = -1$, may now be discussed on the basis of the equations above. All we need to do is replace a^* by $-a^*$ and b^* by $-b^*$. For vanishing off-diagonal elements we then get the condition

$$aq + qa^* = 0 \qquad \Rightarrow \qquad 2q\,\mathrm{Re}(a) = 0; \quad (iq - qi)\,\mathrm{Im}(a) = 0. \tag{10.29}$$

This requires that $\mathrm{Re}(a) = 0$ and that $[i, q] = 0$. The last condition means that q can at most be imaginary, while the first one means that a must be pure imaginary. Further investigation reveals a similar condition on b.

We now turn our attention to the entire space of n_γ Kramers pairs that transform under the irrep γ. We see that operators are represented as $2n_\gamma \times 2n_\gamma$ matrices. By a suitable reordering, the representation matrix for an operator symmetric under time reversal may be brought to the form

$$\boldsymbol{\Omega} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^* & \mathbf{A}^* \end{pmatrix} \tag{10.30}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $n_\gamma \times n_\gamma$ matrices. This matrix may clearly be block-diagonalized by the matrix

$$\mathbf{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{I}_{n_\gamma} & j\mathbf{I}_{n_\gamma} \\ j\mathbf{I}_{n_\gamma} & \mathbf{I}_{n_\gamma} \end{pmatrix} \tag{10.31}$$

giving the result

$$\mathbf{U}^\dagger \boldsymbol{\Omega} \mathbf{U} = \begin{pmatrix} \mathbf{A} + \mathbf{B}j & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^* - \mathbf{B}^* j \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & -k\mathbf{G}k \end{pmatrix}. \tag{10.32}$$

$\mathbf{G}$ is a quaternion Hermitian matrix, and the eigenvalues of $\boldsymbol{\Omega}$ are clearly doubly degenerate.

It may be thought that quaternions could be used in the creation and annihilation operators to define a new basis in which all matrices were block-diagonal in the Kramers pairs. However, because of the noncommutative algebra, the step in which the creation operator is permuted over the matrix element to separate the two does not produce the desired result. Therefore, quaternions are useful only at the matrix algebra stage, and not in the formalism.

10.3 Symmetry of Spinor Components

Up to now in this chapter we have discussed the symmetry properties of spinors as a whole. The spinor itself is a four-component entity in which the rows correspond to spin orientations, the first and third to α spin, and the second and fourth to β spin:

$$\psi = \begin{pmatrix} \phi^{L\alpha} \\ \phi^{L\beta} \\ \phi^{S\alpha} \\ \phi^{S\beta} \end{pmatrix}. \tag{10.33}$$

Partitioning the spinor into large and small components, we can write the 2-spinors in the form

$$\psi^X = \phi^{X\alpha}\alpha + \phi^{X\beta}\beta \tag{10.34}$$

where the spin functions α and β are two-component vectors. These spin functions must transform according to the fermion irreps, and the spatial functions according to the boson irreps. Therefore, we ought to be able to classify the real and imaginary, α and β components of a spinor in terms of the boson irreps and determine the symmetry relationship between the large and small components.

We consider the simplest case where we have a clear division of the spin functions, and that is the group C_2, for which the multiplication table is given, and take α in $E_{1/2}$ and β in $E_{-1/2}$. As $A \otimes E_{1/2} = E_{1/2}$ and $B \otimes E_{-1/2} = E_{1/2}$, a 2-spinor belonging to $E_{1/2}$ must have functions of A symmetry for α spin and B symmetry for β spin. Similarly, as $A \otimes E_{-1/2} = E_{-1/2}$ and $B \otimes E_{1/2} = E_{-1/2}$, a 2-spinor belonging to $E_{-1/2}$ must have functions of B symmetry for α spin and A symmetry for β spin.

Multiplication table for C_2

Symbol	A	B	$E_{1/2}$	$E_{-1/2}$
A	A	B	$E_{1/2}$	$E_{-1/2}$
B		A	$E_{-1/2}$	$E_{1/2}$
$E_{1/2}$			B	A
$E_{-1/2}$				B

For a group such as C_{2v}, which has a doubly degenerate fermion irrep, the group chain to C_2 gives some information, but it does not resolve the division between A_1 and A_2, and B_1 and B_2. We therefore need a more general approach.

We start by splitting each of the four spatial components of the spinor into a real and an imaginary part:

$$\phi^{X\tau} = \mathrm{Re}(\phi^{X\tau}) + i\,\mathrm{Im}(\phi^{X\tau}) \tag{10.35}$$

where X is L or S, and τ is α or β. Each of these parts, $\mathrm{Re}(\phi^{X\tau})$ and $\mathrm{Im}(\phi^{X\tau})$, transforms according to a boson irrep. We wish to derive relations between the various symmetries of the eight such boson contributions to each 4-spinor. For this purpose it is most convenient to look at the large and small components separately. These components form 2-spinors of the form

$$\psi^X = \begin{pmatrix} \mathrm{Re}(\phi^{X\alpha}) + i\mathrm{Im}(\phi^{X\alpha}) \\ \mathrm{Re}(\phi^{X\beta}) + i\mathrm{Im}(\phi^{X\beta}) \end{pmatrix}. \tag{10.36}$$

To elucidate the symmetry properties of the various parts here, we can start from the expressions for elimination of the small component, (4.82) and (4.83),

$$\begin{aligned} c^2(\boldsymbol{\sigma}\cdot\mathbf{p})B^{-1}(\boldsymbol{\sigma}\cdot\mathbf{p})\psi^L &= (E-V)\psi^L \\ B^{-1}(\boldsymbol{\sigma}\cdot\mathbf{p})\psi^L &= \psi^S \end{aligned} \tag{10.37}$$

with

$$B = 2c^2 + E - V. \tag{10.38}$$

In the absence of other external fields, V is totally symmetric under the operations of the molecular symmetry group because it defines the symmetry group. $2c^2$ and E are numbers, which by definition are totally symmetric. B and its inverse are therefore totally symmetric. Moreover, the equation involving only the large component, ψ^L, above tells us that the operator $(\boldsymbol{\sigma}\cdot\mathbf{p})B^{-1}(\boldsymbol{\sigma}\cdot\mathbf{p})$ must also be totally symmetric, because after it has operated on ψ^L we get back something of the same symmetry. Thus we can apply this operator to ψ^L and should regain a 2-spinor where the parts have the same symmetries as ψ^L.

We expand $(\boldsymbol{\sigma}\cdot\mathbf{p})B^{-1}(\boldsymbol{\sigma}\cdot\mathbf{p})$ as

$$\begin{aligned} (\boldsymbol{\sigma}\cdot\mathbf{p})(2c^2+E-V)^{-1}(\boldsymbol{\sigma}\cdot\mathbf{p}) &= \frac{c^2}{2c^2+E-V}\mathbf{p}^2 - \frac{\hbar^2c^2}{(2c^2+E-V)^2} \\ &\quad \times \left[(\nabla V)\cdot\nabla + i\boldsymbol{\sigma}\cdot(\nabla V)\times\nabla\right]. \end{aligned} \tag{10.39}$$

Since V is totally symmetric, any function of V is totally symmetric. $\mathbf{p}^2$ is also totally symmetric. ∇V behaves like a position variable and therefore has the same symmetry as $\mathbf{r}$. ∇ also behaves like a position variable, so the scalar product $(\nabla V)\cdot\nabla$ behaves as r^2, which is totally symmetric. These operators are all multiplied by the unit matrix and so contribute to the real part of the diagonal of the operator matrix. The remaining operator is $(\nabla V)\times\nabla$, which behaves like $\mathbf{r}\times\nabla$. This is essentially an angular momentum operator, and so its symmetry is the same as the vector of rotations $\mathbf{R} = (\hat{R}_x, \hat{R}_y, \hat{R}_z)$.

At this stage we need a notation that enables us to describe symmetry in a convenient and transparent manner. We therefore introduce the symbol $\Gamma(\xi)$ as meaning "something that transforms as the quantity ξ under the operations of the group." This may not appear to be very precise, but if handled with care it turns out to fulfil our needs.

As an example of how this notation works, (10.37) gives rise to the relation

$$\begin{gathered}\Gamma\left((\boldsymbol{\sigma}\cdot\mathbf{p})B^{-1}(\boldsymbol{\sigma}\cdot\mathbf{p})\right)\Gamma\left(\psi^{L}\right)=\Gamma\left(\psi^{L}\right)\\ \Gamma\left(B^{-1}(\boldsymbol{\sigma}\cdot\mathbf{p})\right)\Gamma\left(\psi^{L}\right)=\Gamma\left(\psi^{S}\right).\end{gathered}\tag{10.40}$$

Γ_0 will be used to denote the totally symmetric representation. Based on our discussion above, we have for the symmetries involved

$$\Gamma((\boldsymbol{\sigma}\cdot\mathbf{p})B^{-1}(\boldsymbol{\sigma}\cdot\mathbf{p}))=\begin{pmatrix}\Gamma_0+i\Gamma(\hat{R}_z) & \Gamma(\hat{R}_y)+i\Gamma(\hat{R}_x)\\ \Gamma(\hat{R}_y)+i\Gamma(\hat{R}_x) & \Gamma_0+i\Gamma(\hat{R}_z)\end{pmatrix}.\tag{10.41}$$

As we are only concerned with symmetry, not sign, we will use plus signs in the complex quantities: $\Gamma_0+i\Gamma(\hat{R}_z)$ means "something real that transforms like the totally symmetric irrep plus (or minus) something imaginary that transforms like a rotation about the z axis."

This operator can now be applied to ψ^L, where the symmetries of the separate boson parts are

$$\begin{pmatrix}\Gamma(\mathrm{Re}_{L,\alpha})+i\Gamma(\mathrm{Im}_{L,\alpha})\\ \Gamma(\mathrm{Re}_{L,\beta})+i\Gamma(\mathrm{Im}_{L,\beta})\end{pmatrix}\tag{10.42}$$

in obvious notation. From the equation for the large component we know that in terms of symmetry we must have

$$\Gamma((\boldsymbol{\sigma}\cdot\mathbf{p})B^{-1}(\boldsymbol{\sigma}\cdot\mathbf{p}))\Gamma(\psi^L)=\Gamma(\psi^L).\tag{10.43}$$

When expressed at the level of the various parts of these quantities, the result is

$$\begin{aligned}&\begin{pmatrix}\Gamma_0+i\Gamma(\hat{R}_z) & \Gamma(\hat{R}_y)+i\Gamma(\hat{R}_x)\\ \Gamma(\hat{R}_y)+i\Gamma(\hat{R}_x) & \Gamma_0+i\Gamma(\hat{R}_z)\end{pmatrix}\begin{pmatrix}\Gamma(\mathrm{Re}_{L,\alpha})+i\Gamma(\mathrm{Im}_{L,\alpha})\\ \Gamma(\mathrm{Re}_{L,\beta})+i\Gamma(\mathrm{Im}_{L,\beta})\end{pmatrix}\\ &\quad=\begin{pmatrix}\Gamma(\mathrm{Re}_{L,\alpha})+i\Gamma(\mathrm{Im}_{L,\alpha})\\ \Gamma(\mathrm{Re}_{L,\beta})+i\Gamma(\mathrm{Im}_{L,\beta})\end{pmatrix}.\end{aligned}\tag{10.44}$$

To determine the symmetries of the scalar parts, it is sufficient to consider only the first one, the real part of the alpha component on the right side of the equation, for which we have the symmetry $\Gamma(\mathrm{Re}_{L,\alpha})$. The matrix multiplication yields

$$\Gamma_0\Gamma(\mathrm{Re}_{L,\alpha})+\Gamma(\hat{R}_z)\Gamma(\mathrm{Im}_{L,\alpha})+\Gamma(\hat{R}_y)\Gamma(\mathrm{Re}_{L,\beta})+\Gamma(\hat{R}_x)\Gamma(\mathrm{Im}_{L,\beta})=\Gamma(\mathrm{Re}_{L,\alpha})\tag{10.45}$$

and thus

$$\begin{aligned}\Gamma(\mathrm{Im}_{L,\alpha}) &= \Gamma(\hat{R}_z)\Gamma(\mathrm{Re}_{L,\alpha}),\\ \Gamma(\mathrm{Re}_{L,\beta}) &= \Gamma(\hat{R}_y)\Gamma(\mathrm{Re}_{L,\alpha}),\\ \Gamma(\mathrm{Im}_{L,\beta}) &= \Gamma(\hat{R}_x)\Gamma(\mathrm{Re}_{L,\alpha}).\end{aligned} \tag{10.46}$$

The symmetry of the 2-spinor is then

$$\Gamma(\psi^L) = \begin{pmatrix} \Gamma_0 + i\Gamma(\hat{R}_z) \\ \Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x) \end{pmatrix} \Gamma(\mathrm{Re}_{L,\alpha}). \tag{10.47}$$

Two results follow immediately from this derivation. First, the four parts of the ψ^S 2-spinor must fulfil the same relation, because we could equally well have done an elimination of the large component, analogously to what we did for the large component at the outset. The same argument could then have been carried through for ψ^S. Second, the symmetry of the time-reversed partner of ψ^L follows easily because application of the time-reversal operator just corresponds to a spin flip, effectively swapping the two complex spatial components of the 2-spinor:

$$\hat{\mathcal{K}} \begin{pmatrix} \Gamma_0 + i\Gamma(\hat{R}_z) \\ \Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x) \end{pmatrix} = \begin{pmatrix} \Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x) \\ \Gamma_0 + i\Gamma(\hat{R}_z) \end{pmatrix}. \tag{10.48}$$

Remember that we are not taking the sign into account here: we are purely concerned with the symmetry of the real and imaginary parts of the spinor components.

It remains to determine the symmetry relation between the large and the small component. To do this, we can use the remaining equation from the elimination of the small component, which provides us with a connection between ψ^L and ψ^S. The inverse operator is totally symmetric, so the relations are determined by the operator $c(\boldsymbol{\sigma} \cdot \mathbf{p})$, which in terms of symmetry can be represented as

$$\Gamma(\boldsymbol{\sigma} \cdot \mathbf{p}) = \begin{pmatrix} 0 + i\Gamma(z) & \Gamma(y) + i\Gamma(x) \\ \Gamma(y) + i\Gamma(x) & 0 + i\Gamma(z) \end{pmatrix}, \tag{10.49}$$

where we have to remember that replacing $\mathbf{p}$ by its quantum mechanical operator equivalent introduces a factor i. In symmetry terms we now have

$$\begin{aligned}&\begin{pmatrix} 0 + i\Gamma(z) & \Gamma(y) + i\Gamma(x) \\ \Gamma(y) + i\Gamma(x) & 0 + i\Gamma(z) \end{pmatrix} \begin{pmatrix} \Gamma(\mathrm{Re}_{L,\alpha}) + i\Gamma(\mathrm{Im}_{L,\alpha}) \\ \Gamma(\mathrm{Re}_{L,\beta}) + i\Gamma(\mathrm{Im}_{L,\beta}) \end{pmatrix}\\ &\quad = \begin{pmatrix} \Gamma(\mathrm{Re}_{S,\alpha}) + i\Gamma(\mathrm{Im}_{S,\alpha}) \\ \Gamma(\mathrm{Re}_{S,\beta}) + i\Gamma(\mathrm{Im}_{S,\beta}) \end{pmatrix}.\end{aligned} \tag{10.50}$$

Again we need only consider one function, for which we may write

$$\Gamma(z)\Gamma(\mathrm{Im}_{L,\alpha}) + \Gamma(y)\Gamma(\mathrm{Re}_{L,\beta}) + \Gamma(x)\Gamma(\mathrm{Im}_{L,\beta}) = \Gamma(\mathrm{Re}_{S,\alpha}). \tag{10.51}$$

Substituting for the large-component symmetries in terms of the $\mathrm{Re}_{L,\alpha}$ component, we find

$$\begin{aligned}\Gamma(\mathrm{Re}_{S,\alpha}) &= \left[\Gamma(z)\Gamma(\hat{R}_z) + \Gamma(y)\Gamma(\hat{R}_y) + \Gamma(x)\Gamma(\hat{R}_x)\right]\Gamma(\mathrm{Re}_{L,\alpha}) \\ &= \Gamma(\mathbf{r}\cdot\mathbf{R})\Gamma(\mathrm{Re}_{L,\alpha}).\end{aligned} \tag{10.52}$$

The same symmetry operation connects the other corresponding large- and small-component functions. The symmetry of this operator is given for all groups in table 10.1.

Table 10.1 Symmetry of operators connecting the large- and small-component functions

Group	$\Gamma(\boldsymbol{r})$	$\Gamma(\boldsymbol{R})$	$\Gamma(\boldsymbol{r}\cdot\boldsymbol{R})$
C_1	(A, A, A)	(A, A, A)	A
C_2	(B, B, A)	(B, B, A)	A
$C_n, n > 2$	(E_1, E_1, A)	(E_1, E_1, A)	A
C_{2v}	(B_1, B_2, A_1)	(B_2, B_1, A_2)	A_2
$C_{nv}, n > 2$	(E_1, E_1, A_1)	(E_1, E_1, A_2)	A_2
C_{2h}	(B_u, B_u, A_u)	(B_g, B_g, A_g)	A_u
$C_{2nh}, n > 1$	(E_{1u}, E_{1u}, A_u)	(E_{1g}, E_{1g}, A_g)	A_u
$C_{1h} \equiv C_s$	(A', A', A'')	(A'', A'', A')	A''
$C_{(2n+1)h},\ n > 0$	(E_1', E_1', A'')	(E_1'', E_1'', A')	A''
D_2	(B_3, B_2, B_1)	(B_3, B_2, B_1)	A
$D_n, n > 2$	(E_1, E_1, A_2)	(E_1, E_1, A_2)	A_1
D_{2nd}	(E_1, E_1, B_2)	$(E_{2n-1}, E_{2n-1}, A_2)$	B_1
$D_{(2n+1)d}$	(E_{1u}, E_{1u}, A_{2u})	(E_{1g}, E_{1g}, A_{2g})	A_{1u}
D_{2h}	(B_{3u}, B_{2u}, B_{1u})	(B_{3g}, B_{2g}, B_{1g})	A_u
$D_{2nh}, n > 1$	(E_{1u}, E_{1u}, A_{2u})	(E_{1g}, E_{1g}, A_{2g})	A_{1u}
$D_{(2n+1)h}$	(E_1', E_1', A_2'')	(E_1'', E_1'', A_2')	A_1''
S_{4n}	(E_1, E_1, B)	(E_1, E_1, A)	B
$S_2 \equiv C_i$	(A_u, A_u, A_u)	(A_g, A_g, A_g)	A_u
$S_{4n+2}, n > 0$	(E_{1u}, E_{1u}, A_u)	(E_{1g}, E_{1g}, A_g)	A_u
T	$\mathbf{T}$	$\mathbf{T}$	A
T_h	$\mathbf{T}_u$	$\mathbf{T}_g$	A_u
T_d	$\mathbf{T}_2$	$\mathbf{T}_1$	A_2
O	$\mathbf{T}_1$	$\mathbf{T}_1$	A_1
O_h	$\mathbf{T}_{1u}$	$\mathbf{T}_{1g}$	A_{1u}
I	$\mathbf{T}_1$	$\mathbf{T}_1$	A
I_h	$\mathbf{T}_{1u}$	$\mathbf{T}_{1g}$	A_u

For groups that have only one E irrep, the subscript “1” should be dropped.

To obtain all symmetry 2-spinors for D_{2h} and subgroups, it is only necessary to consider Γ_0 and $\Gamma(xyz)$ for $\Gamma(\mathrm{Re}_{L\alpha})$. The basic spinor symmetry vectors for D_{2h} and subgroups are

$$C_{2v}\colon \begin{pmatrix} A_1 + iA_2 \\ B_1 + iB_2 \end{pmatrix} \qquad D_2\colon \begin{pmatrix} A + iB_1 \\ B_2 + iB_3 \end{pmatrix} \tag{10.53}$$

$$C_2\colon \begin{pmatrix} A \\ B \end{pmatrix} \qquad C_s\colon \begin{pmatrix} A' \\ A'' \end{pmatrix} \tag{10.54}$$

$$C_1\colon \begin{pmatrix} A \\ A \end{pmatrix} \tag{10.55}$$

from which the Kramers partner symmetries can be obtained using the above expression, and the symmetries for the groups with inversion by adding g or u for the inversion operation.

The symmetry relations between the components influence the phases of the atomic 2-spinors that we might use as a basis. For example, consider a $p_{1/2}$ spinor on the axis or the center of symmetry. The spinor can be written

$$\chi(p_{1/2,1/2}) = \frac{1}{\sqrt{3}} \begin{pmatrix} p_z \\ p_x + ip_y \end{pmatrix}. \tag{10.56}$$

This spinor would have the following symmetries in C_{2v} and D_2:

$$C_{2v}\colon \begin{pmatrix} A_1 \\ B_1 + iB_2 \end{pmatrix} \qquad D_2\colon \begin{pmatrix} B_1 \\ iB_2 + B_3 \end{pmatrix}. \tag{10.57}$$

Thus, in D_2 this spinor should be multiplied by i to form a basis function. Of course, the phase is to some extent arbitrary, but the relative phases for each symmetry are fixed. If we did not multiply the $p_{1/2}$ spinor by i in D_2, the one-particle matrices would no longer be real but would have a mixture of real and imaginary elements. This would be undesirable in any implementation.

Further information on double groups and basis functions for representations can be found in appendix D .

10.4 Symmetries of Two-Particle States

For systems with more than one electron, spinors appear as part of an N-particle wave function. It is therefore of interest to describe how the symmetries of the constituent spinors propagate through a many-particle product. Here we will discuss the simplest case, that of a two-particle state. This case illustrates the principles for extension to more particles.

For two-particle states the wave function must transform according to a boson irrep. The simplest case is that of a two-particle state made up of a spinor and its Kramers partner. In the absence of other degeneracies this corresponds to a closed shell, and we would expect this product to transform as the totally symmetric irrep.

From the example of C_2, we see that indeed $E_{1/2} \otimes E_{-1/2} = A$, and also that $E_{1/2} \otimes E_{1/2} = E_{-1/2} \otimes E_{-1/2} = B$. Thus, for the simple groups where the irreps are all singly degenerate, the symmetries of the two-particle states are easily determined from the multiplication table. This is not the case for groups that have doubly degenerate irreps.

For the general two-particle case, the wave function is found as the direct product of the 4-spinors. As before, we concentrate on the large- and small-component 2-spinors, and get

$$\psi_1\psi_2 = \begin{pmatrix} \psi_1^L \\ \psi_1^S \end{pmatrix} \otimes \begin{pmatrix} \psi_2^L \\ \psi_2^S \end{pmatrix} = \begin{pmatrix} \psi_1^L\psi_2^L & \psi_1^L\psi_2^S \\ \psi_1^S\psi_2^L & \psi_1^S\psi_2^S \end{pmatrix}. \tag{10.58}$$

Each product of two 2-spinors consists of four components corresponding to the primitive spin basis {αα, αβ, βα, ββ}.

Concentrating on the relative symmetries of the real and imaginary parts of these blocks in the manner of the preceding section, we get for one such block

$$\psi_1^X\psi_2^Y = \begin{pmatrix} \phi_1^{X\alpha}\phi_2^{Y\alpha}\alpha\alpha & \phi_1^{X\alpha}\phi_2^{Y\beta}\alpha\beta \\ \phi_1^{X\beta}\phi_2^{Y\alpha}\beta\alpha & \phi_1^{X\beta}\phi_2^{Y\beta}\beta\beta \end{pmatrix}, \tag{10.59}$$

where, as before, X and Y may take the values L and S. Here we have explicitly displayed the spin labels that are normally implicit in the vector formalism as the vector basis.[3] Tracing the symmetries of the parts of this *bispinor*, we get

$$\begin{aligned}
&\Gamma(\psi_1^X\psi_2^Y) \\
&= \Gamma(\mathrm{Re}_{X,1\alpha})\Gamma(\mathrm{Re}_{Y,2\alpha}) \begin{pmatrix} \left[\Gamma_0 + i\Gamma(\hat{R}_z)\right]\Gamma(\alpha) \\ \left[\Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x)\right]\Gamma(\beta) \end{pmatrix} \otimes \begin{pmatrix} \left[\Gamma_0 + i\Gamma(\hat{R}_z)\right]\Gamma(\alpha) \\ \left[\Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x)\right]\Gamma(\beta) \end{pmatrix} \\
&= \Gamma(\mathrm{Re}_{X,1\alpha})\Gamma(\mathrm{Re}_{Y,2\alpha}) \begin{pmatrix} \left[\Gamma_0 + i\Gamma(\hat{R}_z)\right]\Gamma(\alpha\alpha) & \left[\Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x)\right]\Gamma(\alpha\beta) \\ \left[\Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x)\right]\Gamma(\beta\alpha) & \left[\Gamma_0 + i\Gamma(\hat{R}_z)\right]\Gamma(\beta\beta) \end{pmatrix},
\end{aligned} \tag{10.60}$$

where we have once again indicated the spin functions explicitly.

---

[3] The notation corresponds to an outer product of the two spinors.

One obvious problem with this bispinor form is that it does not carry time-reversal symmetry. The product of two fermion functions should transform as a boson and thus be symmetric with respect to time reversal. The simple products of spin functions clearly do not do this: $\hat{\mathcal{K}}(\alpha\alpha) = \beta\beta$, not $\alpha\alpha$. This problem can be remedied by making linear combinations of the primitive spin functions. Thus we can change basis to

$$\begin{aligned} &\frac{1}{\sqrt{2}}(\alpha\beta - \beta\alpha), \qquad &&\frac{i}{\sqrt{2}}(\alpha\beta + \beta\alpha), \\ &\frac{1}{\sqrt{2}}(\alpha\alpha + \beta\beta), \qquad &&\frac{i}{\sqrt{2}}(\alpha\alpha - \beta\beta), \end{aligned} \tag{10.61}$$

which are all symmetric under time reversal. We note further that these functions are linear combinations of primitive spin pairs that have the same symmetry for the spatial component, that is, $\alpha\alpha$ with $\beta\beta$ and $\alpha\beta$ with $\beta\alpha$. One of these spin functions is immediately recognizable as the singlet, $S = 0$, spin function:

$$S_0 = \frac{1}{\sqrt{2}}(\alpha\beta - \beta\alpha). \tag{10.62}$$

The other three are connected with the triplet, $S = 1$, state, but in a Cartesian representation:

$$\begin{aligned} S_{1z} &= \frac{1}{\sqrt{2}}(\alpha\beta + \beta\alpha), \\ S_{1x} &= \frac{1}{\sqrt{2}}(\alpha\alpha + \beta\beta), \\ S_{1y} &= \frac{i}{\sqrt{2}}(\alpha\alpha - \beta\beta). \end{aligned} \tag{10.63}$$

We can now determine the symmetry contributions from the spin functions. The S_0 function clearly transforms according to Γ_0. The rotational symmetry of the triplet spin functions follows from their properties as $J = 1$ angular momentum functions. In addition we need to consider the effect of inversion. But the product of two spin functions must always be of even, or gerade, parity.[4] Thus the symmetrized spin basis

---

[4] Actually, the spin functions themselves are of even parity—see Tinkham (1964).

functions of the triplet state transform as the Cartesian rotations $\hat{R}_x$, $\hat{R}_y$, $\hat{R}_z$. The total symmetry of the product of 2-spinors then becomes

$$\begin{aligned}
&\Gamma(\psi_1^X \psi_2^Y) \\
&= \Gamma(\mathrm{Re}_{X,1\alpha})\Gamma(\mathrm{Re}_{Y,2\alpha}) \begin{pmatrix} \left[\Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x)\right]\Gamma(\alpha\beta - \beta\alpha) \\ \left[\Gamma_0 + i\Gamma(\hat{R}_z)\right]\Gamma(\alpha\alpha + \beta\beta) \\ \left[\Gamma_0 + i\Gamma(\hat{R}_z)\right]i\Gamma(\alpha\alpha - \beta\beta) \\ \left[\Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x)\right]i\Gamma(\alpha\beta + \beta\alpha) \end{pmatrix} \\
&= \Gamma(\mathrm{Re}_{X,1\alpha})\Gamma(\mathrm{Re}_{Y,2\alpha}) \begin{pmatrix} \left[\Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x)\right]\Gamma_0 \\ \left[\Gamma_0 + i\Gamma(\hat{R}_z)\right]\Gamma(\hat{R}_x) \\ \left[\Gamma_0 + i\Gamma(\hat{R}_z)\right]\Gamma(\hat{R}_y) \\ \left[\Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x)\right]i\Gamma(\hat{R}_z) \end{pmatrix} \\
&= \Gamma(\mathrm{Re}_{X,1\alpha})\Gamma(\mathrm{Re}_{Y,2\alpha}) \begin{pmatrix} \Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x) \\ \Gamma(\hat{R}_x) + i\Gamma(\hat{R}_y) \\ \Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x) \\ \Gamma(\hat{R}_x) + i\Gamma(\hat{R}_y) \end{pmatrix}.
\end{aligned} \tag{10.64}$$

The end result of this process is a function that might not have a unique boson symmetry. If $\Gamma(\hat{R}_x)$ and $\Gamma(\hat{R}_y)$ belong to different irreps, then the product does not belong to a single irrep. The linear combinations $\hat{R}_x \pm i\hat{R}_y$ belong to different rows of degenerate irreps in many groups. The information given here does not take signs into account, but it can be shown by explicit consideration of the signs that the four terms do indeed belong to the same row of the same irrep, and the product may then belong to one row of a degenerate irrep.

Otherwise, the product can span two irreps. In this case, we can take a linear combination of the product and its time-reversed partner to obtain a function that transforms as an irrep. For example, in C_{2v} the functions that transform as boson irreps are

$$\begin{aligned}
&A_1: \quad E_{1/2}E_{-1/2} - E_{-1/2}E_{1/2} \qquad B_1: \quad E_{1/2}E_{1/2} - E_{-1/2}E_{-1/2} \\
&A_2: \quad E_{1/2}E_{-1/2} + E_{-1/2}E_{1/2} \qquad B_2: \quad E_{1/2}E_{1/2} + E_{-1/2}E_{-1/2}.
\end{aligned} \tag{10.65}$$

There is a simple prescription that can be used to determine the symmetries of product wave functions for groups other than the cubic and icosahedral groups. The basis functions can be represented by angular momentum eigenfunctions on the symmetry axis (or center). Due to the finite order of the principal rotation axis, the angular momentum eigenfunctions with a given value of $m \bmod n$ belong to the same irrep, where m is the azimuthal quantum number and n is the order of the axis. To determine the symmetry of a product, simply add the m quantum numbers and perform the modulo division. The remainder shows which irrep the product belongs to. However, if the result is zero, or half the order of the group for groups with n even, it is necessary to take linear combinations of the product with its time-reversed partner to obtain functions that transform as a single irrep. These rules and information on basis functions are given in appendix D.

10.5 Matrix Elements and Symmetry

Having discussed the symmetry of the spinors and products of spinors, we are now in a position to discuss the symmetries of the one- and two-electron integrals. For computational applications, the integrals over molecular orbitals are often divided into symmetry classes for convenient handling of symmetry. In addition, the consideration of symmetry may produce some simplification in the expressions for the many-electron Hamiltonian. In the relativistic case we must use both point-group and time-reversal symmetry.

The nonrelativistic case is relatively straightforward because the integrals are real for nondegenerate irreps, and can be made real for degenerate irreps by transformation to a real basis. The one-electron integrals f_{pq} must have both p and q belonging to the same irrep because the integral itself must belong to the totally symmetric irrep[5]. Denoting the symmetry species as $I, J, K, L, \ldots$, it is clear that the two-electron integrals fall into one of four symmetry classes: $(II|II)$, $(II|JJ)$, $(IJ|JI)$ and $(IJ|KL)$. In the last class, not all combinations produce the totally symmetric irrep, so this class is restricted to only those combinations that do. Note that the product of boson irreps II belongs to the totally symmetric irrep, and that the product IJ must belong to another irrep.

The Kramers-restricted relativistic case is a little more complicated. First, because the basis is complex, we must use the conjugate representation for the bra in any integral. In the notation used above, we can express this as

$$\langle A \rangle = \langle p|\hat{\Omega}|q\rangle \rightarrow \Gamma(A) = \Gamma(p^{\dagger})\Gamma(\hat{\Omega})\Gamma(q) = \Gamma(\bar{p})\Gamma(\hat{\Omega})\Gamma(q). \qquad (10.66)$$

So if ψ_p and ψ_q are from the same row of the same irrep, the product $\psi_p^{\dagger}\psi_q$ transforms as $\Gamma(\bar{p})\Gamma(q)$. It is this product that contains the symmetric irrep. Thus, for

---

[5] Here we refer to the orbitals or spinors by their indices.

the one-electron integrals over a symmetric operator, such as the Fock matrix f_{pq}, p and q must belong to the same row of the same irrep, just as in the nonrelativistic case.[6]

The symmetry reduction for the one-electron integrals follows from the consideration of time-reversal symmetry:

1. For the quaternion irreps, both f_{pq} and $f_{p\bar{q}}$ are nonzero and complex.
2. For the complex irreps, the second of these is zero, giving a reduction of a factor of 2 in the number of nonzero integrals.
3. For the real irreps, the first matrix element is also real, giving another factor of 2 in the number of independent quantities.
4. Further reductions may arise from the remaining symmetry operations, such as inversion, which provide good fermion quantum numbers.

When considering symmetry reductions, it is important to take into account the number of independent quantities, not simply the number of integrals. If we consider the groups C_1, C_2, and C_{2v}, we can transform f_{pq} and $f_{p\bar{q}}$ into a single quaternion matrix for C_1, and thus the integral count is the same for all three groups for a given basis. However, the number of independent quantities per integral is 4 for C_1, 2 for C_2, and 1 for C_{2v}.

For the two-electron integrals, we want to divide the integrals into symmetry classes, as for the nonrelativistic integrals. We also want to divide the integrals into classes by time reversal symmetry, as we did for the one-electron integrals. Because of the structure of the Kramers-restricted Hamiltonian in terms of the one- and two-particle Kramers replacement operators, we hope to obtain a reduction in the expression for the Hamiltonian from time-reversal symmetry. The classification by time-reversal properties is also important for the construction of the many-electron Hamiltonian matrix, whose symmetry properties we consider in the next section.

First, we consider the classification of the integrals $(pq|rs)$ into symmetry classes, $(II|II)$, $(II|JJ)$, $(IJ|IJ)$, and $(IJ|KL)$. For simplicity we consider only integrals over the Coulomb interaction. To make the classification, we need to know what the symmetry of a spinor overlap density is. We do not need to take the spin into account, as we did for the two-particle wave functions, because the spin is integrated out. Using the tools developed in section 10.3, and the result in (10.47), the symmetry of a charge density can be represented as follows:

$$\begin{aligned}\Gamma(\psi_p^{X\dagger}\psi_q^X) &= \begin{pmatrix}\Gamma_0 + i\Gamma(\hat{R}_z)\\ \Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x)\end{pmatrix}^\dagger \Gamma(\mathrm{Re}_{X,\alpha,p})\Gamma(\mathrm{Re}_{X,\alpha,q})\begin{pmatrix}\Gamma_0 + i\Gamma(\hat{R}_z)\\ \Gamma(\hat{R}_y) + i\Gamma(\hat{R}_x)\end{pmatrix}\\ &= \Gamma(\mathrm{Re}_{X,\alpha,p})\Gamma(\mathrm{Re}_{X,\alpha,q})\left[\Gamma_0 + i\Gamma(\hat{R}_z)\right]. \end{aligned} \tag{10.67}$$

---

[6] A reference to the C_2 multiplication table (page 157) will establish this point.

We note that for the electron density, we must have $p = q$. The total symmetry is $\Gamma(p)\Gamma(\bar{p}) = \Gamma_0$, and so the density is totally symmetric, as it should be.

We have now shown that the symmetry of a spinor overlap density depends on the symmetry of Γ_0 and $\Gamma(\hat{R}_z)$. These two symmetries are the same in the groups C_n, C_{nh}, and S_{2n}, but not in the other groups. If the product of the ket irrep and the conjugate of the bra irrep lies in one of the degenerate boson irreps, the density could have a well-defined symmetry. In analogy to the two-electron wave functions, we can establish the symmetry by subtracting the m_j values for the irreps and performing the modulo division to arrive at an m_ℓ value for a boson irrep.

However, if the product spans nondegenerate boson irreps, the symmetry is not well-defined, and it is the real and imaginary parts of the density that have a well-defined symmetry. This is the case even for $p = q$ and $r = s$, for the groups in which $\Gamma(\hat{R}_z) \neq \Gamma_0$. We can of course make the classification of the integrals into the four classes regardless of whether the densities belong to a particular boson irrep, but if we wish to exploit boson symmetry we have to address this issue.

The extension of these symmetry consideration to the Gaunt or Breit interactions proceeds along similar lines, but is more complicated due to the appearance of the current density and is omitted here.

The reduction of the number of integrals due to time-reversal symmetry follows a similar pattern to that of the one-electron integrals. We consider the case of D_{2h} and subgroups, where all irreps belong to one of the three types: quaternion, complex, or real. The analysis for groups that have more than one irrep type follow.

1. For groups that have only quaternion irreps, all integrals are nonzero and complex.
2. For groups that have only complex irreps, integrals with an odd number of bars are zero. This results in a reduction of a factor of 2 in the number of nonzero integrals.
3. For groups that have only real irreps, the matrix elements are also real, giving another factor of 2 reduction in the number of independent quantities.

Classification by time-reversal relations overlaps the symmetry classification. For groups without quaternion irreps, integrals in the classes $(II|II)$ and $(II|JJ)$ must have an even number of bars for each electron: thus, only integrals of the type $(pq|rs)$ and those to which it is related by time reversal can belong to these classes. Integrals of the type $(p\bar{q}|r\bar{s})$ and $(p\bar{q}|\bar{r}s)$ belong to the $(IJ|IJ)$ or $(IJ|KL)$ classes.

These time-reversal reductions affect the expressions for the second-quantized Kramers-restricted Hamiltonian. Both forms given in chapter 9 contain integrals with an odd number of bars, and in both, these terms vanish if the group has no quaternion irreps. If we want to use the Hamiltonian as expressed in terms of the excitation operators $\hat{E}^s_{pq}$ and $\hat{e}^{s_1 s_2}_{pq,rs}$, such as in a coupled-cluster calculation, the integrals are already partitioned according to the real and imaginary parts of the densities, and these parts have a particular boson symmetry for nondegenerate irreps and belong to a particular row of a degenerate irrep in a real basis. The symmetry reduction in the Hamiltonian is obvious from the structure of the expressions, and follows directly from the time-reversal reductions: half the integrals vanish if there are no quaternion irreps, and another half vanishes if there are no complex irreps. Thus, for the groups that contain only complex irreps, we have

$$\begin{aligned}\hat{\mathcal{H}} = &\sum_{pq} \Big[\operatorname{Re} h_{pq} \hat{E}^{+}_{pq} + i \operatorname{Im} h_{pq} \hat{E}^{-}_{pq} \Big] \\ &+ \frac{1}{2} \sum_{pqrs} \Big[(\operatorname{Re} pq|\operatorname{Re} rs) \hat{e}^{++}_{pq,rs} + i (\operatorname{Re} pq|\operatorname{Im} rs) \hat{e}^{+-}_{pq,rs} \\ &\qquad + i (\operatorname{Im} pq|\operatorname{Re} rs) \hat{e}^{-+}_{pq,rs} - (\operatorname{Im} pq|\operatorname{Im} rs) \hat{e}^{--}_{pq,rs} \\ &\qquad + (\operatorname{Re} p\bar{q}|\operatorname{Re} \bar{r}s) \hat{e}^{++}_{p\bar{q},\bar{r}s} + i (\operatorname{Re} p\bar{q}|\operatorname{Im} \bar{r}s) \hat{e}^{+-}_{p\bar{q},\bar{r}s} \\ &\qquad + i (\operatorname{Im} p\bar{q}|\operatorname{Re} \bar{r}s) \hat{e}^{-+}_{p\bar{q},\bar{r}s} - (\operatorname{Im} p\bar{q}|\operatorname{Im} \bar{r}s) \hat{e}^{--}_{p\bar{q},\bar{r}s} \Big],\end{aligned} \tag{10.68}$$

and for the groups that contain only real irreps we have

$$\begin{aligned}\hat{\mathcal{H}} = &\sum_{pq} \operatorname{Re} h_{pq} \hat{E}^{+}_{pq} \\ &+ \frac{1}{2} \sum_{pqrs} \Big[(\operatorname{Re} pq|\operatorname{Re} rs) \hat{e}^{++}_{pq,rs} - (\operatorname{Im} pq|\operatorname{Im} rs) \hat{e}^{--}_{pq,rs} \\ &\qquad + (\operatorname{Re} p\bar{q}|\operatorname{Re} \bar{r}s) \hat{e}^{++}_{p\bar{q},\bar{r}s} - (\operatorname{Im} p\bar{q}|\operatorname{Im} \bar{r}s) \hat{e}^{--}_{p\bar{q},\bar{r}s} \Big].\end{aligned} \tag{10.69}$$

In each case, the reduction in the types of integrals leads to a reduction in the types of excitation operators that need to be considered. Any implementation must therefore take account of the fact that different numbers and kinds of excitation operators are required for different irrep types. We see here how this symmetry reduction appears naturally when we use the $\hat{E}^{s}_{pq}$ operators. This would not be the case with the $\hat{X}^{s}_{pq}$.

10.6 Time Reversal and Symmetry in the Many-Electron Hamiltonian

We have seen how time-reversal symmetry and double-group symmetry are intimately connected in the matrices of one- and two-electron operators. These two symmetries are just as intimately connected in the many-electron Hamiltonian matrix.

As for the one-electron matrices, we must choose a representation for the basis with which the matrices are represented. We take as our N-particle basis the determinant basis introduced in chapter 9, given in terms of A and B strings, and consider all possible determinants that may be constructed from a given set of Kramers pairs. We group these determinants into subsets with a given value of N_A and N_B, characterized by a pseudo-quantum number M_K,

$$M_K = \tfrac{1}{2}(N_A - N_B). \tag{10.70}$$

In the nonrelativistic case, with α and β strings, $M_K = M_S$. Because the nonrelativistic operators are spin-independent, the Hamiltonian matrix is blocked by M_S. This block diagonalization of the Hamiltonian matrix does not persist in the relativistic case, and in the absence of any point group symmetry, the N-particle basis extends to all M_K values.

After dividing the determinants into subsets defined by their M_K values, we order the subsets from highest to lowest M_K value. The subsets define a partitioning of the Hamiltonian matrix **H** into blocks. Determinants from sets whose M_K values differ by more than two have zero Hamiltonian matrix elements between them, because the excitation between them is more than a two-electron excitation, and the Hamiltonian contains at most two-electron operators. With the arrangement of the determinants in M_K blocks from highest to lowest, **H** is therefore block pentadiagonal. This structure is shown in figures 10.1 and 10.2 for an even and an odd number of electrons.

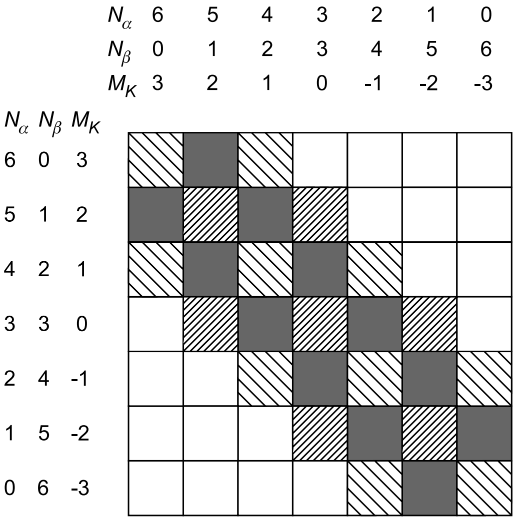


Figure 10.1 Diagram of Hamiltonian block structure for N even. The blocks are ordered by M_K

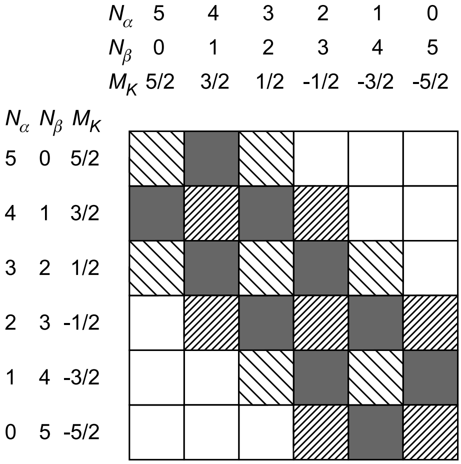


Figure 10.2 Diagram of Hamiltonian block structure for N odd. The blocks are ordered by M_K

We now consider the composition of the Hamiltonian matrix elements H_{PQ} between determinants P and Q on each block diagonal. For convenience we will simply refer to these as diagonals. For the main diagonal, $|M_K(P) - M_K(Q)| = 0$, and there must be the same number of A and B spinors. The matrix elements consist of one-electron integrals of the class h_{tu}, and two-electron integrals of the classes $(tu|vw)$ and $(t\bar{u}|\bar{v}w)$. These classes include all integrals that are related by time-reversal symmetry. For the first off-diagonal, $|M_K(P)-M_K(Q)| = 1$, and one determinant must have an A spinor where the other has a B spinor. The matrix elements are therefore composed of integral classes $h_{t\bar{u}}$ and $(tu|v\bar{w})$. For the second off-diagonal, $|M_K(P)-M_K(Q)| = 2$, and one determinant must have two A spinors where the other has two B spinors. The matrix elements are therefore composed of the integral class $(t\bar{u}|v\bar{w})$.

The classes of two-electron integrals required for the construction of the matrix elements are therefore different for each block diagonal, and the use of time-reversal properties to classify the integrals provides an efficiency in the construction of the Hamiltonian matrix.

To proceed with the analysis of the structure of **H** and the relation between time-reversal and point-group symmetry, we consider the case of D_{2h} and subgroups, the so-called “binary” groups. There are two good reasons for this. First, these groups are particularly convenient for computational implementations. As all the operations are twofold, only a parity factor (or sign bit) is required to describe the effect of an operation on a function. In addition, each group contains only one of the three types of irreps described previously, and it is therefore customary to classify these groups accordingly:

- The groups D_{2h}, D_2, and C_{2v}, whose fermion irreps are all real, will be called *real groups*.
- The groups C_{2h}, C_2, and C_s, whose fermion irreps are all complex, will be called *complex groups*.
- The groups C_i and C_1, whose fermion irreps are all quaternion, will be called *quaternion groups*.

The advantage of having to deal with only one type of fermion irrep will become apparent below.

With the restriction to binary group symmetry, we realize that the integrals on the first off-diagonal are only nonzero for the quaternion groups. For the two other types of binary groups, a spinor and its time-reversed partner belong either to different one-dimensional irreps (complex groups) or to different rows of the same two-dimensional irrep (real groups), and therefore integrals of the types $h_{t\bar{u}}$ and $(tu|v\bar{w})$ must vanish. The elements on the first off-diagonal are represented by the grey shading in the diagrams, and we see that when these disappear for the real and complex groups, the diagonal and second off-diagonal blocks partition into two disjoint, interleaving groups, represented by the striped white and grey shading. As a consequence the determinants may be partitioned into two sets. The relation between these sets depends on the number of electrons:

- For N odd, the N-electron wave function must transform under one of the fermion irreps. In this case time reversal of one set produces the other set. We see this from figure 10.2 by inversion through the center of the diagram.

The two sets of determinants therefore form a basis for the doubly degenerate Kramers pairs of the N-electron states. For the fermion irreps of the binary groups, the only other element of symmetry that can be exploited is inversion, which may be handled in the same way as in nonrelativistic CI theory.

- For N even, the overall symmetry of the wave function must be that of a boson irrep. The application of time reversal maps each set of determinants onto itself. The two sets form bases for states of different boson symmetry, and the set with M_K even contains the basis for the totally symmetric irrep. For the complex binary groups C_2 and C_s there are only two boson irreps, and so the Hamiltonian matrix is fully blocked by the use of Kramers pairs. Thus selection of the parity of M_K determines the state symmetry. The last of the complex groups, C_{2h}, has inversion symmetry, which gives rise to two more boson irreps, but again inversion may be treated by nonrelativistic methods.

Thus, apart from inversion, time-reversal symmetry provides for full symmetry blocking of the Hamiltonian matrix for an odd number of electrons in complex or real groups and for an even number of electrons in complex groups.

For an even number of electrons in real groups, the remaining element of symmetry is introduced by forming a real basis in the manner described in (9.58). From the consideration of the Hamiltonian matrix in this real basis in (9.60), it may be seen that the elements of **H** in the real basis that connect the positive and negative combinations come from the imaginary part of the elements in the determinant basis. However, for the real groups, the Hamiltonian must be real because all the integrals from which it is constructed are real. Therefore, these blocks must be zero, and the Hamiltonian is partitioned into four blocks, defined by the parity of M_K and the sign of the combination of determinants in the real basis. The symmetry of each set of such configuration state functions (CSFs) may be deduced from the discussion of the symmetry of two-particle states above.

The matrix elements between the CSFs in the real basis involve two Hamiltonian matrix elements between determinants, H_{PQ} and $H_{P\bar{Q}}$. However, $H_{P\bar{Q}}$ is only nonzero for $|M_K(P)-M_K(\bar{Q})| \leq 2$, because otherwise the excitation between P and $\bar{Q}$ is more than a two-particle excitation. Thus it is only in the center blocks of the Hamiltonian that the linear combination of matrix elements needs to be taken.

The only remaining case to consider is that of the quaternion groups with N odd. The Hamiltonian matrix may be blocked by the quaternion transformation given previously,

$$\begin{pmatrix} \Phi' \\ \Phi'' \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix} \begin{pmatrix} \Phi \\ \bar{\Phi} \end{pmatrix}, \tag{10.71}$$

producing two quaternion matrices of half the rank, with matrix elements

$$\begin{aligned} H_{P'Q'} &= H_{PQ} - H_{P\bar{Q}}j \\ H_{P''Q''} &= H_{\bar{P}\bar{Q}} - H_{\bar{P}Q}j \\ H_{P'Q''} &= H_{P''Q'} = 0. \end{aligned} \tag{10.72}$$

As for the real groups for N even, the blocks for which both H_{PQ} and $H_{P\bar{Q}}$ are nonzero occur only for $|M_K(P) - M_K(\bar{Q})| \leq 2$. Although the amount of work forming the Hamiltonian or forming its product with a vector would be the same without the transformation, the indexing work is halved, and therefore it may be useful to perform the transformation.

11

Basis-Set Expansions of Relativistic Electronic Wave Functions

There have been several successful applications of the Dirac–Hartree–Fock (DHF) equations to the calculation of numerical electronic wave functions for diatomic molecules (Laaksonen and Grant 1984a, 1984b, Sundholm 1988, 1994, Kullie et al. 1999). However, the use of numerical techniques in relativistic molecular calculations encounters the same difficulties as in the nonrelativistic case, and to proceed to general applications beyond simple diatomic and linear molecules it is necessary to resort to an analytic approximation using a basis set expansion of the wave function. The techniques for such calculations may to a large extent be based on the methods developed for nonrelativistic calculations, but it turns out that the transfer of these methods to the relativistic case requires special considerations. These considerations, as well as the development of the finite basis versions of both the Dirac and DHF equations, form the subject of the present chapter.

In particular, in the early days of relativistic quantum chemistry, attempts to solve the DHF equations in a basis set expansion sometimes led to unexpected results. One of the problems was that some calculations did not tend to the correct nonrelativistic limit. Subsequent investigations revealed that this was caused by inconsistencies in the choice of basis set for the small-component space, and some basic principles of basis-set selection for relativistic calculations were established. The variational stability of the DHF equations in a finite basis has also been a subject of debate. As we show in this chapter, it is possible to establish lower variational bounds, thus ensuring that the iterative solution of the DHF equations does not collapse.

There are two basically different strategies that may be followed when developing a finite basis formulation for relativistic molecular calculations. One possibility is to expand the large and small components of the 4-spinor in a basis of 2-spinors. The alternative is to expand each of the scalar components of the 4-spinor in a scalar basis. Both approaches have their advantages and disadvantages, though the latter approach is obviously the easier one for adapting nonrelativistic methods, which work in real scalar arithmetic. The two approaches differ in the actual implementation and are therefore to some extent mutually exclusive, though there are ways of combining the benefits of both approaches. For purposes of analysis the 2-spinor formulation is more transparent and powerful, and we therefore start with this approach as a basis for discussing the problems mentioned above.

11.1 The Dirac Equation in 2-Spinor Form

In order to analyze the consequences of basis set expansion of the relativistic wave equation for electrons, it is sufficient to consider the one-electron Dirac equation. The only other fundamental complication we foresee when going to a many-electron model is continuum dissolution (Brown–Ravenhall disease), which we have already dealt with in chapter 5 and so need not consider in this context.

We start by writing the 4-spinor ψ in the 2-spinor decomposition

$$\psi = \begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix}. \tag{11.1}$$

A natural choice for the expansion of the wave function would be the atomic 2-spinors of the form

$$\psi^X(\mathbf{r}) = \phi^X(r)\xi_{\kappa,m}(\vartheta,\varphi) = \phi^X(r)\frac{1}{\sqrt{2\ell+1}}\begin{pmatrix} a\sqrt{\ell+\frac{1}{2}+am}\;Y_\ell^{m-1/2} \\ \\ \sqrt{\ell+\frac{1}{2}-am}\;Y_\ell^{m+1/2} \end{pmatrix}, \tag{11.2}$$

but the discussion below is not restricted to this type of basis function, and would in principle apply to any suitable form of 2-spinor basis.

We expand the large and small components in a basis of 2-spinors, $\{\chi_\mu^L, \chi_\mu^S\}$

$$\psi^L = \sum_{\mu=1}^{N_L} \mathrm{a}_\mu^L\, \chi_\mu^L, \quad \psi^S = \sum_{\mu=1}^{N_S} \mathrm{a}_\mu^S\, \chi_\mu^S, \tag{11.3}$$

where we note that N_L and N_S need not be equal. Inserting these expansions in the time-independent Dirac equation, (4.81), we get

$$\begin{aligned} (V-E)\sum_{\mu=1}^{N_L} \mathrm{a}_\mu^L\, \chi_\mu^L \quad &+ \quad c(\boldsymbol{\sigma}\cdot\mathbf{p})\sum_{\mu=1}^{N_S} \mathrm{a}_\mu^S\, \chi_\mu^S = 0 \\ c(\boldsymbol{\sigma}\cdot\mathbf{p})\sum_{\mu=1}^{N_L} \mathrm{a}_\mu^L\, \chi_\mu^L \quad &+ \quad (V-E-2mc^2)\sum_{\mu=1}^{N_S} \mathrm{a}_\mu^S\, \chi_\mu^S = 0. \end{aligned} \tag{11.4}$$

This we can turn into a matrix equation in the expansion coefficients **a**, and we get the one-electron Dirac equation in a basis expansion as

$$\begin{pmatrix} \mathbf{V}^{LL} - E\,\mathbf{S}^{LL} & c\boldsymbol{\Pi}^{LS} \\ c\boldsymbol{\Pi}^{SL} & \mathbf{V}^{SS} - (2mc^2 + E)\,\mathbf{S}^{SS} \end{pmatrix} \begin{pmatrix} \mathbf{a}^L \\ \mathbf{a}^S \end{pmatrix} = 0. \tag{11.5}$$

The matrices appearing here are defined by

$$V^{LL}_{\mu\nu} = \langle \chi^L_\mu | V | \chi^L_\nu \rangle; \quad V^{SS}_{\mu\nu} = \langle \chi^S_\mu | V | \chi^S_\nu \rangle \tag{11.6}$$

$$S^{LL}_{\mu\nu} = \langle \chi^L_\mu | \chi^L_\nu \rangle; \quad S^{SS}_{\mu\nu} = \langle \chi^S_\mu | \chi^S_\nu \rangle \tag{11.7}$$

$$\Pi^{LS}_{\mu\nu} = \langle \chi^L_\mu | -i\hbar\boldsymbol{\sigma}\cdot\nabla | \chi^S_\nu \rangle; \quad \Pi^{SL}_{\mu\nu} = \langle \chi^S_\mu | -i\hbar\boldsymbol{\sigma}\cdot\nabla | \chi^L_\nu \rangle. \tag{11.8}$$

These are the representations of the overlap, the potential energy, and the kinetic energy operators in the expansion basis. Note that $\Pi^{SL}_{\mu\nu} = (\Pi^{LS}_{\nu\mu})^\dagger$, so that

$$\boldsymbol{\Pi} = \begin{pmatrix} \mathbf{0} & \boldsymbol{\Pi}^{LS} \\ \boldsymbol{\Pi}^{SL} & \mathbf{0} \end{pmatrix} \tag{11.9}$$

is Hermitian.

11.2 Kinetic Balance

Elimination (or isolation) of the small component provides a useful basis for discussion of the properties of the Dirac equation. In particular, in this section we want to develop the relationship between the large- and small-component basis sets. We write the matrix 2-spinor Dirac equation in the form of two coupled matrix equations,

$$\begin{gathered} \left[\mathbf{V}^{LL} - E\mathbf{S}^{LL}\right] \mathbf{a}^L + c\,\boldsymbol{\Pi}^{LS}\,\mathbf{a}^S = 0 \\ c\,\boldsymbol{\Pi}^{SL}\mathbf{a}^L + \left[\mathbf{V}^{SS} - (2mc^2 + E)\mathbf{S}^{SS}\right] \mathbf{a}^S = 0. \end{gathered} \tag{11.10}$$

$\mathbf{V}^{SS}$ is negative definite for most atomic and molecular potentials. Assuming that we are looking for solutions above the negative-energy continuum, where $E > -2mc^2$, it is permissible to invert the term in brackets in the second of these equations to get

$$\mathbf{a}^S = \left[(2mc^2 + E)\mathbf{S}^{SS} - \mathbf{V}^{SS}\right]^{-1} c\boldsymbol{\Pi}^{SL}\mathbf{a}^L. \tag{11.11}$$

We now eliminate the small-component expansion coefficients, $\mathbf{a}^S$, from the first equation to obtain

$$\left[\mathbf{V}^{LL} - E\mathbf{S}^{LL}\right] \mathbf{a}^L + c\boldsymbol{\Pi}^{LS}\left[(2mc^2 + E)\mathbf{S}^{SS} - \mathbf{V}^{SS}\right]^{-1} c\boldsymbol{\Pi}^{SL}\mathbf{a}^L = 0. \quad (11.12)$$

This is an equation for the expansion coefficients of the large component only, and corresponds to solving the equations by standard matrix partitioning techniques. For actual calculations we would gain nothing, because we also have to find the expansion coefficients for the small component, but for purposes of analysis this form turns out to be convenient. In order to display the dependence on c more clearly, we can expand the inverse operator in the equation above by using the matrix relation

$$(\mathbf{A} - \mathbf{B})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{A} - \mathbf{B})^{-1}. \quad (11.13)$$

Setting $2mc^2\mathbf{S}^{SS} = \mathbf{A}$ and $\mathbf{V}^{SS} - E\mathbf{S}^{SS} = \mathbf{B}$, we get

$$\begin{aligned}&\left[\mathbf{V}^{LL} - E\mathbf{S}^{LL} + \frac{1}{2m}\boldsymbol{\Pi}^{LS}\left[\mathbf{S}^{SS}\right]^{-1}\boldsymbol{\Pi}^{SL}\right] \mathbf{a}^L \\ &= \frac{1}{2m}\boldsymbol{\Pi}^{LS}\left[\mathbf{S}^{SS}\right]^{-1}\left[\mathbf{V}^{SS} - E\mathbf{S}^{SS}\right]\left[\mathbf{V}^{SS} - (2mc^2 + E)\mathbf{S}^{SS}\right]^{-1}\boldsymbol{\Pi}^{SL}\mathbf{a}^L. \quad (11.14)\end{aligned}$$

Here we have collected terms of order c^0 on the left-hand side of the equation, and terms of order c^{-2} and smaller on the right.

We are now in a position to discuss the first problem mentioned at the beginning of this chapter, which concerns the transition to the nonrelativistic limit. This limit should be approached if we increase c in our equations. Thus, if we let $c \to \infty$ in (11.14), the right-hand side vanishes, and we are left with

$$\left[\mathbf{V}^{LL} - E\mathbf{S}^{LL} + \frac{1}{2m}\boldsymbol{\Pi}^{LS}\left[\mathbf{S}^{SS}\right]^{-1}\boldsymbol{\Pi}^{SL}\right] \mathbf{a}^L = 0. \quad (11.15)$$

By analogy with the nonrelativistic equations, we conclude that the kinetic energy is represented by the matrix product,

$$\mathbf{T}^{LL} = \frac{1}{2m}\boldsymbol{\Pi}^{LS}\left[\mathbf{S}^{SS}\right]^{-1}\boldsymbol{\Pi}^{SL}. \quad (11.16)$$

If we insert the explicit expressions for the matrix elements, we get

$$T_{\mu\nu}^{LL} = -\frac{\hbar^2}{2m}\sum_{\kappa\lambda} \langle\chi_\mu^L|\boldsymbol{\sigma}.\nabla\underline{|\chi_\kappa^S\rangle\,[\mathbf{S}^{SS}]_{\kappa\lambda}^{-1}\,\langle\chi_\lambda^S|}\boldsymbol{\sigma}.\nabla|\chi_\nu^L\rangle. \quad (11.17)$$

The part of this expression that we have underlined constitutes an inner projection onto the small-component basis space. Now, $(\boldsymbol{\sigma}\cdot\nabla)(\boldsymbol{\sigma}\cdot\nabla) = \nabla^2$ according to vector relations we have used previously, and the nonrelativistic version of (11.17) is

$$T_{\mu\nu} = -\frac{\hbar^2}{2m}\langle\chi_\mu^L|\nabla^2|\chi_\nu^L\rangle = -\frac{\hbar^2}{2m}\langle\chi_\mu^L|(\boldsymbol{\sigma}\cdot\nabla)(\boldsymbol{\sigma}\cdot\nabla)|\chi_\nu^L\rangle. \tag{11.18}$$

However, the matrix equivalent of this relation as it appears in (11.17) does not hold unless the small-component basis spans the space $\{\boldsymbol{\sigma}\cdot\nabla\boldsymbol{\chi}_\mu^L\}$ (Dyall et al. 1984). In particular, if we choose

$$\chi_\mu^S = -i\hbar(\boldsymbol{\sigma}\cdot\nabla)\chi_\mu^L = (\boldsymbol{\sigma}\cdot\mathbf{p})\chi_\mu^L \tag{11.19}$$

then $\boldsymbol{\Pi}^{SL} = \mathbf{S}^{SS}$ and thus $\boldsymbol{\Pi}^{LS} = 2m\mathbf{T}^{LL} = 2m\mathbf{T}$.

This condition, which ensures that the kinetic energy is properly represented in the nonrelativistic limit, is known as *kinetic balance* (Stanton and Havriliak 1984). It is an important condition on the basis sets because it defines a relationship between the large- and small-component basis functions that must be satisfied in the nonrelativistic limit.[1]

The consequence of not satisfying this condition was observed in early basis set calculations (Rosicky and Mark 1979, Mark and Rosicky 1980, Mark et al. 1980). It can be shown (Dyall et al. 1984) that the kinetic energy is a maximum when the kinetic balance relation is satisfied, and therefore any approximation that does not fully span the space defined by (11.19) must lower the energy. This is indeed what was observed.

11.3 Variational Bounds

We now have a matrix form of the Dirac equation which tends to the right nonrelativistic limits, but we have no guarantee that the solutions of these equations are well-behaved. In particular, we need to ensure that the finite basis expansion does not produce solutions corresponding to energies below the "positive energy" electronic space, despite all our efforts to cast the equations in a form that should prevent this from occurring (chapter 5 and chapter 8).

The tool we use to analyze this problem is the Rayleigh quotient corresponding to (11.12),

$$R(E) = \mathbf{a}^{L\dagger}\Big[\mathbf{V}^{LL} + c\boldsymbol{\Pi}^{LS}\Big[(2mc^2+E)\mathbf{S}^{SS} - \mathbf{V}^{SS}\Big]^{-1} c\boldsymbol{\Pi}^{SL}\Big]\mathbf{a}^L/\mathbf{a}^{L\dagger}\mathbf{S}^{LL}\mathbf{a}^L. \tag{11.20}$$

To simplify the notation, we normalize the coefficients,

$$\mathbf{a} = \mathbf{a}^L/[\mathbf{a}^{L\dagger}\mathbf{S}^{LL}\mathbf{a}^L]^{1/2}, \tag{11.21}$$

---

[1] This condition was first formulated by Lee and McLean (1982a, 1982b) and Grant (1982).

so that we can write the Rayleigh quotient as

$$R(E) = \mathbf{a}^\dagger \mathbf{V}^{LL}\mathbf{a} \; + \mathbf{a}^\dagger \Big[c\boldsymbol{\Pi}^{LS} \Big[(2mc^2 + E)\mathbf{S}^{SS} - \mathbf{V}^{SS} \Big]^{-1} c\boldsymbol{\Pi}^{SL} \Big] \mathbf{a}. \tag{11.22}$$

We want to know if this quotient is bounded from below for what we expect to be bound-state energies, $0 > E > -2mc^2$. Obviously we are looking for a bound that is greater than $-2mc^2$; if we do not find one it is likely that there is no bound.

The Rayleigh quotient is a function of E because of the elimination of the small component. In order to find a lower bound, we need to remove E and identify a positive term in the quotient, so that we can create an inequality.

First, since $\mathbf{V}^{SS}$ is negative definite, $\mathbf{S}^{SS}$ is positive definite, and $2mc^2 + E > 0$, the inverse in (11.22) is positive definite. Defining a vector $\mathbf{b}$ by

$$\mathbf{b} = \boldsymbol{\Pi}^{SL}\mathbf{a} \tag{11.23}$$

we can write the second term of the Rayleigh quotient as

$$\begin{aligned} R_2 &= c^2 \; \mathbf{a}^\dagger \boldsymbol{\Pi}^{LS} \Big[(2mc^2 + E)\mathbf{S}^{SS} - \mathbf{V}^{SS} \Big]^{-1} \boldsymbol{\Pi}^{SL}\mathbf{a} \\ &= c^2 \mathbf{b}^\dagger \Big[(2mc^2 + E)\mathbf{S}^{SS} - \mathbf{V}^{SS} \Big]^{-1} \mathbf{b}. \end{aligned} \tag{11.24}$$

This term is the expectation of a positive definite matrix—the inverse—and is therefore positive. We can conclude that

$$R(E) > \mathbf{a}^\dagger \mathbf{V}^{LL}\mathbf{a}, \tag{11.25}$$

that is, the energy is bounded from below by the expectation of the potential. This is not a very useful bound, and it might be possible that it is lower than $-2mc^2$.

The term that was neglected to obtain the inequality is large: it contains the entire kinetic energy. We can partition it into two positive terms, one large and one small, using (11.13) with $\mathbf{A} = 2mc^2\mathbf{S}^{SS} - \mathbf{V}^{SS}$ and $\mathbf{B} = -E\mathbf{S}^{SS}$:

$$\begin{aligned} R_2 = {} & c^2 \; \mathbf{b}^\dagger \Big[2mc^2\mathbf{S}^{SS} - \mathbf{V}^{SS} \Big]^{-1} \mathbf{b} \\ & - c^2 E \; \mathbf{b}^\dagger \Big[2mc^2\mathbf{S}^{SS} - \mathbf{V}^{SS} \Big]^{-1} \mathbf{S}^{SS} \Big[(2mc^2 + E)\mathbf{S}^{SS} - \mathbf{V}^{SS} \Big]^{-1} \mathbf{b}. \end{aligned} \tag{11.26}$$

The first term is positive because the matrices in the inverse are both positive definite. The second term is positive because each matrix in the product is positive definite, and $-c^2E$ is also positive because E is negative. Now, E is formally $\mathcal{O}(c^0)$, so the second term is formally $\mathcal{O}(c^{-2})$. The first term is formally $\mathcal{O}(c^0)$, so the first term can be considered “large” and the second term “small.”

If we neglect the second term, we arrive at the inequality for the Rayleigh quotient:

$$R(E) > \mathbf{a}^\dagger \left[\mathbf{V}^{LL} + c^2 \boldsymbol{\Pi}^{LS}\left(2mc^2\mathbf{S}^{SS} - \mathbf{V}^{SS}\right)^{-1}\boldsymbol{\Pi}^{SL}\right]\mathbf{a}. \tag{11.27}$$

The expression on the right is the expectation of the matrix form of the ZORA Hamiltonian, which we will explore in depth in chapter 18. The ZORA Hamiltonian is a variational Hamiltonian and therefore does not suffer from variational collapse.

We can also derive a bound that is related to the expectation of the nonrelativistic Hamiltonian with a relativistic correction. To proceed, we use the matrix relation (11.13) twice to partition the second term of the Rayleigh quotient, R_2, with $\mathbf{A} = 2mc^2\mathbf{S}^{SS}$ and $\mathbf{B} = \mathbf{V}^{SS} - E\mathbf{S}^{SS}$. The first application yields

$$R_2 = \frac{1}{2m}\left[\mathbf{b}^\dagger\left(\mathbf{S}^{SS}\right)^{-1}\mathbf{b} + \mathbf{b}^\dagger\left(\mathbf{S}^{SS}\right)^{-1}\left(\mathbf{V}^{SS} - E\mathbf{S}^{SS}\right)\left[\left(2mc^2 + E\right)\mathbf{S}^{SS} - \mathbf{V}^{SS}\right]^{-1}\mathbf{b}\right]. \tag{11.28}$$

In this expression the first term is again positive, and it is $\mathcal{O}(c^0)$. We retain it so that we may find a better lower bound. Partitioning the inverse again, we get

$$\begin{aligned} R_2 =& \frac{1}{2m}\mathbf{b}^\dagger\left(\mathbf{S}^{SS}\right)^{-1}\mathbf{b} \\ &+ \frac{1}{4m^2c^2}\mathbf{b}^\dagger\left(\mathbf{S}^{SS}\right)^{-1}\left(\mathbf{V}^{SS} - E\mathbf{S}^{SS}\right)\left(\mathbf{S}^{SS}\right)^{-1}\mathbf{b} \\ &+ \frac{1}{4m^2c^2}\mathbf{b}^\dagger\left[\left(\mathbf{S}^{SS}\right)^{-1}\left(\mathbf{V}^{SS} - E\mathbf{S}^{SS}\right)\right]^2\left[\left(2mc^2 + E\right)\mathbf{S}^{SS} - \mathbf{V}^{SS}\right]^{-1}\mathbf{b}. \end{aligned} \tag{11.29}$$

The last term is clearly positive, because the inverse matrix is positive definite and the other term is the square of a matrix. Substituting for **b**, and neglecting the last term, we get the following inequality for the Rayleigh quotient:

$$\begin{aligned} R(E) >& \mathbf{a}^\dagger\left[\mathbf{V}^{LL} + \frac{1}{2m}\boldsymbol{\Pi}^{LS}\left(\mathbf{S}^{SS}\right)^{-1}\boldsymbol{\Pi}^{SL}\right]\mathbf{a} \\ &+ \frac{1}{4m^2c^2}\mathbf{a}^\dagger\boldsymbol{\Pi}^{LS}\left(\mathbf{S}^{SS}\right)^{-1}\left[\mathbf{V}^{SS} - E\mathbf{S}^{SS}\right]\left(\mathbf{S}^{SS}\right)^{-1}\boldsymbol{\Pi}^{SL}\mathbf{a}. \end{aligned} \tag{11.30}$$

If we choose a kinetically balanced basis set according to (11.19), this expression simplifies considerably, to

$$R(E) > \mathbf{a}^\dagger\left[\mathbf{V}^{LL} + \mathbf{T}\right]\mathbf{a} + \frac{1}{4m^2c^2}\mathbf{a}^\dagger\left[\mathbf{V}^{SS} - E\mathbf{S}^{SS}\right]\mathbf{a}. \tag{11.31}$$

The energy is therefore bounded from below by the sum of the expectation of the nonrelativistic Hamiltonian and a term which is $\mathcal{O}(c^{-2})$. Using (11.19) again, we can rewrite the matrix elements in the last term as

$$V_{pq}^{SS} - ES_{pq}^{SS} = \langle \chi_p^L | (\boldsymbol{\sigma}\cdot\mathbf{p})(V - E)(\boldsymbol{\sigma}\cdot\mathbf{p}) | \chi_q^L \rangle. \tag{11.32}$$

We will meet an expression similar to this in chapter 17, as the first-order relativistic perturbation correction to the nonrelativistic energy. We can derive an inequality for these matrix elements by noting that $E^{\mathrm{NR}} > E$, and therefore the expectation value of $-E\mathbf{S}^{SS}$ is greater than the expectation value of $-E^{\mathrm{NR}}\mathbf{S}^{SS}$. Thus,

$$R(E) \;>\; \mathbf{a}^{\dagger}\big[\mathbf{V}^{LL} + \mathbf{T}\big]\mathbf{a} + \frac{1}{4m^2c^2}\mathbf{a}^{\dagger}\big[\mathbf{V}^{SS} - E^{\mathrm{NR}}\mathbf{S}^{SS}\big]\mathbf{a}. \tag{11.33}$$

The Dirac eigenvalue is therefore bounded below by an expression that is the expectation of the nonrelativistic Hamiltonian plus the first-order relativistic perturbation correction to the energy evaluated with the large-component wave function.

The developments of this section show that for energy solutions in the domain of interest to us, the Rayleigh quotient is bounded below, and there is therefore *no* danger of variational collapse when solving the Dirac equation in a kinetically balanced finite basis. For the Dirac–Hartree–Fock equations, the only addition is the electron–electron interaction, which is positive and therefore will not contribute to a variational collapse.

The role of kinetic balance in these bounds is also critical. If the small-component basis set is not related to the large-component basis set by kinetic balance, the large positive term that involves $\boldsymbol{\Pi}$ is reduced toward zero, and the eigenvalues can approach those of the large-component potential as expressed in (11.25).

11.4 Matrix Dirac–Hartree–Fock Equations in a 2-Spinor Basis

The Dirac–Hartree–Fock operator for a single-determinant wave function in terms of atomic spinors has been derived previously (section 8.1), and appears as

$$f_{pq} = h_{pq} + \sum_{i}^{N}\left[(pq|ii) - (pi|iq)\right]. \tag{11.34}$$

While the expansion of h_{pq} follows readily from the development for the Dirac operator above, the electron–electron interaction integrals must be considered separately. We also want to develop a Kramers-restricted Dirac–Hartree–Fock (KR-DHF) theory, but first we develop expressions for the general, Kramers-unrestricted case. In the developments below we follow the practice of giving only the basis function index in the integrals.

For the Coulomb contribution to the Fock operator we have to consider the interaction between charge distributions involving two 4-spinors, for example, $\psi_p{}^{\dagger}\psi_q$. If this distribution is decomposed into 2-spinor form, we get

$$\psi_p^{\dagger}\psi_q = \begin{pmatrix}\psi_p^{L\dagger} & \psi_p^{S\dagger}\end{pmatrix}\begin{pmatrix}\psi_q^{L}\\ \psi_q^{S}\end{pmatrix} = \psi_p^{L\dagger}\psi_q^{L} + \psi_p^{S\dagger}\psi_q^{S}. \tag{11.35}$$

On substituting this expression into the two-electron Coulomb integral, we get

$$(pq|rs) = (p^Lq^L|r^Ls^L) + (p^Lq^L|r^Ss^S) + (p^Sq^S|r^Ls^L) + (p^Sq^S|r^Ss^S) \tag{11.36}$$

To go beyond the Coulomb approximation, we must include the Breit interaction (5.49), given by the operator

$$V^{\mathrm{C}}(0, r_{ij}) = \frac{1}{r_{ij}} - \frac{\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j}{r_{ij}} + \frac{(\boldsymbol{\alpha}_i \times \mathbf{r}_{ij}) \cdot (\boldsymbol{\alpha}_j \times \mathbf{r}_{ij})}{r_{ij}^3}. \tag{11.37}$$

The integrals over this operator depend on densities over $\boldsymbol{\alpha}$ or $\boldsymbol{\alpha} \times \mathbf{r}$. The integrals over the first (Gaunt) term are the same as the regular Coulombic electron repulsion integrals, while the second term requires calculation of integrals that are not encountered in nonrelativistic calculations. If we restrict our discussion to contributions from the Gaunt operator, we get densities of the form

$$\psi_p^\dagger \boldsymbol{\alpha} \psi_q = \left(\psi_p^{L\dagger}, \psi_p^{S\dagger}\right) \begin{pmatrix} \mathbf{0}_2 & \boldsymbol{\sigma} \\ \boldsymbol{\sigma} & \mathbf{0}_2 \end{pmatrix} \begin{pmatrix} \psi_q^L \\ \psi_q^S \end{pmatrix} = \psi_p^{S\dagger} \boldsymbol{\sigma} \psi_q^L + \psi_p^{L\dagger} \boldsymbol{\sigma} \psi_q^S. \tag{11.38}$$

Inserted into the integral, these give

$$\begin{aligned}\left(p\boldsymbol{\alpha}q|r\boldsymbol{\alpha}s\right) &= \left(p^S\boldsymbol{\sigma}q^L|r^L\boldsymbol{\sigma}s^S\right) + \left(p^L\boldsymbol{\sigma}q^S|r^S\boldsymbol{\sigma}s^L\right) + \left(p^S\boldsymbol{\sigma}q^L|r^S\boldsymbol{\sigma}s^L\right) \\ &\quad + \left(p^L\boldsymbol{\sigma}q^S|r^L\boldsymbol{\sigma}s^S\right).\end{aligned} \tag{11.39}$$

We can now introduce the basis set expansion from (11.3). For $\left(p^Lq^L|r^Ls^L\right)$ this yields

$$\left(p^Lq^L|r^Ls^L\right) = \sum_{\mu\nu\kappa\lambda} \mathrm{a}_{\mu p}^{L*}\mathrm{a}_{\nu q}^{L}\mathrm{a}_{\kappa r}^{L*}\mathrm{a}_{\lambda s}^{L}\left(\mu^L\nu^L|\kappa^L\lambda^L\right). \tag{11.40}$$

Similar expressions follow for the other combinations of components, and for the Gaunt integrals. To arrive at a compact expression for the Fock matrix elements, we define the 2-spinor density matrices by

$$\begin{aligned} D_{\mu\nu}^{LL} &= \sum_{i=1}^{N} \mathrm{a}_{\mu i}^{L\dagger}\mathrm{a}_{\nu i}^{L} & D_{\mu\nu}^{LS} &= \sum_{i=1}^{N} \mathrm{a}_{\mu i}^{L\dagger}\mathrm{a}_{\nu i}^{S} \\ D_{\mu\nu}^{SL} &= \sum_{i=1}^{N} \mathrm{a}_{\mu i}^{S\dagger}\mathrm{a}_{\nu i}^{L} & D_{\mu\nu}^{SS} &= \sum_{i=1}^{N} \mathrm{a}_{\mu i}^{S\dagger}\mathrm{a}_{\nu i}^{S}. \end{aligned} \tag{11.41}$$

Note that $D_{\mu\nu}^{LS} = D_{\nu\mu}^{SL\dagger}$, so that

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}^{LL} & \mathbf{D}^{LS} \\ \mathbf{D}^{SL} & \mathbf{D}^{SS} \end{pmatrix} \tag{11.42}$$

is Hermitian.

The DHF matrix can be written in terms of blocks between the components as

$$\mathbf{f} = \begin{pmatrix} \mathbf{f}^{LL} & \mathbf{f}^{LS} \\ \mathbf{f}^{SL} & \mathbf{f}^{SS} \end{pmatrix} \tag{11.43}$$

with the blocks defined by

$$\begin{aligned} f^{LL}_{\mu\nu} &= V^{LL}_{\mu\nu} + \sum_{\kappa\lambda} D^{LL}_{\kappa\lambda} \left\{ \left(\mu^L \nu^L \,|\, \kappa^L \lambda^L \right) - \left(\mu^L \lambda^L \,|\, \kappa^L \nu^L \right) \right\} \\ &\quad + \sum_{\kappa\lambda} D^{SS}_{\kappa\lambda} \left\{ \left(\mu^L \nu^L \,|\, \kappa^S \lambda^S \right) - \left(\mu^L \boldsymbol{\sigma} \lambda^S \,|\, \kappa^S \boldsymbol{\sigma} \nu^L \right) \right\}, \end{aligned} \tag{11.44}$$

$$\begin{aligned} f^{LS}_{\mu\nu} &= c\Pi^{LS}_{\mu\nu} + \sum_{\kappa\lambda} D^{LS}_{\kappa\lambda} \left\{ \left(\mu^L \boldsymbol{\sigma} \nu^S \,|\, \kappa^L \boldsymbol{\sigma} \lambda^S \right) - \left(\mu^L \boldsymbol{\sigma} \lambda^S \,|\, \kappa^L \boldsymbol{\sigma} \nu^S \right) \right\} \\ &\quad + \sum_{\kappa\lambda} D^{SL}_{\kappa\lambda} \left\{ \left(\mu^L \boldsymbol{\sigma} \nu^S \,|\, \kappa^S \boldsymbol{\sigma} \lambda^L \right) - \left(\mu^L \lambda^L \,|\, \kappa^S \nu^S \right) \right\}, \end{aligned} \tag{11.45}$$

$$\begin{aligned} f^{SS}_{\mu\nu} &= V^{SS}_{\mu\nu} - 2c^2 S^{SS}_{\mu\nu} + \sum_{\kappa\lambda} D^{LL}_{\kappa\lambda} \left\{ \left(\mu^S \nu^S \,|\, \kappa^L \lambda^L \right) - \left(\mu^S \boldsymbol{\sigma} \lambda^L \,|\, \kappa^L \boldsymbol{\sigma} \nu^S \right) \right\} \\ &\quad + \sum_{\kappa\lambda} D^{SS}_{\kappa\lambda} \left\{ \left(\mu^S \nu^S \,|\, \kappa^S \lambda^S \right) - \left(\mu^S \lambda^S \,|\, \kappa^S \nu^S \right) \right\}. \end{aligned} \tag{11.46}$$

We can collect the terms into direct and exchange contributions. Here we use the term "direct" rather than the traditional "Coulomb" because we must distinguish the Coulomb interaction, which gives rise to direct and exchange terms, from other interactions, such as the Gaunt interaction, which also give rise to direct and exchange terms. The direct contribution for the Coulomb interaction is in the diagonal blocks,

$$\begin{aligned} J^{XX}_{\mu\nu} &= \sum_{\kappa\lambda} \left[D^{LL}_{\kappa\lambda} \left(\mu^X \nu^X \,|\, \kappa^L \lambda^L \right) + D^{SS}_{\kappa\lambda} \left(\mu^X \nu^X \,|\, \kappa^S \lambda^S \right) \right] \\ &= \sum_i \left(\mu^X \nu^X \,|\, ii \right) = \langle \mu^X \,|\, V^{\mathrm{elec}} \,|\, \nu^X \rangle \end{aligned} \tag{11.47}$$

where the index i runs over occupied molecular spinors. This is the matrix element of the electrostatic Coulomb potential of the electrons. The corresponding direct part of the Gaunt interaction is in the off-diagonal blocks. The sum over these terms is the interaction with the current density,

$$\begin{aligned} J^{XY}_{\mu\nu} &= \sum_{\kappa\lambda} \left[D^{XY}_{\kappa\lambda} \left(\mu^X \boldsymbol{\sigma} \nu^Y \,|\, \kappa^X \boldsymbol{\sigma} \lambda^Y \right) + D^{YX}_{\kappa\lambda} \left(\mu^X \boldsymbol{\sigma} \nu^Y \,|\, \kappa^Y \boldsymbol{\sigma} \lambda^X \right) \right] \\ &= \sum_i \left(\mu^X \boldsymbol{\sigma} \nu^Y \,|\, i \,\boldsymbol{\alpha}\, i \right) = \langle \mu^X \,|\boldsymbol{\sigma} \cdot \mathbf{V}^{\mathrm{current}} \,|\, \nu^Y \rangle. \end{aligned} \tag{11.48}$$

Here we must have $X \neq Y$: if $X = L$ we must have $Y = S$ and vice versa. By contrast to the direct contributions, the exchange contributions from both the Coulomb operator and the Gaunt operator are distributed over all four blocks:

$$K_{\mu\nu}^{XX} = \sum_{\kappa\lambda} \Big[D_{\kappa\lambda}^{XX} \left(\mu^X \lambda^X \,|\, \kappa^X \nu^X \right) + D_{\kappa\lambda}^{YY} \left(\mu^X \boldsymbol{\sigma} \lambda^Y \,|\, \kappa^Y \boldsymbol{\sigma} \nu^X \right) \Big], \tag{11.49}$$

$$K_{\mu\nu}^{XY} = \sum_{\kappa\lambda} \Big[D_{\kappa\lambda}^{YX} \left(\mu^X \lambda^X \,|\, \kappa^Y \nu^Y \right) + D_{\kappa\lambda}^{XY} \left(\mu^X \boldsymbol{\sigma} \lambda^Y \,|\, \kappa^X \boldsymbol{\sigma} \nu^Y \right) \Big]. \tag{11.50}$$

11.5 Kramers-Restricted 2-Spinor Matrix Dirac–Hartree–Fock Equations

So far we have not taken time-reversal symmetry into account. From the preceding chapters, we expect that incorporating time-reversal symmetry in a Kramers-restricted Dirac–Hartree–Fock theory will result in a reduction of the work, and possibly also a reduction in the rank of the Fock matrix. The basis set we will use is a basis set of Kramers pairs. We develop the theory for a closed-shell reference, for which all Kramers pairs are doubly occupied.[2]

We partition both the molecular spinors and the basis set into Kramers pairs. This partitioning reduces to a partitioning of the large and small components into Kramers pairs, $\left(\psi_i^X, \psi_{\bar{i}}^X\right)$ for the molecular spinors and $\left(\chi_\mu^X, \chi_{\bar{\mu}}^X\right)$ for the basis functions, where X is L or S. We cannot assume a priori that the molecular spinors for one of the Kramers partners can be expanded solely in the corresponding basis spinors, and we must therefore retain the full expansion until we have deduced the proper restrictions:

$$\begin{aligned} \psi_i^L &= \sum_{\mu=1}^{n_L} \left(a_{\mu i}^L \, \chi_\mu^L + a_{\bar{\mu} i}^L \, \chi_{\bar{\mu}}^L \right), & \psi_i^S &= \sum_{\mu=1}^{n_S} \left(a_{\mu i}^S \, \chi_\mu^S + a_{\bar{\mu} i}^S \, \chi_{\bar{\mu}}^S \right), \\ \psi_{\bar{i}}^L &= \sum_{\mu=1}^{n_L} \left(a_{\mu \bar{i}}^L \, \chi_\mu^L + a_{\bar{\mu} \bar{i}}^L \, \chi_{\bar{\mu}}^L \right), & \psi_{\bar{i}}^S &= \sum_{\mu=1}^{n_S} \left(a_{\mu \bar{i}}^S \, \chi_\mu^S + a_{\bar{\mu} \bar{i}}^S \, \chi_{\bar{\mu}}^S \right). \end{aligned} \tag{11.51}$$

From time-reversal symmetry, $\hat{\mathcal{K}} \psi_i^X = \psi_{\bar{i}}^X$ and $\hat{\mathcal{K}} \psi_{\bar{i}}^X = -\psi_i^X$, and from these relations we derive expressions for the coefficients

$$a_{\mu i}^{X*} = a_{\bar{\mu} \bar{i}}^X, \qquad a_{\bar{\mu} i}^{X*} = -a_{\mu \bar{i}}^X. \tag{11.52}$$

With the partitioning of the basis set, we can partition each block of each matrix—density, one-electron, and Fock—into four blocks. For the density matrix elements

---

[2] Pioneering work for closed-shell DHF theory was done by Malli and Oreg (1975) and by Hafner (1980).

defined in (11.41),

$$
\begin{aligned}
D^{XY}_{\mu\nu} &= \sum_{i=1}^{n} \mathrm{a}^{X*}_{\mu i}\mathrm{a}^{Y}_{\nu i} + \mathrm{a}^{X*}_{\mu\bar{\imath}}\mathrm{a}^{Y}_{\nu\bar{\imath}} & \qquad D^{XY}_{\mu\bar{\nu}} &= \sum_{i=1}^{n} \mathrm{a}^{X*}_{\mu i}\mathrm{a}^{Y}_{\bar{\nu} i} + \mathrm{a}^{X*}_{\mu\bar{\imath}}\mathrm{a}^{Y}_{\bar{\nu}\bar{\imath}} \\
D^{XY}_{\bar{\mu}\bar{\nu}} &= \sum_{i=1}^{n} \mathrm{a}^{X*}_{\bar{\mu} i}\mathrm{a}^{Y}_{\bar{\nu} i} + \mathrm{a}^{X*}_{\bar{\mu}\bar{\imath}}\mathrm{a}^{Y}_{\bar{\nu}\bar{\imath}} & \qquad D^{XY}_{\bar{\mu}\nu} &= \sum_{i=1}^{n} \mathrm{a}^{X*}_{\bar{\mu} i}\mathrm{a}^{Y}_{\nu i} + \mathrm{a}^{X*}_{\bar{\mu}\bar{\imath}}\mathrm{a}^{Y}_{\nu\bar{\imath}}
\end{aligned}
\tag{11.53}
$$

where $n = N/2$ is the number of occupied Kramers pairs. Using the expressions in (11.52), we can relate these matrix elements, $D^{XY}_{\bar{\mu}\bar{\nu}} = D^{XY*}_{\mu\nu}$ and $D^{XY}_{\bar{\mu}\nu} = -D^{XY*}_{\mu\bar{\nu}}$. These are precisely the relations that we would obtain from the application of (9.10). These relations represent a reduction in the work of a factor of 2, since we only need to construct two of the four possible blocks and can obtain the others from time reversal.

Now that we have expressions for the density matrix elements, we can consider the two-electron terms of the Fock matrix. Because the forms of the contributions to the diagonal and off-diagonal blocks of the Fock matrix for the large and small components are different, we must consider them separately. First, we consider the direct term of the Coulomb interaction,

$$
\begin{aligned}
J^{XX}_{\mu\nu} = \sum_{Y,\kappa\lambda}^{n} \Big[& D^{YY}_{\kappa\lambda}\left(\mu^X\nu^X \,|\, \kappa^Y\lambda^Y\right) + D^{YY}_{\kappa\bar{\lambda}}\left(\mu^X\nu^X \,|\, \kappa^Y\bar{\lambda}^Y\right) \\
& + D^{YY}_{\bar{\lambda}\kappa}\left(\mu^X\nu^X \,|\, \bar{\lambda}^Y\kappa^Y\right) + D^{YY}_{\bar{\lambda}\bar{\kappa}}\left(\mu^X\nu^X \,|\, \bar{\lambda}^Y\bar{\kappa}^Y\right)\Big],
\end{aligned}
\tag{11.54}
$$

where we have swapped the indices κ and λ in the last two terms for the purposes of analysis (which we are free to do since they are summation indices). We can use the time-reversal relations for the integrals and the density matrices to reduce this expression. Since $D^{YY}_{\bar{\lambda}\bar{\kappa}} = D^{YY*}_{\lambda\kappa} = D^{YY}_{\kappa\lambda}$ and $\left(\mu^X\nu^X \,|\, \bar{\lambda}^Y\bar{\kappa}^Y\right) = \left(\mu^X\nu^X \,|\, \kappa^Y\lambda^Y\right)$, the first and fourth terms are equal. However, there is no equality between the integrals in the second and third terms, even though the density matrix elements are conjugates: $D^{YY}_{\bar{\lambda}\kappa} = D^{YY*}_{\kappa\bar{\lambda}}$ but $\left(\mu^X\nu^X \,|\, \bar{\lambda}^Y\kappa^Y\right) \neq -\left(\mu^X\nu^X \,|\, \kappa^Y\bar{\lambda}^Y\right)$. Instead, we can restrict the sum in each term to $\kappa > \lambda$ and obtain a reduction, because $D^{YY}_{\kappa\bar{\lambda}} = -D^{YY}_{\lambda\bar{\kappa}}$ and $\left(\mu^X\nu^X \,|\, \kappa^Y\bar{\lambda}^Y\right) = -\left(\mu^X\nu^X \,|\, \lambda^Y\bar{\kappa}^Y\right)$. With these reductions, the application of time-reversal symmetry for the direct terms leads to

$$
\begin{aligned}
J^{XX}_{\mu\nu} = {} & 2\sum_{Y,\kappa\lambda} D^{YY}_{\kappa\lambda}\left(\mu^X\nu^X \,|\, \kappa^Y\lambda^Y\right) \\
& + 2\sum_{Y,\kappa>\lambda}\Big[D^{YY}_{\kappa\bar{\lambda}}\left(\mu^X\nu^X \,|\, \kappa^Y\bar{\lambda}^Y\right) + D^{YY}_{\bar{\kappa}\lambda}\left(\mu^X\nu^X \,|\, \bar{\kappa}^Y\lambda^Y\right)\Big].
\end{aligned}
\tag{11.55}
$$

The work involved in constructing this part of the Fock matrix is therefore half the work involved in constructing the corresponding part of the unrestricted Fock matrix. This reduction parallels the reduction from spin-unrestricted to spin-restricted Fock matrices in nonrelativistic theory.

Next, we consider the direct term from the Gaunt interaction:

$$\begin{aligned} J_{\mu\nu}^{XY} = \sum_{\kappa\lambda}\Big[& D_{\kappa\lambda}^{XY}\left(\mu^X\boldsymbol{\sigma}\nu^Y\,|\,\kappa^X\boldsymbol{\sigma}\lambda^Y\right) + D_{\kappa\bar{\lambda}}^{XY}\left(\mu^X\boldsymbol{\sigma}\nu^Y\,|\,\kappa^X\boldsymbol{\sigma}\bar{\lambda}^Y\right) \\ & + D_{\bar{\kappa}\lambda}^{XY}\left(\mu^X\boldsymbol{\sigma}\nu^Y\,|\,\bar{\kappa}^X\boldsymbol{\sigma}\lambda^Y\right) + D_{\bar{\kappa}\bar{\lambda}}^{XY}\left(\mu^X\boldsymbol{\sigma}\nu^Y\,|\,\bar{\kappa}^X\boldsymbol{\sigma}\bar{\lambda}^Y\right) \\ & + D_{\lambda\kappa}^{YX}\left(\mu^X\boldsymbol{\sigma}\nu^Y\,|\,\lambda^Y\boldsymbol{\sigma}\kappa^X\right) + D_{\lambda\bar{\kappa}}^{YX}\left(\mu^X\boldsymbol{\sigma}\nu^Y\,|\,\lambda^Y\boldsymbol{\sigma}\bar{\kappa}^X\right) \\ & + D_{\bar{\lambda}\kappa}^{YX}\left(\mu^X\boldsymbol{\sigma}\nu^Y\,|\,\bar{\lambda}^Y\boldsymbol{\sigma}\kappa^X\right) + D_{\bar{\lambda}\bar{\kappa}}^{YX}\left(\mu^X\boldsymbol{\sigma}\nu^Y\,|\,\bar{\lambda}^Y\boldsymbol{\sigma}\bar{\kappa}^X\right)\Big]. \end{aligned} \tag{11.56}$$

Applying the time-reversal operator to the current densities, we derive the relations

$$(\kappa^X\boldsymbol{\sigma}\lambda^Y) = -(\bar{\lambda}^Y\boldsymbol{\sigma}\bar{\kappa}^X), \qquad (\kappa^X\boldsymbol{\sigma}\bar{\lambda}^Y) = (\lambda^Y\boldsymbol{\sigma}\bar{\kappa}^X), \tag{11.57}$$

which we can use in the integrals. Making use of these relations and the relations for the densities, we find that all the integrals cancel: the first and eighth and the fourth and fifth terms cancel due to the first relation, and the second and sixth and the third and seventh terms cancel due to the second. The result is that the direct Gaunt contribution to the Fock matrix vanishes. We can see this by applying the first relation to the molecular density: $(i\boldsymbol{\alpha}i) = -(\bar{\imath}\boldsymbol{\alpha}\bar{\imath})$ and we are adding the two terms and summing over i. It makes sense that these terms cancel because time-reversing the current density should change its sign. The cancellation also exists in the unrestricted case, but it is a numerical cancellation, and it is incomplete because the molecular spinors do not form precisely into Kramers pairs. In the Kramers-restricted case, a single electron outside a closed shell also has no direct Gaunt interaction: it requires two unpaired electrons before this term is nonzero.

The exchange terms do not reduce so easily. Considering first the Coulomb interaction terms for the blocks that are diagonal in the components,

$$\begin{aligned} K_{\mu\nu}^{\mathrm{C},XX} = \sum_{\kappa\lambda}\Big[& D_{\kappa\lambda}^{XX}\left(\mu^X\lambda^X\,|\,\kappa^X\nu^X\right) + D_{\kappa\bar{\lambda}}^{XX}\left(\mu^X\bar{\lambda}^X\,|\,\kappa^X\nu^X\right) \\ & + D_{\bar{\kappa}\lambda}^{XX}\left(\mu^X\lambda^X\,|\,\bar{\kappa}^X\nu^X\right) + D_{\bar{\kappa}\bar{\lambda}}^{XX}\left(\mu^X\bar{\lambda}^X\,|\,\bar{\kappa}^X\nu^X\right)\Big], \end{aligned} \tag{11.58}$$

we see that although we can combine or triangularize the density matrices, there are no relations between the integrals that would give us a reduction in the number of terms. Making the combinations, we get

$$\begin{aligned} K_{\mu\nu}^{\mathrm{C},XX} = & \sum_{\kappa\lambda} D_{\kappa\lambda}^{XX}\Big[\left(\mu^X\lambda^X\,|\,\kappa^X\nu^X\right) + \left(\mu^X\bar{\kappa}^X\,|\,\bar{\lambda}^X\nu^X\right)\Big] \\ & + \sum_{\kappa>\lambda} + D_{\kappa\bar{\lambda}}^{XX}\Big[\left(\mu^X\bar{\lambda}^X\,|\,\kappa^X\nu^X\right) - \left(\mu^X\bar{\kappa}^X\,|\,\lambda^X\nu^X\right)\Big] \\ & + \sum_{\kappa>\lambda} + D_{\bar{\kappa}\lambda}^{XX}\Big[\left(\mu^X\lambda^X\,|\,\bar{\kappa}^X\nu^X\right) - \left(\mu^X\kappa^X\,|\,\bar{\lambda}^X\nu^X\right)\Big]. \end{aligned} \tag{11.59}$$

The reduction in work is not a factor of two, though there is a reduction because the number of multiplications has halved. However, in a conventional disk-based SCF algorithm, the integral sums could be precomputed so that a reduction of a factor of two can be achieved in the construction of the Fock matrix.

The same kinds of reductions occur for the exchange terms from the Gaunt interaction,

$$\begin{aligned}
K_{\mu\nu}^{\mathrm{G},XX} &= \sum_{\kappa\lambda} D_{\kappa\lambda}^{YY}\Big[\big(\mu^X\boldsymbol{\sigma}\lambda^Y\,|\,\kappa^Y\boldsymbol{\sigma}\nu^X\big) + \big(\mu^X\boldsymbol{\sigma}\bar{\kappa}^Y\,|\,\bar{\lambda}^Y\boldsymbol{\sigma}\nu^X\big)\Big] \\
&\quad + \sum_{\kappa>\lambda} +D_{\kappa\bar{\lambda}}^{YY}\Big[\big(\mu^X\boldsymbol{\sigma}\bar{\lambda}^Y\,|\,\kappa^Y\boldsymbol{\sigma}\nu^X\big) - \big(\mu^X\boldsymbol{\sigma}\bar{\kappa}^Y\,|\,\lambda^Y\boldsymbol{\sigma}\nu^X\big)\Big] \\
&\quad + \sum_{\kappa>\lambda} +D_{\bar{\kappa}\lambda}^{YY}\Big[\big(\mu^X\boldsymbol{\sigma}\lambda^Y\,|\,\bar{\kappa}^Y\boldsymbol{\sigma}\nu^X\big) - \big(\mu^X\boldsymbol{\sigma}\kappa^Y\,|\,\bar{\lambda}^Y\boldsymbol{\sigma}\nu^X\big)\Big],
\end{aligned} \tag{11.60}$$

and also for the exchange terms in the large-small and small-large blocks,

$$\begin{aligned}
K_{\mu\nu}^{XY} &= \sum_{\kappa\lambda} D_{\kappa\lambda}^{YX}\Big[\big(\mu^X\lambda^X\,|\,\kappa^Y\nu^Y\big) + \big(\mu^X\boldsymbol{\sigma}\bar{\kappa}^Y\,|\,\bar{\lambda}^X\boldsymbol{\sigma}\nu^Y\big)\Big] \\
&\quad + D_{\lambda\kappa}^{XY}\Big[\big(\mu^X\bar{\lambda}^X\,|\,\bar{\kappa}^Y\nu^Y\big) + \big(\mu^X\boldsymbol{\sigma}\kappa^Y\,|\,\lambda^X\boldsymbol{\sigma}\nu^Y\big)\Big] \\
&\quad + D_{\kappa\bar{\lambda}}^{YX}\Big[\big(\mu^X\bar{\lambda}^X\,|\,\kappa^Y\nu^Y\big) - \big(\mu^X\boldsymbol{\sigma}\kappa^Y\,|\,\bar{\lambda}^X\boldsymbol{\sigma}\nu^Y\big)\Big] \\
&\quad - D_{\lambda\bar{\kappa}}^{XY}\Big[\big(\mu^X\lambda^X\,|\,\bar{\kappa}^Y\nu^Y\big) - \big(\mu^X\boldsymbol{\sigma}\bar{\kappa}^Y\,|\,\lambda^X\boldsymbol{\sigma}\nu^Y\big)\Big].
\end{aligned} \tag{11.61}$$

11.6 Symmetry in the Kramers-Restricted Fock Matrix

The symmetry reductions in nonrelativistic methodology come from spin symmetry and from point-group symmetry. In relativistic methodology, time-reversal symmetry is the equivalent of spin symmetry, but it does not provide the same magnitude of reduction as does spin symmetry. This is due to the presence of spin-dependent terms in the Fock operator. Point-group symmetry is intimately connected with time-reversal symmetry in Kramers-restricted relativistic theory, as we saw in chapter 10.

Point-group symmetry is implemented in nonrelativistic SCF programs in two ways. One way is to construct symmetry-adapted basis functions. In this approach, symmetry is built in to the integrals and hence into the Fock matrix construction (Davidson 1975b, Helgaker and Taylor 1995). The other way, due to Dacre (1970) and Elder (1973), is to construct a partial or "skeleton" Fock matrix from the symmetry-unique integrals over atomic basis functions (the "petite list", Dupuis and King 1977), and symmetrize the Fock matrix afterwards. From experience, there are usually more symmetry-adapted integrals than symmetry-unique atomic integrals. In theory the two should be the same; in practice the thresholds for integral evaluation eliminate many integrals in the petite list. Codes that use both techniques are in existence, but the Dacre and Elder method is probably the more popular.

When it comes to relativistic SCF methods, there are some subtleties in the symmetry reduction. To illustrate this point, we consider first the double group C_2^*. This group has two fermion irreps that are singly degenerate and are complex conjugates. The atomic basis 2-spinors that are on the symmetry axis are already symmetry functions and can be divided between these two irreps, but the atomic basis 2-spinors that are not on the axis are linear combinations of symmetry functions and cannot be classified into the irreps. If we construct symmetry 2-spinors from the atomic 2-spinors and work in a symmetry-adapted basis set, the symmetry reduction is straightforward: matrix elements with an odd number of "barred" functions vanish. As a consequence, the Fock matrix is block-diagonal, and the number of two-electron integrals that must be accumulated into the Fock matrix is halved. These two effects yield a factor of 4 reduction in the work required to construct the Fock matrix. The molecular spinor expansion now only spans one row of the rep for each Kramers partner. The symmetry reduction in the matrices is exactly a factor of 2 because all symmetry-adapted 2-spinors come in Kramers pairs.

In the Dacre and Elder method, the reduction in the integral number comes from the relations between the symmetry-equivalent atoms. Consider two symmetry-equivalent atoms, A and B, each of which has two $s_{1/2}$ 2-spinor basis functions, χ_κ and χ_λ, which we label κ_A and κ_B, and λ_A and λ_B. The relations between the one-electron potential energy integrals (for example) that do not follow from Hermitian conjugation are

$$V_{\kappa_A\kappa_A} = V_{\kappa_B\kappa_B}, \quad V_{\kappa_A\lambda_A} = V_{\kappa_B\lambda_B}, \quad V_{\lambda_A\lambda_A} = V_{\lambda_B\lambda_B}, \quad V_{\kappa_A\lambda_B} = V_{\kappa_B\lambda_A}. \tag{11.62}$$

Similar relations can be written for the "odd-bar" integrals. Relations like these can be used to reduce the number of unique integrals by a factor of 2 and the number of density matrix elements by a factor of 2. However, the odd-bar integrals are not necessarily zero because they are not integrals between symmetry functions, and all the terms in the expressions for the Fock matrix elements remain. The reduction in work comes from the fact that the sums are restricted to sums over symmetry-unique basis functions, and that only the symmetry-unique parts of the Fock matrix are constructed. For matrix elements between functions on the symmetry axis, the reductions follow from the fact that they are symmetry functions, and hence an overall factor of 4 reduction in the work for the two-electron integrals is obtained.

If we introduce a second twofold axis to generate the point group D_2^*, the symmetry reductions for the Dacre and Elder method follow exactly the same lines as for C_2^*. No difference in principle arises from the higher order of the group. This is not the case for the symmetry-adapted method. In D_2^* there is only one doubly degenerate fermion irrep, and the symmetry-adapted 2-spinors divide equally between the rows of the irrep. The expansion of the molecular spinors in the symmetry-adapted 2-spinor basis is restricted to a single row of the irrep, and the odd-bar matrix elements are zero, yielding a factor of 2 reduction in the number of matrix elements, just as for C_2^*. The order of D_2^* is 8, so there should be an extra factor of 2 reduction. This extra factor of 2 results from the fact that the matrix elements and integrals are real, rather than complex. The work is reduced by a factor of 4 relative to C_2^* because the multiplication of two complex numbers is four times the work of the multiplication of two real numbers.

11.7 Kramers-Restricted Open-Shell Methods

In nonrelativistic theory, it is possible to write a single-determinant wave function for closed-shell molecules and for high-spin open-shell molecules. These wave functions form the basis for the restricted Hartree–Fock (RHF) method and the (spin-)unrestricted Hartree–Fock (UHF) methods. The general approach to open-shell RHF optimizes the energy functional as an average over the energies of all the degenerate components of the open-shell state in question (Roothaan and Bagus 1963, Manne 1972).[3] Both restricted and unrestricted Hartree–Fock methods make use of the spin-independence of the Hamiltonian to factorize the Fock matrix.

Single-configuration wave functions are important because they can be used to provide a basis for correlation methods that use a single configuration as a reference function, such as the various open-shell perturbation methods and coupled-cluster methods. The preceding sections have developed the unrestricted Dirac–Hartree–Fock (UDHF) method for a single determinant and the Kramers-restricted Dirac–Hartree–Fock (KRDHF) method for a closed-shell system. The application of time-reversal symmetry does not necessarily result in a block-diagonalization of the Fock matrix as does the application of spin symmetry. It is therefore pertinent to ask the question "Is there an equivalent to the high-spin open-shell Hartree–Fock method in relativistic Kramers-restricted theory?"

The answer to this question is a qualified "yes", and depends on the nature of the wave function. For a single electron outside a closed shell, the Kramers-restricted wave function is a single determinant, and it is relatively easy to define the Kramers-restricted Fock matrices for the closed-shell and open-shell electrons. Using the expressions for the Fock matrix in the molecular basis, (11.34), with indices i and j for doubly occupied Kramers pairs, t for the singly occupied Kramers pair, and a for the empty or virtual Kramers pairs, we can define the three nonredundant sections of the Fock matrix as follows:

$$\begin{aligned}
f_{ia} &= h_{ia} + \sum_j \big[2(ia|jj) - (ij|ja) - (i\bar{j}|\bar{j}a)\big] + \tfrac{1}{2}\big[2(ia|tt) - (it|ta) - (i\bar{t}|\bar{t}a)\big], \\
f_{it} &= h_{it} + \sum_j \big[2(it|jj) - (ij|jt) - (i\bar{j}|\bar{j}t)\big] + (it|tt), \qquad (11.63) \\
f_{ta} &= h_{ta} + \sum_j \big[2(ta|jj) - (tj|ja) - (t\bar{j}|\bar{j}a)\big].
\end{aligned}$$

These expressions are actually derived from the average energy of the two degenerate states. If we were to use only one of the states we would arrive at different expressions for the barred and unbarred spinors. The same is true in nonrelativistic theory. The expressions have also been divided by the relevant occupation number, so that

---

[3] Technically, it is possible to define methods in which other restrictions than those on spin are lifted, but the most common unrestricted method is the spin-unrestricted method, and we follow the common usage of UHF for the spin-unrestricted Hartree–Fock method.

the coefficient of the one-electron integral is unity. Similar expressions can easily be written down for the odd-bar matrix elements.

For molecules with more than a single electron outside a closed shell, the ability to define a single-configuration wave function depends on the molecular symmetry and the symmetry of the spinors. The equivalent of the nonrelativistic high-spin open-shell wave function for two open shells (ignoring the closed shells for the present) can be represented by the configuration $\psi_t\psi_u$. The time-reversed configuration, $\psi_{\bar{t}}\psi_{\bar{u}}$, has the same energy, and the matrix element between the two configurations, $(t\bar{u}|t\bar{u})$, is not in general zero, in contrast to the nonrelativistic case. We can form fixed linear combinations that are symmetric under time reversal,

$$\Psi_\pm = e^{i\eta_\pm}\left(\psi_t\psi_u \pm \psi_{\bar{t}}\psi_{\bar{u}}\right), \tag{11.64}$$

but the matrix element of the many-electron Hamiltonian between them is not necessarily zero—it is in fact the imaginary part of the matrix element between the two configurations. To make matters more complicated, the other two configurations that can be formed from the Kramers pairs t and u, $\psi_t\psi_{\bar{u}}$ and $\psi_{\bar{t}}\psi_u$, also have nonzero matrix elements with the first two configurations, in the absence of symmetry.

Nevertheless, let us analyze the Fock matrices for the wave functions based on the linear combinations. The closed–empty, closed–open and open–empty Fock matrices are

$$\begin{aligned}
f_{ia} &= h_{ia} + \sum_j \big[2(ia|jj) - (ij|ja) - (i\bar{j}|\bar{j}a)\big] + \tfrac{1}{2}\sum_v \big[2(ia|vv) - (iv|va) - (i\bar{v}|\bar{v}a)\big],\\
f_{iw} &= h_{iw} + \sum_j \big[2(iw|jj) - (ij|jw) - (i\bar{j}|\bar{j}w)\big] + \sum_v \big[(iw|vv) - (iv|vw) \mp (i\bar{v}|v\bar{w})\big],\\
f_{wa} &= h_{wa} + \sum_j \big[2(wa|jj) - (wj|ja) - (w\bar{j}|\bar{j}a)\big]\\
&\quad + \sum_v \big[(wa|vv) - (wv|va) \mp (\bar{w}v|\bar{v}a)\big].
\end{aligned} \tag{11.65}$$

In these expressions we have added canceling terms such as $(it|tt)-(it|tt)$ and $(i\bar{u}|\bar{u}u)$ to make it clear that the Fock matrices for the open-shell electrons are equivalent. The sum over v covers the two values t and u, and w can be either t or u. We must also consider the open–open Fock matrices. The off-diagonal matrix element has no contribution from interaction between the open-shell electrons,

$$f_{tu} = h_{tu} + \sum_j \big[2(tu|jj) - (tj|ju) - (t\bar{j}|\bar{j}u)\big], \tag{11.66}$$

but the diagonal term (the orbital energy) does,

$$\begin{aligned}
f_{ww} &= h_{ww} + \sum_j \big[2(ww|jj) - (wj|jw) - (w\bar{j}|\bar{j}w)\big]\\
&\quad + \sum_v \big[(ww|vv) - (wv|vw) \mp (w\bar{v}|v\bar{w})\big].
\end{aligned} \tag{11.67}$$

We would expect the open–open Fock matrix elements to correspond to redundant spinor rotation parameters because the energy should be invariant to rotations of equivalent spinors. However, if we construct the spinor rotation gradient, we see that there is a term that survives in the diagonal of the gradient:

$$g_{ww} = \pm\tfrac{1}{2}\left[(w\bar{v}|v\bar{w}) - (\bar{v}w|\bar{w}v)\right] = \pm i\,\mathrm{Im}(w\bar{v}|v\bar{w}). \tag{11.68}$$

The imaginary part of the integral is not necessarily zero, so the diagonal spinor rotation gradient is not zero. To explain this rather peculiar situation, we must turn to the application of symmetry.

If both components of both open-shell Kramers pairs are in the same nondegenerate irrep (the quaternion case), as is the case when there is no symmetry, all four open-shell determinants must be included in the wave function, and we must resort to a genuine multiconfiguration DHF method.

If the Kramers partners are in different singly degenerate irreps, as for example in the group C_2^*, the configuration space reduces to the two configurations given above. Both configurations belong to the same boson irrep. However, the imaginary part of the integral $(w\bar{v}|v\bar{w})$ is not zero, and for the same reason, the Hamiltonian matrix elements between the two wave functions is not zero. We must still use a multiconfiguration DHF method, albeit in a reduced space of two configurations. The orbital rotation gradient becomes zero only by inclusion of the terms from the configuration interaction problem.

If the Kramers partners are in different rows of the same, doubly degenerate irrep, as in the group C_{2v}^*, then the matrix elements are real and the gradient is zero. Likewise, the Hamiltonian matrix element between the two wave functions is zero, because the two configurations span two boson irreps and the $+$ and $-$ linear combinations have specific boson symmetries. This situation is somewhat similar to the nonrelativistic case of Σ^+ and Σ^- states, which are linear combinations of two determinants, or to the open-shell singlet case.

In groups with rotation axes of higher order, such as C_{4v}^*, the two configurations can belong to different rows of a degenerate boson irrep and do not need to be combined to form symmetry-adapted wave functions. In this case, the integral $(t\bar{u}|t\bar{u})$ is zero, and the Fock matrices reduce to the analogy of the high-spin open-shell nonrelativistic case,

$$\begin{aligned}
f_{ia} &= h_{ia} + \sum_j \left[2(ia|jj) - (ij|ja) - (i\bar{j}|\bar{j}a)\right] \\
&\quad + \tfrac{1}{2}\sum_v \left[2(ia|vv) - (iv|va) - (i\bar{v}|\bar{v}a)\right], \\
f_{iw} &= h_{iw} + \sum_j \left[2(iw|jj) - (ij|jw) - (i\bar{j}|\bar{j}w)\right] + \sum_v \left[(iw|vv) - (iv|vw)\right], \\
f_{wa} &= h_{wa} + \sum_j \left[2(wa|jj) - (wj|ja) - (w\bar{j}|\bar{j}a)\right] + \sum_v \left[(wa|vv) - (wv|va)\right].
\end{aligned} \tag{11.69}$$

As an example, in the group C_{4v}^*, the configuration $(1e_{1/2})^1(2e_{1/2})^1$ has the symmetry E_1, and the time-reversed configuration $(1e_{-1/2})^1(2e_{-1/2})^1$ has the symmetry of the other component of the E boson irrep, E_{-1}. The configuration $(1e_{1/2})^1(1e_{3/2})^1$, however, spans the B_1 and B_2 representations along with $(1e_{-1/2})^1(1e_{-3/2})^1$, and the two configurations must be combined to form symmetry-adapted wave functions.

In this analysis, we only considered the configurations $\psi_t\psi_u$ and $\psi_{\bar{t}}\psi_{\bar{u}}$. A similar analysis could be done with the other pair of configurations, $\psi_t\psi_{\bar{u}}$ and $\psi_{\bar{t}}\psi_u$. However, if ψ_t and ψ_u belong to the same irrep, one of the linear combinations belongs to the totally symmetric irrep, and a multiconfiguration SCF (MCSCF) treatment is mandatory.

The analysis can be extended to more open shells. For three open shells, a configuration and its Kramers partner can belong to the same fermion irrep (as in the case of no symmetry) or to different singly degenerate irreps, or to different rows of the same doubly degenerate irrep. If there are no other configurations that belong to the same irrep or the same row of the same irrep, the analogy of the high-spin open-shell method can be used. For example, in C_{2v}^* the configurations $(1e_{1/2})^1(2e_{1/2})^1(3e_{1/2})^1$ and $(1e_{-1/2})^1(2e_{-1/2})^1(3e_{1/2})^1$ belong to the same row of the fermion irrep, $E_{1/2}$, along with the two other configurations that have two $e_{-1/2}$ spinors occupied, and therefore an MCSCF method must be used. However, in C_{4v}^* the $(1e_{1/2})^1(2e_{1/2})^1(3e_{1/2})^1$ configuration and its Kramers partner are the only configurations from these three Kramers pairs that belong to the $E_{3/2}$ irrep. The remaining configurations built from these three Kramers pairs belong to the $E_{1/2}$ irrep. We can therefore use a single configuration method for the $(1e_{1/2})^1(2e_{1/2})^1(3e_{1/2})^1$ configuration, but we must use a multiconfiguration method for the rest.

The discussion in this section has focused on the behavior of the various open-shell approaches under the symmetry constraints imposed by the double groups. In addition, it should be noted that the degeneracies due to symmetry in the fermion irreps of the double groups are generally lower than those due to spin and boson symmetry in a nonrelativistic calculation. Hence, the need for open-shell calculations to handle real symmetry degeneracies is in general smaller than in the nonrelativistic case.

The conclusion we can draw from the discussion is that many of the systems that would be handled by open-shell RHF calculations in the nonrelativistic case do not really lend themselves to this treatment relativistically. An illustration of this point for the group 14 elements is given in section 12.1. Instead, we must resort to MCSCF methods, which we describe in section 12.9.

11.8 Expansion in Scalar Basis Sets

We have seen that the DHF equations are conveniently analyzed and manipulated in a 2-spinor form. The practical solution of these equations in a finite basis requires the calculation and handling of a large number of integrals. Some effort has been invested in calculating the integrals directly in a 2-spinor basis (Grant and Quiney 2002, Yanai et al. 2002). For nonrelativistic calculations, a substantial fraction of the developmental effort has gone into devising efficient algorithms and techniques for integral-related tasks. If we want to take advantage of these nonrelativistic developments directly, and adapt existing integral codes to our purposes, that is, relativistic molecular calculations, there are two possible roads to follow—calculate integrals in a (real) scalar basis and transform them to 2-spinor integrals, or reformulate the equations in a scalar basis expansion where these integrals can be used directly. In the present section, we will describe the latter approach, leading ultimately to formulation of the DHF equations in a scalar basis set.

We start by considering the one-particle Dirac equation and proceed in a manner analogous to the free-particle solutions of chapter 7. The equation to solve is

$$\left[c\boldsymbol{\alpha}\cdot\mathbf{p} + V - c^2(\boldsymbol{\beta} - \mathbf{I}_4)\right]\psi = E\psi. \tag{11.70}$$

The kinetic energy operator may be expanded in the form

$$c\boldsymbol{\alpha}\cdot\mathbf{p} = c\begin{pmatrix} 0 & 0 & \hat{\Pi}_z & \hat{\Pi}_- \\ 0 & 0 & \hat{\Pi}_+ & -\hat{\Pi}_z \\ \hat{\Pi}_z & \hat{\Pi}_- & 0 & 0 \\ \hat{\Pi}_+ & -\hat{\Pi}_z & 0 & 0 \end{pmatrix} \tag{11.71}$$

where

$$\hat{\Pi}_z = -i\hbar\frac{\partial}{\partial z}, \qquad \hat{\Pi}_\pm = -i\hbar\left(\frac{\partial}{\partial x} \pm i\frac{\partial}{\partial y}\right). \tag{11.72}$$

We write the 4-spinor ψ in the form

$$\psi = \begin{pmatrix} \psi^{L\alpha} \\ \psi^{L\beta} \\ \psi^{S\alpha} \\ \psi^{S\beta} \end{pmatrix}, \tag{11.73}$$

with the $\psi^{X\tau}$ scalar functions that are complex and unnormalized. These functions are expanded in a set of real, normalized, scalar, spatial basis functions χ_μ^X,

$$\psi^{X\tau} = \sum_{\mu=1}^{N_X} \mathrm{a}_\mu^{X\tau}\chi_\mu^X, \tag{11.74}$$

where X is L or S. We use the same basis of spatial functions χ_μ^X for both spin components, but the large- and small-component functions might be different, and there might also be different numbers of large- and small-component basis functions. Inserting the finite-basis expansion into the Dirac equation with the above representation of the kinetic energy operator, we get the matrix Dirac equation in a scalar basis,

$$\begin{pmatrix} \mathbf{V}^{LL} & \mathbf{0} & c\boldsymbol{\Pi}_z^{LS} & c\boldsymbol{\Pi}_-^{LS} \\ \mathbf{0} & \mathbf{V}^{LL} & c\boldsymbol{\Pi}_+^{LS} & -c\boldsymbol{\Pi}_z^{LS} \\ c\boldsymbol{\Pi}_z^{SL} & c\boldsymbol{\Pi}_-^{SL} & \mathbf{V}^{SS} - 2mc^2\mathbf{S}^{SS} & \mathbf{0} \\ c\boldsymbol{\Pi}_+^{SL} & -c\boldsymbol{\Pi}_z^{SL} & \mathbf{0} & \mathbf{V}^{SS} - 2mc^2\mathbf{S}^{SS} \end{pmatrix} \begin{pmatrix} \mathbf{a}^{L\alpha} \\ \mathbf{a}^{L\beta} \\ \mathbf{a}^{S\alpha} \\ \mathbf{a}^{S\beta} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{S}^{LL} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{LL} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}^{SS} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{S}^{SS} \end{pmatrix} \begin{pmatrix} \mathbf{a}^{L\alpha} \\ \mathbf{a}^{L\beta} \\ \mathbf{a}^{S\alpha} \\ \mathbf{a}^{S\beta} \end{pmatrix} E. \tag{11.75}$$

The matrix elements are defined by

$$V_{\mu\nu}^{XX} = \langle \chi_\mu^X | V | \chi_\nu^X \rangle, \qquad S_{\mu\nu}^{XX} = \langle \chi_\mu^X | \chi_\nu^X \rangle,$$

$$\left(\Pi_z^{XY}\right)_{\mu\nu} = \langle \chi_\mu^X | \hat{\Pi}_z | \chi_\nu^Y \rangle, \quad \left(\Pi_\pm^{XY}\right)_{\mu\nu} = \langle \chi_\mu^X | \hat{\Pi}_\pm | \chi_\nu^Y \rangle. \tag{11.76}$$

The kinetic energy matrices obey the relation $\boldsymbol{\Pi}_\pm^{LS} = \left(\boldsymbol{\Pi}_\mp^{SL}\right)^\dagger$, so, despite appearances, the Dirac matrix in the scalar basis given above is Hermitian.

In these equations, the ordering of the components is the conventional order—large components first, then small components, and within each component, α spin then β spin. However, we could equally well have ordered by spin—α spin components first, and then β spin components, and for each spin, large component then small component. This ordering yields the matrix

$$\begin{pmatrix} \mathbf{V}^{LL} & c\boldsymbol{\Pi}_0^{LS} & \mathbf{0} & c\boldsymbol{\Pi}_-^{LS} \\ c\boldsymbol{\Pi}_0^{SL} & \mathbf{V}^{SS} - 2mc^2\mathbf{S}^{SS} & c\boldsymbol{\Pi}_-^{SL} & \mathbf{0} \\ \mathbf{0} & c\boldsymbol{\Pi}_+^{LS} & \mathbf{V}^{LL} & -c\boldsymbol{\Pi}_0^{LS} \\ c\boldsymbol{\Pi}_+^{SL} & \mathbf{0} & -c\boldsymbol{\Pi}_0^{SL} & \mathbf{V}^{SS} - 2mc^2\mathbf{S}^{SS} \end{pmatrix}. \tag{11.77}$$

This form of the matrix equation displays the structure of the matrix representation of an operator that is symmetric under time reversal, given in (10.30)

$$\mathbf{f} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^* & \mathbf{A}^* \end{pmatrix}. \tag{11.78}$$

Based on this ordering, it is possible to rewrite the equations as a quaternion problem, which turns out to provide considerable computational advantages, in particular in the handling of symmetry (Hafner 1980, Saue and Jensen 1999). For the present purposes we will stay with the usual ordering by component size, which is more convenient for displaying the structure of the equations, and also has the advantage of familiarity.

The DHF equations in the scalar basis are developed by adding the electron–electron interaction of the mean field, given for the molecular 4-spinors in (11.34). To derive the contributions to the scalar Fock matrix from the Coulomb interaction, we expand the integrals in the scalar basis,

$$(pq\,|\,rs) = \sum_{XY\tau\upsilon}\sum_{\mu\nu\kappa\lambda}\big(\mathrm{a}_{\mu p}^{X\tau}\big)^{*}\mathrm{a}_{\nu q}^{X\tau}\big(\mathrm{a}_{\kappa r}^{Y\upsilon}\big)^{*}\mathrm{a}_{\lambda s}^{X\upsilon}\big(\mu^{X}\nu^{X}|\kappa^{Y}\lambda^{Y}\big), \tag{11.79}$$

and introduce the appropriate scalar density matrices

$$D_{\kappa\lambda}^{X\tau Y\upsilon} = \sum_{i}\big(\mathrm{a}_{\kappa i}^{X\tau}\big)^{*}\mathrm{a}_{\lambda i}^{Y\upsilon}. \tag{11.80}$$

With these definitions, the component-diagonal blocks of the Fock matrix in the scalar basis are

$$f_{\mu\nu}^{L\tau L\tau} = V_{\mu\nu}^{LL} + \sum_{\kappa\lambda}\sum_{X\upsilon} D_{\kappa\lambda}^{X\upsilon X\upsilon}\big(\mu^{L}\nu^{L}\,|\,\kappa^{X}\lambda^{X}\big) - \sum_{\kappa\lambda} D_{\kappa\lambda}^{L\tau L\tau}\big(\mu^{L}\lambda^{L}\,|\,\kappa^{L}\nu^{L}\big), \tag{11.81}$$

$$f_{\mu\nu}^{X\tau X\upsilon} = -\sum_{\kappa\lambda} D_{\kappa\lambda}^{X\upsilon X\tau}\big(\mu^{X}\lambda^{X}\,|\,\kappa^{X}\nu^{X}\big), \tag{11.82}$$

$$f_{\mu\nu}^{S\tau S\tau} = V_{\mu\nu}^{SS} - 2mc^{2}S_{\mu\nu}^{SS} + \sum_{\kappa\lambda}\sum_{X\upsilon} D_{\kappa\lambda}^{X\upsilon X\upsilon}\big(\mu^{S}\nu^{S}|\kappa^{X}\lambda^{X}\big) - \sum_{\kappa\lambda} D_{\kappa\lambda}^{S\tau S\tau}\big(\mu^{S}\lambda^{S}|\kappa^{S}\nu^{S}\big). \tag{11.83}$$

Likewise, the off-diagonal blocks are

$$f_{\mu\nu}^{X\alpha Y\alpha} = c\big(\Pi_{z}^{XY}\big)_{\mu\nu} - \sum_{\kappa\lambda} D_{\kappa\lambda}^{Y\alpha X\alpha}\big(\mu^{X}\lambda^{X}\,|\,\kappa^{Y}\nu^{Y}\big), \tag{11.84}$$

$$f_{\mu\nu}^{X\beta Y\beta} = -c\big(\Pi_{z}^{XY}\big)_{\mu\nu} - \sum_{\kappa\lambda} D_{\kappa\lambda}^{Y\beta X\beta}\big(\mu^{X}\lambda^{X}\,|\,\kappa^{Y}\nu^{Y}\big), \tag{11.85}$$

$$f_{\mu\nu}^{X\alpha Y\beta} = c\big(\Pi_{+}^{XY}\big)_{\mu\nu} - \sum_{\kappa\lambda} D_{\kappa\lambda}^{Y\beta X\alpha}\big(\mu^{X}\lambda^{X}\,|\,\kappa^{Y}\nu^{Y}\big), \tag{11.86}$$

$$f_{\mu\nu}^{X\beta Y\alpha} = c\big(\Pi_{-}^{XY}\big)_{\mu\nu} - \sum_{\kappa\lambda} D_{\kappa\lambda}^{Y\alpha X\beta}\big(\mu^{X}\lambda^{X}\,|\,\kappa^{Y}\nu^{Y}\big). \tag{11.87}$$

The interaction with the Coulomb potential of the electrons appears with the nuclear potential in the blocks that are diagonal in both the spin and the components, and the exchange contributions are spread between all blocks.

Defining the reduced density matrix as

$$D_{\kappa\lambda}^{XX} = D_{\kappa\lambda}^{X\alpha X\alpha} + D_{\kappa\lambda}^{X\beta X\beta} \tag{11.88}$$

and making use of the fact that the integrals are real, and therefore have the full eightfold permutational symmetry, the large-component diagonal block can be written

$$f_{\mu\nu}^{L\tau L\tau} = V_{\mu\nu}^{LL} + \sum_{\kappa\lambda}\sum_{X} D_{\kappa\lambda}^{XX}\left(\mu^L \nu^L \middle| \kappa^X \lambda^X \right) - \sum_{\kappa\lambda} D_{\kappa\lambda}^{L\tau L\tau}\left(\mu^L \lambda^L \middle| k^L \nu^L \right), \quad (11.89)$$

The analogous expression with the rest mass term can be written down for the small-component diagonal block.

Including the Breit term for the electron–electron interaction in a scalar basis requires extensive additions to a Dirac–Hartree–Fock–Coulomb scheme. It is not possible to achieve the same reductions as for the Coulomb term, and the derivation of the Fock matrix contributions requires considerable bookkeeping. We will not do this in detail, but will provide the development for the Gaunt interaction as we did for the 2-spinor case.

We first expand the current density in 2-spinors,

$$\psi_r \boldsymbol{\alpha} \psi_s = \psi_r^{L\dagger} \boldsymbol{\sigma} \psi_s^S + \psi_r^{S\dagger} \boldsymbol{\sigma} \psi_s^L, \quad (11.90)$$

then expand the 2-spinors in scalar functions and perform the spin integrations that are implicit in the formalism to get

$$\psi_r^{L\dagger} \sigma_x \psi_s^S = \begin{pmatrix} \psi_r^{L\alpha} \\ \psi_r^{L\beta} \end{pmatrix}^\dagger \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \psi_s^{S\alpha} \\ \psi_s^{S\beta} \end{pmatrix} = \psi_r^{L\beta *}\psi_s^{S\alpha} + \psi_r^{L\alpha *}\psi_s^{S\beta}, \quad (11.91)$$

$$\psi_r^{L\dagger} \sigma_y \psi_s^S = \begin{pmatrix} \psi_r^{L\alpha} \\ \psi_r^{L\beta} \end{pmatrix}^\dagger \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \begin{pmatrix} \psi_s^{S\alpha} \\ \psi_s^{S\beta} \end{pmatrix} = i\psi_r^{L\beta *}\psi_s^{S\alpha} - i\psi_r^{L\alpha *}\psi_s^{S\beta}, \quad (11.92)$$

$$\psi_r^{L\dagger} \sigma_z \psi_s^S \Rightarrow \begin{pmatrix} \psi_r^{L\alpha} \\ \psi_r^{L\beta} \end{pmatrix}^\dagger \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \psi_s^{S\alpha} \\ \psi_s^{S\beta} \end{pmatrix} = \psi_r^{L\alpha *}\psi_s^{S\alpha} - \psi_r^{L\beta *}\psi_s^{S\beta}. \quad (11.93)$$

The Gaunt integrals can now be expressed in terms of the scalar functions as

$$\begin{aligned}
&(p\boldsymbol{\alpha}q \,|\, r\boldsymbol{\alpha}s) \\
&\quad = \left(p^{L\beta}q^{S\alpha} + p^{L\alpha}q^{S\beta} + p^{S\beta}q^{L\alpha} + p^{S\alpha}q^{L\beta} \,|\, r^{L\beta}s^{S\alpha} + r^{L\alpha}s^{S\beta} + r^{S\beta}s^{L\alpha} + r^{S\alpha}s^{L\beta}\right) \\
&\qquad - \left(p^{L\beta}q^{S\alpha} - p^{L\alpha}q^{S\beta} + p^{S\beta}q^{L\alpha} - p^{S\alpha}q^{L\beta} \,|\, r^{L\beta}s^{S\alpha} - r^{L\alpha}s^{S\beta} + r^{S\beta}s^{L\alpha} - r^{S\alpha}s^{L\beta}\right) \\
&\qquad + \left(p^{L\alpha}q^{S\alpha} - p^{L\beta}q^{S\beta} + p^{S\alpha}q^{L\alpha} - p^{S\beta}q^{L\beta} \,|\, r^{L\alpha}s^{S\alpha} - r^{L\beta}s^{S\beta} + r^{S\alpha}s^{L\alpha} - r^{S\beta}s^{L\beta}\right) \\
&\quad = 2\left(p^{L\beta}q^{S\alpha} + p^{S\beta}q^{L\alpha} \,|\, r^{L\alpha}s^{S\beta} + r^{S\alpha}s^{L\beta}\right) \\
&\qquad + 2\left(p^{L\alpha}q^{S\beta} + p^{S\alpha}q^{L\beta} \,|\, r^{L\beta}s^{S\alpha} + r^{S\beta}s^{L\alpha}\right) \\
&\qquad + \left(p^{L\alpha}q^{S\alpha} - p^{L\beta}q^{S\beta} + p^{S\alpha}q^{L\alpha} - p^{S\beta}q^{L\beta} \,|\, r^{L\alpha}s^{S\alpha} - r^{L\beta}s^{S\beta} + r^{S\alpha}s^{L\alpha} - r^{S\beta}s^{L\beta}\right).
\end{aligned} \quad (11.94)$$

Expanding the scalar functions in the scalar basis set, we can derive the contributions to the Fock matrices (with $X \neq Y$ and $\tau \neq \upsilon$),

$$\begin{aligned}
f_{\mu\nu}^{X\tau X\tau} &\leftarrow -\sum_{\kappa\lambda} \left(2D_{\kappa\lambda}^{Y\upsilon Y\upsilon} + D_{\kappa\lambda}^{Y\tau Y\tau}\right)\left(\mu^X\lambda^Y \,|\, \kappa^Y\nu^X\right), && (11.95)\\
f_{\mu\nu}^{X\tau X\upsilon} &\leftarrow \sum_{\kappa\lambda} D_{\kappa\lambda}^{Y\upsilon Y\tau}\left(\mu^X\lambda^Y \,|\, \kappa^Y\nu^X\right), && (11.96)\\
f_{\mu\nu}^{X\tau Y\tau} &\leftarrow \sum_{\kappa\lambda} \Big[\left(D_{\kappa\lambda}^{L\tau S\tau} - D_{\kappa\lambda}^{L\upsilon S\upsilon}\right)\left(\mu^X\nu^Y \,|\, \kappa^L\lambda^S\right) \\
&\qquad + \left(D_{\kappa\lambda}^{S\tau L\tau} - D_{\kappa\lambda}^{S\upsilon L\upsilon}\right)\left(\mu^X\nu^Y \,|\, \kappa^S\lambda^L\right) \\
&\qquad - \left(2D_{\kappa\lambda}^{X\upsilon Y\upsilon} + D_{\kappa\lambda}^{X\tau Y\tau}\right)\left(\mu^X\lambda^Y \,|\, \kappa^X\nu^Y\right)\Big], && (11.97)\\
f_{\mu\nu}^{X\tau Y\upsilon} &\leftarrow \sum_{\kappa\lambda} \Big[2D_{\kappa\lambda}^{L\upsilon S\tau}\left(\mu^X\nu^Y \,|\, \kappa^L\lambda^S\right) + 2D_{\kappa\lambda}^{S\upsilon L\tau}\left(\mu^X\nu^Y \,|\, \kappa^S\lambda^L\right) \\
&\qquad + D_{\kappa\lambda}^{X\upsilon Y\tau}\left(\mu^X\lambda^Y \,|\, \kappa^X\nu^Y\right)\Big]. && (11.98)
\end{aligned}$$

The expressions for the Fock matrix elements can be further reduced if we apply time-reversal symmetry. We assume a closed-shell wave function for this purpose. From the expansion of the spinors in the scalar functions, we can derive the following relations for the coefficients:

$$a_{\mu\bar{\imath}}^{X\beta} = \left(a_{\mu i}^{X\alpha}\right)^*, \quad a_{\mu\bar{\imath}}^{X\alpha} = -\left(a_{\mu i}^{X\beta}\right)^*. \tag{11.99}$$

Using these expressions in the density matrices, we get

$$D_{\kappa\lambda}^{X\beta Y\beta} = \left(D_{\kappa\lambda}^{Y\alpha X\alpha}\right)^*, \quad D_{\kappa\lambda}^{X\alpha Y\beta} = -\left(D_{\kappa\lambda}^{X\beta Y\alpha}\right)^*. \tag{11.100}$$

The same relations hold for the Fock matrix elements,

$$f_{\kappa\lambda}^{X\beta Y\beta} = \left(f_{\kappa\lambda}^{Y\alpha X\alpha}\right)^*, \quad f_{\kappa\lambda}^{X\alpha Y\beta} = -\left(f_{\kappa\lambda}^{X\beta Y\alpha}\right)^*. \tag{11.101}$$

We can also show that the direct contribution from the Gaunt interaction vanishes, as we saw in the 2-spinor formalism, and we are left with

$$\begin{aligned}
f_{\mu\nu}^{X\tau Y\tau} &\leftarrow -\sum_{\kappa\lambda} \left(2D_{\kappa\lambda}^{X\upsilon Y\upsilon} + D_{\kappa\lambda}^{X\tau Y\tau}\right)\left(\mu^X\lambda^Y \,|\, \kappa^X\nu^Y\right), && (11.102)\\
f_{\mu\nu}^{X\tau Y\upsilon} &\leftarrow \sum_{\kappa\lambda} D_{\kappa\lambda}^{X\upsilon Y\tau}\left(\mu^X\lambda^Y \,|\, \kappa^X\nu^Y\right). && (11.103)
\end{aligned}$$

11.9 Basis Set Choice and Design

Up to now the basis sets have appeared mostly as formal entities. However, the success of any computational scheme is crucially dependent on the size and suitability of the basis sets used. In this section, we discuss the various considerations that influence the choice of a good basis set for relativistic four-component calculations. It is obvious that we want basis sets that are generally applicable, give a good approximation of the exact functions with a minimum number of terms in the expansion, and provide for fast and easy calculation of matrix elements.

Today, the overwhelming majority of quantum chemical calculations are done using basis sets of Gaussian functions, that is, functions of the type

$$\chi_\mu(\mathbf{r}) = f(\mathbf{r})e^{-\zeta_\mu r^2}. \tag{11.104}$$

While this is by no means the only possible choice for relativistic four-component calculations, it is definitely the most efficient and convenient. It permits programmers to exploit an extensive technology that has been refined and tested through years of development of nonrelativistic methods. The usual complaint about the poor behavior of Gaussian functions close to nuclei is less severe for relativistic calculations, where nuclei of finite size are normally used and not point nuclei. As we discussed in chapter 7, Gaussians actually are particularly suitable for describing the wave function close to a nucleus of finite size.

It is tempting to assume that the basis for the large component should be quite close to the nonrelativistic basis. While this holds for lighter elements, it is normally not accurate enough for heavy elements. Some of the reasons for this are:

1. High exponents are required for an accurate description of the relativistically contracted inner shells.
2. The $2p_{1/2}$ orbital has considerable density close to the nucleus, requiring more high exponents than the nonrelativistic $2p$.
3. Orbitals with $\ell > 0$ are spin–orbit split into two components, which may have their maxima quite far apart, for example, 13 pm for the $6p$ of Rn and 35 pm for the $7p$ of element 117.

All this means that nonrelativistic basis sets at least need to be considerably modified for high-quality four-component calculations involving heavy elements, and, preferably, basis sets should be developed explicitly for such calculations.

It should also be borne in mind that the large-component basis is not totally independent of the small-component basis, and that the size of the total basis set will be determined by the kinetic balance requirement. If we use the relation derived previously in (11.19),

$$\chi_\mu^S \propto (\boldsymbol{\sigma} \cdot \mathbf{p})\chi_\mu^L, \tag{11.105}$$

we can easily demonstrate the consequences for a Gaussian basis. We define a spherical Gaussian 2-spinor for the large component by

$$\chi_\mu^L = N_{\mu n} r^{n-1} e^{-\zeta_\mu r^2} \xi_{\kappa,m}(\vartheta,\varphi) = (1/r)R_{\mu\kappa}^L(r)\xi_{\kappa,m}(\vartheta,\varphi). \tag{11.106}$$

Operating with $(\boldsymbol{\sigma} \cdot \mathbf{p})$ changes the sign of κ in the angular function, leaving the radial function, for which kinetic balance requires that

$$R^S_{\mu\kappa}(r) \propto \left[\frac{\mathrm{d}}{\mathrm{d}r} + \frac{\kappa + 1}{r}\right] R^L_{\mu\kappa}(r). \tag{11.107}$$

Application of this relation to the radial function yields

$$R^S_{\mu\kappa}(r) \propto N_{\mu\kappa n}\big[(n+\kappa)r^{n-1} - 2\zeta_\mu r^{n+1}\big]e^{-\zeta_\mu r^2} = \left[\frac{n+\kappa}{r} - 2\zeta_\mu r\right] R^L_{\mu\kappa}(r). \tag{11.108}$$

Normally, a spherical Gaussian basis set is made up of functions with minimum n-value, that is, $1s, 2p, 3d, \ldots$. We now get two different cases depending on whether κ is positive or negative:

CASE 1 $\kappa < 0$

This is the case for $1s_{1/2}$, $2p_{3/2}$, $3d_{5/2}, \ldots$. Then, $n = \ell + 1 = -\kappa$ and kinetic balance requires that

$$R^S_{\mu\kappa}(r) \propto -2\zeta_\mu r R^L_{\mu\kappa}(r). \tag{11.109}$$

This and the fact that $(\boldsymbol{\sigma} \cdot \mathbf{p})$ changes the sign of κ shows that a $1s_{1/2}$ function with exponent ζ_μ must be kinetically balanced by a $2p_{1/2}$ function with the same exponent, $2p_{3/2}$ must be balanced by $3d_{3/2}$, $3d_{5/2}$ by $4f_{5/2}$, and so on.

CASE 2 $\kappa > 0$

This is the case for $2p_{1/2}$, $3d_{3/2}$, $4f_{5/2}, \ldots$. Then $n = \kappa + 1$, and we must have

$$R^S_{\mu\kappa}(r) \propto \left[\frac{2\kappa + 1}{r} - 2\zeta_\mu r\right] R^L_{\mu\kappa}(r). \tag{11.110}$$

With the change in sign of κ, we see that a $2p_{1/2}$ function must be balanced by a linear combination of a $1s_{1/2}$ function and a $3s_{1/2}$ function. As mentioned above, functions with $n > \ell + 1$ (such as the $3s$) are not normally used in nonrelativistic basis sets, because linear combinations of $n = \ell + 1$ functions cover the same space. However, we see here that in four-component relativistic calculations the $n = \ell + 3$ functions are important for kinetic balance and must be included in the small-component basis. It is, of course, possible to gamble that the $n = \ell + 1$ functions can do the job also in this case, but it should be remembered that deviation from kinetic balance affects the bounds on the eigenvalues, and even a small deviation could have serious consequences for the energetics of a molecule. So for this case a $3d_{3/2}$ function is balanced by a linear combination of $2p_{3/2}$ and $4p_{3/2}$, a $4f_{5/2}$ by $3d_{5/2}$ and $5d_{5/2}$, and so on. These $n > \ell + 1$ functions are present by default in Cartesian basis sets, so any integral evaluation algorithm that uses Cartesian functions (and most do) can be easily adapted to produce the required functions.

If the $n > \ell + 1$ functions are included in the basis set, it is tempting to relax the requirement of a fixed linear combination in (11.110). Instead, the two functions could be treated as separate, independent small-component basis functions. Thus, while we would always have

$$\left\{R^{S}_{\mu\kappa}(r)\right\} \propto \left\{rR^{L}_{\mu\kappa}(r)\right\}; \quad \kappa < 0, \tag{11.111}$$

for negative κ, for positive κ we have a choice of either *restricted kinetic balance* (RKB) where

$$\left\{R^{S}_{\mu\kappa}(r)\right\} \propto \left\{\left[(1+2\kappa)r^{-1} - 2\zeta_{\mu}r\right]R^{L}_{\mu\kappa}(r)\right\}; \quad \kappa > 0, \tag{11.112}$$

or *unrestricted kinetic balance* (UKB) where

$$\left\{R^{S}_{\mu\kappa}(r)\right\} \propto \left\{rR^{L}_{\mu\kappa}(r)\right\} \cup \left\{r^{-1}R^{L}_{\mu\kappa}(r)\right\}; \quad \kappa > 0. \tag{11.113}$$

In reality, the computational work involved in either approach is about the same—the same number of integrals must be calculated. The difference shows up in solving the DHF equations, where the UKB Fock matrix is larger than the RKB Fock matrix. This has two consequences: the computational time for diagonalization of the Fock matrix is larger in UKB, and the extra functions generate a number of extra solutions. These solutions arise from small-component pieces that do not really have a matching large-component partner and consequently have a zero kinetic energy. We would therefore expect these solutions to cluster around $-2mc^2$, and experience shows that these solutions normally do not interfere much with the electronic solution space. However, for calculation of properties that depend directly on the small-component space, these extra solutions should be removed.

When it comes to contracted basis sets, kinetic balance strictly applied to the contracted large component can lead to problems. While it would be possible to apply the kinetic balance relation to derive a small-component basis from a set of large-component contracted basis functions, this procedure has been shown to be unsuitable in practice (Visscher et al. 1991). The best approach for generating contracted basis sets for relativistic four-component calculations has been to start with an uncontracted large-component basis, and to construct a small-component basis from this basis using kinetic balance. This set is then used in an uncontracted DHF calculation for the atom in question, yielding large- and small-component atomic functions that are kinetically balanced by virtue of the DHF equations. These atomic functions may then be used to select contracted basis functions for large and for small components. This technique has been termed *atomic kinetic balance* or *atomic balance*.

The preceding discussion has been based on a 2-spinor basis expansion of the wave function. For calculations in a scalar basis, the situation is somewhat more complicated than for 2-spinors. One major advantage in using a scalar basis is that integrals may be evaluated using real arithmetic, leaving the complex parts for the coefficients. This choice also allows us to use Cartesian Gaussians, the standard basis sets for nonrelativistic quantum chemistry. The gradient of a Cartesian Gaussian is of the form

$$\nabla x^i y^j z^k e^{-\zeta_\mu r^2} = x^i y^j z^k e^{-\zeta_\mu r^2} (i/x - 2\zeta_\mu x,\, j/y - 2\zeta_\mu y,\, k/z - 2\zeta_\mu z), \quad (11.114)$$

and κ is no longer a good quantum number, as could be expected. Also, integrals over Cartesian Gaussians are usually evaluated over all the $(\ell+1)(\ell+2)/2$ components of a given function, not just the $2\ell+1$ spherical harmonic components for that ℓ value. For example, integrals for $3d$ orbitals are evaluated over the Cartesian components d_{xx}, d_{xy}, d_{xz}, d_{yy}, d_{yz}, and d_{zz}. This corresponds to evaluating integrals over the spherical components $d_{-2}, d_{-1}, d_0, d_{+1}, d_{+2}$, plus the combination $d_{xx} + d_{yy} + d_{zz}$, which is a $3s$ function. Thus with a scalar basis it is natural to use a UKB approach, but again care should be taken to remove extraneous solutions.

The decision of whether to work with 2-spinors or a scalar spin–orbital basis must be made at an early stage of computer program construction because it affects all stages of the SCF process: evaluation of the integrals, construction of the Fock matrix, and solution of the SCF equations. However, at each stage, the scalar spin–orbital basis can be transformed to the 2-spinor basis. Transformation of the integrals to a 2-spinor basis is not particularly difficult: it is similar in principle to the transformation from Cartesians to spherical harmonics. Some efforts have been made to develop new algorithms in which these transformations are incorporated, and RKB is implemented from the start in the 2-spinor basis (Quiney et al. 1999, 2002, Yanai et al. 2002).

The main problem to be addressed in the choice between scalar and 2-spinor basis sets is that of linear dependence, which can affect the convergence of the SCF procedure. Several issues arise from the basis set choice that have consequences for linear dependence. The first of these is the choice of RKB or UKB in the representation of the small component of the $j = \ell - 1/2$ spinors. UKB has more severe linear dependence problems than RKB, due to the overrepresentation of the small component. However, RKB requires some manipulation of the integrals, because the nl and $(n+2)l$ integrals are not generated as part of the same angular momentum shell. This issue affects both scalar and 2-spinor basis sets.

The second issue is that, in a scalar basis, the set of spin–orbitals formed by taking the direct product of the scalar basis and the spin functions spans both j values for a given angular momentum. This is not a particular problem for the large component—in fact in an uncontracted basis set it is an advantage—but it does present some problems in the small-component basis. For a large-component s set, the small-component p set generated by kinetic balance forms spin–orbitals that span both the $p_{1/2}$ and the $p_{3/2}$ space. Only the $p_{1/2}$ spinors are needed for the small component of the large-component s functions. The $p_{3/2}$ spinors have no corresponding large component and produce SCF solutions that gather around $-2mc^2$. These solutions can be removed, just as for UKB. However, the small component generated from a large-component d set consists of p and f functions. The p functions span the same space as the p functions generated from the large-component s set, and the result can be linear dependence. There are standard techniques for eliminating linearly dependent functions from the basis set. Another solution to this problem is to make the large-component d exponents a subset of the s exponents, or, more generally, to constrain the functions for even ℓ to have the same exponents and the functions for odd ℓ to have the same exponents. These basis sets have been termed *dual family* basis sets, and they usually require a few more functions for each angular momentum higher than p than for fully optimized basis sets. There is a trade-off, because having common functions in the small component reduces the amount of work needed to evaluate the integrals. Even more efficiency could be obtained by using the same exponents for all angular momenta, as is done in the well-tempered basis sets (Huzinaga and Klobukowski 1985). Then, the integrals could be used for both large and small components. However, not much attention has been given to efficient integral algorithms for this kind of basis set.

The discussion above applies to uncontracted basis sets. Contracted basis sets present a few further problems. To properly represent the spin–orbit splitting, the two spin–orbit components should be contracted separately. The contraction is now j-dependent, rather than ℓ-dependent, and can only be represented directly in a 2-spinor basis. The problem is not now confined to the small component. If the large-component scalar basis set includes contractions for both spin–orbit components, the product of the contracted basis functions for each spin–orbit component with the spin functions generates a representation for both spin–orbit components. Thus there is a duplication of the basis set that is close to linearly dependent, and some kind of scheme to project out linearly dependent components, either numerically or by conversion to a 2-spinor basis, is mandatory. The same applies to the small component. For example, the contracted p sets for the large-component s and d sets both span the same space, but because of the contraction the d-generated set cannot be made a subset of the s-generated set, even if a dual family basis set is used.

One consequence of the choice between 2-spinor and scalar basis sets is that the results of DHF calculations with the two are not equivalent. The extra functions in the small-component space affect the spinor eigenvalues and hence the total energy and the ionization potential. These differences are likely to be minor for most of chemistry. If contracted basis sets are used, however, the duplication of large-component functions for the spin–orbit components in the scalar basis set provides extra flexibility that is not present for a 2-spinor basis set, and the valence properties could be significantly different.

In addition to these practical considerations, there has been considerable discussion of the formal considerations. The Dirac equation provides wave functions for both electrons and positrons. In the absence of any field, the solutions are related by charge conjugation, and there should be equal numbers of electron and positron functions and consequently equal numbers of large- and small-component basis functions. If we use a finite Gaussian basis set to represent the solutions, the basis set should also display charge-conjugation symmetry so that the representation of electron and positron solutions is equivalent. Kinetically balanced basis sets do not in fact display this symmetry: for that, the basis functions would have to satisfy both the relations

$$\chi_\mu^S \propto (\boldsymbol{\sigma} \cdot \mathbf{p})\chi_\mu^L, \quad \chi_\mu^L \propto (\boldsymbol{\sigma} \cdot \mathbf{p})\chi_\mu^S. \tag{11.115}$$

The first relation is the normal kinetic balance relation; the second is the kinetic balance relation resulting from elimination of the large component. To satisfy both of these relations, the basis functions must be solutions of

$$\nabla^2\chi = \lambda\chi, \tag{11.116}$$

which Gaussian functions are manifestly not. However, the principle that the electron and positron solutions should be equally represented in the basis set has been strongly argued for (Grant and Quiney 1988). The extra solutions that appear when this is not the case are unphysical. This principle is satisfied by 2-spinor RKB basis sets, but not in general by other basis sets. For properties that directly involve the small component, such as magnetic properties, basis sets with a 1:1 ratio of large- and small-component functions are preferred.

There exists a wide selection of exponents for Gaussian basis sets for nonrelativistic calculations, although most of these are for lighter elements which for most purposes do not require a relativistic treatment. For four-component relativistic calculations, nonrelativistic basis sets can be used for lighter atoms, but as the relativistic effects of orbital contraction and spin–orbit splitting increase in importance, these nonrelativistic basis sets become inadequate. In some measure the orbital contraction for inner orbitals is counteracted by the use of a finite nucleus, which tends to “push out” the inner parts of the spinors. A major concern is the $2p_{1/2}$ space (Matsuoka and Okada 1989): due to the s-character of the small component at least two extra functions relative to the nonrelativistic basis are needed for the $6p$ block to reduce the error in the energy to $0.5E_\mathrm{h}$.

The alternative to patching up a nonrelativistic basis set for four-component calculations is to derive sets of exponents specifically for use in relativistic calculations. This is usually done by optimizing the exponents directly in relativistic atomic calculations, for example by minimizing the energy. An alternative is to use an even-tempered basis, where the exponents form a geometric series such that

$$\zeta_k = \zeta_1\beta^{k-1}. \tag{11.117}$$

Energy-minimized sets yield lower energies than even-tempered sets for the same number of functions. However, the energy minimization can result in energies lower than the exact result because the bound on the energy is not the exact energy but, for kinetically balanced basis sets, is $\mathcal{O}(c^{-4})$ lower than the exact energy. This is not a variational collapse because there is a bound, but has been termed a *prolapse* (Fægri 2001). Systematic sequences of even-tempered basis sets do not generally suffer from this problem, and converge from above on the exact energy. For properties that are sensitive to the description of the region near the nucleus, even-tempered basis sets are preferred over energy-optimized basis sets because of their better coverage of that region of space.

The requirements for a proper description of the $p_{1/2}$ space, discussed above, point to another dilemma in the construction of relativistic basis sets. Nonrelativistic sets are *ℓ-based*, that is, exponents are derived for functions of s type, p type, d type, and so on. However, the tight functions required for $p_{1/2}$ spinors are not needed for the $p_{3/2}$ spinors, which have little amplitude so close to the nucleus. An alternative is to derive *j-based* basis sets, where all functions of the same j value have the same exponent. In this scheme, $1s_{1/2}$ and $2p_{1/2}$ sets would be of the same size and have common exponents, as would $2p_{3/2}$ and $3d_{3/2}$, $3d_{5/2}$ and $4f_{5/2}$, and so on. The justification for j-based sets is that for heavy elements, the small components of the $j = \ell - 1/2$ spinors penetrate as close to the nucleus as the large component of the $j = \ell + 1/2$ spinors, and therefore the basis set demands are the same. The use of a j-based basis set in scalar expansions would be somewhat complicated. For 2-spinor expansions, the use of such j-based sets requires few additional measures. The j-based sets would also alleviate another problem with ℓ-based sets: due to the large spin–orbit splitting in very heavy atoms, ℓ-based sets have to be rather extensive in the valence space in order to describe this splitting correctly. To achieve accuracy comparable to a double-zeta nonrelativistic set for a highly split ℓ-value, for example, might require at least a triple-zeta description of the valence space. The development of basis sets has been reviewed by Fægri and Dyall (2002).

11.10 Comparison of Nonrelativistic and Relativistic SCF Methods

Compared to the relatively simple matrix elements of the nonrelativistic Fock matrix, the expressions given in the sections above indicate the degree of increased complexity involved in a relativistic calculation. The critical issue is of course the logistics of the two-electron integrals. The following observations on the number of integrals can be made for a system that has no spatial symmetry using a 2-spinor RKB basis for the relativistic calculation.

- The spin can be factored out of the nonrelativistic integrals. In the relativistic integrals, the spin is coupled in. Therefore, there is a factor of 2 increase for each basis function, that is, a total factor of $2^4 = 16$. However, time-reversal symmetry reduces the relativistic integral count by a factor of 4. The net increase from the replacement of spin symmetry with time-reversal symmetry is therefore a factor of 4.
- The integrals and the basis are in general complex for relativistic calculations, so there is a factor of 2 increase for the loss of one degree of permutational symmetry relative to the nonrelativistic case.
- If the small-component basis is the same size as the large-component basis in a relativistic calculation, there is a factor of 2 increase for each basis function, that is, a factor of $2^4 = 16$. However, integrals with an odd number of large and small components never appear, reducing this factor by 2. Half of the remaining integrals are integrals over the Gaunt interaction, and if only the Coulomb interaction is used, there is another factor of 2 reduction. The net increase due to the small component is therefore a factor of 4 in a Dirac–Coulomb SCF calculation.

Thus, for a Dirac–Coulomb SCF calculation in a 2-spinor basis there are 32 times as many integrals as in an equivalent nonrelativistic SCF calculation. However, the integrals are complex in the general case so that there are 64 times as many pieces of data. Given that complex arithmetic involves 4 times as many operations as real arithmetic, the operation count for such a Dirac–Coulomb SCF calculation is 128 times that of a nonrelativistic SCF calculation. In modern computers, where a complex multiply can be initiated each clock cycle, this is not a disadvantage, and the operation count would reduce to 32 times. We should point out here that this accounting has been done for an orthodox list-driven SCF process. In integral-driven or direct SCF processes, the basis set dependence is lower than N^4. Also, modern techniques for large molecular systems employ various integral approximations that can further reduce the power of the basis set dependence. These techniques can be carried over to relativistic integral evaluation.

If we use a scalar basis instead, the large-component scalar basis can be essentially the same as the nonrelativistic basis. It is the small-component basis size that must be determined. Again, we can make the following observations on the relative sizes of the basis sets and the number of integrals.

- The large-component basis must be slightly larger than for the nonrelativistic calculation of the same quality. (This applies also to 2-spinor basis sets.)
- The principle of kinetic balance requires that the small component basis contains roughly 2.5 times as many functions as the large-component basis.
- The number of $(LL|SS)$ and $(LS|LS)$ integrals is then about 13 times the number of nonrelativistic integrals.
- The number of $(SS|SS)$ integrals is then about 39 times the number of nonrelativistic integrals.
- The total increase in the number of integrals for a Dirac–Coulomb calculation over a nonrelativistic calculation is about a factor of 50.

Thus, the scalar basis involves about 20% fewer real quantities than the 2-spinor basis, and therefore less work in the Fock matrix construction. This applies to an *uncontracted* basis set.

In a *contracted* scalar basis set, the core spin–orbit splitting must be accounted for, and the calculations become a little more difficult to do directly. The worst case for basis set size would be a basis set in which all functions are contracted, for example, a generally contracted basis set or an atomic natural orbital (ANO) basis set. Then, the large-component basis set would double in size, apart from the s functions. The small-component basis set, however, would not change, provided the contractions were properly restricted: for instance, the s component was eliminated from the d functions for the small component of the $p_{3/2}$ contracted functions.

Allowing for a 40% increase in the large-component basis size, the $(LL|LL)$ set would increase by a factor of 4 and the $(LL|SS)$ set by a factor of 2, giving $4 + 25 + 39 = 68$ as the overall factor, which is a little more than for the 2-spinor basis. If the increase is more like 70%, the $(LL|LL)$ set would increase by a factor of 9 and the $(LL|SS)$ set by a factor of 3, giving $9 + 39 + 39 = 87$. The 2-spinor method therefore has a size advantage in a contracted basis. If the small-component contractions were not properly restricted, however, the integral count in these two cases increases uniformly by a factor of 4 or of 9, which is much larger than in the nonrelativistic case.

By far the largest amount of work is spent on the $(SS|SS)$ integrals. Although their contribution to the total energy is of order c^{-4}, neglecting them does lead to significant errors in calculated molecular properties. Physical considerations as well as practical experience indicates that the main contributions to these integrals are strongly localized in the core regions of the atoms. A very good approximation is to include only those integrals that involve one-center small-component densities. This approximation reduces the number of $(SS|SS)$ integrals by a factor that is approximately the square of the number of atoms. It could also be applied to the $(LL|SS)$ integrals for a reduction of a factor that is approximately the number of atoms.

A further approximation eliminates the $(SS|SS)$ integrals entirely. Visscher (1997). has shown that a good approximation is to replace the contribution of the $(SS|SS)$ integrals with the Coulombic repulsion between the atomic small-component electron densities. The contribution to the energy is of the form

$$\varDelta E = \sum_{A} \varDelta E_A^{SS} + \sum_{A>B} \frac{q_A^S q_B^S}{R_{AB}}. \tag{11.118}$$

The first term is a term that accounts for the neglect of the $(SS|SS)$ integrals on the separate atoms. This term is constant for all practical purposes, and may be neglected. The other term provides the Coulomb repulsion between the small component charges on different atoms, and is essential for a correct description of the electron repulsion. Care should be exercised when using this approximation in calculations of properties that are particularly sensitive to small-component densities or the near-nucleus environment. A less severe approximation would be to fit the spherically averaged small-component density to a set of s functions and include their effect with the nuclear potential.

The conclusion is that use of screening methods and soundly based approximations can reduce the work by a large amount, and so the cost of efficient relativistic calculations is not so much larger than that of nonrelativistic calculations—perhaps a factor of 10 or less.

12

Correlation Methods

It is well known from nonrelativistic quantum chemistry that mean-field methods, such as the Hartree–Fock (HF) model, provide mainly qualitative insights into the electronic structure and bonding of molecules. To obtain reliable results of "chemical accuracy" usually requires models that go beyond the mean field and account for electron correlation. There is no reason to expect that the mean-field approach should perform significantly better in this respect for the relativistic case, and so we are led to develop schemes for introducing correlation into our models for relativistic quantum chemistry.

There is no fundamental change in the concept of correlation between relativistic and nonrelativistic quantum chemistry: in both cases, correlation describes the difference between a mean-field description, which forms the reference state for the correlation method, and the exact description. We can also define dynamical and nondynamical correlation in both cases. There is in fact no formal difference between a nonrelativistic spin–orbital-based formalism and a relativistic spinor-based formalism. Thus we should be able to transfer most of the schemes for post-Hartree–Fock calculations to a relativistic post-Dirac–Hartree–Fock model. Several such schemes have been implemented and applied in a range of calculations. The main technical differences to consider are those arising from having to deal with integrals that are complex, and the need to replace algorithms that exploit the nonrelativistic spin symmetry by schemes that use time-reversal and double-group symmetry.

In addition to these technical differences, however, there are differences of content between relativistic and nonrelativistic methods. The division between dynamical and nondynamical correlation is complicated by the presence of the spin–orbit interaction, which creates near-degeneracies that are not present in the nonrelativistic theory. The existence of the negative-energy states of relativistic theory raise the question of whether they should be included in the correlation treatment. The first two sections of this chapter are devoted to a discussion of these issues.

The main challenges in the rest of this chapter are to handle the presence of complex integrals and to exploit time-reversal symmetry. For this latter purpose we will resort to a Kramers-pair basis, as discussed in chapter 9. We will assume that the reader is familiar with the details of the nonrelativistic versions of the various methods that we discuss. If not, information about these may be found in several good texts, such as Jørgensen and Simons (1981), Szabo and Ostlund (1989), Helgaker et al. (2000), and Jensen (1999). Thus, we will concentrate mainly on features that arise specifically from the relativistic formulation of these methods. We start by describing two single reference methods, Møller–Plesset second-order perturbation theory (MP2) and the coupled-cluster (CC) model. We then discuss aspects of the CI expansion for four-component calculations, and we conclude with some remarks about MCSCF schemes.

In nonrelativistic quantum chemistry, there has been an impressive development in methods and algorithms to make post-HF calculations economically feasible. The rapid development of computers has greatly increased the scope of these methods. Still, the cost of doing these calculations rises rapidly with the size of the system. Relativistic methods in quantum chemistry will have one of its main impact areas in applications to molecules containing heavy elements. This makes the challenge of performing relativistic post-DHF calculations formidable. Not only do these molecules have large numbers of electrons, but valence energy levels are also usually quite closely spaced, and the outer core often needs to be correlated for quantitative accuracy. This means that “trivialities” such as operation counts and resource demands become even more important for relativistic methods than in the nonrelativistic case. We therefore provide some assessment of the cost of performing relativistic calculations compared with that of performing nonrelativistic calculations.

12.1 The Reference State

In nonrelativistic mean-field calculations, the orbitals are often sufficiently separated in energy that the reference state in a correlated calculation can be chosen to be a single determinant. Most closed-shell molecules fall into this category, and a considerable number of high-spin open-shell species can be well represented by a single determinant. The validity of the single-determinant reference can be established by measures such as the weight of the reference in a configuration interaction expansion or the t_1 diagnostic in coupled-cluster calculations.

In relativistic calculations, the spinors are not necessarily so well separated, due to the spin–orbit interaction. As an example of the effect of spin–orbit interaction, we choose the atoms of group 14—C, Si, Ge, Sn, Pb—which in a nonrelativistic picture have the valence configuration np^2, and the ground state is $np^2(^3P)$ in LS coupling. In a relativistic model the np manifold splits into the nondegenerate sets of $np_{1/2}$ and $np_{3/2}$ spinors. If we apply a simple Aufbau principle, we would end up with the state $2p^2_{1/2}2p^0_{3/2}(J = 0)$ for the relativistic ground state of C. If we expand the $2p_{1/2}$ spinor into spin–orbitals, we find that this state is $\frac{2}{3}\,2p^2(^3P)$ and $\frac{1}{3}\,2p^2(^1S)$, which we know is wrong because the spin–orbit splitting in C is very small and the wave function should not deviate so much from the LS-coupled state. We are left with no choice but to include more than one configuration in the wave function. The other possible choice, the $2p^0_{1/2}2p^2_{3/2}(J = 0)$ configuration, is $\frac{1}{3}\,2p^2(^3P)$ and $\frac{2}{3}\,2p^2(^1S)$. Mixing this with the other configuration allows the correct zeroth-order description of the ground state of the C atom to be obtained: a state that is mostly ^3P but with a small amount of ^1S.

For the mean-field spinors we have two choices: we can either use some restricted Hartree–Fock average, and thereby sacrifice the variational treatment of the spin–orbit splitting, or we can resort to a small multiconfigurational SCF calculation with the $np_{1/2}$ and $np_{3/2}$ orbitals as the active space. It may be argued that the spin–orbit coupling in C is small ($2.9 \times 10^{-4} E_\mathrm{h}$), and that the loss in using averaged $2p$ spinors probably would not introduce any serious errors. This has in fact been found to be a satisfactory approach in many atomic calculations. But the problem would become more serious as we go to the heavier elements of group 14. By the time we reach Pb, at atomic number 82, the spin–orbit splitting has increased to $5.5 \times 10^{-2} E_\mathrm{h}$, or more than 140 kJ/mol, and the ground state is closer to what would be obtained by the simple Aufbau filling of the p spinors.

Thus, while the ground state of these elements can be represented as a single, high-spin determinant in nonrelativistic theory, they must be represented by at least two determinants in relativistic theory. So for the relativistic case we may find ourselves forced to use methods beyond simple DHF to obtain reference states for cases that nonrelativistically could be comfortably treated in a mean-field model.

This problem becomes more acute when we lower the symmetry. For the atomic case, where the spin–orbit splitting is large, a single determinant might well be an adequate reference, even if it artificially fixes the mixing due to the spin–orbit effect. But what happens in a molecule?

We can analyze this problem formally by considering a doubly occupied Kramers pair. For simplicity we drop the small component and work with 2-spinor functions in a group that has no degenerate irreps. We also represent the wave function as a simple Hartree product. The extension to a determinantal function, 4-spinor functions, and groups with degeneracies is straightforward, but it would unnecessarily complicate the analysis. We write the wave function as

$$\Psi = \psi\bar{\psi} = \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \begin{pmatrix} -\phi_3 + i\phi_4 \\ \phi_1 - i\phi_2 \end{pmatrix} \tag{12.1}$$

where ϕ_1, ϕ_2, ϕ_3, and ϕ_4 are scalar functions, which consist of a coefficient multiplying a function that can be considered to be a molecular orbital. Expanding the spinor product into its components and spin functions, we get the following linear combination of states:

$$\begin{aligned} \Psi = {} & [\phi_1\phi_1 + \phi_2\phi_2 - \phi_3\phi_3 - \phi_4\phi_4]\, \eta(\alpha)\eta(\beta) \\ & - i\,[\phi_1\phi_2 - \phi_3\phi_4]\,[\eta(\alpha)\eta(\beta) - \eta(\beta)\eta(\alpha)] \\ & - [\phi_1\phi_3 + \phi_2\phi_4]\,[\eta(\alpha)\eta(\alpha) + \eta(\beta)\eta(\beta)] \\ & + i\,[\phi_1\phi_4 - \phi_2\phi_3]\,[\eta(\alpha)\eta(\alpha) - \eta(\beta)\eta(\beta)]\,. \end{aligned} \tag{12.2}$$

Using Hartree products, a state now appears as a product of two spatial functions and a spin function that is an eigenfunction of $\hat{S}^2$. Thus the first four states are nonrelativistic closed-shell singlet states; the next two are open-shell singlet states, and the remaining eight are triplet states. If this were a nonrelativistic wave function with independent coefficients for each of these states, the wave function would include elements of non-dynamical correlation, for both the singlet and the triplet components (and would be separable into spin components). In the relativistic SCF procedure we have in fact only four independent coefficients, but the elements of nonrelativistic nondynamical correlation remain, and the optimum wave function is a balance between the self-consistent field effects, the nondynamical correlation effects, and the spin–orbit effects, constrained by the form of the spinors. The constraints could force the single-determinant relativistic SCF process to a solution that was not close to the physical solution. In the general case, to remove the constraints requires an MCSCF calculation. There are of course many molecules for which the relativistic SCF procedure provides a quite satisfactory single reference—such as for AuH, where only one of the four components in the bonding large-component 2-spinor has significant amplitude. However, the choice of correlation method in a relativistic calculation must take into account the possibility of a multiconfigurational reference.

12.2 The No-Pair Approximation

When deriving mean-field methods, we had to consider the role of the negative-energy states in the variation of the wave function. Now that we are considering correlation, we must examine the role of these functions in the N-particle wave function.

The most general approach to variational calculations is represented by the MCSCF wave function, with the simultaneous optimization of both the one-particle and the N-particle functions. The MCSCF wave function may be written (Jørgensen and Simons 1981)

$$\Psi = e^{-\hat{\kappa}} \sum_{\mu} C_\mu \Phi_\mu \tag{12.3}$$

where the Φ_μ are single determinants and $e^{-\hat{\kappa}}$ performs rotations between the one-particle functions (orbitals or spinors).[1] The variational parameters are the vectors $\boldsymbol{\kappa}$ and $\mathbf{C}$. The set $\{\Phi_\mu\}$ contains determinants that may be derived by exciting electrons from orbitals that are occupied in some reference determinant to virtual orbitals, that is, orbitals that are not occupied in the reference state.

$$\{\Phi_\mu\} = \{\Phi_0, \Phi_i^a, \Phi_{ij}^{ab}, \ldots\} \tag{12.4}$$

where

$$\Phi_i^a = a_a{}^\dagger a_i \; \Phi_0, \quad \Phi_{ij}^{ab} = a_a{}^\dagger a_b{}^\dagger a_j \; a_i \; \Phi_0, \quad \ldots . \tag{12.5}$$

Here and in the following, we will use indices $i, j, k, \ldots$ to denote one-particle functions occupied in the reference determinant, and $a, b, c, \ldots$ to denote virtual one-particle functions. In the nonrelativistic case, these one-particle functions are spin–orbitals, normally derived as eigenvalues of the Hartree–Fock operator. For the relativistic case, the

---

[1] In this chapter we use the opposite sign convention for $\hat{\kappa}$ to that used in chapter 8.

corresponding one-particle basis functions are the spinors derived from the solution of the Dirac–Hartree–Fock equation, including the spinors that describe negative-energy solutions, that is, solutions with eigenvalues less than $-2mc^2$.

As discussed in chapter 5, the treatment of the negative-energy states is crucial if we are to establish a consistent model for relativistic many-electron systems. In this connection we have to decide what role these negative-energy spinors should have in our variational space. A moment's consideration should convince us that we do not want our variational space of N-particle functions to include determinants where positive-energy spinors have been replaced by negative-energy spinors. Such a choice would open the possibility of variational collapse, due to the fact that there would be states either below the ground state or degenerate with it, as discussed in chapter 5. Likewise, inclusion of such determinants in a perturbation expansion risks introduction of intruder states that would cause the perturbation method to fail. So the conclusion must be that the N-particle space should be made up of determinants containing positive-energy spinors only. This is frequently referred to as the *no-pair approximation.*

While this no-pair approximation is chemically reasonable and convenient to work with, it does have consequences for our use of the one-particle variational space. For the nonrelativistic case, the orbital rotation space is the electronic spin–orbitals, and we can indicate this explicitly in the expression for the MCSCF wave function as

$$\Psi = e^{-\hat{\kappa}_{ee}} \sum_{\mu} C_{\mu} \Phi_{\mu} \tag{12.6}$$

where $e^{-\hat{\kappa}_{ee}}$ includes rotations for the entire spin–orbital space. In the limit of a full CI, that is, the set $\{\Phi_{\mu}\}$ includes all possible excitations using the one-particle basis, the spin–orbital rotations become redundant, and the operator $e^{-\hat{\kappa}_{ee}}$ may be omitted. The corresponding expression for the relativistic MCSCF function is

$$\Psi = e^{-\hat{\kappa}_{ee}-\hat{\kappa}_{ep}} \sum_{\mu} C_{\mu} \Phi_{\mu} \tag{12.7}$$

with $e^{-\hat{\kappa}_{ep}}$ accounting for mixing in of negative-energy one-particle functions. This orbital rotation operator will never become redundant as long as we use the no-pair approximation. Thus, to be complete, any no-pair CI expansion of a relativistic wave function should include spinor optimization over the entire one-particle basis.

Keeping the above considerations in mind, and adhering to the no-pair approximation, we should be able to recast most of the nonrelativistic post-HF schemes into a relativistic post-DHF form.

12.3 Integral Transformations

As long as one stays with "traditional" methods for calculation of correlation energies, it is necessary to perform a transformation of the one- and two-electron integrals into the molecular spinor basis. While Møller-Plesset perturbation expansions can be cast in a semi-direct form that does not require a complete integral transformation, it is not so easy for other correlation methods. The alternatives are to resort to approximate methods, density functional methods, or methods that incorporate interelectronic distances, the so-called r_{12} methods. In this chapter, we concentrate on the traditional approaches.

The integral transformation itself[2] is straightforward, but it must be broken into pieces because of the large and small components. For the Coulomb interaction, the two-electron integral transformation can be written

$$(pq|rs) = \sum_{XY} \sum_{\mu\nu\kappa\lambda} \mathrm{a}_{\mu p}^{X*} \mathrm{a}_{\nu q}^{X} \mathrm{a}_{\kappa r}^{Y*} \mathrm{a}_{\lambda s}^{Y} (\mu^X \nu^X | \kappa^Y \lambda^Y). \tag{12.8}$$

Because we have restricted the N-particle space to states composed of positive-energy spinors, the transformation reduces the number of integrals by a factor of 4. Just as in the nonrelativistic case, the transformation over the spinor indices can be performed as a sequence of four matrix multiplications, and as a result the cost of the integral transformation scales as n^5, where n is the number of basis functions.

Mean-field calculations for large systems are generally carried out within a direct SCF scheme, and therefore the number of integrals that actually need to be calculated is much smaller, and the scaling with basis set size for these processes has been shown to be closer to $n^{2.2}$ rather than n^4. The integral transformation is therefore more expensive than the SCF step, but apart from second-order perturbation theory, most correlated methods are at least $\mathcal{O}(n^6)$.

We can estimate the increased relative cost of relativistic integral transformations over nonrelativistic integral transformations as follows. For this purpose we assume that the integral transformation is performed with one triangular pair index and one square pair index to exploit the efficiency of matrix multiplications. We also assume a nonrelativistic basis of size n that matches the relativistic 2-spinor Kramers-pair basis.

- The operation count for the nonrelativistic case (multiply and add) is $2n^5$.
- The relativistic transformation must include integral classes $(LL|LL)$, $(LL|SS)$, $(SS|LL)$, and $(SS|SS)$. Without any further efficiencies, this results in a factor of 4 increase in the operation count.
- In the optimal implementation of time-reversal symmetry, there are four classes of integrals over molecular spinors, $(pq|rs)$, $(p\bar{q}|\bar{r}s)$, $(p\bar{q}|r\bar{s})$, and $(pq|r\bar{s})$, each of which scales the same. Without further efficiencies, this also results in a factor of 4 increase in the operation count.
- The integrals are complex, introducing another factor of 4, an overall factor of 64, and a total operation count of $128n^5$.
- However, in the first half-transformation, we can exploit the common indices, and add the half-transformed integrals to yield $(pq|LL)$, $(pq|SS)$, $(p\bar{q}|LL)$, and $(p\bar{q}|SS)$ integrals, with an operation count of $24n^5$. The operation count for the second half-transformation is also $24n^5$, yielding a total of $48n^5$.

The Kramers-restricted integral transformation is therefore 24 times more expensive than the nonrelativistic integral transformation.

---

[2] Relativistic integral transformations were first discussed by Esser et al. (1981).

Here we have only considered the transformation of integrals from the 2-spinor basis. The transformation of integrals from a scalar basis is a little more complicated because there are sums over spin components, but it is no different in principle.

12.4 Kramers-Restricted Møller–Plesset Perturbation Theory

The use of low-order perturbation theory is probably the cheapest and conceptually simplest method for including correlation effects in a quantum-chemical calculation while maintaining a minimum of formal rigor. In particular, Møller–Plesset perturbation expansions to various orders (commonly denoted MPn) have seen widespread use. For our purposes it is sufficient to discuss only the MP2 expansion, which is the lowest order that contributes beyond the mean-field approximation.

The Møller–Plesset zeroth-order Hamiltonian is just a sum over Fock operators, which can be expressed in second quantization as

$$\hat{\mathcal{H}}_0 = \sum_i a_i{}^\dagger a_i \; \epsilon_i . \tag{12.9}$$

This Hamiltonian gives a zeroth-order energy that is a sum of orbital energies, and therefore double-counts the electron repulsion. The perturbation is the difference between the mean-field electron interaction and the full electron–electron interaction operator, the so-called *fluctuation potential*. Thus, the first-order contribution to the energy just corrects for the overcounting of electron repulsion in $E^{(0)}$ and we recover the energy from the mean-field calculation

$$E^{\mathrm{MF}} = E^{(0)} + E^{(1)}. \tag{12.10}$$

The general expression for the second-order energy with a single-determinant reference is

$$E^{(2)} = \frac{1}{4} \sum_{i,j} \sum_{a,b} \frac{\left|(ia||jb)\right|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} \tag{12.11}$$

where the ϵ are eigenvalues of the Fock matrix and $(ia||jb)$ are the antisymmetrized two-electron integrals

$$(ia||jb) = (ia|jb) - (ib|ja). \tag{12.12}$$

This expression applies equally to the nonrelativistic case expressed in terms of spin–orbitals and the relativistic case expressed in terms of spinors. In nonrelativistic MP2, the sums over i, j and a, b span the range of occupied and virtual one-electron functions, respectively. In the relativistic formulation the sum over virtual functions can only include spinors with positive energies if we are to stay within the no-pair approximation.

The usual zeroth-order wave function for an MP2 expansion is a closed-shell single determinant. For the nonrelativistic case we can exploit this fact and spin symmetry in order to reduce the expression for $E^{(2)}$ to one involving only unique integrals over real spatial orbitals:

$$E_{\mathrm{NR}}^{(2)} = \sum_{i,j} \sum_{a,b} \frac{[2(ia|jb) - (ib|ja)](bj|ia)}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b}. \tag{12.13}$$

The sums extend over all orbitals. However, use has been made of the symmetry in the indices i and j to collect the terms.

For the relativistic case, we use time-reversal symmetry instead of nonrelativistic spin symmetry to effect a reduction, and we introduce a set of spinors that are grouped as Kramers pairs. If we exploit time reversal on the integrals, we find that

$$(ia||jb) = (\bar{\imath}\bar{a}||\bar{\jmath}\bar{b})^*, \tag{12.14}$$

and we can use the bar-reversal operator to derive expressions for all the other integrals. Making use of these relations results in a reduction of a factor of 2 in the number of terms in the sum for the second-order energy. If we also exploit the equivalence of the index pairs a and b, and i and j, we arrive at the following expression for the Kramers-restricted second-order energy:

$$\begin{aligned} E^{(2)} = &\frac{1}{2} \sum_{i,j} \sum_{a,b} \frac{\left|(ia||jb)\right|^2 + 2\left|(ia||\bar{\jmath}\bar{b})\right|^2 + \left|(i\bar{a}||j\bar{b})\right|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} \\ &+ \sum_{i,j} \sum_{a,b} \frac{\left|(ia||\bar{\jmath}b)\right|^2 + \left|(ia||j\bar{b})\right|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b}. \end{aligned} \tag{12.15}$$

This is not quite a reduction of a factor of 4 that would be expected from the integral equivalences under time reversal.

Comparing this expression with the nonrelativistic expression, we see that there are five times as many terms. This does not translate into a factor of 5 in the amount of work. For that, we need to consider the number of floating-point operations. Designating the number of occupied orbitals O and the number of virtual orbitals N, there are O^2V^2 terms in each sum, both relativistic and nonrelativistic.

In the nonrelativistic expression, there are two adds and one multiply to form the numerator. To form the denominator, the orbital eigenvalues can be added in pairs to form arrays of order O^2 and V^2. The work to assemble the final sum of orbital energies is then one add per denominator. Finally, there is one divide and one add per term to accumulate the energy. The total operation count for the nonrelativistic MP2 expression given above is therefore $6O^2V^2$.

In the relativistic expression, the work required to form the denominator is the same as for the nonrelativistic expression, but only needs to be done once for all sums. The integrals are complex, so each numerator term requires the equivalent of three adds and two multiplies. The five numerators can be added and then the division and accumulation performed. The total operation count is $31O^2V^2$. The relativistic to nonrelativistic ratio is therefore a little more than 5.

The expressions given above have complete sums over all indices, which can be exploited to simplify the implementation. However, we can reduce the operation count by restricting the summation ranges and exploiting the index symmetry. The relativistic expression reduces to

$$E^{(2)} = 2\sum_{i>j}\sum_{a>b} \frac{\left|(ia||jb)\right|^2 + \left|(i\bar{a}||j\bar{b})\right|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} + \sum_{i\geq j}(2-\delta_{ij})\sum_{a,b} \frac{\left|(ia||\bar{j}\bar{b})\right|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b}$$

$$+2\sum_{i,j}\sum_{a>b} \frac{\left|(ia||\bar{j}b)\right|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} + 2\sum_{i>j}\sum_{a,b} \frac{\left|(ia||j\bar{b})\right|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b}. \qquad (12.16)$$

The operation count for this expression is $12.5O^2V^2$. The nonrelativistic expression given above cannot be reduced in this way, but we can use the above relativistic expression and treat the unbarred and barred spinors as α and β spin–orbitals. We are left with the expression

$$E^{(2)}_{\text{NR}} = 2\sum_{i>j}\sum_{a>b} \frac{\left|(ia||jb)\right|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} + \sum_{i\geq j}(2-\delta_{ij})\sum_{a,b} \frac{\left|(ia|jb)\right|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b}. \qquad (12.17)$$

The operation count for this expression is $1.5O^2V^2$, and the relativistic to nonrelativistic ratio is now a little more than 8. The index restrictions therefore benefit the relativistic MP2 method less than the nonrelativistic method, but the ratio is still much less than that for SCF calculations.

Symmetry reductions for the relativistic MP2 method follow very similar lines to those for the DHF method. For the complex reps (singly degenerate in pairs) the integrals with an odd number of bars vanish, while for the real irreps (doubly degenerate), the integrals become real. Both result in reductions of a factor of approximately 2.

One great advantage of the MP2 method is that it may be cast in a semi-direct or integral-driven algorithm where classes of integrals are recalculated rather than stored. The general experience is that a relativistic MP2 calculation requires roughly the same computing time as a mean-field calculation.

The Møller–Plesset perturbation expansion is based on a given set of one-electron functions, which do not change during the calculation. For the relativistic case, this means that in the no-pair approximation spinor relaxation through mixing of positive- and negative-energy solutions is sacrificed, as discussed in the previous section. While this usually is of little importance for the calculation of correlation energies and molecular geometries through perturbation theory, care must be exercised in the calculation of properties from a perturbation expansion.

As mentioned above, the MP2 expansion usually starts from a closed-shell, single-determinant zeroth-order wave function. For the nonrelativistic case, several schemes for extending this to open-shell system have been proposed and applied. Most of these schemes can also be cast in an equivalent relativistic version. There are, however, some peculiarities of open-shell Kramers-restricted MP2 when the open shell is spin–orbit split, particularly if one is aiming at calculating energies including basis-set superposition error (BSSE). Suppose we are calculating the dissociation energy of a diatomic molecule such as Cl_2 by the counterpoise method. The Cl atom calculation with the other Cl as a ghost does not have the full atomic symmetry. The $p_{1/2}$ and $p_{3/2}$ spinors with $\omega = 1/2$ mix and the basis set on the ghost atom contributes a greater energy lowering to the σ component than to the π component. This affects the calculated spin–orbit splitting and changes the energetics.

The presence of the ghost basis is not the only problem with open-shell relativistic MP2 calculations. In nonrelativistic calculations there is a symmetry breaking due to the lack of spherical averaging, but it is small and not very orientation-dependent for reasonable basis sets. In relativistic calculations the symmetry breaking involves the spin–orbit interaction, and affects the core spinors indirectly. These may rotate significantly because the energy differences between the spin–orbit components are much smaller than the energy differences between the shells. Thus the symmetry breaking at the SCF level can propagate into the core. In subsequent MP2 calculations the resulting eigenvalue shifts can cause an increased energy difference between the two nominally identical states. The only remedy is to use methods that permit orbital relaxation and in which the core is also allowed to relax.

For more than one open shell, the wave function can consist of two or more determinants. To do MP2 calculations for dynamic correlation on these systems requires a multiconfiguration perturbation theory. The various theories proposed can be reasonably straightforwardly extended to the relativistic case.

12.5 Kramers-Restricted Coupled-Cluster Expansions

Another correlation method commonly used in nonrelativistic quantum chemistry is the coupled-cluster (CC) method. In this method the wave function is developed by applying an exponential *wave operator* to an N-particle reference function,

$$\Psi = e^{\hat{T}}\, \Phi_0 \tag{12.18}$$

where the operator $\hat{T}$ is a sum of products of excitation operators $\hat{E}$ and amplitudes (or expansion coefficients) t,

$$\begin{aligned} \hat{T} &= \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \cdots \\ &= \sum_i \sum_a t_i^a\, \hat{E}_i^a + \sum_{ij} \sum_{ab} t_{ij}^{ab}\, \hat{E}_{ij}^{ab} + \sum_{ijk} \sum_{abc} t_{ijk}^{abc}\, \hat{E}_{ijk}^{abc} + \cdots . \end{aligned} \tag{12.19}$$

Note that these excitation operators differ from those of section 9.1. They are given in terms of one-particle creation and annihilation operators as

$$\hat{E}_i^a = a_a{}^{\dagger} a_i; \qquad \hat{E}_{ij}^{ab} = a_a{}^{\dagger} a_b{}^{\dagger} a_j\, a_i; \qquad \hat{E}_{ijk}^{abc} = a_a{}^{\dagger} a_b{}^{\dagger} a_c{}^{\dagger} a_k\, a_j\, a_i, \ldots . \tag{12.20}$$

To simplify the appearance of the equations, we invoke the Einstein summation convention for this section, in which a summation over repeated indices is made. The expression for $\hat{T}$ reduces to

$$\hat{T} = t_i^a \, \hat{E}_i^a + t_{ij}^{ab} \, \hat{E}_{ij}^{ab} + t_{ijk}^{abc} \, \hat{E}_{ijk}^{abc} + \cdots . \tag{12.21}$$

The CCSD method (coupled cluster with single and double excitations) truncates the expansion of $\hat{T}$ after the double excitation term, $\hat{T}_2$. The t_2 amplitudes obey the relations

$$t_{ij}^{ab} = -t_{ji}^{ab} = -t_{ij}^{ba} = t_{ji}^{ba}, \tag{12.22}$$

that is, they are antisymmetric to index interchange in the occupied spinor space and the virtual spinor space.

If the reference is a closed-shell determinant, it is symmetric under time reversal. We require that the coupled-cluster wave function is also symmetric under time reversal, because the wave function is nondegenerate and has an even number of electrons. Then

$$\hat{\mathcal{K}} \, \Psi = \hat{\mathcal{K}} \, e^{\hat{T}} \, \Phi_0 = \Psi = e^{\hat{T}} \, \hat{\mathcal{K}} \, \Phi_0. \tag{12.23}$$

Thus $\hat{\mathcal{K}}$ commutes with the wave operator and the wave operator is therefore symmetric under time reversal. Because of the structure of the wave operator, each term $\hat{T}_i$ must be symmetric under time reversal. We now introduce a Kramers-pair basis, and apply the time-reversal operator to the $\hat{T}_1$ term, to obtain

$$\begin{aligned}
\hat{\mathcal{K}} \, \hat{T}_1 &= \hat{\mathcal{K}} \left[t_i^a \, \hat{E}_i^a + t_{\bar{\imath}}^a \, \hat{E}_{\bar{\imath}}^a + t_i^{\bar{a}} \, \hat{E}_i^{\bar{a}} + t_{\bar{\imath}}^{\bar{a}} \, \hat{E}_{\bar{\imath}}^{\bar{a}} \right] \\
&= \hat{\mathcal{K}} \, t_i^a \, a_a{}^{\dagger} a_i \; + \hat{\mathcal{K}} \, t_{\bar{\imath}}^a \, a_a{}^{\dagger} a_{\bar{\imath}} \; + \hat{\mathcal{K}} \, t_i^{\bar{a}} \, a_{\bar{a}}{}^{\dagger} a_i \; + \hat{\mathcal{K}} \, t_{\bar{\imath}}^{\bar{a}} \, a_{\bar{a}}{}^{\dagger} a_{\bar{\imath}} \\
&= t_i^{a*} \, a_{\bar{a}}{}^{\dagger} a_{\bar{\imath}} \; - t_{\bar{\imath}}^{a*} \, a_{\bar{a}}{}^{\dagger} a_i \; - t_i^{\bar{a}*} \, a_a{}^{\dagger} a_{\bar{\imath}} \; + t_{\bar{\imath}}^{\bar{a}*} \, a_a{}^{\dagger} a_i \\
&= t_i^{a*} \, \hat{E}_{\bar{\imath}}^{\bar{a}} - t_{\bar{\imath}}^{a*} \, \hat{E}_i^{\bar{a}} - t_i^{\bar{a}*} \, \hat{E}_{\bar{\imath}}^a + t_{\bar{\imath}}^{\bar{a}*} \, \hat{E}_i^a
\end{aligned} \tag{12.24}$$

which gives the relations

$$t_i^a = t_{\bar{\imath}}^{\bar{a}*}; \qquad t_{\bar{\imath}}^a = -t_i^{\bar{a}*}. \tag{12.25}$$

These are the normal relations expected from time reversal of a one-electron operator. Similarly, for the two-electron amplitudes,

$$t_{ij}^{ab} = t_{\bar{\imath}\bar{\jmath}}^{\bar{a}\bar{b}*}; \qquad t_{ij}^{\bar{a}\bar{b}} = t_{\bar{\imath}\bar{\jmath}}^{ab*}; \qquad t_{i\bar{\jmath}}^{a\bar{b}} = t_{\bar{\imath}j}^{\bar{a}b*}; \qquad t_{ij}^{a\bar{b}} = -t_{\bar{\imath}\bar{\jmath}}^{\bar{a}b*}. \tag{12.26}$$

If we use these equalities between the various excitation amplitudes of the same rank, we can simplify the expressions somewhat. For $\hat{T}_1$ we get

$$\begin{aligned}\hat{T}_1 &= t_i^a\, a_a{}^\dagger a_i + t_{\bar{i}}^a\, a_a{}^\dagger a_{\bar{i}} - t_{\bar{i}}^{a*}\, a_{\bar{a}}{}^\dagger a_i + t_i^{a*}\, a_{\bar{a}}{}^\dagger a_{\bar{i}} \\ &= (\mathrm{Re}\, t_i^a + i \mathrm{Im}\, t_i^a)\, a_a{}^\dagger a_i + (\mathrm{Re}\, t_{\bar{i}}^a + i \mathrm{Im}\, t_{\bar{i}}^a)\, a_a{}^\dagger a_{\bar{i}} \\ &\quad - (\mathrm{Re}\, t_{\bar{i}}^a - i \mathrm{Im}\, t_{\bar{i}}^a)\, a_{\bar{a}}{}^\dagger a_i + (\mathrm{Re}\, t_i^a - i \mathrm{Im}\, t_i^a)\, a_{\bar{a}}{}^\dagger a_{\bar{i}}\,.\end{aligned} \tag{12.27}$$

We can use the excitation operators defined in (9.14)

$$\hat{E}^{\pm}_{pq} = a_p{}^\dagger a_q \pm a_{\bar{p}}{}^\dagger a_{\bar{q}}\ , \qquad \hat{E}^{\pm}_{\bar{p}q} = a_{\bar{p}}{}^\dagger a_q \mp a_p{}^\dagger a_{\bar{q}}\ , \tag{12.28}$$

to simplify the final expression for $\hat{T}_1$ to

$$\hat{T}_1 = \mathrm{Re}\, t_i^a\ E^+_{ai}\ +\ i \mathrm{Im}\, t_i^a\ E^-_{ai}\ +\ \mathrm{Re}\, t_{\bar{i}}^a\ E^+_{a\bar{i}}\ +\ i \mathrm{Im}\, t_{\bar{i}}^a\ E^-_{a\bar{i}}. \tag{12.29}$$

Here we have to use the operators $\hat{E}^{\pm}_{pq}$ and not the $\hat{X}^{\pm}_{pq}$. The reason is that to ensure that the commutator expansions in coupled-cluster theory truncate, the wave operator must be expressed in terms of excitation operators between two disjoint one-particle spaces. This expression and the corresponding one for $\hat{T}_2$ form the basis for the Kramers-restricted CCSD (KRCCSD) method (Visscher et al. 1995).

As for the SCF method, the main reductions due to double-group symmetry follow from the nature of the irreps. For quaternion irreps, all amplitudes are in principle nonzero; for complex and real irreps the odd-bar amplitudes vanish, giving a reduction of a factor of 2; and for real irreps the imaginary part of the amplitudes is zero, giving another reduction of a factor of 2.

At this point, we pause to consider the cost of the relativistic KRCCSD method compared with the nonrelativistic spin-restricted CCSD method.

- The fact that the integrals are complex in the relativistic version gives a factor of 4 increase in the number of operations.
- The scaling of the CCSD equations is $\mathcal{O}(n^6)$, where n is the number of basis functions. Thus the increase due to lack of spin symmetry in the relativistic case should be a factor of $2^6 = 64$.
- The reduction due to the optimal implementation of time-reversal symmetry is a factor of 8.

The KRCCSD method is therefore a factor of 32 times more expensive than the nonrelativistic CCSD method.

Before proceeding to open-shell theory, it is worth noting that CCSD properly treats the nondynamical effects that are missing in a single-determinant reference function, which were discussed in section 12.1. This is because the coupled-cluster wave function is an infinite-order expansion to the given excitation level: the coefficients of the determinants that complete the reference expansion *and* all the excitations from these are included and optimized in the coupled-cluster wave function. Also, the presence of single excitations accounts for the orbital relaxation that would correct the distortion of the reference determinant.

12.6 Open-Shell Kramers-Restricted Coupled-Cluster Expansions

Open-shell coupled-cluster theory can be formulated in terms of the one-particle spinor or spin–orbital basis. However, spin-restricted or Kramers-restricted open-shell theories are complicated by the ambiguous role of the open-shell orbital or Kramers pair. We develop here the basic outline of open-shell Kramers-restricted coupled-cluster theory.

We consider first the case of a single electron outside a closed-shell core, for which the reference N-particle functions form a Kramers pair of states

$$\hat{\mathcal{K}}\Phi_0 = \bar{\Phi}_0. \tag{12.30}$$

Because the open-shell spinor is different for the two partners, the Kramers-unrestricted wave operator cannot be the same. At the least it must differ in the terms that involve the open-shell Kramers pair. Hence we must have two excitation operators, $\hat{T}$ and $\bar{T}$, such that

$$\Psi = e^{\hat{T}}\,\Phi_0, \qquad \bar{\Psi} = e^{\bar{T}}\,\bar{\Phi}_0. \tag{12.31}$$

(where we have omitted the operator “hat” over $\bar{T}$ for typographical reasons). Since the correlated wave functions must be related by time reversal, we can derive a relation for the wave operators:

$$e^{\bar{T}} = \hat{\mathcal{K}}e^{\hat{T}}. \tag{12.32}$$

This relation gives us a connection between the t amplitudes and the $\bar{t}$ amplitudes, but it does not provide any relation between t amplitudes for the Kramers partners such as we get for the closed-shell case. This stands to reason, because the amplitudes represent the configuration mixing due to correlation, and we cannot expect the correlation to be the same for α and β spin in an open-shell doublet. The incorporation of spin–orbit interaction makes no change to this picture, in which the Pauli repulsion between spin–orbitals of the same spin is transferred to spinors of the same irrep row.

The only way to get a wave operator that is symmetric under time-reversal symmetry is to impose the restriction from the beginning. While this fixes the relations between the amplitudes, it also forces the occupied and the unoccupied Kramers components of the open shell to be treated equivalently. This equivalence is what introduces the ambiguity in the treatment of the open shell: the open-shell Kramers pair must behave as both a particle and a hole, and the result is that the truncated commutator expansions in the coupled-cluster equations are much longer than in closed-shell theory. The alternative is to use an unrestricted wave operator with the Kramers-restricted spinors. The use of the latter provides some reduction in the work due to the relations between the integrals, but not a full reduction (Visscher et al. 1996).

The same considerations also apply to the case of two open shells where the product of the fermion irreps for the open shells belongs to a doubly degenerate boson irrep. In this case the reference is a single determinant, related to its partner by the time-reversal operator. Because there is no symmetry between the open shells, we cannot derive relations between the amplitudes for Kramers partners.

However, if the product of the fermion irreps for the open shells belongs to a singly degenerate boson irrep, the wave function is symmetric under time reversal, and hence there ought to be a way of devising a Kramers-restricted method. We write the reference as

$$\Phi_0 = (\Phi + \bar{\Phi})/\sqrt{2} \tag{12.33}$$

where Φ and $\bar{\Phi}$ are the two open-shell determinants related by

$$\hat{\mathcal{K}}\Phi = \bar{\Phi}; \qquad \hat{\mathcal{K}}\bar{\Phi} = \Phi. \tag{12.34}$$

The correlated wave function is

$$\Psi = e^{\hat{T}}\Phi_0. \tag{12.35}$$

Applying the time-reversal operator, we find that

$$\left[\hat{\mathcal{K}}, e^{\hat{T}}\right] = 0 \tag{12.36}$$

because $\hat{\mathcal{K}}\Psi = \Psi$ and $\hat{\mathcal{K}}\Phi_0 = \Phi_0$. The relations between the amplitudes follow, just as for the closed-shell case. However, the same problem exists as for the open-shell doublet case: the $\hat{T}$ operator contains excitations into and out of the two open shells.

One way to address this problem is to treat the two configurations as a model space, and use the approach of Jeziorski and Monkhorst (1981) to determine the correlated wave function and energy. In this approach, the reference determinants are correlated independently with the unrestricted wave operator form, except that excitations internal to the open-shell space are excluded. Thus, the $\hat{T}_1$ operator is

$$\hat{T}_1 = t_i^a \hat{E}_i^a + t_i^u \hat{E}_i^u + t_v^a \hat{E}_v^a \tag{12.37}$$

where u and v are occupied and unoccupied open-shell spinor indices, but it does not include the term $t_u^v \hat{E}_u^v$. The two reference determinants are related by time-reversal symmetry according to (12.34), and give rise to two correlated functions,

$$\Psi = e^{\hat{T}}\,\Phi; \qquad \bar{\Psi} = e^{\bar{T}}\,\bar{\Phi}; \tag{12.38}$$

which are also related by time reversal,

$$\hat{\mathcal{K}}\Psi = \bar{\Psi}; \qquad \hat{\mathcal{K}}\bar{\Psi} = \Psi. \tag{12.39}$$

It follows that $\hat{\mathcal{K}}\hat{T} = \bar{T}$. We therefore only need to determine a single set of amplitudes in an unrestricted CC calculation with a Kramers-restricted basis. The energy is obtained from a 2 × 2 diagonalization in the correlated model space. Since the two correlated reference functions are degenerate, no diagonalization is strictly necessary.

As a bonus, we get the energy for both of the states that are formed by taking the sum or difference of the reference functions,

$$\Psi_\pm = (\Psi \pm \bar{\Psi})/\sqrt{2}. \tag{12.40}$$

Substituting for the correlated references, we have

$$\hat{\mathcal{H}}(e^{\hat{T}}\Phi \ \pm \ e^{\bar{T}}\bar{\Phi}) = E_\pm(e^{\hat{T}}\Phi \ \pm \ e^{\bar{T}}\bar{\Phi}) \tag{12.41}$$

Projecting on the left with $\Phi e^{-\hat{T}}$, we get

$$\langle \Phi | e^{-\hat{T}}\hat{\mathcal{H}}e^{\hat{T}} | \Phi \rangle \ \pm \ \langle \Phi | e^{-\hat{T}}\hat{\mathcal{H}}e^{\bar{T}} | \bar{\Phi} \rangle = E_\pm \ (1 \ \pm \ \langle \Phi | e^{-\hat{T}}e^{\bar{T}} | \bar{\Phi} \rangle). \tag{12.42}$$

Because Φ and $\bar{\Phi}$ are related by a double excitation in the open-shell space, and because we left excitations within this space out of the excitation operators, the second part of the normalization term is zero, and the energy is given by the left side of the equation. This technique can be used for open-shell Kramers pairs belonging to complex or real irreps, but not to quaternion irreps. In the last case, there are four determinants that are composed of the open-shell spinors, and even though they occur in pairs related by time-reversal symmetry, the Hamiltonian operator connects all four. In the case of complex irreps, the absolute value of the off-diagonal matrix element must be taken, because it will in general be complex.

For systems with more than two open shells, it is in general necessary to resort to multireference methods. This section has dealt only with state-specific coupled-cluster methods, also known as state-universal methods or Hilbert space methods, for which a considerable amount of effort has been expended on nonrelativistic multireference methods.[3] The alternative, which is much more suited to multireference problems, is the valence-universal or Fock space method, which has been developed for relativistic systems by Kaldor and coworkers (Eliav and Kaldor 1996, Eliav et al. 1994, 1998, Visscher et al. 2001).

12.7 Configuration Interaction Expansions

We have already pointed out that the situations in which a single-determinant wave function can be used as a reference for a correlated method are much fewer in relativistic theory than in nonrelativistic theory. We must therefore resort to methods that do not assume a single-determinant reference. Whether we treat dynamical correlation perturbationally or variationally, it is usually the case that we need to obtain eigenfunctions of some N-particle Hamiltonian, and so we turn to CI methods.

Once a set of transformed integrals is available, there is a large variety of nonrelativistic methods to solve the CI problem. Most of these can also be applied to

---

[3] See for example the book by Hoffmann and Dyall (2002).

the relativistic case with small modifications. In following sections, we examine in more detail the relativistic CI and MCSCF methods. Some of the earliest work in this area was done by Esser and coworkers (Esser et al. 1981, Esser 1984a, 1984b) who developed multireference CI methods taking time-reversal symmetry into account, and developing the machinery of the Graphical Unitary Group Approach (GUGA) for the relativistic case.

We begin in this section with a brief overview of the main principles of modern CI methods. Here, we assume that we want a single eigenstate, but it is easy to generalize to a set of eigenstates.

In the early days of CI calculations, the Hamiltonian matrix was constructed and diagonalized directly to yield all the eigenvectors and eigenvalues. The fact that diagonalization scales as M^3, where M is the number of determinants in the CI expansion, soon exposed the limitations of this approach. The limitations are even more severe in the relativistic case because of the greater number of determinants for the same one-particle space. Moreover, it is rarely the case that one is interested in all M eigenstates. Iterative methods were therefore developed for extracting the lowest few eigenstates by Davidson (1975a), which reduced the scaling from M^3 to M^2. The method works as follows.

The CI wave function Ψ is expanded in a set of N-particle functions Φ_i

$$\Psi = \sum_i C_i \Phi_i \tag{12.43}$$

giving rise to the matrix eigenvalue problem

$$\mathbf{H}\,\mathbf{C} = E\,\mathbf{C} \tag{12.44}$$

where $\mathbf{H}$ is the Hamiltonian matrix, $\mathbf{C}$ is the coefficient (column) vector, and E is the eigenvalue. The essence of the Davidson method is to expand the coefficient vector in a set of basis vectors that are developed with some physical motivation, as follows.

The first step is to make a guess, $\mathbf{C}_0$, at the eigenvector of interest. For instance, for the ground state we may use the reference configuration as a first guess. Using this guess, we generate the vector[4] $\mathbf{a}$, given by

$$\mathbf{a} = \mathbf{H}\,\mathbf{C}_0. \tag{12.45}$$

We can project this vector onto the original guess $\mathbf{C}_0$

$$D = \mathbf{C}_0^T \cdot \mathbf{a} = \mathbf{C}_0^T\,\mathbf{H}\,\mathbf{C}_0 = E_0. \tag{12.46}$$

---

[4] This vector is normally called $\boldsymbol{\sigma}$, but to avoid confusion with the Pauli matrices we have chosen a different notation in the present context. Likewise, the residue vector $\mathbf{b}$ commonly appears as $\mathbf{r}$.

D is in fact the energy of the initial guess vector. We can partition the eigenvector, the energy, and the Hamiltonian as we would in perturbation theory (but without the perturbation parameter),

$$\mathbf{H} = \mathbf{H}_0 + \mathbf{H}'; \qquad E = E_0 + \delta E; \qquad \mathbf{C} = \mathbf{C}_0 + \delta\mathbf{C}, \tag{12.47}$$

where $\mathbf{H}_0$ is the diagonal of $\mathbf{H}$. This is usually a good approximation because $\mathbf{H}$ is diagonally dominant. Here, however, we will not be using perturbation theory directly, but rather deriving a correction vector that we can use to develop a basis for solving the eigenvalue equation. Substituting into the eigenvalue equation and rearranging, we can write

$$(\mathbf{H} - E_0)\,\mathbf{C}_0 = (E_0 - \mathbf{H}_0)\,\delta\mathbf{C} \;+\; \delta E\,\mathbf{C}_0 \;+\; (\delta E - \mathbf{H}')\,\delta\mathbf{C}. \tag{12.48}$$

We define the left-hand side as the *residue vector* **b** by

$$\mathbf{b} = \mathbf{a} - E_0\mathbf{C}_0 \tag{12.49}$$

This vector is orthogonal to $\mathbf{C}_0$. If we ignore the second term on the right-hand side of (12.48) because of its orthogonality to **b**, and ignore the third term because it is small, we get an expression for the correction vector $\mathbf{q} = \delta\mathbf{C}$,

$$\mathbf{q} = (E_0 - \mathbf{H}_0)^{-1}\mathbf{b} \tag{12.50}$$

or in terms of its matrix elements

$$q_i = b_i/(H_{ii} - D). \tag{12.51}$$

Note that **q** is essentially the first-order correction to the wave function in the perturbation problem represented by the partitioning above. We now orthogonalize this vector to $\mathbf{C}_0$, yielding a new vector $\mathbf{C}_1$. This new vector in turn may be used to calculate a new **a** vector, and the **H** matrix is now diagonalized over the space of the vectors $\mathbf{C}_0$ and $\mathbf{C}_1$ and a new correction vector calculated. The algorithm runs iteratively until the norm of the residue vector falls below a given threshold. Each iteration can be considered as adding the next order in perturbation theory.

The efficiency supplied by the Davidson method is that the main work is in the matrix–vector multiplications, which scales as M^2, rather than the M^3 of direct diagonalization. The biggest problem is the storage of the Hamiltonian matrix, which can be written to disk and read in row by row, or in batches of matrix elements if it is sparse. Thus, we do not need to keep the Hamiltonian matrix in memory to obtain its eigenvectors.

It was not long before it was realized that even evaluating and storing the Hamiltonian placed too severe limits on the size of the CI expansion. For a million determinants, the number of Hamiltonian matrix elements is 10^{12}—far too large to store on disk and read in each iteration, even if the matrix was sparse and the sparsity could be exploited. The Hamiltonian is, however, composed of integrals over the molecular orbital basis, of which there are n^4, where n is the number of basis functions. This number is usually a lot smaller than the number of Hamiltonian matrix elements, so a method in which the integrals were stored and used directly to construct the product of the Hamiltonian matrix with a vector would be preferable. This is achieved in the so-called "direct CI" method, developed by Siegbahn (1984) in a CSF basis and extended to a determinant basis by Knowles and Handy (1984).

The key to the method is to use the second-quantized version of the Hamiltonian, and express the Ith component of the **b** vector as follows:

$$\begin{aligned} b_I &= \sum_J \left[\sum_p q \langle I \,|\, \hat{E}_{pq} \,|\, J \rangle h_{pq} + \sum_{pqrs} \langle I \,|\, \hat{e}_{pqrs} \,|\, J \rangle (pq|rs) \right] C_J \\ &= \sum_{pq} h_{pq} \langle I \,|\, \hat{E}_{pq} \,|\, J \rangle \; C_J + \sum_{pqrs} (pq|rs) \sum_J \langle I \,|\, \hat{e}_{pqrs} \,|\, J \rangle \; C_J . \end{aligned} \tag{12.52}$$

The formation of the **b** vector therefore reduces to a contraction of the one- and two-electron integrals with the one- and two-particle density matrices,

$$D^{IJ}_{pq} = \langle I \,|\, \hat{E}_{pq} \,|\, J \rangle ; \qquad P^{IJ}_{pqrs} = \langle I \,|\, \hat{e}_{pqrs} \,|\, J \rangle \tag{12.53}$$

Unfortunately, both of these matrices are much larger than the Hamiltonian matrix itself—they are of order M^2n^2 and M^2n^4—so we haven't gained anything unless these matrices can be evaluated extremely efficiently, and preferably not stored.

The first step is to expand the two-particle density matrix in terms of the one-particle density matrix, which we can do by expanding $\hat{e}_{pqrs}$ and inserting a resolution of the identity:

$$\begin{aligned} P^{IJ}_{pqrs} &= \langle I \,|\, \hat{E}_{pq} \hat{E}_{rs} - \delta_{qr} \hat{E}_{ps} \,|\, J \rangle \\ &= \sum_K \langle I \,|\, \hat{E}_{pq} \,|\, K \rangle \langle K \,|\, \hat{E}_{rs} \,|\, J \rangle - \delta_{qr} \langle I \,|\, \hat{E}_{ps} \,|\, J \rangle \\ &= \sum_K D^{IK}_{pq} D^{KJ}_{rs} - \delta_{qr} D^{IJ}_{ps} . \end{aligned} \tag{12.54}$$

At the expense of an extra summation, which introduces a step of order M^2, we have reduced the two-particle density matrix to a sum involving one-particle density matrices. The CI problem therefore reduces to the problem of evaluating the one-particle density matrices efficiently.

It is at this point that the N-particle basis comes into play. If we choose a determinant basis, or can transform the basis to a set of determinants, we can also factorize the states, and further reduce the CI problem. Since we want to apply this method to the relativistic case, we take as our N-particle basis the determinant basis introduced in chapter 9, given in terms of A and B strings, which we will here consider to take the role of the α and β strings of nonrelativistic theory. We also have to expand the excitation operators into their spin components,

$$\hat{E}_{pq} = a_p{}^{\dagger} a_q \; + a_{\bar{p}}{}^{\dagger} a_{\bar{q}} \; = \hat{E}^A_{pq} + \hat{E}^B_{pq} \tag{12.55}$$

For a given number of α and β electrons, we can write the coefficient vector as a matrix, indexed by A strings and B strings,

$$\mathbf{C} \equiv \mathbf{C}(I^A, J^B) \tag{12.56}$$

where I^A indexes A strings and J^B indexes B strings.

We now expand the states I and J in terms of the strings. The one-particle density matrix becomes

$$\begin{aligned} D^{IJ}_{pq} &= \langle I^A I^B \,|\, \hat{E}^A_{pq} + \hat{E}^B_{pq} \,|\, J^A J^B \rangle \\ &= \delta_{I^B, J^B} \langle I^A \,|\, \hat{E}^A_{pq} \,|\, J^A \rangle + \delta_{I^A, J^A} \langle I^B \,|\, \hat{E}^B_{pq} \,|\, J^B \rangle \\ &\equiv \delta_{I^B, J^B} D^{IJ,A}_{pq} + \delta_{I^A, J^A} D^{IJ,B}_{pq}. \end{aligned} \tag{12.57}$$

If $n_\alpha = n_\beta$, the number of strings is $\sqrt{M}$, and the density matrices over the strings are now only of order $n^2 M$, which is easily manageable, especially in a loop over the orbital indices. Moreover, from Slater's rules, these density matrices are sparse and reduce to a sign that comes from the commutation of the creation and annihilation operators. They are therefore very easy to evaluate and can be calculated as needed.

The transition to the relativistic case is straightforward, but it brings with it extra considerations and some complications. The structure of the relativistic determinants in terms of A and B strings is identical to the structure of the nonrelativistic determinants, so there is no change there. Likewise, the creation and annihilation operators have the same behavior whether we interpret them as operators for α and β spin–orbitals or members of a Kramers pair. We can also replace the $\hat{E}$ and $\hat{e}$ operators with the $\hat{X}$ and $\hat{x}$ operators. What differs is the integrals, the types of excitation operators, and the structure of the Hamiltonian. The relativistic case has classes of integrals that are zero in the corresponding nonrelativistic case, which correspond to operators with odd numbers of bars for a given electron coordinate. The Hamiltonian is block-pentadiagonal and covers the full range of determinants, whereas in the nonrelativistic case the off-diagonal blocks are zero and we only need to consider a single block.

The main complication arises when we consider the off-diagonal blocks. The blocks on the first off-diagonal have contributions from the odd-bar integrals and operators and the blocks on the second off-diagonal have contributions from integrals with two bars. The $\hat{x}$ operators in both cases can be reduced to expressions that contain $X^{\mathrm{s}}_{p\bar{q}}$ and $X^{\mathrm{s}}_{\bar{p}q}$ operators. These operators connect A and B strings, and so we lose the benefit of the factorization of the determinants into A and B strings.

What we can do about this complication is to factorize the density matrix elements into matrix elements over a single creation or annihilation operator. For the one-particle density matrix over $X^{\pm}_{p\bar{q}}$, the factorization falls out easily:

$$\begin{aligned} D^{IJ,\pm}_{p\bar{q}} &= \langle I\,|\,\hat{X}^{\pm}_{p\bar{q}}\,|\,J\,\rangle = \langle I^A I^B\,|a_p{}^{\dagger} a_{\bar{q}} \mp a_q{}^{\dagger} a_{\bar{p}}\ \,|\,J^A J^B\,\rangle \\ &= \Big[\langle I^A\,|a_p{}^{\dagger}\,|\,J^A\,\rangle\langle I^B\,|a_{\bar{q}}\ \,|\,J^B\,\rangle \\ &\quad \mp \langle I^A\,|a_q{}^{\dagger}\,|\,J^A\,\rangle\langle I^B\,|a_{\bar{p}}\ \,|\,J^B\,\rangle\Big]. \end{aligned} \tag{12.58}$$

We can also reduce the density matrices for the $X^{\pm}_{pq}$ operators with the introduction of another resolution of the identity:

$$\begin{aligned} D^{IJ,\pm}_{pq} &= \langle I\,|\,\hat{X}^{\pm}_{pq}\,|\,J\,\rangle = \langle I^A I^B\,|a_p{}^{\dagger} a_q \pm a_{\bar{q}}{}^{\dagger} a_{\bar{p}}\ \,|\,J^A J^B\,\rangle \\ &= \Big[\delta_{I^B,J^B}\sum_{K^A}\langle I^A\,|a_p{}^{\dagger}\,|\,K^A\,\rangle\langle K^A\,|a_q|\,J^A\,\rangle \\ &\quad \pm \delta_{I^A,J^A}\sum_{K^B}\langle I^B\,|a_{\bar{q}}{}^{\dagger}\,|\,K^B\,\rangle\langle K^B\,|a_{\bar{p}}\,|\,J^B\,\rangle\Big]. \end{aligned} \tag{12.59}$$

The strings on each side of these matrix elements differ by one electron, and for each string on one side of the matrix element there is only one string that is connected to it by the creation or annihilation operator. This matrix is therefore very sparse: instead of being order nM, it is only of order $n\sqrt{M}$. The matrix elements themselves must have the value ±1, determined entirely by the number of permutations required to move the extra electron to one end of the string. All that is needed to evaluate the matrix elements is a list of pointers from the parent string to the daughter string for each annihilation operator and the sign change for the permutation. The number of quantities stored is therefore only $2n\sqrt{M}$ integers. This approach was pioneered in the nonrelativistic context by Knowles and Werner (1988). In atomic structure theory, these matrix elements are called "coefficients of fractional parentage." Here we can simply call them *parentage coefficients*, because they describe the connection between a parent and a daughter state.

We expanded the $\hat{X}$ operators into the primitive products of creation and annihilation operators to achieve this factorization. Instead of first expanding the $\hat{x}$ in terms of $\hat{X}$, we can expand them directly in the primitive products of creation and annihilation operators, to obtain the sums of products of parentage coefficients. As an example, consider the second off-diagonal block, for which the two-particle density matrix factorizes naturally:

$$\langle I^A I^B\,|a_p{}^{\dagger} a_r{}^{\dagger} a_{\bar{s}}\ a_{\bar{q}}\ \,|\,J^A J^B\,\rangle = \langle I^A\,|a_p{}^{\dagger} a_r{}^{\dagger}\,|\,J^A\,\rangle\langle I^B\,|a_{\bar{s}}\ a_{\bar{q}}\ \,|\,J^B\,\rangle. \tag{12.60}$$

The coefficients in this expression are "grandparentage" coefficients, and could be stored instead of using the parentage coefficients, at a cost of $2n^2\sqrt{M}$. The contribution to the **b** vector is then

$$b(I^A, I^B) = \sum_{pqrs} (p\bar{q}|r\bar{s}) \sum_{J^A J^B} \langle I^A | a_p{}^\dagger a_r{}^\dagger | J^A \rangle \langle I^B | a_{\bar{s}}\, a_{\bar{q}}\ | J^B \rangle\ \ C(J^A, J^B). \tag{12.61}$$

Evaluation of these contributions can be done at a cost of the order of $n^4 M$ operations.

The second complication in relativistic direct CI arises because we have to consider all possible determinants, from those represented entirely by A strings through products of A and B strings, to those represented entirely by B strings. The most efficient strategy may vary from one block of the Hamiltonian to the next. However, each block can be treated independently, and therefore the work can be performed in pieces that do not exceed the size of the corresponding nonrelativistic problem.

12.8 The Cost of Configuration Interaction Methods

Earlier we compared the cost of nonrelativistic and relativistic methods for perturbation and coupled-cluster calculations, and also the cost of transforming the integrals. We now turn to the cost of CI methods, and consider the case of a full CI, which is the basis of the complete active space (CAS) SCF method, and of a singles and doubles CI (SDCI) calculation, which is often used for dynamic correlation from a given active space.

For this purpose we use the determinant basis introduced earlier (chapter 9), given in terms of A and B strings, and consider all possible determinants that may be constructed from a given set of Kramers pairs. We group these determinants into subsets with a given value of N_A and N_B, characterized by $M_K = \frac{1}{2}(N_A - N_B)$.

In the nonrelativistic case, the CI expansion is limited to a given $M_K \equiv M_S$ value due to spin factorization. In this case the length of a full CI expansion, in the absence of symmetry, is $\binom{n}{N_\alpha}\binom{n}{N_\beta}$, where n is the number of active orbitals. In the Kramers-restricted relativistic case, the CI includes all possible M_K values because of the spin–orbit interaction. The length of a full CI expansion is therefore $\binom{2n}{N}$.

The lengths of the N-particle bases for relativistic and nonrelativistic calculations for the case of 12 electrons in 12 active orbitals is presented in table 12.1, which clearly demonstrates the size problems connected with the relativistic full CI expansion. In the best possible case, for $M_S = 0$, the relativistic full CI vector is only three times the length of the nonrelativistic vector.

The immediate consequence of this increased number of determinants is an increase of the size of the Hamiltonian matrix. For the $M_S = 0$ case, the relativistic Hamiltonian matrix has nine times as many elements as in the nonrelativistic Hamiltonian matrix. We would expect the work in multiplying the Hamiltonian matrix by a vector (the time-consuming step in modern CI procedures, as discussed in the previous section) to increase by a factor of 36:9 from the size of the Hamiltonian matrix and 4 from the complex arithmetic.

In section 10.6, we showed that the relativistic Hamiltonian is block pentadiagonal, so the scaling of the number of matrix elements is actually less than the square of the

Table 12.1 Length of *N*-particle basis for 12 electrons in 12 orbitals

M_S	*Determinants*	*Ratio R/NR*
0	853776	3
1	627264	4
2	245025	11
3	48400	55
4	4356	629
5	144	18778
6	1	2704156

Symmetry is not used. The second column is the number of determinants in the nonrelativistic (NR) expansion. The number of determinants in the relativistic (R) expansion is the same for each M_S at 2704156.

R/NR ratio—in this case about 75% of the ratio. The symmetry of the Hamiltonian about both diagonals is also found in the nonrelativistic case by a suitable ordering of the determinants, so the gains from Hermiticity and time reversal on the matrix itself are the same in both cases.

The scaling can be reduced by constructing a real basis according to (9.58). The Hamiltonian matrix is still block pentadiagonal and symmetric but it lacks the symmetry about the other diagonal. The reduction in work is therefore only a factor of 2. Taking into account the 75% due to the block pentadiagonal character of the Hamiltonian, the overall increase in work is about a factor of 14.

The second example of the cost of CI methods is the case of double excitations from a closed-shell determinant with no spatial symmetry. In the nonrelativistic case, there are four excitations for each pair of occupied orbitals i,j and virtual orbitals a,b. These arise from singlet and triplet coupling of the holes in the occupied orbitals, which must match the coupling of the virtual orbitals, to give an overall spin singlet. In the relativistic case, all of the 16 possible excitations from an occupied Kramers pair to a virtual Kramers pair survive, because in the absence of spatial symmetry all products of an even number of spinors generate the A representation. The expansion length is therefore four times longer in the relativistic case. The extra excitations correspond to products of the singlet and triplet spin functions that yield an overall spin triplet and spin quintet, and these are connected by the spin–orbit interaction. The Hamiltonian matrix is 16 times larger, and the work is therefore 64 times the work of a nonrelativistic CI, calculation, due to the complex arithmetic. Following the same line of reasoning as for the full CI construction of a real basis reduces the work to a factor of 32. This is the same factor as for CCSD calculations, which is not surprising.

12.9 Relativistic Multiconfiguration Self-Consistent Field Theory

The special importance of multiconfiguration SCF (MCSCF) methods in relativistic calculations was discussed in the introduction to this chapter. But quite apart from this, we know that MCSCF methods are very useful in nonrelativistic quantum chemistry, where various brands of the method, such as *complete active space* (CAS) SCF, *restricted active space* (RAS) SCF, and the *general valence bond* (GVB) methods, have had considerable success and impact. It is therefore clearly desirable to develop a relativistic approach to MCSCF calculations. It is obvious that the conclusions about relativistic multireference expansions may be applied directly to the CI part of the MCSCF case. In addition to this, we must also give special consideration to the treatment of time-reversal symmetry as well as the possibility of complex expansion and spinor rotation parameters. Here we will only touch rather briefly on some of the underlying principles of relativistic MCSCF calculations. We will not provide extensive technical descriptions in terms of formulas and algorithms for a Kramers-restricted MCSCF scheme as these may be found in Jensen et al. (1996).

The general MCSCF iteration starts with an orthonormal one-particle basis and a wave function calculated through a number of previous iterations (or from some suitable initial guess). If we assume that the process has completed iteration number k, the wave function at this stage may be written as $\Psi^{(k)}$. The (optimal) MCSCF function is derived from this by means of orbital rotations in the one-electron basis and expansion in the set of N-electron states. In parametrized form we may write this as

$$\Psi(\mathbf{C}\boldsymbol{\kappa}) = e^{-\hat{\kappa}} \left[\Psi^{(k)} + \hat{\mathcal{P}} \sum_{\mu} C_{\mu} \Phi_{\mu} \right] S^{-1/2} \tag{12.62}$$

where $\boldsymbol{\kappa}$ represents the orbital rotations, $\mathbf{C}$ the expansion coefficients for the N-particle basis, and $\hat{\mathcal{P}}$ is a projection operator that maintains the N-particle variational space orthogonal to the current wave function, that is

$$\hat{\mathcal{P}} = 1 - |\Psi^{(k)}\rangle\langle\Psi^{(k)}|. \tag{12.63}$$

S is a normalization factor arising from the variation in the N-particle space

$$S = 1 + \sum_{\mu,\nu} C_{\mu}^{*} C_{\nu} \langle \Phi_{\mu} | \Phi_{\nu} \rangle. \tag{12.64}$$

In line with what we have said previously, we must in general expect the orbital mixing coefficients κ_{rs} to be complex and the $\boldsymbol{\kappa}$ matrices to be anti-Hermitian. In a Kramers-restricted formalism the operator $e^{-\hat{\kappa}}$ must be invariant under time reversal, and we can incorporate this requirement by expressing the elements of $\boldsymbol{\kappa}$ in terms of the $\hat{X}^{\pm}$ operators:

$$\begin{aligned} \hat{\kappa} &= \sum_{r,s} \left(\kappa_{rs} \hat{X}^{-}_{rs} + \tfrac{1}{2} \kappa_{r\bar{s}} \hat{X}^{-}_{r\bar{s}} + \tfrac{1}{2} \kappa_{\bar{r}s} \hat{X}^{-}_{\bar{r}s} \right) \\ &= \sum_{r<s} \left(\kappa_{rs} \hat{X}^{-}_{rs} + \kappa_{r\bar{s}} \hat{X}^{-}_{r\bar{s}} - \kappa^{*}_{rs} \hat{X}^{-}_{sr} - \kappa^{*}_{r\bar{s}} \hat{X}^{-}_{\bar{s}r} \right) + \tfrac{1}{2} \sum_{r} \left(\kappa_{r\bar{r}} \hat{X}^{-}_{r\bar{r}} - \kappa^{*}_{r\bar{r}} \hat{X}^{-}_{\bar{r}r} \right). \end{aligned} \tag{12.65}$$

Note that with an orthonormal one-particle basis, and time-reversal symmetry imposed in this manner on the variational space, it is possible to use unconstrained optimization techniques for parameter variation. Further simplifications of the $\boldsymbol{\kappa}$ matrix arises for the case that we have real or complex groups, as described previously. Redundant rotations are, of course, excluded from the variational space. The nonredundant rotations include rotations between occupied spinors and negative-energy spinors, as discussed in chapter 8. However, if a good starting guess is available, such as from a DHF calculation, neglecting the rotations with the negative-energy spinors is likely to be a good approximation.

We now collect the variational parameters in a column vector $\boldsymbol{\lambda}$, that is

$$\boldsymbol{\lambda} = \left[\mathbf{C}, \boldsymbol{\kappa}, \mathbf{C}^*, \boldsymbol{\kappa}^*\right]^T, \tag{12.66}$$

where we have also included the complex conjugates of the parameters to account for the two degrees of variational freedom represented by the real and imaginary parts of a complex number. In this variational space our current wave function $\Psi^{(k)}$ corresponds to $\boldsymbol{\lambda} = [\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}]^T$. The total energy of the MCSCF wave function may now be written as

$$E(\boldsymbol{\lambda}) = \mathcal{E}(\boldsymbol{\lambda}) S^{-1} \tag{12.67}$$

where, in agreement with standard practice (Jørgensen and Simons 1981), we can expand $\mathcal{E}(\boldsymbol{\lambda})$ as

$$\begin{aligned}\mathcal{E}(\boldsymbol{\lambda}) &= \Big\langle \Big(\sum_{\mu} C_\mu \Phi_\mu\Big)\mathcal{P} + \Psi^{(k)} \;\Big|\; e^{\hat{\kappa}} \hat{H} e^{-\hat{\kappa}} \;\Big|\; \Psi^{(k)} + \mathcal{P}\Big(\sum_{\nu} C_\nu \Phi_\nu\Big) \Big\rangle \\ &= \Big\langle \Big(\sum_{\mu} C_\mu \Phi_\mu\Big)\mathcal{P} + \Psi^{(k)} \;\Big|\; \hat{H} + \big[\,\hat{\kappa}, \hat{H}\big] + \tfrac{1}{2}\big[\,\hat{\kappa}, \big[\,\hat{\kappa}, \hat{H}\big]\big] + \cdots \;\Big|\; \Psi^{(k)} \\ &\quad + \mathcal{P}\Big(\sum_{\nu} C_\nu \Phi_\nu\Big) \Big\rangle .\end{aligned} \tag{12.68}$$

This formulation requires that the integrals are transformed to the one-electron basis of the current iteration.

We can expand the total energy in a Taylor series in the variational parameter $\boldsymbol{\lambda}$ as

$$E(\boldsymbol{\lambda}) = E^{[0]} + \boldsymbol{\lambda}^\dagger \mathbf{E}^{[1]} + \tfrac{1}{2}\boldsymbol{\lambda}^\dagger \mathbf{E}^{[2]} \boldsymbol{\lambda} + \cdots . \tag{12.69}$$

Here

$$\begin{aligned} E^{[0]} &= \langle \Psi^{(k)} | \hat{H} | \Psi^{(k)} \rangle \\ \mathbf{E}^{[1]} &= \left(\frac{\partial E}{\partial \boldsymbol{\lambda}}\right)_{\boldsymbol{\lambda}=\mathbf{0}} \\ \mathbf{E}^{[2]} &= \left(\frac{\partial^2 E}{\partial \boldsymbol{\lambda}^2}\right)_{\boldsymbol{\lambda}=\mathbf{0}} . \end{aligned} \tag{12.70}$$

We can write $\mathbf{E}^{[1]}$ as a column vector explicitly in terms of the variational parameters,

$$\mathbf{E}^{[1]T} = \left(\frac{\partial E}{\partial \mathbf{C}}, \frac{\partial E}{\partial \boldsymbol{\kappa}}, \frac{\partial E}{\partial \mathbf{C}^*}, \frac{\partial E}{\partial \boldsymbol{\kappa}^*} \right)_{[\mathbf{C},\boldsymbol{\kappa}]=[\mathbf{0},\mathbf{0}]}, \tag{12.71}$$

and similarly $\mathbf{E}^{[2]}$ expands to a 4 × 4 matrix. The elements of $\mathbf{E}^{[1]}$ and $\mathbf{E}^{[2]}$ may be found by straightforward differentiation of the expression for the total energy. For $\mathbf{E}^{[1]}$ we get contributions of the type

$$\begin{aligned}
E_{\mu}^{[1]*} &= \left.\frac{\partial E}{\partial C_{\mu}^*}\right|_{\lambda=0} = \langle \Phi_\mu \mid \hat{H} - E^{[0]} \mid \Psi^{(k)} \rangle \\
E_{rs}^{[1]*} &= \left.\frac{\partial E}{\partial \kappa_{rs}^*}\right|_{\lambda=0} = -\langle \Psi^{(k)} \mid \left[\hat{X}_{sr}^{-}, \hat{H}\right] \mid \Psi^{(k)} \rangle \\
E_{r\bar{s}}^{[1]*} &= \left.\frac{\partial E}{\partial \kappa_{r\bar{s}}^*}\right|_{\lambda=0} = -\langle \Psi^{(k)} \mid \left[\hat{X}_{\bar{s}r}^{-}, \hat{H}\right] \mid \Psi^{(k)} \rangle
\end{aligned} \tag{12.72}$$

where $C_{\nu}^{(k)}$ is the coefficient of CSF Φ_ν in the reference wave function, $\Psi^{(k)}$. We can easily identify in this expression the residue vector from (12.49), which is constructed in a Davidson diagonalization procedure. Likewise the derivatives with respect to the rotation parameters can be written in terms of generalized Fock matrices

$$\begin{aligned}
f_{ut} = &\sum_{q} \left[D_{tq}^{+} h_{uq} + D_{t\bar{q}}^{+} h_{u\bar{q}} \right] + \sum_{qrs} \left[P_{tq,rs}^{++} (uq|rs) + P_{t\bar{q},rs}^{++} (u\bar{q}|rs) \right] \\
&+ \tfrac{1}{2} \sum_{qrs} \left[P_{t\bar{q},\bar{r}s}^{++} (u\bar{q}|\bar{r}s) + P_{t\bar{q},r\bar{s}}^{++} (u\bar{q}|r\bar{s}) + P_{tq,\bar{r}s}^{++} (uq|\bar{r}s) \right. \\
&\left. + P_{tq,r\bar{s}}^{++} (uq|r\bar{s}) \right],
\end{aligned} \tag{12.73}$$

where D_{tq}^{+} and $P_{tq,rs}^{++}$ are the one- and two-particle Kramers-restricted density matrices,

$$D_{tq}^{+} = \langle \Psi^{(k)} | \hat{X}_{tq}^{+} | \Psi^{(k)} \rangle; \qquad P_{tq,rs}^{++} \langle \Psi^{(k)} | \hat{x}_{tq,rs}^{++} | \Psi^{(k)} \rangle. \tag{12.74}$$

It remains to determine the optimum $\boldsymbol{\lambda}$, that is the best values of the variational parameters. There are a number of unconstrained optimization procedures that may be used for this purpose. One possibility is the second-order restricted-step optimization, which takes the Taylor expansion of $E(\boldsymbol{\lambda})$ to second order in $\boldsymbol{\lambda}$ as a starting point. Then the optimal step may be given as

$$\boldsymbol{\lambda}(\xi) = -(\mathbf{E}^{[2]} - \xi \mathbf{I})^{-1} \mathbf{E}^{[1]} \tag{12.75}$$

where ξ is a scaling parameter. If ξ is set equal to 0, this expression is just the optimal step from the Newton–Raphson procedure, but the most important function of ξ is to scale $\boldsymbol{\lambda}$ such that $||\boldsymbol{\lambda}(\xi)|| \leq h$, where h is a trust radius for the optimization in the present iteration.

In nonrelativistic MCSCF schemes applied to ground states of atomic and molecular systems, the goal of the optimization process is to locate a global minimum on the energy surface spanned by the variational parameters. However, for the relativistic case there is a set of negative-energy solutions with energy below the electronic ground state. Therefore, the energy must be maximized with respect to rotations between positive-energy and negative-energy spinors and minimized with respect to all other variational parameters. This means that the MCSCF energy of the electronic ground state really corresponds to a saddle point in the spinor optimization space. The convergence of second-order optimization methods is only guaranteed in searches for minima, but in this case the energy gap between positive- and negative-energy solutions is so large that convergence of the optimization process should be quite safe also for electronic ground states. With a good starting guess, the rotations between positive-energy and negative-energy spinors could safely be neglected, and then the optimization process reduces to a minimization. This approximation corresponds to a projection onto the positive-energy states of the starting guess.

With these provisions the optimization process may proceed in analogy with any of the various schemes developed for nonrelativistic MCSCF. Jensen et al. (1996) have shown in detail how this may be done for one particular algorithm—the norm-extended optimization. The only added complication for the relativistic case arises from the need to use complex arithmetic. The implementation of time-reversal and double-group symmetry follows from the discussions of the symmetry of Fock matrices and of the relativistic many-electron Hamiltonian in earlier chapters.

13

Molecular Properties

Strictly speaking, in quantum mechanics a measurable property is defined as an observable connected to a self-adjoint operator. However, in common usage the term molecular property is loosely taken to mean any physical attribute of a molecule, preferably amenable to experimental measurement. Common examples of properties of interest to chemists are molecular structure, thermodynamic quantities, spectroscopic transition energies and intensities, and various electric and magnetic moments. The amenability to experiment may exist only in principle—one of the strong points of modern computational chemistry is the possibility of studying phenomena occurring under conditions that lie beyond the present experimental capabilities. Sometimes, differential effects between different theoretical models are also regarded as properties: thus the correlation energy is generally considered to be the difference between the Hartree–Fock energy and the energy obtained from a complete many-electron treatment (e.g. full CI or MBPT to all orders). At best only the latter of these is accessible to experiment. Similarly, certain relativistic effects (e.g. bond contraction) only appear as the difference between results from a relativistic and a nonrelativistic calculation.

The calculation of molecular properties in a relativistic framework follows the same principles as for the nonrelativistic case once a wave function or electron density of adequate quality is available. Our aim here is therefore not to provide explicit expressions and formulas for the calculation of a more or less complete catalog of properties. However, in relativistic calculations of molecular properties there are some aspects of the theory that warrant special care and consideration. In particular, we need to know how to handle features such as Lorentz invariance, gauge invariance, and negative-energy states. Moreover, the electric and magnetic fields appear as natural parts of the relativistic Hamiltonian, and we therefore expect that properties involving these may require a different treatment from the nonrelativistic case where terms involving external fields are grafted onto the nonrelativistic Hamiltonian, often based on some reduction or approximation from the relativistic case. For a more comprehensive introduction to the various approaches to molecular properties, we refer the reader to the extensive, mostly nonrelativistic, literature on the subject (Jørgensen and Simons 1986, Hinchliffe 1987, Swanström and Hegelund 1975).

A convenient, if not very rigorous, classification of properties is to divide them into two groups—*intrinsic* and *response* properties. Intrinsic properties are those that derive directly from a field-free Hamiltonian (except for internal electrostatic interaction) and include thermodynamic properties, molecular structure, electronic excitation energies, vibrational energy levels, and so on. Response properties, on the other hand, measure the response of the molecule to an applied electromagnetic field: here we find the various electric and magnetic moments, polarizabilities, NMR shielding and coupling, spectral intensities, and so on. Below, we give a brief discussion of intrinsic properties within a relativistic framework, then we demonstrate how the relativistic theory applies to external electromagnetic fields. For this chapter we do not use atomic units but include physical constants such as $\hbar, e, m$ explicitly, as these give a more direct demonstration of the dimensionality of the properties involved.

13.1 Intrinsic Properties

Up to this point in our treatment of four-component methods, the nuclear framework of the molecule has been of interest mainly as an electrostatic potential determining the behavior of the electrons. If we are going to interpret the positions of the nuclei as a molecular structure, then we should pause briefly and reconsider some of the assumptions made on the way to make sure that the mathematical and physical quantities we are using really cover the concepts that we need in our discussion of chemical properties. Furthermore, we should ensure that our usage and interpretation of these quantities is consistent with relativity within reasonable, well-controlled approximations. The standard quantum chemical model with electrons surrounding a rigid network of nuclear charges, and where the nuclear positions define a potential surface, has been very successful. Under favorable circumstances, calculations within this model are able to predict results in reasonable, and sometimes excellent, agreement with experimental measurements. Unfortunately, the chain of reasoning leading to this model is far from perfect. Even for the nonrelativistic case there are difficulties, and introducing relativity further adds to these problems. The theoretical background for these difficulties extends far beyond what can reasonably be included here.[1] Here we will therefore take a qualitative approach, follow the chain of progressive approximations and point out where the weaknesses and conceptual difficulties are.

In a strict reductionist view there would be no structure. In the standard model of particle theory there would be a system composed of quarks and electrons, and we would have to resort to the full machinery of the standard model for a proper description. Clearly, this would be a case of serious overkill for chemical purposes: at the energies available in chemical reactions, the nuclear forces are not accessible, and internal structure of nuclei becomes uninteresting. So from extensive experience we know that we can consider the nuclei as single, positively charged particles, although we do need to take their finite size into account, as discussed previously—and possibly

---

[1] Some of this background has been summarized by Primas (1980).

some details of their shape, spin, or charge distribution for properties that involve interaction with the nucleus.

An assembly of nuclei and electrons could be described very accurately within QED. There would still just be a cluster of particles, our molecule, and any structure would have to arise out of the dynamics of the system. For reasons pointed out earlier, QED—if viable at all—would be a very expensive path to calculation of the electronic structure and chemical properties of molecules. For electrons, we circumvented this problem by going to a many-particle treatment based on the Dirac equation, as discussed in chapter 5, and we could presumably do the same here for our cluster of electrons and nuclei. In doing this, we choose a Hamiltonian description of the system, but alternative approaches based on a Lagrangian formalism are also possible. In this process we draw a formal distinction between the molecule and the electromagnetic field, which leaves us with the normal Coulomb interactions between the particles in the molecule and the radiation field as an entity external to the molecule.

At this stage we also encounter a difficulty arising from the nature of the relativistic equations. For the nonrelativistic case we would have no problems composing a Schrödinger equation for the cluster of electrons and nuclei. For the relativistic case, however, only particles of spin $\frac{1}{2}$ are described by the Dirac equation. So while we could set up a total Dirac Hamiltonian accounting for all the spin $\frac{1}{2}$ particles of the cluster, we would have to resort to other types of Hamiltonians for describing any nuclei of spin other than $\frac{1}{2}$ (Greiner 1997). The problem must be mainly formal because we know that as far as electronic structure calculations are concerned nuclei behave the same regardless of nuclear spin—at least until we consider hyperfine interactions and NMR parameters. Moreover, the nucleons are spin $\frac{1}{2}$ particles and, at least formally, we could regard the nucleus as a collection of Dirac particles. For simplicity, we will therefore assume that the particles in our cluster are all Dirac particles, and that it is possible to set up an overall Dirac equation describing the behavior of these particles. Using the Pauli Hamiltonian, Bethe and Salpeter (1957) have shown that the error in doing this should be at most c^{-2}.

To make progress towards the solution of this many-particle Dirac equation, we would like to separate out translational and rotational motion and work in a molecule-fixed coordinate system. This is a difficult task even in nonrelativistic quantum mechanics. For the relativistic case it has some extra complications. In the nonrelativistic case, it is mainly the separation of the rotational degrees of freedom that creates problems (Sutcliffe 1980). These problems are further aggravated in the relativistic case by the need to maintain Lorentz invariance (Aaberge 1977). For the relativistic case the transformation to a center-of-mass system also has problems. Due to the equivalence between mass and energy, the interaction between two relativistic particles changes their mass, which in turn affects the transformation to a center-of-mass frame of reference. One of the theories that deal with this effect leads to an extra term in the nuclear potential between two particles. The effect is small: for Au_2 it has been estimated to lead to a slight bond shortening (Grelland 1980). It is not clear how this theory can be extended to systems of more than two heavy particles. For now, the best we can do is to assume that the nonrelativistic procedure for separating out the rotational and translational degrees of freedom also holds for the relativistic description to a good approximation and provides us with a molecule-fixed coordinate system.

Within these approximations, we now have a cluster of particles described in a mass-centered, nonrotating system. How does molecular structure arise out of this? At this point the issues we face are common to both relativistic and nonrelativistic quantum chemistry. To introduce the concept of structure, we have to mathematically divide the particle cluster into a nuclear and an electronic part. In solving the electronic part of the equation, the other, nuclear, part is treated as a classical semi-rigid framework that we can adjust parametrically. The technique most commonly used to achieve this separation is the Born–Oppenheimer approximation, which may be regarded as the lowest order of an expansion about a system of infinite nuclear masses. The problems inherent in the application of this approximation are the same both for relativistic and nonrelativistic models. Whether the same measures should also be taken when the approximation starts to break down has not, to our knowledge, been explored.

Within the Born–Oppenheimer approximation, we still need to know that the nuclear position parameters really correspond to the distances and angles of a classical molecular framework. Our choice of the Coulomb gauge ensures this—the nuclear positions only appear in the electron–nucleus interaction terms, and the derivation of this potential from relativistic field theory shows us that it is indeed the quantities of normal 3-space that appear here. Thus, any potential surface that we might calculate on the basis of the Born–Oppenheimer-separated electronic molecular Dirac equation is indeed spanned by the variations of molecular structural parameters in the usual meaning.

Having recovered the potential surface from the solutions of the Born–Oppenheimer electronic problem, we can now proceed to solve the equation for nuclear motion. The Dirac-type equation for the nuclei can easily be reduced to the corresponding nonrelativistic equation by following the same reduction as we did for taking the nonrelativistic limit of the Dirac equation in section 4.6. Doing this, we abandon all pretense of Lorentz invariance for this part of the system, but we know from experiment that the nuclear relative motion in molecules takes place at rather low energies where relativistic effects may safely be neglected.

Through this somewhat qualitative chain of reasoning, we have recovered the concept of chemical structure closely connected to that arising from the nonrelativistic treatment. The same reasoning would, of course, apply to both the ground and excited states, and therefore we would also recover the concept of excitation energies. Due to spin–orbit splitting and relativistic shifts of total energies, the relativistic spectra would differ from those calculated in the nonrelativistic approximation, but the underlying concepts would be the same.

Thus, for intrinsic properties of molecules, the relativistic treatment in a four-component formalism can be made completely analogous to the nonrelativistic treatment. This conclusion is based on a series of approximations, some of which may be hard to justify stringently. We can take some comfort in the fact that the models introduced along the way are physically simple and intuitive, such as the separation of rotational motion, and the Born–Oppenheimer approximation, and that most of the neglected effects may quite reasonably be expected to remain insignificant, or at the least, treatable by similar methods to the nonrelativistic case. The success of the final model in producing results that agree with experiment helps ease any doubts we might have about the underlying theory, but it should not disguise the approximate nature of the model we are using. We do expect the treatment to fail for those cases where the nonrelativistic model also has problems, such as in the treatment of “floppy molecules” where the Born–Oppenheimer approximation does not hold. In addition, we expect that there may be problems arising from the specific relativistic features of the chain of approximations above. For intrinsic properties this may include terms from the frame transformation that would affect geometries, as in the case of Au_2 quoted above. For response properties we need to be especially wary of any steps that involve neglect of strict gauge invariance.

In addition to scattering and diffraction methods for structure determination, important experimental probes for intrinsic properties are vibrational and rotational spectroscopy. Rotational spectra will be affected by a relativistic reduction of bond length, which will reduce the moments of inertia. This lowers the rotational constant, and we should expect a relativistic red-shift of the rotational spectrum. For vibrational spectroscopy, the situation is less clear—relativistic effects may strengthen as well as weaken bonds. Thus effects of relativity on vibrational spectroscopy depend very much on the system under consideration. A further discussion of these effects is therefore postponed to chapter 22. For the diffraction and scattering techniques, relativistic effects are absorbed into atomic scattering parameters and structure factors and are thus not a primary concern of relativistic quantum chemistry.

Finally, the shapes and topology of potential surfaces are also affected by relativity. This will influence reaction energies and reaction rates. Here, spin–orbit splitting can play an important role. While the splitting may be energetically small, it still becomes significant as two surfaces approach and could affect barriers and rates. This is a large field in itself, and will only be represented by a few examples in chapter 22.

13.2 Electric Properties

Most studies of molecular properties are founded on the response of the molecule to electromagnetic fields. A relativistic electronic system is described by the time-independent Dirac equation

$$c\boldsymbol{\alpha} \cdot (-i\hbar\nabla + e\mathbf{A})\,\psi + \beta mc^2\psi - e\phi\psi = E\psi \tag{13.1}$$

where the electromagnetic field appears directly. We need to examine the theoretical description of molecular response properties to ensure that this is consistent with special relativity within the approximations and assumptions discussed in the previous section. We will start by examining the case of a time-independent electric field assuming the Coulomb gauge:

$$\nabla \cdot \mathbf{A} = 0. \tag{13.2}$$

We will mostly stay with the simple form of the Dirac equation above, although this only strictly applies to one-electron molecules. The extension to many-electron systems poses no major difficulties.

The electric properties are deduced from the response of the system to an electric field $\mathbf{E}$, which is related to the scalar potential ϕ by

$$\mathbf{E} = -\nabla\phi. \tag{13.3}$$

Thus, in terms of this field, we can write the Dirac equation as

$$\hat{\mathcal{H}}\psi = \left[c\boldsymbol{\alpha}\cdot(-i\hbar\nabla + e\mathbf{A}) + \beta mc^2\psi + e\int \mathbf{E}\cdot d\mathbf{r}\right]\psi = E\psi. \tag{13.4}$$

The simplest electric property of a neutral molecule is the dipole moment, which may be defined in terms of the derivative of the energy at $\mathbf{E} = 0$ as

$$\boldsymbol{\mu} = -\frac{\mathrm{d}E}{\mathrm{d}\mathbf{E}} \tag{13.5}$$

using the standard definition of differentiation by a vector. In this case, the field is a uniform field, $|\mathbf{E}| = \text{const}$.

According to the Hellmann–Feynman theorem, we have

$$\frac{\mathrm{d}E}{\mathrm{d}\mathbf{E}} = \left\langle \frac{\partial\hat{\mathcal{H}}}{\partial\mathbf{E}} \right\rangle \tag{13.6}$$

and the dipole moment may therefore be expressed as

$$\boldsymbol{\mu} = -\left\langle \frac{\partial\hat{\mathcal{H}}}{\partial\mathbf{E}} \right\rangle = \left\langle \frac{\partial}{\partial\mathbf{E}}\left[e\int \mathbf{E}\cdot d\mathbf{r}\right] \right\rangle = \langle e\mathbf{r}\rangle. \tag{13.7}$$

Thus, the dipole moment expression as derived from the Dirac equation is the same as for the nonrelativistic case.

The potential ϕ may be expanded in a power series about a point $\mathbf{r}_0$. The first terms of this expansion are

$$\begin{aligned}\phi(\mathbf{r}) &= \phi(\mathbf{r}_0) + (\nabla\phi)_0\cdot(\mathbf{r}-\mathbf{r}_0) + \frac{1}{2}(\mathbf{r}-\mathbf{r}_0)^T\cdot(\nabla(\nabla\phi))_0\cdot(\mathbf{r}-\mathbf{r}_0) + \cdots \\ &= \phi(\mathbf{r}_0) - \mathbf{E}(\mathbf{r}_0)\cdot(\mathbf{r}-\mathbf{r}_0) - \frac{1}{2}(\mathbf{r}-\mathbf{r}_0)^T\cdot(\nabla\mathbf{E})_0\cdot(\mathbf{r}-\mathbf{r}_0) + \cdots\end{aligned} \tag{13.8}$$

where $\nabla\mathbf{E}$ is a tensor, the field gradient tensor. Using tensor notation we may also write the second term of this expansion as

$$\frac{1}{2}(\mathbf{r}-\mathbf{r}_0)^T\cdot(\nabla\mathbf{E})_0\cdot(\mathbf{r}-\mathbf{r}_0) = \frac{1}{2}\left[(\mathbf{r}-\mathbf{r}_0)\otimes(\mathbf{r}-\mathbf{r}_0)\right]\odot(\nabla\mathbf{E})_0 = \frac{1}{2}\mathbf{Q}\odot(\nabla\mathbf{E})_0. \tag{13.9}$$

Here we have introduced $\mathbf{Q}$, the tensor of second electric moments, which is related to the quadrupole moment of the electron distribution, and the scalar and vector product operators for two tensors, $\odot$ and $\otimes$. The expansion may, of course, be carried to higher-order terms to produce higher-order electric moments. But we note again that the expressions we get are analogous to the nonrelativistic case.

We have here considered the molecular system and the field as one unit described by one Dirac equation. In the usual experimental situation a molecule is somehow placed in an external field, and it is this field that may be controlled for measurements of the system response. In this situation it is natural to divide the total field into internal and external contributions

$$\phi = \phi^{\text{int}} + \phi^{\text{ext}} \tag{13.10}$$

which corresponds to the fields

$$-\nabla\phi = \mathbf{E} = \mathbf{E}^{\text{int}} + \mathbf{E}^{\text{ext}} = \nabla\phi^{\text{int}} + \nabla\phi^{\text{ext}}. \tag{13.11}$$

However, for the usual molecular case, the external field will influence the internal field. Even within the Born–Oppenheimer approximation, $\mathbf{E}^{\text{ext}}$ will distort the electron distribution, thereby modifying the interactions in $\mathbf{E}^{\text{int}}$. We can take this into account by expressing the energy of the system as a function of the field. We use the Taylor expansion

$$E(\mathbf{E}) = E_0 + \left[\frac{\partial E}{\partial \mathbf{E}^{\text{ext}}}\right]_0 \cdot \mathbf{E}^{\text{ext}} + \frac{1}{2}\mathbf{E}^{\text{ext}} \cdot \left[\frac{\partial^2 E}{\partial \mathbf{E}^{\text{ext}}\partial \mathbf{E}^{\text{ext}}}\right]_0 \cdot \mathbf{E}^{\text{ext}} + \cdots \tag{13.12}$$

where the subscript zero denotes that the derivatives are evaluated at $\mathbf{E}^{\text{ext}} = 0$. The definition of the dipole moment now gives

$$\boldsymbol{\mu} = -\frac{\mathrm{d}E}{\mathrm{d}\mathbf{E}} = -\left[\frac{\partial E}{\partial \mathbf{E}^{\text{ext}}}\right]_0 - \left[\frac{\partial^2 E}{\partial \mathbf{E}^{\text{ext}}\partial \mathbf{E}^{\text{ext}}}\right]_0 \cdot \mathbf{E}^{\text{ext}} - \cdots. \tag{13.13}$$

The first term in this series is obviously just the permanent dipole moment of the molecule. The coefficient of the second term is the polarizability of the molecule, which measures the change in dipole moment induced by the external field. Higher terms in the series will yield higher hyperpolarizabilities. Analogous expressions may be derived for the quadrupole moment.

For a static electric field, the Dirac equation thus provides expressions for various electric properties that are completely analogous to those obtained from nonrelativistic theory. The further calculation of these properties can proceed as for the nonrelativistic case, only using the appropriate relativistic versions of perturbation theory or response theory.

13.3 Gauge Invariance and Finite Basis Sets

Before moving on to magnetic properties, we need to reexamine the role of gauge invariance for molecular calculations. As shown in chapter 3, the introduction of the vector potential **A** through

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{13.14}$$

is arbitrary in the sense that any term with a vanishing curl may be added to **A** without changing **B**. As the energy of the system must depend on the magnetic field, gauge transformations that leave **B** invariant cannot affect the total energy. We will show that this has consequences for the form of the wave function.

The Dirac equation for an electron in a vector potential **A** may be written

$$c\boldsymbol{\alpha} \cdot (-i\hbar\nabla + e\mathbf{A})\,\psi + (\beta mc^2 - E - e\phi)\psi = 0. \tag{13.15}$$

A gauge transformation $\mathbf{A} \rightarrow \mathbf{A}' = \mathbf{A} + \nabla\Lambda$ would be expected to cause a change in the wave function, say $\psi \rightarrow \psi'$. We assume that this may be written as

$$\psi' = \psi Q(\Lambda, \mathbf{r}) \tag{13.16}$$

for some function Q that remains to be determined. The function ψ' must satisfy the equation

$$c\boldsymbol{\alpha} \cdot (-i\hbar\nabla + e\mathbf{A} + e\nabla\Lambda)\,\psi' + (\beta mc^2 - E - e\phi)\psi' = 0. \tag{13.17}$$

Because ψ is a solution for the system without the gauge term, (13.15), this expression reduces to

$$c\boldsymbol{\alpha} \cdot [(-i\hbar\nabla Q) + eQ\nabla\Lambda]\,\psi = 0. \tag{13.18}$$

This equation is satisfied if

$$\nabla Q = (ie/\hbar)Q\nabla\Lambda \tag{13.19}$$

and it is easily seen that this leads to the solution

$$Q = \exp(ie\Lambda/\hbar). \tag{13.20}$$

Thus, the wave function must incorporate the gauge term exponentially in order for the energy to remain invariant under gauge transformations. We should add that the same factor must be included also for the nonrelativistic treatment of the magnetic fields. For a more general derivation for the nonrelativistic case, the reader should consult the book by Sakurai (1967).

The gauge invariance requirement has profound consequences for finite basis calculations. Quiney (2002) has pointed out that even when there is no magnetic field present, a gauge transformation may cause problems. For the simple case of $\Lambda = z$, the transformation of the Hamiltonian is

$$\hat{\mathcal{H}} \rightarrow \hat{\mathcal{H}}' = \hat{\mathcal{H}} + c\alpha_z. \tag{13.21}$$

This new Hamiltonian, $\hat{\mathcal{H}}'$ must have the same eigenvalue spectrum as $\hat{\mathcal{H}}$ as no additional fields have been introduced. However, Quiney has performed calculations on the hydrogen-like Ne^{9+} ion in a finite basis and found that the eigenvalues do indeed change under this gauge transformation. Using a basis of 30 s and 30 p Gaussian basis functions he showed that not only did the additional term change the eigenvalues, but also the fourfold degeneracy of the $p_{3/2}$ level was reduced to two twofold degeneracies. (The gauge term did not remove Kramers degeneracies.) The proper spectrum could only be restored by the addition of a considerable number of higher angular momentum functions to the basis set. The reason for this can be seen by remembering that with the invariance factor included we have for this case

$$\psi' = exp(ie\Lambda/\hbar)\psi. \tag{13.22}$$

Expanding the exponential as

$$exp(iez/\hbar) = 1 + \frac{iez}{\hbar} - \frac{e^2z^2}{\hbar^2} + \cdots \tag{13.23}$$

we see that this introduces higher angular momentum character into the wave function, and this cannot be described using a basis of only s and p functions. In a complete basis, this problem would disappear. Unfortunately, molecular quantum chemistry calculations are carried out with quite limited basis sets, normally chosen on the basis of calculations on atoms with no magnetic field present. The use of additional high angular momentum basis functions would increase the cost of the calculations considerably.

This problem of gauge is not unique to relativistic quantum chemistry: the same problem arises in nonrelativistic theory. What this points to is that we must take care to use a consistent gauge in all calculations. Most formulas in quantum chemistry are derived using the Coulomb gauge, more or less implicitly. This general usage aids in maintaining consistency of gauge, even though it may not be given much thought. We will also adopt the Coulomb gauge where it is relevant.

Calculations on molecules also face an additional problem in the choice of *gauge origin*. Over molecular dimensions, an applied magnetic field can be considered uniform and constant. In this case, the vector potential may be written as

$$\mathbf{A} = \frac{1}{2}\mathbf{B} \times \mathbf{r}. \tag{13.24}$$

The choice of origin for the coordinate system that describes $\mathbf{r}$ is arbitrary, and while the induction $\mathbf{B}$ must remain constant, a shift of origin will change $\mathbf{A}$. If the origin is moved such that the new coordinate $\mathbf{r}' = \mathbf{r} + \Delta\mathbf{r}$, the vector potential changes to

$$\mathbf{A} \rightarrow \mathbf{A}' = \frac{1}{2}\mathbf{B} \times \mathbf{r}' = \frac{1}{2}\mathbf{B} \times \mathbf{r} + \frac{1}{2}\mathbf{B} \times \Delta\mathbf{r} = \mathbf{A} + \Delta\mathbf{A}. \tag{13.25}$$

With the new origin, we must still have

$$\mathbf{B} = \nabla \times \mathbf{A}' = \nabla \times \mathbf{A} + \nabla \times \Delta\mathbf{A} \tag{13.26}$$

showing that $\Delta\mathbf{A}$ must be of the form $\nabla\Lambda$. With $\mathbf{B}$ and $\Delta\mathbf{r}$ constant, we get

$$\nabla\Lambda = \Delta\mathbf{A} = \frac{1}{2}\mathbf{B} \times \Delta\mathbf{r} = \nabla\left[\frac{1}{2}(\mathbf{B} \times \Delta\mathbf{r}) \cdot \mathbf{r}\right]. \tag{13.27}$$

For atoms the natural choice of gauge origin is the atomic nucleus. For a molecule in a uniform magnetic field the choice of gauge origin may not be obvious, and can at most be on one atomic center. To account for this one may use basis functions of the form

$$\chi = \chi_0 \exp\left[\frac{1}{2}(\mathbf{B} \times \Delta\mathbf{r}) \cdot \mathbf{r}\right] \tag{13.28}$$

where χ_0 is the normal basis function derived from field-free calculations. Such functions are often called *London orbitals*, after F. London who introduced these in a study of ring currents in aromatic compounds (London 1973). Ruud et al. (1993) have demonstrated in a series of calculations on small molecules how the use of London orbitals drastically improves the convergence of the calculated magnetizability as a function of basis set size.

13.4 Magnetic Properties

Our discussion of electric properties showed that at least for time-independent fields the operators derived for the relativistic case were just the four-component analogs of the nonrelativistic operators. For magnetic relativistic property operators we do not expect the connection to be so simple due to the fact that the field appears in different forms in the two versions of the Hamiltonian. For the relativistic case, we again write the Hamiltonian as

$$\hat{\mathcal{H}} = c\boldsymbol{\alpha} \cdot (\mathbf{p} + e\mathbf{A}) + \beta mc^2 - e\phi. \tag{13.29}$$

For the nonrelativistic case we may use the spin Hamiltonian discussed in chapter 4,

$$\hat{\mathcal{H}}^{\mathrm{NR}} = [\boldsymbol{\sigma} \cdot (\mathbf{p} + e\mathbf{A})]^2 - e\phi. \tag{13.30}$$

We see that the vector potentials appear to different order in the two Hamiltonians. The form of any property operator derived as a response to changes in the vector potential (or the induction) is therefore likely to depend on which Hamiltonian is used as a starting point.

We will demonstrate this for the magnetic dipole moment, **m**, and for simplicity we assume that we have a uniform field, that is, (13.24) holds. For the relativistic case, this yields the Hamiltonian

$$\hat{\mathcal{H}} = c\boldsymbol{\alpha} \cdot \left[\mathbf{p} + \frac{e}{2}(\mathbf{B} \times \mathbf{r})\right] + \beta mc^2 - e\phi = c\boldsymbol{\alpha} \cdot \mathbf{p} - \frac{ce}{2}(\boldsymbol{\alpha} \times \mathbf{r}) \cdot \mathbf{B} + \beta mc^2 - e\phi. \tag{13.31}$$

We define the magnetic dipole moment **m** via the first-order response to changes in the magnetic induction

$$\mathbf{m} = -\left[\frac{\mathrm{d}E}{\mathrm{d}\mathbf{B}}\right]_{\mathbf{B}=0} = -\left\langle \frac{\partial \hat{\mathcal{H}}}{\partial \mathbf{B}} \right\rangle_{\mathbf{B}=0} \tag{13.32}$$

where we again use the Hellmann–Feynman theorem. Differentiating, we get

$$\mathbf{m} = \frac{ce}{2}\langle \psi \,|\, \boldsymbol{\alpha} \times \mathbf{r} \,|\, \psi \rangle = \frac{1}{2}\langle \psi \,|\, \mathbf{j} \times \mathbf{r} \,|\, \psi \rangle \tag{13.33}$$

where **j** is the current density. If we expand the wave function in 2-spinors, we get

$$\mathbf{m} = \frac{ce}{2}\left[\langle \psi^S \,|\, \boldsymbol{\sigma} \times \mathbf{r} \,|\, \psi^L \rangle + \langle \psi^L \,|\, \boldsymbol{\sigma} \times \mathbf{r} \,|\, \psi^S \rangle\right]. \tag{13.34}$$

We see that the magnetic dipole operator connects large and small components, but the prefactor c ensures that the results are of the same magnitude as expectation values between two large components. We can demonstrate this point by substituting for the small component using the approximate expression from the nonrelativistic limit

$$\psi^S = \frac{c}{2mc^2}\boldsymbol{\sigma} \cdot \mathbf{p}\psi^L. \tag{13.35}$$

This yields the magnetic moment as

$$\mathbf{m} = \frac{e}{4m}\left[\langle \boldsymbol{\sigma} \cdot \mathbf{p}\psi^L \,|\, \boldsymbol{\sigma} \times \mathbf{r} \,|\, \psi^L \rangle + \langle \psi^L \,|\, \boldsymbol{\sigma} \times \mathbf{r} \,|\, \boldsymbol{\sigma} \cdot \mathbf{p}\psi^L \rangle\right]. \tag{13.36}$$

Evaluation of this expression is straightforward but somewhat lengthy, and we just give the result

$$\mathbf{m} = \frac{e}{2m}\langle \psi^L \,|\, -\mathbf{r} \times \mathbf{p} - \hbar\boldsymbol{\sigma} \,|\, \psi^L \rangle = \langle \psi^L \,|\, -\gamma(\boldsymbol{\ell} + 2\mathbf{s}) \,|\, \psi^L \rangle \tag{13.37}$$

where γ is the gyromagnetic ratio of the electron. This is the same result as we get by substituting for the induction in the nonrelativistic spin Hamiltonian, (4.22),

$$\hat{\mathcal{H}}^{\mathrm{NR}} = -\frac{\hbar^2}{2m}\nabla^2 + \frac{e}{2m}(\mathbf{B} \times \mathbf{r}) \cdot \mathbf{p} + \frac{e^2}{8m}(\mathbf{B} \times \mathbf{r})^2 + \frac{e\hbar}{2m}\boldsymbol{\sigma} \cdot \mathbf{B} - e\phi \tag{13.38}$$

and applying the response definition for **m**. The gyromagnetic ratio of the electron is altered in the third decimal place by QED effects.

Atoms and molecules may also have internal sources of magnetic fields. Many isotopes of atomic nuclei have magnetic moments, and these give rise to local fields. The energy shifts caused by these local fields are the most important of the effects leading to *hyperfine structure* of the energy spectrum, the spin–orbit splittings accounting for the *fine structure*. A nucleus K with spin $\mathbf{I}_K$ has a magnetic moment given by

$$\mathbf{m}_K = \gamma_K \hbar \mathbf{I}_K \tag{13.39}$$

where γ_K is the gyromagnetic ratio for that isotope. For simplicity, we will assume a point nucleus throughout this part of the derivation, but we return to the effects of finite nuclear size in a later section. If we regard the nucleus as a magnetic point dipole, the corresponding vector potential may be written as (Jackson 1975)

$$\mathbf{A}_k(\mathbf{r}_k) = \frac{\mu_0(\mathbf{m}_K \times \mathbf{r}_K)}{4\pi r_K^3} = \frac{\mu_0 \gamma_K \hbar (\mathbf{I}_K \times \mathbf{r}_K)}{4\pi r_K^3} \tag{13.40}$$

where $\mathbf{r}_K$ is the radial coordinate in a coordinate system centered on nucleus K. For measurements in which there is an applied external magnetic field in addition to the nuclear field, the Hamiltonian is

$$\begin{aligned}\hat{\mathcal{H}} &= c\boldsymbol{\alpha}\cdot\left[\mathbf{p}+\frac{e}{2}(\mathbf{B}\times\mathbf{r})+e\sum_K \frac{\mu_0\gamma_K\hbar(\mathbf{I}_K\times\mathbf{r}_K)}{4\pi r_K^3}\right]+\beta mc^2-e\phi \\ &= c\boldsymbol{\alpha}\cdot\mathbf{p}-\frac{ce}{2}\mathbf{B}\cdot(\boldsymbol{\alpha}\times\mathbf{r})-\frac{ce\mu_0\hbar}{4\pi}\sum_K\frac{\gamma_K\mathbf{I}_K\cdot(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3}+\beta mc^2-e\phi.\end{aligned} \tag{13.41}$$

We can now define a *hyperfine coupling tensor* $\mathbf{g}_K$ for nucleus K by

$$\mathbf{g}_K = \left[\frac{\mathrm{d}E}{\mathrm{d}\mathbf{I}_K}\right]_{\mathbf{B}=0,\mathbf{I}_K=0} \tag{13.42}$$

and immediately find that

$$\mathbf{g}_K = -\frac{ce\mu_0\hbar\gamma_K}{4\pi}\left\langle \frac{(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3}\right\rangle. \tag{13.43}$$

For neutral closed shell molecules this term vanishes, but for radicals and ions it becomes important and is exploited in *electron spin resonance* (ESR) spectroscopy. For hyperfine interactions that might be observed in closed shell, neutral molecules, we must turn to higher-order effects. This in turn requires that we elaborate somewhat on our approach to properties in terms of response theory, which we do in the next section.

13.5 Second-Order Properties

For a first-order property and a fully variational wave function, we have used the Hellmann–Feynman theorem

$$\frac{\mathrm{d}E}{\mathrm{d}\mathbf{X}} = \left\langle \frac{\partial \hat{\mathcal{H}}}{\partial \mathbf{X}} \right\rangle \tag{13.44}$$

to obtain the first derivatives of the energy with respect to the response parameters **X**. For second-order properties we have to obtain the necessary derivatives in a slightly more roundabout manner. Let us assume that our Hamiltonian can be written in terms of the parameters **X** in the form

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{H}}_1 \cdot \mathbf{X} + \mathbf{X}^T \cdot \hat{\mathcal{H}}_2 \cdot \mathbf{X} + \cdots . \tag{13.45}$$

We note that

$$\hat{\mathcal{H}}_1 = \frac{\mathrm{d}\hat{\mathcal{H}}}{\mathrm{d}\mathbf{X}}; \qquad 2\hat{\mathcal{H}}_2 = \frac{\mathrm{d}^2\hat{\mathcal{H}}}{\mathrm{d}\mathbf{X}^2}. \tag{13.46}$$

Here and for the remainder of the section we assume that the derivatives are evaluated at $\mathbf{X} = 0$.

The total energy of the system may be expanded in a Taylor series around $\mathbf{X} = 0$,

$$E = E_0 + \frac{\mathrm{d}E}{\mathrm{d}\mathbf{X}} \cdot \mathbf{X} + \frac{1}{2}\mathbf{X}^T \cdot \frac{\mathrm{d}^2E}{\mathrm{d}\mathbf{X}^2} \cdot \mathbf{X} + \cdots . \tag{13.47}$$

We know from time-independent nondegenerate perturbation theory that the energy may also be written as

$$E = E_0 + \langle 0|\hat{\mathcal{H}}_1 \cdot \mathbf{X}|0\rangle + \langle 0|\mathbf{X}^T \cdot \hat{\mathcal{H}}_2 \cdot \mathbf{X}|0\rangle + \sum_{k\neq 0} \frac{\langle 0|\hat{\mathcal{H}}_1 \cdot \mathbf{X}|k\rangle\langle k|\hat{\mathcal{H}}_1 \cdot \mathbf{X}|0\rangle}{E_0 - E_k} + \cdots \tag{13.48}$$

where $\{|k\rangle\}$ is a set of intermediate states. Comparing the perturbation expansion term by term with the Taylor series, we can easily identify the first derivative as

$$\frac{\mathrm{d}E}{\mathrm{d}\mathbf{X}} = \langle 0|\hat{\mathcal{H}}_1|0\rangle = \langle 0|\frac{\mathrm{d}\hat{\mathcal{H}}}{\mathrm{d}\mathbf{X}}|0\rangle \tag{13.49}$$

which is nothing but the Hellmann–Feynman theorem. The second derivative is obtained from

$$\frac{1}{2}\frac{\mathrm{d}^2E}{\mathrm{d}\mathbf{X}^2} = \langle 0|\hat{\mathcal{H}}_2|0\rangle + \sum_{k\neq 0} \frac{\langle 0|\hat{\mathcal{H}}_1|k\rangle\langle k|\hat{\mathcal{H}}_1|0\rangle}{E_0 - E_k} \tag{13.50}$$

or

$$\frac{\mathrm{d}^2 E}{\mathrm{d}\mathbf{X}^2} = \langle 0|\frac{d^2\hat{\mathcal{H}}}{d\mathbf{X}^2}|0\rangle + 2\sum_{k\neq 0}\langle 0|\frac{\mathrm{d}\hat{\mathcal{H}}}{\mathrm{d}\mathbf{X}}|k\rangle\langle k|\frac{\mathrm{d}\hat{\mathcal{H}}}{\mathrm{d}\mathbf{X}}|0\rangle/(E_0 - E_k). \tag{13.51}$$

This derivation provides us with an expression for the second derivative of the energy with respect to the response parameter **X**. However, the seemingly innocent perturbation expression above hides some subtleties that turn out to be of great importance. The question arises, "How do we choose the intermediate states $|k\rangle$?" As pointed out in chapter 5, in the Dirac picture the vacuum is filled with electrons in negative-energy states, and the definition of the vacuum is therefore dependent on the Hamiltonian. If we change the Hamiltonian, we will in general change the negative-energy states, and thereby the vacuum. Thus, the vacuum for the unperturbed Hamiltonian with $\mathbf{X} = 0$ is different from the vacuum for the Hamiltonian that incorporates a perturbation. In doing a quantum chemical calculation, we are interested only in the electrons with positive energies, and we somewhat arbitrarily set the energy contribution from the vacuum to zero.

Another way of looking at this is to say that we do a calculation on the entire system with M electrons in positive-energy states and N electrons in negative-energy states, and then we subtract the energy of a system of N electrons in negative-energy states to obtain the energy of the M-electron system. If our starting point is a calculation of this type for the unperturbed Hamiltonian, then we must account for the changes to the vacuum when we include the perturbation. We write $|v\rangle$ for positive-energy unoccupied states and $|n\rangle$ for (filled) negative-energy states, and find the total energy change for the $N + M$-particle system due to the perturbation as

$$E^{(2)}_{N+M} = \sum_{n,v}\frac{\langle n|\hat{\mathcal{H}}_1\cdot\mathbf{X}|v\rangle\langle v|\hat{\mathcal{H}}_1\cdot\mathbf{X}|n\rangle}{E_n - E_v} + \sum_{v}\frac{\langle 0|\hat{\mathcal{H}}_1\cdot\mathbf{X}|v\rangle\langle v|\hat{\mathcal{H}}_1\cdot\mathbf{X}|0\rangle}{E_0 - E_v}. \tag{13.52}$$

The energy correction to the vacuum is similar, but now $|0\rangle$ is an unoccupied state, accessible as a virtual state for the negative-energy electrons:

$$E^{(2)}_{N} = \sum_{n,v}\frac{\langle n|\hat{\mathcal{H}}_1\cdot\mathbf{X}|v\rangle\langle v|\hat{\mathcal{H}}_1\cdot\mathbf{X}|n\rangle}{E_n - E_v} + \sum_{n}\frac{\langle n|\hat{\mathcal{H}}_1\cdot\mathbf{X}|0\rangle\langle 0|\hat{\mathcal{H}}_1\cdot\mathbf{X}|n\rangle}{E_n - E_0}. \tag{13.53}$$

Finally, we can calculate the energy of the M-electron system as the difference of the two energies above

$$E^{(2)}_{M} = E^{(2)}_{N+M} - E^{(2)}_{N} = \sum_{v}\frac{\langle 0|\hat{\mathcal{H}}_1\cdot\mathbf{X}|v\rangle\langle v|\hat{\mathcal{H}}_1\cdot\mathbf{X}|0\rangle}{E_0 - E_v} + \sum_{n}\frac{\langle 0|\hat{\mathcal{H}}_1\cdot\mathbf{X}|n\rangle\langle n|\hat{\mathcal{H}}_1\cdot\mathbf{X}|0\rangle}{E_0 - E_n}. \tag{13.54}$$

The conclusion of this is that we must include the negative-energy states in the perturbation sum over intermediate states. In a simple spinor picture this means that the perturbation sums also include those spinors with energies below $-2mc^2$ (using an energy-shifted Hamiltonian). This conclusion is exactly the same as if we had regarded the negative-energy states simply as virtual spinors to be included in the perturbation sum. One might naïvely think that contributions from these terms would be unimportant because of the large energy differences in the denominators. However, the magnitude of the contribution depends on the nature of the perturbation, whether this is from an electric or magnetic field.

For electric properties the operators have the same form as in nonrelativistic theory, and appear on the diagonal of the Dirac matrix. The perturbation matrix elements therefore involve one-electron matrix elements of the form

$$\langle \psi_0^L \,|\, V^{\text{ext}} \,|\, \psi_p^L \rangle + \langle \psi_0^S \,|\, V^{\text{ext}} \,|\, \psi_p^S \rangle. \tag{13.55}$$

If p is a positive-energy spinor, this matrix element is $\mathcal{O}(c^0)$, and the contribution from the sum over positive-energy states is also $\mathcal{O}(c^0)$. If p is a negative-energy spinor, this matrix element is $\mathcal{O}(c^{-1})$, and the contribution from the sum over negative-energy states is $\mathcal{O}(c^{-4})$. The lowest-order relativistic corrections to second-order electric properties, of $\mathcal{O}(c^{-2})$, therefore arise solely from the sum over positive-energy states.

The situation is different for magnetic properties. Here, the perturbation matrix elements involve one-electron matrix elements of the form

$$\langle \psi_0^L \,|\, c\boldsymbol{\sigma} \cdot \mathbf{A} \,|\, \psi_p^S \rangle + \langle \psi_0^S \,|\, c\boldsymbol{\sigma} \cdot \mathbf{A} \,|\, \psi_p^L \rangle \tag{13.56}$$

If p is a positive-energy spinor, this matrix element is $\mathcal{O}(c^0)$, because the small component is $\mathcal{O}(c^{-1})$. The contribution from the sum over positive-energy states is therefore also $\mathcal{O}(c^0)$. However, if p is a negative-energy spinor, this matrix element is $\mathcal{O}(c^1)$, because the small component for a negative-energy spinor is $\mathcal{O}(c^0)$. The contribution from the sum over negative-energy states is therefore also $\mathcal{O}(c^0)$.

Due to the sign of the denominator, the sum over negative-energy states is positive, whereas the sum over positive-energy states is negative. The difference in sign between these two terms means that they must cancel to some extent. Consequently, the large and small components and the positive- and negative-energy states must be well represented to avoid introducing significant error into the result. One approach that addresses this problem is given in a later section.

The necessity for summing over the negative-energy states raises the question of whether we should include QED renormalization effects. The answer is, in general, "yes": a rigorous treatment would involve virtual pairs and renormalization with counter-terms to remove the divergences in the perturbation sums, which is beyond the scope of relativistic quantum chemistry at present. For electric perturbations, the effect of renormalization is small, because the contribution is already $\mathcal{O}(c^{-4})$. It is therefore reasonable to adopt the "no-pair" approximation for electric perturbations. The same is true for second-order perturbation treatments of the correlation energy, where sums over the negative-energy states are omitted. For magnetic perturbations, the sums over negative-energy states contribute at $\mathcal{O}(c^0)$, and it is pertinent to ask whether QED renormalization is required. In the alternative treatment of magnetic effects in section 13.7, the magnetic perturbations appear in the same way as the electric perturbations, and the effect of renormalization is therefore likely to be small. However, it should be noted that QED effects change the electron gyromagnetic ratio in the third decimal place, from the integer value of 2 derived from the Dirac equation. QED effects are also important for magnetic hyperfine interactions in atoms.

Second-order properties are often evaluated using coupled-perturbed Hartree–Fock (CPHF) theory. The CPHF wave function is essentially the first-order perturbed wave function, which, as we saw above, must include the negative-energy states. Thus, in the relativistic case, the CPHF method must include both the positive- and negative-energy states.

13.6 NMR Parameters

Nuclear magnetic resonance (NMR) spectroscopy is based on the measurement of the magnetic moments of atomic nuclei. The nuclei are subjected to a strong external field, and this splits the normally degenerate nuclear spin states. For the simplest case of a spin $\frac{1}{2}$ nucleus, we would get a low-energy state with the spin aligned in the direction of the field, and a high-energy state with the spin in the opposite direction. By applying electromagnetic radiation, we can induce transitions between the two states and measure the transition energy.

NMR spectra can be described using an effective Hamiltonian of the form

$$\hat{\mathcal{H}} = -\sum_{K} \gamma_K \hbar \mathbf{B} \cdot (\mathbf{1} - \boldsymbol{\sigma}_K) \cdot \mathbf{I}_K + \sum_{K<L} \gamma_K \gamma_L \hbar^2 \mathbf{I}_K \cdot (\mathbf{D}_{KL} + \mathbf{K}_{KL}) \cdot \mathbf{I}_L, \tag{13.57}$$

which operates in the space of nuclear spin functions. Here $\boldsymbol{\sigma}_K$ is the shielding tensor for nucleus K.[2] $\mathbf{D}_{KL}$ is the direct spin–spin coupling tensor, which accounts for the direct interaction of nuclear magnetic dipoles, and $\mathbf{K}_{KL}$ is the reduced indirect spin–spin coupling tensor, which accounts for the coupling of nuclear spins mediated by the interaction with the surrounding electrons. It turns out that the contribution from the $\mathbf{D}_{KL}$ term vanishes in the rotational averaging necessary to account for the rapid tumbling of molecules in NMR experiments, and so $\mathbf{K}_{KL}$ is the quantity of immediate interest. This is related to the indirect spin–spin coupling tensor $\mathbf{J}_{KL}$ normally derived from NMR experiments by

$$\mathbf{J}_{KL} = \gamma_K \gamma_L \hbar^2 \mathbf{K}_{KL}. \tag{13.58}$$

---

[2] The use of $\boldsymbol{\sigma}_K$ for the shielding tensor is conventional but unfortunate in this connection due to the notational similarity to the Pauli $\boldsymbol{\sigma}$ matrices. Here, the shielding tensor will always have a capital letter subscript, and it only appears in this section, so the chances of confusion should be minimal.

To obtain the quantities $\boldsymbol{\sigma}_K$ and $\mathbf{K}_{KL}$ from electronic structure calculations, we turn again to response theory. From the form of the effective Hamiltonian it is evident that

$$\boldsymbol{\sigma}_K = \mathbf{1} + \frac{1}{\hbar\gamma_K}\frac{\mathrm{d}^2 E}{\mathrm{d}\mathbf{B}\,\mathrm{d}\mathbf{I}_K} \tag{13.59}$$

and

$$\mathbf{D}_{KL} + \mathbf{K}_{KL} = \frac{1}{\gamma_K\gamma_L\hbar^2}\frac{\mathrm{d}^2 E}{\mathrm{d}\mathbf{I}_K\,\mathrm{d}\mathbf{I}_L}. \tag{13.60}$$

Again, all derivatives are evaluated at $\mathbf{B} = 0, \mathbf{I}_K = 0, \mathbf{I}_L = 0$.

If we use the Hamiltonian (13.41), we have no terms that are quadratic in the response parameters $\mathbf{B}$ and $\mathbf{I}_K$, and the only contribution to the second-order perturbation energy comes from the sum over states

$$\begin{aligned} E^{(2)} = \sum_{k\neq 0}\Bigg[& \big\langle 0 \big| \frac{ce}{2}\mathbf{B}\cdot(\boldsymbol{\alpha}\times\mathbf{r}) + \frac{ce\mu_0\hbar}{4\pi}\sum_K \frac{\gamma_K \mathbf{I}_K\cdot(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3} \big| k \big\rangle \\ & \times \big\langle k \big| \frac{ce}{2}\mathbf{B}\cdot(\boldsymbol{\alpha}\times\mathbf{r}) + \frac{ce\mu_0\hbar}{4\pi}\sum_K \frac{\gamma_K \mathbf{I}_K\cdot(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3} \big| 0 \big\rangle / (E_0 - E_k)\Bigg] \end{aligned} \tag{13.61}$$

where as we have seen in the previous section, the sum over k must also include the negative-energy states. If we now carry out the required differentiations, we get the shielding tensor as

$$\begin{aligned} \boldsymbol{\sigma}_K = \mathbf{1} + \frac{\gamma_K c^2 e^2 \mu_0 \hbar}{4\pi}\sum_{k\neq 0}\Bigg[& \big\langle 0 \,\big|\, (\boldsymbol{\alpha}\times\mathbf{r}) \,\big|\, k \big\rangle \big\langle k \,\big|\, \frac{(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3} \,\big|\, 0 \big\rangle \\ & + \big\langle 0 \,\big|\, \frac{(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3} \,\big|\, k \big\rangle \big\langle k \,\big|\, (\boldsymbol{\alpha}\times\mathbf{r}) \,\big|\, 0 \big\rangle \Bigg] / (E_0 - E_k) \end{aligned} \tag{13.62}$$

and the spin–spin coupling tensors as

$$\begin{aligned} \mathbf{D}_{KL} + \mathbf{K}_{KL} = \frac{\gamma_K\gamma_L c^2 e^2 \mu_0^2 \hbar^2}{8\pi^2}\sum_{k\neq 0}\Bigg[& \big\langle 0 \,\big|\, \frac{(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3} \,\big|\, k \big\rangle \big\langle k \,\big|\, \frac{(\boldsymbol{\alpha}\times\mathbf{r}_L)}{r_L^3} \,\big|\, 0 \big\rangle \\ & + \big\langle 0 \,\big|\, \frac{(\boldsymbol{\alpha}\times\mathbf{r}_L)}{r_L^3} \,\big|\, k \big\rangle \big\langle k \,\big|\, \frac{(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3} \,\big|\, 0 \big\rangle \Bigg] / (E_0 - E_k). \end{aligned} \tag{13.63}$$

The first term in the sum for the shielding constants may be split into the contribution from the negative-energy states and the positive unoccupied states

$$\begin{aligned}
&\sum_{k\neq 0}\left[\langle\, 0 \mid (\boldsymbol{\alpha}\times\mathbf{r}) \mid k\,\rangle\langle\, k \mid \frac{(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3} \mid 0\,\rangle\right]/(E_0 - E_k)\\
&\quad=\sum_{n}\left[\langle\, 0 \mid (\boldsymbol{\alpha}\times\mathbf{r}) \mid n\,\rangle\langle\, n \mid \frac{(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3} \mid 0\,\rangle\right]/(E_0 - E_n)\\
&\quad+\sum_{v}\left[\langle\, 0 \mid (\boldsymbol{\alpha}\times\mathbf{r}) \mid v\,\rangle\langle\, v \mid \frac{(\boldsymbol{\alpha}\times\mathbf{r}_K)}{r_K^3} \mid 0\,\rangle\right]/(E_0 - E_v)
\end{aligned}\tag{13.64}$$

in the notation of the previous section. As pointed out there, the two types of contributions to the sum would be expected to be of equal magnitude, and thus the negative-energy states turn out to play a crucial role in the calculation of NMR parameters within a four-component formalism. The sum over positive-energy states may be interpreted as corresponding to the *paramagnetic* contribution in the nonrelativistic formulation, whereas the other sum corresponds to the *diamagnetic* contribution.

An advantage of the relativistic formulation is that the expressions are considerably simpler than those arising in the nonrelativistic treatment of NMR parameters. For spin–spin coupling, the nonrelativistic Hamiltonian involves the perturbation (Ramsey 1953)

$$\hat{\mathcal{H}}' = \frac{e^2A^2}{2m} + \frac{2C\mathbf{I}\cdot\mathbf{L}_K}{r_K^3} + \frac{3C(\mathbf{I}\cdot\mathbf{r}_K)(\boldsymbol{\sigma}\cdot\mathbf{r}_K) - \mathbf{I}\cdot\boldsymbol{\sigma}}{r_K^3} + \frac{8C\pi}{3}\mathbf{I}\cdot\boldsymbol{\sigma}\delta(\mathbf{r}_K) \tag{13.65}$$

where C is a constant. Of course, the machinery involved in generating the four-component wave functions is computationally considerably heavier than the nonrelativistic calculations.

In actual practice, the perturbation expressions given above are better suited for displaying the structural features of the theory than for use in large-scale calculations. For nonrelativistic calculations, the most efficient approach to the ab initio calculation of NMR parameters has been variational perturbation theory (Helgaker et al. 1999). This, or equivalent approaches such as second-order propagator theory, will probably also turn out to be the optimal choice for the relativistic case.

Finally, we note that the term

$$\frac{\mathrm{d}^2E}{\mathrm{d}\mathbf{B}^2} \tag{13.66}$$

may also be calculated. This represents a higher-order contribution to the magnetizability, but is of little interest for NMR spectroscopy.

13.7 Alternative Treatment of Magnetic Interactions

The nonrelativistic Hamiltonian can—as we have shown—be derived by elimination of the small component from the relativistic Hamiltonian, including magnetic terms.

However, the nonrelativistic term in $\mathbf{A}^2$ arises from the relativistic second-order perturbation expressions given above: like the correspondence between the nonrelativistic and relativistic kinetic energy discussed in section 11.2, there is a projection operator involving the small component (or in this case, a resolvent). If the small component is poorly represented, the cancellation between the sums over positive- and negative-energy states could show considerable error.

Kutzelnigg (2003) has shown how the sum over negative-energy states can be largely avoided. This sum arises because $\boldsymbol{\alpha}$ connects large and small components. If we could make the perturbation operator block diagonal, but *not* the unperturbed Dirac operator, the sum over negative-energy states might be eliminated, or at worst reduced to contributions that were $\mathcal{O}(c^{-4})$.

To perform the diagonalization, we use the unitary transformation

$$\hat{\mathcal{U}} = e^{\lambda\tau} \tag{13.67}$$

with

$$\tau = -\frac{e}{2mc}\beta\boldsymbol{\alpha}\cdot\mathbf{A}, \tag{13.68}$$

and apply this transformation to the perturbed Dirac Hamiltonian,

$$\hat{\mathcal{H}}' = \hat{\mathcal{U}}^\dagger\,(\hat{\mathcal{H}}_0 + \hat{\mathcal{H}}_1)\,\hat{\mathcal{U}} \tag{13.69}$$

where

$$\hat{\mathcal{H}}_0 = c\boldsymbol{\alpha}\cdot\mathbf{p} + (\beta - \mathbf{I})c^2 + V; \qquad \hat{\mathcal{H}}_1 = ec\boldsymbol{\alpha}\cdot\mathbf{A}. \tag{13.70}$$

There is of course a corresponding transformation to the wave function,

$$\psi' = \hat{\mathcal{U}}^\dagger\psi. \tag{13.71}$$

The parameter λ has been introduced so that an expansion can be made, but it can also be considered as a field strength parameter.

We now use the expansion of the exponential operators in a commutator series (see section 8.1), and collect terms of various orders to define perturbed operators:

$$\begin{aligned}
\hat{\mathcal{H}}'_0 &= \hat{\mathcal{H}}_0 \\
\hat{\mathcal{H}}'_1 &= \hat{\mathcal{H}}_1 + \left[\hat{\mathcal{H}}_0, \tau\right] = \frac{e}{2m}\beta[\boldsymbol{\alpha}\cdot\mathbf{p}, \boldsymbol{\alpha}\cdot\mathbf{A}]_+ = \frac{e}{2m}\beta(2\mathbf{A}\cdot\mathbf{p} + \hbar\boldsymbol{\Sigma}\cdot\mathbf{B}) \\
\hat{\mathcal{H}}'_2 &= \left[\hat{\mathcal{H}}_1, \tau\right] + \tfrac{1}{2}\left[\left[\hat{\mathcal{H}}_0, \tau\right], \tau\right] = \frac{e^2}{2m}\beta\mathbf{A}^2 - \frac{e^2}{4m^2c}[\mathbf{A}\cdot\mathbf{p}, \boldsymbol{\alpha}\cdot\mathbf{A}]_+.
\end{aligned} \tag{13.72}$$

Already we see that we have obtained expressions that are much more like the nonrelativistic expressions for the magnetic perturbations, except that they are multiplied by a factor of β. The second-order perturbation contains the diamagnetic term $\mathbf{A}^2$ and a relativistic correction term, whose expectation is $\mathcal{O}(c^{-2})$ because of the $\boldsymbol{\alpha}$ matrices. The diamagnetic term is now simply

$$\left(\langle \psi^L | e^2\mathbf{A}^2 | \psi^L \rangle - \langle \psi^S | e^2\mathbf{A}^2 | \psi^S \rangle\right)/2m. \tag{13.73}$$

Formally, we have made an expansion in λ. To arrive at the properties, both first- and second-order, we differentiated with respect to the field strength. We can consider λ as a field strength parameter, taking the scale factor out of **A**. However, the transformation is also an expansion in powers of $1/c$, which appears in the perturbation operator.

This transformation thus provides a simple and robust expression for the diamagnetic shielding term, but it has some drawbacks. It makes gauge transformations more complicated: we now have

$$\exp\left[-i\Lambda - \frac{1}{2mc}\beta\boldsymbol{\alpha}\cdot\nabla\Lambda\right] \tag{13.74}$$

as the gauge transformation. A further complication is that, while the transformation commutes with the Coulomb interaction, it does not commute with the magnetic interactions. Taking the Gaunt interaction as an example, the transformation yields two-particle operators that must be added to the perturbed Hamiltonian at various orders in λ. The contributions to operators of order c^{-2} are

$$\begin{aligned}\hat{\mathcal{H}}_1'(i,j) &\leftarrow -\frac{e^3}{4\pi\epsilon_0 mc}\left(\beta_i\boldsymbol{\alpha}_j\cdot\mathbf{A}_i + \beta_j\boldsymbol{\alpha}_i\cdot\mathbf{A}_j\right)/r_{ij}\\ \hat{\mathcal{H}}_2'(i,j) &\leftarrow -\frac{e^4}{4\pi\epsilon_0 m^2c^2}\beta_i\beta_j\mathbf{A}_i\cdot\mathbf{A}_j/r_{ij}.\end{aligned} \tag{13.75}$$

The two-electron terms of $\hat{\mathcal{H}}_1'$ contribute to first-order properties (that is, linear in **A**) at $\mathcal{O}(c^{-2})$, but contribute to second-order properties (that is, quadratic in **A**) only at $\mathcal{O}(c^{-4})$ and can be safely neglected there. The two-electron terms of $\hat{\mathcal{H}}_2'$, however, contribute at $\mathcal{O}(c^{-2})$ to second-order properties. Thus, while the primary contributions to the magnetic interactions will come from terms that are $\mathcal{O}(c^0)$, the evaluation of the terms of $\mathcal{O}(c^{-2})$, which are the relativistic corrections to the property operators, will be complicated by these two-electron contributions.

13.8 Finite Nucleus Effects on Properties

Thus far in our discussion of relativistic expressions for properties we have assumed that the nuclei are represented by point charges. However, schemes for actual calculation of relativistic wave functions normally use nuclei with finite size in order to avoid problems with the weak singularity of the Dirac equation at the nucleus—and also because the nucleus really does have a finite size. The use of a point nucleus to calculate properties therefore appears somewhat inconsistent. At the very least we should know what errors we incur by using a point nucleus, and we will therefore discuss the low-order effects of finite nuclear size for electric and magnetic fields.

The source of the electric field can be an externally applied field, or it can originate in the components of the nuclear potential that are not included in the internal component of the field (that is, the nuclear potential V). Such components arise from the nonspherical nature of the nucleus, the lowest-order term of which is the quadrupole moment. The implementation of a finite-nuclear model is quite straightforward: we simply expand the nuclear charge distribution in a series:

$$\rho_{\mathrm{nuc}}(\mathbf{r}) = \rho_0 \exp[-\zeta_{\mathrm{nuc}} r^2] + \mathbf{Q}_{\mathrm{nuc}} \odot \boldsymbol{\rho}_2(\mathbf{r}) + \cdots \tag{13.76}$$

where the first term is the spherical component of the nuclear charge distribution introduced in chapter 7. The tensor function $\boldsymbol{\rho}_2(\mathbf{r})$ can be factorized in spherical coordinates. For this purpose we introduce the modified spherical harmonic[3]

$$C_{kq}(\vartheta, \phi) = \sqrt{\frac{4\pi}{2k+1}} Y_{kq}(\vartheta, \phi). \tag{13.77}$$

The collection of all the C_{kq} for a given value of k then yields a vector $\mathbf{C}_k$ with $2k+1$ components. For the tensor function above we get

$$\boldsymbol{\rho}_2(\mathbf{r}) = \rho_2(r)\mathbf{C}_2(\vartheta, \varphi). \tag{13.78}$$

The form of the radial function depends on the nuclear structure giving rise to the nonspherical nuclear shape, but given the very simple models of the nucleus that are currently used in relativistic quantum chemistry, it is probably very little restriction[4] to set $\rho_2 \propto r^2 \rho_0 \exp[-\zeta_{\mathrm{nuc}} r^2]$. The integrals over $\boldsymbol{\rho}_2$ can be readily evaluated with standard two-electron integral code as $(d|ab)$ integrals, just as the spherical part of the finite nuclear potential can be evaluated as $(s|ab)$ integrals. Here, d and s refer to the angular momentum of the charge distribution, and a and b to basis functions. This approach can be applied to both relativistic and nonrelativistic calculations with a finite nucleus. For the relativistic case, symmetry considerations show that spinors must have $j \geq 3/2$ to contribute to these integrals. These spinors have small amplitudes in the nuclear region, and thus the correction to the nuclear quadrupole moment due to a finite nuclear model is likely to be small.

We next consider the effect of finite nuclear size on the nuclear spin Hamiltonian. The electric moments were derived by considering the Coulomb interaction of the nuclear charge density, expanded in a multipole series, with the electrons. By analogy, the magnetic moments are derived by considering the Gaunt interaction of the nucleus with the electrons. It is at this point that we must consider, at least as a formal entity, the nuclear wave function, and from it obtain a nuclear spin density that interacts with the electron spin density.

---

[3] Also known as a spherical tensor.

[4] The power of r is necessary to give the correct form for reduction to the point nucleus expression, and can be justified from a consideration of the part of the nuclear wave function that gives rise to the nonspherical terms.

We write the nuclear wave function (actually, the wave function for the protons because the interaction is electromagnetic) as a single determinant, and the wave function for the nucleus and the electrons as a product of determinants. This means that there are no exchange contributions to the interaction, only direct contributions. From our consideration of time-reversal symmetry applied to the Gaunt interaction, the interaction must come from the unpaired protons and the unpaired electrons. For the purpose of developing the theory, we assume that there is only a single unpaired proton in the nucleus. The general case can be derived by summing over unpaired nucleons.

The contribution from the electron–proton Gaunt interaction will be proportional to the expectation value of the operator $(\boldsymbol{\alpha}_e \cdot \boldsymbol{\alpha}_n) r_{en}^{-1}$, where indices e and n refer to electron and proton, respectively. The nuclear part of this expectation value should appear as a vector potential $\mathbf{A}^{\mathrm{nuc}}$ in the manner that any vector potential appears in the Hamiltonian:

$$c\boldsymbol{\alpha} \cdot (\mathbf{p} + e\mathbf{A}^{\mathrm{nuc}}) = c\boldsymbol{\alpha} \cdot \mathbf{p} + ce\boldsymbol{\alpha} \cdot \mathbf{A}^{\mathrm{nuc}}. \tag{13.79}$$

Thus, to find $\mathbf{A}^{\mathrm{nuc}}$ we must develop the nuclear Gaunt contribution to the Hamiltonian and identify the factor forming a scalar product with $\boldsymbol{\alpha}$. Below we sketch how to proceed.

We first introduce the wave function for the proton. The angular part can be represented as a 2-spinor angular momentum eigenfunction and the radial part as a simple Gaussian,

$$\psi^{\mathrm{nuc}} = e^{-\eta r_n^2} \begin{pmatrix} N^L \, r_n^{\ell} \, \chi_{\kappa,m}(\Omega_n) \\ N^S \, r_n^{\ell-a} \, \chi_{-\kappa,m}(\Omega_n) \end{pmatrix}, \tag{13.80}$$

where $a = \operatorname{sgn} \kappa$. The contribution from the electron–proton Gaunt interaction to the Hamiltonian will be a term proportional to

$$\langle \psi^{\mathrm{nuc}} | \frac{\boldsymbol{\alpha}_e \cdot \boldsymbol{\alpha}_n}{r_{en}} | \psi^{\mathrm{nuc}} \rangle = \boldsymbol{\alpha}_e \cdot \langle \psi^{\mathrm{nuc}} | \frac{\boldsymbol{\alpha}_n}{r_{en}} | \psi^{\mathrm{nuc}} \rangle. \tag{13.81}$$

From this we deduce that the corresponding magnetic field may be written as

$$\mathbf{A}^{\mathrm{nuc}} \propto N^L N^S \int r_n^{2\ell-a} \frac{1}{r_{en}} e^{-2\eta r_n^2} (\chi_{\kappa,m}^{\dagger} \boldsymbol{\sigma} \chi_{-\kappa,m} + \chi_{-\kappa,m}^{\dagger} \boldsymbol{\sigma} \chi_{\kappa,m}) d\tau_n. \tag{13.82}$$

The r_{en}^{-1} term may be expanded in terms of spherical harmonics. Using the vectors $\mathbf{C}_k$ introduced for the electric field, this expansion may be written as

$$\frac{1}{r_{en}} = \sum_k U_k(r_e, r_n) \mathbf{C}_k(\Omega_n) \cdot \mathbf{C}_k(\Omega_e) \tag{13.83}$$

where $U_k(r_e, r_n)$ is a function of the two variables r_e and r_n only, eliminating r_{en}. Using this expression, we may recast the integral above in the form

$$\mathbf{A}^{\text{nuc}} \propto \sum_k N^L N^S \int dr_n r_n^{2\ell-a} U_k(r_e, r_n) e^{-2\eta r_n^2} \big[\langle \kappa, m \,|\, \boldsymbol{\sigma} \mathbf{C}_k(\Omega_n) \,|\, -\kappa, m \rangle \\ + \langle -\kappa, m \,|\, \boldsymbol{\sigma} \mathbf{C}_k(\Omega_n) \,|\, \kappa, m \rangle\big] \cdot \mathbf{C}_k(\Omega_e). \tag{13.84}$$

Evaluation of this integral is straightforward. The radial integrations reduce to a combination of functions $F_m(\eta r^2)$, where $F_m(x) = \int_0^1 t^{2m} \exp(-xt^2)\mathrm{d}t$. These are the standard integrals that appear in any two-electron integral expression. We can extract from these functions a term r^{-k-1} and write the radial integral in the form $r^{-k-1} f_k(r, \eta)$. The angular integrals follow from standard angular momentum algebra. Collecting these into the vector $\boldsymbol{\omega}_k$, we may write the contribution to the Hamiltonian as

$$\hat{\mathcal{H}}^{\text{mag}} \propto \boldsymbol{\alpha} \cdot \sum_k \boldsymbol{\omega}_k \otimes \frac{\mathbf{C}_k(\Omega)}{r^{k+1}} f_k(r, \eta). \tag{13.85}$$

This expression is starting to resemble the point nucleus expression. For any proton j value, there is a term with $k = 1$, which we can write as

$$\boldsymbol{\alpha} \cdot \frac{\boldsymbol{\omega}_1 \times \mathbf{r}}{r^3} f_k(r, \eta). \tag{13.86}$$

In the limit $\eta \to \infty$, $f_k(r, \eta)$, with the associated normalization factors, becomes a constant and the $k = 1$ component reduces to the point nucleus expression in (13.40). However it should be noted that for proton j values greater than 1/2, there are components higher than the magnetic dipole term and that involve higher negative powers of r. For most purposes it is safe to neglect these higher moments and retain only the lowest, dipole term.

Unlike the electric quadrupole moment, contributions to the magnetic interactions with the nucleus can come from (electron) spinors with $j = 1/2$, and thus the magnitude of the finite nucleus effect is likely to be larger than for the electric quadrupole moment.

13.9 Parity-Violating Interactions

We conclude this chapter with a look at some more exotic properties, at least from the point of view of mainstream chemistry. In a 1949 article celebrating Einstein's 70th birthday, Dirac (1949) suggested that the laws of nature might not be invariant with respect to space inversion or time reversal. Special relativity only requires that physical laws be invariant with respect to the position and velocity of the observer, and any change in these can be effected though a series of (infinitesimal) transformations that do not involve reflections of time or space. Experimental evidence for processes that do not conserve parity under space inversion, *P-odd* processes, was eventually observed in nuclear β decay, contributing in turn to the development of the standard model for electroweak interactions. According to this, the P-odd interactions are mediated by the weak nuclear force.

Spatial inversion has an important position in chemistry as the operation that connects two different enantiomers of a chiral molecule. Biochemically it is observed that for living organisms only L-amino acids are present in proteins, and that DNA and RNA are built up from D-sugars. In the wake of the discovery of P-odd processes, suggestions have been made that there may be a connection between this type of interaction and the natural selection of only one enantiomeric form for biochemical processes. It is possible to envision some interaction between molecular structure and the weak force that would favor one of the enantiomers energetically.

To see how this might occur, we look at the operator that would be responsible for breaking the energy degeneracy between two enantiomers. This has the form

$$\hat{\mathcal{H}}^{\mathrm{P}} = \frac{G_F}{2\sqrt{2}} \sum_{N,i} Q_{W,N} \gamma_i^5 \rho(\mathbf{r}_i) \tag{13.87}$$

where the sum is over nuclei, N, and electrons, i. G_F is the Fermi electroweak coupling constant, $Q_{W,N}$ is the weak charge of the nucleus, $\rho(\mathbf{r})$ is the normalized nucleon density, and γ^5 is the chirality operator, which in the representation used by us corresponds to the 4×4 matrix

$$\gamma^5 = \begin{pmatrix} \mathbf{0}_2 & \mathbf{I}_2 \\ \mathbf{I}_2 & \mathbf{0}_2 \end{pmatrix}. \tag{13.88}$$

The chirality operator is odd under spatial inversion, exhibiting what is called *pseudoscalar* behavior, that is, it is a scalar with the transformation properties of a vector. For a molecule with two enantiomeric forms, A and B, with respective wave functions Ψ_A and Ψ_B, we expect space inversion, represented by the operator $\hat{I}$, to connect the two forms such that

$$\hat{I}\,\Psi_A = \Psi_B. \tag{13.89}$$

For a matrix element of a many-electron chirality operator we have

$$\hat{I}\,\langle\Psi_A|\sum_i \gamma_i^5 \,|\,\Psi_A\rangle = -\langle\Psi_B|\sum_i \gamma_i^5 \,|\,\Psi_B\rangle. \tag{13.90}$$

Thus the two enantiomers should differ in energy by

$$\delta E = 2\left|\langle\Psi_A|\hat{\mathcal{H}}^{\mathrm{P}}\,|\,\Psi_A\rangle\right|. \tag{13.91}$$

The energy difference between the two enantiomeric forms of alanine has been calculated to be of order 10^{-17} kJ/mol (Lærdahl et al. 2000) and this is most probably too small to be a decisive factor in the natural predominance of one enantiomer over another, unless coupled to some (unknown) amplification mechanism. Still, it is of interest to be able to demonstrate these P-odd interactions for a chiral molecule, and the calculations can help here in establishing the sensitivity required for their experimental detection. Calculations have been made for halogen-substituted methanes such as CHFClBr and CHFClI (Schwerdtfeger et al. 2002), and the energy shifts found are again of the order of 10^{-17} kJ/mol. The magnitude of this effect is, however, strongly dependent on the nuclear charge, and can be shown to scale as Z^5 or higher. Therefore, the experimental work is more likely to succeed with studies of systems with heavy elements. From the calculated values it is evident that present experimental sensitivity probably is not sufficient to discern these P-odd effects.

There is only one known example of processes that are not invariant under time-reversal, T-odd processes. This is the decay of the neutral K^0 meson. This type of interaction is not encompassed by the standard model, but several possible mechanisms have been suggested. Some of these also predict PT-odd processes. From symmetry considerations it may be shown that a PT-odd interaction would result in an experimental signature characteristic of an effective electric dipole moment. While there are a number of ways such an effective electric dipole moment might arise, the simplest would be if the electron or a nucleon possessed a nonvanishing electric dipole moment. These processes definitely belong more to the realm of particle physics than to chemistry, but it has been recognized (Sandars 1967) that polar molecules containing heavy elements would be favorable for experiments looking for PT-odd interactions. Thus there is also a challenge to relativistic molecular quantum mechanical calculations in predicting these interactions.

If the electron is assumed to have an electric dipole moment of d_e, then the interaction with an electric field **E** is given as

$$\hat{\mathcal{H}}_d^{\mathrm{NR}} = -d_e \boldsymbol{\sigma} \cdot \mathbf{E}. \tag{13.92}$$

In nonrelativistic quantum theory it may be shown that this interaction would give a null result because the charge distribution of the system would react to the presence of an applied field by adjusting such that the net force on the system vanishes (*Schiff's theorem*). In relativistic theory this restriction is lifted and the interaction now takes the form

$$\hat{\mathcal{H}}_d = -d_e \beta \boldsymbol{\Sigma} \cdot \mathbf{E} \tag{13.93}$$

where β is the usual β matrix from the Dirac equation and $\boldsymbol{\Sigma}$ is the usual 4×4 spin matrix. We can exploit the fact that the system must still adjust to the field such that no net force is experienced, and thus

$$\langle -d_e \boldsymbol{\Sigma} \cdot \mathbf{E} \rangle = 0. \tag{13.94}$$

If we subtract the term giving this null contribution from the relativistic operator, we get

$$\hat{\mathcal{H}}_d = -d_e (\beta - \mathbf{I}_4) \boldsymbol{\Sigma} \cdot \mathbf{E} \tag{13.95}$$

where the matrix $(\beta - \mathbf{I}_4)$ takes the form

$$(\beta - \mathbf{I}_4) = \begin{pmatrix} \mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & -2\mathbf{I}_2 \end{pmatrix}. \tag{13.96}$$

Thus the evaluation of this interaction involves only the small component densities. This places quite heavy demands on the accuracy of the calculations. Because the small components normally do not contribute greatly to the energy, variationally based calculations will be less sensitive to the quality of this part of the wave function. This must be taken into account, and the calculations designed accordingly.

As demonstrated by the presence of the $\boldsymbol{\Sigma}$ term, only paramagnetic molecules would be sensitive to this type of interaction involving the electron electric dipole moment. One of the most favorable molecules for experiments is thought to be YbF. Several calculations have been performed on YbF, and the agreement is reasonable. The calculations show that the value of this interaction, if it exists, is such that experiments with a sensitivity corresponding to an electron electric dipole moment of $10^{-28}e$ cm should be able to detect it. Such a sensitivity is reportedly within reach of present experimental techniques.

TlF is considered one of the most favorable molecules for discovering a possible dipole moment of the proton. For a nucleon with an electric dipole moment the effective interaction in a polar molecule would be

$$\hat{\mathcal{H}}^{\text{eff}} = -d\boldsymbol{\Sigma}_N \cdot \boldsymbol{\lambda} \tag{13.97}$$

where $\boldsymbol{\Sigma}_N$ is the nucleon spin operator and $\boldsymbol{\lambda}$ is a vector along the molecular axis. There are several possible contributions to such an effective dipole moment d. The most important is expected to be the *volume effect*, which arises from the interaction between the electric dipole moment of the (heavy) nucleus and the electric field of the electrons assuming that the charge and dipole distributions of the nucleons in the system differ. Through a fairly straightforward derivation (Quiney et al. 1998), this contribution d^v may be shown to be

$$d^v = d_p X R \tag{13.98}$$

where d_p is the electric dipole moment of the proton, R is a nuclear structure factor and X is the electronic contribution

$$X = \lim_{r_n \to 0} \frac{1}{r_n^2} \langle \Psi_e | \sum_i \frac{\cos\theta_i}{r_i^2} | \Psi_e \rangle_{r_n} \tag{13.99}$$

where Ψ_e is the electronic wave function, the sum is over electrons, and the subscript r_n indicates that the radial integration covers the interval $[0, r_n]$. If we assume a DHF wave function with spinors ψ_j, we can write X as

$$\begin{aligned} X &= \sum_j X_j \\ &= \sum_j \lim_{r_n \to 0} \frac{1}{r_n^2} \int_0^{2\pi} \int_0^{\pi} \int_0^{r_n} \psi_j^\dagger(\mathbf{r}) \psi_j(\mathbf{r}) \sin\theta \cos\theta dr\, d\theta\, d\phi. \end{aligned} \tag{13.100}$$

As $r_n \to 0$ this only involves the numerical values of the 4-spinor amplitudes at $r = 0$.

For a system with one heavy and one light nucleus (such as TlF), the contributions will be dominated by the heavy atom, and we may analyze the X_j in terms of the contributions from a one-center expansion of the spinor. We can write

$$\psi_j = \sum_\kappa |\kappa, m\rangle = |-1, m\rangle + |1, m\rangle + |-2, m\rangle + \cdots \tag{13.101}$$

(m being a good quantum number in a diatomic molecule). The most critical contributions to X_j turn out to be those involving the $|-1, 1/2\rangle$ and the $|1, 1/2\rangle$ components of this expansion. Using the atomic expansion for small r discussed in section 7.5, we can show that this contribution is proportional to a simple product of the lowest-order expansion coefficients

$$X_j^{-1,1} \propto p_0^- p_0^+ + q_0^- q_0^+ \tag{13.102}$$

where the superscript signs refer to coefficients arising from the expansion of the $|-1, 1/2\rangle$ or the $|1, 1/2\rangle$ components respectively. From the ratios of the p_0 and q_0, it follows that

$$\begin{aligned} X_j^{-1,1} &\propto p_0^- p_0^+ - \frac{v_1 + E}{v_1 + E + 2c^2} p_0^- p_0^+ \\ &= p_0^- p_0^+ \left[1 - \frac{v_1 + E}{v_1 + E + 2c^2}\right] \approx \frac{2c^2}{v_1} p_0^- p_0^+ \end{aligned} \tag{13.103}$$

where the last approximation follows from the fact that v_1 is very large—for Tl it is close to 10^6—and thus $v_1 >> c^2$.

The main point to note here is the extremely delicate cancellation involved between the $p_0^- p_0^+$ and $q_0^- q_0^+$ terms. In reality the matrix elements X_j are calculated not from the atomic expansion, but from DHF spinors. If we are to succeed, our analysis above has clearly indicated that the ratio between the small and large component for small r is crucial. Again this puts high demands on the quality of the calculations, and also on the basis sets used. Thus, in the TlF calculation of Quiney et al. (1998), energy-optimized basis sets proved to be unsuitable and had to be replaced by large even-tempered basis sets, even if the total molecular energy obtained using the energy-optimized set was lower. This provides another example of how the variational process may not adequately describe regions that are of crucial importance for the calculation of properties. As the ratio of the lowest-order expansion coefficients is known for the atoms, one can monitor the quality of the basis set by comparing the results from calculations on the heavy atom only with the theoretical values of these expansion coefficients ratios.

For each of the three properties discussed in this section, there are also other interactions that may contribute, but the ones presented here are believed to be dominant. We have chosen these because each one illustrates a particular point: the P-odd interaction connects to the important chemical concept of chirality, the electron electric dipole moment illustrates a property that depends solely on the small-component density, placing heavy demands on the quality of the small-component wave function, and the proton electric dipole moment demonstrates a case where the ratios of the large and small components in the nuclear region is crucial.

There is an almost unlimited number of properties that might have been discussed further, including QED effects, nuclear multipole moments, and various polarizabilities. The aim of this chapter has not been to give a complete review, but rather to provide an introduction to some of the features present and issues that arise in the calculation of properties within a fully relativistic model. The treatment of properties within more approximate models will be discussed to some extent in connection with the presentation of these models in the following chapters.

14

Density Functional Approaches to Relativistic Quantum Mechanics

The wave function is an elusive and somewhat mysterious object. Nobody has ever observed the wave function directly: rather, its existence is inferred from the various experiments whose outcome is most rationally explained using a wave function interpretation of quantum mechanics[1]. Further, the N-particle wave function is a rather complicated construction, depending on $3N$ spatial coordinates as well as N spin coordinates, correlated in a manner that almost defies description. By contrast, the electron density of an N-electron system is a much simpler quantity, described by three spatial coordinates and even accessible to experiment. In terms of the wave function, the electron density is expressed as

$$\rho(\mathbf{r}) = N \int \Psi^*(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) d\mathbf{r}_2 d\mathbf{r}_3 \ldots \mathrm{d}\mathbf{r}_N \tag{14.1}$$

where the sum over spin coordinates is implicit. It might be much more convenient to have a theory based on the electron density rather than the wave function. The description would be much simpler, and with a greatly reduced (and constant) number of variables, the calculation of the electron density would hopefully be faster and less demanding. We also note that given the correct ground state density, we should be able to calculate any observable quantity of a stationary system. The answer to these hopes is density functional theory, or DFT. Over the past decade, DFT has become one of the most widely used tools of the computational chemist, and in particular for systems of some size. This success has come despite complaints about arbitrary parametrization of potentials, and laments about the absence of a universal principle (other than comparison with experiment) that can guide improvements in the way the variational principle has led the development of wave-function-based methods. We do not intend to pursue that particular discussion, but we note as a historical fact that

---

[1] In the strict quantum mechanical sense, this nonobservation is only natural because the wave function is not a quantum mechanical observable.

many important early contributions to relativistic quantum chemistry were made using DFT-like methods. Furthermore, there is every reason to try to extend the success of nonrelativistic DFT methods to the relativistic domain. We suspect that their potential for conquering a sizable part of this field is at least as large as it has been in the nonrelativistic domain.

One requirement for a successful extension of DFT to the relativistic domain is a reformulation of the basic theorems of DFT within a relativistic framework. We therefore start this chapter with a brief review of basic nonrelativistic DFT, before we show how these ideas appear in a relativistic setting.

14.1 A Brief Review of Nonrelativistic Density Functional Theory

Modern DFT is founded on the Hohenberg–Kohn theorems and the Kohn–Sham equations. These are presented in detail in textbooks as well as in the review literature on the subject (Parr and Yang 1989, Koch and Holthausen 2001, Eschrig 1996, Gross and Kurth 1994, Salahub et al. 1994). However, to set the stage for a discussion of the relativistic case, a brief summary of the nonrelativistic foundations serves as a convenient starting point.

We write the time-independent Hamiltonian for the N-electron system as

$$\hat{\mathcal{H}} = \hat{T} + U + V \tag{14.2}$$

where $\hat{T}$ is the contribution from the kinetic energy, U is the potential of the interaction between the N electrons, and V is the potential due to external fields. In the simple Born–Oppenheimer molecule, V is the potential from the fixed framework of nuclear charges. For a given N, it is the term V that gives the system its “identity.” As an example, in a 10-electron system, it is the V term that determines whether the system is a Ne atom or an HF or H_2O molecule. The two other terms are universal for all 10-electron systems. This observation sets the stage for the first Hohenberg–Kohn theorem. Writing

$$V = \sum_{i=1}^{N} v(\mathbf{r}_i), \tag{14.3}$$

this theorem states that: *The potential $v(\mathbf{r})$ is completely determined by the electron density $\rho(\mathbf{r})$ to within an additive constant.*

To prove this theorem we must now show a one-to-one correspondence between $v(\mathbf{r})$ and $\rho(\mathbf{r})$, and we do this using the wave function as an intermediary. Assuming a nondegenerate ground state with energy E_0, the Schrödinger equation

$$\hat{\mathcal{H}}\Psi = E_0\Psi \tag{14.4}$$

defines a mapping from each $v(\mathbf{r})$ to the ground state wave function Ψ. From this wave function, we can find the ground state density $\rho(\mathbf{r})$ according to (14.1). To establish the reverse mapping from $\rho(\mathbf{r})$ to $v(\mathbf{r})$, we note that a mapping from the wave function to the potential is provided (formally) by solving the Schrödinger equation for the potential, given a wave function. It remains to show that there is a unique mapping from $\rho(\mathbf{r})$ to Ψ, which is equivalent to showing that two different ground states Ψ and Ψ' arising from different potentials $v(\mathbf{r})$ and $v'(\mathbf{r})$ always give different ground state densities. Calling the corresponding Hamiltonians $\hat{\mathcal{H}}$ and $\hat{\mathcal{H}}'$, E_0 is given by

$$\begin{aligned} E_0 = \langle \Psi|\hat{\mathcal{H}}|\Psi\rangle < \langle \Psi'|\hat{\mathcal{H}}|\Psi'\rangle &= \langle \Psi'|\hat{\mathcal{H}}' - V' + V|\Psi'\rangle \\ &= E_0' + \int \rho'(\mathbf{r}) \left[v'(\mathbf{r}) - v(\mathbf{r})\right] \mathrm{d}\mathbf{r}. \end{aligned} \tag{14.5}$$

Likewise, we can show that

$$E_0' < E_0 - \int \rho(\mathbf{r}) \left[v'(\mathbf{r}) - v(\mathbf{r})\right] \mathrm{d}\mathbf{r}. \tag{14.6}$$

Adding these two inequalities we get

$$E_0 + E_0' < E_0' + E_0 + \int \left[\rho'(\mathbf{r}) - \rho(\mathbf{r})\right] \left[v'(\mathbf{r}) - v(\mathbf{r})\right] \mathrm{d}\mathbf{r}. \tag{14.7}$$

By hypothesis, $v'(\mathbf{r})$ and $v(\mathbf{r})$ are different. If we assume that $\rho'(\mathbf{r}) = \rho(\mathbf{r})$ we get the self-contradiction

$$E_0 + E_0' < E_0' + E_0. \tag{14.8}$$

Thus, we conclude that two different densities must correspond to two different potentials, and a nondegenerate ground state can correspond to one and only one potential.

The second Hohenberg–Kohn theorem establishes a variational principle for the energy as a functional of the density: *If* $\rho(\mathbf{r})$ *is the density arising from the solution of the N-electron Schrödinger equation*

$$\hat{\mathcal{H}}\Psi = E[\rho]\Psi \tag{14.9}$$

then for any density $\rho'(\mathbf{r}) \neq \rho(\mathbf{r})$ *that satisfies*

$$\int \rho'(\mathbf{r})\mathrm{d}\mathbf{r} = N \tag{14.10}$$

we must have $E[\rho'] > E[\rho]$.

This theorem follows readily from the 1:1 correspondence between Ψ and ρ and the Rayleigh–Ritz variational principle. By incorporating the N-electron constraint on the density using a Lagrangian multiplier μ, we can formulate the stationarity condition at the minimum as

$$\delta\left\{E[\rho]-\mu\left[\int\rho(\mathbf{r})\mathrm{d}\mathbf{r}-N\right]\right\}=0. \tag{14.11}$$

Substituting for the energy functional in this expression

$$\begin{aligned}E[\rho]&=\langle\Psi|\hat{\mathcal{H}}|\Psi\rangle=\langle\Psi|\hat{T}+\hat{U}+V|\Psi\rangle=\langle\Psi|\hat{T}+\hat{U}|\Psi\rangle+\int\rho(\mathbf{r})v(\mathbf{r})\mathrm{d}\mathbf{r}\\&=F[\rho]+\int\rho(\mathbf{r})v(\mathbf{r})\mathrm{d}\mathbf{r}\end{aligned} \tag{14.12}$$

the stationarity condition may be recast in the Euler–Lagrange form

$$\mu=\frac{\delta E[\rho]}{\delta\rho(\mathbf{r})}=v(\mathbf{r})+\frac{\delta F[\rho]}{\delta\rho(\mathbf{r})} \tag{14.13}$$

where $F[\rho]$ is the universal functional for all systems of N particles with Coulomb interactions.

While this equation is quite simple in form, it hides the fact that the universal functional $F[\rho]$ is not available in explicit form. Numerous schemes have been formulated for a direct optimization of the density based on these stationarity conditions, but these methods have not really been competitive. The most efficient approach has been to invoke a quasi-independent-particle approximation, formulated in the Kohn–Sham equations.

To arrive at the Kohn–Sham equations, we assume the existence of a set of orbitals $\{\phi_i(\mathbf{r})\}$ such that the density may be expressed in terms of these orbitals as

$$\rho(\mathbf{r})=\sum_{i=1}^{N}\phi_i(\mathbf{r})^*\phi_i(\mathbf{r}). \tag{14.14}$$

For a set of noninteracting particles, the term $\hat{U}$ vanishes and the Hamiltonian is

$$\hat{\mathcal{H}}^{\mathrm{ni}}=\hat{T}^{\mathrm{ni}}+V^{\mathrm{ni}}=\sum_{i=1}^{N}\left(-\frac{1}{2}\nabla^2+v(\mathbf{r}_i)\right). \tag{14.15}$$

The ground state of this system is described by the determinantal wave function

$$\Psi^{\mathrm{ni}}=|\phi_1^{\mathrm{ni}}(\mathbf{r}_1)\phi_2^{\mathrm{ni}}(\mathbf{r}_2)\ldots\phi_N^{\mathrm{ni}}(\mathbf{r}_N)| \tag{14.16}$$

where $\{\phi_i^{\mathrm{ni}}(\mathbf{r}_i);\, i=1,2,\ldots,N\}$ are the solutions of the equation

$$\left(-\frac{1}{2}\nabla^2+v(\mathbf{r}_i)\right)\phi_i^{\mathrm{ni}}(\mathbf{r}_i)=\epsilon_i^{\mathrm{ni}}\phi_i^{\mathrm{ni}}(\mathbf{r}_i) \tag{14.17}$$

corresponding to the lowest N eigenvalues.

In a system of interacting electrons, the term $\hat{U}$ must be included in the Hamiltonian. A large part of this is the Coulomb interaction, and we also expect this to contain some sort of exchange interaction as well as more obscure contributions due to correlation. We define an effective single-particle potential by

$$\int v^{\mathrm{eff,i}}(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} = \int v(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} + \frac{1}{2}\int\int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' + E^{\mathrm{xc,i}}[\rho] \tag{14.18}$$

where the superscript i indicates that these are for the interacting system. Here we have indicated explicitly the two "simple" terms—the external potential and the direct Coulomb interaction between the charge densities. The third term takes care of other exchange and correlation contributions to the interaction potential.

However, the introduction of particle interactions in the system may also influence the functional form of the kinetic energy term. It is tempting to assume that it remains the same as in the noninteracting case, $\hat{T}^{\mathrm{ni}}$, but we really have no assurance that this is so. Thus we write the exact kinetic energy functional for the interacting system as

$$\hat{T}[\rho] = \hat{T}^{\mathrm{ni}}[\rho] + \Delta\hat{T}[\rho] \tag{14.19}$$

where the last term incorporates any changes in the exact kinetic energy functional due to the electron–electron interaction. Collecting all terms, we can express the energy functional for the interacting system as

$$E[\rho] = \hat{T}^{\mathrm{ni}}[\rho] + \int v(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} + \frac{1}{2}\int\int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' + E^{\mathrm{xc}}[\rho] \tag{14.20}$$

where the correction to the kinetic energy functional has been included in the exchange-correlation energy functional

$$E^{\mathrm{xc}}[\rho] = E^{\mathrm{xc,i}}[\rho] + \Delta\hat{T}[\rho]. \tag{14.21}$$

The Euler–Lagrange equation now becomes

$$\mu = \frac{\delta E[\rho]}{\delta\rho(\mathbf{r})} = \frac{\delta\hat{T}^{\mathrm{ni}}[\rho]}{\delta\rho(\mathbf{r})} + v(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' + v^{\mathrm{xc}}(\mathbf{r}) \tag{14.22}$$

where the exchange-correlation potential is defined by

$$v^{\mathrm{xc}}(\mathbf{r}) = \frac{\delta E^{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})}. \tag{14.23}$$

We note that the last three terms of the Euler–Lagrange equation above are similar to the effective potential $v^{\mathrm{eff,i}}(\mathbf{r})$ introduced above, except that now we have included the correction to the kinetic energy functional. We can write the Euler–Lagrange equation as

$$\mu = \frac{\delta \hat{T}^{\mathrm{ni}}[\rho]}{\delta \rho(\mathbf{r})} + v^{\mathrm{eff}}(\mathbf{r})\rho(\mathbf{r}). \tag{14.24}$$

But the Euler–Lagrange equation for the noninteracting particles was

$$\mu = \frac{\delta \hat{T}^{\mathrm{ni}}[\rho]}{\delta \rho(\mathbf{r})} + v(\mathbf{r})\rho(\mathbf{r}) \tag{14.25}$$

and this was equivalent to the set of single-particle Schrödinger equations (14.17). By analogy we conclude that for the interacting particles we can find the orbitals from the equations

$$\left(-\frac{1}{2}\nabla^2 + v^{\mathrm{eff}}(\mathbf{r}_i)\right)\phi_i(\mathbf{r}_i) = \epsilon_i \phi_i(\mathbf{r}_i). \tag{14.26}$$

These are the Kohn-Sham equations, and we note that the density derived from (14.14) automatically fulfills the N-particle constraint, because the orbitals are the normalized, orthogonal solutions of an eigenvalue equation.

14.2 The Local Density and Local Exchange Approximations

In the derivation of the Kohn–Sham equations we have hidden a number of difficulties in the exchange-correlation potential, $v^{\mathrm{xc}}(\mathbf{r})$. Indeed, the success of DFT depends on finding an accurate and convenient form of this potential. There is an extensive literature discussing the merits of various potentials, and good accounts of these may be found elsewhere (Koch and Holthausen 2001). Here, we restrict the discussion to the local density approximation (LDA), because it provides a link to another approximation that has been used extensively in relativistic atomic and molecular calculations and which predates the Kohn–Sham equations.

In this approximation, we assume that the exchange-correlation energy is a local function of the density. Following Parr and Yang (1989) we write the exchange-correlation energy per electron as $\epsilon^{\mathrm{xc}}(\rho)$, and the entire energy contribution from this term may be written as

$$E^{\mathrm{xc,LDA}}[\rho] = \int \rho(\mathbf{r})\epsilon^{\mathrm{xc}}(\rho)d\mathbf{r} \tag{14.27}$$

with the corresponding potential

$$v^{\mathrm{xc,LDA}}(\mathbf{r}) = \epsilon^{\mathrm{xc}}(\rho) + \rho(\mathbf{r})\frac{\delta \epsilon^{\mathrm{xc}}(\rho)}{\delta \rho(\mathbf{r})}. \tag{14.28}$$

We also assume separability of exchange and correlation such that

$$\epsilon^{\mathrm{xc}}(\rho) = \epsilon^{\mathrm{x}}(\rho) + \epsilon^{\mathrm{c}}(\rho). \tag{14.29}$$

There is one N-electron system where it is possible to derive a closed, local form of the exchange energy as a function of the density. This is the uniform electron gas. For this model system, Dirac has derived an expression for the exchange energy using statistical considerations. The result is

$$\epsilon^{\mathrm{x}}(\rho) = -\frac{3}{4}\left(\frac{3\rho}{\pi}\right)^{1/3} \tag{14.30}$$

and if we insert this expression into (14.28), the exchange contribution to the potential becomes

$$v^{\mathrm{x,LDA}}(\mathbf{r}) = -\left(\frac{3\rho}{\pi}\right)^{1/3}. \tag{14.31}$$

For the correlation contribution no analytic form has been found, but extensive Monte Carlo calculations have provided values of $\epsilon^{\mathrm{c}}(\rho)$ over a wide range of densities. These values have been fitted analytically, giving a very good approximation of the correlation energy of the uniform electron gas. The Kohn–Sham equations now take the form

$$\left(-\frac{1}{2}\nabla^2 + v(\mathbf{r}_i) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' - \left(\frac{3\rho}{\pi}\right)^{1/3} + v^{\mathrm{c}}(\mathbf{r}_i)\right)\phi_i(\mathbf{r}_i) = \epsilon_i\phi_i(\mathbf{r}_i). \tag{14.32}$$

Local density approximations have been used in quantum chemistry and atomic physics also for other purposes. If we write the Hartree–Fock equations in the form

$$\left(-\frac{1}{2}\nabla^2 + v(\mathbf{r}_i) + \sum_j\left(\hat{J}_j - \hat{K}_j\right)\right)\phi = \epsilon\phi \tag{14.33}$$

the only nonlocal operator is the exchange operator $\hat{K}_j$. In the early days of computational efforts, the evaluation of this term was a major obstacle. Slater therefore proposed that the exchange interaction should be approximated by a local exchange potential of the form

$$v^{\mathrm{X}\alpha}(\mathbf{r}) = -\frac{3}{2}\alpha\left(\frac{3\rho}{\pi}\right)^{1/3} \tag{14.34}$$

where α is a parameter introduced to provide some flexibility. This forms the basis of the $X\alpha$ model, where the resulting Hartree–Fock–Slater equations for the orbitals are written

$$\left(-\frac{1}{2}\nabla^2 + v(\mathbf{r}_i) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' - \frac{3}{2}\alpha\left(\frac{3\rho}{\pi}\right)^{1/3}\right)\phi = \epsilon\phi \tag{14.35}$$

which for the value $\alpha = \frac{2}{3}$ is identical to the LDA Kohn–Sham equations without the correlation potential. In actual calculations using Slater exchange, α was frequently given the value 0.7, a number arrived at from experience by comparison with experimental results.

14.3 The Hohenberg–Kohn Theorem for Relativistic *N*-Particle Systems

The extension of the Hohenberg–Kohn theorem to the case of a relativistic N-particle system was first made by Rajagopal and Callaway (Rajagopal 1978, Rajagopal and Callaway 1973) for the inhomogeneous electron gas. More recently, a thorough discussion of this extension has been presented by Engel and coworkers (Engel 2002, Engel and Dreizler 1996, Engel et al. 1995). A rigorous treatment requires the use of quantum field theory beyond the level we want to pursue in this text. We therefore adopt a more pragmatic view and use a simplified approach that demonstrates the most important features. We then briefly describe the additional considerations required for a rigorous treatment.

Let us start with the usual N-electron Dirac Hamiltonian including the external fields. The energy contribution from the external field is given by

$$E^{\text{ext}} = \langle\Psi|\sum_i ec\boldsymbol{\alpha}(\mathbf{r}_i)\cdot\mathbf{A}(\mathbf{r}_i)|\Psi\rangle - \langle\Psi|\sum_i e\phi(\mathbf{r}_i)|\Psi\rangle. \tag{14.36}$$

With our previous expressions for the current density and charge density in (4.41) and (4.42) we may write this energy contribution as

$$e\langle\mathbf{j}\cdot\mathbf{A}\rangle - e\langle\rho\phi\rangle = e\langle\mathrm{j}\cdot\mathrm{A}\rangle \tag{14.37}$$

with j and A as the four-vector representation of the current and the field. The expectation value of the Dirac Hamiltonian may now be written as

$$\langle\Psi|\hat{\mathcal{H}}|\Psi\rangle = \hat{T}^{\mathrm{D}} + U^{\mathrm{R}} + e\langle\mathrm{j}\cdot\mathrm{A}\rangle \tag{14.38}$$

where $\hat{T}^{\mathrm{D}}$ derives from the expectation value of the Dirac free-particle operators, and U^{R} from the relativistic electron–electron interaction term (in the approximation of our choice). This expression demonstrates that in the relativistic model, the four-current j plays the same role as the density does in the nonrelativistic case.

Starting from these considerations, we can repeat the reasoning of the nonrelativistic Hohenberg–Kohn theorem in a relativistic setting. We thus need to show the 1:1

correspondence between j and A. We assume the existence of two different fields A and A$'$ yielding the Hamiltonians $\hat{\mathcal{H}}$ and $\hat{\mathcal{H}}'$, respectively, such that

$$\hat{\mathcal{H}}\Psi = E\Psi \ \ ; \ \ \hat{\mathcal{H}}'\Psi' = E'\Psi'. \tag{14.39}$$

If we follow the lines of the nonrelativistic proof, we need to establish the inequality

$$E = \langle\Psi|\hat{\mathcal{H}}|\Psi\rangle < \langle\Psi'|\hat{\mathcal{H}}|\Psi'\rangle. \tag{14.40}$$

However, even with the assumption of a nondegenerate ground state, this is problematic. The nonrelativistic argument is based on the Ritz variational principle, but Engel and Dreizler (1996) point out that there is no rigorous minimum principle for the renormalized energies of field theoretical systems. To proceed, they argue the existence of a similar minimum principle for the relativistic case by noting first that the renormalized energies must approach values that satisfy the Ritz principle in the nonrelativistic limit, and second that real systems are stable and most accurately described by QED. This at least would appear to make the existence of a minimum principle probable also for the relativistic case. If we believe this, we may proceed as before and write

$$E < \langle\Psi'|\hat{\mathcal{H}}'|\Psi'\rangle + e\langle\, \mathrm{j}' \cdot \mathrm{A}\rangle - e\langle\, \mathrm{j}' \cdot \mathrm{A}'\rangle = E' + e\langle\, \mathrm{j}' \cdot \mathrm{A}\rangle - e\langle\, \mathrm{j}' \cdot \mathrm{A}'\rangle. \tag{14.41}$$

Correspondingly, we get

$$E' < E + e\langle\, \mathrm{j} \cdot \mathrm{A}'\rangle - e\langle\, \mathrm{j} \cdot \mathrm{A}\rangle. \tag{14.42}$$

We add the two inequalities and get

$$E + E' < E' + E + e\langle(\mathrm{j}' - \mathrm{j}) \cdot (\mathrm{A} - \mathrm{A}')\rangle, \tag{14.43}$$

and making the assumption $\mathrm{j}' = \mathrm{j}$ we again get a contradiction. The problem with this derivation is that while we have used the assumption of a nondegenerate ground state, we have not considered possible gauge transformations of the external field. The Hamiltonian must be assumed gauge invariant, and thus if A and A$'$ only differ by a gauge transformation, a similar development leads not to the inequality above but to the equality

$$2E = 2E + e\langle(\mathrm{j}' - \mathrm{j}) \cdot (\mathrm{A} - \mathrm{A}')\rangle, \tag{14.44}$$

indicating that for this case indeed $\mathrm{j}' = \mathrm{j}$. We must therefore conclude that there is a 1:1 correspondence between the four-current and a class of external potentials that only differ by a gauge transformation. The ground-state energy becomes a functional of the four-current j

$$E[\mathrm{j}\,] = \langle\Psi[\mathrm{j}\,]|\hat{\mathcal{H}}|\Psi[\,\mathrm{j}\,]\rangle \tag{14.45}$$

and with the minimum principle introduced above, we get the variational equation for the optimal value of j

$$\frac{\delta}{\delta \mathrm{j}(\mathbf{x})}\left\{E[\mathrm{j}]-\mu\int\rho(\mathbf{x})\mathrm{d}\mathbf{x}\right\}=0. \tag{14.46}$$

This outlines the main steps in developing a relativistic Hohenberg–Kohn theorem. The requirements for a stringent derivation go considerably beyond this. A complete derivation would involve:

- Additional terms in the Hamiltonian to describe the free radiation as well as interaction with the transverse photons (giving rise to the Breit interaction).
- Renormalization or introduction of counter terms to avoid the divergences arising from the infinite Dirac sea in the vacuum and from ultraviolet (UV) divergences of QED.
- A more rigorous treatment of gauge invariance.

For a further exploration of these points, the reader should consult the article by Engel (2002).

From the minimum principle we may go on to obtain the relativistic Kohn–Sham equations

$$\left\{c\boldsymbol{\alpha}\cdot\mathbf{p}+mc^2(\beta-\mathbf{I}_4)+v^{\mathrm{eff}}[\mathrm{j}]\right\}\psi(\mathbf{x})=\epsilon\psi(\mathbf{x}) \tag{14.47}$$

where the effective potential may be written in the form

$$\begin{aligned} v^{\mathrm{eff}}[\mathbf{j},\rho]=&-\left(e\phi(\mathbf{r})+e^2\int\frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}'+\frac{\delta E^{\mathrm{xc}}}{\delta\rho(\mathbf{r})}\right)\\ &-c\boldsymbol{\alpha}\cdot\left(e\mathbf{A}(\mathbf{r})+e^2\int\frac{\rho(\mathbf{j}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}'+\frac{\delta E^{\mathrm{xc}}}{\delta\mathbf{j}(\mathbf{r})}\right)\end{aligned} \tag{14.48}$$

which explicitly shows the contributions from the charge density and current density parts of the four-current. These equations are also known as the Dirac–Kohn–Sham equations.

14.4 Density Functional Theory and the Dirac–Coulomb Hamiltonian

As described previously, in practical computational schemes of relativistic quantum chemistry the Dirac–Coulomb Hamiltonian is frequently preferred. The use of this Hamiltonian, which retains only the Coulomb interaction between the electrons, omitting the Breit term and higher-order corrections, is justified by the argument that these terms are small and the effect on most properties of chemical interest is insignificant. The validity of this argument appears to be supported by actual applications. Saue and Helgaker (2002) have shown that when this Hamiltonian is used in the absence of magnetic fields, it is consistent to reduce the relativistic DFT to a theory based on functionals of the charge density only, avoiding reference to the current density.

To show this, we note that in the absence of external magnetic fields, the external potential reduces to the scalar term

$$V^{\text{ext}}[\rho] = \int \rho(\mathbf{r})v(\mathbf{r})\mathrm{d}\mathbf{r}. \tag{14.49}$$

Thus, the only way the current density **j** can contribute to the energy would be through the electron–electron interaction. From (5.49) the operator for this interaction is

$$\begin{aligned} V^{\mathrm{C}}(0, r_{ij}) &= \frac{1}{r_{ij}} - \frac{\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j}{r_{ij}} + \frac{(\boldsymbol{\alpha}_i \times \mathbf{r}_{ij}) \cdot (\boldsymbol{\alpha}_j \times \mathbf{r}_{ij})}{2r_{ij}^3} \\ &= g^{\mathrm{C}}(\mathbf{r}_i, \mathbf{r}_j) + g^{\mathrm{G}}(\mathbf{r}_i, \mathbf{r}_j) + g^{\text{gauge}}(\mathbf{r}_i, \mathbf{r}_j). \end{aligned} \tag{14.50}$$

Disregarding for the moment the gauge term, we consider the contribution to the functional U from the two other terms. This may be written as

$$\begin{aligned} U^{\mathrm{CG}}[\mathrm{j}] &= \frac{1}{2}\sum_{i,j}\langle\Psi|\frac{1}{r_{ij}}|\Psi\rangle - \frac{1}{2}\sum_{i,j}\langle\Psi|\frac{\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j}{r_{ij}}|\Psi\rangle \\ &= \frac{1}{2}\int \frac{\rho_2(\mathbf{r}_1, \mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 - \frac{1}{2c^2}\int \frac{\mathbf{j}_2(\mathbf{r}_1, \mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 \end{aligned} \tag{14.51}$$

where we have introduced the two-particle charge density

$$\rho_2(\mathbf{r}_1, \mathbf{r}_2) = N(N-1)\int \Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_N)^{\dagger}\Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_N)d\mathbf{r}_3 \ldots d\mathbf{r}_N \tag{14.52}$$

and an analogous two-particle current density

$$\begin{aligned} \mathbf{j}_2(\mathbf{r}_1, \mathbf{r}_2) = N(N-1)\int & \Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_N)^{\dagger}c^2\boldsymbol{\alpha}_1 \\ & \cdot \boldsymbol{\alpha}_2\Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_N)d\mathbf{r}_3 \ldots d\mathbf{r}_N. \end{aligned} \tag{14.53}$$

The two-particle density may be rewritten in the form

$$\rho_2(\mathbf{r}_1, \mathbf{r}_2) = \rho(\mathbf{r}_1)\rho(\mathbf{r}_2) + \rho(\mathbf{r}_1)\hat{h}^{\mathrm{xc}}(\mathbf{r}_1; \mathbf{r}_2) \tag{14.54}$$

using the (formal) operator $\hat{h}^{\mathrm{xc}}$, which gives rise to the exchange-correlation hole. With this expression, the contribution from the charge density to the electron–electron interaction functional $U^{\mathrm{CG}}[\mathrm{j}]$ becomes

$$\frac{1}{2}\int\frac{\rho_2(\mathbf{r}_1,\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2=\frac{1}{2}\int\frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2+\frac{1}{2}\int\frac{\rho(\mathbf{r}_1)\hat{h}^{\mathrm{xc}}(\mathbf{r}_1,\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2. \tag{14.55}$$

The first term on the right-hand side may be recognized as the Hartree term (or direct interaction term) J and the other is a nonclassical term containing the effects of exchange and correlation. The same formal rewriting may, of course, be done for the current density term

$$\mathbf{j}_2(\mathbf{r}_1,\mathbf{r}_2)=\mathbf{j}(\mathbf{r}_1)\mathbf{j}(\mathbf{r}_2)+\mathbf{j}(\mathbf{r}_1)\hat{h}^{\mathrm{jxc}}(\mathbf{r}_1;\mathbf{r}_2) \tag{14.56}$$

where $\hat{h}^{\mathrm{jxc}}$ is the exchange-correlation hole operator for the current density. The contributions to $U^{\mathrm{CG}}[\mathrm{j}]$ are analogous to those from the charge density.

If we stop to reflect on this situation, it becomes clear that in the absence of external vector potentials the current density only contributes to the energy functional through the Gaunt term (or in the general case, the Breit term). Thus, if we use the Dirac–Coulomb Hamiltonian, it is reasonable and consistent to regard the exchange-correlation potential as a function of the charge density alone, omitting all references to the current density. The neglect of the current density is in any case likely to cause only small errors in practice, because for closed shells only the exchange part is nonzero. However, the spin or current density is required for the proper description of excitations, time-dependent properties and static magnetic properties.

Even with this approximation, we cannot assume that the relativistic exchange-correlation functional is the same as the nonrelativistic functional, though we might reasonably suppose that an expansion of the relativistic functional will yield the nonrelativistic functional with correction terms. This proves to be the case for the exchange energy of a relativistic uniform electron gas, for which we can write the exchange energy per electron as

$$\epsilon^{\mathrm{x,RLDA}}(\rho)=\epsilon^{\mathrm{x,NRLDA}}(\rho)\left[1-\frac{2}{3}\delta^2+\mathcal{O}(\delta^4)\right] \tag{14.57}$$

where the parameter δ is given by

$$\delta=\frac{\left(3\pi^2\rho\right)^{1/3}}{mc} \tag{14.58}$$

and $\epsilon^{\mathrm{x,NRLDA}}$ is the nonrelativistic local exchange energy given in (14.30). Part of this correction comes from a two-electron Darwin correction and part is due to the transverse (Breit) interaction. If we are neglecting the corrections to the Coulomb interaction as small, we can also safely neglect this correction and use nonrelativistic exchange-correlation functionals.

Starting from the Dirac–Coulomb approximation, a set of Dirac–Kohn–Sham equations may again be derived. In chapter 8, a spinor-rotation procedure was used to derive the relativistic Fock operator. A similar procedure applied to the present case shows that the gradient of the energy has elements the form

$$g_{ai} = -h_{ia} - \sum_{j} (ia|jj) - \int v^{\mathrm{xc}}[\rho]\psi_i{}^{\dagger}\psi_a. \tag{14.59}$$

For the optimal set of Dirac–Kohn–Sham spinors the gradient must disappear, and this can be achieved by rotating to the spinor set that diagonalizes the Dirac–Kohn–Sham matrix $\mathbf{f}^{\mathrm{DKS}}$ with elements

$$f_{ia}^{\mathrm{DKS}} = -g_{ai}. \tag{14.60}$$

The Dirac–Kohn–Sham equation can be expressed in terms of the large and small components as

$$\begin{pmatrix} V^{\mathrm{nuc}} + J + v^{\mathrm{xc}} & c(\boldsymbol{\sigma} \cdot \mathbf{p}) \\ c(\boldsymbol{\sigma} \cdot \mathbf{p}) & V^{\mathrm{nuc}} + J + v^{\mathrm{xc}} - 2mc^2 \end{pmatrix} \begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} = E \begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} \tag{14.61}$$

where the external potential $v(\mathbf{r})$ has been replaced by V^{nuc} for the purpose of connection with the developments in other parts of this book.

For the majority of quantum chemical problems suitable for a DFT treatment, the accuracy ambitions are compatible with use of the Dirac–Coulomb Hamiltonian. The development above shows that in those cases one can with good conscience borrow freely from the exchange-correlation functionals of ordinary nonrelativistic DFT. This includes the use of gradient-corrected and hybrid functionals, which are easily implemented within the formalism above. There exist also explicitly relativistic exchange-correlation functionals, which may be used if a more consistently relativistic approach is desirable. For quantum chemical applications it is not clear that such functionals offer any real advantages over the standard nonrelativistic functionals. In open-shell systems, spin polarization is important, and nonrelativistic calculations rely on spin-density functionals. The application of these functionals in a relativistic context is not straightforward. The discussion of the merits of the various exchange-correlation functionals and spin-density functionals is beyond the scope of this book, and the reader is referred to the DFT specialist literature for a more extensive treatment of these subjects (Koch and Holthausen 2001, Eschrig 1996, Engel 2002). For heavy-element calculations it should be remembered that the nonrelativistic functionals are parametrized for light elements and might therefore be less accurate for heavy elements.

The implementation of relativistic DFT methods for molecular calculations follows by straightforward extension of nonrelativistic implementations. The major features of the implementation are the tabulation of the exchange-correlation potential on a grid, and subsequent numerical integration to obtain the Fock matrix elements. With modern computing technology, the time taken to evaluate the exchange-correlation potential, even with gradient corrections, is relatively minor. The evaluation of the Coulomb potential is often aided by an expansion of the density in an auxiliary basis set, which reduces the formal scaling of the method with the number of basis functions to n^3.

Due to these techniques, nonrelativistic DFT has established itself as a very convenient and economical method for performing calculations that account for essential parts of the correlation energy. As such it has become the method of choice for a sizable community of users, in particular those who require moderately accurate results for large molecular systems. The ease of implementation has meant that several existing four-component relativistic programs (DIRAC 2004, BERTHA 2002, BDF 1997, UTChem: Yanai et al. 2001a, 2001b) and a number of two-component or one-component relativistic programs (ADF 2005, MAGIC 2000, Rösch et al. 2002) now include density functional methods. A brief description of some of the earlier approaches to relativistic DFT is given in appendix J.

Part IV

Approximations to the Dirac Equation

15

Spin Separation and the Modified Dirac Equation

In the preceding chapters, the theory for calculations based on the Dirac equation has been laid out in some detail. The discussion of the methods included a comparison with equivalent nonrelativistic methods, from which it is apparent that four-component calculations will be considerably more expensive than the corresponding nonrelativistic calculations—perhaps two orders of magnitude more expensive. For this reason, there have been many methods developed that make approximations to the Dirac equation, and it is to these that we turn in this part of the book.[1]

There are two elements of the Dirac equation that contribute to the large amount of work: the presence of the small component of the wave function and the spin dependence of the Hamiltonian. The small component is primarily responsible for the large number of two-electron integrals which, as will be seen later, contain all the lowest-order relativistic corrections to the electron–electron interaction. The spin dependence is incorporated through the kinetic energy operator and the correction to the electronic Coulomb interaction, and also through the coupling of the spin and orbital angular momenta in the atomic 2-spinors, which form a natural basis set for the solution of the Dirac equation.

Spin separation has obvious advantages from a computational perspective. As we will show for several spin-free approaches below, a spin-free Hamiltonian is generally real, and therefore real spin–orbitals may be employed for the large and small components. The spin can then be factorized out and spin-restricted Hartree–Fock methods used to generate the one-electron functions. In the post-SCF stage, where the no-pair approximation is invoked, the transformation of the integrals from the atomic to the molecular basis produces a set of real molecular integrals that are indistinguishable from a set of nonrelativistic MO integrals, and therefore all the

---

[1] It should be pointed out at the outset that the Dirac–Coulomb(–Breit) Hamiltonian already contains approximations to the full QED Hamiltonian (if one exists!): therefore, we are considering further approximations, with this Hamiltonian as the most rigorous reference point.

nonrelativistic correlation methods may be employed without modification to obtain relativistic spin-free correlated wave functions.

In most cases, spin–free relativistic effects dominate the relativistic corrections to electronic structure. We will show later that in a perturbation expansion based on the nonrelativistic wave function, the spin-free effects for a closed-shell system enter in first order, whereas the spin-dependent effects make their first contribution in second order.[2] Thus a reasonable approach to the treatment of relativistic effects is to include the spin-free effects fully and treat the spin-dependent effects as a perturbation. We discuss the latter task in chapter 21. In this chapter, we will examine a modification to the Dirac equation that permits the spin-free and spin-dependent terms to be separated (Kutzelnigg 1984, Dyall 1994). This separation is exact, in that no approximations have been made to obtain the separation, and therefore results obtained with the modified Dirac equation are identical to those obtained with the unmodified Dirac equation. The advantage of the separation is the identification of the "genuine" spin-dependent terms and the possibility of their omission in approximate calculations. This development also provides a basis for discussion and analysis of spin-free and spin-dependent operators in other approximations.

15.1 The Modified Dirac Equation

We will work with the Dirac equation in 2-spinor form,

$$(V - E)\psi^L + c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^S = 0 \tag{15.1a}$$

$$c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L + (V - E - 2mc^2)\psi^S = 0, \tag{15.1b}$$

where we still use Hartree atomic units but have chosen to make the m-dependence explicit. The problem of separating out the spin-dependent terms in these equations is that the spin appears in the kinetic energy operator, here represented in terms of $(\boldsymbol{\sigma} \cdot \mathbf{p})$. The tool we will use is the Dirac relation (4.14), which was introduced in chapter 4 and is repeated here:

$$(\boldsymbol{\sigma} \cdot \mathbf{u})(\boldsymbol{\sigma} \cdot \mathbf{v}) = \mathbf{u} \cdot \mathbf{v} + i\boldsymbol{\sigma} \cdot \mathbf{u} \times \mathbf{v}. \tag{15.2}$$

The second term on the right-hand side vanishes if $\mathbf{u} = \mathbf{v}$, and hence the spin dependence of the product is eliminated. To make use of this relation, we need to introduce another scalar product involving $\boldsymbol{\sigma}$ into (15.1). This is achieved as follows.

The operator $(\boldsymbol{\sigma} \cdot \mathbf{p})$ appears in the relation between the small and large components,

$$\psi^S = (2mc^2 - V + E)^{-1}\, c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L. \tag{15.3}$$

---

[2] For open shell systems, the z component of the spin–orbit interaction can contribute at first order since it preserves m_s.

It is this operator that determines the symmetry characteristics of the small component relative to the large component: the premultiplying factor belongs to the totally symmetric irrep. We will therefore introduce a *pseudo-large component*, ϕ^L, defined by

$$2mc\ \psi^S \equiv (\boldsymbol{\sigma}\cdot\mathbf{p})\phi^L. \tag{15.4}$$

This is a valid definition, as it only requires that the small component be integrable, which is certainly the case since ψ^S must be square integrable in order to normalize the Dirac wave function. The pseudo-large component now has the same symmetry properties as the large component. The nonrelativistic limit of the pseudo-large component is the large component, since

$$\lim_{c\to\infty} 2mc\ \psi^S = (\boldsymbol{\sigma}\cdot\mathbf{p})\psi^L. \tag{15.5}$$

At any finite value of c, however, the pseudo-large component differs from the large component.

Substituting the definition of the pseudo-large component into (15.1), and multiplying (15.1b) by $(\boldsymbol{\sigma}\cdot\mathbf{p})/2mc$, we get the equations

$$\begin{aligned}&(V-E)\psi^L + \frac{1}{2m}(\boldsymbol{\sigma}\cdot\mathbf{p})(\boldsymbol{\sigma}\cdot\mathbf{p})\phi^L = 0\\ &\frac{1}{2m}(\boldsymbol{\sigma}\cdot\mathbf{p})(\boldsymbol{\sigma}\cdot\mathbf{p})\psi^L + \frac{1}{4m^2c^2}(\boldsymbol{\sigma}\cdot\mathbf{p})(V-E-2mc^2)(\boldsymbol{\sigma}\cdot\mathbf{p})\phi^L = 0.\end{aligned} \tag{15.6}$$

We now make use of the Dirac relation (4.14) to eliminate the spin dependence where possible. The kinetic energy term is

$$\frac{1}{2m}(\boldsymbol{\sigma}\cdot\mathbf{p})(\boldsymbol{\sigma}\cdot\mathbf{p}) = \frac{\mathbf{p}\cdot\mathbf{p}}{2m} = \hat{T}, \tag{15.7}$$

which is the nonrelativistic kinetic energy operator. The energy and rest mass terms in the second equation are constants, so the same operation may be performed on these terms. The result is the modified Dirac equation

$$(V-E)\psi^L\ +\ \hat{T}\phi^L = 0 \tag{15.8a}$$

$$\hat{T}\psi^L + \left[\frac{(\boldsymbol{\sigma}\cdot\mathbf{p})V(\boldsymbol{\sigma}\cdot\mathbf{p})}{4m^2c^2}\ -\ \hat{T} - \frac{E}{2mc^2}\hat{T}\right]\phi^L = 0. \tag{15.8b}$$

The only term remaining that has any spin dependence is the term involving the potential in the second equation, and this can also be separated out using the Dirac relation,

$$(\boldsymbol{\sigma}\cdot\mathbf{p})V(\boldsymbol{\sigma}\cdot\mathbf{p}) = \mathbf{p}V\cdot\mathbf{p} + i\boldsymbol{\sigma}\cdot\mathbf{p}V\times\mathbf{p}. \tag{15.9}$$

It is now plain that the “real” spin dependence in the Dirac equation is not in the kinetic energy but in the potential energy for the small component—a fact that is hidden in the standard form of the Dirac equation but that becomes plain in the modified form. In an atomic system the potential is spherically symmetric, and we may write the spin-dependent term as

$$i\boldsymbol{\sigma}\cdot\mathbf{p}V\times\mathbf{p} = \frac{1}{r}\frac{\partial V}{\partial r}\hbar\boldsymbol{\sigma}\cdot\mathbf{r}\times\mathbf{p} = \frac{2}{r}\frac{\partial V}{\partial r}\mathbf{s}\cdot\boldsymbol{\ell}. \tag{15.10}$$

By performing the spin separation, we have obtained a term that involves the interaction of the spin and the orbital angular momentum—a *spin–orbit* interaction.

From (15.8) we can extract a modified Dirac Hamiltonian that consists of a spin-free and a spin-dependent term,

$$\tilde{\mathcal{H}}^{\mathrm{D}} = \begin{pmatrix} V & \hat{T} \\ \hat{T} & \left[(\mathbf{p}V\cdot\mathbf{p}) - \hat{T}\right]/4m^2c^2 \end{pmatrix} + \begin{pmatrix} \mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & i\boldsymbol{\sigma}\cdot\mathbf{p}V\times\mathbf{p}/4m^2c^2 \end{pmatrix}, \tag{15.11}$$

a modified wave function,

$$\tilde{\psi} = \begin{pmatrix} \psi^L \\ \phi^L \end{pmatrix}, \tag{15.12}$$

and a modified metric that now differs from unity,

$$\tilde{G} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \hat{T}/2mc^2 \end{pmatrix}, \tag{15.13}$$

due to the definition of the pseudo-large component in (15.4). (See appendix E for details on metric changes.) If we use both the spin-free and the spin-dependent parts, this modified Hamiltonian will yield the same results as the unmodified Hamiltonian. The separation of the spin dependence is therefore exact: we have made no approximations to get to this point.

The spin-free modified Dirac Hamiltonian is obtained by simply omitting the second term on the right-hand side of (15.11):

$$\tilde{\mathcal{H}}^{\mathrm{sf}} = \begin{pmatrix} V & \hat{T} \\ \hat{T} & \left[(\mathbf{p}V\cdot\mathbf{p}) - \hat{T}\right]/4m^2c^2 \end{pmatrix}, \tag{15.14}$$

This spin-free Hamiltonian is not exact in the sense that it returns the results of the original Dirac Hamiltonian. What we have achieved is the separation of the spin-free and spin-dependent terms in the Dirac equation without further approximation: there is no truncation of either the spin-free or the spin-dependent part.

It should be noted that the choice of the pseudo-large component is not the only choice that could be made to effect the spin separation (Sadlej and Snijders 1994, Visscher and Saue 2000). We could have multiplied $(\boldsymbol{\sigma}\cdot\mathbf{p})$ by any function of the coordinates to obtain a separation. What makes this choice unique (up to a scaling factor) is that the metric is spin free. Any other choice results in a metric that has a spin-free part and a spin-dependent part.

In the next section, we will examine some properties of the solutions of the spin-free modified Dirac equation, and we then proceed to a closer inspection of the one-electron operators in the modified formalism before treating the two-electron terms.

15.2 Solutions of the Spin-Free Modified Dirac Equation

Two critical questions about the removal of the spin dependence from the Dirac equation are, “What effect does this have on the eigenvalue spectrum?” and “How is the wave function modified?” These questions may be answered by a similar process to that used for the atomic Dirac equation in chapter 7. The second question has as an auxiliary question, “How does the wave function behave in the nonrelativistic limit?”

The eigenvalue spectrum of (15.14) may be deduced from the behavior of the equations in the region far from the nuclei, where $V \to 0$ and consequently all terms involving V are negligible. The spin-free modified Dirac equation in this region may be written (in Hartree atomic units)

$$\begin{aligned} \nabla^2 \phi^L &= -2mE\psi^L \\ \nabla^2 \psi^L &= \left[1 + \frac{E}{2mc^2}\right] \nabla^2 \phi^L. \end{aligned} \tag{15.15}$$

Substituting for $\nabla^2 \phi^L$ from the first of these equations into the second gives the equation

$$\nabla^2 \psi^L = -2mE \left[1 + \frac{E}{2mc^2}\right] \psi^L, \tag{15.16}$$

whose solutions are oscillatory for $E > 0$ and $E < -2mc^2$ and exponentially decaying (bound) for $0 > E > -2mc^2$, just as they are for the original Dirac equation. This is no surprise: the spin-dependent terms arise from the potential, which is zero at large distances from the nuclei. The spin-free modified Dirac equation and the original Dirac equation are equivalent in this region, and therefore the long-range behavior of the wave function is the same. It also follows that the same reinterpretation of the negative-energy states is required for the spin-free modified Dirac equation as for the original Dirac equation.

In the region close to one of the nuclei the wave function behaves like the atomic solution for that particular nucleus. The boundary conditions are important here because there is a term involving the gradient of the potential, which for a point nucleus gives an operator $Z\mathbf{r}/r^3$ whose behavior could be problematic. To investigate the behavior near the nuclei we follow the approach adopted in chapter 7 and write the wave function as

$$\begin{pmatrix} \psi^L \\ \phi^L \end{pmatrix} = \begin{pmatrix} r^{-1} P(r) \xi \\ r^{-1} \tilde{Q}(r) \xi \end{pmatrix} \tag{15.17}$$

where ξ contains angular and spin variables.[3] Here we use $\tilde{Q}$ for the pseudo-large component: the tilde is added to distinguish it from the small component in chapter 7. We can now derive a radial spin-free equation for an atom:

$$\begin{aligned}
&\frac{d^2\tilde{Q}}{dr^2} - \frac{\ell(\ell+1)}{r^2}\tilde{Q} + 2m(E-V)P = 0 \\
&\frac{d^2P}{dr^2} - \frac{\ell(\ell+1)}{r^2}P - \left[1 - \frac{1}{2mc^2}(E-V)\right]\left[\frac{d^2\tilde{Q}}{dr^2} - \frac{\ell(\ell+1)}{r^2}\tilde{Q}\right] \\
&\quad + \frac{1}{2mc^2}\frac{dV}{dr}\left[\frac{d\tilde{Q}}{dr} - \frac{\tilde{Q}}{r}\right] = 0.
\end{aligned} \tag{15.18}$$

We expand P and $\tilde{Q}$ in power series

$$\begin{aligned}
P &= p_0 r^{\gamma} + p_1 r^{\gamma+1} + p_2 r^{\gamma+2} + \cdots , \\
\tilde{Q} &= q_0 r^{\eta} + q_1 r^{\eta+1} + q_2 r^{\eta+2} + \cdots ,
\end{aligned} \tag{15.19}$$

and making use of the expansion of the nuclear potential in (7.53) we substitute into (15.18) to obtain for the lowest terms

$$\begin{aligned}
&[\eta(\eta-1)-\ell(\ell+1)]q_0 r^{\eta-2} + 2mv_0 p_0 r^{\gamma-1} + 2m[(E+v_1)p_0 + v_0 p_1]r^{\gamma} + \cdots = 0, \\
&[\gamma(\gamma-1)-\ell(\ell+1)]p_0 r^{\gamma-2} + \frac{1}{2mc^2}v_0[\eta(\eta-1)-\ell(\ell+1)]q_0 r^{\eta-3} \\
&\quad - \left[1 - \frac{1}{2mc^2}(E+v_1)\right][\eta(\eta-1)-\ell(\ell+1)]q_0 r^{\eta-2} \\
&\quad + \frac{1}{2mc^2}\left[v_0(\eta-1)q_0 r^{\eta-3} + v_0\eta q_1 r^{\eta-2} + [v_2(\eta-1)q_0 + v_0(\eta+1)q_2]r^{\eta-1} + \cdots\right] = 0.
\end{aligned} \tag{15.20}$$

The exponents γ and η are obtained by equating the lowest terms in these expressions. Here also the point nucleus and the finite nucleus cases must be considered separately because they have different expansions of the potential.

For a point nucleus, $V = -Z/r$, that is, $v_0 = Z$, we equate the lowest terms in the first equation, which are of order $r^{\eta-2}$ from the expansion of $\tilde{Q}$ and $r^{\gamma-1}$ from the expansion of P, yielding $\eta = \gamma + 1$. The same result is obtained from the second equation. Thus $\phi^L \sim r\psi^L$ near the origin, and the potentially problematic term that behaves as $r^{-2}d\tilde{Q}/dr$ is reduced to a term that behaves as $r^{-2}P$, which is the same as for the angular momentum term. The exponent γ is non-integral as it is in the

---

[3] Note that it is the same angular and spin function for both large and pseudo-large components.

solutions of the original Dirac equation, but it has a different value for $\ell > 0$, given by the equation

$$[\gamma^2 - \ell^2][\gamma^2 - (\ell + 1)^2] - c^{-2}Z^2[\gamma^2 - \ell(\ell + 1)] = 0. \tag{15.21}$$

For a finite nucleus, where $v_0 = 0$, the lowest term in the first equation yields $\eta = \ell + 1$, and on substitution into the second equation we find that $\gamma = \ell + 1$, and thus both large and pseudo-large components have the same behavior at the origin. There is no difficulty with the terms involving the derivative of the potential because the nuclear charge distribution is no longer singular, and in fact the differentiation produces a *higher* power of r rather than a lower power because the first term in the series expansion of the potential, which is independent of r, vanishes.

In the nonrelativistic limit, (15.8) reduces to

$$\begin{aligned} (V - E)\psi^L + \hat{T}\phi^L &= 0 \\ \hat{T}\psi^L - \hat{T}\phi^L &= 0. \end{aligned} \tag{15.22}$$

Substituting the second equation into the first gives the Schrödinger equation, with ψ^L the nonrelativistic wave function. The second of these equations implies that ϕ^L also becomes the nonrelativistic wave function. For a finite nucleus this is indeed the case, but for a point nucleus it is necessary to take care with the nonrelativistic limit, as we did for the Dirac equation. Naïvely setting $c \to \infty$ in the series expansions above would still yield the result $\phi^L \sim r\psi^L$ near the nucleus, not $\phi^L \sim \psi^L$, which is what we would expect from the equation above. To illustrate, consider the ground state of a hydrogenic atom, for which the radial part of ψ^S goes as $r^{\gamma-1}e^{-Zr}$, with $\gamma^2 = \kappa^2 - Z^2/c^2$ and $\kappa = -1$. (In the following, the normalization will be neglected for simplicity.) The radial part of ϕ^L is given by

$$\phi_r^L \sim \int r^{\gamma-1}e^{-Zr}dr. \tag{15.23}$$

Integrating by parts to obtain the series expansion

$$\phi_r^L \sim r^\gamma e^{-Zr}\left(1 + \frac{Zr}{\gamma + 1} + \cdots\right) \tag{15.24}$$

gives a leading term of r^γ whose nonrelativistic limit leads to the same result as obtained from the power series expansion above, $\phi^L \sim r\psi^L$. Instead, it is necessary to obtain the asymptotic expansion from (15.23),

$$\phi_r^L \sim r^{\gamma-1}e^{-Zr}\left(1 + \frac{\gamma - 1}{Zr} + \cdots\right), \qquad r > 0. \tag{15.25}$$

Taking the limit $c \to \infty$ gives the correct nonrelativistic wave function, but the limit is valid only for $r > 0$, that is, the nonrelativistic limit is obtained only asymptotically.

It is apparently this discontinuity that causes the problems in defining the nonrelativistic limit. This point has been raised by Kutzelnigg (1989) in his discussion of perturbation theories based on the Dirac equation, where he shows that the limits $c \to \infty$ and $r \to 0$ do not necessarily commute, and for certain kinds of relativistic Hamiltonians there is no well-defined nonrelativistic limit. No such problem occurs for the finite nucleus.

15.3 Modified One-Electron Operators

One way of viewing the modification of the Dirac Hamiltonian is to consider that the wave function has been written as the product of a transformation operator $\hat{\mathcal{T}}$ and a new wave function:

$$\psi = \hat{\mathcal{T}}\tilde{\psi} \equiv \begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & (\boldsymbol{\sigma}\cdot\mathbf{p})/2mc \end{pmatrix} \begin{pmatrix} \psi^L \\ \phi^L \end{pmatrix}. \tag{15.26}$$

This transformation has to be applied to the left and the right of the Dirac Hamiltonian to obtain the modified Hamiltonian. The same applies to operators for various molecular properties, which must also be modified: the unmodified form is simply multiplied on the left and the right by the transformation operator $\hat{\mathcal{T}}$ to obtain the modified form. We consider here both an operator defined by a scalar potential and an operator defined by a vector potential.

For a scalar potential W the modified form is exactly analogous to the nuclear potential V, and separates into a spin-free and a spin-dependent term,

$$\begin{aligned}\tilde{W} &= \begin{pmatrix} W & \mathbf{0}_2 \\ \mathbf{0}_2 & (\boldsymbol{\sigma}\cdot\mathbf{p})W(\boldsymbol{\sigma}\cdot\mathbf{p})/4m^2c^2 \end{pmatrix} \\ &= \begin{pmatrix} W & \mathbf{0}_2 \\ \mathbf{0}_2 & (\mathbf{p}W\cdot\mathbf{p})/4m^2c^2 \end{pmatrix} + \begin{pmatrix} \mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & (i\boldsymbol{\sigma}\cdot\mathbf{p}W\times\mathbf{p})/4m^2c^2 \end{pmatrix}.\end{aligned} \tag{15.27}$$

For a vector potential $\mathbf{A}$ the unmodified operator is $\hat{Y} = ec\boldsymbol{\alpha}\cdot\mathbf{A}$, and the modified operator is

$$\begin{aligned}\tilde{Y} &= \frac{e}{2m}\begin{pmatrix} \mathbf{0}_2 & (\boldsymbol{\sigma}\cdot\mathbf{A})(\boldsymbol{\sigma}\cdot\mathbf{p}) \\ (\boldsymbol{\sigma}\cdot\mathbf{p})(\boldsymbol{\sigma}\cdot\mathbf{A}) & \mathbf{0}_2 \end{pmatrix} \\ &= \frac{e}{2m}\begin{pmatrix} \mathbf{0}_2 & \mathbf{A}\cdot\mathbf{p} \\ \mathbf{p}\cdot\mathbf{A} & \mathbf{0}_2 \end{pmatrix} + \frac{e}{2m}\begin{pmatrix} \mathbf{0}_2 & i\boldsymbol{\sigma}\cdot\mathbf{A}\times\mathbf{p} \\ i\boldsymbol{\sigma}\cdot\mathbf{p}\times\mathbf{A} & \mathbf{0}_2 \end{pmatrix}.\end{aligned} \tag{15.28}$$

If we are working in the Coulomb gauge, then $\nabla\cdot\mathbf{A} = 0$. We also have $\nabla\times\mathbf{A} = \mathbf{B}$, and the operator may be rewritten as

$$\tilde{Y} = \frac{e}{2m}\begin{pmatrix} \mathbf{0}_2 & \mathbf{I}_2 \\ \mathbf{I}_2 & \mathbf{0}_2 \end{pmatrix}\mathbf{A}\cdot\mathbf{p} + \frac{ie}{2m}\begin{pmatrix} \mathbf{0}_2 & \mathbf{I}_2 \\ -\mathbf{I}_2 & \mathbf{0}_2 \end{pmatrix}\boldsymbol{\Sigma}\cdot\mathbf{A}\times\mathbf{p} + \frac{e}{2m}\begin{pmatrix} \mathbf{0}_2 & \mathbf{0}_2 \\ \boldsymbol{\sigma}\cdot\mathbf{B} & \mathbf{0}_2 \end{pmatrix}. \tag{15.29}$$

This operator looks a little unsymmetrical due to the term involving the magnetic field. The lack of symmetry is apparent rather than real, because the unmodified operator had no such asymmetry. The spin-free term in $\hat{W}$ is also unsymmetrical in form, if we apply the momentum operator to the potential:

$$\mathbf{p}W \cdot \mathbf{p} = -\hbar^2 \left[(\nabla W) \cdot \nabla + W\nabla^2 \right]. \tag{15.30}$$

In practice, it is not the bare operators that we are dealing with, but the integrals, which can be written as

$$\begin{aligned} \tilde{W}_{ab} = \langle \psi_a^L | W | \psi_b^L \rangle - \frac{\hbar^2}{4m^2c^2} \langle \phi_a^L | (\nabla W) \cdot \nabla | \phi_b^L \rangle - \frac{\hbar^2}{4m^2c^2} \langle \phi_a^L | W \nabla^2 | \phi_b^L \rangle \\ + \frac{\hbar}{4m^2c^2} \langle \phi_a^L | \boldsymbol{\sigma} \cdot (\nabla W) \times \mathbf{p} | \phi_b^L \rangle \end{aligned} \tag{15.31}$$

for the scalar potential, and

$$\begin{aligned} \tilde{Y}_{ab} = -\frac{ie\hbar}{2m} \big[\langle \psi_a^L | \mathbf{A} \cdot \nabla | \phi_b^L \rangle + \langle \phi_a^L | \mathbf{A} \cdot \nabla | \psi_b^L \rangle \big] \\ + \frac{e\hbar}{2m} \big[\langle \psi_a^L | \boldsymbol{\sigma} \cdot \mathbf{A} \times \nabla | \phi_b^L \rangle - \langle \phi_a^L | \boldsymbol{\sigma} \cdot \mathbf{A} \times \nabla | \psi_b^L \rangle \\ + \langle \phi_a^L | \boldsymbol{\sigma} \cdot \mathbf{B} | \psi_b^L \rangle \big] \end{aligned} \tag{15.32}$$

for the vector potential. The basis functions here are in general 2-spinors. The reason the unsymmetrical form appears is because we are making all of the momentum operators operate to the right. It is a simple exercise in integration by parts to show that, despite the asymmetry in the operators, the integral matrices are Hermitian. For computational purposes, the most useful way of evaluating the integrals is to make the gradient (or momentum) operator always operate on the basis functions rather than the operator, for then the "normal" form of the operators may be used. With this approach, the scalar potential integrals are

$$\tilde{W}_{ab} = \langle \psi_a^L | W | \psi_b^L \rangle + \frac{\hbar^2}{4m^2c^2} \langle \nabla \phi_a^L | W \cdot | \nabla \phi_b^L \rangle - \frac{i\hbar^2}{4m^2c^2} \langle \nabla \phi_a^L | \cdot W \boldsymbol{\sigma} \times | \nabla \phi_b^L \rangle \tag{15.33}$$

and the integrals over the vector potential are

$$\begin{aligned} \tilde{Y}_{ab} = -\frac{ie\hbar}{2m} \big[\langle \psi_a^L | \mathbf{A} \cdot | \nabla \phi_b^L \rangle + \langle \nabla \phi_a^L | \cdot \mathbf{A} | \psi_b^L \rangle \big] \\ + \frac{e\hbar}{2m} \big[\langle \psi_a^L | \boldsymbol{\sigma} \cdot \mathbf{A} \times | \nabla \phi_b^L \rangle - \langle \nabla \phi_a^L | \times \mathbf{A} \cdot \boldsymbol{\sigma} | \psi_b^L \rangle \big]. \end{aligned} \tag{15.34}$$

These are manifestly symmetric integrals.

While for practical purposes these last expressions are of more use, the unsymmetrical form (15.32) gives more insight into the contributions to the integral. In (15.34) the interaction of the spin with the magnetic field is obscured. If we insert into (15.32) the nonrelativistic limit of the wave function, $\phi \to \psi$, the first and second terms become identical and the third and fourth terms cancel, to give an integral

$$\tilde{Y}_{ab}^{\mathrm{NR}} = \frac{e}{m}\langle \psi_a^{\mathrm{NR}} | \mathbf{A}\cdot\mathbf{p} | \psi_b^{\mathrm{NR}} \rangle + \frac{e\hbar}{2m}\langle \psi_a^{\mathrm{NR}} | \boldsymbol{\sigma}\cdot\mathbf{B} | \psi_b^{\mathrm{NR}} \rangle. \tag{15.35}$$

For the case of a uniform magnetic field, $\mathbf{A} = \frac{1}{2}\mathbf{B}\times\mathbf{r}$, the scalar operator can be rewritten as $\mathbf{A}\cdot\mathbf{p} = \frac{1}{2}\mathbf{B}\cdot\boldsymbol{\ell}$, and with $\hbar\boldsymbol{\sigma} = 2\mathbf{s}$ the integral becomes

$$\tilde{Y}_{ab}^{\mathrm{NR}} = \frac{e}{2m}\langle \psi_a^{\mathrm{NR}} | (\boldsymbol{\ell} + 2\mathbf{s})\cdot\mathbf{B} | \psi_b^{\mathrm{NR}} \rangle. \tag{15.36}$$

Here we see that the spin-dependent term of the modified vector potential is in fact describing the interaction of the spin with the magnetic field, and the scalar term is describing the interaction of the orbital angular momentum with the magnetic field. These terms are the spin and orbital Zeeman terms respectively. In (15.32) there are also relativistic corrections that arise from the difference between ϕ and ψ.

One further step is usually taken in the evaluation of the integrals (Vahtras et al. 1992). The basis functions are functions of the relative coordinates of the atomic center, $\mathbf{r} - \mathbf{R}_A$, and the differentiation can therefore be transferred from the electronic to the nuclear coordinate, because

$$\frac{\partial\phi_a}{\partial x} = -\frac{\partial\phi_a}{\partial X_A}. \tag{15.37}$$

The differentiation can be taken outside the integral, converting it into a derivative integral with respect to nuclear displacements. The scalar potential integrals are then

$$\begin{aligned}\tilde{W}_{ab} = \langle \psi_a^L | W | \psi_b^L \rangle &+ \frac{\hbar^2}{4m^2c^2}\nabla_A\cdot\nabla_B\langle \phi_a^L | W | \phi_b^L \rangle \\ &- \frac{i\hbar^2}{4m^2c^2}\nabla_A\times\nabla_B\cdot\langle \phi_a^L | W\boldsymbol{\sigma} | \phi_b^L \rangle\end{aligned} \tag{15.38}$$

and the integrals over the vector potential are

$$\begin{aligned}\tilde{Y}_{ab} = &-\frac{ie\hbar}{2m}\big[\nabla_B\cdot\langle \psi_a^L | \mathbf{A} | \phi_b^L \rangle + \nabla_A\cdot\langle \phi_a^L | \mathbf{A} | \psi_b^L \rangle\big] \\ &+ \frac{e\hbar}{2m}\big[\nabla_B\cdot\langle \psi_a^L | \boldsymbol{\sigma}\times\mathbf{A} | \phi_b^L \rangle + \nabla_A\cdot\langle \phi_a^L | \boldsymbol{\sigma}\times\mathbf{A} | \psi_b^L \rangle\big].\end{aligned} \tag{15.39}$$

It is also possible to use the transformed magnetic perturbation operator of section 13.7. The use of this operator is discussed in section 19.5.

15.4 Modified Two-Electron Operators

So far we have only considered the modification of the one-electron terms. The modified two-electron terms are derived in a straightforward manner from a consideration of the two-electron integrals, in much the same way as the one-electron property operators were derived. To proceed we define the projection operators onto the large (positive sign) and small (negative sign) components,

$$\hat{\mathcal{P}}^{\pm} = \frac{1}{2}\big[\mathbf{I}_4 \pm \beta\big] \equiv \hat{\mathcal{P}}^{+} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{0}_2 \end{pmatrix}, \quad \hat{\mathcal{P}}^{-} = \begin{pmatrix} \mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{I}_2 \end{pmatrix}. \tag{15.40}$$

These operators are idempotent, as all projection operators should be: $\hat{\mathcal{P}}^{\pm}\hat{\mathcal{P}}^{\pm} = \hat{\mathcal{P}}^{\pm}$. In addition, they commute with the four-component spin operator $\boldsymbol{\Sigma}$—a property that is important in the developments below. Using these projection operators, the transformation operator for electron i can be written as

$$\hat{\mathcal{T}}_i = \hat{\mathcal{P}}_i^{+} + \frac{1}{2mc}(\boldsymbol{\Sigma}\cdot\mathbf{p})\hat{\mathcal{P}}_i^{-}. \tag{15.41}$$

With this definition, we can write the modified two-electron operators as

$$\tilde{g}_{ij} = \hat{\mathcal{T}}_i\, \hat{\mathcal{T}}_j\, g_{ij}\, \hat{\mathcal{T}}_j\, \hat{\mathcal{T}}_i. \tag{15.42}$$

The three cases of interest for g_{ij} are the Coulomb interaction, the Gaunt interaction, and the Breit interaction.

The unmodified Coulomb interaction has no spin-dependent terms, and we can straightforwardly apply the transformation operator and use the idempotency of the projection operators to obtain the modified Coulomb operator,

$$\begin{aligned} \tilde{g}^{C} &= \hat{\mathcal{P}}_i^{+}\hat{\mathcal{P}}_j^{+}\ (\hat{g}_{ij}^{00}) + \frac{1}{4m^2c^2}\hat{\mathcal{P}}_i^{+}\hat{\mathcal{P}}_j^{-}\ (\hat{g}_{ij}^{02}) + \frac{1}{4m^2c^2}\hat{\mathcal{P}}_i^{-}\hat{\mathcal{P}}_j^{+}\ (\hat{g}_{ij}^{20}) \\ &\quad + \frac{1}{16m^4c^4}\hat{\mathcal{P}}_i^{-}\hat{\mathcal{P}}_j^{-}\ (\hat{g}_{ij}^{22}). \end{aligned} \tag{15.43}$$

Here, we have defined four operators with superscripts that describe the dependence on $\boldsymbol{\Sigma}$ and $\mathbf{p}$ of the densities involving electrons i and j. These four operators are

$$\hat{g}_{ij}^{00} = \frac{1}{r_{ij}} \tag{15.44a}$$

$$\hat{g}_{ij}^{02} = (\boldsymbol{\Sigma}_j\cdot\mathbf{p}_j)\ \frac{1}{r_{ij}}\ (\boldsymbol{\Sigma}_j\cdot\mathbf{p}_j) \tag{15.44b}$$

$$\hat{g}_{ij}^{20} = (\boldsymbol{\Sigma}_i\cdot\mathbf{p}_i)\ \frac{1}{r_{ij}}\ (\boldsymbol{\Sigma}_i\cdot\mathbf{p}_i) \tag{15.44c}$$

$$\hat{g}_{ij}^{22} = (\boldsymbol{\Sigma}_i\cdot\mathbf{p}_i)(\boldsymbol{\Sigma}_j\cdot\mathbf{p}_j)\ \frac{1}{r_{ij}}\ (\boldsymbol{\Sigma}_j\cdot\mathbf{p}_j)(\boldsymbol{\Sigma}_i\cdot\mathbf{p}_i). \tag{15.44d}$$

The first of the four operators is the normal Coulomb operator. The second and third are related by index interchange, and have their analogies in the one-electron terms:

$$\hat{g}_{ij}^{20} = \mathbf{p}_i \frac{1}{r_{ij}} \cdot \mathbf{p}_i + \hbar \boldsymbol{\Sigma}_i \cdot \left(\nabla_i \frac{1}{r_{ij}} \right) \times \mathbf{p}_i . \tag{15.45}$$

The gradient term can be written

$$\nabla_i \frac{1}{r_{ij}} = -\frac{\mathbf{r}_{ij}}{r_{ij}^3} \tag{15.46}$$

to show that the spin-dependent term is a spin–orbit interaction,

$$\hbar \boldsymbol{\Sigma}_i \cdot \left(\nabla_i \frac{1}{r_{ij}} \right) \times \mathbf{p}_i = -\frac{1}{r_{ij}^3} \hbar \boldsymbol{\Sigma}_i \cdot \mathbf{r}_{ij} \times \mathbf{p}_i = -\frac{2}{r_{ij}^3} \mathbf{s}_i \cdot \boldsymbol{\ell}_{ij} . \tag{15.47}$$

We can compare this expression with the one-electron spin–orbit term from the nuclear potential, which for a point nucleus is

$$\hbar \boldsymbol{\sigma}_i \cdot \left(\nabla_i \frac{-Z}{r_i} \right) \times \mathbf{p}_i = +\frac{2Z}{r_i^3} \mathbf{s}_i \cdot \boldsymbol{\ell}_i . \tag{15.48}$$

The two-electron term is the interaction of the spin of electron i with its orbital angular momentum around electron j, sometimes called the *spin-own-orbit* interaction. It is of the opposite sign to the nuclear term and therefore acts as a screening term. The fourth operator in (15.43) contains four terms, which are rather difficult to represent cleanly. Instead, it is simpler to consider the integrals of these operators in terms of derivatives with respect to nuclear coordinates, as we did in the previous section. Then, we have the four integrals

$$\begin{aligned}
g_{abcd}^{00} &= (ab|cd) \\
g_{abcd}^{20} &= (\nabla_A \cdot \nabla_B)(\tilde{a}\tilde{b}|cd) + i(\nabla_A \times \nabla_B \cdot)(\tilde{a}\boldsymbol{\sigma}\tilde{b}|cd) \\
g_{abcd}^{02} &= (\nabla_C \cdot \nabla_D)(ab|\tilde{c}\tilde{d}) + i(\nabla_C \times \nabla_D \cdot)(ab|\tilde{c}\boldsymbol{\sigma}\tilde{d}) \\
g_{abcd}^{22} &= (\nabla_A \cdot \nabla_B)(\nabla_C \cdot \nabla_D)(\tilde{a}\tilde{b}|\tilde{c}\tilde{d}) + i(\nabla_A \cdot \nabla_B)(\nabla_C \times \nabla_D \cdot)(\tilde{a}\tilde{b}|\tilde{c}\boldsymbol{\sigma}\tilde{d}) \\
&\quad + i(\nabla_C \cdot \nabla_D)(\nabla_A \times \nabla_B \cdot)(\tilde{a}\boldsymbol{\sigma}\tilde{b}|\tilde{c}\tilde{d}) \\
&\quad - (\nabla_A \times \nabla_B \cdot)(\nabla_C \times \nabla_D \cdot)(\tilde{a}\boldsymbol{\sigma}\tilde{b}|\tilde{c}\boldsymbol{\sigma}\tilde{d})
\end{aligned} \tag{15.49}$$

where the scalar products in the last integral are taken to apply to the appropriate spin density in the integral, and the tildes distinguish the pseudo-large component from the large component. The appearance of $\boldsymbol{\sigma}$ rather than $\boldsymbol{\Sigma}$ in these integrals is due to the fact that the projection operators have been applied. We have also omitted $\hbar$ in these expressions.

The modified operators for the Gaunt and Breit interactions can be derived in an analogous manner. The derivation is somewhat more involved than for the Coulomb interaction due to the presence of the α matrices. Here we derive the Gaunt terms, but the gauge term for the Breit interaction is considerably more complicated (and the derivation may be found in appendix F).

We start the derivation of the modified Gaunt interaction by examining the effect of sandwiching $\boldsymbol{\alpha}$ between two projection operators:

$$\begin{aligned}
\hat{\mathcal{P}}^+\boldsymbol{\alpha}\hat{\mathcal{P}}^+ &= 0 \\
\hat{\mathcal{P}}^+\boldsymbol{\alpha}\hat{\mathcal{P}}^- &= \begin{pmatrix} \mathbf{0}_2 & \boldsymbol{\sigma} \\ \mathbf{0}_2 & \mathbf{0}_2 \end{pmatrix} = \boldsymbol{\Sigma}\begin{pmatrix} \mathbf{0}_2 & \mathbf{I}_2 \\ \mathbf{0}_2 & \mathbf{0}_2 \end{pmatrix} \\
\hat{\mathcal{P}}^-\boldsymbol{\alpha}\hat{\mathcal{P}}^+ &= \begin{pmatrix} \mathbf{0}_2 & \mathbf{0}_2 \\ \boldsymbol{\sigma} & \mathbf{0}_2 \end{pmatrix} = \boldsymbol{\Sigma}\begin{pmatrix} \mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{I}_2 & \mathbf{0}_2 \end{pmatrix} \\
\hat{\mathcal{P}}^-\boldsymbol{\alpha}\hat{\mathcal{P}}^- &= 0.
\end{aligned} \tag{15.50}$$

This result makes sense because the α matrices connect large and small components. The two new matrices introduced in the second and third equations are similar to the projection operators, but they also perform an interchange of the large and small components. We define the "raising" and the "lowering" operators in the component space,

$$\begin{aligned}
\hat{\mathcal{Q}}^+ &= \begin{pmatrix} \mathbf{0}_2 & \mathbf{I}_2 \\ \mathbf{0}_2 & \mathbf{0}_2 \end{pmatrix}, \qquad \hat{\mathcal{Q}}^+\begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} = \begin{pmatrix} \psi^S \\ \mathbf{0}_2 \end{pmatrix}; \\
\hat{\mathcal{Q}}^- &= \begin{pmatrix} \mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{I}_2 & \mathbf{0}_2 \end{pmatrix}, \qquad \hat{\mathcal{Q}}^-\begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} = \begin{pmatrix} \mathbf{0}_2 \\ \psi^L \end{pmatrix}.
\end{aligned} \tag{15.51}$$

We can therefore write

$$\hat{\mathcal{P}}^+\boldsymbol{\alpha}\hat{\mathcal{P}}^- = \hat{\mathcal{Q}}^+\boldsymbol{\Sigma}, \qquad \hat{\mathcal{P}}^-\boldsymbol{\alpha}\hat{\mathcal{P}}^+ = \hat{\mathcal{Q}}^-\boldsymbol{\Sigma}. \tag{15.52}$$

Note that $\hat{\mathcal{Q}}^\pm$ commutes with $\boldsymbol{\Sigma}$.

Inserting $g_{ij} = -\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j / r_{ij}$ into (15.42) and making use of the above definitions, we get the modified Gaunt interaction,

$$\tilde{g}^{\mathrm{G}} = -\frac{1}{4m^2c^2}\left[\hat{\mathcal{Q}}_i^+\hat{\mathcal{Q}}_j^+\,(g_{ij}^{++}) + \hat{\mathcal{Q}}_i^+\hat{\mathcal{Q}}_j^-\,(g_{ij}^{+-}) + \hat{\mathcal{Q}}_i^-\hat{\mathcal{Q}}_j^+\,(g_{ij}^{-+}) + \hat{\mathcal{Q}}_i^-\hat{\mathcal{Q}}_j^-\,(g_{ij}^{--})\right], \tag{15.53}$$

where the superscripts on the g operators have been chosen to match the product of projection operators. Each of these operators contains four $\boldsymbol{\Sigma}$ matrices, two for electron i and two for electron j, in three scalar products. The reduction to spin-free and spin-dependent parts using the Dirac relation (4.14) follows a similar pattern for each of these operators:

$$
\begin{aligned}
g_{ij}^{++} &= \frac{1}{r_{ij}}(\boldsymbol{\Sigma}_i \cdot \boldsymbol{\Sigma}_j)(\boldsymbol{\Sigma}_i \cdot \mathbf{p}_i)(\boldsymbol{\Sigma}_j \cdot \mathbf{p}_j) \\
&= \frac{1}{r_{ij}}\big[\mathbf{p}_i \cdot \mathbf{p}_j - i(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)\cdot(\mathbf{p}_i \times \mathbf{p}_j) - (\boldsymbol{\Sigma}_i \times \mathbf{p}_i)\cdot(\boldsymbol{\Sigma}_j \times \mathbf{p}_j)\big],
\end{aligned} \tag{15.54a}
$$

$$
\begin{aligned}
g_{ij}^{+-} &= (\boldsymbol{\Sigma}_j \cdot \mathbf{p}_j)\frac{1}{r_{ij}}(\boldsymbol{\Sigma}_i \cdot \boldsymbol{\Sigma}_j)(\boldsymbol{\Sigma}_i \cdot \mathbf{p}_i) \\
&= \big[\mathbf{p}_j \frac{1}{r_{ij}} \cdot \mathbf{p}_i + i(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)\cdot(\mathbf{p}_j \frac{1}{r_{ij}} \times \mathbf{p}_i) + (\boldsymbol{\Sigma}_j \times \mathbf{p}_j \frac{1}{r_{ij}})\cdot(\boldsymbol{\Sigma}_i \times \mathbf{p}_i)\big],
\end{aligned} \tag{15.54b}
$$

$$
\begin{aligned}
g_{ij}^{-+} &= (\boldsymbol{\Sigma}_i \cdot \mathbf{p}_i)(\boldsymbol{\Sigma}_i \cdot \boldsymbol{\Sigma}_j)\frac{1}{r_{ij}}(\boldsymbol{\Sigma}_j \cdot \mathbf{p}_j) \\
&= \big[\mathbf{p}_i \frac{1}{r_{ij}} \cdot \mathbf{p}_j + i(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)\cdot(\mathbf{p}_i \frac{1}{r_{ij}} \times \mathbf{p}_j) + (\boldsymbol{\Sigma}_i \times \mathbf{p}_i \frac{1}{r_{ij}})\cdot(\boldsymbol{\Sigma}_j \times \mathbf{p}_j)\big],
\end{aligned} \tag{15.54c}
$$

$$
\begin{aligned}
g_{ij}^{--} &= (\boldsymbol{\Sigma}_i \cdot \mathbf{p}_i)(\boldsymbol{\Sigma}_j \cdot \mathbf{p}_j)(\boldsymbol{\Sigma}_i \cdot \boldsymbol{\Sigma}_j)\frac{1}{r_{ij}} \\
&= \big[\mathbf{p}_i \cdot \mathbf{p}_j + i(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)\cdot(\mathbf{p}_i \times \mathbf{p}_j) - (\boldsymbol{\Sigma}_i \times \mathbf{p}_i)\cdot(\boldsymbol{\Sigma}_j \times \mathbf{p}_j)\big]\frac{1}{r_{ij}}.
\end{aligned} \tag{15.54d}
$$

Each operator thus reduces to a spin-free part involving a scalar product of momentum operators, a spin–orbit term involving the spin of both electrons, and a spin–spin interaction term. Because of the projection operators, no further simplifications can be made in general in these expressions apart from rearrangement of the scalar and vector products into other forms.

We may collect the integrals over the operators according to the three kinds of term. The spin-free integrals are

$$
g_{abcd}^{\mathrm{G,sf}} = -\nabla_A \cdot \nabla_C(\tilde{a}b|\tilde{c}d) + \nabla_A \cdot \nabla_D(\tilde{a}b|c\tilde{d}) + \nabla_B \cdot \nabla_C(a\tilde{b}|\tilde{c}d) - \nabla_B \cdot \nabla_D(a\tilde{b}|c\tilde{d}), \tag{15.55}
$$

the spin–orbit integrals are

$$
\begin{aligned}
g_{abcd}^{\mathrm{G,so}} = &-i\nabla_A \times \nabla_C[(\tilde{a}\boldsymbol{\sigma}b|\tilde{c}d) - (\tilde{a}b|\tilde{c}\boldsymbol{\sigma}d)] + i\nabla_A \times \nabla_D[(\tilde{a}\boldsymbol{\sigma}b|c\tilde{d}) + (\tilde{a}b|c\boldsymbol{\sigma}\tilde{d})] \\
&+ i\nabla_B \times \nabla_C[(a\boldsymbol{\sigma}\tilde{b}|\tilde{c}d) + (a\tilde{b}|\tilde{c}\boldsymbol{\sigma}d)] - i\nabla_B \times \nabla_D[(a\boldsymbol{\sigma}\tilde{b}|c\tilde{d}) - (a\tilde{b}|c\boldsymbol{\sigma}\tilde{d})],
\end{aligned} \tag{15.56}
$$

and the spin–spin integrals are

$$
\begin{aligned}
g_{abcd}^{\mathrm{G,ss}} = {}&(\nabla_A\times)\cdot(\nabla_C\times)(\tilde{a}\boldsymbol{\sigma}b|\tilde{c}\boldsymbol{\sigma}d) - (\nabla_A\times)\cdot(\nabla_D\times)(\tilde{a}\boldsymbol{\sigma}b|c\boldsymbol{\sigma}\tilde{d}) \\
&- (\nabla_B\times)\cdot(\nabla_C\times)[(a\boldsymbol{\sigma}\tilde{b}|\tilde{c}\boldsymbol{\sigma}d)] + (\nabla_B\times)\cdot(\nabla_D\times)[(a\boldsymbol{\sigma}\tilde{b}|c\boldsymbol{\sigma}\tilde{d})],
\end{aligned} \tag{15.57}
$$

where the vector products apply to the appropriate spin density and the scalar product is taken between the resultant two vector products. These integrals are in reality integrals over 4-spinors.

In terms of the large and pseudo-large components, the integrals can be classified into three basic integral types,

$$\begin{aligned}&\nabla_A \cdot \nabla_C (\tilde{a}b|\tilde{c}d),\\&\nabla_A \times \nabla_C (\tilde{a}\boldsymbol{\sigma}b|\tilde{c}d),\\&(\nabla_A \times)\cdot(\nabla_C \times)(\tilde{a}\boldsymbol{\sigma}b|\tilde{c}\boldsymbol{\sigma}d).\end{aligned} \tag{15.58}$$

The scalar quadruple product in the spin–spin operator can be rewritten as

$$(\boldsymbol{\sigma}_i \times \mathbf{p}_i)\cdot(\boldsymbol{\sigma}_j \times \mathbf{p}_j) = (\boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j)(\mathbf{p}_i \cdot \mathbf{p}_j) - (\boldsymbol{\sigma}_i \cdot \mathbf{p}_j)(\boldsymbol{\sigma}_j \cdot \mathbf{p}_i) \tag{15.59}$$

to remove the vector product. The integrals from the first term on the right-hand side are then identical in form to the spin-free integrals, except that they are taken over spin densities instead of charge densities. The second term is not related to any of the previous terms. Ultimately, the integrals involve combinations of all nine Cartesian second-derivative integrals.

The gauge term, which is the difference between the Gaunt interaction and the Breit interaction, produces a spin-free operator that can be interpreted as an orbit–orbit interaction. Thus, both the Gaunt interaction and the gauge term of the Breit interaction give rise to spin-free contributions to the modified Dirac operator. We will use the developments of this section in chapter 17 to derive the Breit–Pauli Hamiltonian.

15.5 Practical Implications of Spin Separation

One important advantage of the modified Dirac equation is that, since the large component and the pseudo-large component have the same symmetry, we can use the same primitive basis set for both. However, if we want to use a contracted basis set, the contraction coefficients for these functions will differ. We will therefore distinguish the basis sets for the components when we expand them, which we do now:

$$\psi^L = \sum_k a_k^L \chi_k^L; \quad \phi^L = \sum_k a_k^P \chi_k^P \tag{15.60}$$

We use the superscript P for the pseudo-large component, which represents both the initial letter, and also the fact that it is related to the small component through the momentum operator $\mathbf{p}$. The one-electron modified Dirac equation in this basis set is then

$$\begin{pmatrix} \mathrm{V}^{LL} & \mathrm{T}^{LP} \\ \mathrm{T}^{PL} & \mathrm{V}^{PP}/4m^2c^2 - \mathrm{T}^{PP} \end{pmatrix} \begin{pmatrix} \mathrm{a}^L \\ \mathrm{a}^P \end{pmatrix} = \begin{pmatrix} \mathrm{S}^{LL} & 0 \\ 0 & \mathrm{T}^{PP}/2mc^2 \end{pmatrix} \begin{pmatrix} \mathrm{a}^L \\ \mathrm{a}^P \end{pmatrix} E \tag{15.61}$$

There are three kinetic energy matrices, all of which can be calculated with the same primitive integrals, contracted appropriately. V^{PP} is the matrix of $(\boldsymbol{\sigma} \cdot \mathbf{p})V(\boldsymbol{\sigma} \cdot \mathbf{p})$, or of $\mathbf{p}V \cdot \mathbf{p}$ if we are choosing the spin-free version. The integrals over the potential can be evaluated as derivatives of regular integrals, just as was done for the electric properties in section 15.3. Likewise, the two-electron integrals can be evaluated as derivative integrals, and their form follows from (15.49).

At the beginning of this chapter it was asserted that a considerable amount of work could be saved by performing a spin separation in the Dirac equation, using the spin-free part variationally, then treating the spin-dependent terms as a perturbation. We are now in a position to assess the reduction of work that would actually be obtained by performing a spin-free calculation.

We also want to compare the amount of work in a spin-free modified Dirac calculation with that in a nonrelativistic calculation. One of the reasons given for the development of relativistic effective core potentials (RECPs) was the expense of four-component Dirac–Fock calculations. RECPs are often used in spin-free form, but the comparison is made with the spin-dependent unmodified Dirac approach. The comparison of the cost of a spin-free all-electron calculation with a nonrelativistic calculation will give a more realistic indication of the relative cost of the incorporation of spin-free relativistic effects.

We will evaluate the numbers of integrals required for a calculation with the unmodified Dirac Hamiltonian, and compare them with the number of integrals required for a calculation with the spin-free modified Dirac Hamiltonian, and with the number required for a nonrelativistic calculation. The spin-free Hamiltonian is formed by summing all the spin-free terms defined above, but we will consider the Coulomb term and the Gaunt and Breit terms separately. For the purpose of this evaluation, we make the following assumptions and definitions:

- The system has no spatial symmetry.
- There is a 1:1 ratio of large- and small-component functions, that is, the restricted kinetic balance relation applies between the primitive large- and small-component 2-spinor basis sets.
- The number of orbitals in the basis set is n, that is, there are n Kramers pairs in the unmodified approach, and n real orbitals in the spin-free modified approach.

In the analysis of the numbers of integrals, there are several distinctions that need to be made. First, the integrals over the unmodified Dirac Hamiltonian are given in a 2-spinor basis, whereas the spin-free modified Dirac integrals are given in a scalar basis. Second, the large and small components will have different basis sets in the unmodified Dirac method, regardless of whether the basis is contracted or not, whereas in the spin-free modified Dirac method, the *primitive* basis sets are the same for the large and pseudo-large components, so that only in a contracted basis are the large and pseudo-large basis sets different.

We start by comparing the one-electron matrices, of which there are three types to consider:

- Overlap matrices: $\mathbf{S}^{LL}$ and $\mathbf{S}^{SS}$ in the unmodified Dirac method, $\mathbf{S}^{LL}$ and $\mathbf{T}^{PP}$ in the modified Dirac method, $\mathbf{S}$ in the nonrelativistic method.

- Potential energy matrices: $\mathbf{V}^{LL}$ and $\mathbf{V}^{SS}$ in the unmodified Dirac method, $\mathbf{V}^{LL}$ and $\mathbf{V}^{PP}$ in the modified Dirac method, $\mathbf{V}$ in the nonrelativistic method.
- Kinetic energy matrices: $\boldsymbol{\Pi}^{LS}$ in the unmodified Dirac method, $\mathbf{T}^{LP}$ in the modified Dirac method, $\mathbf{T}$ in the nonrelativistic method.

The overlap and potential energy matrices are matrices of dimension n^2 for each component, both in the modified and unmodified Dirac method, and both are Hermitian. The unmodified Dirac matrices are quaternion matrices, and there are $n(2n-1)$ unique quantities (real numbers) in each matrix. The modified spin-free Dirac matrices are real, with $n(n+1)/2$ unique quantities. The reduction in the number of unique quantities is a factor of approximately 4.

The kinetic energy matrix is a full matrix of dimension n^2 in both cases. As for the other two, the unmodified Dirac kinetic energy matrix is a quaternion matrix, with $4n^2$ unique quantities, whereas the spin-free modified Dirac kinetic energy matrix is a real matrix with n^2 unique quantities, and the resultant reduction is a factor of 4. However, if an uncontracted basis is used, the spin-free modified Dirac kinetic energy matrix is symmetric, and is the same as the pseudo-large-component overlap matrix.

Overall, there are approximately $12n^2$ unique quantities in the unmodified Dirac one-electron matrices. There are $3n^2$ unique quantities in the modified spin-free Dirac one-electron matrices if the basis is contracted, and $2n^2$ quantities if the basis is uncontracted. In a nonrelativistic calculation, where there is no small component, the number of unique quantities is approximately $3/2n^2$. Thus, the spin-free modified Dirac method has twice as many one-electron integrals in a contracted calculation and only 33% more integrals in an uncontracted calculation than a nonrelativistic calculation.

Similar considerations apply to the two-electron Coulomb integrals. The densities scale in exactly the same way as the one-electron integrals, and therefore the reduction from the unmodified Dirac integral set to the spin-free modified Dirac integral set is a factor of 16. Similarly, the increase from the nonrelativistic integral set is a factor of 4. In other words, where the full Dirac–Coulomb integral set is 64 times the nonrelativistic set, and therefore the SCF calculation is 64 times the cost, a spin-free modified Dirac-Coulomb SCF calculation is only four times the cost of a nonrelativistic calculation. It is no surprise that the same scaling factors apply to the integrals from the Gaunt and the Breit interactions, which can be combined into a single integral set since the two have the same permutational symmetry. A spin-free modified Dirac–Breit SCF calculation is then eight times the cost of a nonrelativistic SCF calculation. These factors apply regardless of whether the basis functions are contracted or uncontracted, so that as far as the integral count is concerned, contraction is a minor issue.

The costs evaluated so far relate only to the number of integrals. The expense of evaluating the integrals is also a consideration. Because the integrals over the pseudo-large component involve gradients, the cost is essentially the same as that of the small-component integrals. Thus, the integral cost does not change on going from the original Dirac equation to the modified Dirac equation.

Another important issue in performing the spin separation is whether this would give any advantage in a calculation that included all the spin-dependent terms. We might then choose to work in the modified representation of the Dirac equation rather than the unmodified representation.

The spin-free one-electron integrals comprise a quarter of the number of integrals from the unmodified Dirac operators. The remaining integrals are the spin–orbit integrals, which would be generated and stored as the spatial integrals $\nabla_A \times \nabla_B \langle \tilde{a} \,|\, V \,|\, \tilde{b} \rangle$. For each index pair ab there are three integrals, and these are antisymmetric to interchange of the indices, making a total of $3n(n-1)/2$ integrals. When added to the spin-free pseudo-large-component potential energy integrals, this gives $n(2n-1)$ unique integrals, which is exactly the same as for the unmodified small-component potential energy integrals. The number of large-component potential energy integrals, kinetic energy and overlap integrals is the same as before. If an uncontracted basis set is used, the count of modified integrals is approximately $7/2n^2$.

Before drawing any conclusions about the one-electron matrices, we make the following observation. If a contracted basis set is to be used in a spin-dependent calculation, the spin-orbit components of the atomic functions with $\ell > 0$ should be contracted separately, rather than using a single function determined, for example, in a spin-free atomic calculation. Otherwise, there is a risk of a large contraction error in the calculations, particularly for heavy elements where the spin–orbit components can have significantly different radial functions, even in the valence shell. Under these circumstances the number of modified integrals for functions with $\ell > 0$ is quadrupled in the modified approach, and the advantage of the representation in terms of scalar functions is to a large extent lost. We could of course transform the modified integrals to a 2-spinor basis, but then there would be the same number as in the unmodified approach, and any advantages gained by the modified approach would again be lost.

Turning to the two-electron terms, we find a similar situation. We consider only the Coulomb interaction, where the spin-free integrals constituted 1/16th of the full unmodified set, or four times the nonrelativistic set. The spin–orbit integrals are antisymmetric to the interchange of the two indices involved in the vector product, and symmetric to the interchange of the other two indices. We therefore have $n(n-1)/2 * n(n+1)/2$ integrals for each spin–orbit component, which amounts to six times the nonrelativistic set. The integrals from four pseudo-large-component functions amount to six times the nonrelativistic set for those involving a single spin operator, and nine times for those involving two spin operators. The overall factor for the spin-dependent integrals is 21. Summing all the contributions gives a factor of 25 relative to the nonrelativistic set. This is a factor of about 2.5 reduction from the full relativistic set, which has 64 times as many independent quantities as in the nonrelativistic integral set.

In a contracted basis set, the advantage is again lost when the spin–orbit components are contracted separately. The doubling of the number of basis functions for $\ell > 0$ means that there will be sixteen times as many integrals for these basis functions. In a typical basis set there would therefore be no advantage from the point of view of integral storage in keeping the integrals in modified Dirac form over scalar basis functions.

16

Unitary Transformations of the Dirac Hamiltonian

The separation of the spin-dependent terms in the Dirac Hamiltonian enables us to make an approximation in which the spin-free terms are included in the orbital optimization and the spin-dependent terms may be treated later as a perturbation. In this process, the parameter space required to treat the large and small components has not changed. Even with the extraction of $(\boldsymbol{\sigma} \cdot \mathbf{p})$ from the small component, we still have to calculate integrals involving $\mathbf{p}\phi^L$, which essentially regenerates the original small component space and so the integral work has not really changed. What has been achieved is the ability to use the machinery of spin algebra from nonrelativistic theory, but we are left with a large and a small component.

The obvious next step is to separate the large and small components, or the positive- and negative-energy states. The small component can be eliminated from the Dirac equation by algebraic manipulation, but this leaves the energy in the denominator. It would be preferable to obtain an energy-independent Hamiltonian that acted only on positive-energy states and that could therefore be represented as two-component spinors. If, following this separation, it were possible to separate out the spin-free and spin-dependent terms, we would have a spin-free Hamiltonian that would operate on a one-component wave function, and we would then be able to use all the machinery of nonrelativistic quantum chemistry but with modified one- and two-electron integrals.

The matrix form of the Dirac Hamiltonian suggests that we should seek a unitary transformation that will make it diagonal with respect to the large- and small-component spinor spaces. Such a transformation is called a Foldy–Wouthuysen transformation (Foldy and Wouthuysen 1950). Although in their original paper only the free-particle transformation was derived, together with an iterative decoupling procedure that will be described later in this chapter, the term Foldy–Wouthuysen transformation has come to mean any unitary transformation that decouples the large and small components, either exactly or approximately, and we will use it in this sense.

16.1 The Foldy–Wouthuysen Transformation

We start from the time-independent Dirac equation in two-component form,

$$\hat{\mathcal{H}}^{\mathrm{D}} = \begin{pmatrix} V & c(\boldsymbol{\sigma}\cdot\mathbf{p}) \\ c(\boldsymbol{\sigma}\cdot\mathbf{p}) & V-2mc^2 \end{pmatrix}. \tag{16.1}$$

Here we are not considering the time dependence, which was done in the original paper by Foldy and Wouthuysen (1950). It turns out that the results of the transformation are essentially the same. We are looking for a transformation $\hat{\Omega}$ that transforms a 4-spinor to a 2-spinor, such that

$$\hat{\Omega}\begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} = \begin{pmatrix} \psi^{\mathrm{FW}} \\ 0 \end{pmatrix}. \tag{16.2}$$

This means that the lower row elements of $\hat{\Omega}$ must fulfill the relation

$$\hat{\Omega}_{21}\psi^L + \hat{\Omega}_{22}\psi^S = 0. \tag{16.3}$$

To achieve this, we introduce the operator $\hat{\mathcal{X}}$, which connects ψ^L and ψ^S

$$\psi^S = \hat{\mathcal{X}}\psi^L. \tag{16.4}$$

This relation indicates that a suitable form of $\hat{\Omega}$ might be

$$\hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{11} & \hat{\Omega}_{12} \\ -\hat{\mathcal{X}} & 1 \end{pmatrix}. \tag{16.5}$$

For an orthogonal transformation we must have

$$\hat{\Omega} = \begin{pmatrix} 1 & \hat{\mathcal{X}}^\dagger \\ -\hat{\mathcal{X}} & 1 \end{pmatrix}. \tag{16.6}$$

To make this transformation unitary, we determine the norm from

$$\hat{\Omega}^\dagger\hat{\Omega} = \begin{pmatrix} 1 & -\hat{\mathcal{X}}^\dagger \\ \hat{\mathcal{X}} & 1 \end{pmatrix}\begin{pmatrix} 1 & \hat{\mathcal{X}}^\dagger \\ -\hat{\mathcal{X}} & 1 \end{pmatrix} = \begin{pmatrix} 1+\hat{\mathcal{X}}^\dagger\hat{\mathcal{X}} & 0 \\ 0 & 1+\hat{\mathcal{X}}\hat{\mathcal{X}}^\dagger \end{pmatrix}. \tag{16.7}$$

We thus find that a formal decoupling of the large and small components may be effected by the unitary Foldy–Wouthuysen transformation (see Kutzelnigg 1989),

$$\hat{\mathcal{U}} = \begin{pmatrix} \dfrac{1}{\sqrt{1+\hat{\mathcal{X}}^\dagger\hat{\mathcal{X}}}} & \dfrac{1}{\sqrt{1+\hat{\mathcal{X}}^\dagger\hat{\mathcal{X}}}}\hat{\mathcal{X}}^\dagger \\ -\dfrac{1}{\sqrt{1+\hat{\mathcal{X}}\hat{\mathcal{X}}^\dagger}}\hat{\mathcal{X}} & \dfrac{1}{\sqrt{1+\hat{\mathcal{X}}\hat{\mathcal{X}}^\dagger}} \end{pmatrix}. \tag{16.8}$$

Applying this transformation to the Dirac Hamiltonian (16.1) we get the formal decoupling

$$\hat{\mathcal{U}}\hat{\mathcal{H}}^{\mathrm{D}}\hat{\mathcal{U}}^{-1} = \begin{pmatrix} \hat{\mathcal{H}}_{+} & \mathbf{0} \\ \mathbf{0} & \hat{\mathcal{H}}_{-} \end{pmatrix}, \tag{16.9}$$

where the Hamiltonian for the positive-energy states is

$$\hat{\mathcal{H}}^{+} = \frac{1}{\sqrt{1+\hat{\mathcal{X}}^{\dagger}\hat{\mathcal{X}}}}\left[V + c(\boldsymbol{\sigma}\cdot\mathbf{p})\hat{\mathcal{X}} + c\hat{\mathcal{X}}^{\dagger}(\boldsymbol{\sigma}\cdot\mathbf{p}) + \hat{\mathcal{X}}^{\dagger}(V - 2mc^{2})\hat{\mathcal{X}}\right]\frac{1}{\sqrt{1+\hat{\mathcal{X}}^{\dagger}\hat{\mathcal{X}}}}. \tag{16.10}$$

This operator operates on 2-spinors. The wave function for the positive-energy states is

$$\psi^{\mathrm{FW}} = \frac{1}{\sqrt{1+\hat{\mathcal{X}}^{\dagger}\hat{\mathcal{X}}}}\left(\psi^{L} + \hat{\mathcal{X}}^{\dagger}\psi^{S}\right). \tag{16.11}$$

Substituting for ψ^{S} using (16.4) we get an expression for the 2-component transformed wave function in terms of the large component,

$$\psi^{\mathrm{FW}} = \sqrt{1+\hat{\mathcal{X}}^{\dagger}\hat{\mathcal{X}}}\;\psi^{L}. \tag{16.12}$$

We may derive an expression for $\hat{\mathcal{X}}$ from the off-diagonal element of the transformed Hamiltonian, which must be zero. With some rearrangement, this expression is

$$2mc^{2}\hat{\mathcal{X}} = c(\boldsymbol{\sigma}\cdot\mathbf{p}) + \left[V, \hat{\mathcal{X}}\right] - \hat{\mathcal{X}}c(\boldsymbol{\sigma}\cdot\mathbf{p})\hat{\mathcal{X}}. \tag{16.13}$$

The solution of this equation is not known in analytic form for a general potential because it is clear that $\hat{\mathcal{X}}$ must at least involve the operator $c(\boldsymbol{\sigma}\cdot\mathbf{p})$ and therefore does not commute with V. It should also be noted that $\hat{\mathcal{X}}$ does not commute with $\hat{\mathcal{X}}^{\dagger}$. This may be verified by considering the relationship between the large and small component for any given eigenstate of the Dirac Hamiltonian with energy E, for which we may write

$$\hat{\mathcal{X}} = (2mc^{2} + E - V)^{-1}c(\boldsymbol{\sigma}\cdot\mathbf{p}); \qquad \hat{\mathcal{X}}^{\dagger} = c(\boldsymbol{\sigma}\cdot\mathbf{p})(2mc^{2} + E - V)^{-1}. \tag{16.14}$$

The order of the operators $\hat{\mathcal{X}}$ and $\hat{\mathcal{X}}^{\dagger}$ must therefore be maintained in any manipulations.

Even though we have an equation for $\hat{\mathcal{X}}$, we are no better off for it, because of the complicated nature of the equation. If we make the approximation $\hat{\mathcal{X}} = (\boldsymbol{\sigma}\cdot\mathbf{p})/2mc$, which is essentially the nonrelativistic limit of $\hat{\mathcal{X}}$, and commence an iterative process using (16.13), the first cycle will generate derivatives of the potential from the second term and third derivative operators from the third term. The next cycle will generate third derivatives of the potential and fifth derivative operators. This process therefore generates highly singular operators, and it is difficult to see how it could be otherwise.

There appears to be no alternative: we may have to be content with some sort of approximation.

There is one situation in which the solution for $\hat{\mathcal{X}}$ is known, and that is for a constant potential, such as for a free particle. The second term in (16.13) disappears and it may be assumed that $\hat{\mathcal{X}}$ is a function of $(\boldsymbol{\sigma} \cdot \mathbf{p})$ only and therefore commutes with $(\boldsymbol{\sigma} \cdot \mathbf{p})$. The quadratic equation for $\hat{\mathcal{X}}$ is now

$$c(\boldsymbol{\sigma} \cdot \mathbf{p})\big[\hat{\mathcal{X}}^2 - 1\big] + 2mc^2\hat{\mathcal{X}} = 0. \tag{16.15}$$

Substituting $\hat{\mathcal{X}} = c(\boldsymbol{\sigma} \cdot \mathbf{p})/\hat{Y}$ we get a simple equation

$$\hat{Y}^2 - 2mc^2\hat{Y} - c^2p^2 = 0 \tag{16.16}$$

that has the solution $\hat{Y} = mc^2 + \sqrt{m^2c^4 + c^2p^2}$, and thus

$$\hat{\mathcal{X}} = \frac{c(\boldsymbol{\sigma} \cdot \mathbf{p})}{mc^2 + \sqrt{m^2c^4 + c^2p^2}} = \frac{c(\boldsymbol{\sigma} \cdot \mathbf{p})}{mc^2 + \hat{E}_p} \tag{16.17}$$

where the square root is recognized as the classical energy of a relativistic free particle and defines the free-particle Dirac energy operator:

$$\hat{E}_p = \sqrt{m^2c^4 + c^2p^2}. \tag{16.18}$$

We recall that it was the desire to find an expansion of this square root operator that led to the development of the Dirac equation (see chapter 4). We see also that the assumption that $\hat{\mathcal{X}}$ commutes with $(\boldsymbol{\sigma} \cdot \mathbf{p})$ was justified. The free-particle Foldy–Wouthuysen transformation can now be written

$$\hat{\mathcal{U}} = \frac{1}{\sqrt{2\hat{E}_p(\hat{E}_p + mc^2)}}\begin{pmatrix} \hat{E}_p + mc^2 & c(\boldsymbol{\sigma} \cdot \mathbf{p}) \\ -c(\boldsymbol{\sigma} \cdot \mathbf{p}) & \hat{E}_p + mc^2 \end{pmatrix} = \sqrt{\frac{\hat{E}_p + mc^2}{2\hat{E}_p}}\left[\mathbf{I}_4 + \frac{c\beta\boldsymbol{\alpha} \cdot \mathbf{p}}{\hat{E}_p + mc^2}\right]. \tag{16.19}$$

The transformed free-particle Hamiltonian is

$$\hat{\mathcal{H}}^{\text{FPFW}} = \beta\hat{E}_p = \begin{pmatrix} \hat{E}_p & \mathbf{0} \\ \mathbf{0} & -\hat{E}_p \end{pmatrix}. \tag{16.20}$$

The operators in the transformation now involve square roots of the square root operator, and it is much easier to use these in momentum space than in position space, where their interpretation is problematic.

It is clear from this section that in eliminating the small component by a Foldy–Wouthuysen transformation we have arrived at operators that are much more difficult to handle than those of either the original Dirac equation or the modified

Dirac equation. This method of eliminating the small component is not a procedure that leads to a simplification. It does, however, have some motivation, both physical and practical. First, it projects out the negative-energy states, and leaves a Hamiltonian that may have a variational lower bound, avoiding the potential problem of variational collapse. Second, it removes from explicit consideration the small component, and with the use of the Dirac relation (4.14) it yields a one-component operator that can be used in nonrelativistic computer programs.

One final note is necessary before proceeding to some approximations. The definition of $\hat{\mathcal{X}}$, and consequently of the transformed Hamiltonian, depends on the potential. We have considered here only a scalar potential, but it is straightforward to include a vector potential as well. This dependence on the potential means that any other operator to which we apply the transformation—such as a property operator—will not in general be diagonal. Calculation of properties in the transformed representation using the two-component wave function defined in (16.12) will therefore be approximate, even if the energy and wave function are exact. This issue will be considered further in the sections on properties.

16.2 Approximate Foldy–Wouthuysen Transformations

If it is not possible to obtain the Foldy–Wouthuysen transformation in closed form, the best that we might be able to do is to find a transformation that decouples the large and small components, or the positive- and negative-energy states, to some order in a suitable coupling parameter. This means that our approximations will all be based on perturbation theory in a general sense. Before examining some specific transformations, we need to develop the formal theory of transformations so that we can have some information about the order to which the components are decoupled.

To do this, we first note that the operators in the Dirac equation can be divided into even and odd operators, $\hat{\mathcal{E}}$ and $\hat{\mathcal{O}}$. Even operators connect components of the same type—large with large, small with small—and so only have elements on the diagonal blocks of the 4×4 matrices. They are sometimes called *diagonal operators*. An operator that is composed entirely of even operators therefore has no coupling between large and small components. Odd operators connect the large and small components, and are sometimes called *nondiagonal* or *off-diagonal operators*. It is the odd operators that we wish to remove from the Dirac Hamiltonian in order to bring it into a diagonal, and hence decoupled, form. In the Dirac Hamiltonian, the potential and βmc^2 are even operators and $\boldsymbol{\alpha} \cdot \mathbf{p}$ is an odd operator. The Hamiltonian can be written symbolically as

$$\hat{\mathcal{H}} = \beta mc^2 + \hat{\mathcal{O}} + \hat{\mathcal{E}} \tag{16.21}$$

where the rest mass term has been retained explicitly for later convenience. The commutation properties of the even and odd operators are

$$[\hat{\mathcal{E}}, \beta] = 0; \qquad [\hat{\mathcal{O}}, \beta]_+ = 0, \tag{16.22}$$

that is, even operators commute with β, odd operators anticommute.

We now make a unitary transformation of the wave function using an exponential ansatz,

$$\psi_1 = e^{i\hat{\mathcal{S}}}\psi \tag{16.23}$$

where $\hat{\mathcal{S}}$ is Hermitian. The transformed Hamiltonian is

$$\hat{\mathcal{H}}_1 = e^{i\hat{\mathcal{S}}}\hat{\mathcal{H}}e^{-i\hat{\mathcal{S}}} \tag{16.24}$$

and we may use the exponential expansion to write the new Hamiltonian in terms of the commutator series,

$$\hat{\mathcal{H}}_1 = \hat{\mathcal{H}} + i[\hat{\mathcal{S}}, \hat{\mathcal{H}}] - \frac{1}{2}[\hat{\mathcal{S}}, [\hat{\mathcal{S}}, \hat{\mathcal{H}}]] - \frac{i}{3!}[\hat{\mathcal{S}}, [\hat{\mathcal{S}}, [\hat{\mathcal{S}}, \hat{\mathcal{H}}]]] + \ldots, \tag{16.25}$$

involving powers of the operator $\hat{\mathcal{S}}$, which can therefore be used to define an expansion parameter. The operator $\hat{\mathcal{S}}$ is chosen to make some commutator or commutators cancel with the terms from the Dirac Hamiltonian, thereby reducing the contribution of odd operators to the transformed Hamiltonian by some power of the expansion parameter.

The obvious expansion parameter is $1/c$, which should generate the nonrelativistic Hamiltonian and operators for relativistic corrections to various orders. In terms of this parameter, the three operators in the Dirac Hamiltonian, βmc^2, $\hat{\mathcal{O}} = c\boldsymbol{\alpha}\cdot\mathbf{p}$, and $\hat{\mathcal{E}} = V$, are of orders -2, -1, and 0, respectively. It is for this reason that the rest mass term was not included in the even operator. If we choose

$$\hat{\mathcal{S}} = -i\beta\hat{\mathcal{O}}/2mc^2 = -i\beta\boldsymbol{\alpha}\cdot\mathbf{p}/2mc \tag{16.26}$$

we have an operator that has $1/c$ as the expansion parameter. Considering the first commutator, we find

$$\begin{aligned} i\big[\hat{\mathcal{S}}, \beta mc^2\big] &= -\hat{\mathcal{O}}, \\ i\big[\hat{\mathcal{S}}, \hat{\mathcal{O}}\big] &= \beta\hat{\mathcal{O}}^2/mc^2, \\ i\big[\hat{\mathcal{S}}, \hat{\mathcal{E}}\big] &= \beta\big[\hat{\mathcal{O}}, \hat{\mathcal{E}}\big]/2mc^2, \end{aligned} \tag{16.27}$$

where we have left the even operator—the potential—unspecified for the present. These three terms are of order -1, 0, and $+1$ in the expansion parameter, respectively. The first of them cancels the odd operator in the Dirac Hamiltonian. The second term is an even term, but the third term is an odd term. Before we draw any conclusions, we must consider the next commutator, which will reduce the order of the Dirac operators by 2, contributing terms of orders 0, 1, and 2 in the expansion parameter:

$$\begin{aligned} \big[\hat{\mathcal{S}}, \big[\hat{\mathcal{S}}, \beta mc^2\big]\big] &= \beta\hat{\mathcal{O}}^2/mc^2, \\ \big[\hat{\mathcal{S}}, \big[\hat{\mathcal{S}}, \hat{\mathcal{O}}\big]\big] &= \hat{\mathcal{O}}^3/m^2c^4, \\ \big[\hat{\mathcal{S}}, \big[\hat{\mathcal{S}}, \hat{\mathcal{E}}\big]\big] &= \big[\hat{\mathcal{O}}, \big[\hat{\mathcal{O}}, \hat{\mathcal{E}}\big]\big]/4m^2c^4. \end{aligned} \tag{16.28}$$

In fact, we also need to consider the next commutator to obtain the full term of order 1, which is the highest-order odd operator in the transformed Hamiltonian—and if we wish to go further and obtain all the terms of order 2 to determine the lowest-order relativistic correction to the Hamiltonian we need to consider the next two commutators. Iterating the previous results, we find that

$$\begin{aligned} -i[\hat{\mathcal{S}},[\hat{\mathcal{S}},[\hat{\mathcal{S}},\beta mc^2]]] &= \hat{\mathcal{O}}^3/m^2c^4, \\ -i[\hat{\mathcal{S}},[\hat{\mathcal{S}},[\hat{\mathcal{S}},\hat{\mathcal{O}}]]] &= -\beta\hat{\mathcal{O}}^4/m^3c^6, \\ -i[\hat{\mathcal{S}},[\hat{\mathcal{S}},[\hat{\mathcal{S}},\hat{\mathcal{E}}]]] &= -\beta[\hat{\mathcal{O}},[\hat{\mathcal{O}},[\hat{\mathcal{O}},\hat{\mathcal{E}}]]]/8m^3c^6, \end{aligned} \tag{16.29}$$

$$\begin{aligned} [\hat{\mathcal{S}},[\hat{\mathcal{S}},[\hat{\mathcal{S}},[\hat{\mathcal{S}},\beta mc^2]]]] &= \beta\hat{\mathcal{O}}^4/m^3c^6, \\ [\hat{\mathcal{S}},[\hat{\mathcal{S}},[\hat{\mathcal{S}},[\hat{\mathcal{S}},\hat{\mathcal{O}}]]]] &= \hat{\mathcal{O}}^5/m^4c^8, \\ [\hat{\mathcal{S}},[\hat{\mathcal{S}},[\hat{\mathcal{S}},[\hat{\mathcal{S}},\hat{\mathcal{E}}]]]] &= [\hat{\mathcal{O}},[\hat{\mathcal{O}},[\hat{\mathcal{O}},[\hat{\mathcal{O}},\hat{\mathcal{E}}]]]]/16m^4c^8. \end{aligned} \tag{16.30}$$

With these expressions, the transformed Hamiltonian correct to order 2 in $1/c$ is

$$\begin{aligned} \hat{\mathcal{H}}_1 = \beta mc^2 &+ \frac{1}{2mc^2}\beta\hat{\mathcal{O}}^2 - \frac{1}{3m^2c^4}\hat{\mathcal{O}}^3 - \frac{1}{8m^3c^6}\beta\hat{\mathcal{O}}^4 + \dots \\ &+ \hat{\mathcal{E}} + \frac{1}{2mc^2}\beta[\hat{\mathcal{O}},\hat{\mathcal{E}}] - \frac{1}{8m^2c^4}[\hat{\mathcal{O}},[\hat{\mathcal{O}},\hat{\mathcal{E}}]] + \dots . \end{aligned} \tag{16.31}$$

The odd operator of order -1 in $1/c$ has now been removed, and the lowest-order even operators are of order -2 and 0. Substituting for $\hat{\mathcal{O}}$ and $\hat{\mathcal{E}}$ the term of order 0 is just

$$\hat{\mathcal{H}}^{(0)} = \beta\frac{\mathbf{p}^2}{2m} + V = \begin{pmatrix} T+V & 0 \\ 0 & -T+V \end{pmatrix}. \tag{16.32}$$

The portion that operates on the electron solutions is just the nonrelativistic Hamiltonian. This should be no surprise, because we are applying a decoupling procedure in powers of $1/c$.

The highest-order odd operator is now of order 1. To remove it, we repeat the process of defining odd and even operators and selecting $\hat{\mathcal{S}}$ to eliminate the lowest term in the new odd operator. The new odd and even operators are

$$\begin{aligned} \hat{\mathcal{O}}_1 &= -\frac{1}{3m^2c^4}\hat{\mathcal{O}}^3 + \frac{1}{2mc^2}\beta[\hat{\mathcal{O}},\hat{\mathcal{E}}] + \dots, \\ \hat{\mathcal{E}}_1 &= \hat{\mathcal{E}} - \frac{1}{8m^2c^4}[\hat{\mathcal{O}},[\hat{\mathcal{O}},\hat{\mathcal{E}}]] + \frac{1}{2mc^2}\beta\hat{\mathcal{O}}^2 - \frac{1}{8m^3c^6}\beta\hat{\mathcal{O}}^4 + \dots, \end{aligned} \tag{16.33}$$

and we take $\hat{\mathcal{S}} \equiv \hat{\mathcal{S}}_1$ from the lowest terms of $\hat{\mathcal{O}}_1$,

$$\hat{\mathcal{S}}_1 = \frac{i}{6m^3c^6}\hat{\mathcal{O}}^3 - \frac{i}{4m^2c^4}[\hat{\mathcal{O}},\hat{\mathcal{E}}]. \tag{16.34}$$

$\hat{\mathcal{S}}_1$ is now of order 3. The commutator relations derived above hold for the new operators as well. The commutator with βmc^2 produces a term that cancels the odd operator of order 1. The next commutator, with $\hat{\mathcal{O}}_1$, generates a term of order 4, which we neglect since we are concerned with an expansion to order 2. The commutator with $\hat{\mathcal{E}}_1$ produces terms of which the highest order is 3, which again we neglect. The Hamiltonian correct to order 2 is therefore given by the lowest terms of $\hat{\mathcal{E}}_1$,

$$\begin{aligned}\hat{\mathcal{H}}^{(2)} &= \hat{\mathcal{E}} - \frac{1}{8m^2c^4}[\hat{\mathcal{O}}, [\hat{\mathcal{O}}, \hat{\mathcal{E}}]] + \frac{1}{2mc^2}\beta\hat{\mathcal{O}}^2 - \frac{1}{8m^3c^6}\beta\hat{\mathcal{O}}^4 \\ &= \beta\left(mc^2 + \frac{\mathbf{p}^2}{2m} - \frac{\mathbf{p}^4}{8m^3c^2}\right) + V - \frac{1}{8m^2c^2}\big[\boldsymbol{\alpha}\cdot\mathbf{p}, \big[\boldsymbol{\alpha}\cdot\mathbf{p}, V\big]\big].\end{aligned} \tag{16.35}$$

Here we see a relativistic correction to the kinetic energy, which is commonly called the *mass–velocity* operator, and a relativistic correction to the potential. The latter term reduces using the equivalent of (4.14) to

$$\big[\boldsymbol{\alpha}\cdot\mathbf{p}, \big[\boldsymbol{\alpha}\cdot\mathbf{p}, V\big]\big] = -\hbar^2(\nabla^2 V) + 2\hbar\boldsymbol{\Sigma}\cdot\mathbf{p}\times(\nabla V) \tag{16.36}$$

where as before

$$\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\sigma} & \mathbf{0}_2 \\ \mathbf{0}_2 & \boldsymbol{\sigma}\end{pmatrix}. \tag{16.37}$$

The scalar operator is called the *Darwin* operator and the spin-dependent operator the *spin–orbit* operator. We will meet these again in the chapter on perturbation theory (chapter 17). The Hamiltonian correct to order 2 in $1/c$ can then be written as

$$\hat{\mathcal{H}}^{(2)} = \beta\left(mc^2 + \frac{\mathbf{p}^2}{2m} - \frac{\mathbf{p}^4}{8m^3c^2}\right) + V + \frac{\hbar^2(\nabla^2 V)}{8m^2c^2} + \frac{\hbar\boldsymbol{\Sigma}\cdot(\nabla V)\times\mathbf{p}}{4m^2c^2}. \tag{16.38}$$

This is the four-component form of the *Pauli Hamiltonian*.

It is clear that this approach of successive transformations in the expansion parameter $1/c$ yields operators that involve higher and higher powers of $\mathbf{p}$: the commutator series is in fact a series in $\mathbf{p}/mc$. Powers of the momentum operator higher than 2 are not bounded, and when operating on the potential V produce highly singular operators. This makes it problematic to use any but the lowest-order terms in a calculation, and since the mass–velocity operator is unbounded from below, this form of the relativistic corrections must be used in perturbation theory only.

16.3 The Douglas–Kroll Transformation

If the expansion in powers of $\mathbf{p}/mc$ produces operators that can only be used in perturbation theory, is there an expansion that will produce operators that can be used variationally? Such a transformation would have to provide an expansion in powers of the potential energy rather than the momentum. The Douglas–Kroll transformation

(Douglas and Kroll 1974) is one such transformation, and its application has been developed by Hess and coworkers (Hess 1985, 1986, Hess et al. 1986).

The first step is the application of the free-particle Foldy–Wouthuysen transformation to the Dirac operator. As in the previous section, we use the Dirac Hamiltonian without subtracting the rest mass,

$$\hat{\mathcal{H}}^{\mathrm{D}} = c\boldsymbol{\alpha}\cdot\mathbf{p} + \beta mc^2 + V. \tag{16.39}$$

Defining the operators

$$\begin{aligned}\hat{\mathcal{A}} &= \left(\frac{\hat{E}_p + mc^2}{2\hat{E}_p}\right)^{1/2} = \mathcal{O}(c^0),\\ \hat{\mathcal{D}} &= \frac{1}{\hat{E}_p + mc^2} = \mathcal{O}(c^{-2})\\ \hat{\mathcal{R}} &= \hat{\mathcal{D}}c\boldsymbol{\alpha}\cdot\mathbf{p} = \mathcal{O}(c^{-1})\end{aligned} \tag{16.40}$$

we may write the free-particle Foldy–Wouthuysen transformation of (16.19) as

$$\hat{\mathcal{U}}^{\mathrm{FPFW}} = \hat{\mathcal{A}}(\mathbf{I} + \beta\hat{\mathcal{R}}). \tag{16.41}$$

Applying this to the Dirac Hamiltonian, and comparing to the free-particle transformed Hamiltonian of (16.20) above, we see that the only additional terms are those involving the potential

$$\begin{aligned}\hat{\mathcal{H}}_1 &= \hat{\mathcal{U}}^{\mathrm{FPFW}}\hat{\mathcal{H}}^{\mathrm{D}}(\hat{\mathcal{U}}^{\mathrm{FPFW}})^{-1} = \beta\hat{E}_p + \hat{\mathcal{A}}\left(V + \beta[\hat{\mathcal{R}}, V] + \hat{\mathcal{R}}V\hat{\mathcal{R}}\right)\hat{\mathcal{A}}\\ &= \begin{pmatrix}\hat{E}_p & 0\\ 0 & -\hat{E}_p\end{pmatrix} + \hat{\mathcal{A}}\begin{pmatrix}V + \hat{\mathcal{R}}_2V\hat{\mathcal{R}}_2 & [\hat{\mathcal{R}}_2, V]\\ [V, \hat{\mathcal{R}}_2] & V + \hat{\mathcal{R}}_2V\hat{\mathcal{R}}_2\end{pmatrix}\hat{\mathcal{A}}\end{aligned} \tag{16.42}$$

where we have used the fact that β anticommutes with $\hat{\mathcal{R}}$, and the operator $\hat{\mathcal{R}}_2$ is the 2-spinor version of $\hat{\mathcal{R}}$

$$\hat{\mathcal{R}}_2 = \hat{\mathcal{D}}c\boldsymbol{\sigma}\cdot\mathbf{p}. \tag{16.43}$$

We can split the $\hat{\mathcal{R}}_2V\hat{\mathcal{R}}_2$ term into scalar and spin–orbit operators, which makes it possible to define a spin-free one-electron Hamiltonian. The two-component Hamiltonian for the positive-energy states can then be written

$$\begin{aligned}\hat{\mathcal{H}}_1^{+} &= \hat{E}_p + \hat{\mathcal{A}}\Big(V + c^2\hat{\mathcal{D}}\mathbf{p}V\cdot\mathbf{p}\hat{\mathcal{D}} + c^2\hat{\mathcal{D}}\hbar\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{p}\hat{\mathcal{D}}\Big)\hat{\mathcal{A}}\\ &= \hat{E}_p + \hat{\mathcal{A}}V\hat{\mathcal{A}} + \frac{1}{m^2c^2}\hat{\mathcal{Q}}\Big(\mathbf{p}V\cdot\mathbf{p} + \hbar\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{p}\Big)\hat{\mathcal{Q}}\end{aligned} \tag{16.44}$$

where we have introduced the dimensionless kinematic factor $\hat{\mathcal{Q}}$,

$$\hat{\mathcal{Q}} = mc^2 \hat{\mathcal{D}} \hat{\mathcal{A}} = \frac{mc^2}{[2\hat{E}_p(\hat{E}_p + mc^2)]^{1/2}} = \mathcal{O}(c^0). \tag{16.45}$$

The Hamiltonian $\hat{\mathcal{H}}_1^+$ is bounded from below, and can therefore be used variationally. The spin-free Hamiltonian is

$$\hat{\mathcal{H}}_1^{\mathrm{sf}} = \hat{E}_p + \hat{\mathcal{A}} V \hat{\mathcal{A}} + \frac{1}{m^2c^2} \hat{\mathcal{Q}} \mathbf{p} V \cdot \mathbf{p} \hat{\mathcal{Q}}. \tag{16.46}$$

Provided we can find some way of evaluating the effect of the operators $\hat{\mathcal{A}}$ and $\hat{\mathcal{Q}}$ on the wave function, the matrix elements of this Hamiltonian would not be too difficult to calculate. This question is addressed later on in this chapter. But is this a useful place to stop the transformation process, and how accurate are the results we get from this Hamiltonian?

For purposes of further analysis we partition $\hat{\mathcal{H}}_1$ into the kinetic term $\beta \hat{E}_p$, which only involves the momentum, and an even and an odd operator, both of which depend on the potential,

$$\begin{aligned} \hat{\mathcal{E}}_1 &= \hat{\mathcal{A}}(V + \hat{\mathcal{R}} V \hat{\mathcal{R}}) \hat{\mathcal{A}}, \\ \hat{\mathcal{O}}_1 &= \beta \hat{\mathcal{A}} [V, \hat{\mathcal{R}}] \hat{\mathcal{A}}. \end{aligned} \tag{16.47}$$

Comparing the transformed Hamiltonian with the original Dirac Hamiltonian, we note that the new even operator including the kinetic energy has terms of $\mathcal{O}(c^2)$, $\mathcal{O}(c^0)$, and $\mathcal{O}(c^{-2})$, whereas the original even operator has terms of $\mathcal{O}(c^2)$ and $\mathcal{O}(c^0)$. The new odd operator is $\mathcal{O}(c^{-1})$, whereas the original is $\mathcal{O}(c^1)$. The coupling between the large and small components has therefore been reduced by $\mathcal{O}(c^2)$. It is still larger than the relativistic correction to the potential, which is essentially the term of $\mathcal{O}(c^{-2})$, $\hat{\mathcal{A}} \hat{\mathcal{R}} V \hat{\mathcal{R}} \hat{\mathcal{A}}$. However, the influence of the off-diagonal term is only $\mathcal{O}(c^{-4})$, as may be seen by the following perturbation analysis.

If we use this partitioning of the transformed Hamiltonian to define a perturbation theory, with a zeroth-order Hamiltonian $\hat{\mathcal{H}}_0 = \beta \hat{E}_p + \hat{\mathcal{E}}_1$ and a perturbation $\hat{\mathcal{H}}' = \hat{\mathcal{O}}_1$, the energy expressions for the first few terms are

$$\begin{aligned} E^{(0)} &= \langle 0 | \beta \hat{E}_p + \hat{\mathcal{E}}_1 | 0 \rangle, \\ E^{(1)} &= \langle 0 | \hat{\mathcal{O}}_1 | 0 \rangle, \\ E^{(2)} &= \sum_{k_-} \frac{\langle 0 | \hat{\mathcal{O}}_1 | k_- \rangle \langle k_- | \hat{\mathcal{O}}_1 | 0 \rangle}{E_0 - E_{k_-}} + \sum_{k_+} \frac{\langle 0 | \hat{\mathcal{O}}_1 | k_+ \rangle \langle k_+ | \hat{\mathcal{O}}_1 | 0 \rangle}{E_0 - E_{k_+}}, \end{aligned} \tag{16.48}$$

where we have split the sum in the second-order energy into sums over the positive- and negative-energy states. The zeroth-order wave function has a nonzero large component but a zero small component. Since the perturbation connects large and small components, the first-order energy must be zero and the first-order wave function must involve a sum over the negative-energy states only, which have zero large components and nonzero small components. The sum over positive-energy states in the second-order energy is therefore zero. The denominator in the negative-energy sum is $\mathcal{O}(c^{-2})$ and must be positive. The perturbation operator is $\mathcal{O}(c^{-1})$, making the second-order energy $\mathcal{O}(c^{-4})$ and positive. The zeroth-order Hamiltonian therefore contains the lowest-order relativistic corrections, those of $\mathcal{O}(c^{-2})$, and produces energies that are too low.

It is instructive to compare the 2-component Hamiltonian $\hat{\mathcal{H}}_1^+$ with the Pauli Hamiltonian, which was derived in the previous section. The term $\beta\hat{E}_p$ contains the relativistic free-particle energy, which is well-behaved for all values of the momentum. $\hat{E}_p$, which is the kinetic energy operator for the positive-energy states, is a positive definite operator. In the Pauli Hamiltonian we see that this operator is expanded in powers of $\mathbf{p}/mc$, which does not converge if $\mathbf{p}/mc > 1$—a situation that will occur in any potential if the electron is sufficiently close to the nucleus. As mentioned above, the mass–velocity term is not bounded from below and so cannot be used variationally.

To compare the operator $\hat{\mathcal{E}}_1$ with the terms from the Pauli Hamiltonian, we need to expand the various kinematic factors in powers of $1/c$ up to second order, as follows:

$$\hat{\mathcal{A}} = 1 - \frac{p^2}{8m^2c^2} + \mathcal{O}(c^{-4}); \qquad \hat{\mathcal{Q}} = \frac{1}{2} + \mathcal{O}(c^{-2}). \tag{16.49}$$

Substituting these expressions into $\hat{\mathcal{H}}_1^+$, we get

$$\hat{\mathcal{E}}_1^+ = V - \frac{1}{8m^2c^2}\big[p^2, V\big]_+ + \frac{1}{4m^2c^2}\big(\mathbf{p}V \cdot \mathbf{p} + i\boldsymbol{\sigma} \cdot (\mathbf{p}V) \times \mathbf{p}\big) + \mathcal{O}(c^{-4}). \tag{16.50}$$

Expanding and collecting terms, we get the Darwin and spin–orbit operators from the Pauli Hamiltonian. The transformed Hamiltonian therefore contains all the terms of the Pauli Hamiltonian—as might be expected since it is correct to $\mathcal{O}(c^{-2})$.

The variational problems with the Pauli Hamiltonian stem not only from the mass–velocity operator, which is negative, but also from the spin–orbit operator, which behaves as $1/r^3$ and can be negative. In the Hamiltonian $\hat{\mathcal{H}}_1^+$, as $\mathbf{p}$ becomes larger the kinematic factors $\hat{\mathcal{A}}$ and $\hat{\mathcal{D}}$ become progressively smaller, with the result that the potential energy terms are reduced as the momentum increases. In the large momentum limit, when the electron is close to the nucleus, $\hat{E}_p \to cp$, $\hat{\mathcal{A}} \to 1/2$, and $\hat{\mathcal{R}} \to \boldsymbol{\alpha}\cdot\mathbf{p}/p$. The relativistic correction to the potential is no more singular than the potential itself in this limit and therefore will support bound states. In the small momentum limit, when the electron is far from the nucleus, the potential goes as $1/r^3$ and is therefore a short-range potential. It can be seen that the kinematic factors provide a cutoff to the potential that is absent in the Pauli approximation and that permits variational calculations with the free-particle Foldy–Wouthuysen transformed Hamiltonian.

The Hamiltonian $\hat{\mathcal{H}}_1^+$ has been used by Almlöf et al. (1985), and also by Buenker et al. (1984). As we showed in the perturbation analysis above, the energies in this approximation are too low at $\mathcal{O}(c^{-4})$. To remedy this deficiency, we may apply a transformation derived by Douglas and Kroll (1974) to further reduce the magnitude of the odd operator. Instead of using an exponential ansatz for a unitary operator, the unitary operator is written as

$$\hat{\mathcal{U}}_1 = (1 + \hat{\mathcal{W}}_1^2)^{1/2} + \hat{\mathcal{W}}_1 \tag{16.51}$$

where $\hat{\mathcal{W}}_1$ is anti-Hermitian. A power series in $\hat{\mathcal{W}}_1$ is used to expand $\hat{\mathcal{U}}_1$,

$$\hat{\mathcal{U}}_1 = 1 + \hat{\mathcal{W}}_1 + \frac{1}{2}\hat{\mathcal{W}}_1^2 + \frac{1}{8}\hat{\mathcal{W}}_1^4 + \dots . \tag{16.52}$$

Note that the expansion is the same as the exponential ansatz up to the second-order term, so that for a low-order approximation it makes no difference which we use.

We now apply this transformation to $\hat{\mathcal{H}}_1$ from (16.42),

$$\begin{aligned}\hat{\mathcal{H}}_2 &= \hat{\mathcal{U}}_1 \hat{\mathcal{H}}_1 \hat{\mathcal{U}}_1^{-1} = \hat{\mathcal{H}}_1 + [\hat{\mathcal{W}}_1, \hat{\mathcal{H}}_1] + \frac{1}{2}[\hat{\mathcal{W}}_1, [\hat{\mathcal{W}}_1, \hat{\mathcal{H}}_1]] + \dots \\ &= \beta \hat{E}_p + \hat{\mathcal{E}}_1 + \hat{\mathcal{O}}_1 + [\hat{\mathcal{W}}_1, \beta \hat{E}_p] + [\hat{\mathcal{W}}_1, \hat{\mathcal{E}}_1] + [\hat{\mathcal{W}}_1, \hat{\mathcal{O}}_1] \\ &\quad + \frac{1}{2}[\hat{\mathcal{W}}_1, [\hat{\mathcal{W}}_1, \beta \hat{E}_p]] + \dots ,\end{aligned} \tag{16.53}$$

and choose the highest-order term to cancel $\hat{\mathcal{O}}_1$,

$$\hat{\mathcal{O}}_1 + [\hat{\mathcal{W}}_1, \beta \hat{E}_p] = 0. \tag{16.54}$$

Since $\beta \hat{E}_p$ is an even operator, $\hat{\mathcal{W}}_1$ must be odd and therefore must anticommute with β. Multiplying through by β and extracting β from the commutator, we may write the equation for $\hat{\mathcal{W}}_1$ as

$$\hat{\mathcal{W}}_1 \hat{E}_p + \hat{E}_p \hat{\mathcal{W}}_1 = \beta \hat{\mathcal{O}}_1. \tag{16.55}$$

The solution to this equation is most easily expressed in momentum space, where $\hat{\mathcal{W}}_1$ is an integral operator with the kernel

$$\hat{\mathcal{W}}_1(\mathbf{p}, \mathbf{p}') = \frac{\hat{\mathcal{A}} \hat{\mathcal{R}} V(\mathbf{p}, \mathbf{p}') \hat{\mathcal{A}}' - \hat{\mathcal{A}} V(\mathbf{p}, \mathbf{p}') \hat{\mathcal{R}}' \hat{\mathcal{A}}'}{(\hat{E}_p + \hat{E}_{p'})}. \tag{16.56}$$

The potential V in momentum space also becomes an integral operator. The kernel of this operator for an atom is

$$V(\mathbf{p}, \mathbf{p}') = -Ze^2 \left(\frac{2\hbar}{\pi}\right)^{1/2} \frac{1}{|\mathbf{p} - \mathbf{p}'|^2}. \tag{16.57}$$

The factor in parentheses depends on the definition of the Fourier transform; here we have adopted the convention (Messiah 1961)

$$f(\mathbf{p}) = (2\pi\hbar)^{3/2} \int d^3\mathbf{r} e^{-i\mathbf{p}\cdot\mathbf{r}/\hbar} f(\mathbf{r}). \qquad (16.58)$$

$\hat{\mathcal{W}}_1$ is a linear function of the potential energy, so the expansion of the transformation operator is an expansion in powers of the potential: in fact, since $\hat{E}_p \sim mc^2$ and $\hat{\mathcal{O}}_1 \sim \mathbf{p}V/2mc$, $\hat{\mathcal{W}}_1 \sim \mathbf{p}V/4m^2c^3$. Both $\hat{\mathcal{O}}_1$ and $\hat{\mathcal{E}}_1$ are of first order in the potential.

Having eliminated the first-order odd term, the remaining terms must be of order 2 or higher in the potential. There are three terms of order 2, $[\hat{\mathcal{W}}_1, \hat{\mathcal{E}}_1]$, $[\hat{\mathcal{W}}_1, \hat{\mathcal{O}}_1]$, and $\frac{1}{2}[\hat{\mathcal{W}}_1, [\hat{\mathcal{W}}_1, \beta\hat{E}_p]]$. The first of these is odd, and may be removed by a further transformation. To obtain the Hamiltonian correct to second order it is not necessary to do this because another transformation cannot generate any more terms of second order. The remaining two terms give the second-order correction to the Hamiltonian. From (16.54), the first of these is

$$[\hat{\mathcal{W}}_1, \hat{\mathcal{O}}_1] = -[\hat{\mathcal{W}}_1, [\hat{\mathcal{W}}_1, \beta\hat{E}_p]]. \qquad (16.59)$$

The Hamiltonian correct to second order is then

$$\hat{\mathcal{H}}_2 = \beta\hat{E}_p + \hat{\mathcal{E}}_1 - \frac{1}{2}[\hat{\mathcal{W}}_1, [\hat{\mathcal{W}}_1, \beta\hat{E}_p]]. \qquad (16.60)$$

Extracting the β matrix from the last double commutator, the Hamiltonian may be written

$$\hat{\mathcal{H}}_2 = \beta\hat{E}_p + \hat{\mathcal{E}}_1 - \beta\left(\hat{\mathcal{W}}_1\hat{E}_p\hat{\mathcal{W}}_1 + \frac{1}{2}[\hat{\mathcal{W}}_1^2, \hat{E}_p]_+\right). \qquad (16.61)$$

Further details on the content of this second-order operator may be found in appendix G, where it is shown that the term in parentheses is negative, and therefore the contribution to the energy is positive.

We can classify the Douglas–Kroll expansion in terms of the leading power of c as well as in terms of V. The off-diagonal operator in the zeroth order (free-particle Foldy–Wouthuysen) expansion is $\mathcal{O}(c^{-1})$. Its elimination in the first Douglas–Kroll transformation leaves an odd operator of $\mathcal{O}(c^{-3})$, and even terms that are correct to $\mathcal{O}(c^{-4})$. When this odd operator is also eliminated, by a second Douglas–Kroll transformation, the remaining odd operator is of $\mathcal{O}(c^{-5})$, and the even terms are correct to $\mathcal{O}(c^{-6})$. Thus the Douglas–Kroll transformation can be seen as an expansion in powers of V/c^2.

One issue raised by the use of an expansion in powers of the potential is that of electric gauge invariance. If we add a constant to the potential, we should obtain a constant shift in the energy, if the potential is gauge-invariant. Terms that are of second order in the potential would be expected to give rise to a quadratic term in the added constant. Looking at the free-particle Foldy–Wouthuysen transformed Hamiltonian, (16.42), it is apparent that the added constant only survives in the even operator: the odd operator involves a commutator that eliminates the constant. Consequently, the higher-order Douglas–Kroll transformations do not contain the constant, and since they are unitary transformations (as is the free-particle Foldy–Wouthuysen transformation) the constant may simply be taken out. The transformed Hamiltonians are therefore all gauge-invariant.

The second-order one-electron Douglas–Kroll Hamiltonian has found wide application in quantum chemistry programs through approximations that are discussed in the next two sections. Although it is a considerable improvement on the first-order Hamiltonian, for some heavy elements the error is significant. Hamiltonians through fifth order have been derived by Nakajima and Hirao (2000). The third-order Hamiltonian is given by

$$\hat{\mathcal{H}}_3 = \hat{\mathcal{H}}_2 + \big[\hat{\mathcal{W}}_1, \big[\hat{\mathcal{W}}_1, \hat{\mathcal{E}}_1\big]\big]. \tag{16.62}$$

Notice that the additional term only involves $\hat{\mathcal{W}}_1$. This is an instance of the familiar "$2n+1$ rule" of perturbation theory. Here, the operators up to $\hat{\mathcal{W}}_n$ are all that are needed to determine the Hamiltonian of order $2n+1$. Higher-order transformations have also been derived and examined by Wolf et al. (2002), to which the reader is referred for details. The Douglas–Kroll Hamiltonian of order n is often written as $\hat{\mathcal{H}}_{\mathrm{DK}n}$ or $\hat{\mathcal{H}}^{\mathrm{DK}n}$.

16.4 Two-Electron Terms and the Douglas–Kroll–Hess Approximation

So far we have made no reference to the two-electron terms of the Hamiltonian. Performing the free-particle Foldy–Wouthuysen transformation took the potential from an even operator to a combination of even and odd operators. The same will be true of the two-electron terms, except that a transformation must be performed on both electron coordinates. For the Coulomb interaction the resulting operator is

$$\hat{\mathcal{U}}_i\hat{\mathcal{U}}_j g_{ij}\hat{\mathcal{U}}_j^{-1}\hat{\mathcal{U}}_i^{-1} = \hat{\mathcal{A}}_i\hat{\mathcal{A}}_j\big[g_{ij} + \hat{\mathcal{R}}_i g_{ij}\hat{\mathcal{R}}_i + \hat{\mathcal{R}}_j g_{ij}\hat{\mathcal{R}}_j + \hat{\mathcal{R}}_i\hat{\mathcal{R}}_j g_{ij}\hat{\mathcal{R}}_j\hat{\mathcal{R}}_i\big]\hat{\mathcal{A}}_j\hat{\mathcal{A}}_i. \tag{16.63}$$

To be entirely consistent with the one-particle terms, a second transformation should be applied to this operator. However, it has been found that this first transformation of the two-electron operators is about as important in scalar relativistic calculations as the transformations of the one-electron operators up to fifth order (Wolf et al. 2002). The second transformation of the two-electron operators is therefore unlikely to be of great significance. The reason is the strength of the nuclear potential, which is a factor of Z larger than the electron–electron interaction and is attractive rather than repulsive.

The transformed two-electron operator bears a striking resemblance to the operator from the modified Dirac equation given in (15.43). We need only define $\psi^L = \hat{\mathcal{A}}\psi^{\mathrm{FW}}$ and $\phi^L = [c^2/(\hat{E}_p + mc^2)]\hat{\mathcal{A}}\psi^{\mathrm{FW}}$ and the identity is complete. The analysis of the terms of the modified Dirac equation into scalar and spin–orbit terms in section 15.4 can then be transferred directly to the above equation. The kinematic factors are reintroduced at the end to obtain the final expressions.

The resemblance between the operators is no accident. In the modified Dirac equation, the small component has been modified by extracting $(\boldsymbol{\sigma} \cdot \mathbf{p})/2mc$ to define a pseudo-large component. In the Douglas–Kroll transformation, the small component has been eliminated to a certain order by unitary transformations. In both cases, the two-electron operators include products of $(\boldsymbol{\sigma} \cdot \mathbf{p})$, which essentially regenerate the original small component. The physical content of the Dirac equation does not really change in the transformation, and hence the work done to evaluate the relativistic electron–electron interaction does not decrease either. The main advantage of the Douglas–Kroll-transformed Hamiltonian is that the spin-free version can be fairly readily implemented in nonrelativistic programs by substituting the transformed one- and two-electron integrals.

Even though the number of transformed integrals is no larger than in a nonrelativistic calculation, the cost of the integral evaluation remains. What we would like is an approximation that is no more severe than the truncation of the transformed one-electron operator and that reduces the integral evaluation work.

We can extract the untransformed two-electron operator from (16.63), as follows:

$$\hat{\mathcal{U}}_i\hat{\mathcal{U}}_j g_{ij}\hat{\mathcal{U}}_j^{-1}\hat{\mathcal{U}}_i^{-1} = g_{ij} + \big[\hat{\mathcal{U}}_i, g_{ij}\big]\hat{\mathcal{U}}_i^{-1} + \big[\hat{\mathcal{U}}_j, g_{ij}\big]\hat{\mathcal{U}}_j^{-1} + \big[\hat{\mathcal{U}}_i, \big[\hat{\mathcal{U}}_j, g_{ij}\big]\big]\hat{\mathcal{U}}_j^{-1}\hat{\mathcal{U}}_i^{-1}. \tag{16.64}$$

The commutators in this expression must yield operators of $\mathcal{O}(c^{-2})$ and $\mathcal{O}(c^{-4})$. However, the order in $1/c$ does not tell us anything about their size. If the commutators are small, we could neglect them and use the untransformed Coulomb operator. Evaluating these commutators involves some lengthy algebra, the techniques for which are outlined in appendix H. Even without this algebra, though, it is obvious that the spin-dependent operators are entirely contained within the commutators. Neglecting the commutators can therefore only be done in a spin-free Hamiltonian. This approach was pioneered by Hess (1985, 1986), and the spin-free Douglas–Kroll-transformed Hamiltonian with the bare Coulomb operator is termed the Douglas–Kroll–Hess (DKH) Hamiltonian.

The neglect of the commutators of the transformation with the two-electron operator eliminates the extra two-electron integral work for the relativistic corrections entirely. All that is left is a set of modified one-electron integrals, for which the amount of work required is negligible in comparison to the two-electron work. The work relative to a nonrelativistic calculation is now also negligible. Tests comparing the results of this approximation to the two-electron integrals with results using the full expression for the two-electron integrals (Samzow et al. 1992, Park and Almlöf 1994) indicate that the approximation is very good for lighter elements, and even for the Au atom. However, calculations for $\mathrm{Au_2}$ and $\mathrm{Pt_2}$ suggest that there is a non-negligible contribution from the correction to the bare Coulomb operator. Nevertheless, this approximation has been widely used with considerable success.

16.5 Implementation of the Douglas–Kroll Transformation

The remaining issue to be considered is that of implementation. Performing all the integrations in momentum space is a tedious business, even given that the operators are now algebraic functions of the momentum. Indeed, the complicated functions of the momentum make it difficult to obtain closed-form expressions for the momentum-space integrals, and therefore one would have to resort to numerical integration. However, the task of evaluating the integrals becomes quite easy if a matrix representation of the operators in momentum space is used. This is the approach adopted by Hess and coworkers.

The evaluation of the matrix elements of the momentum-dependent operators is achieved using a theorem for functions of a matrix. If we have a matrix **B** whose eigenvalues b form a diagonal matrix represented by $[b]$ and whose eigenvectors are **X**, any function of the matrix can be expressed in terms of the function of the eigenvalues by

$$f(\mathbf{B}) = \mathbf{X}[f(b)]\mathbf{X}^{\dagger}. \tag{16.65}$$

If we choose **B** to be the matrix of p^2, it is straightforward to generate the matrix of any algebraic function of p^2—for example, the kinematic factors $\hat{\mathcal{A}}$ and $\hat{\mathcal{D}}$. If we transform the integrals to the basis of the eigenvectors of the matrix of p^2, all the kinematic factors become diagonal matrices, and it then becomes simple to apply these to the remaining nondiagonal matrices, which are those of the potentials V and $(\boldsymbol{\sigma} \cdot \mathbf{p})V(\boldsymbol{\sigma} \cdot \mathbf{p})$. Moreover, since the operator p^2 belongs to the totally symmetric irreducible representation (irrep) of any group, any function of p^2 also belongs to the totally symmetric irrep. The expansion basis functions can then be chosen to have the same symmetry as the basis functions on which the kinematic factors are operating.

The key issue here is the accuracy of the representation of p^2. If the basis set is too small, there could be a serious loss of accuracy. However, for a reasonably large primitive basis, the same basis could be used for the representation of p^2 as for the molecular calculations. Even more usefully, since the kinematic factors do not change the symmetry of the atomic basis functions, they can be used to redefine the contraction coefficients. This redefinition essentially generates a contracted basis set in the modified Dirac representation,

$$\mathbf{a}^L = \mathbf{S}^{-1}\mathbf{A}\mathbf{a}, \qquad \mathbf{a}^{P'} = (1/mc)\mathbf{S}^{-1}\mathbf{Q}\mathbf{a}, \tag{16.66}$$

where **A** is the matrix of $\hat{\mathcal{A}}$, **Q** is the matrix of $\hat{\mathcal{Q}}$, **a** is the original matrix of contraction coefficients, and $\mathbf{a}^L$ and $\mathbf{a}^{P'}$ are the coefficients of the large and pseudo-large components. The prime indicates that we have absorbed the factor of $1/2mc$ into the pseudo-large component. The integrals of the first-order transformed potential energy in this basis are

$$\bar{V}_{ij} \simeq \langle a_i^L | V | a_j^L \rangle + \langle a_i^{P'} | (\boldsymbol{\sigma} \cdot \mathbf{p}) V (\boldsymbol{\sigma} \cdot \mathbf{p}) | a_j^{P'} \rangle. \tag{16.67}$$

Redefining the contraction coefficients does involve a further level of approximation, but for a reasonably large basis set this approximation should not be too severe.

One point of caution in this method is that for a sufficiently dense basis for the representation of p^2, there may be problems with linear dependence, and it is possible for numerical rounding to accumulate to such an extent in the sequence of matrix operations needed to define the operators that the final integrals may have significant inaccuracies.

16.6 The Barysz–Sadlej–Snijders Transformation

The complicated nature of the operator $\hat{\mathcal{W}}_1$ raises the question of whether a simpler alternative can be found. Barysz, Sadlej, and Snijders (1997) devised an approach which starts from the free-particle Foldy–Wouthuysen transformation, just as the Douglas–Kroll transformation does. Whereas the Douglas–Kroll approach seeks to eliminate the lowest-order odd term from the transformed Hamiltonian, their approach seeks to be correct to a particular order in $1/c^2$, and it provides a ready means for defining a sequence of approximations of increasing order in $1/c^2$. It is important to note that, while the expansion in powers of $1/c^2$ that resulted from the Foldy–Wouthuysen transformation in section 16.2 generates highly singular operators, this is not true per se of expansions in $1/c^2$. What multiplies $1/c^2$ is all-important. In the case of the Foldy–Wouthuysen transformation it is p^2/m^2, and due to the fact that p can become infinite the series is not convergent. If $1/c^2$ multiplies a well-defined operator then an expansion might indeed be found that is well-behaved.

The first stages of the procedure are precisely the same as for the exact Foldy–Wouthuysen transformation, which we will follow here. The definitions in the paper by Barysz et al. differ by a sign in the lower row of the transformation matrix, but the results obviously do not depend on this sign. We write the transformation matrix in similar form to the formal unitary Foldy–Wouthuysen transformation:

$$\hat{\mathcal{U}}_1 = \begin{pmatrix} \dfrac{1}{\sqrt{1+\hat{\mathcal{X}}_1{}^\dagger\hat{\mathcal{X}}_1}} & \dfrac{1}{\sqrt{1+\hat{\mathcal{X}}_1{}^\dagger\hat{\mathcal{X}}_1}}\hat{\mathcal{X}}_1{}^\dagger \\ -\dfrac{1}{\sqrt{1+\hat{\mathcal{X}}_1\hat{\mathcal{X}}_1{}^\dagger}}\hat{\mathcal{X}}_1 & \dfrac{1}{\sqrt{1+\hat{\mathcal{X}}_1\hat{\mathcal{X}}_1{}^\dagger}} \end{pmatrix}. \tag{16.68}$$

The operator $\hat{\mathcal{X}}_1$ is defined by setting the off-diagonal term of the Hamiltonian to zero, resulting in the operator equation

$$(\hat{\mathcal{H}}_1)_{22}\hat{\mathcal{X}}_1 = -(\hat{\mathcal{H}}_1)_{21} + \hat{\mathcal{X}}_1(\hat{\mathcal{H}}_1)_{11} + \hat{\mathcal{X}}_1(\hat{\mathcal{H}}_1)_{12}\hat{\mathcal{X}}_1. \tag{16.69}$$

$\hat{\mathcal{H}}_1$ is defined by the free-particle Foldy–Wouthuysen transformation, (16.42). For the purpose of better analyzing the transformation and the resultant Hamiltonians, we will make some new definitions of various operators, extracting powers of $1/c$. First, the free-particle energy will be written

$$\hat{E}_p = mc^2\hat{e}_p \tag{16.70}$$

using the dimensionless quantity

$$\hat{e}_p = \sqrt{1 + p^2/m^2c^2}. \tag{16.71}$$

Second, we will shift the energy scale by subtracting the rest energy, and define the kinetic energy operator by

$$\hat{T}_p = mc^2(\hat{e}_p - 1) = \hat{E}_p - mc^2. \tag{16.72}$$

The operator $\hat{\mathcal{A}}$ needs no redefinition because it involves a ratio of quantities of the same order in c. The operator $\hat{\mathcal{R}}_2$ is redefined in terms of another operator $\hat{\mathcal{B}}$ that is of order c^0. $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}$ may be written in terms of $\hat{e}_p$ and other operators as

$$\hat{\mathcal{A}} = \sqrt{\frac{\hat{e}_p + 1}{2\hat{e}_p}}, \qquad \hat{\mathcal{B}} = \frac{(\boldsymbol{\sigma} \cdot \mathbf{p})}{\hat{e}_p + 1} = \frac{1}{mc}\hat{\mathcal{R}}_2. \tag{16.73}$$

Using these definitions, we may write the blocks of the Hamiltonian after the free-particle Foldy–Wouthuysen transformation as

$$\begin{aligned}
(\hat{\mathcal{H}}_1)_{11} &= \hat{T}_p + \hat{\mathcal{A}}\Big(V + \frac{1}{m^2c^2}\hat{\mathcal{B}}V\hat{\mathcal{B}}\Big)\hat{\mathcal{A}}, \\
(\hat{\mathcal{H}}_1)_{21} &= \frac{1}{mc}\hat{\mathcal{A}}[V, \hat{\mathcal{B}}]\hat{\mathcal{A}} = (\hat{\mathcal{H}}_1)_{12}{}^{\dagger}, \\
(\hat{\mathcal{H}}_1)_{22} &= -2mc^2 - \hat{T}_p + \hat{\mathcal{A}}\Big(V + \frac{1}{m^2c^2}\hat{\mathcal{B}}V\hat{\mathcal{B}}\Big)\hat{\mathcal{A}}.
\end{aligned} \tag{16.74}$$

Equation (16.69) can now solved by an iterative procedure, making use of the orders of the various terms in powers of $1/c$ to define a procedure that should be convergent. Since $(\hat{\mathcal{H}}_1)_{22}$ is of order c^2 and $(\hat{\mathcal{H}}_1)_{21}$ is of order c^{-1}, the maximum order of $\hat{\mathcal{X}}_1$ is c^{-3} (as was found for $\hat{\mathcal{W}}_1$ in the Douglas–Kroll transformation). The three terms on the right-hand side of (16.69) are then of order c^{-1}, c^{-3}, and c^{-7}, respectively. The iteration is started by neglecting the second and third terms of (16.69). As $(\hat{\mathcal{H}}_1)_{22}$ is of order c^2, this gives a transformation that is correct to order c^{-3}:

$$\begin{aligned}
\hat{\mathcal{X}}_1^{(0)} &= -\big[(\hat{\mathcal{H}}_1)_{22}\big]^{-1}(\hat{\mathcal{H}}_1)_{21} \\
&= \frac{1}{2m^2c^3}\left(1 + \frac{1}{2mc^2}(\hat{T}_p - \hat{\mathcal{A}}(V + \hat{\mathcal{B}}V\hat{\mathcal{B}}/m^2c^2)\hat{\mathcal{A}})\right)^{-1}\hat{\mathcal{A}}[V, \hat{\mathcal{B}}]\hat{\mathcal{A}} \\
&= \left(\frac{1}{2m^2c^3} - \frac{1}{4m^3c^5}(\hat{T}_p - \hat{\mathcal{A}}V\hat{\mathcal{A}}) + \ldots\right)\hat{\mathcal{A}}[V, \hat{\mathcal{B}}]\hat{\mathcal{A}},
\end{aligned} \tag{16.75}$$

where in the last line we have expanded the inverse operator. To obtain a Hamiltonian correct to $\mathcal{O}(c^{-4})$ we only need the transformation correct to order $1/c^3$, which results in

$$\hat{\mathcal{U}}_1 \simeq \begin{pmatrix} 1 & \hat{\mathcal{A}}[\hat{\mathcal{B}}, V]\hat{\mathcal{A}}/2m^2c^3 \\ \hat{\mathcal{A}}[\hat{\mathcal{B}}, V]\hat{\mathcal{A}}/2m^2c^3 & 1 \end{pmatrix}. \tag{16.76}$$

The transformed Hamiltonian is now

$$\hat{\mathcal{H}}_2 = \hat{\mathcal{U}}_1 \hat{\mathcal{H}}_1 \hat{\mathcal{U}}_1{}^{\dagger}. \tag{16.77}$$

The upper left block of this transformed Hamiltonian is the Hamiltonian for positive-energy states to second order,

$$\hat{\mathcal{H}}_2^{+} = (\hat{\mathcal{H}}_2)_{11} = \hat{T}_p + \hat{\mathcal{A}}(V + \hat{\mathcal{B}}V\hat{\mathcal{B}}/m^2c^2)\hat{\mathcal{A}} - \hat{\mathcal{A}}[V, \hat{\mathcal{B}}]\hat{\mathcal{A}}^2[V, \hat{\mathcal{B}}]\hat{\mathcal{A}}/2m^3c^4. \tag{16.78}$$

This is the equivalent of the second-order Douglas–Kroll operator, but it only involves operators that have been defined in the free-particle Foldy–Wouthuysen transformation. As for the Douglas–Kroll transformed Hamiltonian, spin separation may be achieved with the use of the Dirac relation to define a spin-free relativistic Hamiltonian, and an approximation in which the transformation of the two-electron integrals is neglected, as in the Douglas–Kroll–Hess method, may also be defined. Implementation of this approximation can be carried out in the same way as for the Douglas–Kroll approximation: both approximations involve the evaluation of kinematic factors, which may be done by matrix methods.

16.7 Transformation of Electric Property Operators

The transformed Hamiltonians that we have derived allow us to calculate intrinsic molecular properties, such as geometries and harmonic frequencies. We would like to be able to calculate response properties as well, with wave functions derived from the transformed Hamiltonian. If we used a method such as the Douglas–Kroll–Hess method, it would be tempting to simply evaluate the property using the nonrelativistic property operators and the transformed wave function. As we saw in section 15.3, the property operators can have a relativistic correction, and for properties sensitive to the environment close to the nuclei where the relativistic effects are strong, these corrections are likely to be significant. To ensure that we do not omit important effects, we must derive a transformed property operator, starting from the Dirac form of the property operator.

It would also be tempting to use the transformation from the unperturbed Dirac Hamiltonian to derive the property operator. Again, doing so risks the neglect of terms that might be significant. The proper starting point is to derive a transformation for the perturbed Hamiltonian and then to attempt to separate the transformed Hamiltonian into a zeroth-order part and a perturbation. We already have the zeroth-order part from the transformation of the unperturbed Hamiltonian. What might prove to be difficult is getting an expression for the perturbation.

For electric perturbations, deriving the appropriate perturbation operator turns out to be relatively straightforward because the four-component operator is even. The rest of this section is devoted to derivation and discussion of transformed electric perturbation operators. Magnetic perturbations are a different story, because the four-component operator is odd and therefore enters into the lowest order of the transformation. We will deal with magnetic perturbations in the next section.

As well as deriving the appropriate forms of the electric perturbation operators, we would like to be able to express them in terms of the nonrelativistic operators plus a relativistic correction. This we can do by writing $\hat{\mathcal{H}}^{\rm FW}$, the transformed Hamiltonian including the perturbation, as a sum of the untransformed operator $\hat{\mathcal{H}}$ plus a correction,

$$\hat{\mathcal{H}}^{\rm FW} = \hat{\mathcal{U}}\hat{\mathcal{H}}\hat{\mathcal{U}}^{-1} = \hat{\mathcal{H}} + [\hat{\mathcal{U}}, \hat{\mathcal{H}}]\hat{\mathcal{U}}^{-1}. \tag{16.79}$$

The commutator will (in general) change the order of the perturbation with respect to some expansion parameter and will introduce cross-terms between the unperturbed Hamiltonian and the perturbation.

The electric perturbation operator may be represented as

$$\hat{\mathcal{H}}' = W\mathbf{I}_4 \tag{16.80}$$

where W is the scalar potential corresponding to the perturbation.[1] The electric perturbation operators behave in exactly the same way as the nuclear Coulomb potential, so we may simply substitute $V + W$ for V in all the expressions given above to derive the transformations including the perturbation, and hence obtain the transformed perturbation operators.

In the Pauli approximation, the Hamiltonian correct to second order is linear in the potential, and the separation of the zeroth-order Hamiltonian from the perturbation is trivial. The perturbation operator is

$$\hat{W}^{\rm Pauli} = W + \frac{\hbar^2}{8m^2c^2}(\nabla^2 W) + \frac{\hbar}{4m^2c^2}\boldsymbol{\sigma} \cdot (\nabla W) \times \mathbf{p} \tag{16.81}$$

which has a spin-free and a spin-dependent relativistic correction to the nonrelativistic operator.

External electric fields transform as the regular solid spherical harmonics, $r^\ell Y_{\ell m}(\vartheta, \varphi)$, for which the Laplacian vanishes, $\nabla^2 r^\ell Y_{\ell m}(\vartheta, \varphi) = 0$. There is therefore *no* spin-free correction of $\mathcal{O}(c^{-2})$ to the property operator, and the operator can be written as

$$\hat{W}^{\rm ext,Pauli} = W + \frac{\hbar}{4m^2c^2}\boldsymbol{\sigma} \cdot (\nabla W) \times \mathbf{p}. \tag{16.82}$$

For closed-shell molecules, the spin-dependent correction will only affect the property in second order. As an example of the operators, the perturbation for a constant electric field $e\mathbf{E} \cdot \mathbf{r}$ is

$$\hat{W}^{\rm E,Pauli} = e\mathbf{E} \cdot \left(\mathbf{r} - \frac{\hbar}{4m^2c^2}\boldsymbol{\sigma} \times \mathbf{p}\right). \tag{16.83}$$

Point nuclear electric fields transform as the irregular solid spherical harmonics $r^{-(\ell+1)}Y_{\ell m}(\vartheta, \varphi)$, for which the Laplacian is a generalized delta function. For finite

---

[1] In this section, we use W for the electric perturbation. It should not be confused with the operators $\hat{\mathcal{W}}_n$ used in the Douglas–Kroll transformation.

nuclei the Laplacian does not vanish either. The spin-free relativistic correction for nuclear electric fields is therefore nonzero.

The absence of a spin-free relativistic correction to the external field property operator means that we could in fact use the nonrelativistic operator with the transformed wave function for external electric field properties. For nuclear electric fields, this would not be a reliable procedure because the effects of relativity are most pronounced near the nucleus.

The other transformations that we have considered, the Douglas–Kroll transformation and the Barysz–Sadlej–Snijders transformation, start with a free-particle Foldy–Wouthuysen transformation. This transformation is independent of V and it is a simple matter to separate the perturbation. The transformed perturbation operator can be written down directly from (16.44)

$$\hat{W}^{\mathrm{FPFW}} = \hat{\mathcal{A}}(W + \hat{\mathcal{R}}_2 W \hat{\mathcal{R}}_2)\hat{\mathcal{A}} = \hat{\mathcal{A}} W \hat{\mathcal{A}} + \frac{1}{m^2c^2}\hat{\mathcal{Q}}(\boldsymbol{\sigma}\cdot\mathbf{p})W(\boldsymbol{\sigma}\cdot\mathbf{p})\hat{\mathcal{Q}}. \tag{16.84}$$

Implementation of this operator is no more difficult than implementing the nuclear Coulomb potential, especially if the momentum operators are made to act on the wave function.

Before considering how to separate this operator into a nonrelativistic operator and a relativistic correction, we should examine the effect of the next transformation, $\hat{\mathcal{U}}_1$. In both the Douglas–Kroll transformation and the Barysz–Sadlej–Snijders transformation, the result of using $\hat{\mathcal{U}}_1$ is of second order in the potential. When we substitute $V + W$ for V, we would expect to get terms in V^2, terms in VW, and terms in W^2. This transformation therefore produces a perturbation operator that is linear in the field strength and one that is quadratic in the field strength. Obviously, higher-order transformations will yield higher powers of the field strength.

To illustrate, we use the second-order Barysz–Sadlej–Snijders transformation, which is more transparent than the Douglas–Kroll transformation. Introducing a perturbation parameter λ, the Hamiltonian including the electric perturbation is

$$\begin{aligned}
\hat{\mathcal{H}}_2^+ &= \hat{T}_p + \hat{\mathcal{A}}(V + \lambda W + \hat{\mathcal{B}}(V + \lambda W)\hat{\mathcal{B}}/m^2c^2)\hat{\mathcal{A}} \\
&\quad - \hat{\mathcal{A}}[V + \lambda W, \hat{\mathcal{B}}]\hat{\mathcal{A}}^2[V + \lambda W, \hat{\mathcal{B}}]\hat{\mathcal{A}}/2m^3c^4 \\
&= \hat{T}_p + \hat{\mathcal{A}}(V + \hat{\mathcal{B}}V\hat{\mathcal{B}}/m^2c^2)\hat{\mathcal{A}} - \hat{\mathcal{A}}[V, \hat{\mathcal{B}}]\hat{\mathcal{A}}^2[V, \hat{\mathcal{B}}]\hat{\mathcal{A}}/2m^3c^4 \\
&\quad + \lambda\Big(\hat{\mathcal{A}}(W + \hat{\mathcal{B}}W\hat{\mathcal{B}}/m^2c^2)\hat{\mathcal{A}} - \hat{\mathcal{A}}[V, \hat{\mathcal{B}}]\hat{\mathcal{A}}^2[W, \hat{\mathcal{B}}]\hat{\mathcal{A}}/2m^3c^4 \\
&\quad - \hat{\mathcal{A}}[W, \hat{\mathcal{B}}]\hat{\mathcal{A}}^2[V, \hat{\mathcal{B}}]\hat{\mathcal{A}}/2m^3c^4\Big) \\
&\quad - \lambda^2\hat{\mathcal{A}}[W, \hat{\mathcal{B}}]\hat{\mathcal{A}}^2[W, \hat{\mathcal{B}}]\hat{\mathcal{A}}/2m^3c^4.
\end{aligned} \tag{16.85}$$

The terms that come from the second-order transformation are of $\mathcal{O}(c^{-4})$, and for external field perturbations we can expect these terms to be small. For nuclear electric field perturbations, these terms could be large and should not be neglected.

We would now like to express the electric perturbations as an untransformed operator plus a correction, using (16.79), and we will consider only the free-particle

Foldy–Wouthuysen transformation, (16.19), which may be cast in the form

$$\hat{\mathcal{U}}^{\mathrm{FPFW}} = \begin{pmatrix} \hat{\mathcal{A}} & \hat{\mathcal{Q}}(\boldsymbol{\sigma}\cdot\mathbf{p})/mc \\ -\hat{\mathcal{Q}}(\boldsymbol{\sigma}\cdot\mathbf{p})/mc & \hat{\mathcal{A}} \end{pmatrix} \tag{16.86}$$

using the operators $\hat{\mathcal{A}}$ and $\hat{\mathcal{Q}} = mc^2\hat{\mathcal{D}}\hat{\mathcal{A}}$ introduced previously. The perturbation operator is given in (16.84). The first term is easily transformed into a commutator expression,

$$\hat{\mathcal{A}}W\hat{\mathcal{A}} = [\hat{\mathcal{A}}, W]\hat{\mathcal{A}} + W\hat{\mathcal{A}}\hat{\mathcal{A}}. \tag{16.87}$$

The term involving the $\hat{\mathcal{Q}}$ operators may be expanded using the Dirac relation (4.14)

$$\begin{aligned} \hat{\mathcal{Q}}(\boldsymbol{\sigma}\cdot\mathbf{p})W(\boldsymbol{\sigma}\cdot\mathbf{p})\hat{\mathcal{Q}} &= \hat{\mathcal{Q}}W(\boldsymbol{\sigma}\cdot\mathbf{p})(\boldsymbol{\sigma}\cdot\mathbf{p})\hat{\mathcal{Q}} + \hat{\mathcal{Q}}[\boldsymbol{\sigma}\cdot(\mathbf{p}W)](\boldsymbol{\sigma}\cdot\mathbf{p})\hat{\mathcal{Q}} \\ &= \hat{\mathcal{Q}}W\mathbf{p}^2\hat{\mathcal{Q}} + \hat{\mathcal{Q}}[(\mathbf{p}W)\cdot\mathbf{p}]\hat{\mathcal{Q}} + \hat{\mathcal{Q}}i\boldsymbol{\sigma}\cdot[(\mathbf{p}W)\times\mathbf{p}]\hat{\mathcal{Q}}. \end{aligned} \tag{16.88}$$

These three terms may each be turned into commutator expressions, as was done for the terms in $\hat{\mathcal{A}}$ above. Using the fact that the transformation is unitary and therefore

$$W\hat{\mathcal{A}}\hat{\mathcal{A}} + W\frac{\mathbf{p}^2}{m^2c^2}\hat{\mathcal{Q}}\hat{\mathcal{Q}} = W \tag{16.89}$$

we arrive at an expression of the desired form

$$\begin{aligned} \hat{W}^{\mathrm{FPFW}} &= W + [\hat{\mathcal{A}}, W]\hat{\mathcal{A}} + [\hat{\mathcal{Q}}, W]\frac{\mathbf{p}^2}{m^2c^2}\hat{\mathcal{Q}} \\ &\quad + \frac{1}{m^2c^2}(\mathbf{p}W)\cdot\mathbf{p}\hat{\mathcal{Q}}^2 + \frac{1}{m^2c^2}[\hat{\mathcal{Q}}, (\mathbf{p}W)\cdot\mathbf{p}]\hat{\mathcal{Q}} \\ &\quad + \frac{\hbar}{m^2c^2}\boldsymbol{\sigma}\cdot(\nabla W)\times\mathbf{p}\hat{\mathcal{Q}}^2 + \frac{\hbar}{m^2c^2}[\hat{\mathcal{Q}}, \boldsymbol{\sigma}\cdot(\nabla W)\times\mathbf{p}]\hat{\mathcal{Q}}. \end{aligned} \tag{16.90}$$

Although we have an expression with W plus some correction term, we still have commutators of $\hat{\mathcal{A}}$ and $\hat{\mathcal{Q}}$ with W that are not exactly transparent. Using the techniques described in appendix H, we can simplify these expressions to arrive at an expression expanded in leading powers of $1/c^2$ to $\mathcal{O}(c^{-4})$:

$$\begin{aligned} \hat{W}^{\mathrm{FPFW}} &= W + \frac{\hbar^2}{m^2c^2}(\nabla^2 W)\frac{1}{4\hat{e}_p(1+\hat{e}_p)} + \frac{\hbar}{m^2c^2}\boldsymbol{\sigma}\cdot(\nabla W)\times\mathbf{p}\frac{1}{2\hat{e}_p(1+\hat{e}_p)} \\ &\quad + \frac{\hbar^4}{m^4c^4}[\nabla^2, (\nabla^2 W)]\frac{6\hat{e}_p^2+4\hat{e}_p+1}{32\hat{e}_p^4(1+\hat{e}_p)^2} + \frac{\hbar^4}{m^4c^4}[\nabla^2, (\nabla W)\cdot\nabla]\frac{2\hat{e}_p^2+2\hat{e}_p+1}{16\hat{e}_p^4(1+\hat{e}_p)^2} \\ &\quad + \frac{\hbar^3}{m^4c^4}[\nabla^2, \boldsymbol{\sigma}\cdot(\nabla W)\times\mathbf{p}]\frac{2\hat{e}_p+1}{8\hat{e}_p^3(1+\hat{e}_p)^2}. \end{aligned} \tag{16.91}$$

The terms of leading order $1/(mc)^2$ comprise a Darwin-like term and a spin–orbit term, and reduce to the Pauli expressions in the limit $\hat{e}_p \to 1$. These terms are therefore regularized, just like the nuclear potential terms. The leading spin-free term is a term in $\nabla^2 W$, which vanishes for external electric fields, just as it does in the Pauli Hamiltonian.

Expressions for the corrections due to the Douglas–Kroll and Barysz–Sadlej–Snijders transformations may be derived in the same way. However, since these operators are only correct to $\mathcal{O}(c^{-4})$ and have their leading term at $\mathcal{O}(c^{-4})$, we only need the lowest-order corrections. The Barysz–Sadlej–Snijders correction is

$$\Delta \hat{W}^{\mathrm{BSS}} = \frac{\hbar^2}{2m^3c^4} \left(2(\nabla V)\cdot(\nabla W) + (\nabla W)^2\right) \frac{1}{4\hat{e}_p^2}. \tag{16.92}$$

This is a scalar term: the spin-dependent terms that are second order in the potential have a higher leading power of $1/(mc)^2$. The Douglas–Kroll correction including the nuclear potential term can be derived from the lowest-order part of the $\hat{\mathcal{W}}_1^2$ term,

$$\hat{\mathcal{W}}_1^2 = \hbar^2 c^2 \hat{\mathcal{A}}\hat{\mathcal{D}}\nabla \left(\tilde{V}(\mathbf{p},\mathbf{p}') + \tilde{W}(\mathbf{p},\mathbf{p}')\right)\cdot(\hat{\mathcal{A}}')^2\hat{\mathcal{D}}'\nabla\left(\tilde{V}(\mathbf{p}',\mathbf{p}'') + \tilde{W}(\mathbf{p}',\mathbf{p}'')\right)\hat{\mathcal{A}}'' + \ldots, \tag{16.93}$$

where $\tilde{W}$ is similarly defined to $\tilde{V}$. Insertion of the factors of $\hat{e}_p$ is straightforward. As an example from this class of transformation, the Barysz–Sadlej–Snijders operator correct to $\mathcal{O}(c^{-4})$ for a constant electric field is

$$\hat{W}^{\mathrm{BSS,E}} = e\mathbf{E}\cdot\left(\mathbf{r} + \frac{\hbar}{m^2c^2}\,\frac{\boldsymbol{\sigma}\times\mathbf{p}}{2\hat{e}_p(1+\hat{e}_p)} + \frac{\hbar^2}{2m^3c^4}\,(2\nabla V + e\mathbf{E})\frac{1}{4\hat{e}_p^2}\right). \tag{16.94}$$

The fact that there is a relativistic correction to the property operators is a manifestation of what has been called the “picture change” (see for example Kellö and Sadlej 1998). The Dirac operator and the corresponding Schrödinger operator do not have the same meaning. For the example given above, the position operator **r** does not have the same content in the Dirac picture as in the Schrödinger picture. More information on the interpretation of the picture change can be found in Moss (1973), for example.

16.8 Transformation of Magnetic Property Operators

If applying a transformation derived from the decoupling of the unperturbed Dirac Hamiltonian was potentially unreliable for electric properties, for magnetic properties it simply will not work. The reason is that magnetic perturbations enter through the vector potential **A** and thus are odd operators,

$$\hat{\mathcal{H}}' = ec\boldsymbol{\alpha}\cdot\mathbf{A}. \tag{16.95}$$

None of the transformations we have considered will transform this operator completely into an even operator. If we use the exponential ansatz of (16.23), the first term in the commutator series is the untransformed Hamiltonian. The first commutator is meant to cancel the odd term in the untransformed Hamiltonian, but if we use the definition of $\hat{\mathcal{S}}$ in (16.26), only $c\boldsymbol{\alpha} \cdot \mathbf{p}$ is canceled, and not $ec\boldsymbol{\alpha} \cdot \mathbf{A}$. This makes sense: $\mathbf{A}$ does not appear anywhere in the transformation. We fare no better with the free-particle Foldy–Wouthuysen transformation, which yields

$$\hat{\mathcal{H}}_1' = ec\hat{\mathcal{A}}\left(\boldsymbol{\alpha} \cdot \mathbf{A} + \beta\Big[\hat{\mathcal{R}}, \boldsymbol{\alpha} \cdot \mathbf{A}\Big]_+ - \hat{\mathcal{R}}\boldsymbol{\alpha} \cdot \mathbf{A}\hat{\mathcal{R}}\right)\hat{\mathcal{A}}. \tag{16.96}$$

Only the second term of this operator is even; the other two are odd.

We must therefore derive a transformation that includes the vector potential. In whichever way we define the sequence of transformations, we are using a momentum-dependent operator in the first step—and indeed in subsequent steps as well. The vector potential $\mathbf{A}$ enters the Dirac equation in the same way as the momentum $\mathbf{p}$. Consequently, any transformation that contains the momentum must also contain the vector potential. The correct procedure is to replace $\mathbf{p}$ with $\boldsymbol{\pi} = \mathbf{p} + e\mathbf{A}$ in the transformation, and transform the Hamiltonian with this new transformation operator.

From our experience to date with the transformations, we can immediately foresee a problem. If the transformation is some complicated function of the momentum, we might not be able to separate out the perturbation from the zeroth-order Hamiltonian. This would be unfortunate, because magnetic operators break Kramers symmetry and we would be forced to perform calculations without spin (or time-reversal) symmetry. We might also be forced to perform finite-field calculations. We will address this problem as it arises.

First we consider the Pauli Hamiltonian, and replace $\mathbf{p}$ with $\mathbf{p} + e\mathbf{A}$ in the operator $\hat{\mathcal{S}}$, which defines the first-order transformation, (16.26). Because $\hat{\mathcal{S}}$ is linear in the momentum, it is easy to separate the perturbation and the zeroth-order Hamiltonian. The odd operator in the Dirac Hamiltonian with the inclusion of the vector potential is now

$$\hat{\mathcal{O}} = c\boldsymbol{\alpha} \cdot (\mathbf{p} + e\mathbf{A}). \tag{16.97}$$

The transformation operator $\hat{\mathcal{S}}$ becomes

$$\hat{\mathcal{S}} = -i\beta\boldsymbol{\alpha} \cdot (\mathbf{p} + e\mathbf{A})/2mc. \tag{16.98}$$

The substitution does not affect the formal expressions for the commutators, and we may replace the operators in (16.35) with their perturbed forms.

Considering first the term involving the square of $\hat{\mathcal{O}}$, and making use of the Dirac relation, we have

$$\begin{aligned}\frac{\hat{\mathcal{O}}^2}{2mc^2} &= \frac{1}{2m}\boldsymbol{\alpha} \cdot (\mathbf{p} + e\mathbf{A})\boldsymbol{\alpha} \cdot (\mathbf{p} + e\mathbf{A}) \\ &= \frac{\mathbf{p}^2}{2m} + \frac{e}{2m}(\mathbf{p} \cdot \mathbf{A} + \mathbf{A} \cdot \mathbf{p}) + \frac{e^2}{2m}\mathbf{A}^2 + \frac{ie}{2m}\boldsymbol{\Sigma} \cdot (\mathbf{p} \times \mathbf{A} + \mathbf{A} \times \mathbf{p}).\end{aligned} \tag{16.99}$$

Here, we have dropped the factor of β because we are only interested in the positive-energy states. The first term in this equation is the only one we obtained without the vector potential. The remaining terms clearly all belong to the perturbation supplied by the vector potential, and one of them is quadratic in the vector potential. In the Coulomb gauge $\nabla \cdot \mathbf{A} = 0$, and from Maxwell's equations $\nabla \times \mathbf{A} = \mathbf{B}$, so we may rewrite this expression as

$$\frac{\hat{\mathcal{O}}^2}{2mc^2} = \hat{T} + \frac{e}{m}\mathbf{A} \cdot \mathbf{p} + \frac{e^2}{2m}\mathbf{A}^2 + \frac{e\hbar}{2m}\boldsymbol{\Sigma} \cdot \mathbf{B}. \tag{16.100}$$

The introduction of a vector potential therefore introduces *at order* c^0 the interaction of the spin with a magnetic field, as well as the scalar operators. For a constant magnetic field, the second term may be rewritten as $(e/2m)\mathbf{B} \cdot \mathbf{r} \times \mathbf{p}$, and with the replacement of $\hbar\boldsymbol{\Sigma}$ with $2\mathbf{s}$ we have

$$\frac{\hat{\mathcal{O}}^2}{2mc^2} = \hat{T} + \frac{e^2}{2m}\mathbf{A}^2 + \frac{e}{2m}(2\mathbf{s} + \boldsymbol{\ell}) \cdot \mathbf{B}. \tag{16.101}$$

The magnetic field term gives rise to the Zeeman effect. These terms all arise from the nonrelativistic spin Hamiltonian, (4.22).

The relativistic corrections to the magnetic perturbation come from the remaining two terms in (16.35). The commutator term involving the even operator produces a cross-term between the nuclear potential and the magnetic perturbation:

$$-\frac{1}{8m^2c^4}[\hat{\mathcal{O}}, [\hat{\mathcal{O}}, \hat{\mathcal{E}}]] = \frac{\hbar^2}{8m^2c^2}(\nabla^2 V) + \frac{\hbar}{4m^2c^2}\boldsymbol{\Sigma} \cdot (\nabla V) \times (\mathbf{p} + e\mathbf{A}). \tag{16.102}$$

There is no spin-free part to the cross-term, only a spin–orbit part. For a constant magnetic field, $\mathbf{A} = \frac{1}{2}\mathbf{B} \times \mathbf{r}$, and a point nuclear potential for which $\nabla V = (eZ/r^3)\mathbf{r}$, the cross-term becomes

$$\begin{aligned} \frac{e\hbar}{4m^2c^2}\boldsymbol{\Sigma} \cdot (\nabla V) \times \mathbf{A} &= \frac{e^2 Z\hbar}{4m^2c^2r^3}\boldsymbol{\Sigma} \times \mathbf{r} \cdot \mathbf{B} \times \mathbf{r} \\ &= \frac{e^2\hbar}{4m^2c^2}\frac{Z}{r}\left[\boldsymbol{\Sigma} \cdot \mathbf{B} - \frac{(\boldsymbol{\Sigma} \cdot \mathbf{r})(\mathbf{B} \cdot \mathbf{r})}{r^2}\right]. \end{aligned} \tag{16.103}$$

Here we have expanded the scalar quadruple product. The dominant term is the first term, which is a relativistic correction to the spin Zeeman term. The corrected spin Zeeman operator is

$$\frac{e\hbar}{2m}\boldsymbol{\Sigma} \cdot \mathbf{B}\left(1 + \frac{1}{2mc^2}\frac{eZ}{r}\right) = \frac{e\hbar}{2m}\boldsymbol{\Sigma} \cdot \mathbf{B}\left(1 - \frac{V}{2mc^2}\right). \tag{16.104}$$

The term in $\hat{\mathcal{O}}^4$ from (16.35) produces kinematic corrections and terms that are of higher order in $\mathbf{A}$:

$$\begin{aligned}
-\frac{1}{8m^3c^6}\hat{\mathcal{O}}^4 &= -\frac{1}{8m^3c^2}\left[\mathbf{p}^2 + 2e\mathbf{A}\cdot\mathbf{p} + e^2\mathbf{A}^2 + e\hbar\boldsymbol{\Sigma}\cdot\mathbf{B}\right]^2 \\
&= -\frac{1}{8m^3c^2}\Big[\mathbf{p}^4 + 2e(\mathbf{p}^2\mathbf{A}\cdot\mathbf{p} + \mathbf{A}\cdot\mathbf{p}\mathbf{p}^2) + e\hbar(\boldsymbol{\Sigma}\cdot\mathbf{B}\mathbf{p}^2 + \mathbf{p}^2\boldsymbol{\Sigma}\cdot\mathbf{B}) \\
&\quad + e^2\left(\mathbf{p}^2\mathbf{A}^2 + 4(\mathbf{A}\cdot\mathbf{p})^2 + \mathbf{A}^2\mathbf{p}^2\right) + e^2\hbar^2(\boldsymbol{\Sigma}\cdot\mathbf{B})^2 \\
&\quad + 2e^2\hbar(\boldsymbol{\Sigma}\cdot\mathbf{B}\mathbf{A}\cdot\mathbf{p} + \mathbf{A}\cdot\mathbf{p}\boldsymbol{\Sigma}\cdot\mathbf{B}) \\
&\quad + 2e^3\hbar\boldsymbol{\Sigma}\cdot\mathbf{B}\mathbf{A}^2 + e^4\mathbf{A}^4\Big]. \qquad (16.105)
\end{aligned}$$

Again, only the $\mathbf{p}^4$ operator comes from the unperturbed Hamiltonian, and all the remaining terms originate in the perturbation. We can therefore make a clean separation of the zeroth-order Hamiltonian and the perturbation from the external field and treat the field-dependent terms analytically. This term provides the *kinetic* relativistic corrections to the magnetic field perturbation. Of most interest are the linear terms. It is easy to show that $\mathbf{p}^2$ commutes with $\mathbf{A}\cdot\mathbf{p}$ for a constant magnetic field. The Zeeman term with the kinetic relativistic correction is then

$$\frac{e}{2m}(\boldsymbol{\ell} + 2\mathbf{s})\cdot\mathbf{B}\left(1 - \frac{\mathbf{p}^2}{2m^2c^2}\right). \qquad (16.106)$$

Notice that the kinetic relativistic correction applies to both the spin and orbital Zeeman terms, but the relativistic correction from the potential only affects the spin Zeeman term. These corrections are only the first in a truncated expansion, and for a better description of relativistic corrections we turn to the free-particle Foldy–Wouthuysen transformation and the higher transformations that are based on it.

The magnetic terms for the free-particle Foldy–Wouthuysen transformation are more complicated because the operator is not linear in the momentum. It is possible nevertheless to define the field-dependent transformation by the same procedure used in section 16.1. We arrive at a two-component field-dependent operator corresponding to $\hat{e}_p$, which we will call $\hat{\epsilon}_p$,

$$\hat{\epsilon}_p = \sqrt{\mathbf{I} + (\boldsymbol{\sigma}\cdot\boldsymbol{\pi})^2/m^2c^2} = \sqrt{\mathbf{I} + (\boldsymbol{\sigma}\cdot(\mathbf{p} + e\mathbf{A}))^2/m^2c^2}. \qquad (16.107)$$

Note that we cannot simply replace $(\boldsymbol{\sigma}\cdot\boldsymbol{\pi})^2$ with $\boldsymbol{\pi}^2$ because the interaction of the spin with the magnetic field would not be included.

We may now substitute $\hat{\epsilon}_p$ for $\hat{e}_p$ in the kinematic factors in the transformed Hamiltonian of (16.44), and we will add a prime to indicate that they include the vector potential. In addition, we have to substitute $\boldsymbol{\sigma}\cdot\boldsymbol{\pi}$ for $\boldsymbol{\sigma}\cdot\mathbf{p}$ in $\hat{\mathcal{R}}$ (or $\hat{\mathcal{B}}$). The result is

$$\begin{aligned}
\hat{\mathcal{H}}_1^{+\prime} &= mc^2\hat{\epsilon}_p + \hat{\mathcal{A}}'V\hat{\mathcal{A}}' + \frac{1}{m^2c^2}\hat{\mathcal{Q}}'\boldsymbol{\sigma}\cdot\boldsymbol{\pi}V\boldsymbol{\sigma}\cdot\boldsymbol{\pi}\hat{\mathcal{Q}}' \\
&= mc^2\hat{\epsilon}_p + \hat{\mathcal{A}}'V\hat{\mathcal{A}}' + \frac{1}{m^2c^2}\hat{\mathcal{Q}}'\Big(\mathbf{p}V\cdot\mathbf{p} + \hbar\boldsymbol{\sigma}\cdot\nabla V\times\mathbf{p}\Big)\hat{\mathcal{Q}}'
\end{aligned}$$

$$
\begin{aligned}
&+\frac{e}{m^2c^2}\hat{\mathcal{Q}}'\Big(V\mathbf{A}\cdot\mathbf{p}+\mathbf{A}\cdot\mathbf{p}V+\hbar\boldsymbol{\sigma}\cdot\mathbf{B}V+\hbar\boldsymbol{\sigma}\cdot\nabla V\times\mathbf{A}\Big)\hat{\mathcal{Q}}'\\
&+\frac{e^2}{m^2c^2}\hat{\mathcal{Q}}'V\mathbf{A}^2\hat{\mathcal{Q}}'
\end{aligned}
\tag{16.108}
$$

where we have expanded the terms involving $\boldsymbol{\sigma}\cdot\boldsymbol{\pi}$ from $\hat{\mathcal{B}}$. The expansion separates out some of the vector potential terms, but we still have a problem because the vector potential is tied up in complicated ways in the kinematic factors.

To proceed, we will have to expand this expression in powers of the perturbation. A general approach to this problem is given in appendix H, and here we will only quote the results for the terms linear in the vector potential and correct to lowest order in an expansion in $1/m^2c^2$:

$$
\begin{aligned}
\hat{\mathcal{H}}_1^{+\prime}&=mc^2\hat{e}_p+\hat{\mathcal{A}}V\hat{\mathcal{A}}+\frac{1}{m^2c^2}\hat{\mathcal{Q}}\Big(\mathbf{p}V\cdot\mathbf{p}+\hbar\boldsymbol{\sigma}\cdot\nabla V\times\mathbf{p}\Big)\hat{\mathcal{Q}}\\
&+\frac{e}{2m}(2\mathbf{A}\cdot\mathbf{p}+\hbar\boldsymbol{\sigma}\cdot\mathbf{B})\frac{1}{\hat{e}_p}+\frac{e}{m^2c^2}\hat{\mathcal{Q}}\hbar\boldsymbol{\sigma}\cdot\nabla V\times\mathbf{A}\hat{\mathcal{Q}}+\dots\,.
\end{aligned}
\tag{16.109}
$$

The perturbation operator to lowest order is the last line of this expression. For a constant magnetic field and a point nucleus, this perturbation operator is

$$
\frac{e}{2m}(\boldsymbol{\ell}+2\mathbf{s})\cdot\mathbf{B}\frac{1}{\hat{e}_p}+\frac{e^2}{m^2c^2}\hat{\mathcal{Q}}\frac{Z}{r}\left[2\mathbf{s}\cdot\mathbf{B}-\frac{2(\mathbf{s}\cdot\mathbf{r})(\mathbf{B}\cdot\mathbf{r})}{r^2}\right]\hat{\mathcal{Q}}.
\tag{16.110}
$$

As for the Pauli Hamiltonian, there is a kinematic correction to the entire Zeeman term, but only the spin Zeeman term has a relativistic correction from the potential. Expanding the kinematic factors to order $\mathbf{p}^2/m^2c^2$, we have $\hat{e}_p\simeq 1-\mathbf{p}^2/2m^2c^2$ and $\hat{\mathcal{Q}}\simeq 1/2$, and on substitution this perturbation operator reduces to the Pauli operator to $\mathcal{O}(c^{-2})$.

Terms of higher leading order in $1/m^2c^2$ arise from the free-particle Foldy–Wouthuysen transformation; the terms of leading order $1/m^4c^4$ are given in appendix H. The terms from the corrections in the Douglas-Kroll and Barysz–Sadlej–Snijders transformations are of leading order $1/m^6c^6$, and for most purposes can therefore be neglected.

The complicated nature of the magnetic operators means that the transformed Hamiltonians based on the free-particle Foldy–Wouthuysen transformation are not particularly well suited to calculation of magnetic properties. An alternative is to use the transformed magnetic operator developed in section 13.7, which is also expanded in powers of $1/c^2$, but in terms of even operators, which behave like electric field operators.

17

Perturbation Methods

Perturbation theory has been one of the most frequently used and most powerful tools of quantum mechanics. The very foundations of relativistic quantum theory—quantum electrodynamics—are perturbative in nature. Many-body perturbation theory has been used for electron correlation treatments since the early days of quantum chemistry, and in more recent times multireference perturbation theories have been developed to provide quantitative or semiquantitative information in very complex systems.

In the beginnings of relativistic quantum mechanics, perturbation methods based on an expansion in powers of the fine structure constant, $\alpha = 1/c$, were used extensively to obtain operators that would provide a connection with nonrelativistic quantum mechanics and permit some evaluation of relativistic corrections, in days well before the advent of the computer. This seems a reasonable approach, considering the small size of the fine structure constant—and for light elements it has been found to work remarkably well. Relativity *is* a small perturbation for a good portion of the periodic table.

Perturbation expansions have their limitations, however, and as well as successes, there have been failures due to the highly singular or unbounded nature of the operators in the perturbation expansions. Therefore, in recent times other perturbation approaches have been developed to provide alternatives to the standard Breit–Pauli approach.

This chapter is devoted to the development of perturbation expansions in powers of $1/c$ from the Dirac equation. In the previous chapter, the Pauli Hamiltonian was developed using the Foldy–Wouthuysen transformation. While this is an elegant method, it is probably simpler to make the derivation from the elimination of the small component with expansion of the denominator, and it is this approach that we use here. Another convenient approach is to make use of the modified Dirac equation in the limit of equality of the large and pseudo-large components. This approach enables us to draw on results from the modified Dirac approach in developing the two-electron terms of the Breit–Pauli Hamiltonian. We then demonstrate how the use of perturbation theory for relativistic corrections requires that multiple perturbation theory be employed for correlation effects and for properties. The last sections of this chapter are devoted to the direct perturbation approach of Rutkowski and coworkers (Rutkowski 1986, Rutkowski et al. 1988, 1992) and Kutzelnigg (1989, 1990), in which the small component is retained in the expansion. These techniques circumvent some of the mathematical difficulties inherent in Pauli perturbation theory.

Before we begin, a note about terminology is needed. Conventionally, the expansion parameter used for the relativistic perturbation is $1/c$, to maintain the connection between the order of perturbation theory and the powers of c. In the relativistic Hamiltonian and the relativistic expansion of the energy, the powers of $1/c$ are all even.[1] Thus, the corrections labeled with a "2" are $\mathcal{O}(c^{-2})$ and are the lowest-order perturbation corrections. In terms of an ordinary perturbation expansion these would be called first-order corrections. When a comparison with other perturbations is needed, we will try to make the order in perturbation theory clear as well as the order in $1/c$. Otherwise, the relativistic corrections are labeled with the power of $1/c$ as "second-order," "fourth-order," and so on.

17.1 The Pauli Hamiltonian

To develop the Pauli Hamiltonian via perturbation theory we start from the Dirac equation in two-component form,

$$(V - E)\psi^L + c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^S = 0 \tag{17.1a}$$

$$c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L + (V - E - 2mc^2)\psi^S = 0. \tag{17.1b}$$

The first step is to eliminate the small component using the second of these equations,

$$\psi^S = (2mc^2 - V + E)^{-1}\, c(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L, \tag{17.2}$$

and then substitute it into the first, to get

$$(V - E)\psi^L + \frac{1}{2m}(\boldsymbol{\sigma} \cdot \mathbf{p})\left[1 - \frac{V - E}{2mc^2}\right]^{-1}(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L = 0. \tag{17.3}$$

Here, a factor of $2mc^2$ has been taken out of the denominator term of (17.2). As an equation in its own right, this equation has two major problems. First, it is nonlinear in the energy, which appears in the denominator. Second, it has the potential in the denominator also, making it hard to evaluate integrals. Since we are interested in a perturbation theory that starts from a nonrelativistic Hamiltonian, the obvious thing to do is to expand the expression in square brackets in powers of $1/2mc^2$. The validity of this expansion depends on $|V - E|$ being less than $2mc^2$—a point to which we will return later.

---

[1] The exception is the Lamb shift, which we treat briefly and is $\mathcal{O}(c^{-3})$.

Performing the expansion of the term in square brackets,

$$\left[1 - \frac{V-E}{2mc^2}\right]^{-1} = 1 + \frac{(V-E)}{2mc^2} + \frac{(V-E)^2}{4m^2c^4} + \cdots, \tag{17.4}$$

and substituting into (17.3), we get the operator $\hat{\mathcal{H}}^{\mathrm{ESC}}$, defined by

$$\left[\hat{\mathcal{H}}^{\mathrm{ESC}} - E\right]\psi^L = \left[V - E + \frac{1}{2m}(\boldsymbol{\sigma}\cdot\mathbf{p})(\boldsymbol{\sigma}\cdot\mathbf{p}) + \frac{1}{4m^2c^2}(\boldsymbol{\sigma}\cdot\mathbf{p})(V-E)(\boldsymbol{\sigma}\cdot\mathbf{p}) + \ldots\right]\psi^L = 0. \tag{17.5}$$

From the Dirac relation (4.14), $(\boldsymbol{\sigma}\cdot\mathbf{p})(\boldsymbol{\sigma}\cdot\mathbf{p}) = \mathbf{p}^2$, and the first term in the series gives $\mathbf{p}^2/2m = \hat{T}$, the nonrelativistic kinetic energy operator. Making use of the Dirac relation in the second term gives

$$(\boldsymbol{\sigma}\cdot\mathbf{p})(V-E)(\boldsymbol{\sigma}\cdot\mathbf{p}) = (V-E)\mathbf{p}^2 + \hbar[\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{p} - i(\nabla V)\cdot\mathbf{p}]. \tag{17.6}$$

There is a clear connection of these operators with those of the modified Dirac equation, except that they are operating on the large component ψ^L rather than on the pseudo-large component ϕ^L.

The development is not complete, because the large component of the Dirac wave function is not normalized: it is the four-component wave function that is normalized. Of course, in the limit $c \to \infty$, the large component goes over to the nonrelativistic wave function and is therefore normalized, but we are interested in finite c and a perturbation expansion correct to $\mathcal{O}(c^{-2})$. We therefore need to make the replacement

$$\psi^N = \hat{\mathcal{O}}\psi^L \tag{17.7}$$

where ψ^N is now normalized. Writing the normalization condition as

$$1 = \int d\tau\,\psi^{N\dagger}\psi^N = \int d\tau\,\psi^{L\dagger}\hat{\mathcal{O}}^\dagger\hat{\mathcal{O}}\psi^L = \int d\tau\left[\psi^{L\dagger}\psi^L + \psi^{S\dagger}\psi^S\right] \tag{17.8}$$

and inserting the relation between the large and small components, we find

$$\hat{\mathcal{O}}^\dagger\hat{\mathcal{O}} = 1 + \frac{1}{4m^2c^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\left[1 - \frac{V-E}{2mc^2}\right]^{-2}(\boldsymbol{\sigma}\cdot\mathbf{p}). \tag{17.9}$$

The simplest solution to this equation is to require $\hat{\mathcal{O}}$ to be Hermitian, and take the square root, to produce an operator

$$\hat{\mathcal{O}} = \left[1 + \frac{1}{4m^2c^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\left[1 - \frac{V-E}{2mc^2}\right]^{-2}(\boldsymbol{\sigma}\cdot\mathbf{p})\right]^{1/2}. \tag{17.10}$$

This operator is expanded in a double power series, first for the square root and then for the inverse square operator:

$$\begin{aligned}\hat{\mathcal{O}} &= 1 + \frac{1}{8m^2c^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\left[1 + \frac{2(V-E)}{2mc^2} + \cdots\right](\boldsymbol{\sigma}\cdot\mathbf{p}) + \cdots \\ &= 1 + \frac{1}{8m^2c^2}\mathbf{p}^2 + \mathcal{O}\big(c^{-4}\big).\end{aligned} \tag{17.11}$$

The inverse of the operator is simply the inverse square root, and has the expansion

$$\hat{\mathcal{O}}^{-1} = 1 - \frac{1}{8m^2c^2}\mathbf{p}^2 + \mathcal{O}\big(c^{-4}\big). \tag{17.12}$$

The usual procedure for applying a transformation $\hat{\mathcal{O}}$ would be to insert $\hat{\mathcal{O}}^{-1}\hat{\mathcal{O}}$ between the wave function and the Hamiltonian and premultiply by $\hat{\mathcal{O}}$. Using (17.5) this procedure gives

$$\hat{\mathcal{O}}\hat{\mathcal{H}}^{\mathrm{ESC}}\hat{\mathcal{O}}^{-1}\hat{\mathcal{O}}\psi^L = E\hat{\mathcal{O}}\psi^L = E\psi^N. \tag{17.13}$$

Unfortunately, this leads to some problems and inconsistencies, not the least of which is that the energy is still present in the Hamiltonian. These may be circumvented by premultiplying (17.13) with $\hat{\mathcal{O}}^{-2}$ to obtain

$$\hat{\mathcal{O}}^{-1}\hat{\mathcal{H}}^{\mathrm{ESC}}\hat{\mathcal{O}}^{-1}\psi^N = E\hat{\mathcal{O}}^{-2}\psi^N. \tag{17.14}$$

We write the Hamiltonian to $\mathcal{O}(c^{-2})$ as

$$\hat{\mathcal{H}}^{\mathrm{ESC}} = V + \hat{T} + \frac{1}{4m^2c^2}(\boldsymbol{\sigma}\cdot\mathbf{p})(V-E)(\boldsymbol{\sigma}\cdot\mathbf{p}) \tag{17.15}$$

from which, upon application of $\hat{\mathcal{O}}^{-1}$, we get

$$\begin{aligned}&\left(\hat{T} + V + \frac{1}{4m^2c^2}\Big((\boldsymbol{\sigma}\cdot\mathbf{p})V(\boldsymbol{\sigma}\cdot\mathbf{p}) - E\mathbf{p}^2 - \hat{T}\mathbf{p}^2 - \tfrac{1}{2}(\mathbf{p}^2V + V\mathbf{p}^2)\Big)\right)\psi^N \\ &\quad = \left(E - \frac{1}{4m^2c^2}E\mathbf{p}^2\right)\psi^N.\end{aligned} \tag{17.16}$$

Now, the term involving the energy on the left-hand side cancels with an identical term on the right, and the energy dependence is eliminated. The last term on the left can be expanded to

$$\tfrac{1}{2}\big(\mathbf{p}^2V + V\mathbf{p}^2\big) = -\hbar^2\big(\tfrac{1}{2}(\nabla^2V) + (\nabla V)\cdot\nabla + V\nabla^2\big) \tag{17.17}$$

and the first term in the inner parentheses expands to

$$(\boldsymbol{\sigma}\cdot\mathbf{p})V(\boldsymbol{\sigma}\cdot\mathbf{p}) = -\hbar^2\big((\nabla V)\cdot\nabla + V\nabla^2\big) + \hbar\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{p}. \tag{17.18}$$

With the cancellations from these two expressions, we get the Pauli Hamiltonian

$$\hat{\mathcal{H}}^{\text{Pauli}} = \hat{T} + V - \frac{1}{8m^3c^2}\mathbf{p}^4 + \frac{\hbar^2}{8m^2c^2}(\nabla^2 V) + \frac{\hbar}{4m^2c^2}\boldsymbol{\sigma} \cdot (\nabla V) \times \mathbf{p}, \tag{17.19}$$

which we met in the previous chapter. For a point nuclear potential in Hartree atomic units,

$$\begin{aligned} V &= -\frac{Z}{r} \\ \nabla V &= \frac{Z}{r^3}\mathbf{r} \\ \nabla^2 V &= 4\pi Z \delta(\mathbf{r}) \end{aligned} \tag{17.20}$$

and noting that $\hbar\boldsymbol{\sigma} = 2\mathbf{s}$ and $\boldsymbol{\ell} = \mathbf{r} \times \mathbf{p}$, the Pauli Hamiltonian is

$$\hat{\mathcal{H}}^{\text{Pauli}} = \hat{T} + V - \frac{1}{8m^3c^2}\mathbf{p}^4 + \frac{\pi\hbar^2 Z}{2m^2c^2}\delta(\mathbf{r}) + \frac{Z}{2m^2c^2r^3}\mathbf{s} \cdot \boldsymbol{\ell}. \tag{17.21}$$

The first of the relativistic correction terms is called the *mass–velocity* operator. If we expand the square root operator in the classical relativistic Hamiltonian for a free particle, we find

$$mc^2\sqrt{1 + \left(\frac{\mathbf{p}}{mc}\right)^2} = mc^2 + \frac{\mathbf{p}^2}{2m} - \frac{\mathbf{p}^4}{8m^3c^2} + \cdots . \tag{17.22}$$

The mass–velocity term is therefore the lowest-order term from the relativistic Hamiltonian that comes from the variation of the mass with the velocity. The second relativistic term in the Pauli Hamiltonian is called the *Darwin* operator, and has no classical analogue. Due to the presence of the Dirac delta function, the only contributions for an atom come from s functions. The third term is the *spin–orbit* term, resulting from the interaction of the spin of the electron with its orbital angular momentum around the nucleus. This operator is identical to the spin–orbit operator of the modified Dirac equation.

It has been stressed in many places that this operator should *only* be used in perturbation theory—even though we have made no assumptions about the nature of the wave function and its relation to the nonrelativistic wave function, except that it must be normalized. The reason is not hard to find: it is because of the mass–velocity operator. If we were to try to minimize the energy of a system with the Pauli Hamiltonian, there would be a region close to the nucleus where the mass–velocity term would dominate over the kinetic energy term. For a classical particle with zero energy, this would happen for $r < (Z/2mc^2)a_0$. Since the kinetic energy is positive in this region, and the mass–velocity term is negative, the effect would be to cause the wave function to collapse, both spatially toward the nucleus and in energy toward $-\infty$. The Darwin term, being positive, provides a balance to the mass–velocity term, but the different magnitudes of the two terms still produces a collapse.

We may also view this from the point of view of the expansion of the classical relativistic energy. For any atom there will always be a region of space in which $|\mathbf{p}| > mc$, and therefore the expansion of the square root will not be convergent in this region. Some have drawn the conclusion from this problem that one should not use the Pauli Hamiltonian at all (Farazdel and Smith 1986). This point of view has been disputed, and from the many calculations performed with the mass–velocity and Darwin operators—albeit as a first-order perturbation (second order in $1/c$)—it would appear that the objection to the use of the operator solely on the grounds of the expansion is not justified. What is important is that, even though the operator expansion itself is invalid in a small region of space around the nuclei, the expectation values of the operators appear to be well-behaved because of the form of the wave function in that region and because of the very small volume of that region. The wave function is nonrelativistic: it does not have the relativistic contraction and so is smaller near the nucleus than the Dirac wave function. The small size of the coefficients of the basis functions near the nucleus serve as a cutoff to the large relativistic operators.

Nonetheless, care needs to be exercised in calculations of the one-electron scalar relativistic correction by perturbation theory using these operators. It has been found that the magnitude of the correction is quite sensitive to the contraction of the basis set. This follows from the fact that the operators weight the region near the nucleus. If a basis function is taken out of the contraction the nonrelativistic wave function may not change very much, but a small change in the coefficient of the core function can have a big effect on the relativistic correction.

17.2 The Breit–Pauli Hamiltonian

The success of the first-order Pauli results for one-electron systems suggests that it is worth considering the two-electron terms also—and in fact the two-electron terms are necessary to obtain accurate fine-structure splittings for light elements. We may derive the necessary expressions by setting $\phi^L = \psi^L$ in the modified Dirac equation and introducing the renormalization term. In doing this, the operators in the modified Dirac equation that operated on different functions now operate on the same function and may be collected. An alternative is to apply the transformation derived in the previous chapter. We will only consider terms of order $1/c^2$: thus, the terms from the $(SS|SS)$ integrals are neglected.

From the Coulomb interaction we get the operator

$$\begin{aligned}\hat{g}^{\mathrm{C}} = \frac{1}{2}\sum_{i\neq j}\Bigg\{\frac{1}{r_{ij}} - \frac{1}{8m^2c^2}\left[\mathbf{p}_i^2\frac{1}{r_{ij}} + \frac{1}{r_{ij}}\mathbf{p}_i^2 + \mathbf{p}_j^2\frac{1}{r_{ij}} + \frac{1}{r_{ij}}\mathbf{p}_j^2\right] \\ + \frac{1}{4m^2c^2}\left[(\boldsymbol{\sigma}_i\cdot\mathbf{p}_i)\frac{1}{r_{ij}}(\boldsymbol{\sigma}_i\cdot\mathbf{p}_i) + (\boldsymbol{\sigma}_j\cdot\mathbf{p}_j)\frac{1}{r_{ij}}(\boldsymbol{\sigma}_j\cdot\mathbf{p}_j)\right]\Bigg\}.\end{aligned} \tag{17.23}$$

The terms in the first square bracket come from the renormalization operator. The terms in the second square bracket are not renormalized because they are already of order $1/c^2$. The renormalization terms cancel with the terms in the second square bracket, in exactly the same way as for the one-electron operators, to give a two-electron Darwin term and a two-electron spin–orbit term:

$$\begin{aligned}\hat{g}^{\mathrm{C}} = \frac{1}{2}\sum_{i\neq j}\Bigg\{\frac{1}{r_{ij}} + \frac{\hbar^2}{8m^2c^2}\left[(\nabla_i^2\frac{1}{r_{ij}}) + (\nabla_j^2\frac{1}{r_{ij}})\right] \\ +\frac{\hbar}{4m^2c^2}\left[\boldsymbol{\sigma}_i\cdot(\nabla_i\frac{1}{r_{ij}})\times\mathbf{p}_i + \boldsymbol{\sigma}_j\cdot(\nabla_j\frac{1}{r_{ij}})\times\mathbf{p}_j\right]\Bigg\}.\end{aligned} \tag{17.24}$$

Since the operators are symmetric in i and j we may reindex and combine the terms, and substitute $\nabla_i^2(1/r_{ij}) = -4\pi\delta(r_{ij})$ and $\nabla_i(1/r_{ij}) = -\mathbf{r}_{ij}/r_{ij}^3$ to give the operator

$$\hat{g}^{\mathrm{C}} = \frac{1}{2}\sum_{i\neq j}\left[\frac{1}{r_{ij}} - \frac{\pi\hbar^2}{m^2c^2}\delta(\mathbf{r}_{ij}) - \frac{\hbar}{2m^2c^2r_{ij}^3}\boldsymbol{\sigma}_i\cdot\mathbf{r}_{ij}\times\mathbf{p}_i\right]. \tag{17.25}$$

As for the one-electron term we may write $\boldsymbol{\ell}_{ij} = \mathbf{r}_{ij}\times\mathbf{p}_i$ and $\hbar\boldsymbol{\sigma} = 2\mathbf{s}$, and hence

$$\hat{g}^{\mathrm{C}} = \frac{1}{2}\sum_{i\neq j}\left[\frac{1}{r_{ij}} - \frac{\pi\hbar^2}{m^2c^2}\delta(\mathbf{r}_{ij}) - \frac{1}{m^2c^2r_{ij}^3}\mathbf{s}_i\cdot\boldsymbol{\ell}_{ij}\right]. \tag{17.26}$$

The spin–orbit term involves the interaction of the spin of the electron with its own orbital angular momentum around the other electrons, and is often called the spin–own–orbit interaction or spin–same–orbit interaction.

We now turn to the Gaunt interaction, and use the terms from the modified Dirac representation in (15.54) to derive the Breit–Pauli operators. These terms need no renormalization, because they are all of order $1/c^2$. The three classes of operators defined in (15.54) are considered in turn.

The first class contains the spin-free operators,

$$\begin{aligned}&\mathbf{p}_i\cdot\mathbf{p}_j\frac{1}{r_{ij}} + \mathbf{p}_i\frac{1}{r_{ij}}\cdot\mathbf{p}_j + \mathbf{p}_j\frac{1}{r_{ij}}\cdot\mathbf{p}_i + \frac{1}{r_{ij}}\mathbf{p}_i\cdot\mathbf{p}_j \\ &= \frac{4}{r_{ij}}\mathbf{p}_i\cdot\mathbf{p}_j + \frac{2i\hbar}{r_{ij}^3}\mathbf{r}_{ij}\cdot\mathbf{p}_j - \frac{2i\hbar}{r_{ij}^3}\mathbf{r}_{ij}\cdot\mathbf{p}_i - \hbar^2\big(\nabla_i\cdot\nabla_j\frac{1}{r_{ij}}\big).\end{aligned} \tag{17.27}$$

Since $\nabla_i(1/r_{ij}) = -\nabla_j(1/r_{ij})$, the term in parentheses may be rewritten as $-\nabla_i^2(1/r_{ij})$, which yields a delta function. Introducing the factor of $-1/4m^2c^2$ gives

$$\hat{g}^{\mathrm{G,sf}} = -\frac{1}{2m^2c^2}\sum_{i\neq j}\left[\frac{1}{r_{ij}}\mathbf{p}_i\cdot\mathbf{p}_j - \frac{i\hbar}{2r_{ij}^3}\mathbf{r}_{ij}\cdot\mathbf{p}_i + \frac{i\hbar}{2r_{ij}^3}\mathbf{r}_{ij}\cdot\mathbf{p}_j - \pi\hbar^2\delta(r_{ij})\right]. \tag{17.28}$$

These operators will be considered again later, when all the spin-free terms will be accumulated. We note here only that the Darwin term from the Gaunt interaction cancels the Darwin term from the Coulomb interaction, so that the Coulomb–Gaunt interaction has no two-electron Darwin term, but it does nevertheless have a scalar relativistic interaction.

In the second class, which contains the spin–orbit operators, the only terms that survive are those in which the momentum operator operates on $1/r_{ij}$:

$$\begin{aligned} & i(\boldsymbol{\sigma}_i - \boldsymbol{\sigma}_j) \cdot (\mathbf{p}_i \times \mathbf{p}_j)\frac{1}{r_{ij}} + i(\boldsymbol{\sigma}_i + \boldsymbol{\sigma}_j) \cdot \left(\mathbf{p}_i \frac{1}{r_{ij}} \times \mathbf{p}_j\right) \\ & \quad + i(\boldsymbol{\sigma}_i + \boldsymbol{\sigma}_j) \cdot \left(\mathbf{p}_j \frac{1}{r_{ij}} \times \mathbf{p}_i\right) - i\frac{1}{r_{ij}}(\boldsymbol{\sigma}_i - \boldsymbol{\sigma}_j) \cdot (\mathbf{p}_i \times \mathbf{p}_j) \\ & = \hbar(\boldsymbol{\sigma}_i - \boldsymbol{\sigma}_j) \cdot \left(\mathbf{p}_i \times \nabla_j \frac{1}{r_{ij}}\right) + \hbar(\boldsymbol{\sigma}_i - \boldsymbol{\sigma}_j) \cdot \left(\nabla_i \frac{1}{r_{ij}} \times \mathbf{p}_j\right) \\ & \quad + \hbar(\boldsymbol{\sigma}_i + \boldsymbol{\sigma}_j) \cdot \left(\nabla_i \frac{1}{r_{ij}} \times \mathbf{p}_j\right) + \hbar(\boldsymbol{\sigma}_i + \boldsymbol{\sigma}_j) \cdot \left(\nabla_j \frac{1}{r_{ij}} \times \mathbf{p}_i\right) \\ & = \frac{2}{r_{ij}^3}\hbar\boldsymbol{\sigma}_j \cdot (\mathbf{r}_{ij} \times \mathbf{p}_i) - \frac{2}{r_{ij}^3}\hbar\boldsymbol{\sigma}_i \cdot (\mathbf{r}_{ij} \times \mathbf{p}_j). \end{aligned} \tag{17.29}$$

Introducing the factor of $-1/4m^2c^2$ and combining the two terms with index interchange gives

$$\hat{g}^{\mathrm{G,so}} = -\frac{1}{m^2c^2}\sum_{i \neq j} \frac{1}{r_{ij}^3}\hbar\boldsymbol{\sigma}_j \cdot (\mathbf{r}_{ij} \times \mathbf{p}_i). \tag{17.30}$$

This term involves the interaction of the spin of electron j with the orbital angular momentum of electron i around electron j, and is called the *spin–other–orbit* interaction.

The third class comprises the spin–spin operators,

$$\begin{aligned} & -(\boldsymbol{\sigma}_i \times \mathbf{p}_i) \cdot (\boldsymbol{\sigma}_j \times \mathbf{p}_j)\frac{1}{r_{ij}} + \left(\boldsymbol{\sigma}_i \times \mathbf{p}_i \frac{1}{r_{ij}}\right) \cdot (\boldsymbol{\sigma}_j \times \mathbf{p}_j) \\ & + \left(\boldsymbol{\sigma}_j \times \mathbf{p}_j \frac{1}{r_{ij}}\right) \cdot (\boldsymbol{\sigma}_i \times \mathbf{p}_i) - \frac{1}{r_{ij}}(\boldsymbol{\sigma}_i \times \mathbf{p}_i) \cdot (\boldsymbol{\sigma}_j \times \mathbf{p}_j). \end{aligned} \tag{17.31}$$

As for the second class, there is a cancellation of many of the terms. In fact the only one that survives is the contribution from the first term in which both momentum operators operate on $1/r_{ij}$. Expansion of the scalar quadruple product gives

$$-(\boldsymbol{\sigma}_i \times \mathbf{p}_i) \cdot (\boldsymbol{\sigma}_j \times \mathbf{p}_j)\frac{1}{r_{ij}} = \hbar^2(\boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j)\left(\nabla_i \cdot \nabla_j \frac{1}{r_{ij}}\right) - \hbar^2\left(\boldsymbol{\sigma}_i \cdot \nabla_j\right)\left(\boldsymbol{\sigma}_j \cdot \nabla_i \frac{1}{r_{ij}}\right). \tag{17.32}$$

The spatial part of the first term can be rewritten in terms of a delta function, as was done for the scalar operators. In the second term, we need the individual derivatives of $1/r_{ij}$ with respect to the coordinates. Provided $r \neq 0$, we have

$$\frac{\partial^2}{\partial q_i \partial q'_j} \frac{1}{r_{ij}} = \delta_{q,q'} \frac{1}{r_{ij}^3} - \frac{3 q_{ij} q'_{ij}}{r_{ij}^5} \tag{17.33}$$

where q and q' are any of x, y, and z. For $r = 0$, however, we have

$$\sum_q \frac{\partial^2}{\partial q_i \partial q_j} \frac{1}{r_{ij}} = 4\pi \delta(\mathbf{r}_{ij}). \tag{17.34}$$

We may deduce from the fact that the delta function must be isotropic that

$$\frac{\partial^2}{\partial q_i \partial q_j} \frac{1}{r_{ij}} = \frac{4\pi}{3} \delta(\mathbf{r}_{ij}). \tag{17.35}$$

Combining these two gives an expression for all r,

$$\frac{\partial^2}{\partial q_i \partial q'_j} \frac{1}{r_{ij}} = \delta_{q,q'} \left[\frac{1}{r_{ij}^3} + \frac{4\pi}{3} \delta(\mathbf{r}_{ij}) \right] - \frac{3 q_{ij} q'_{ij}}{r_{ij}^5}. \tag{17.36}$$

The spin–spin operator may now be written as

$$\hat{g}^{\mathrm{G,ss}} = \frac{\hbar^2}{8m^2c^2} \sum_{i \neq j} \left[\frac{\boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j}{r_{ij}^3} - \frac{3(\boldsymbol{\sigma}_i \cdot \mathbf{r}_{ij})(\boldsymbol{\sigma}_j \cdot \mathbf{r}_{ij})}{r_{ij}^5} - \frac{8\pi}{3} (\boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j) \delta(\mathbf{r}_{ij}) \right]. \tag{17.37}$$

The delta function term is known as the *Fermi contact* interaction, the remaining terms comprise the *spin–spin dipole* interaction.

The contributions to the two-electron operator from the gauge term of the Breit interaction may be developed in the same manner, using the representation for the modified Dirac operator. The derivation is more complicated, and details may be found in appendix I. The only term that contributes to $\mathcal{O}(c^{-2})$ is a spin-free term:

$$\begin{aligned} \hat{g}^{\mathrm{gauge,sf}} = \frac{1}{2m^2c^2} \sum_{i \neq j} \Bigg[& \frac{1}{2r_{ij}} (\mathbf{p}_i \cdot \mathbf{p}_j) - \frac{1}{2r_{ij}^3} (\mathbf{r}_{ij} (\mathbf{r}_{ij} \cdot \mathbf{p}_j) \cdot \mathbf{p}_i) \\ & - \frac{i\hbar}{2r_{ij}^3} (\mathbf{r}_{ij} \cdot \mathbf{p}_i) + \frac{i\hbar}{2r_{ij}^3} (\mathbf{r}_{ij} \cdot \mathbf{p}_j) - \pi \hbar^2 \delta(\mathbf{r}_{ij}) \Bigg]. \end{aligned} \tag{17.38}$$

The first two terms combine to give the *orbit–orbit* interaction, mentioned at the end of section 15.4, but they have been written in the present form so that the cancellation with the terms in the Gaunt interaction is more obvious.

We may now combine all the spin-free contributions from the Coulomb, Gaunt, and gauge terms. As noted before, the delta functions from the Coulomb and Gaunt interactions cancel—or alternatively, the delta functions from the Gaunt and the gauge terms cancel. The third and fourth terms from the expression for the gauge term cancel with the corresponding terms from the Gaunt interaction, and the first term partially cancels, leaving

$$\hat{g}^{\mathrm{sf}} = \frac{1}{2}\sum_{i\neq j}\left[\frac{1}{r_{ij}} - \frac{1}{2m^2c^2}\left(\frac{1}{r_{ij}}\mathbf{p}_i\cdot\mathbf{p}_j + \frac{1}{r_{ij}^3}(\mathbf{r}_{ij}(\mathbf{r}_{ij}\cdot\mathbf{p}_j)\cdot\mathbf{p}_i)\right) - \frac{\pi\hbar^2}{m^2c^2}\delta(\mathbf{r}_{ij})\right]. \tag{17.39}$$

Note that we would have a very different operator if we did not include the contributions from the gauge term. However, this says little about the magnitude of the gauge term, which is usually small.

Combining all the terms, the final form of the Breit–Pauli Hamiltonian for a point nucleus can be written as follows:

$$\begin{aligned}\hat{\mathcal{H}}^{\mathrm{BP}} &= \sum_i\left[\frac{\mathbf{p}_i^2}{2m} - \frac{Z}{r_i} - \frac{\mathbf{p}_i^4}{8m^3c^2} + \frac{\pi\hbar^2 Z}{2m^2c^2}\delta(\mathbf{r}_i) + \frac{Z}{2m^2c^2r^3}\mathbf{s}_i\cdot\boldsymbol{\ell}_i\right]\\ &+\frac{1}{2}\sum_{i\neq j}\left[\frac{1}{r_{ij}} - \frac{1}{2m^2c^2}\left(\frac{1}{r_{ij}}\mathbf{p}_i\cdot\mathbf{p}_j + \frac{1}{r_{ij}^3}(\mathbf{r}_{ij}(\mathbf{r}_{ij}\cdot\mathbf{p}_j)\cdot\mathbf{p}_i)\right) - \frac{\pi\hbar^2}{m^2c^2}\delta(\mathbf{r}_{ij})\right.\\ &\qquad - \frac{1}{m^2c^2r_{ij}^3}(\mathbf{s}_i + 2\mathbf{s}_j)\cdot\mathbf{r}_{ij}\times\mathbf{p}_i - \frac{8\pi}{3m^2c^2}(\mathbf{s}_i\cdot\mathbf{s}_j)\delta(\mathbf{r}_{ij})\\ &\qquad\left. + \frac{1}{m^2c^2}\left(\frac{(\mathbf{s}_i\cdot\mathbf{s}_j)}{r_{ij}^3} - \frac{3(\mathbf{s}_i\cdot\mathbf{r}_{ij})(\mathbf{s}_j\cdot\mathbf{r}_{ij})}{r_{ij}^5}\right)\right].\end{aligned} \tag{17.40}$$

This expression is for a single atom: for a molecule, we would need to sum the terms involving the nuclear potential over the nuclei in the molecule. These are the nuclear potential itself, and the one-electron Darwin and spin–orbit terms.

17.3 Perturbative Treatment of the Lamb Shift

In chapter 5 we mentioned the Lamb shift in passing and the difficulty of calculating it due to the renormalizations required. In fact, there is a well-developed perturbation theory of the Lamb shift in the same framework as Breit–Pauli theory. We do not propose to derive the expressions here, for which the reader is referred to Bethe and Salpeter (1957). Instead, we report the results for the lowest-order terms, which turn out to be expressible as corrections to the Darwin and spin–orbit one-electron operators. The combined operators may be written

$$\hat{\mathcal{H}}^{\mathrm{Pauli+Lamb}} = \hat{T} + V - \frac{1}{8m^3c^2}\mathbf{p}^4 - \frac{\hbar^2}{8m^2c^2}(1+f_1)(\nabla^2 V) + \frac{\hbar}{4m^2c^2}(1+f_2)\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{p} \tag{17.41}$$

where f_1 and f_2 are quantities given by

$$f_1 = \frac{8}{3\pi c}\left[\ln\frac{mc^2}{2K_0} + \frac{19}{30}\right], \qquad f_2 = \frac{1}{\pi c}. \tag{17.42}$$

The Bethe logarithm, $\ln K_0$, is a number that is independent of Z but depends on n, the principal quantum number of the atomic electron. This would be awkward for quantum chemical calculations, but since the most important differential effects for chemistry would come from the valence shell, a good approximation is to use the value for the valence shell (Pyykkö et al. 2001). Bethe logarithms have been tabulated by Drake and Swainson (1990). The correction to the spin–orbit interaction is small, being less than 0.25%, but the correction to the Darwin term varies from nearly 10% for hydrogen down to 1% for heavier elements. There is also a perturbative correction to the two-electron Darwin term, which we write as

$$\hat{g}_{ij}^{\text{Darwin+Lamb}} = -\frac{\pi}{m^2c^2}\left[1 + \frac{14}{3\pi c}\ln c\right]\delta(\mathbf{r}_{ij}). \tag{17.43}$$

The correction is a little more than 5%.

Unfortunately, the one-electron formulas cannot be used for very high Z because the perturbation expression diverges from the real value reasonably soon. A safe limit is probably Ar, $Z = 18$. A semiempirical extension suggested by Pyykkö et al. (2001) is to use the fully relativistic values of the hydrogenic self-energy term, which are given in a number of places as a tabulation of factors $F_{n\kappa}(Z/c)$, for which the self-energy is given by

$$E_{n\kappa}^{\text{SE}} = \frac{Z^4}{\pi n^3 c^3} F_{n\kappa}(Z/c). \tag{17.44}$$

The multiplying factor is clearly related to the expectation of the Dirac delta function or of the spin–orbit operator, both of which scale as Z^3/n^3, and hence we can derive a correction factor for the Darwin operator of

$$f_1 = \frac{2}{\pi c}F_{ns}(Z/c) - \frac{8}{15\pi c} \tag{17.45}$$

and a correction to the spin–orbit term for p electrons of

$$f_2 = \frac{4}{\pi c}\left[F_{np_{3/2}}(Z/c) - F_{np_{1/2}}(Z/c)\right]. \tag{17.46}$$

A more elaborate treatment of the Lamb shift for general polyatomic molecules is not at present available.

17.4 Multiple Perturbation Theory for Many-Electron Systems and Properties

In the previous two sections, we have presented the Breit–Pauli perturbation Hamiltonian for one- and two-electron relativistic corrections of order $1/c^2$ to the nonrelativistic Hamiltonian. But there is a problem for many-electron systems. For the perturbation theory to be valid, the reference wave function must be an eigenfunction of the zeroth-order Hamiltonian. If we take this to be the nonrelativistic Hamiltonian and the perturbation parameter to be $1/c$, we do not have the exact solutions of the zeroth-order equation.

To overcome this problem, we can use double perturbation theory. The zeroth-order Hamiltonian is then normally some one-particle operator, such as the Fock operator, and the two perturbations are correlation and relativity. Formally, we write the Hamiltonian as

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \lambda \hat{\mathcal{H}}_{10} + \mu^2 \hat{\mathcal{H}}_{02} \tag{17.47}$$

where λ is the perturbation parameter for correlation and $\mu = 1/c$ is the perturbation parameter for relativity. The expansion in μ is an expansion in even powers. Here, we will call the lowest-order term, which is $\mathcal{O}(c^{-2})$, the "first-order" correction, to make the correspondence with the correlation perturbation clear.

The correlation and relativistic Hamiltonians can be written as

$$\begin{aligned} \hat{\mathcal{H}}_{10} &= \hat{\mathcal{H}}^{\mathrm{NR}} - \hat{\mathcal{F}}^{\mathrm{NR}}, \\ \hat{\mathcal{H}}_{02} &= \hat{\mathcal{H}}^{\mathrm{rel}}. \end{aligned} \tag{17.48}$$

In principle, we should have an infinite series in μ because there are relativistic corrections to the Hamiltonian at all orders in μ, but since we are developing a method appropriate for the Breit–Pauli approximation, we need only consider the lowest-order terms. Technically, we ought to write $\hat{\mathcal{H}}_0$ as $\hat{\mathcal{H}}_{00}$ to indicate that it is of zeroth order in both perturbations, but we will omit the second zero.

The energy and the wave function are expanded in infinite series in both perturbations,

$$\begin{aligned} E &= \sum_{k,m=0}^{\infty} \lambda^k \mu^m E^{(km)}, \\ \Psi &= \sum_{k,m=0}^{\infty} \lambda^k \mu^m \Psi^{(km)}, \end{aligned} \tag{17.49}$$

and m is even. We use intermediate normalization,

$$\langle \Psi^{(00)} | \Psi \rangle = 1, \qquad \Rightarrow \qquad \langle \Psi^{(00)} | \Psi^{(00)} \rangle = 1, \qquad \langle \Psi^{(00)} | \Psi^{(km)} \rangle = 0. \tag{17.50}$$

The lowest-order correction in λ simply yields the Hartree–Fock energy (for a closed-shell reference),

$$E^{\mathrm{HF}} = E^{(00)} + E^{(10)}. \tag{17.51}$$

The lowest-order correction in μ is the expectation of the relativistic Hamiltonian over the reference wave function,

$$E^{(02)} = \langle \Psi^{(00)} \,|\, \hat{\mathcal{H}}^{\mathrm{rel}} \,|\, \Psi^{(00)} \rangle. \tag{17.52}$$

This is the correction to the Hartree–Fock energy. Obviously, only the terms that are diagonal in the spin and are spatially totally symmetric will contribute. These terms are the scalar relativistic corrections: the mass–velocity term, the one- and two-electron Darwin terms, and the orbit–orbit term. Other terms such as the spin–spin term and the z component of the spin–orbit interaction contribute for open-shell systems, where the spin is nonzero.

However, this is not the entire correction of order c^{-2}. There are mixed terms of $\mathcal{O}(c^{-2})$ that involve the correlation perturbation. The lowest order of these is

$$E^{(12)} = \langle \Psi^{(00)} \,|\, \hat{\mathcal{H}}_{02} \,|\, \Psi^{(10)} \rangle + \langle \Psi^{(00)} \,|\, \hat{\mathcal{H}}_{10} \,|\, \Psi^{(02)} \rangle. \tag{17.53}$$

If we expand the lowest-order perturbed wave functions for both perturbations in a set of determinants Φ_k we can write this correction as

$$E^{(12)} = 2\,\mathrm{Re} \sum_{k \neq 0} \frac{\langle \Phi_0 \,|\, \hat{\mathcal{H}}_{02} \,|\, \Phi_k \rangle \langle \Phi_k \,|\, \hat{\mathcal{H}}_{10} \,|\, \Phi_0 \rangle}{E_0 - E_k}. \tag{17.54}$$

There are several observations we may make about this correction. First, as for the pure relativistic correction, it vanishes for all except the scalar relativistic terms because the correlation perturbation is totally symmetric and hence the determinants Φ_k must have the same symmetry as the ground state.

Second, if only the one-electron mass–velocity and Darwin terms are included in the relativistic perturbation—a common choice since they are relatively easy to evaluate—the relativistic first-order wave function (which is $\mathcal{O}(c^{-2})$) must consist only of the Brillouin single excitations. But these have a zero matrix element of the correlation perturbation with the ground state, and hence the correction is zero. It would then be necessary to include the two-electron terms to get a correction at this order to the ground state. This is true for both closed-shell and open-shell systems.

An alternative way of analyzing this term is to consider that the first-order correlated wave function for a closed-shell system consists entirely of doubly excited configurations, and therefore the matrix elements of the one-electron relativistic operators between the reference and these configurations will be zero. The contribution of these operators will therefore only enter through the second-order wave function in either perturbation. We should not assume that this contribution will be small, however: the relativistic perturbation is large in the core region and could produce significant effects even at higher order.

Third, since this term mixes relativity and correlation, a nonzero value means that relativity and correlation are not additive. If the term is small, the degree of nonadditivity will be small, and it may be negligible for the purposes of the calculation in hand. The fact that the lowest cross-term is zero for the one-electron scalar relativistic corrections implies that at least for light elements the neglect of the cross-term is reasonably well justified.

The spin-dependent relativistic effects only enter for a closed-shell system at $\mathcal{O}(c^{-4})$ because the operators change the spatial or the spin symmetry, or both. The treatment of the spin–orbit interaction will be deferred to chapter 21, where the symmetry aspects can be treated in fuller detail.

We now turn to the treatment of properties, for which we introduce a third perturbation, along with an associated parameter that is related to the field strength of the perturbation. As we saw in the previous chapter, the perturbation may also have a relativistic correction. Working only at the lowest order again, the Hamiltonian can be written

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \lambda\hat{\mathcal{H}}_{100} + \mu^2\hat{\mathcal{H}}_{020} + \eta\hat{\mathcal{H}}_{001} + \mu^2\eta\hat{\mathcal{H}}_{021}. \tag{17.55}$$

Expanding the energy and wave function in the three parameters,

$$\begin{aligned} E &= \sum_{k,m,i=0}^{\infty} \lambda^k \mu^m \eta^i E^{(kmi)} \\ \Psi &= \sum_{k,m,i=0}^{\infty} \lambda^k \mu^m \eta^i \Psi^{(kmi)}, \end{aligned} \tag{17.56}$$

with m even, and using intermediate normalization, we differentiate with respect to η at $\eta = 0$ to obtain the first-order property,

$$\left.\frac{\partial E}{\partial \eta}\right|_{\eta=0} = \sum_{k,m=0}^{\infty} \lambda^k \mu^m E^{(km1)}. \tag{17.57}$$

We are interested in the relativistic correction to the property, of which the lowest-order term is $E^{(021)}$:

$$\begin{aligned} E^{(021)} &= \langle \Psi^{(000)} | \hat{\mathcal{H}}_{021} | \Psi^{(000)} \rangle + \langle \Psi^{(000)} | \hat{\mathcal{H}}_{001} | \Psi^{(020)} \rangle + \langle \Psi^{(000)} | \hat{\mathcal{H}}_{020} | \Psi^{(001)} \rangle \\ &= \langle \Phi_0 | \hat{\mathcal{H}}_{021} | \Phi_0 \rangle + 2\,\mathrm{Re} \sum_{k\neq 0} \frac{\langle \Phi_0 | \hat{\mathcal{H}}_{001} | \Phi_k \rangle \langle \Phi_k | \hat{\mathcal{H}}_{020} | \Phi_0 \rangle}{E_0 - E_k}. \end{aligned} \tag{17.58}$$

Apart from the expectation of the relativistic correction to the property operator, we get a mixed expression, like the mixed relativistic–correlation correction above. The scalar relativistic correction to the property operator is zero for external electric fields but nonzero otherwise, as we saw in chapter 16. If the property is a first-order property, that is, $\langle \Psi^{(000)} | \hat{\mathcal{H}}_{001} | \Psi^{(000)} \rangle$ does not vanish, the mixed correction will also be nonzero because both the scalar relativistic operator and the property operator have the same symmetry.

The lowest-order mixed relativistic–correlation correction to the property is

$$\begin{aligned}
E^{(121)} &= \langle \Psi^{(000)} \,|\, \hat{\mathcal{H}}_{021} \,|\, \Psi^{(100)} \rangle + \langle \Psi^{(000)} \,|\, \hat{\mathcal{H}}_{001} \,|\, \Psi^{(120)} \rangle \\
&\quad + \langle \Psi^{(000)} \,|\, \hat{\mathcal{H}}_{020} \,|\, \Psi^{(101)} \rangle + \langle \Psi^{(000)} \,|\, \hat{\mathcal{H}}_{100} \,|\, \Psi^{(021)} \rangle \\
&= \sum_{k\neq 0} \frac{\langle \Phi_0 \,|\, \hat{\mathcal{H}}_{021} \,|\, \Phi_k \rangle \langle \Phi_k \,|\, \hat{\mathcal{H}}_{100} \,|\, \Phi_0 \rangle}{E_0 - E_k} \\
&\quad + 2\,\mathrm{Re} \sum_{k,m\neq 0} \frac{\langle \Phi_0 \,|\, \hat{\mathcal{H}}_{001} \,|\, \Phi_k \rangle \langle \Phi_k \,|\, \hat{\mathcal{H}}_{020} - E^{(020)} \,|\, \Phi_m \rangle \langle \Phi_m \,|\, \hat{\mathcal{H}}_{100} \,|\, \Phi_0 \rangle}{(E_0 - E_k)(E_0 - E_m)} \\
&\quad + 2\,\mathrm{Re} \sum_{k,m\neq 0} \frac{\langle \Phi_0 \,|\, \hat{\mathcal{H}}_{020} \,|\, \Phi_k \rangle \langle \Phi_k \,|\, \hat{\mathcal{H}}_{100} - E^{(100)} \,|\, \Phi_m \rangle \langle \Phi_m \,|\, \hat{\mathcal{H}}_{001} \,|\, \Phi_0 \rangle}{(E_0 - E_k)(E_0 - E_m)} \\
&\quad + 2\,\mathrm{Re} \sum_{k,m\neq 0} \frac{\langle \Phi_0 \,|\, \hat{\mathcal{H}}_{100} \,|\, \Phi_k \rangle \langle \Phi_k \,|\, \hat{\mathcal{H}}_{001} - E^{(001)} \,|\, \Phi_m \rangle \langle \Phi_m \,|\, \hat{\mathcal{H}}_{020} \,|\, \Phi_0 \rangle}{(E_0 - E_k)(E_0 - E_m)} .
\end{aligned} \tag{17.59}$$

The first sum is zero for a closed-shell system because $\hat{\mathcal{H}}_{021}$ is a one-electron operator and the correlation perturbation only connects the reference state to doubly excited configurations. None of the remaining sums is necessarily zero. The mixed relativistic–correlation correction to a property can therefore be nonzero even if the corresponding correction to the energy is zero.

Spin-dependent corrections to an electric property for a closed-shell system only enter at $\mathcal{O}(c^{-4})$, but of course they may enter at $\mathcal{O}(c^{-2})$ for an open-shell system. For magnetic properties, relativistic corrections to the interaction with a magnetic field can enter at $\mathcal{O}(c^{-2})$ because both the perturbation operator and the relativistic correction operator have a spin dependence.

The multiple perturbation approach to relativistic corrections has been developed and applied extensively by Kellö and Sadlej et al. (1990, 1998), from whose papers the reader may gain additional information.

17.5 Direct Perturbation Theory

If we proceed from the elimination of the small component, the expansion of the Dirac equation in powers of $1/c^2$ to give the Pauli Hamiltonian produces operators for which the integrals are readily evaluated but are highly singular. Moreover, there are doubts about the validity of the expansion in the region close to the nucleus. Is there an alternative that would give simple integrals but still have a valid expansion? Such an alternative has been developed by Rutkowski and coworkers (1986, 1988, 1992) and Kutzelnigg (1989, 1990), and is termed *direct perturbation theory* (DPT).

We may consider making the perturbation partitioning in the Dirac equation itself, *without* eliminating the small component, but first we have to write the Dirac equation in a form involving $1/c$ rather than c. This is simply achieved by redefining the small component, much as we did in the modified Dirac equation. Here we make the replacement

$$\phi^S = c\psi^S, \tag{17.60}$$

which may be regarded as a transformation on the original Dirac wave function,

$$\begin{pmatrix}\psi^L \\ \psi^S\end{pmatrix} = \begin{pmatrix}\mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & 1/c\mathbf{I}_2\end{pmatrix}\begin{pmatrix}\psi^L \\ \phi^S\end{pmatrix}. \tag{17.61}$$

Applying the transformation to the Dirac Hamiltonian, we may now partition it into a zeroth-order term and a perturbation of $\mathcal{O}(c^{-2})$,

$$\hat{\mathcal{H}}'_D = \begin{pmatrix}V\mathbf{I}_2 & (\boldsymbol{\sigma}\cdot\mathbf{p}) \\ (\boldsymbol{\sigma}\cdot\mathbf{p}) & -2m\mathbf{I}_2\end{pmatrix} + \frac{1}{c^2}\begin{pmatrix}\mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & V\mathbf{I}_2\end{pmatrix} = \hat{\mathcal{H}}_0 + \frac{1}{c^2}\hat{\mathcal{H}}_2. \tag{17.62}$$

The transformed metric may similarly be partitioned into a zeroth-order term and a perturbation,

$$\hat{\mathcal{G}} = \begin{pmatrix}\mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{0}_2\end{pmatrix} + \frac{1}{c^2}\begin{pmatrix}\mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{I}_2\end{pmatrix} = \hat{\mathcal{G}}_0 + \frac{1}{c^2}\hat{\mathcal{G}}_2. \tag{17.63}$$

Here we are explicitly taking into account the fact that the metric has a relativistic correction, a feature that in the Pauli approximation was avoided with the consequence of the appearance of pathological operators. The metric in fact is a combination of projectors onto the large and small components. Using the definitions from (15.40),

$$\hat{\mathcal{P}}^+ = (\mathbf{I}_4 + \beta)/2 = \begin{pmatrix}\mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{0}_2\end{pmatrix}, \qquad \hat{\mathcal{P}}^- = (\mathbf{I}_4 - \beta)/2 = \begin{pmatrix}\mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{I}_2\end{pmatrix}, \tag{17.64}$$

we see that the metric is simply

$$\hat{\mathcal{G}} = \hat{\mathcal{P}}^+ + \frac{1}{c^2}\hat{\mathcal{P}}^-. \tag{17.65}$$

Using these projection operators, the perturbation $\hat{\mathcal{H}}_2$ may be written in the form

$$\hat{\mathcal{H}}_2 = \hat{\mathcal{P}}^- V \hat{\mathcal{P}}^- = \hat{\mathcal{P}}^- V. \tag{17.66}$$

The zeroth-order Hamiltonian is the Lévy-Leblond Hamiltonian (Lévy-Leblond 1967, 1970). For small components chosen to obey

$$\phi^S = (\boldsymbol{\sigma}\cdot\mathbf{p})\psi^L/2m, \tag{17.67}$$

that is, to satisfy the kinetic balance condition of section 11.2, the Lévy-Leblond equation becomes bounded below, with the same solutions as the Schrödinger equation. These solutions are the positive-energy solutions. Negative-energy solutions exist but these are pathological: their form may be inferred by making the redefinition of the small component in the Dirac solutions for the atom and taking the limit $c \to \infty$.

We now have an apparently well-behaved operator, in which no expansion has been made that is invalid at small r. That is not to say that the magnitude of the perturbation is always small, because we still have V/c^2 as the perturbation, which becomes infinite at $r = 0$ for a point nucleus. The rate of convergence of the perturbation series will still be related to this feature. However, the fact that the perturbation operator is large in some region of space does not mean that the integrals over the operator are also large in value, as discussed above in relation to the Pauli Hamiltonian.

We start by writing the transformed Dirac equation as

$$\left(\hat{\mathcal{H}}_0 + \frac{1}{c^2}\hat{\mathcal{H}}_2\right)\begin{pmatrix}\psi^L\\ \phi^S\end{pmatrix} = E\left(\hat{\mathcal{G}}_0 + \frac{1}{c^2}\hat{\mathcal{G}}_2\right)\begin{pmatrix}\psi^L\\ \phi^S\end{pmatrix} \tag{17.68}$$

with the zeroth-order equation as

$$\hat{\mathcal{H}}_0\begin{pmatrix}\psi^{L(0)}\\ \phi^{S(0)}\end{pmatrix} = E_0\begin{pmatrix}\psi^{L(0)}\\ \phi^{S(0)}\end{pmatrix}. \tag{17.69}$$

The perturbation parameter is chosen to be $\mu = 1/c$, and the expansion performed in even powers. The metric must also be included in the perturbation expansion, so we have

$$(\hat{\mathcal{H}}_0 + \mu^2\hat{\mathcal{H}}_2)\sum_{k=0}^{\infty}\mu^{2k}\psi^{(2k)} = (\hat{\mathcal{G}}_0 + \mu^2\hat{\mathcal{G}}_2)\sum_{m=0}^{\infty}\mu^{2m}\psi^{(2m)}\sum_{k=0}^{\infty}\mu^{2k}E^{(2k)}. \tag{17.70}$$

Writing out the expressions up to fourth order, we have

$$\begin{aligned}
&\left(\hat{\mathcal{H}}_0 - \hat{\mathcal{G}}_0E^{(0)}\right)\psi^{(0)} = 0,\\
&\left(\hat{\mathcal{H}}_0 - \hat{\mathcal{G}}_0E^{(0)}\right)\psi^{(2)} = \left(\hat{\mathcal{G}}_0E^{(2)} + \hat{\mathcal{G}}_2E^{(0)} - \hat{\mathcal{H}}_2\right)\psi^{(0)},\\
&\left(\hat{\mathcal{H}}_0 - \hat{\mathcal{G}}_0E^{(0)}\right)\psi^{(4)} = \left(\hat{\mathcal{G}}_0E^{(2)} + \hat{\mathcal{G}}_2E^{(0)} - \hat{\mathcal{H}}_2\right)\psi^{(2)} + \left(\hat{\mathcal{G}}_0E^{(4)} + \hat{\mathcal{G}}_2E^{(2)}\right)\psi^{(0)}.
\end{aligned} \tag{17.71}$$

The zeroth-order equation has been discussed above. The kinetic balance relation holds between the large and small components of the zeroth-order wave function. Intermediate normalization has been chosen for the zeroth-order wave function,

$$\langle\,\psi^{(0)}\,|\hat{\mathcal{G}}_0|\,\psi\,\rangle = 1. \tag{17.72}$$

It follows from this choice that

$$\begin{aligned} \langle \psi^{(0)} | \hat{\mathcal{G}}_0 | \psi^{(0)} \rangle &= 1, \quad \langle \psi^{(0)} | \hat{\mathcal{G}}_0 | \psi^{(2k)} \rangle = 0 \\ \langle \psi^{L(0)} | \psi^{L(0)} \rangle &= 1, \quad \langle \psi^{L(2k)} | \psi^{L(0)} \rangle = 0. \end{aligned} \tag{17.73}$$

that is, the nonrelativistic wave function is normalized to unity and is orthogonal to the higher corrections to the large component.

Premultiplying (17.71) by $\psi^{(0)\dagger}$ and integrating we get the relativistic corrections to the energy,

$$\begin{aligned} &E^{(2)} \langle \psi^{(0)} | \hat{\mathcal{G}}_0 | \psi^{(0)} \rangle = \langle \psi^{(0)} | \hat{\mathcal{H}}_2 - E^{(0)} \hat{\mathcal{G}}_2 | \psi^{(0)} \rangle, \\ &E^{(4)} \langle \psi^{(0)} | \hat{\mathcal{G}}_0 | \psi^{(0)} \rangle = \langle \psi^{(0)} | \hat{\mathcal{H}}_2 - E^{(0)} \hat{\mathcal{G}}_2 - E^{(2)} \hat{\mathcal{G}}_0 | \psi^{(2)} \rangle - \langle \psi^{(0)} | E^{(2)} \hat{\mathcal{G}}_2 | \psi^{(0)} \rangle. \end{aligned} \tag{17.74}$$

Making use of the intermediate normalization relations and the definitions of the operators, we can write the energies as

$$\begin{aligned} E^{(2)} &= \langle \phi^{S(0)} | V - E^{(0)} | \phi^{S(0)} \rangle, \\ E^{(4)} &= \langle \phi^{S(0)} | V - E^{(0)} | \phi^{S(2)} \rangle - E^{(2)} \langle \phi^{S(0)} | \phi^{S(0)} \rangle. \end{aligned} \tag{17.75}$$

Now, with the relation between the large and small components for the zeroth-order wave function, we may write the second-order energy as

$$E^{(2)} = \frac{1}{4m^2} \langle \psi^{L(0)} | (\boldsymbol{\sigma} \cdot \mathbf{p})(V - E^{(0)})(\boldsymbol{\sigma} \cdot \mathbf{p}) | \psi^{L(0)} \rangle. \tag{17.76}$$

This is precisely the same energy as we would have obtained from (17.5) if we had expanded *both* the energy *and* the wave function in a perturbation series, without trying to define an energy-independent Hamiltonian. It is also the same as the Pauli energy, a point that may be demonstrated as follows:

$$\begin{aligned} &\langle \psi^{L(0)} | (\boldsymbol{\sigma} \cdot \mathbf{p}) V (\boldsymbol{\sigma} \cdot \mathbf{p}) | \psi^{L(0)} \rangle - E^{(0)} \langle \psi^{L(0)} | \mathbf{p}^2 | \psi^{L(0)} \rangle \\ &\quad = \langle \psi^{L(0)} | \mathbf{p} V \cdot \mathbf{p} + \hbar \boldsymbol{\sigma} \cdot (\nabla V) \times \mathbf{p} | \psi^{L(0)} \rangle - \langle \psi^{L(0)} | \hat{T} + V | \psi^{L(0)} \rangle \langle \psi^{L(0)} | \mathbf{p}^2 | \psi^{L(0)} \rangle \\ &\quad = \langle \psi^{L(0)} | \mathbf{p} V \cdot \mathbf{p} + \hbar \boldsymbol{\sigma} \cdot (\nabla V) \times \mathbf{p} | \psi^{L(0)} \rangle - \langle \psi^{L(0)} | \mathbf{p}^4 / 2m + V \mathbf{p}^2 | \psi^{L(0)} \rangle. \end{aligned} \tag{17.77}$$

The last step follows from the fact that $\psi^{L(0)}$ is an eigenfunction of $\hat{T} + V$, and we may replace the projector onto the ground state $| \psi^{L(0)} \rangle \langle \psi^{L(0)} |$ by a projector onto the complete set of states, and hence remove it. With the addition of the factor of $1/4m^2c^2$, the mass–velocity and spin–orbit terms are now obvious in this expression, and the remaining two terms can be shown by integration by parts to be equal to the Darwin term. However, the equivalence of the DPT and Pauli energy depends on the validity of the removal of the ground state projector. This will not be the case in general, and we conclude that it applies strictly only to exact wave functions. We return to the distinction between exact and approximate wave functions later.

The fact that we have obtained the Pauli energy without using an expansion of the inverse or square root of an operator in powers of $1/c^2$ means that, regardless of the validity of this expansion, the expectation value may nevertheless have a convergent series expansion (Kutzelnigg 1990). We know already that the expansion of the Dirac energy of a hydrogenic atom in powers of Z/c is convergent for $Z < c$. The perturbation expansion of the energy must therefore also be convergent. The problem lies in the form of the operators. In Pauli theory one obtains highly singular operators that give divergent terms, but these terms nevertheless, by carefully controlled cancellation, may be made to sum to the correct finite result (Morrison and Moss 1980) at least up to $\mathcal{O}(c^{-6})$. In DPT, on the other hand, the operators are much more well-behaved.

We can trace this difference a little further by looking at the higher-order energies. Here, we will only be concerned with the momentum terms. The energy at each order $2k$ contains a term involving the energy at the next lowest order,

$$E^{(2k)} = E^{(2k-2)} \langle \psi^{L(0)} | \mathbf{p}^2/4m^2 | \psi^{L(0)} \rangle + \cdots . \tag{17.78}$$

Substituting successively for the energy at each order we get

$$E^{(2k)} = E^{(0)} \Big(\langle \psi^{L(0)} | \mathbf{p}^2/4m^2 | \psi^{L(0)} \rangle \Big)^k + \cdots . \tag{17.79}$$

With the factor of $1/c^{2k}$ we see that we have an expression involving the kth power of the expectation value of $\mathbf{p}^2/m^2c^2$. Because $\psi^{L(0)}$ is not an eigenfunction of $\mathbf{p}^2$ we cannot make the same substitution as we did above, replacing the projector $| \psi^{L(0)} \rangle \langle \psi^{L(0)} |$ by the infinite expansion and remove it to obtain the expectation of the kth power of $\mathbf{p}^2/m^2c^2$. That is,

$$\Big(\langle \psi^{L(0)} | \mathbf{p}^2/4m^2 | \psi^{L(0)} \rangle \Big)^k \neq \langle \psi^{L(0)} | \mathbf{p}^{2k}/(2m)^{2k} | \psi^{L(0)} \rangle. \tag{17.80}$$

So, in the convergent theory, the high powers of the momentum are evaluated only as powers of the expectation of $\mathbf{p}^2$, which is always finite.

The singular nature of the Pauli operators can in fact be traced to the normalization of the wave function. When we renormalized the large component, we introduced operators that canceled the energy-dependent term and produced the mass–velocity term. This connection between the singular operators and the normalization has been made in another way by Kutzelnigg (1999).

Since the operators in direct perturbation theory are not singular, we can continue to evaluate energies to higher order. The evaluation of the fourth-order (and sixth-order) energy requires the second-order perturbed wave function, for which the equation is given in component form by

$$\begin{aligned} \big(V - E^{(0)}\big)\, \psi^{L(2)} + (\boldsymbol{\sigma} \cdot \mathbf{p}) \phi^{S(2)} &= E^{(2)} \psi^{(0)} \\ (\boldsymbol{\sigma} \cdot \mathbf{p})\, \psi^{L(2)} - 2m\phi^{S(2)} + \big(V - E^{(0)}\big)\, \phi^{S(0)} &= 0. \end{aligned} \tag{17.81}$$

We may eliminate $\phi^{S(2)}$ from the first equation using the second and use the relation between $\phi^{S(0)}$ and $\psi^{L(0)}$ to yield

$$\left(\hat{T} + V - E^{(0)}\right) \psi^{L(2)} = \left[E^{(2)} - \frac{1}{4m^2} (\boldsymbol{\sigma} \cdot \mathbf{p}) \left(V - E^{(0)}\right) (\boldsymbol{\sigma} \cdot \mathbf{p}) \right] \psi^{L(0)} \tag{17.82}$$

on which we still have to impose orthogonality of $\psi^{L(2)}$ to $\psi^{L(0)}$. We now have a second-order differential equation to solve.

However, for a hydrogenic atom with a point nucleus we know the solution to the Dirac equation, and we can expand it in powers of $1/c$. Although this may be done in general, we consider here only the 1*s* ground state. The critical term is $r^{\gamma-1}$, which we may write as $e^{(\gamma-1)\ln r}$. Expanding γ and then expanding the exponential, we get

$$r^{\gamma-1} = 1 - \frac{Z^2}{2c^2} \ln r - \frac{Z^4}{8c^4} \ln r(1 - \ln r) + \cdots . \tag{17.83}$$

To obtain the full expression for the wave function, we must also expand the normalization term, (7.40), but the radial properties of the expansion do not depend on it. This expression must be multiplied by the exponential term e^{-Zr} to obtain the wave function. The expansion of $r^{\gamma-1}$ is convergent for all r because the exponential has an infinite radius of convergence. This does not mean that the convergence is uniform. The convergence will be slow for very large and very small r. At large r this is not a problem because of the factor of e^{-Zr}, but at small r there will be slow convergence. Taking the term of $\mathcal{O}(c^{-2})$ for the second-order wave function, we find that apart from a normalization, $\psi^{L(2)} = \ln r \; e^{-Zr}$. The correctness of this expression may be verified by substitution of this and $\psi^{L(0)} = e^{-Zr}$ into (17.82).

We see then that the lowest-order relativistic correction to the wave function involves a term that is logarithmic in r, and hence is singular at the origin, and in fact more singular than the Dirac wave function. Any expansion in terms of the complete set of zeroth-order functions is bound to be problematic because it cannot adequately represent the logarithmic singularity. It might then be concluded that, although we may obtain an expansion for a hydrogenic atom because we know the form of the wave function, the method will not be of practical value beyond second order in $1/c$ for more complex systems. Such a conclusion would be premature, as we will see in the next section. This problem of course only exists for a point nucleus. For a finite nucleus there is no singularity and the wave function is well-behaved at the origin.

17.6 Stationary Direct Perturbation Theory

The problem raised at the end of the last section, that the exact solutions to the second-order perturbation equations are not known except for very simple cases, is not the only problem in the application of direct perturbation theory. The major problem is that the exact solutions of the zeroth-order problem, that is, of the Schrödinger equation, are not known except for the same simple cases. As a consequence, the zeroth-order wave function is not the eigenfunction of the zeroth-order Hamiltonian and the perturbation theory is not strictly valid. This is true not only for direct perturbation theory but also for Breit–Pauli theory. One way of addressing this problem is to partition the zeroth-order Hamiltonian and use multiple perturbation theory, as described in section 17.4.

Another way of averting this problem is to confine the expansion space for the perturbed wave functions to a set of approximate functions, for which the matrix of the zeroth-order Hamiltonian may be made diagonal. The perturbation expansion must then be done on the matrix level, not on the operator level. In essence, we are performing a projection onto a finite space, which is what we are always doing when we solve in a finite basis set.

It is possible to proceed a step further than a simple expansion, and make the energy stationary with respect to variations of the approximate functions. This is already implicit in the derivation of the zeroth-order wave function: diagonalization of the zeroth-order Hamiltonian is a variational procedure that makes the energy stationary with respect to changes in the wave function. The same can be done in higher order, leading to a *stationary* perturbation theory. Stationary direct perturbation theory (SDPT) for relativistic corrections was developed by Rutkowski and coworkers (Rutkowski 1986, Rutkowski et al. 1988, 1992) and by Kutzelnigg (1996). In this section, we will develop this theory, considering the zeroth-order problem as well as the higher orders since the zeroth-order problem has some peculiarities of its own. We will use the modified Dirac representation, which makes some of the analysis easier.

In this representation both the Hamiltonian and the metric partition transparently into zeroth-order operators and a perturbation,

$$\begin{aligned}\hat{\mathcal{H}}' &= \hat{\mathcal{H}}'_0 + \frac{1}{c^2}\hat{\mathcal{H}}'_2 = \begin{pmatrix} V & \hat{T} \\ \hat{T} & -\hat{T} \end{pmatrix} + \frac{1}{c^2}\begin{pmatrix} 0 & 0 \\ 0 & \hat{V}^p \end{pmatrix} \\ \hat{\mathcal{G}}' &= \hat{\mathcal{G}}'_0 + \frac{1}{c^2}\hat{\mathcal{G}}'_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{c^2}\begin{pmatrix} 0 & 0 \\ 0 & \hat{T}/2m \end{pmatrix}\end{aligned} \tag{17.84}$$

where the primes have been added to signify the use of modified operators and where $\hat{V}^p = (\boldsymbol{\sigma}\cdot\mathbf{p})V(\boldsymbol{\sigma}\cdot\mathbf{p})/4m^2$.

What we wish to do is to make the energy stationary at various orders of perturbation theory by expanding the expectation of the Dirac operator,

$$\begin{aligned}E &= \langle\tilde{\psi}|\hat{\mathcal{H}}'|\tilde{\psi}\rangle \\ &= \langle\psi^L|V|\psi^L\rangle + \langle\psi^L|\hat{T}|\phi^L\rangle + \langle\phi^L|\hat{T}|\psi^L\rangle - \langle\phi^L|\hat{T}|\phi^L\rangle \\ &\quad + \frac{1}{c^2}\langle\phi^L|\hat{V}^p|\phi^L\rangle\end{aligned} \tag{17.85}$$

subject to unitary normalization conditions,

$$\langle\psi^L|\psi^L\rangle + \frac{1}{2mc^2}\langle\phi^L|\hat{T}|\phi^L\rangle = 1. \tag{17.86}$$

It is useful to rewrite the zeroth-order Hamiltonian as

$$\hat{\mathcal{H}}_0' = \begin{pmatrix} \hat{T}+V & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} -\hat{T} & \hat{T} \\ \hat{T} & -\hat{T} \end{pmatrix} = \begin{pmatrix} \hat{T}+V & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} 1 \\ -1 \end{pmatrix} \hat{T} \begin{pmatrix} 1 & -1 \end{pmatrix}, \quad (17.87)$$

which can be seen to be a nonrelativistic Hamiltonian operating only on the large component and a kinetic correction term. The form of the last term shows that the expectation value always appears with a difference between ψ^L and ϕ^L, which we define to be

$$\chi^{L(n)} = \begin{pmatrix} 1 & -1 \end{pmatrix} \psi^{(n)} = \psi^{L(n)} - \phi^{L(n)}. \quad (17.88)$$

Thus, the kinetic corrections can all be expressed in terms of χ^L.

Developing the perturbation expansion gives the same basic expressions as before for the various orders of the energy, but in terms of the modified operators:

$$\begin{aligned} E^{(0)} &= \langle \psi^{L(0)} | \hat{T} + V | \psi^{L(0)} \rangle - \langle \chi^{L(0)} | \hat{T} | \chi^{L(0)} \rangle, \\ E^{(2)} &= 2\,\mathrm{Re}\langle \psi^{L(0)} | \hat{T} + V | \psi^{L(2)} \rangle - 2\,\mathrm{Re}\langle \chi^{L(0)} | \hat{T} | \chi^{L(2)} \rangle + \langle \phi^{L(0)} | \hat{V}^p | \phi^{L(0)} \rangle, \\ E^{(4)} &= 2\,\mathrm{Re}\langle \psi^{L(4)} | \hat{T} + V | \psi^{L(0)} \rangle - 2\,\mathrm{Re}\langle \chi^{L(4)} | \hat{T} | \chi^{L(0)} \rangle \\ &\quad + \langle \psi^{L(2)} | \hat{T} + V | \psi^{L(2)} \rangle - \langle \chi^{L(2)} | \hat{T} | \chi^{L(2)} \rangle + 2\,\mathrm{Re}\langle \phi^{L(0)} | \hat{V}^p | \phi^{L(2)} \rangle. \end{aligned} \quad (17.89)$$

The normalization conditions at the various orders of perturbation are

$$\begin{aligned} \langle \psi^{L(0)} | \psi^{L(0)} \rangle &= 1, \\ 2\,\mathrm{Re}\langle \psi^{L(0)} | \psi^{L(2)} \rangle + \langle \phi^{L(0)} | \hat{T}/2m | \phi^{L(0)} \rangle &= 0, \\ 2\,\mathrm{Re}\langle \psi^{L(0)} | \psi^{L(4)} \rangle + \langle \psi^{L(2)} | \psi^{L(2)} \rangle + 2\,\mathrm{Re}\langle \phi^{L(0)} | \hat{T}/2m | \phi^{L(2)} \rangle &= 0. \end{aligned} \quad (17.90)$$

This is not quite the same as the intermediate normalization conditions above, but necessary because the normalization would otherwise be expanded with the energy.

We want to consider variations of the wave function at the different orders. The functional to vary for the zeroth-order wave function is obviously the zeroth-order energy, with the zeroth-order orthogonality condition as a constraint. Note that there is no orthogonality condition involving $\chi^{(0)}$, which is determined solely from the variation equations. The zeroth-order variation equation is then

$$\delta \Big(\langle \psi^{(0)} | \hat{\mathcal{H}}_0' | \psi^{(0)} \rangle - E^{(0)} \langle \psi^{(0)} | \hat{\mathcal{G}}_0' | \psi^{(0)} \rangle \Big) = 0 \quad (17.91)$$

where $E^{(0)}$ is the Lagrange multiplier for the constraint. The zeroth-order functional is the expression in parentheses, which we can write as

$$F_0(\psi^{L(0)}, \chi^{L(0)}) = \langle \psi^{L(0)} | \hat{T} + V - E^{(0)} | \psi^{L(0)} \rangle - \langle \chi^{L(0)} | \hat{T} | \chi^{L(0)} \rangle. \quad (17.92)$$

The first term is positive since $E^{(0)}$ is a lower bound to the approximate energy. The second term is negative. Because of this, it is not possible to simply perform a minimization of the functional with respect to arbitrary variations in $\psi^{L(0)}$ and $\chi^{L(0)}$ without getting variational collapse. Instead, stationarity conditions are applied to the two parts independently, leading to

$$\begin{aligned} \langle \delta\psi^{L(0)} \,|\, \hat{T} + V - E^{(0)} \,|\, \psi^{L(0)} \rangle &= 0, \\ \langle \delta\chi^{L(0)} \,|\, \hat{T} \,|\, \chi^{L(0)} \rangle &= 0. \end{aligned} \tag{17.93}$$

We are solving these equations in a finite basis set, so the first equation gives the matrix form of the Schrödinger equation. The second gives the matrix version of the Laplace equation. We require all eigenvalues of this matrix to be zero, which means that the matrix itself must be zero. Now $\hat{T}$ is a positive semi-definite operator, and in a finite basis the diagonal elements are zero only if the function is zero. Thus, for all functions in the variational space, only the trivial solution $\chi^{L(0)} = 0$ is admissible. Consequently, $\psi^{L(0)} = \phi^{L(0)}$ for all solutions of the zeroth-order problem.

The implication of the above arguments is that we have taken the maximum of the second part of F_0 and the minimum of the first part. F_0 is called the Lévy-Leblond functional because the solutions of the Lévy-Leblond equation make this functional stationary. The Lagrange multiplier $E^{(0)}$ is simply the nonrelativistic energy.

We may make use of the variation conditions and the normalization condition to reduce the second-order energy expression. Provided $\psi^{L(2)}$ is restricted to the variational space of $\psi^{L(0)}$, we may write

$$2\,\mathrm{Re}\langle \psi^{L(2)} \,|\, \hat{T} + V \,|\, \psi^{L(0)} \rangle = 2E^{(0)}\,\mathrm{Re}\langle \psi^{L(2)} \,|\, \psi^{L(0)} \rangle. \tag{17.94}$$

The normalization condition may now be used to reduce the right-hand side to an expression involving only $\phi^{L(0)}$. With $\chi^{L(0)} = 0$, the second-order energy is

$$E^{(2)} = \langle \phi^{L(0)} \,|\, \hat{V}^{p} - E^{(0)}\hat{T}/2m \,|\, \phi^{L(0)} \rangle = \langle \psi^{L(0)} \,|\, (\boldsymbol{\sigma}\cdot\mathbf{p})(V - E^{(0)})(\boldsymbol{\sigma}\cdot\mathbf{p})/4m^2 \,|\, \psi^{L(0)} \rangle \tag{17.95}$$

which we recognize as the same expression as in the exact theory.

The variation equation for the second-order wave function comes from the fourth-order energy. If we were to make the variation of the second-order wave function in the second-order energy expression we would only get the zeroth-order equation again, so we need a functional that is quadratic in the second-order wave function. Variation of the fourth-order energy subject to the normalization conditions gives the equation

$$\begin{aligned} \delta\Big(&2\,\mathrm{Re}\langle \psi^{L(4)} \,|\, \hat{T} + V \,|\, \psi^{L(0)} \rangle - 2\,\mathrm{Re}\langle \chi^{L(4)} \,|\, \hat{T} \,|\, \chi^{L(0)} \rangle \\ &+ \langle \psi^{L(2)} \,|\, \hat{T} + V \,|\, \psi^{L(2)} \rangle - \langle \chi^{L(2)} \,|\, \hat{T} \,|\, \chi^{L(2)} \rangle + 2\,\mathrm{Re}\langle \phi^{L(0)} \,|\, \hat{V}^{p} \,|\, \phi^{L(2)} \rangle \\ &- E^{(2)}\Big[\langle \phi^{L(0)} \,|\, \hat{T}/2m \,|\, \phi^{L(0)} \rangle + 2\,\mathrm{Re}\langle \psi^{L(2)} \,|\, \psi^{L(0)} \rangle\Big] \\ &- E^{(0)}\Big[\langle \psi^{L(2)} \,|\, \psi^{L(2)} \rangle + 2\,\mathrm{Re}\langle \phi^{L(2)} \,|\, \hat{T}/2m \,|\, \phi^{L(0)} \rangle + 2\,\mathrm{Re}\langle \psi^{L(4)} \,|\, \psi^{L(0)} \rangle\Big]\Big) = 0. \end{aligned} \tag{17.96}$$

The functional, which is called the Rutkowski–Hylleraas functional, is the expression in parentheses. The fourth-order wave function can be eliminated by invoking the zeroth-order variational conditions, with $\psi^{L(4)}$ restricted to the variational space of $\psi^{L(0)}$. This causes the first and last terms to cancel, and the second term is zero due to $\chi^{L(0)} = 0$. We may also insert $\psi^{L(0)} = \phi^{L(0)}$, and replace $\phi^{L(2)}$ with $\psi^{L(2)} - \chi^{L(2)}$, to obtain

$$\begin{aligned} F_4 = &\langle \psi^{L(2)} \,|\, \hat{T} + V - E^{(0)} \,|\, \psi^{L(2)} \rangle + 2\,\mathrm{Re}\langle \psi^{L(2)} \,|\, \hat{V}^p - E^{(0)}\hat{T}/2m - E^{(2)} \,|\, \psi^{L(0)} \rangle \\ &- \langle \chi^{L(2)} \,|\, \hat{T} \,|\, \chi^{L(2)} \rangle - 2\,\mathrm{Re}\langle \chi^{L(2)} \,|\, \hat{V}^p - E^{(0)}\hat{T}/2m \,|\, \psi^{L(0)} \rangle \\ &- E^{(2)}\langle \psi^{L(0)} \,|\, \hat{T}/2m \,|\, \psi^{L(0)} \rangle. \end{aligned} \tag{17.97}$$

It can be seen that this functional divides into a part that depends on $\psi^{L(2)}$, a part that depends on $\chi^{L(2)}$, and a part that depends on neither.

$$F_4(\psi^{L(2)},\ \chi^{L(2)}) = F_{4a}(\psi^{L(2)}) + F_{4b}(\chi^{L(2)}) + F_{4c}. \tag{17.98}$$

Making the functional stationary with respect to variations in the components of the second-order wave function $\psi^{L(2)}$ and $\chi^{L(2)}$ we find

$$\begin{gathered} \langle \delta\psi^{L(2)} \,|\, (\hat{T} + V - E^{(0)})\psi^{L(2)} + (\hat{V}^p - E^{(0)}\hat{T}/2m - E^{(2)})\psi^{L(0)} \rangle = 0 \\ \langle \delta\chi^{L(2)} \,|\, \hat{T}\chi^{L(2)} + (\hat{V}^p - E^{(0)}\hat{T}/2m)\psi^{L(0)} \rangle = 0. \end{gathered} \tag{17.99}$$

This applies to all possible variations. If we let $\delta\psi^{L(2)} = \psi^{L(0)}$, the first variation equation gives us the same expression as before for the second-order energy $E^{(2)}$, which functions here as a Lagrange multiplier. If we let $\delta\psi^{L(2)} = \psi^{L(2)}$, then we get the expression

$$\langle \psi^{L(2)} \,|\, \hat{T} + V - E^{(0)} \,|\, \psi^{L(2)} \rangle = -\,\mathrm{Re}\langle \psi^{L(2)} \,|\, \hat{V}^p - E^{(0)}\hat{T}/2m - E^{(2)} \,|\, \psi^{L(0)} \rangle. \tag{17.100}$$

The right-hand side must be real because the left-hand side is real. Using this expression in the first part of the functional we find that

$$F_{4a} = -\langle \psi^{L(2)} \,|\, \hat{T} + V - E^{(0)} \,|\, \psi^{L(2)} \rangle. \tag{17.101}$$

The integral must be positive, so this term has a negative value. In the second variation equation we let $\delta\chi^{L(2)} = \chi^{L(2)}$ and find that

$$\langle \chi^{L(2)} \,|\, \hat{T} \,|\, \chi^{L(2)} \rangle = -\,\mathrm{Re}\langle \chi^{L(2)} \,|\, \hat{V}^p - E^{(0)}\hat{T}/2m \,|\, \psi^{L(0)} \rangle \tag{17.102}$$

from which the second part of the functional can be written

$$F_{4b} = \langle \chi^{L(2)} \,|\, \hat{T} \,|\, \chi^{L(2)} \rangle \tag{17.103}$$

which must be positive. The third part must also be positive because $E^{(2)}$ is negative. With these expressions we may finally write the fourth-order energy as

$$\begin{aligned} E^{(4)} = &-\langle \psi^{L(2)} \,|\, \hat{T} + V - E^{(0)} \,|\, \psi^{L(2)} \rangle - E^{(2)} \langle \psi^{L(0)} \,|\, \hat{T}/2m \,|\, \psi^{L(0)} \rangle \\ &+ \langle \chi^{L(2)} \,|\, \hat{T} \,|\, \chi^{L(2)} \rangle. \end{aligned} \tag{17.104}$$

Like the zeroth-order functional, the fourth-order functional has both negative and positive terms. The question is, do these have bounds, and if so what are the conditions? To determine the bounds, we consider an approximate and an exact solution to the variation equations for the second-order wave function. Placing a tilde over the approximate solution, we can write the difference for the first functional as

$$\begin{aligned} F_{4a}(\tilde{\psi}^{L(2)}) - F_{4a}(\psi^{L(2)}) = &\langle \tilde{\psi}^{L(2)} \,|\, \hat{T} + V - E^{(0)} \,|\, \tilde{\psi}^{L(2)} \rangle \\ &+ 2\,\mathrm{Re}\langle \tilde{\psi}^{L(2)} \,|\, \hat{V}^{p} - E^{(0)} \hat{T}/2m - E^{(2)} \,|\, \psi^{L(0)} \rangle \\ &+ \langle \psi^{L(2)} \,|\, \hat{T} + V - E^{(0)} \,|\, \psi^{L(2)} \rangle. \end{aligned} \tag{17.105}$$

Now the variation equation must be satisfied for $\delta\psi^{L(2)} = \tilde{\psi}^{L(2)}$, so the second term can be written

$$\langle \tilde{\psi}^{L(2)} \,|\, \hat{V}^{p} - E^{(0)} \hat{T}/2m - E^{(2)} \,|\, \psi^{L(0)} \rangle = -\langle \tilde{\psi}^{L(2)} \,|\, \hat{T} + V - E^{(0)} \,|\, \psi^{L(2)} \rangle \tag{17.106}$$

and hence the difference in functionals becomes

$$F_{4a}(\tilde{\psi}^{L(2)}) - F_{4a}(\psi^{L(2)}) = \langle \tilde{\psi}^{L(2)} - \psi^{L(2)} \,|\, \hat{T} + V - E^{(0)} \,|\, \tilde{\psi}^{L(2)} - \psi^{L(2)} \rangle. \tag{17.107}$$

This must be a positive quantity, so F_{4a}, which is negative, is bounded from below by the exact value. By a similar procedure we may show that

$$F_{4b}(\tilde{\chi}^{L(2)}) - F_{4b}(\chi^{L(2)}) = -\langle \tilde{\chi}^{L(2)} - \chi^{L(2)} \,|\, \hat{T} \,|\, \tilde{\chi}^{L(2)} - \chi^{L(2)} \rangle, \tag{17.108}$$

which is negative. As for the zeroth-order functional, we may not perform an unconstrained variation of each of the parts without risk of collapse. However, if we choose $\chi^{L(2)}$ to satisfy the variation equation exactly, that is,

$$\hat{T}\chi^{L(2)} + (\hat{V}^{p} - E^{(0)} \hat{T}/2m)\psi^{L(0)} = 0 \tag{17.109}$$

then the fourth-order energy is bounded from below. This choice is problematic for a point nucleus because it makes $\phi^{L(2)}$ too singular at the origin and consequently the sixth-order energy becomes infinite (Kutzelnigg 1996). In this respect the choice made for $\chi^{L(2)}$ differs markedly from the choice for $\chi^{L(0)}$. Here, it is better to expand $\phi^{L(2)}$ in the space of functions used to expand $\phi^{L(0)}$ and use the variation condition in matrix form.

Several comments are worth making at this point. First, the expressions for the energy from the stationary perturbation theory are exactly the same as one would obtain from the normal procedure of expanding the wave function in the zeroth-order solutions. The use of the stationary theory nevertheless has some advantages, particularly in the analysis of the energy and wave function. The use of the zeroth-order solutions is not the only choice that can be made, and in fact some choices give a lower bound to the energy functionals. The second comment is that using the set of zeroth-order solutions as an expansion basis is not in fact a serious restriction. From the chapter on the modified Dirac equation (chapter 15) we see that we may expand both large and pseudo-large components in the same basis, and in a finite basis the set of all solutions for the large component can be used to represent the pseudo-large component without further approximation. There is therefore no need to use the pathological negative-energy states from the Lévy-Leblond equation for the perturbation expansion. Third, the use of a finite nuclear size gets rid of many of the problems of the expansion, such as the attempt to represent the logarithmic terms in the exact second-order wave function. There is now no singularity at the nucleus and the large and pseudo-large components are finite and have the same behavior at the origin.

17.7 Stationary Direct Perturbation Theory for Many-Electron Systems

In section 17.4, we developed a double perturbation theory approach for many-electron systems, and we would expect to do the same for direct perturbation theory. The development follows entirely analogous lines to that section. However, it is possible to use the stationarity principle on other functionals than the one-electron energy, which was implicit in the developments of the previous section. Stationary direct perturbation theory has been developed for DHF energies (Kutzelnigg et al. 1995) and for multiconfiguration DHF energies (Sundholm and Ottschofski 1997). In this section, we wish to illustrate the general approach for the closed-shell DHF energy. The reader may follow the analogy for other energy expressions. The development will follow the lines of the previous section rather than that of the original papers, where the DPT of the eigenvalues of the Fock matrix is given first.

The energy expression is the usual (unrestricted) expression for a closed-shell state,

$$\begin{aligned} E^{\mathrm{DHF}} &= \sum_i \langle \psi_i | \hat{\mathcal{H}}' | \psi_i \rangle + \tfrac{1}{2} \sum_{ij} \big(\langle \psi_i \psi_j | g' | \psi_i \psi_j \rangle - \langle \psi_i \psi_j | g' | \psi_j \psi_i \rangle \big) \\ &= \sum_i \langle \psi_i | \hat{\mathcal{H}}' + \tfrac{1}{2}(\hat{\mathcal{J}}' - \hat{\mathcal{K}}') | \psi_i \rangle \end{aligned} \tag{17.110}$$

where the primes indicate that we are using the modified Dirac operators. $\hat{\mathcal{J}}'$ and $\hat{\mathcal{K}}'$ are the usual direct and exchange operators in terms of modified Dirac 4-spinors.

This energy is to be expanded as a perturbation series in $1/c$ and the appropriate functionals varied subject to the unitary normalization conditions to arrive at stationarity conditions. The stationarity conditions naturally lead to the DHF equations, so we expect the variation of the perturbation functionals to involve an expansion of the DHF operators.

To perform the variation of the energies at various orders, we need the expansion of the one- and two-electron modified Dirac operators in terms of the components, which are found in (17.84) and (15.43). From these we can define two-component direct and exchange operators,

$$\begin{aligned}
\hat{\mathcal{J}}^{kl}_{p0} &= \sum_r \int d\mathbf{r}\ \psi^{L(k)\dagger}_r(j)\hat{g}^{p0}_{ij}\ \psi^{L(l)}_r(j)/(2m)^p \\
\hat{\mathcal{J}}^{kl}_{p2} &= \sum_r \int d\mathbf{r}\ \phi^{L(k)\dagger}_r(j)\hat{g}^{p2}_{ij}\ \phi^{L(l)}_r(j)/(2m)^{p+2}
\end{aligned} \tag{17.111}$$

$$\begin{aligned}
\hat{\mathcal{K}}^{kl}_{00}\psi^L(i) &= \sum_r \int d\mathbf{r}\ \psi^{L(k)\dagger}_r(j)\hat{g}^{00}_{ij}\ \psi^L(j)\ \psi^{L(l)}_r(i) \\
\hat{\mathcal{K}}^{kl}_{02}\psi^L(i) &= \sum_r \int d\mathbf{r}\ \phi^{L(k)\dagger}_r(j)\hat{g}^{02}_{ij}\ \phi^L(j)\ \psi^{L(l)}_r(i)/4m^2 \\
\hat{\mathcal{K}}^{kl}_{20}\phi^L(i) &= \sum_r \int d\mathbf{r}\ \psi^{L(k)\dagger}_r(j)\hat{g}^{20}_{ij}\ \psi^L(j)\ \phi^{L(l)}_r(i)/4m^2 \\
\hat{\mathcal{K}}^{kl}_{22}\phi^L(i) &= \sum_r \int d\mathbf{r}\ \phi^{L(k)\dagger}_r(j)\hat{g}^{22}_{ij}\ \phi^L(j)\ \phi^{L(l)}_r(i)/16m^4.
\end{aligned} \tag{17.112}$$

The superscripts give the order in perturbation theory of the spinor, the subscripts correspond to the superscripts on the operators and determine whether ψ or ϕ is used for the density, and may take the values 0 and 2. The second and third of the exchange operators exchange the component types as well as the electron indices; this is merely a convenient form of expression that reflects the fact that the spinors contributing to the density for any electron in the Dirac–Coulomb approximation must have the same component type.

As before, we expand the energy and the one-electron functions in powers of $1/c$, which gives the energy expressions to fourth order,

$$E^{(0)} = \sum_i \Big(\langle \psi^{L(0)}_i | \hat{T} + V + \tfrac{1}{2}(\hat{\mathcal{J}}^{00}_{00} - \hat{\mathcal{K}}^{00}_{00}) | \psi^{L(0)}_i \rangle - \langle \chi^{L(0)}_i | \hat{T} | \chi^{L(0)}_i \rangle \Big), \tag{17.113}$$

$$\begin{aligned}
E^{(2)} = \sum_i \Big(& 2\,\mathrm{Re}\langle \psi^{L(2)}_i | \hat{T} + V + \hat{\mathcal{J}}^{00}_{00} - \hat{\mathcal{K}}^{00}_{00} | \psi^{L(0)}_i \rangle - 2\,\mathrm{Re}\langle \chi^{L(2)} | \hat{T} | \chi^{L(0)} \rangle \\
& + \langle \phi^{L(0)}_i | \hat{V}^p + \hat{\mathcal{J}}^{00}_{20} - \hat{\mathcal{K}}^{00}_{20} | \phi^{L(0)}_i \rangle \Big),
\end{aligned} \tag{17.114}$$

$$\begin{aligned}
E^{(4)} = \sum_i \Big(& 2\,\mathrm{Re}\langle \psi_i^{L(4)} \,|\, \hat{T} + V + \hat{\mathcal{J}}_{00}^{00} - \hat{\mathcal{K}}_{00}^{00} \,|\, \psi_i^{(0)} \rangle - 2\,\mathrm{Re}\langle \chi_i^{L(4)} \,|\, \hat{T} \,|\, \chi_i^{L(0)} \rangle \\
& + \langle \psi_i^{L(2)} \,|\, \hat{T} + V + \hat{\mathcal{J}}_{00}^{00} - \hat{\mathcal{K}}_{00}^{00} \,|\, \psi_i^{L(2)} \rangle - \langle \chi_i^{L(2)} \,|\, \hat{T} \,|\, \chi_i^{L(2)} \rangle \\
& + \mathrm{Re}\langle \psi_i^{L(2)} \,|\, \hat{\mathcal{J}}_{00}^{20} - \hat{\mathcal{K}}_{00}^{20} + \hat{\mathcal{J}}_{00}^{02} - \hat{\mathcal{K}}_{00}^{02} + 2(\hat{\mathcal{J}}_{02}^{00} - \hat{\mathcal{K}}_{02}^{00}) \,|\, \psi_i^{L(0)} \rangle \\
& + 2\,\mathrm{Re}\langle \phi_i^{L(2)} \,|\, \hat{V}^p + \hat{\mathcal{J}}_{20}^{00} - \hat{\mathcal{K}}_{20}^{00} \,|\, \phi_i^{L(0)} \rangle + \tfrac{1}{2}\langle \phi_i^{L(0)} \,|\, \hat{\mathcal{J}}_{22}^{00} - \hat{\mathcal{K}}_{22}^{00} \,|\, \phi_i^{L(0)} \rangle \Big),
\end{aligned} \tag{17.115}$$

and as before, we vary the zeroth- and fourth-order energy subject to the unitary normalization conditions (17.90). Using $\epsilon^{(0)}$ as the Lagrange multiplier for this constraint, the zeroth-order variation equations in component form are

$$\begin{gathered}
\langle \delta\psi_i^{L(0)} \,|\, \hat{T} + V + \hat{\mathcal{J}}_{00}^{00} - \hat{\mathcal{K}}_{00}^{00} - \epsilon_i^{(0)} \,|\, \psi_i^{L(0)} \rangle = 0, \\
\langle \delta\chi_i^{L(0)} \,|\, \hat{T} \,|\, \chi_i^{L(0)} \rangle = 0.
\end{gathered} \tag{17.116}$$

The second of these is the same as for the one-electron case, and the first has merely added the Coulomb and exchange operators to the one-electron operator to give the Hartree–Fock equation for spinor i. As before, all variations are valid, and we may use these equations and the unitary normalization conditions (17.90) to reexpress the second-order energy as

$$E^{(2)} = \sum_i \langle \phi_i^{L(0)} \,|\, \hat{V}^p - \epsilon_i^{(0)} \hat{T}/2m + \hat{\mathcal{J}}_{20}^{00} - \hat{\mathcal{K}}_{20}^{00} \,|\, \phi_i^{L(0)} \rangle. \tag{17.117}$$

Since $\phi^{L(0)} = \psi^{L(0)}$, the second-order energy is simply an expectation value over the nonrelativistic wave function. The one-electron term is, as we would expect, the term we obtained in the one-electron theory.

The variation of the second-order wave function is done using the fourth-order energy, from which we first eliminate the fourth-order wave function using the zeroth-order variation equation with $\delta\psi^{L(0)} = \psi^{L(4)}$ and the normalization conditions. The fourth-order functional is then

$$\begin{aligned}
F_4 = \sum_i \Big(& \langle \psi_i^{L(2)} \,|\, \hat{T} + V - \epsilon_i^{(0)} + \hat{\mathcal{J}}_{00}^{00} - \hat{\mathcal{K}}_{00}^{00} \,|\, \psi_i^{L(2)} \rangle - \langle \chi_i^{L(2)} \,|\, \hat{T} \,|\, \chi_i^{L(2)} \rangle \\
& + \mathrm{Re}\langle \psi_i^{L(2)} \,|\, -2\epsilon_i^{(2)} + \hat{\mathcal{J}}_{00}^{20} - \hat{\mathcal{K}}_{00}^{20} + \hat{\mathcal{J}}_{00}^{02} - \hat{\mathcal{K}}_{00}^{02} + 2(\hat{\mathcal{J}}_{02}^{00} - \hat{\mathcal{K}}_{02}^{00}) \,|\, \psi_i^{L(0)} \rangle \\
& + 2\,\mathrm{Re}\langle \phi_i^{L(2)} \,|\, \hat{V}^p - \epsilon_i^{(0)} \hat{T}/2m + \hat{\mathcal{J}}_{20}^{00} - \hat{\mathcal{K}}_{20}^{00} \,|\, \phi_i^{L(0)} \rangle \\
& + \langle \phi_i^{L(0)} \,|\, -\epsilon_i^{(2)} \hat{T}/2m + \tfrac{1}{2}(\hat{\mathcal{J}}_{22}^{00} - \hat{\mathcal{K}}_{22}^{00}) \,|\, \phi_i^{L(0)} \rangle \Big).
\end{aligned} \tag{17.118}$$

Substituting $\psi^{L(2)} - \chi^{L(2)}$ for $\phi^{L(2)}$, the variation equations in component form are

$$\begin{aligned}
&\langle \delta\psi_i^{L(2)} \,|\, \hat{T} + V - \epsilon_i^{(0)} + \hat{\mathcal{J}}^{00}_{00} - \hat{\mathcal{K}}^{00}_{00} \,|\, \psi_i^{L(2)} \rangle + \langle \delta\psi_i^{L(2)} \,|\, \hat{V}^p - \epsilon_i^{(0)} \hat{T}/2m - \epsilon_i^{(2)} \,|\, \psi_i^{L(0)} \rangle \\
&\quad + \langle \delta\psi_i^{L(2)} \,|\, \hat{\mathcal{J}}^{20}_{00} - \hat{\mathcal{K}}^{20}_{00} + \hat{\mathcal{J}}^{02}_{00} - \hat{\mathcal{K}}^{02}_{00} + \hat{\mathcal{J}}^{00}_{02} - \hat{\mathcal{K}}^{00}_{02} + \hat{\mathcal{J}}^{00}_{20} - \hat{\mathcal{K}}^{00}_{20} \,|\, \psi_i^{L(0)} \rangle = 0, \\
&\langle \delta\chi_i^{L(2)} \,|\, \hat{T} \,|\, \chi_i^{L(2)} \rangle + \langle \delta\chi_i^{L(2)} \,|\, \hat{V}^p - \epsilon_i^{(0)} \hat{T}/2m + \hat{\mathcal{J}}^{00}_{20} - \hat{\mathcal{K}}^{00}_{20} \,|\, \phi_i^{L(0)} \rangle = 0.
\end{aligned} \tag{17.119}$$

With $\delta\psi^{L(2)} = \psi^{L(2)}$ and $\delta\chi^{L(2)} = \chi^{L(2)}$ these equations yield expressions that may be used to simplify the fourth-order energy functional to

$$\begin{aligned}
F_4 = \sum_i \Big(&- \langle \psi_i^{L(2)} \,|\, \hat{T} + V - \epsilon_i^{(0)} + \hat{\mathcal{J}}^{00}_{00} - \hat{\mathcal{K}}^{00}_{00} \,|\, \psi_i^{L(2)} \rangle + \langle \chi_i^{L(2)} \,|\, \hat{T} \,|\, \chi_i^{L(2)} \rangle \\
&- \mathrm{Re} \langle \psi_i^{L(2)} \,|\, \hat{\mathcal{J}}^{20}_{00} - \hat{\mathcal{K}}^{20}_{00} + \hat{\mathcal{J}}^{02}_{00} - \hat{\mathcal{K}}^{02}_{00} \,|\, \psi_i^{L(0)} \rangle \\
&+ \langle \phi_i^{L(0)} \,|\, - \epsilon_i^{(2)} \hat{T}/2m + \tfrac{1}{2} (\hat{\mathcal{J}}^{00}_{22} - \hat{\mathcal{K}}^{00}_{22}) \,|\, \phi_i^{L(0)} \rangle \Big).
\end{aligned} \tag{17.120}$$

Other expressions may be obtained by eliminating different terms in the original expression. To complete the development, we need the expansion of the orbital eigenvalues to second order. The development follows exactly the same lines as the one-electron development that allows us to write

$$\begin{aligned}
\epsilon_i^{(0)} &= \langle \psi^{L(0)} \,|\, \hat{T} + V + \hat{\mathcal{J}}^{00}_{00} - \hat{\mathcal{K}}^{00}_{00} \,|\, \psi_i^{L(0)} \rangle \\
\epsilon_i^{(2)} &= \langle \phi_i^{L(0)} \,|\, \hat{V}^p - \epsilon_i^{(0)} \hat{T}/2m + \hat{\mathcal{J}}^{00}_{20} - \hat{\mathcal{K}}^{00}_{20} \,|\, \phi_i^{L(0)} \rangle \\
&\quad + \langle \psi_i^{L(0)} \,|\, \hat{\mathcal{J}}^{20}_{00} - \hat{\mathcal{K}}^{20}_{00} + \hat{\mathcal{J}}^{02}_{00} - \hat{\mathcal{K}}^{02}_{00} + \hat{\mathcal{J}}^{00}_{02} - \hat{\mathcal{K}}^{00}_{02} \,|\, \psi_i^{L(0)} \rangle.
\end{aligned} \tag{17.121}$$

Alternatively, the eigenvalues may be obtained from the zeroth- and second-order variation equations. We can now use the simplified fourth-order energy functional to define a set of equations to solve in our finite basis for the second-order wave function.

17.8 Direct Perturbation Theory of Properties

The calculation of properties using direct perturbation theory follows exactly the same lines as we used for Breit–Pauli theory. As we noted above, stationary direct perturbation theory leads to precisely the same equations we would have obtained by simply expanding the perturbed wave functions in the set of eigenfunctions of the zeroth-order Hamiltonian, and on this basis we proceed with the development of multiple direct perturbation theory for properties.

The main difference between the direct and the Breit–Pauli perturbation schemes is that in the former we have to expand the metric as well as the Hamiltonian. To illustrate the differences between the two we will confine ourselves to the one-electron case, so we will be dealing with a double perturbation expansion in relativity and the applied field. We will perform the development in parallel, to compare the two schemes.

As in section 17.4, we use the parameters μ and η for relativity and the external field, with $\mu = 1/c$ and η the field strength. The Hamiltonian we write as

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \mu^2 \hat{\mathcal{H}}_{20} + \eta \hat{\mathcal{H}}_{01} + \mu^2 \eta \hat{\mathcal{H}}_{21}. \tag{17.122}$$

The two perturbations will have a different form for direct and Breit–Pauli perturbation theory. The wave function also will be different in form: two components in Breit–Pauli theory and four components in direct perturbation theory. In addition, the metric must be expanded in the direct scheme,

$$\hat{\mathcal{G}} = \hat{\mathcal{G}}_0 + \mu^2 \hat{\mathcal{G}}_2. \tag{17.123}$$

The energy and wave function are expanded in both perturbation parameters,

$$\begin{aligned} E &= \sum_{m,i=0}^{\infty} \mu^{2m} \eta^i E^{(2mi)} \\ \Psi &= \sum_{m,i=0}^{\infty} \mu^{2m} \eta^i \Psi^{(2mi)}. \end{aligned} \tag{17.124}$$

Writing out the lowest-order expressions, we have

$$\begin{aligned} \text{BP} \ &: \quad \hat{\mathcal{H}}_0 \Psi^{(00)} = E^{(00)} \Psi^{(00)} \\ \text{DPT} &: \quad \hat{\mathcal{H}}_0 \Psi^{(00)} = E^{(00)} \hat{\mathcal{G}}_0 \Psi^{(00)} \end{aligned} \tag{17.125}$$

$$\begin{aligned} \text{BP} \ &: \quad \hat{\mathcal{H}}_0 \Psi^{(01)} + \hat{\mathcal{H}}_{01} \Psi^{(00)} = E^{(00)} \Psi^{(01)} + E^{(01)} \Psi^{(00)} \\ \text{DPT} &: \quad \hat{\mathcal{H}}_0 \Psi^{(01)} + \hat{\mathcal{H}}_{01} \Psi^{(00)} = E^{(00)} \hat{\mathcal{G}}_0 \Psi^{(01)} + E^{(01)} \hat{\mathcal{G}}_0 \Psi^{(00)} \end{aligned} \tag{17.126}$$

$$\begin{aligned} \text{BP} \ &: \quad \hat{\mathcal{H}}_0 \Psi^{(20)} + \hat{\mathcal{H}}_{20} \Psi^{(00)} = E^{(00)} \Psi^{(20)} + E^{(20)} \Psi^{(00)} \\ \text{DPT} &: \quad \hat{\mathcal{H}}_0 \Psi^{(20)} + \hat{\mathcal{H}}_{20} \Psi^{(00)} = E^{(00)} \hat{\mathcal{G}}_0 \Psi^{(20)} + E^{(00)} \hat{\mathcal{G}}_2 \Psi^{(00)} + E^{(20)} \hat{\mathcal{G}}_0 \Psi^{(00)} \end{aligned} \tag{17.127}$$

$$\begin{aligned} \text{BP} \ &: \quad \hat{\mathcal{H}}_0 \Psi^{(21)} + \hat{\mathcal{H}}_{20} \Psi^{(01)} + \hat{\mathcal{H}}_{01} \Psi^{(20)} + \hat{\mathcal{H}}_{21} \Psi^{(00)} \\ &\qquad = E^{(00)} \Psi^{(21)} + E^{(20)} \Psi^{(01)} + E^{(01)} \Psi^{(20)} + E^{(21)} \Psi^{(00)} \\ \text{DPT} &: \quad \hat{\mathcal{H}}_0 \Psi^{(21)} + \hat{\mathcal{H}}_{20} \Psi^{(01)} + \hat{\mathcal{H}}_{01} \Psi^{(20)} + \hat{\mathcal{H}}_{21} \Psi^{(00)} \\ &\qquad = E^{(00)} \hat{\mathcal{G}}_0 \Psi^{(21)} + E^{(00)} \hat{\mathcal{G}}_2 \Psi^{(01)} + E^{(20)} \hat{\mathcal{G}}_0 \Psi^{(01)} \\ &\qquad\quad + E^{(01)} \hat{\mathcal{G}}_0 \Psi^{(20)} + E^{(01)} \hat{\mathcal{G}}_2 \Psi^{(00)} + E^{(21)} \hat{\mathcal{G}}_0 \Psi^{(01)}. \end{aligned} \tag{17.128}$$

The first two equations are equivalent, yielding the nonrelativistic energy and the first-order external field term,

$$E^{(00)} = \langle \Psi^{(00)} | \hat{\mathcal{H}}_{00} | \Psi^{(00)} \rangle, \tag{17.129}$$

$$E^{(01)} = \langle \Psi^{(00)} | \hat{\mathcal{H}}_{01} | \Psi^{(00)} \rangle. \tag{17.130}$$

The third equation gives the relativistic correction to the energy,

$$\begin{aligned} \text{BP} \quad : \quad & E^{(20)} = \langle \Psi^{(00)} | \hat{\mathcal{H}}_{20} | \Psi^{(00)} \rangle \\ \text{DPT} : \quad & E^{(20)} = \langle \Psi^{(00)} | \hat{\mathcal{H}}_{20} - E^{(00)} \hat{\mathcal{G}}_2 | \Psi^{(00)} \rangle. \end{aligned} \tag{17.131}$$

As we saw in section 17.5, these are the same for exact wave functions, but differ for inexact wave functions. The fourth equation gives the lowest-order relativistic correction to the property,

$$\begin{aligned} \text{BP} \quad : \quad & \mathrm{E}^{(21)} = \langle \Psi^{(00)} | \hat{\mathcal{H}}_{01} | \Psi^{(20)} \rangle + \langle \Psi^{(00)} | \hat{\mathcal{H}}_{20} | \Psi^{(01)} \rangle + \langle \Psi^{(00)} | \hat{\mathcal{H}}_{21} | \Psi^{(00)} \rangle \\ \text{DPT} : \quad & \mathrm{E}^{(21)} = \langle \Psi^{(00)} | \hat{\mathcal{H}}_{01} | \Psi^{(20)} \rangle + \langle \Psi^{(00)} | \hat{\mathcal{H}}_{20} - E^{(00)} \hat{\mathcal{G}}_2 | \Psi^{(01)} \rangle \\ & \quad + \langle \Psi^{(00)} | \hat{\mathcal{H}}_{21} - E^{(01)} \hat{\mathcal{G}}_2 | \Psi^{(00)} \rangle. \end{aligned} \tag{17.132}$$

So far we have not specified the perturbation operator, nor have we expanded the DPT wave function in components, which is necessary for the comparison of the two theories. This we proceed to do next, considering both electric and magnetic perturbations. We will focus on $E^{(21)}$ since it is here that we may see differences between BP and DPT.

The operator for the electric perturbation, W, behaves in the same way as the potential. Expanding the field perturbation equation (17.126) into its components in DPT, we find

$$\begin{aligned} (V - E^{(00)})\psi^{L(01)} + (\boldsymbol{\sigma} \cdot \mathbf{p})\phi^{S(01)} + W\psi^{L(00)} &= E^{(01)}\psi^{L(00)} \\ (\boldsymbol{\sigma} \cdot \mathbf{p})\psi^{L(01)} - 2m\phi^{S(01)} &= 0. \end{aligned} \tag{17.133}$$

The second equation is the same as for the zeroth-order equation, and we immediately arrive at the nonrelativistic equation by substituting the second into the first,

$$(\hat{T} + V - E^{(00)})\psi^{L(01)} + W\psi^{L(00)} = E^{(01)}\psi^{L(00)}, \tag{17.134}$$

which is the same as in Pauli theory. Next we expand the DPT equation for the relativistic correction to the perturbation using intermediate normalization for the large component and making use of substitutions where possible to arrive at

$$E^{(21)} = \langle \psi^{L(00)} | W | \psi^{L(20)} \rangle + \langle \psi^{L(00)} | (\boldsymbol{\sigma} \cdot \mathbf{p})(V - E^{(00)})(\boldsymbol{\sigma} \cdot \mathbf{p}) | \psi^{L(01)} \rangle / 4m^2 \\ + \langle \psi^{L(00)} | (\boldsymbol{\sigma} \cdot \mathbf{p})(W - E^{(01)})(\boldsymbol{\sigma} \cdot \mathbf{p}) | \psi^{L(00)} \rangle / 4m^2. \tag{17.135}$$

The equivalent Pauli expression is

$$E^{(21)} = \langle \psi^{(00)} | W | \psi^{(20)} \rangle + \langle \psi^{(00)} | -\mathbf{p}^4/2m + [(\boldsymbol{\sigma} \cdot \mathbf{p}), \boldsymbol{\sigma} \cdot (\mathbf{p}V)] | \psi^{(01)} \rangle / 4m^2 \\ + \langle \psi^{(00)} | [(\boldsymbol{\sigma} \cdot \mathbf{p}), \boldsymbol{\sigma} \cdot (\mathbf{p}W)] | \psi^{(00)} \rangle / 4m^2. \tag{17.136}$$

In each expression, the first two terms are equivalent, due to the interchange theorem of double perturbation theory (Dalgarno and Stewart 1956). This may be verified by making use of the equations for the two perturbations—but note that it is true only for exact wave functions.

We can show that the second term in the DPT expression is the same as the second term in the Pauli expression, by substituting $\langle \psi^{L(00)} | \hat{T} + V | \psi^{L(00)} \rangle$ for $E^{(00)}$ as we did for the relativistic correction to the energy and removing the projector onto $\psi^{L(00)}$. This is again only true for exact wave functions. Because of the equivalence of the first and second terms, this means that the first term in each expression is also the same for exact wave functions. Any differences must therefore be in the third term. We can no longer convert the DPT operator into a form that is equivalent to the Pauli operator because $\psi^{(00)}$ is not an eigenfunction of W. To demonstrate, we write the third integral in the DPT expression as

$$\langle \psi^{L(00)} | (\boldsymbol{\sigma} \cdot \mathbf{p})(W - E^{(01)})(\boldsymbol{\sigma} \cdot \mathbf{p}) | \psi^{L(00)} \rangle \\ = \langle \psi^{L(00)} | (\boldsymbol{\sigma} \cdot \mathbf{p}) W (\boldsymbol{\sigma} \cdot \mathbf{p}) | \psi^{L(00)} \rangle - \langle \psi^{L(00)} | \mathbf{p}^2 | \psi^{L(00)} \rangle \langle \psi^{L(00)} | W | \psi^{L(00)} \rangle. \tag{17.137}$$

If we could remove the projector $| \psi^{L(00)} \rangle \langle \psi^{L(00)} |$ this expression would rearrange into the Pauli expression with integration by parts, but the presence of the projector ensures that the DPT and Pauli values of this term are different, even for exact wave functions.

It should be noted that the equivalence for exact wave functions is further restricted by the form of the perturbation operator. For external field perturbations, where the operator is not singular, there is no problem, but for point nuclear fields, where the operator *is* singular, the integration by parts leaves a boundary term that does not vanish.

Now, if we expand the perturbed wave functions in both cases in the set of zeroth-order solutions, we find again that the first and second terms are equivalent in each case. Provided we use the exact ground state wave function, the two theories produce the same values for these terms, but if we are using a finite expansion, the results are no longer the same.

A similar situation arises for the magnetic perturbations. Here there is a formal difference, in that the perturbation operator has a relativistic correction in the Pauli theory but in DPT there is no relativistic correction: the operator is simply

$$\hat{\mathcal{H}}_{01} = e \begin{pmatrix} \mathbf{0}_2 & (\boldsymbol{\sigma} \cdot \mathbf{A}) \\ (\boldsymbol{\sigma} \cdot \mathbf{A}) & \mathbf{0}_2 \end{pmatrix}. \tag{17.138}$$

Now we may write the expansion of (17.126) as

$$\begin{aligned} (V - E^{(00)})\psi^{L(01)} + (\boldsymbol{\sigma} \cdot \mathbf{p})\phi^{S(01)} + e(\boldsymbol{\sigma} \cdot \mathbf{A})\phi^{S(00)} &= E^{(01)}\psi^{L(00)} \\ (\boldsymbol{\sigma} \cdot \mathbf{p})\psi^{L(01)} - 2m\phi^{S(01)} + e(\boldsymbol{\sigma} \cdot \mathbf{A})\psi^{L(00)} &= 0. \end{aligned} \tag{17.139}$$

Substitution of the second equation into the first gives the usual expression for the terms linear in **A**,

$$(\hat{T} + V - E^{(00)})\psi^{L(01)} + e\big[(\boldsymbol{\sigma} \cdot \mathbf{p}), (\boldsymbol{\sigma} \cdot \mathbf{A})\big]_+/2m\psi^{L(00)} = E^{(01)}\psi^{L(00)}. \tag{17.140}$$

Expanding the DPT expression for $E^{(21)}$ in components and substituting for the small components we arrive at

$$\begin{aligned} E^{(21)} = &\langle \psi^{L(00)} \,|\, (\boldsymbol{\sigma}\cdot\mathbf{p})(V - E^{(00)})(\boldsymbol{\sigma}\cdot\mathbf{p}) \,|\, \psi^{L(01)} \rangle/4m^2 \\ &+ e\langle \psi^{L(00)} \,|\, \big[(\boldsymbol{\sigma}\cdot\mathbf{A}), (\boldsymbol{\sigma}\cdot\mathbf{p})\big]_+ \,|\, \psi^{L(20)} \rangle/2m \\ &+ e\langle \psi^{L(00)} \,|\, (\boldsymbol{\sigma}\cdot\mathbf{p})(V - E^{(00)})(\boldsymbol{\sigma}\cdot\mathbf{A}) + (\boldsymbol{\sigma}\cdot\mathbf{A})(V - E^{(00)})(\boldsymbol{\sigma}\cdot\mathbf{p}) \,|\, \psi^{L(00)} \rangle/4m^2 \\ &- e\langle \psi^{L(00)} \,|\, \big[(\boldsymbol{\sigma}\cdot\mathbf{A}), (\boldsymbol{\sigma}\cdot\mathbf{p})\big]_+ \,|\, \psi^{L(00)} \rangle\langle \psi^{L(00)} \,|\, \mathbf{p}^2 \,|\, \psi^{L(00)} \rangle/8m^3. \end{aligned} \tag{17.141}$$

The corresponding expression for the Pauli Hamiltonian is

$$\begin{aligned} E^{(21)} = &\langle \psi^{(00)} \,|\, -\mathbf{p}^4/2m + \big[(\boldsymbol{\sigma} \cdot \mathbf{p}), \boldsymbol{\sigma} \cdot (\mathbf{p}V)\big] \,|\, \psi^{(01)} \rangle/4m^2 \\ &+ e\langle \psi^{(00)} \,|\, \big[(\boldsymbol{\sigma} \cdot \mathbf{A}), (\boldsymbol{\sigma} \cdot \mathbf{p})\big]_+ \,|\, \psi^{(20)} \rangle/2m \\ &+ e\langle \psi^{(00)} \,|\, \big[(\boldsymbol{\sigma} \cdot \mathbf{A}), \boldsymbol{\sigma} \cdot (\mathbf{p}V)\big] \,|\, \psi^{(00)} \rangle/8m^2 \\ &- e\langle \psi^{(00)} \,|\, \big[\big[(\boldsymbol{\sigma} \cdot \mathbf{A}), (\boldsymbol{\sigma} \cdot \mathbf{p})\big]_+, \mathbf{p}^2\big]_+ \,|\, \psi^{(00)} \rangle/8m^3. \end{aligned} \tag{17.142}$$

The first two terms in the DPT expression have obvious equivalents to the Pauli expressions under the same conditions as given above for the electric perturbation. It is the third and fourth terms that provide the equivalents of the relativistic corrections to the Pauli operator. The spin–orbit correction clearly comes from the third term. The kinetic correction is divided between the third and fourth terms. We can certainly make the substitution for $E^{(00)}$ in the third term for exact wave functions and rearrange to give half the kinetic correction, but the other half involves a projector and this we cannot avoid.

The conclusion is that there are real differences between Pauli and direct perturbation theory for the lowest-order relativistic corrections to both electric and magnetic properties, which do not vanish even for exact wave functions. Direct perturbation theory is in general to be preferred because it is convergent and therefore can be used to higher order, in contrast to Breit–Pauli theory.

18

Regular Approximations

Perturbation theory is a useful tool for evaluating small corrections to a system, and as we noted in the preceding chapter, relativity is a small correction for much of the periodic table. If we can use perturbation theory based on an expansion in $1/c$ we assign much of the work associated with a more complete relativistic treatment to the end of an otherwise nonrelativistic calculation. The problems with the Pauli Hamiltonian—the singular operators and the questionable validity of an expansion in powers of $\mathbf{p}/mc$—are essentially circumvented by the use of direct perturbation theory. For systems containing heavier atoms it is necessary to go to higher order in $1/c$ perturbation theory, and possibly even abandon perturbation theory altogether. If we could perform an expansion that yielded a zeroth-order Hamiltonian incorporating relativistic effects to some degree and that was manifestly convergent, it might be possible to use perturbation theory to low order for heavy elements.

If we wish to incorporate some level of relativistic effects into the zeroth-order Hamiltonian, we cannot start from Pauli perturbation theory or direct perturbation theory. But can we find an alternative expansion that contains relativistic corrections and is valid for all r: that is, can we derive a regular expansion that is convergent for all reasonable values of the parameters? The expansion we consider in this chapter has roots in the work by Chang, Pélissier, and Durand (1986) and Heully et al. (1986), which was developed further by van Lenthe et al. (1993, 1994). These last authors coined the term "regular approximation" because of the properties of the expansion.

The Pauli expansion results from taking $2mc^2$ out of the denominator of the equation for the elimination of the small component (ESC). The problem with this is that both E and V can potentially be larger in magnitude than $2mc^2$ and so the expansion is not valid in some region of space. In particular, there is always a region close to the nucleus where $|V - E|/2mc^2 > 1$. An alternative operator to extract from the denominator is the operator $2mc^2 - V$, which is always positive definite for the nuclear potential and is always greater than $2mc^2$. Moreover, since the potential becomes infinite at a point nucleus, having V in the denominator regularizes the expansion at the nucleus. With this choice we may eliminate the small component by writing

$$(V - E)\psi^L + \frac{1}{2m}(\boldsymbol{\sigma} \cdot \mathbf{p})\frac{2mc^2}{2mc^2 - V}\left[1 + \frac{E}{2mc^2 - V}\right]^{-1}(\boldsymbol{\sigma} \cdot \mathbf{p})\psi^L = 0. \qquad (18.1)$$

The power series expansion of the inverse has now been converted from one that depends critically for its convergence on the magnitude of the potential to one that depends on the magnitude of the energy. The minimum magnitude of the potential is zero, so the power series is now valid everywhere for energies $|E| < 2mc^2$. This range covers all the electron bound states, and free (continuum) states up to an energy of $2mc^2$. For most quantum chemical purposes, such a restriction on the virtual orbitals will present little if any problem. For energies outside this range, there is a region from zero to some value of r inside which the series is convergent. Contrast this to the Pauli approximation, where it is the region from zero to some value of r inside which the series is *di*vergent. Since this is the region where relativistic effects are most important, the new expansion ought to be superior to the Pauli expansion.

This expansion has some further interesting properties. Close to the nuclei where the potential is large, it is an expansion in inverse powers of the potential. The closer we are to the nucleus, the smaller the inverse is and the better the expansion becomes. Far from the nucleus, where the potential is effectively zero, the expansion becomes an expansion in powers of $1/c^2$. The magnitude of the terms is also dependent on the energy, and so the expansion will be more rapidly convergent for small energies than for large energies—a useful property for chemistry, which is concerned with the small energies of the valence region. Finally, not the least of the interesting properties of this expansion, is the zeroth-order Hamiltonian, which we examine in the next section.

18.1 The CPD or ZORA Hamiltonian

In the development of the Pauli Hamiltonian in section 17.1, truncation of the power series expansion of the inverse operator after the first term yielded the nonrelativistic Hamiltonian. In (18.1), the zeroth-order term is the Hamiltonian first developed by Chang, Pélissier, and Durand (1986), often referred to as the CPD Hamiltonian. The name given by van Lenthe et al. is the zeroth-order regular approximation, ZORA, which we will adopt here. The zeroth-order Hamiltonian is

$$\hat{\mathcal{H}}^{\text{ZORA}} = V + (\boldsymbol{\sigma} \cdot \mathbf{p})\frac{c^2}{2mc^2 - V}(\boldsymbol{\sigma} \cdot \mathbf{p}). \qquad (18.2)$$

The fact that this Hamiltonian includes relativistic corrections can be demonstrated by expanding $(2mc^2 - V)^{-1}$ in a series, to give

$$\hat{\mathcal{H}}^{\text{ZORA}} = V + \frac{1}{2m}(\boldsymbol{\sigma} \cdot \mathbf{p})\left[1 + \frac{V}{2mc^2} + \frac{V^2}{4m^2c^4} + \dots\right](\boldsymbol{\sigma} \cdot \mathbf{p}). \qquad (18.3)$$

The first term is simply $\hat{T}$, the nonrelativistic kinetic energy. Commuting $(\boldsymbol{\sigma} \cdot \mathbf{p})$ to the right and using the Dirac relation (4.14), the second term gives

$$(\boldsymbol{\sigma} \cdot \mathbf{p})(V)(\boldsymbol{\sigma} \cdot \mathbf{p}) = V\mathbf{p}^2 + (\mathbf{p}V) \cdot \mathbf{p} + \hbar\boldsymbol{\sigma} \cdot (\nabla V) \times \mathbf{p}. \tag{18.4}$$

This operator contains the Pauli spin–orbit operator and part, but not all, of the spin-free relativistic correction. What is missing, compared with the $1/c^2$ expansion in (17.5), is the term involving the energy, $(\boldsymbol{\sigma} \cdot \mathbf{p})(-E_0)(\boldsymbol{\sigma} \cdot \mathbf{p})$, which gives rise to the mass–velocity correction and part of the Darwin term. We conclude from this that the ZORA Hamiltonian should reproduce spin–orbit splittings well but will be deficient in the spin-free relativistic corrections. This is potentially serious because the spin-free relativistic corrections are usually the largest. Since the missing term is positive for bound states, we conclude that ZORA will give energies that are too low. This conclusion will be verified later.

We could also partition $\hat{\mathcal{H}}^{\text{ZORA}}$ by commuting $(\boldsymbol{\sigma} \cdot \mathbf{p})$ to the right, to give

$$\hat{\mathcal{H}}^{\text{ZORA}} = V + \frac{2mc^2}{2mc^2 - V}\hat{T} + \frac{c^2}{(2mc^2 - V)^2}\big[(\mathbf{p}V) \cdot \mathbf{p} - \hbar\boldsymbol{\sigma} \cdot (\nabla V) \times \mathbf{p}\big]. \tag{18.5}$$

Here we see a modified kinetic energy term that is cut off near the nuclei, a spin-free relativistic correction, and a spin–orbit term, both of which are regularized and behave as $1/r$ for small r. We may compare this with the regularization of the free-particle Foldy–Wouthuysen or Douglas–Kroll transformed Hamiltonian of section 16.3. The regularization clearly corrects the overestimation of relativistic effects that plagues the Pauli Hamiltonian. There is another consequence of the small r behavior. Since the relativistic correction operator behaves like $1/r$, it ought to be possible to use $\hat{\mathcal{H}}^{\text{ZORA}}$ variationally—and in fact we may demonstrate that there is a variational lower bound.

There are several ways to demonstrate the lower bound. The spectrum of an atomic or a molecular Hamiltonian in which the potential goes to zero at large distances can be determined by considering the eigenstates in this same region of space. At large distances, where the potential is effectively zero, the factor $c^2/(2mc^2 - V)$ reduces to $1/2m$ and the ZORA Hamiltonian reduces to the nonrelativistic Schrödinger Hamiltonian. It therefore has the same eigenvalue spectrum, with oscillatory (continuum) states for $E > 0$ and exponentially decaying (bound) states for $E < 0$. This means that the lowest state must be a bound state, unlike the states of the Dirac Hamiltonian, where the lowest state is a negative-energy continuum state.

We can obtain more detailed information, not only on the spectrum but also on a number of properties, from the relationship between the ZORA and the Dirac equations. For an atomic Coulomb potential (i.e., for a point nucleus) we can write the Dirac equation with the small component eliminated as

$$\left[-\frac{Z}{r} + (\boldsymbol{\sigma} \cdot \mathbf{p})\ \frac{c^2}{2mc^2 + E^{\text{D}} + Z/r}(\boldsymbol{\sigma} \cdot \mathbf{p})\right]\psi^L(r) = E^{\text{D}}\psi^L(r). \tag{18.6}$$

We now define a scaled coordinate $r' = \lambda r$, where λ is an energy-dependent parameter. For the momentum operator we have $\mathbf{p} = \lambda \mathbf{p}'$. Substituting into (18.6) we get

$$\left[-\frac{Z}{r'} + (\boldsymbol{\sigma} \cdot \mathbf{p}') \, \frac{c^2}{(2mc^2 + E^{\mathrm{D}})/\lambda + Z/r'} \, (\boldsymbol{\sigma} \cdot \mathbf{p}') \right] \psi^L \left(\frac{r'}{\lambda} \right) = \frac{E^{\mathrm{D}}}{\lambda} \, \psi^L \left(\frac{r'}{\lambda} \right). \tag{18.7}$$

If we choose $\lambda = (2mc^2 + E^{\mathrm{D}})/2mc^2$ and substitute into (18.7) we get

$$\left[-\frac{Z}{r'} + (\boldsymbol{\sigma} \cdot \mathbf{p}') \, \frac{c^2}{2mc^2 + Z/r'} (\boldsymbol{\sigma} \cdot \mathbf{p}') \right] \psi^L \left(\frac{r'}{\lambda} \right) = \frac{2mc^2 E^{\mathrm{D}}}{2mc^2 + E^{\mathrm{D}}} \, \psi^L \left(\frac{r'}{\lambda} \right). \tag{18.8}$$

This is the ZORA equation for an atomic Coulomb potential with $r^{\mathrm{ZORA}} = r'$, and the ZORA wave function is the large component of the Dirac wave function in the scaled coordinate system. Note that even though we have used a single nucleus in the derivation we could equally well have used a sum over nuclei for the nuclear potential. The results therefore apply to molecules as well as atoms, but we have to scale the nuclear coordinates as well as the electronic coordinates. Clearly, for most "atomic" values of E^{D} λ will be close to 1, but for heavy atoms the deviation will be significant.

From (18.8) we get a relation between the ZORA and the Dirac eigenvalues,

$$E^{\mathrm{ZORA}} = \frac{2mc^2 E^{\mathrm{D}}}{2mc^2 + E^{\mathrm{D}}} = E^{\mathrm{D}} \left[1 + \frac{E^{\mathrm{D}}}{2mc^2} \right]^{-1} = 2mc^2 \left[1 + \frac{2mc^2}{E^{\mathrm{D}}} \right]^{-1}. \tag{18.9}$$

Since we know the behavior of the Dirac spectrum, this immediately gives us information about the ZORA spectrum. For positive energies E^{D}, the ZORA energy is always smaller than the Dirac energy. As $E^{\mathrm{D}} \to \infty$ the ZORA energy approaches $2mc^2$. The positive-energy spectrum is therefore compressed from the interval $(0, \infty)$ to $(0, 2mc^2)$. For the Dirac bound states, which lie in the interval $(-mc^2, 0)$, the ZORA energies spread out to the interval $(-2mc^2, 0)$, lying below the Dirac energies. This is what we concluded earlier from qualitative considerations. If the Dirac energy were to approach $-2mc^2$ from above, the corresponding ZORA energy would approach $-\infty$. The Dirac negative-energy states get translated to appear *above* the positive-energy states: they are mapped from $(-\infty, -2mc^2)$ to $(2mc^2, \infty)$. The singular point at $E^{\mathrm{D}} = -2mc^2$ is one for which the large component is zero and the ZORA wave function corresponds to the trivial solution of the eigenvalue equation, $\psi^{\mathrm{ZORA}} = 0$. The ZORA equation is therefore bounded from below and is variational—the ground state is the variational minimum. A more rigorous proof has been provided by van Lenthe (1996).

In the derivation of the ZORA equation, we made the assumption that $|E| < 2mc^2$ so that we could do the expansion of the inverse operator. However, the ZORA equation has a valid spectrum for all E. It is therefore not necessarily the case that the use of a truncated expansion is invalid outside the strict radius of convergence. We could equally have made a partitioning using the operator relation

$$(A + B)^{-1} = A^{-1} - (A + B)^{-1} B A^{-1}. \tag{18.10}$$

This partitioning is valid for all values of the operators provided the inverses exist. For the case we have just considered, that is, $A = 2mc^2 - V$, $B = E$ with $V = -Z/r$, the partitioning is valid for all $E > -2mc^2$. For $E < -2mc^2$ the left-hand side is always singular for some value of r, as pointed out in chapter 4; however, the partitioning remains valid. As a further note, it can be shown (van Lenthe 1996) that the ZORA Hamiltonian is only bounded below for $Z < c$, due to the fact that the Dirac equation has no bound solutions for $Z > c$ for a point nucleus.

To extend the bound to a general potential, we simply add a term to the Coulomb potential of the nucleus,

$$V(r) = -\frac{Z}{r} + V_1(r), \tag{18.11}$$

where $V_1(r)$ is positive for all r but $V(r)$ is always negative. This will be the case for a finite nuclear size correction, or for the electron repulsion in a Hartree–Fock potential, for example—the most important cases for relativistic quantum chemistry. We now partition the inverse using (18.10), giving

$$\begin{aligned} \hat{\mathcal{H}}^{\mathrm{ZORA}} &= -\frac{Z}{r} + (\boldsymbol{\sigma} \cdot \mathbf{p}) \frac{c^2}{2mc^2 + Z/r} (\boldsymbol{\sigma} \cdot \mathbf{p}) \\ &\quad + V_1 + (\boldsymbol{\sigma} \cdot \mathbf{p}) \frac{c^2 V_1}{(2mc^2 + Z/r)(2mc^2 + Z/r - V_1)} (\boldsymbol{\sigma} \cdot \mathbf{p}). \end{aligned} \tag{18.12}$$

The first two terms are the point nucleus ZORA Hamiltonian. In the last term, the expression between the $(\boldsymbol{\sigma} \cdot \mathbf{p})$ operators is positive because $V_1 < 2mc^2 + Z/r$, and therefore the expectation value of this operator is positive. We can conclude that the energy for a general potential is always greater than for the bare Coulomb potential. Even if V_1 is negative somewhere, such as in a negative ion or a polar molecule, the last two terms may still have a positive expectation value and the bound will still be valid.

It is interesting to note that for the bound states, the Dirac eigenvalue is an upper bound to the ZORA eigenvalue, which is bounded from below—even though the Dirac operator itself is not bounded from below. What the latter statement really means is that the lowest Dirac eigenvalue has no bound, because it is in the negative continuum. The boundedness of the electron states is a different question, and here we see that we can provide a rigorous bound for the Dirac bound-state eigenvalues.

It has probably not escaped the attention of the reader that methods based on the ZORA Hamiltonian will be difficult to implement because the potential appears in the denominator. There are two possible approaches that could be taken to the evaluation of the inverse terms. One is to use numerical integration techniques where the potentials are tabulated on a grid. Formation of a function of the potential such as an inverse power is then a trivial problem. This kind of approach is suited to density functional techniques where the exchange-correlation potential is usually evaluated on a grid, and tabulation of the Coulomb potential on the same grid is also feasible. The second is to use matrix techniques and perform a matrix inversion (Filatov 2002). This essentially involves the insertion of the resolution of the identity in various places. The resolution for the inverse powers, however, has to be performed in the small-component basis set as there is always a $(\boldsymbol{\sigma} \cdot \mathbf{p})$ connecting the inverse operators with the large-component function. The elimination of the small component therefore does not eliminate the small-component basis set.

A second and more serious consequence of the appearance of the potential in the denominator is that the ZORA Hamiltonian is not invariant to the choice of electric gauge. Adding a constant to the potential should result in the addition of a constant to the energy, which is indeed the case for the Dirac equation. For ZORA, the relation between the ZORA and the Dirac eigenvalue for a one-electron system is given by (18.9). If we add a constant, Δ, to the Dirac energy in this equation, we get

$$\begin{aligned} E^{\mathrm{ZORA}}(V+\Delta) &= \frac{2mc^2(E^{\mathrm{D}}+\Delta)}{2mc^2+E^{\mathrm{D}}+\Delta} \\ &= E^{\mathrm{ZORA}}+\Delta\left(1-\frac{E^{\mathrm{ZORA}}}{2mc^2}\right)\left(1-\frac{E^{\mathrm{ZORA}}(V+\Delta)}{2mc^2}\right). \end{aligned} \tag{18.13}$$

Since E^{ZORA} and $E^{\mathrm{ZORA}}(V+\Delta)$ are negative for $\Delta < |E^{\mathrm{D}}|$, the factor multiplying Δ is positive and greater than 1. Therefore, the addition of a constant to the potential in ZORA does *not* result in a constant energy shift: the ZORA energy has an *electric gauge-dependence*, that is, it depends on where the zero of the potential is chosen.

This gauge dependence turns out to be serious for calculations of energies of most processes because it is almost invariably the case that the potential changes. The problem is most serious for the core orbitals because the factor multiplying Δ increases as the magnitude of the eigenvalues increase. In molecular processes, the tail of the potentials from neighboring atoms provides an almost constant, negative shift, which results in a net attractive potential. The shift in the core eigenvalues is not insignificant (van Lenthe et al. 1994).

Several methods have been proposed to alleviate this problem. The first is to freeze the core orbitals for some suitable state, such as the isolated atom, and perform the calculation in the valence space only, where the gauge problem is not as serious because the eigenvalues are smaller (van Lenthe et al. 1994). Such restrictions may prove to be too limiting for widespread applications—but no less serious than those made in pseudopotential or model potential methods, where the core is frozen. The second is to freeze the potential that appears in the denominator. To do this some approximations must be made. Several criteria for a valid frozen potential have been proposed by van Wüllen (1998) which are that the potential (a) has the correct behavior near the nuclei, (b) does not depend on the orbitals, and (c) has no contribution from distant atoms or molecules. In addition, the potential must represent the real system fairly well or it will be of no value. Probably the most obvious choice are to take a superposition of atomic potentials or to construct the potential from a superposition of atomic densities. These two differ for a density functional method only in the exchange-correlation potential. A third approach is to add terms from neighboring atoms to the potential in the denominator to cancel the erroneous attraction due to the neighboring centers (Filatov and Cremer 2005). A fourth approach, which scales the eigenvalues, is presented in a later section.

The distortion of the ZORA eigenstates from the Dirac eigenstates proves not to be insignificant for chemical purposes. It is easy to show from (18.9) that the ZORA energy differs from the Dirac energy by $\mathcal{O}(c^{-2})$, which is of the same order as the lowest relativistic correction to the nonrelativistic Hamiltonian. Although a lot can be gained from the ZORA equation, for any sort of quantitative results it is usually necessary to go beyond zeroth order. This could be achieved either by perturbation theory or by seeking a correction that can be treated variationally. We will consider both possibilities: the perturbative corrections in the next section and the variational corrections in section 18.3.

18.2 Perturbative Corrections to the ZORA Hamiltonian

There are several ways in which we can develop a perturbation series for the ZORA equation. The first is simply to ignore the normalization—a perfectly valid procedure since the wave function is only defined up to a multiplicative constant. This we will do later in the present section. The second is to follow the same procedure as in the development of the Pauli Hamiltonian in chapter 17, and the third is to start from the Foldy–Wouthuysen transformation, as in the development of chapter 16. The last two of these both explicitly involve the normalization. We will commence here with the procedure used in chapter 17.

Just as in the Pauli approximation, we will expand the normalization operator in a series. The normalized wave function is given by (17.7), and the normalization operator from (17.10) expressed in the new expansion parameter is

$$\hat{\mathcal{O}} = \left[1 + (\boldsymbol{\sigma} \cdot \mathbf{p}) \frac{c^2}{(2mc^2 - V)^2} \left[1 + \frac{E}{2mc^2 - V} \right]^{-2} (\boldsymbol{\sigma} \cdot \mathbf{p}) \right]^{1/2} . \tag{18.14}$$

The expansion of this operator is again a double series, one for the square root and one for the inverse square. Before we proceed, we must consider what we are to use as a perturbation parameter. The power series expansion of the inverse is in $E/(2mc^2 - V)$, and if we neglected the normalization we would simply multiply this term by a formal perturbation parameter. Here however we want to include the normalization and eliminate the energy dependence, so we are really considering an expansion in $1/(2mc^2 - V)$, of which the terms up to order 1 comprise the zeroth-order Hamiltonian. The expansion of the normalization operator (and its inverse) therefore gives us

$$\begin{aligned} \hat{\mathcal{O}} &= 1 + \frac{1}{2} (\boldsymbol{\sigma} \cdot \mathbf{p}) \frac{c^2}{(2mc^2 - V)^2} (\boldsymbol{\sigma} \cdot \mathbf{p}) + \dots ; \\ \hat{\mathcal{O}}^{-1} &= 1 - \frac{1}{2} (\boldsymbol{\sigma} \cdot \mathbf{p}) \frac{c^2}{(2mc^2 - V)^2} (\boldsymbol{\sigma} \cdot \mathbf{p}) + \dots . \end{aligned} \tag{18.15}$$

Following the Pauli development from the elimination of the small component, we may write

$$\begin{aligned}\hat{O}^{-1}\hat{\mathcal{H}}^{ESC}\hat{O}^{-1}\psi^N = \Bigg[\hat{\mathcal{H}}^{\text{ZORA}} - \frac{1}{2}\Big[(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p}), \hat{\mathcal{H}}^{\text{ZORA}}\Big]_+ \\ -E(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p}) + \dots\Bigg]\psi^N \\ = E\Bigg[1 - (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p}) + \dots\Bigg]\psi^N.\end{aligned} \tag{18.16}$$

The energy-dependent term on the left cancels with the corresponding term on the right, and we end up with

$$\begin{aligned}\hat{\mathcal{H}}^{\text{RA}} &= \hat{\mathcal{H}}^{\text{ZORA}} - \frac{1}{2}\Big[(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p}), \hat{\mathcal{H}}^{\text{ZORA}}\Big]_+ + \dots \\ &= \hat{\mathcal{H}}^{\text{ZORA}} + \hat{\mathcal{H}}^{\text{FORA}} + \dots .\end{aligned} \tag{18.17}$$

FORA stands for first-order regular approximation, and is the first-order correction to the ZORA Hamiltonian.

What is the magnitude of the first-order perturbation correction obtained with $\hat{\mathcal{H}}^{\text{FORA}}$? We may write it as

$$\begin{aligned}E^{(1)} &= -\frac{1}{2}\langle\psi^{\text{ZORA}}|\Big[(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p}), \hat{\mathcal{H}}^{\text{ZORA}}\Big]_+|\psi^{\text{ZORA}}\rangle \\ &= -E^{\text{ZORA}}\langle\psi^{\text{ZORA}}|(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})|\psi^{\text{ZORA}}\rangle.\end{aligned} \tag{18.18}$$

To evaluate the integral, we may make use of the scaled coordinate system that was used to relate the Dirac and the ZORA equations in the previous section to show that

$$\langle\psi^{\text{ZORA}}|(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})|\psi^{\text{ZORA}}\rangle = \frac{\langle\psi^S|\psi^S\rangle}{\langle\psi^L|\psi^L\rangle}. \tag{18.19}$$

From the solutions of the Dirac equation for a hydrogenic atom in chapter 7 we can show that the ratio of the norms of the small and large components is $-E^{\text{D}}/(2mc^2 + E^{\text{D}})$, which is equal to $-E^{\text{ZORA}}/2mc^2$. The first-order energy is thus

$$E^{\text{FORA}} = \big(E^{\text{ZORA}}\big)^2/2mc^2. \tag{18.20}$$

With the use of (18.9) we may write the sum of the zeroth- and first-order energies in terms of the Dirac energy as

$$E^{(0+1)} = E^{\mathrm{ZORA}}\left(1 + \frac{E^{\mathrm{ZORA}}}{2mc^2}\right) = E^{\mathrm{D}}\left[1 - \left(\frac{E^{\mathrm{D}}}{2mc^2 + E^{\mathrm{D}}}\right)^2\right]. \tag{18.21}$$

The Dirac eigenvalue is the shifted eigenvalue in which the negative continuum starts at $-2mc^2$, and which has $-Z^2/2$ as its leading term. So, $E^{(0+1)}$ is correct to $\mathcal{O}(c^{-2})$, and the first-order correction has removed the error of $\mathcal{O}(c^{-2})$ in the ZORA energy.

There is a slight inconsistency in the preceding development of the perturbation, which is that the labeling of the orders of perturbation in powers of $1/(2mc^2 - V)$ does not strictly work because there are already terms of two different orders in the zeroth-order Hamiltonian. What we must do is to label the terms in each of the expansions of inverse or square root operators. It is in fact easier to define the perturbation series if we start from the unnormalized equation (18.1) and simply multiply $E/(2mc^2 - V)$ by the formal perturbation parameter, to give

$$\left[V + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{2mc^2 - V}(\boldsymbol{\sigma}\cdot\mathbf{p})\right]\psi^L + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{2mc^2 - V}\sum_{k=1}^{\infty}\left[\frac{-\lambda E}{2mc^2 - V}\right]^k(\boldsymbol{\sigma}\cdot\mathbf{p})\psi^L$$
$$= E\psi^L. \tag{18.22}$$

The zeroth-order Hamiltonian is obviously $\hat{\mathcal{H}}^{\mathrm{ZORA}}$. Expanding ψ^L and E in a series in λ and collecting terms we have

$$\begin{aligned}
\lambda^0:&\quad \hat{\mathcal{H}}^{\mathrm{ZORA}}\psi^{(0)} = E^{(0)}\psi^{(0)};\\
\lambda^1:&\quad \hat{\mathcal{H}}^{\mathrm{ZORA}}\psi^{(1)} - (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})E^{(0)}\psi^{(0)} = E^{(0)}\psi^{(1)} + E^{(1)}\psi^{(0)};\\
\lambda^2:&\quad \hat{\mathcal{H}}^{\mathrm{ZORA}}\psi^{(2)} - (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\left(E^{(0)}\psi^{(1)} + E^{(1)}\psi^{(0)}\right)\\
&\quad\quad + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^3}(\boldsymbol{\sigma}\cdot\mathbf{p})\left(E^{(0)}\right)^2\psi^{(0)}\\
&\quad\quad = E^{(0)}\psi^{(2)} + E^{(1)}\psi^{(1)} + E^{(2)}\psi^{(0)}.
\end{aligned} \tag{18.23}$$

Premultiplying the first-order equation by $\psi^{(0)\dagger}$ and integrating, we get the first-order energy expression

$$E^{(1)} = -E^{(0)}\langle\psi^{(0)}|(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})|\psi^{(0)}\rangle/\langle\psi^{(0)}|\psi^{(0)}\rangle, \tag{18.24}$$

which is precisely the FORA energy obtained above, on the assumption that $\psi^{(0)}$ is normalized. If we expand the operator $c^2/(2mc^2-V)^2$ in powers of $V/2mc^2$,

$$\begin{aligned} E^{(1)} &= -E^{(0)}\langle\,\psi^{(0)}\,|\,(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{1}{4m^2c^2}\left[1+\frac{V}{mc^2}+\ldots\right](\boldsymbol{\sigma}\cdot\mathbf{p})\,|\,\psi^{(0)}\,\rangle \\ &= -E^{(0)}\langle\,\psi^{(0)}\,|\,(\boldsymbol{\sigma}\cdot\mathbf{p})(\boldsymbol{\sigma}\cdot\mathbf{p})\,|\,\psi^{(0)}\,\rangle/4m^2c^2+\ldots, \end{aligned} \tag{18.25}$$

we see that the first-order energy provides the term that was missing from the ZORA Hamiltonian expansion in (18.3). This missing term contains the rest of the relativistic correction of $\mathcal{O}(c^{-2})$.

Assuming that the wave functions in each order are orthogonal, the second-order energy can be written as

$$\begin{aligned} E^{(2)} = \frac{\left(E^{(1)}\right)^2}{E^{(0)}} &+ \left(E^{(0)}\right)^2\langle\,\psi^{(0)}\,|\,(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2-V)^3}(\boldsymbol{\sigma}\cdot\mathbf{p})\,|\,\psi^{(0)}\,\rangle \\ &- E^{(0)}\langle\,\psi^{(0)}\,|\,(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2-V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\,|\,\psi^{(1)}\,\rangle \end{aligned} \tag{18.26}$$

where we have used (18.24) to eliminate one of the integrals. We can expand the first-order wave function in the zeroth-order states as usual to arrive at an expression for the last term, with the result for the second-order energy

$$\begin{aligned} E^{(2)} = \frac{\left(E^{(1)}\right)^2}{E^{(0)}} &+ \left(E^{(0)}\right)^2\langle\,\psi^{(0)}\,|\,(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2-V)^3}(\boldsymbol{\sigma}\cdot\mathbf{p})\,|\,\psi^{(0)}\,\rangle \\ &- E^{(0)}\sum_k\left|\langle\,\psi^{(0)}\,|\,(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2-V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\,|\,\psi_k^{(0)}\,\rangle\right|^2/(E_k^{(0)}-E^{(0)}). \end{aligned} \tag{18.27}$$

The first term in this expression is negative and the second and third terms are positive. It is consequently difficult to say a priori whether the second-order energy is positive or negative.

The magnitude of the perturbation corrections is related to the magnitude of the integral $\langle\,\psi^{(0)}\,|\,(\boldsymbol{\sigma}\cdot\mathbf{p})c^2/(2mc^2-V)^{k+1}(\boldsymbol{\sigma}\cdot\mathbf{p})\,|\,\psi^{(0)}\,\rangle$. The value of $2mc^2-V$ will always be greater than $2mc^2$, and thus the integral will always be less than $\langle\,\psi^{(0)}\,|\,\hat{T}\,|\,\psi^{(0)}\,\rangle/(2mc^2)^k$. The terms in the perturbation will be less than $(E/2mc^2)^k\langle\hat{T}\rangle$, which will at least converge for $|E|<2mc^2$.

One final approach to the development of a regular perturbation series is worth examining, and that is the renormalization perturbation theory of Sadlej et al. (1994, 1995). This approach follows the line of direct perturbation theory, but the formal perturbation is different. Writing the Dirac equation with $1/c$ extracted from the small component according to (17.60) and replacing $1/c^2$ in the metric with μ^2, we have

$$\begin{pmatrix} V & (\boldsymbol{\sigma}\cdot\mathbf{p}) \\ (\boldsymbol{\sigma}\cdot\mathbf{p}) & V/c^2-2m \end{pmatrix}\begin{pmatrix} \psi^L \\ \phi^S \end{pmatrix} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mu^2\mathbf{I}_2 \end{pmatrix}\begin{pmatrix} \psi^L \\ \phi^S \end{pmatrix}E. \tag{18.28}$$

Now, making a perturbation series in μ^2 with the physical value $\mu = 1/c$, the zeroth-order equation is

$$\begin{pmatrix} V & (\boldsymbol{\sigma}\cdot\mathbf{p}) \\ (\boldsymbol{\sigma}\cdot\mathbf{p}) & V/c^2 - 2m \end{pmatrix} \begin{pmatrix} \psi^L \\ \phi^S \end{pmatrix} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{0}_2 \end{pmatrix} \begin{pmatrix} \psi^L \\ \phi^S \end{pmatrix} E. \tag{18.29}$$

This is similar to the Lévy-Leblond equation, in that there is a zero in the metric and the negative-energy solutions would be pathological. From the second row in the matrix equation we have

$$\phi^S = \frac{c^2}{2mc^2 - V}(\boldsymbol{\sigma}\cdot\mathbf{p})\,\psi^L \tag{18.30}$$

which on substitution into the first row gives the ZORA equation.

To consider the perturbation expansion we write formally

$$\hat{\mathcal{H}}'\psi = (\hat{\mathcal{G}}_0' + \mu^2\hat{\mathcal{G}}_2')\psi E \tag{18.31}$$

where the primes indicate the extraction of $1/c$ from the original Dirac Hamiltonian, as in section 17.5. Making the expansion in even powers of μ, the lowest-order perturbation equation is

$$\hat{\mathcal{H}}'\psi^{(2)} = \hat{\mathcal{G}}_0'\psi^{(0)}E^{(2)} + \hat{\mathcal{G}}_0'\psi^{(2)}E^{(0)} + \hat{\mathcal{G}}_2'\psi^{(0)}E^{(0)}. \tag{18.32}$$

Premultiplying by $\psi^{(0)\dagger}$ and integrating,

$$E^{(2)} = -\langle\psi^{(0)}|\hat{\mathcal{G}}_2'|\psi^{(0)}\rangle E^{(0)}, \tag{18.33}$$

which on substitution of (18.30) for the zeroth-order wave function gives us, as we would expect, the FORA energy. Expressions for higher orders are easily derived.

18.3 Nonperturbative Improvements of the ZORA Equation

The perturbation series discussed so far start from the ZORA Hamiltonian as the zeroth-order approximation. We already know that this Hamiltonian has relativistic corrections in it but is missing some terms of $\mathcal{O}(c^{-2})$. These terms in fact come from the correction to the metric, as we saw in the previous section. Is there a way to obtain an improved zeroth-order approximation? In this section, we consider two possible improvements.

Our line of development here is to return to the exact Foldy–Wouthuysen transformation, presented in section 16.1 and used above to develop a perturbation series. There, we chose $1/(2mc^2 - V)$ as the perturbation and expanded both the square root and the inverse powers. Here, we choose the perturbation parameter as $E/(2mc^2 - V)$

and essentially expand only the inverse powers. The zeroth-order approximation corresponds to setting E to zero in the operator $\hat{\mathcal{X}}$,

$$\hat{\mathcal{X}}_0 = (2mc^2 - V)^{-1} c(\boldsymbol{\sigma}\cdot\mathbf{p}). \tag{18.34}$$

The zeroth-order Hamiltonian in this approximation is

$$\begin{aligned}\hat{\mathcal{H}}_0 &= \frac{1}{\sqrt{1+\hat{\mathcal{X}}_0{}^{\dagger}\hat{\mathcal{X}}_0}}\Big[V + c(\boldsymbol{\sigma}\cdot\mathbf{p})\hat{\mathcal{X}}_0 + c\hat{\mathcal{X}}_0{}^{\dagger}(\boldsymbol{\sigma}\cdot\mathbf{p}) + \hat{\mathcal{X}}_0{}^{\dagger}(V - 2mc^2)\hat{\mathcal{X}}_0\Big]\frac{1}{\sqrt{1+\hat{\mathcal{X}}_0{}^{\dagger}\hat{\mathcal{X}}_0}} \\ &= \left[1 + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2-V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\right]^{-1/2}\left[V + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{2mc^2-V}(\boldsymbol{\sigma}\cdot\mathbf{p})\right] \\ &\quad\times\left[1 + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2-V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\right]^{-1/2}.\end{aligned} \tag{18.35}$$

The operator in the second bracket is the ZORA Hamiltonian, and it is sandwiched by normalization operators. If we expand these operators as we did above, we get the FORA Hamiltonian as the first term. The higher terms differ, however, because the final energy in the previous series must be the Dirac energy, whereas here it is the energy for the approximate Hamiltonian. Inclusion of the normalization terms corresponds to a resummation of certain parts of the ZORA perturbation series to infinite order, and the name coined by Dyall and van Lenthe (1999) is IORA—infinite-order regular approximation.

The IORA Hamiltonian still presents some problems, because there are momentum operators in the inverse square roots of the normalization operators. Defining a modified wave function,

$$\phi = \left[1 + \hat{\mathcal{X}}_0{}^{\dagger}\hat{\mathcal{X}}_0\right]^{-1/2}\psi, \tag{18.36}$$

the zeroth-order approximate equation for the positive energy states becomes

$$\hat{\mathcal{H}}^{\mathrm{ZORA}}\phi = E\left[1 + \hat{\mathcal{X}}_0{}^{\dagger}\hat{\mathcal{X}}_0\right]\phi, \tag{18.37}$$

which is an equation with a modified metric. Now all the momentum operators are in the numerator, and it is as easy to solve this equation as the ZORA equation. The extra term in the metric is the operator that appears in the FORA Hamiltonian, and if we were to expand this equation in a perturbation series we would see that the FORA correction is included. The IORA energy is therefore correct to $\mathcal{O}(c^{-2})$. Another way of looking at this equation is that we have taken the lowest-order perturbation correction from (18.22) and included it in the zeroth-order Hamiltonian. Since the term is linear in the energy this presents no problem. Before continuing, it should be noted that it is not trivial to obtain ψ from ϕ because of the square root.

The spectrum of the IORA Hamiltonian can be derived by considering the asymptotic region for large r, where $V \to 0$ and the IORA equation goes to

$$\hat{T}\phi = E\left[1 + \frac{\hat{T}}{2mc^2}\right]\phi. \tag{18.38}$$

Rearranging this equation we find that the wave function is $\phi = Ne^{\mathbf{k}\cdot\mathbf{r}}$, where N is a normalization constant and k is given by

$$k^2 = -E\frac{4m^2c^2}{2mc^2 - E}. \tag{18.39}$$

For $E < 0$, k^2 is positive, k is real, and choosing $k < 0$ we get wave functions that are exponentially decaying, that is, they are bound-state wave functions.[1] For $0 < E < +2mc^2$, k^2 is negative, k is imaginary, and the wave functions are oscillatory. This is the continuum region. For $E > +2mc^2$, k^2 is again positive and this region can only contain bound states. We will see later that this region is empty. The curious result is an inversion of the Dirac spectrum, which has a region of bound states between 0 and $-2mc^2$, flanked on either side by a region of free or continuum states. The IORA spectrum has a region of free or continuum states between 0 and $+2mc^2$ with bound-state regions on either side.

What is the relation between the IORA energies and the ZORA and Dirac energies? There is a correspondence at $E = 0$ and we expect that the correspondence continues in the vicinity of this point. Unlike the ZORA equation, we cannot perform a scaling to obtain a relation with the Dirac ESC equation, and therefore we cannot obtain a direct relation with the Dirac eigenvalues. What we can do is to make use of the Rayleigh quotient for (18.37) to obtain a relation between the ZORA and IORA eigenvalues, since ZORA and IORA have the same Hamiltonian but a different metric. For an arbitrary wave function ψ,

$$E(\psi) = \frac{\langle\psi|\hat{\mathcal{H}}^{\mathrm{ZORA}}|\psi\rangle}{\langle\psi|1 + \hat{\mathcal{X}}_0^{\dagger}\hat{\mathcal{X}}_0|\psi\rangle}. \tag{18.40}$$

If we take ψ to be the IORA wave function but normalize it on the unit metric (which is the ZORA metric), the energy $E(\psi)$ is E^{IORA}, and the denominator is greater than 1. The numerator is the expectation of the ZORA Hamiltonian for a wave function that is not the ground state wave function and therefore $\langle\psi|\hat{\mathcal{H}}^{\mathrm{ZORA}}|\psi\rangle > E^{\mathrm{ZORA}}$. Since both energies are negative this means that $E^{\mathrm{IORA}} > E^{\mathrm{ZORA}}$, and the IORA energy is bounded below by the ZORA energy.

We may also evaluate the quotient with the ZORA wave function. In that case, the numerator is the ZORA energy, and the denominator contains the term we evaluated for the FORA energy. We are interested in the entire range of eigenvalues, so we must use

---

[1] Choosing $k \geq 0$ results in unphysical solutions.

the absolute value of this term, $|-E^D/2mc^2 - E^D|$. The denominator of the Rayleigh quotient can then be expressed as

$$\begin{aligned} 1+\left|\frac{-E^D}{2mc^2+E^D}\right| &= 1-\frac{E^D}{2mc^2+E^D} = \frac{2mc^2}{2mc^2+E^D}, \quad 0 > E^D > -2mc^2; \\ &= 1+\frac{E^D}{2mc^2+E^D} = \frac{2mc^2+2E^D}{2mc^2+E^D}, \quad E^D < -2mc^2 \text{ and } E^D > 0. \end{aligned} \tag{18.41}$$

With the expression for the ZORA energy in terms of the Dirac energy from (18.9), the Rayleigh quotient is

$$\begin{aligned} E(\psi^{\text{ZORA}}) &= \frac{2mc^2E^D}{2mc^2+E^D}\cdot\frac{2mc^2+E^D}{2mc^2} = E^D, \quad 0 > E^D > -2mc^2; \\ &= \frac{2mc^2E^D}{2mc^2+E^D}\cdot\frac{2mc^2+E^D}{2mc^2+2E^D} = \frac{mc^2E^D}{mc^2+E^D}, \quad E^D < -2mc^2, E^D > 0. \end{aligned} \tag{18.42}$$

Since the IORA equation is variational and the ZORA wave function is not the ground state wave function for the IORA Hamiltonian, the quotient must be an upper bound to the IORA energy for the bound states. We can summarize as follows:

$$E^{\text{D}} > E^{\text{IORA}} > E^{\text{ZORA}}. \tag{18.43}$$

For the states with $E^D > 0$, the spectrum is compressed to the range $0 < E^{\text{IORA}} < +mc^2$, half the range of the corresponding states for ZORA. The negative-energy states of the Dirac spectrum, $E^D < -2mc^2$, lie in the range $+mc^2 < E^{\text{IORA}} < +2mc^2$, and there are no states that lie in the region above $+2mc^2$.

To improve on the IORA energies, a perturbation series may be developed based on the IORA equation. This is most simply done by comparing with the perturbation series for ZORA, (18.22). As we mentioned above, IORA corresponds to putting the term that is linear in the energy into the metric to make it part of the zeroth-order equation. Because of this, the first-order energy and the first-order wave function are zero. The second-order energy is

$$E^{(2)} = \left(E^{(0)}\right)^2 \langle \phi^{(0)} | (\boldsymbol{\sigma}\cdot\mathbf{p}) \frac{c^2}{(2mc^2-V)^3} (\boldsymbol{\sigma}\cdot\mathbf{p}) | \phi^{(0)} \rangle, \tag{18.44}$$

which is positive and should bring the IORA energy closer to the Dirac energy. Other improvements have been developed by Filatov and Cremer (2003).

IORA has one practical problem compared with ZORA. For a hydrogenic atom, where the potential is fixed, the metric is also fixed, but for a general potential, such as a Hartree–Fock or Kohn–Sham potential, the metric will change as the density changes. This means a reorthogonalization in each iteration. One way to circumvent this problem is to reevaluate the potential only at certain intervals: for instance, one might start by solving the ZORA equation and then use the converged potential to construct a new metric, then converge this approximation and repeat until the potential does not change. Another option is to use the potential of the nuclei or the superposition of atomic Coulomb potentials in the metric, much like van Wüllen proposed to solve the gauge problem. This would also have the advantage of solving the gauge problem for IORA, which is still present at $\mathcal{O}(c^{-4})$ but is less serious than in ZORA.

The fact that the Rayleigh quotient for the ZORA wave function yields the Dirac eigenvalue logically leads into another, even simpler, approach than IORA to the improvement of ZORA, and one that preceded IORA: the scaled ZORA method (van Lenthe et al. 1994). It is the renormalization terms in the IORA Hamiltonian that present the main difficulty. But for any given function, renormalization may be achieved by a simple scaling. We therefore make the approximations

$$\begin{gathered}\left[1+\hat{\mathcal{X}}_0{}^{\dagger}\hat{\mathcal{X}}_0\right]^{-1/2}\psi_i \simeq \left[1+\langle\,\psi_i\,|\,\hat{\mathcal{X}}_0{}^{\dagger}\hat{\mathcal{X}}_0\,|\,\psi_i\,\rangle\right]^{-1/2}\psi_i,\\ \psi_i{}^{\dagger}\hat{\mathcal{X}}_0{}^{\dagger}\hat{\mathcal{X}}_0\psi_i \simeq \langle\,\psi_i\,|\,\hat{\mathcal{X}}_0{}^{\dagger}\hat{\mathcal{X}}_0\,|\,\psi_i\,\rangle\psi_i{}^{\dagger}\psi_i.\end{gathered} \tag{18.45}$$

The renormalization terms commute with the ZORA Hamiltonian and simply give a scaling factor. The IORA equation now becomes

$$\hat{\mathcal{H}}^{\text{ZORA}}\psi_i = E_i\left[1+\langle\,\psi_i\,|\,\hat{\mathcal{X}}_0{}^{\dagger}\hat{\mathcal{X}}_0\,|\,\psi_i\,\rangle\right]\psi_i, \tag{18.46}$$

which is the ZORA equation with a scaled energy. It is therefore only necessary to solve the ZORA equation and scale the eigenvalues to get the energies.

The scaled ZORA energies for a one-electron system can easily be evaluated. We already know that the integral in the metric is equal to $-E^{\text{D}}/(2mc^2+E^{D})$. The scaling factor is therefore

$$1+\langle\,\psi_i\,|\,\hat{\mathcal{X}}_0{}^{\dagger}\hat{\mathcal{X}}_0\,|\,\psi_i\,\rangle = \frac{2mc^2}{2mc^2+E^{\text{D}}}. \tag{18.47}$$

Since the ZORA eigenvalue is $2mc^2E^{\text{D}}/(2mc^2+E^{\text{D}})$, the scaled ZORA eigenvalue E_i must be simply E^{D}, the Dirac eigenvalue, as we suggested above.[2] Due to this fact, the scaled ZORA energy is gauge-independent, but only for a hydrogenic system. For other systems, there will be a small gauge dependence.

Finally, we mention the regular modified Dirac equation of Sadlej and Snijders (1994), which is exact and follows the lines of chapter 15. The regular transformation of the small component,

$$\phi^L = (2mc^2-V)^{-1}c(\boldsymbol{\sigma}\cdot\mathbf{p})\psi^S, \tag{18.48}$$

---

[2] This relation applies only to the Dirac bound states, which have energies $0 > E^{\text{D}} > -2mc^2$. For continuum states, the eigenvalue is given by (18.43).

yields the regular modified Dirac equation,

$$\begin{pmatrix} V & \hat{\mathcal{H}}^{\text{ZORA}} \\ \hat{\mathcal{H}}^{\text{ZORA}} & -\hat{\mathcal{H}}^{\text{ZORA}} \end{pmatrix} \begin{pmatrix} \psi^L \\ \phi^L \end{pmatrix} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \hat{\mathcal{X}}_0{}^\dagger \hat{\mathcal{X}}_0 \end{pmatrix} \begin{pmatrix} \psi^L \\ \phi^L \end{pmatrix} E. \tag{18.49}$$

This equation is exact just as the modified Dirac equation of chapter 15 is exact. It can also be separated into spin-free and spin-dependent terms, but now the separation must be done in both the Hamiltonian and the metric. Visscher and van Lenthe (1999) have shown that the spin separation gives different results for the two modified equations, and therefore the spin separation is not unique. This regular modified Dirac equation can be used in renormalization perturbation theory, with ZORA as the zeroth-order Hamiltonian.

18.4 Many-Electron Systems

While one-electron systems provide a good formal testing ground for an approximate theory, for quantum chemistry we need a theory that encompasses many-electron systems. Formally, we can treat the regular approximations as a Foldy–Wouthuysen transformation with a particular choice of $\hat{\mathcal{X}}$, and then we can write the transformed two-electron operator as

$$\hat{g}_{ij}^{\text{RA}} = \hat{\mathcal{O}}_i^{-1} \hat{\mathcal{O}}_j^{-1} \left(g_{ij} + \hat{\mathcal{X}}_i{}^\dagger g_{ij} \hat{\mathcal{X}}_i + \hat{\mathcal{X}}_j{}^\dagger g_{ij} \hat{\mathcal{X}}_j + \hat{\mathcal{X}}_i{}^\dagger \hat{\mathcal{X}}_j{}^\dagger g_{ij} \hat{\mathcal{X}}_j \hat{\mathcal{X}}_i \right) \hat{\mathcal{O}}_j^{-1} \hat{\mathcal{O}}_i^{-1} \tag{18.50}$$

where the normalization operator $\hat{\mathcal{O}}$ is

$$\hat{\mathcal{O}} = \sqrt{1 + \hat{\mathcal{X}}^\dagger \hat{\mathcal{X}}}. \tag{18.51}$$

The operator $\hat{\mathcal{X}}$ is given in (18.34) which we repeat here without the subscript:

$$\hat{\mathcal{X}} = (2mc^2 - V)^{-1} c(\boldsymbol{\sigma} \cdot \mathbf{p}). \tag{18.52}$$

This operator is very similar to the $\hat{\mathcal{R}}_2$ operator of the free-particle Foldy–Wouthuysen transformation: it has a regularizing factor multiplied by $(\boldsymbol{\sigma} \cdot \mathbf{p})$. We can derive spin-free and spin-dependent operators from the Coulomb, Gaunt, and Breit interactions in an entirely analogous fashion to the Foldy–Wouthuysen transformation. As an example, the two-electron spin–orbit interaction in the regular approximation is

$$\hat{g}^{\text{RA,so}} = \sum_{i \neq j} \hat{\mathcal{O}}_i^{-1} \hat{\mathcal{O}}_j^{-1} \; \hbar \boldsymbol{\sigma}_i \cdot \nabla_i \left(\frac{1}{r_{ij}} \frac{c^2}{(2mc^2 - V_i)^2} \right) \times \mathbf{p}_i \; \hat{\mathcal{O}}_j^{-1} \hat{\mathcal{O}}_i^{-1}. \tag{18.53}$$

This operator is complicated by the inclusion of the regularization terms, which must be differentiated in addition to $1/r_{ij}$.

The biggest problem that we now face is that the potential V is in the denominator, and we have to decide what to include in V. Obviously, including $g_{ij} = 1/r_{ij}$ would make the two-electron terms impossibly complicated, even if we were tabulating V on a grid. We are therefore forced to make some kind of approximation.

We could of course proceed as we did for the Douglas–Kroll transformation and use a transformation that depends only on the nuclear potential. This would remove the awkwardness of having $1/r_{ij}$ in the denominator, but we still have the product of $c^2/(2mc^2 - V_i)^2$ with $1/r_{ij}$ to deal with. If we are only interested in spin-free relativistic effects, we could neglect the transformation of the electron–electron interaction, as we did in the Douglas–Kroll–Hess approximation. This approximation yields the Hamiltonian

$$\hat{\mathcal{H}}^{\text{RA}} = \sum_i \hat{h}_i^{\text{RA}} + \tfrac{1}{2}\sum_{i \neq j} g_{ij}, \tag{18.54}$$

where

$$\hat{h}_i^{\text{RA}} = \hat{\mathcal{O}}_i^{-1}\left[V_i^{\text{nuc}} + (\boldsymbol{\sigma}_i \cdot \mathbf{p}_i)\frac{c^2}{2mc^2 - V_i^{\text{nuc}}}(\boldsymbol{\sigma}_i \cdot \mathbf{p}_i)\right]\hat{\mathcal{O}}_i^{-1} \tag{18.55}$$

and we can choose $\hat{\mathcal{O}}_i$ for the particular regular approximation we want to use—ZORA, scaled ZORA, or IORA.

If we wish to use IORA in this approximation and absorb the normalization factors into the wave function, the two-electron operator is no longer g_{ij} but

$$\hat{g}_{ij}^{\text{modIORA}} = \hat{\mathcal{O}}_i \hat{\mathcal{O}}_j g_{ij} \hat{\mathcal{O}}_j \hat{\mathcal{O}}_i. \tag{18.56}$$

This operator is not so convenient. We cannot simply ignore the normalization operators, because they change the magnitude of the electron–electron interaction, which accounts for some of the relativistic effects. It would then be necessary to abandon the approximation and include the effects arising from the transformation of g_{ij} with $\hat{\mathcal{X}}$. The alternative is to evaluate the normalization operators in a matrix representation, which is relatively easy since they are one-electron operators.

If we want to construct a more rigorous approach to the electron–electron interaction we need to address the inclusion of at least part of the electron–electron interaction in the denominator. In a many-electron system, using just the nuclear potential for V in the denominator neglects the screening from the electrons, with the result that the denominator is too large and the regularized operators are too small. The energies are consequently too low. This can be seen by setting V_1 to zero in the last term of (18.12) which is positive.

The question is how do we include the two-electron contributions to V? The operator $\hat{\mathcal{X}}$ was derived by neglecting the energy in the exact expression for the small component in terms of the large component. Since this expression was derived from the Dirac equation, we should turn to a mean-field Dirac equation to derive an expression that includes the two-electron contributions. Because much of the development of regular approximations has been made in the context of density functional theory, we will consider both the Dirac–Hartree–Fock (DHF) equation and the Dirac–Kohn–Sham (DKS) equation as starting points. We write these equations in unrestricted spinor form as

$$\text{DHF}: \begin{pmatrix} V^{\text{nuc}}+J^{\text{tot}}-\hat{K}^{LL} & c(\boldsymbol{\sigma}\cdot\mathbf{p})-\hat{K}^{LS} \\ c(\boldsymbol{\sigma}\cdot\mathbf{p})-\hat{K}^{SL} & V^{\text{nuc}}+J^{\text{tot}}-\hat{K}^{SS}-2mc^2 \end{pmatrix}\begin{pmatrix}\psi^L\\ \psi^S\end{pmatrix}=E\begin{pmatrix}\psi^L\\ \psi^S\end{pmatrix} \qquad (18.57)$$

$$\text{DKS}: \begin{pmatrix} V^{\text{nuc}}+J^{\text{tot}}+V^{\text{xc}} & c(\boldsymbol{\sigma}\cdot\mathbf{p}) \\ c(\boldsymbol{\sigma}\cdot\mathbf{p}) & V^{\text{nuc}}+J^{\text{tot}}+V^{\text{xc}}-2mc^2 \end{pmatrix}\begin{pmatrix}\psi^L\\ \psi^S\end{pmatrix}=E\begin{pmatrix}\psi^L\\ \psi^S\end{pmatrix} \qquad (18.58)$$

where V^{nuc} is the nuclear potential, $J^{\text{tot}}=J^{LL}+J^{SS}$ is the electronic Coulomb potential and V^{xc} is the exchange-correlation potential. To simplify the expressions a little, we write the total Coulomb potential as

$$V^{\text{Coul}}=V^{\text{nuc}}+J^{\text{tot}}. \qquad (18.59)$$

The potential in the Dirac–Kohn–Sham equation is diagonal, so we can write

$$V^{\text{DKS}}=V^{\text{nuc}}+J^{\text{tot}}+V^{\text{xc}}. \qquad (18.60)$$

The Dirac–Kohn–Sham equation reduces to

$$\begin{pmatrix} V^{\text{DKS}} & c(\boldsymbol{\sigma}\cdot\mathbf{p}) \\ c(\boldsymbol{\sigma}\cdot\mathbf{p}) & V^{\text{DKS}}-2mc^2 \end{pmatrix}\begin{pmatrix}\psi^L\\ \psi^S\end{pmatrix}=E\begin{pmatrix}\psi^L\\ \psi^S\end{pmatrix}, \qquad (18.61)$$

which is of the same form as the one-particle Dirac equation. This suggests that we can simply replace V with V^{DKS} in the one-electron ZORA equation. Unfortunately, it is not quite so simple: there are some hidden complexities in this substitution that we will address below. The exchange potential in the Dirac–Hartree–Fock equation is nonlocal and nondiagonal, but we can still eliminate the small component and make a regular approximation by neglecting E. In this way, we arrive at the operator $\hat{\mathcal{X}}_0$ from the approximate Foldy–Wouthuysen transformation for both equations:

$$\hat{\mathcal{X}}_0^{\text{DHF}}=\left[2mc^2-V^{\text{Coul}}+\hat{K}^{SS}\right]^{-1}\left[c(\boldsymbol{\sigma}\cdot\mathbf{p})-\hat{K}^{SL}\right]; \qquad (18.62)$$

$$\hat{\mathcal{X}}_0^{\text{DKS}}=\left[2mc^2-V^{\text{DKS}}\right]^{-1}c(\boldsymbol{\sigma}\cdot\mathbf{p}). \qquad (18.63)$$

With these definitions we can write the transformed one-electron Fock operators in analogy to (18.35) as

$$\hat{f}^{\text{RegDHF}}=\left[\hat{\mathcal{O}}_0^{\text{DHF}}\right]^{-1}\Big[V^{\text{Coul}}-\hat{K}^{LL}+\left(c(\boldsymbol{\sigma}\cdot\mathbf{p})-\hat{K}^{SL\dagger}\right)$$
$$\times\left(2mc^2-V^{\text{Coul}}+\hat{K}^{SS}\right)^{-1}\left(c(\boldsymbol{\sigma}\cdot\mathbf{p})-\hat{K}^{SL}\right)\Big]\left[\hat{\mathcal{O}}_0^{\text{DHF}}\right]^{-1} \qquad (18.64)$$

$$\hat{f}^{\text{RegDKS}} = \left[\hat{\mathcal{O}}_0^{\text{DKS}}\right]^{-1} \left[V^{\text{DKS}} + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{2mc^2 - V^{\text{DKS}}}(\boldsymbol{\sigma}\cdot\mathbf{p})\right]\left[\hat{\mathcal{O}}_0^{\text{DKS}}\right]^{-1} \tag{18.65}$$

where $\hat{\mathcal{O}}_0 = \sqrt{1 + \hat{\mathcal{X}}_0{}^\dagger \hat{\mathcal{X}}_0}$ is the normalization term in each case. These equations as they stand represent the IORA approximation: scaled ZORA is obtained by replacing $\hat{\mathcal{X}}_0{}^\dagger \hat{\mathcal{X}}_0$ by its expectation value, and ZORA is obtained by replacing $\hat{\mathcal{O}}_0$ with 1.

So far we have made no approximations to the Coulomb and exchange terms in the potential, which are still expressed in terms of the large and small components. In an actual calculation we would not have the large and small components, only the approximate transformed wave function. The potentials must therefore be expressed in terms of the transformed wave functions and the transformed operators, taking into account that these are not exact.

We first consider the relationships between the original and transformed wave functions and the effects of the approximations. The relationships for a general transformation defined by (16.8), which is unitary by construction, are

$$\begin{pmatrix} \psi^{\text{FW}} \\ \delta^{\text{FW}} \end{pmatrix} = \begin{pmatrix} \hat{\mathcal{O}}_+^{-1}\left[\psi^L + \hat{\mathcal{X}}^\dagger \psi^S\right] \\ \hat{\mathcal{O}}_-^{-1}\left[-\hat{\mathcal{X}}\psi^L + \psi^S\right] \end{pmatrix}; \quad \begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} = \begin{pmatrix} \hat{\mathcal{O}}_+^{-1}\psi^{\text{FW}} - \hat{\mathcal{X}}^\dagger \hat{\mathcal{O}}_-^{-1}\delta^{\text{FW}} \\ \hat{\mathcal{X}}\hat{\mathcal{O}}_+^{-1}\psi^{\text{FW}} + \hat{\mathcal{O}}_-\delta^{\text{FW}} \end{pmatrix} \tag{18.66}$$

where $\hat{\mathcal{O}}_+ = \sqrt{1 + \hat{\mathcal{X}}^\dagger \hat{\mathcal{X}}} \equiv \hat{\mathcal{O}}$ and $\hat{\mathcal{O}}_- = \sqrt{1 + \hat{\mathcal{X}}\hat{\mathcal{X}}^\dagger}$. We will henceforth use $\hat{\mathcal{X}}$ for any approximation, but when a regular approximation is intended, one of the $\hat{\mathcal{X}}_0$ expressions given above should be substituted.

The approximation we make in the transformation process is to neglect both the off-diagonal terms in the transformed Hamiltonian and the residual coupling in the wave function. At the same time as neglecting δ^{FW}, we also renormalize ψ^{FW}. This will introduce a scaling factor, which we will absorb into the definition of ψ^{FW} and $\hat{\mathcal{O}}$. With these approximations, the relations between the wave functions become

$$\psi^{\text{FW}} \simeq \hat{\mathcal{O}}^{-1}\left[\psi^L + \hat{\mathcal{X}}^\dagger \psi^S\right]; \quad \begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} \simeq \begin{pmatrix} \hat{\mathcal{O}}^{-1}\psi^{\text{FW}} \\ \hat{\mathcal{X}}\hat{\mathcal{O}}^{-1}\psi^{\text{FW}} \end{pmatrix}. \tag{18.67}$$

The second of these relations also implies that we are setting $\psi^S = \hat{\mathcal{X}}\psi^L$. The effect of this approximation can be seen if we construct the density from the transformed wave function,

$$\begin{aligned} \rho^{\text{FW}} = \psi^{\text{FW}\dagger}\psi^{\text{FW}} &= \left[\psi^{L\dagger} + \psi^{S\dagger}\hat{\mathcal{X}}\right]\left[1 + \hat{\mathcal{X}}^\dagger \hat{\mathcal{X}}\right]^{-1}\left[\psi^L + \hat{\mathcal{X}}^\dagger \psi^S\right] \\ &= \left[\psi^{L\dagger} + \psi^{S\dagger}\hat{\mathcal{X}}\right]\left[1 + \hat{\mathcal{X}}^\dagger \hat{\mathcal{X}}\right]^{-1}\left[\psi^L + \hat{\mathcal{X}}^\dagger \hat{\mathcal{X}}\psi^L\right] \\ &= \left[\psi^{L\dagger} + \psi^{S\dagger}\hat{\mathcal{X}}\right]\psi^L \\ &= \psi^{L\dagger}\psi^L + \psi^{S\dagger}\psi^S = \rho \end{aligned} \tag{18.68}$$

where we have used $\psi^S = \hat{\mathcal{X}}\psi^L$ twice. The exact transformed density is $\rho = \psi^{\mathrm{FW}\dagger}\psi^{\mathrm{FW}} + \delta^{\mathrm{FW}\dagger}\delta^{\mathrm{FW}}$, so the approximations in (18.67) correspond to the neglect of the second term in the density and the renormalization of the first.

We turn now to consider the potentials. It would be tempting to apply (18.68) immediately to the electronic Coulomb potential, since it can be expressed in terms of the integral of the density with the electron–electron interaction. We must be careful because, as we saw in (18.50), the transformation introduces operators that do not necessarily commute with the interaction. The electronic Coulomb potential is

$$
\begin{aligned}
J^{\mathrm{tot}} &= \sum_j \int d\mathbf{r} \left[\psi_j^{L\dagger}\, g\, \psi_j^L + \psi_j^{S\dagger}\, g\, \psi_j^S \right] \\
&= \sum_j \int d\mathbf{r}_j\, \psi_j^{\mathrm{FW}} \hat{\mathcal{O}}_j^{-1} \left[g + \hat{\mathcal{X}}_j^\dagger\, g\, \hat{\mathcal{X}}_j \right] \hat{\mathcal{O}}_j^{-1} \psi_j^{\mathrm{FW}}.
\end{aligned}
\tag{18.69}
$$

If we assume that the electron–electron interaction g commutes with $\hat{\mathcal{X}}$ then we may write the Coulomb potential as

$$
J^{\mathrm{tot,sf}} = \sum_j \int d\mathbf{r}\, g\, \psi_j^{\mathrm{FW}} \hat{\mathcal{O}}_j^{-1} \left[1 + \hat{\mathcal{X}}_j^\dagger \hat{\mathcal{X}}_j \right] \hat{\mathcal{O}}_j^{-1} \psi_j^{\mathrm{FW}} = \sum_j \int d\mathbf{r}_j\, g\, \rho_j^{\mathrm{FW}}. \tag{18.70}
$$

The neglect of the commutator of the transformation with the two-electron operator is precisely what we did above to obtain a spin-free relativistic equation, and is therefore only valid in a spin-free formalism.

The problem with this approximation is that we have only neglected the commutator for one electron coordinate. The commutator for the other electron coordinate is embedded in the mean-field equation, and it still gives rise to a spin-free and a spin–orbit operator. Neglecting for the moment the normalization terms, if we use the partitioning in (18.5) and substitute $V^{\mathrm{Coul}} = V^{\mathrm{nuc}} + J^{\mathrm{tot,sf}}$ for V, we get

$$
\begin{aligned}
\hat{f}^{\mathrm{Coul}} &= V^{\mathrm{nuc}} + J^{\mathrm{tot,sf}} + \frac{2mc^2}{2mc^2 - V^{\mathrm{Coul}}} \hat{T} \\
&\quad + \frac{c^2}{\left(2mc^2 - V^{\mathrm{Coul}}\right)^2} \left[\mathbf{p}\left(V^{\mathrm{nuc}} + J^{\mathrm{tot,sf}}\right) \cdot \mathbf{p} + \hbar\boldsymbol{\sigma} \cdot \nabla\left(V^{\mathrm{nuc}} + J^{\mathrm{tot,sf}}\right) \times \mathbf{p} \right],
\end{aligned}
\tag{18.71}
$$

where we have expanded out only the numerator terms involving V^{Coul}. Even with the spin-free approximation to the Coulomb potential, the last term includes a contribution from the two-electron spin–orbit interaction and from a spin-free relativistic correction to the potential. This term includes only half of the contribution from these operators. The other half was discarded when we neglected the commutator in the approximation to J^{tot}.

We see then that the contributions from the regularized two-electron relativistic operators to the Fock operator are split between two terms. If we want the screening contribution in the denominators, we cannot write the Coulomb potential as an integral over the density. What we must do is construct a density that includes the $\hat{\mathcal{X}}$ operators, and not approximate it. For a spin-free calculation this density can be given in terms of the modified wave function of (18.36) as

$$\tilde{\rho}^{\mathrm{FW}} = \sum_j \phi_j^{\mathrm{FW}\dagger}\phi_j^{\mathrm{FW}} + (\mathbf{p}\phi_j^{\mathrm{FW}})^\dagger \cdot \frac{c^2}{(2mc^2 - V_j)^2}(\mathbf{p}\phi_j^{\mathrm{FW}}). \tag{18.72}$$

For the spin–orbit interaction we must add a spin-density term to this expression. In essence, we are reconstructing the small component and calculating the sum of the large- and small-component densities. In a density-functional calculation, the density that we must construct is therefore no longer simply the density from the transformed wave function.

In the Dirac–Hartree–Fock case, we have to consider in addition the inclusion of the exchange operators in the transformation. It can quite reasonably be argued that, given the other approximation, neglecting them is reasonable. Since the small component is $\mathcal{O}(c^{-1})$, the relative size of $\hat{K}^{SS}$ in the denominator term is $\mathcal{O}(c^{-4})$ and therefore negligible, especially since the IORA approximation itself has errors of $\mathcal{O}(c^{-4})$. Likewise, $\hat{K}^{LS}$ is $\mathcal{O}(c^{-2})$ smaller than $c(\boldsymbol{\sigma}\cdot\mathbf{p})$, and with its neglect, $\hat{\mathcal{X}}_0^{\mathrm{DHF}}$ reduces to

$$\hat{\mathcal{X}}_0^{\mathrm{DHF}} \simeq \left[2mc^2 - V^{\mathrm{nuc}} - J^{\mathrm{tot}}\right]^{-1} c(\boldsymbol{\sigma}\cdot\mathbf{p}). \tag{18.73}$$

The approximate transformed Fock operator is then

$$\begin{aligned}\hat{f}^{\mathrm{RegDHF}} \simeq \left[\hat{\mathcal{O}}_0^{\mathrm{DHF}}\right]^{-1} \Big[& V^{\mathrm{nuc}} + J^{\mathrm{tot}} - \hat{K}^{LL} \\ & + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{2mc^2 - V^{\mathrm{nuc}} - J^{\mathrm{tot}}}(\boldsymbol{\sigma}\cdot\mathbf{p})\Big]\left[\hat{\mathcal{O}}_0^{\mathrm{DHF}}\right]^{-1}.\end{aligned} \tag{18.74}$$

If we make use of the modified wave function ϕ^{FW}, which is approximately equal to ψ^L, we can write the transformed Fock operator entirely in terms of the transformed wave function:

$$\hat{f}^{\mathrm{RegDHF}\prime} \simeq V^{\mathrm{nuc}} + J^{\mathrm{tot}} - \hat{K} + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{2mc^2 - V^{\mathrm{nuc}} - J^{\mathrm{tot}}}(\boldsymbol{\sigma}\cdot\mathbf{p}) \tag{18.75}$$

where

$$\hat{K}\phi_i^{\mathrm{FW}} = \sum_j \left[\int d\mathbf{r}\; \phi_j^{\mathrm{FW}}\; g\; \phi_i^{\mathrm{FW}}\right]\psi_j^{\mathrm{FW}}. \tag{18.76}$$

Along with the construction of J^{tot} using (18.70) (and ignoring the inconsistency) and a spin-free approximation, the Fock matrix built from this operator would involve no more integrals than in a nonrelativistic calculation. This approximation was first proposed and investigated by Faas et al. (1995).

An alternative to the operator approach is to start from the matrix equations (Filatov 2002). Then the elimination the small-component, the construction of the transformation and the transformed Fock matrix are all straightforward. There is no difficulty with interpretation because the inverse of a matrix is well defined. The matrix to be inverted is positive definite so it presents no numerical problems. The drawback of a matrix method is that the basis set for the small component must be used, at least to construct the potentials that appear in the inverse. In that case, the same number of integrals is required as in the full Dirac–Hartree–Fock method, and there is no reduction in the integral work or the construction of the Fock matrix.

It should also be pointed out that the presence of the momentum operators in the ZORA Hamiltonian generates the small-component basis set. If a numerical integration scheme is employed, the small component must be tabulated on the grid as well as the large component. Thus, no matter whether a matrix approach or a grid-based approach is used, the small component is present either explicitly or implicitly.

18.5 Properties in the Regular Approximations

Since we can regard the regular approximation as a Foldy–Wouthuysen transformation, we should be able to use a similar line of development for the property operators as in chapter 16. There we found that it was relatively easy to deal with the electric perturbations because they appeared as powers of the perturbation operator, with a linear operator at lowest order. The magnetic perturbations, on the other hand, were much more complicated because the transformation involved complicated functions of the momentum, which must be replaced with the expression that includes the vector potential.

In the regular approximations, the situation is reversed. Here, the scalar potential is in the denominator of the transformation operator $\hat{\mathcal{X}}$, whereas the vector potential is in the numerator. The electric perturbations are therefore more complicated, and we will address these first. In what follows, we will consider only the one-electron operators and ignore two-electron contributions.

As in previous chapters, we represent the electric perturbation by a scalar potential W, which is added to the potential V. We must perform the transformation on the perturbed system and then try to separate the perturbation. The perturbed transformation operator $\hat{\mathcal{X}}$ is

$$\hat{\mathcal{X}} = (2mc^2 - V - W)^{-1} c(\boldsymbol{\sigma}\cdot\mathbf{p}) \tag{18.77}$$

and the normalization operator, $\hat{\mathcal{O}}$ is

$$\hat{\mathcal{O}} = \sqrt{1+\hat{\mathcal{X}}^\dagger\hat{\mathcal{X}}} = \left[1+(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2-V-W)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\right]^{1/2}. \tag{18.78}$$

If we work in the unnormalized representation and use the modified wave function ϕ^{FW} of (18.36), we do not have to deal with the square root, but rather with a metric operator.

We can define the perturbed Hamiltonian and the perturbed metric as

$$\hat{\mathcal{H}}' = V + W + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V - W)}(\boldsymbol{\sigma}\cdot\mathbf{p}), \tag{18.79}$$

$$\hat{\mathcal{G}}' = 1 + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V - W)^2}(\boldsymbol{\sigma}\cdot\mathbf{p}). \tag{18.80}$$

In this representation, we will have to work with a perturbation in both the Hamiltonian and the metric. For the ZORA Hamiltonian, of course, the metric is unity, and the metric perturbations can be disregarded. We can use the partitioning of (18.12) to separate the perturbation from the zeroth-order Hamiltonian, with the result

$$\begin{aligned}\hat{\mathcal{H}}' &= V + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{2mc^2 - V}(\boldsymbol{\sigma}\cdot\mathbf{p}) + W + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2 W}{(2mc^2 - V)(2mc^2 - V - W)}(\boldsymbol{\sigma}\cdot\mathbf{p})\\ &= \hat{\mathcal{H}}_0 + \hat{\mathcal{H}}_1.\end{aligned} \tag{18.81}$$

The perturbation separates nicely into a nonrelativistic perturbation and a relativistic correction, which is $\mathcal{O}(c^{-2})$. Likewise, the perturbed metric can be separated into a zeroth-order term and a perturbation:

$$\begin{aligned}\hat{\mathcal{G}}' &= 1 + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2 - V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p}) + (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{2c^2 W}{(2mc^2 - V)(2mc^2 - V - W)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\\ &\quad - (\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2 W^2}{(2mc^2 - V)^2(2mc^2 - V - W)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\\ &= \hat{\mathcal{G}}_0 + \hat{\mathcal{G}}_1.\end{aligned} \tag{18.82}$$

The perturbation to the metric has two relativistic corrections, which are $\mathcal{O}(c^{-4})$ and $\mathcal{O}(c^{-6})$. Since IORA has energy errors of $\mathcal{O}(c^{-4})$, the perturbation to the metric could justifiably be neglected.

To compare the relativistic correction to the perturbation with the Pauli expression, we let the denominators go to $2mc^2$, with the result

$$\hat{\mathcal{H}}_1 \rightarrow W + \frac{1}{4m^2c^2}(\boldsymbol{\sigma}\cdot\mathbf{p})W(\boldsymbol{\sigma}\cdot\mathbf{p}). \tag{18.83}$$

This is not exactly the same as the Pauli operator, but we must remember that the relativistic correction to the property is a second-order property in the Pauli approximation and includes a relativistic correction to the wave function. However, by applying the Dirac relation we can see that it contains the spin–orbit correction to the property and a spin-free correction as well.

With the perturbation operators defined, we can write down the first-order energy as

$$\begin{aligned}E^{(1)} &= \langle \phi^{(0)} | \hat{\mathcal{H}}_1 | \phi^{(0)} \rangle - E^{(0)} \langle \phi^{(0)} | \hat{\mathcal{G}}_1 | \phi^{(0)} \rangle \\ &= \langle \phi^{(0)} | W | \phi^{(0)} \rangle + \langle \phi^{(0)} | (\boldsymbol{\sigma}\cdot\mathbf{p}) \frac{c^2 W}{(2mc^2 - V)(2mc^2 - V - W)} (\boldsymbol{\sigma}\cdot\mathbf{p}) | \phi^{(0)} \rangle \\ &\quad - E^{(0)} \langle \phi^{(0)} | (\boldsymbol{\sigma}\cdot\mathbf{p}) \frac{2c^2 W}{(2mc^2 - V)(2mc^2 - V - W)^2} (\boldsymbol{\sigma}\cdot\mathbf{p}) | \phi^{(0)} \rangle \\ &\quad + E^{(0)} \langle \phi^{(0)} | (\boldsymbol{\sigma}\cdot\mathbf{p}) \frac{c^2 W^2}{(2mc^2 - V)^2 (2mc^2 - V - W)^2} (\boldsymbol{\sigma}\cdot\mathbf{p}) | \phi^{(0)} \rangle. \end{aligned} \tag{18.84}$$

Evaluation of the first-order property would be done in the same way as the energy, either by tabulation on a grid or by matrix manipulation.

If we work in the unnormalized representation, the magnetic perturbations are simple, because $(\boldsymbol{\sigma}\cdot\mathbf{p})$ only occurs in the numerator. The perturbed Hamiltonian is

$$\begin{aligned}\hat{\mathcal{H}}' &= V + \boldsymbol{\sigma}\cdot(\mathbf{p} + e\mathbf{A}) \frac{c^2}{2mc^2 - V} \boldsymbol{\sigma}\cdot(\mathbf{p} + e\mathbf{A}) \\ &= V + (\boldsymbol{\sigma}\cdot\mathbf{p}) \frac{c^2}{2mc^2 - V} (\boldsymbol{\sigma}\cdot\mathbf{p}) + e(\boldsymbol{\sigma}\cdot\mathbf{A}) \frac{c^2}{2mc^2 - V} (\boldsymbol{\sigma}\cdot\mathbf{p}) \\ &\quad + e(\boldsymbol{\sigma}\cdot\mathbf{p}) \frac{c^2}{2mc^2 - V} (\boldsymbol{\sigma}\cdot\mathbf{A}) + e(\boldsymbol{\sigma}\cdot\mathbf{A}) \frac{c^2}{2mc^2 - V} (\boldsymbol{\sigma}\cdot\mathbf{A}) \\ &= \hat{\mathcal{H}}_0 + \hat{\mathcal{H}}_1. \end{aligned} \tag{18.85}$$

Using the Dirac relation (4.14), we can write the perturbation as

$$\begin{aligned}\hat{\mathcal{H}}_1 &= \frac{c^2}{2mc^2 - V} \left[2e\mathbf{A}\cdot\mathbf{p} + e\hbar\boldsymbol{\sigma}\cdot\mathbf{B} + e^2\mathbf{A}^2 \right] \\ &\quad + \frac{c^2}{(2mc^2 - V)^2} \left[e(\mathbf{p}V)\cdot\mathbf{A} + e\hbar\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{A} \right]. \end{aligned} \tag{18.86}$$

The first term is the Pauli operator regularized, which can be seen by setting V to zero in the denominator. The second term contains a spin-free and a spin–orbit relativistic correction to the operator. The spin–orbit correction is the Pauli operator regularized. There is no spin-free relativistic correction to the Pauli operator, but as before we must realize that the relativistic correction to the property is a second-order perturbation, which includes a relativistic correction to the wave function.

In the same way the perturbed metric can be separated into a zeroth-order term and a perturbation:

$$\begin{aligned}\hat{\mathcal{G}}' &= 1+\boldsymbol{\sigma}\cdot(\mathbf{p}+e\mathbf{A})\frac{c^2}{(2mc^2-V)^2}\boldsymbol{\sigma}\cdot(\mathbf{p}+e\mathbf{A})\\ &= 1+(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2-V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})+e(\boldsymbol{\sigma}\cdot\mathbf{A})\frac{c^2}{(2mc^2-V)^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\\ &\quad +e(\boldsymbol{\sigma}\cdot\mathbf{p})\frac{c^2}{(2mc^2-V)^2}(\boldsymbol{\sigma}\cdot\mathbf{A})+e^2(\boldsymbol{\sigma}\cdot\mathbf{A})\frac{c^2}{(2mc^2-V)^2}(\boldsymbol{\sigma}\cdot\mathbf{A})\\ &= \hat{\mathcal{G}}_0+\hat{\mathcal{G}}_1.\end{aligned} \tag{18.87}$$

Using the Dirac relation (4.14) again, we can write the perturbation to the metric as

$$\begin{aligned}\hat{\mathcal{G}}_1 = &\frac{c^2}{(2mc^2-V)^2}\left[2e\mathbf{A}\cdot\mathbf{p}+e\hbar\boldsymbol{\sigma}\cdot\mathbf{B}+e^2\mathbf{A}^2\right]\\ &+\frac{2c^2}{(2mc^2-V)^3}\left[e(\mathbf{p}V)\cdot\mathbf{A}+e\hbar\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{A}\right].\end{aligned} \tag{18.88}$$

The two contributions to the metric corrections are $\mathcal{O}(c^{-2})$ and $\mathcal{O}(c^{-4})$. When we evaluate the first-order energy there are now two contributions of $\mathcal{O}(c^{-2})$, one from the metric and one from the Hamiltonian. These contribute to the kinetic correction to the magnetic perturbation in the Pauli approximation. The first-order energy expression is

$$\begin{aligned}E^{(1)} &= \langle\phi^{(0)}|\hat{\mathcal{H}}_1|\phi^{(0)}\rangle-E^{(0)}\langle\phi^{(0)}|\hat{\mathcal{G}}_1|\phi^{(0)}\rangle\\ &= \langle\phi^{(0)}|\frac{c^2}{(2mc^2-V)}\left[2e\mathbf{A}\cdot\mathbf{p}+e\hbar\boldsymbol{\sigma}\cdot\mathbf{B}+e^2\mathbf{A}^2\right]|\phi^{(0)}\rangle\\ &\quad +\langle\phi^{(0)}|\frac{c^2}{(2mc^2-V)^2}\left[e(\mathbf{p}V)\cdot\mathbf{A}+e\hbar\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{A}\right]|\phi^{(0)}\rangle\\ &\quad -E^{(0)}\langle\phi^{(0)}|\frac{c^2}{(2mc^2-V)^2}\left[2e\mathbf{A}\cdot\mathbf{p}+e\hbar\boldsymbol{\sigma}\cdot\mathbf{B}+e^2\mathbf{A}^2\right]|\phi^{(0)}\rangle\\ &\quad -E^{(0)}\langle\phi^{(0)}|\frac{2c^2}{(2mc^2-V)^3}\left[e(\mathbf{p}V)\cdot\mathbf{A}+e\hbar\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{A}\right]|\phi^{(0)}\rangle.\end{aligned} \tag{18.89}$$

19

Matrix Approximations

In quantum chemistry, regardless of which operators we choose for the Hamiltonian, we almost invariably implement our chosen method in a finite basis set. The Douglas–Kroll and Barysz–Sadlej–Snijders methods in the end required a matrix representation of the momentum-dependent operators in the implementation, and the regular methods usually end up with a basis set, even if the potentials are tabulated on a grid. Why not start with a matrix representation of the Dirac equation and perform transformations on the Dirac matrix rather than doing operator transformations, for which the matrix elements are difficult to evaluate analytically? It is almost always much easier to do manipulations with matrices of operators than with the operators themselves. Provided proper account is taken in the basis sets of the correct relationships between the range and the domain of the operators (Dyall et al. 1984), matrix manipulations can be performed with little or no approximation beyond the matrix representation itself.

In this chapter, we explore the use of matrix approximations. We will work within the framework of the modified Dirac equation (15.8), which we may write as

$$\tilde{\mathcal{H}}^D \begin{pmatrix} \psi^L \\ \phi^L \end{pmatrix} = \tilde{G} \begin{pmatrix} \psi^L \\ \phi^L \end{pmatrix} E. \tag{19.1}$$

This equation has the advantages that the large and pseudo-large components have the same symmetry and can be expanded in the same basis set, and the way in which the speed of light appears makes it easy to identify relativistic terms of different orders. We expand the large and pseudo-large components in a basis set

$$\psi^L = \sum_k a_k^L \chi_k; \qquad \phi^L = \sum_k a_k^P \chi_k \tag{19.2}$$

and obtain the matrix modified Dirac equation

$$\begin{pmatrix} \mathbf{V} & \mathbf{T} \\ \mathbf{T} & \mathbf{V}^p/4m^2c^2 - \mathbf{T} \end{pmatrix} \begin{pmatrix} \mathbf{a}^L \\ \mathbf{a}^P \end{pmatrix} = \begin{pmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}/2mc^2 \end{pmatrix} \begin{pmatrix} \mathbf{a}^L \\ \mathbf{a}^P \end{pmatrix} E \tag{19.3}$$

where **V** and **T** are the usual matrices of the nuclear potential and kinetic energies, **S** is the overlap matrix of the basis functions with elements $S_{ij} = \langle \chi_i | \chi_j \rangle$, and $\mathbf{V}^p$ is the matrix of $(\boldsymbol{\sigma} \cdot \mathbf{p}) V (\boldsymbol{\sigma} \cdot \mathbf{p})$, or of $\mathbf{p}V \cdot \mathbf{p}$ if we are choosing the spin-free version. For the purposes of the development it does not matter which we choose at this point. The Hamiltonian matrix will be denoted **H** and the metric (or overlap) matrix **G**:

$$\mathbf{H} = \begin{pmatrix} \mathbf{V} & \mathbf{T} \\ \mathbf{T} & \mathbf{V}^p/4m^2c^2 - \mathbf{T} \end{pmatrix}; \qquad \mathbf{G} = \begin{pmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}/2mc^2 \end{pmatrix}. \tag{19.4}$$

Note that, since we are working in a nonorthogonal basis, we always have to take account of the overlaps. We therefore need to insert the appropriate overlap, or *metric*, matrix wherever it is needed.

The process of making approximations starts with either a partitioning of the Hamiltonian and the metric, as in direct perturbation theory, or, for variational approximations, the elimination of the small component. This can be done either directly or via a Foldy–Wouthuysen transformation. Here, we are interested first and foremost in variational approximations, so we will focus on the elimination of the small component and the Foldy–Wouthuysen transformation. Before considering the approximations, we first outline some theory for the exact solutions.

19.1 The Matrix Elimination of the Small Components

The theory for the matrix elimination of the small component follows closely the lines of the operator elimination of the small component, but there are some subtleties that arise from the fact that we have a metric matrix that is not the unit matrix. We will first perform the direct elimination, and subsequently discuss the Foldy–Wouthuysen transformation.

In the direct elimination, we invert the second line of (19.3) to obtain a relation between the large- and pseudo-large-component coefficient vectors,

$$\mathbf{a}^P = \left[\mathbf{T}(1 + E/2mc^2) - \mathbf{V}^p/4m^2c^2\right]^{-1} \mathbf{T}\mathbf{a}^L \equiv \mathbf{X}\mathbf{a}^L. \tag{19.5}$$

This is a valid procedure for $E > -2mc^2$ because $\mathbf{V}^p$ is negative definite and **T** is positive definite. Substituting into the first line of (19.3), we get an equation for the large-component coefficient vector,

$$\left[\mathbf{V} + \mathbf{T}\left[\mathbf{T}(1 + E/2mc^2) - \mathbf{V}^p/4m^2c^2\right]^{-1} \mathbf{T}\right] \mathbf{a}^L = \mathbf{S}\mathbf{a}^L E. \tag{19.6}$$

This equation can be solved iteratively to obtain the eigenvalue E and eigenvector $\mathbf{a}^L$.

There is a problem with this equation, which turns out to be a major one. We have to perform the inversion and solution for each eigenstate because the energy is in the denominator and each eigenstate has a different energy. This means that each eigenvector corresponds to a different Hamiltonian matrix, and the eigenvectors $\mathbf{a}^L$ are not orthogonal on the metric matrix, which here is just $\mathbf{S}$. The same is true for the operator elimination of the small component: the large components are not orthogonal on the unit metric. Since we are ultimately interested in a procedure for many-electron systems, requiring iterative solutions for the eigenvalue problem, we would like to solve the matrix equation only once to obtain all eigenstates and have a properly orthonormal set.

The issue of different energies in the denominator of (19.6) is a soluble problem. Consider the whole set of exact, positive-energy solutions, whose large- and pseudo-large-component vectors are $\mathbf{A}^L$ and $\mathbf{A}^P$ and whose eigenvalues are collected into a diagonal matrix $\mathbf{E}$. The matrix equations are

$$\mathbf{V}\mathbf{A}^L + \mathbf{T}\mathbf{A}^P = \mathbf{S}\mathbf{A}^L\mathbf{E} \tag{19.7a}$$

$$\mathbf{T}\mathbf{A}^L + (\mathbf{V}^p/4m^2c^2 - \mathbf{T})\mathbf{A}^P = \mathbf{T}/2mc^2\mathbf{A}^P\mathbf{E}. \tag{19.7b}$$

Substituting $\mathbf{A}^P = \mathbf{X}\mathbf{A}^L$, we have to use the first equation to define $\mathbf{X}$ rather than the second,

$$\mathbf{X} = \mathbf{A}^P(\mathbf{A}^L)^{-1} = \mathbf{T}^{-1}\left[\mathbf{S}\mathbf{A}^L\mathbf{E}(\mathbf{A}^L)^{-1} - \mathbf{V}\right]. \tag{19.8}$$

The matrix $\mathbf{A}^L$ consists of linearly independent vectors and hence is invertible. Having already used the first equation we must now use the second to perform the elimination, which we do with a premultiplication by $\mathbf{X}^\dagger$ to give

$$\left[(\mathbf{X}^\dagger\mathbf{T} - \mathbf{X}^\dagger\mathbf{T}\mathbf{X}) + \frac{1}{4m^2c^2}\mathbf{X}^\dagger\mathbf{V}^p\mathbf{X}\right]\mathbf{A}^L = \frac{1}{2mc^2}\mathbf{X}^\dagger\mathbf{T}\mathbf{X}\mathbf{A}^L\mathbf{E}. \tag{19.9}$$

We now have an equation that in principle applies to all solutions. But the problem of orthonormality remains. The vectors $\mathbf{A}^L$ are the large-component vectors, and these are not orthonormal on the metric in this equation either.

There is fortunately a solution to this dilemma. We need only substitute $\mathbf{A}^P = \mathbf{X}\mathbf{A}^L$ into (19.7a) and add it to this equation to obtain an equation with the proper metric,

$$\left[(\mathbf{T}\mathbf{X} + \mathbf{X}^\dagger\mathbf{T} - \mathbf{X}^\dagger\mathbf{T}\mathbf{X}) + (\mathbf{V} + \frac{1}{4m^2c^2}\mathbf{X}^\dagger\mathbf{V}^p\mathbf{X})\right]\mathbf{A}^L = \left(\mathbf{S} + \frac{1}{2mc^2}\mathbf{X}^\dagger\mathbf{T}\mathbf{X}\right)\mathbf{A}^L\mathbf{E}. \tag{19.10}$$

The eigenvector orthogonality is given by the expectation value of the metric over $\mathbf{A}^L$:

$$\mathbf{A}^{L\dagger}\left(\mathbf{S} + \frac{1}{2mc^2}\mathbf{X}^\dagger\mathbf{T}\mathbf{X}\right)\mathbf{A}^L. \tag{19.11}$$

But $\mathbf{A}^L$ and $\mathbf{A}^P$ come from exact solutions of the modified Dirac equation, for which we have the orthonormality relation

$$\begin{aligned}\mathbf{I} &= \begin{pmatrix}\mathbf{A}^{L\dagger} & \mathbf{A}^{P\dagger}\end{pmatrix} \mathbf{G} \begin{pmatrix}\mathbf{A}^L \\ \mathbf{A}^P\end{pmatrix} = \begin{pmatrix}\mathbf{A}^{L\dagger} & \mathbf{A}^{P\dagger}\end{pmatrix} \begin{pmatrix}\mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}/2mc^2\end{pmatrix} \begin{pmatrix}\mathbf{A}^L \\ \mathbf{A}^P\end{pmatrix} \\ &= \mathbf{A}^{L\dagger}\mathbf{S}\mathbf{A}^L + \mathbf{A}^{P\dagger}\mathbf{T}\mathbf{A}^P/2mc^2\end{aligned} \tag{19.12}$$

Substituting $\mathbf{A}^P = \mathbf{X}\mathbf{A}^L$, we get

$$\mathbf{A}^{L\dagger}(\mathbf{S} + \frac{1}{2mc^2}\mathbf{X}^\dagger\mathbf{T}\mathbf{X})\mathbf{A}^L = \mathbf{I}, \tag{19.13}$$

proving the orthonormality of $\mathbf{A}^L$ on this metric. Because of the properties of the metric, (19.10) has been called *normalized elimination of the small component*, or NESC (Dyall 1997). The usual method defined in (19.6), which we write here with the use of $\mathbf{X}$,

$$(\mathbf{V} + \mathbf{T}\mathbf{X})\mathbf{A}^L = \mathbf{S}\mathbf{A}^L\mathbf{E}, \tag{19.14}$$

has been termed *unnormalized elimination of the small component*, or UESC.

Both equations, UESC and NESC, fit naturally into a nonrelativistic environment. We have a modified potential energy matrix, a modified kinetic energy matrix, and a modified overlap or metric matrix. For UESC these are

$$\begin{aligned}\tilde{\mathbf{V}}^{\text{UESC}} &= \mathbf{V} \\ \tilde{\mathbf{T}}^{\text{UESC}} &= \mathbf{T}\mathbf{X} \\ \tilde{\mathbf{S}}^{\text{UESC}} &= \mathbf{S}\end{aligned} \tag{19.15}$$

which differ from the nonrelativistic matrices only in the kinetic energy[1]. For NESC the matrices are

$$\begin{aligned}\tilde{\mathbf{V}}^{\text{NESC}} &= \mathbf{V} + \mathbf{X}^\dagger\mathbf{V}^p\mathbf{X}/4m^2c^2 \\ \tilde{\mathbf{T}}^{\text{NESC}} &= \mathbf{T}\mathbf{X} + \mathbf{X}^\dagger\mathbf{T} - \mathbf{X}^\dagger\mathbf{T}\mathbf{X} = \mathbf{T} - (\mathbf{I} - \mathbf{X})^\dagger\mathbf{T}(\mathbf{I} - \mathbf{X}) \\ \tilde{\mathbf{S}}^{\text{NESC}} &= \mathbf{S} + \mathbf{X}^\dagger\mathbf{T}\mathbf{X}/2mc^2\end{aligned} \tag{19.16}$$

which have relativistic corrections for all three matrices. The potential and the metric in both cases have the correct nonrelativistic limit. To obtain the correct limit for the kinetic energy, we should have $\mathbf{X} \to \mathbf{I}$. Clearly this is so from (19.5), but for the entire set we must use (19.8). In the nonrelativistic limit,

$$(\mathbf{T} + \mathbf{V})\mathbf{A}^L = \mathbf{S}\mathbf{A}^L\mathbf{E}, \tag{19.17}$$

---

[1] The UESC kinetic energy matrix is not Hermitian for the **X** derived for all eigenvalues (Kutzelnigg and Liu 2005).

and substitution into (19.8) gives us the desired nonrelativistic limit, $\mathbf{X} = \mathbf{I}$. It follows that for the physical value of c, $\mathbf{X} = \mathbf{I} + \mathcal{O}(c^{-2})$. With this relation, we can examine the relativistic corrections to the matrices. In UESC, all the relativistic corrections are lumped into a correction to the kinetic energy that is $\mathcal{O}(c^{-2})$, but in NESC, the potential and metric both have relativistic corrections that are $\mathcal{O}(c^{-2})$. The kinetic energy correction is $\mathcal{O}(c^{-4})$, and is negative. There is an interesting parallel here with direct perturbation theory. The relativistic corrections to the operators in direct perturbation theory are in the potential and the metric, and there is no kinetic energy correction. In NESC, there are explicit relativistic corrections to the potential and the metric. **X** represents the correction to the wave function, and hence the kinetic energy correction is a pure wave function effect.

19.2 Properties of the NESC and UESC Equations

There are other differences between UESC and NESC that give further reasons for preferring the latter. To examine these differences, we use perturbation theory and exploit the fact that NESC may in fact be obtained by an unnormalized Foldy–Wouthuysen transformation of the Dirac Hamiltonian and metric matrices. The transformation can be written in terms of **X** as

$$\mathbf{U}' = \begin{pmatrix} \mathbf{I} & \mathbf{X}^{\dagger} \\ -\mathbf{T}\mathbf{X}\mathbf{S}^{-1}/2mc^2 & \mathbf{I} \end{pmatrix} \tag{19.18}$$

where the prime indicates that the matrix is orthogonal but not unitary. The matrix **X** can be arbitrarily chosen here—it is not necessarily the matrix defined in (19.8) but it will be taken to be the appropriate matrix for the method, NESC or UESC. The lower left corner of $\mathbf{U}'$ follows from the requirement that use of this transformation on the metric matrix **G** should preserve the block-diagonal structure, a property that will be useful later on.

The transformed Hamiltonian and metric are symbolically represented as

$$\mathbf{H}' = \begin{pmatrix} \mathbf{H}'_{++} & \mathbf{H}'_{+-} \\ \mathbf{H}'_{-+} & \mathbf{H}'_{--} \end{pmatrix} \qquad \mathbf{G}' = \begin{pmatrix} \mathbf{G}'_{++} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}'_{--} \end{pmatrix}, \tag{19.19}$$

with

$$\begin{aligned} \mathbf{H}'_{++} &= \mathbf{V} + \mathbf{X}^{\dagger}\mathbf{V}^{p}\mathbf{X}/4m^2c^2 + \mathbf{T}\mathbf{X} + \mathbf{X}^{\dagger}\mathbf{T} - \mathbf{X}^{\dagger}\mathbf{T}\mathbf{X}, \\ \mathbf{H}'_{-+} &= -\mathbf{T}\mathbf{X}\mathbf{S}^{-1}(\mathbf{V} + \mathbf{T}\mathbf{X}) + \mathbf{T} - \mathbf{T}\mathbf{X} + \mathbf{V}^{p}\mathbf{X}/4m^2c^2 = (\mathbf{H}'_{+-})^{\dagger}, \\ \mathbf{H}'_{--} &= \mathbf{V}^{p}/4m^2c^2 + \mathbf{X}\mathbf{V}\mathbf{X}^{\dagger} - \mathbf{T} - \mathbf{X}\mathbf{T} - \mathbf{T}\mathbf{X}^{\dagger}, \end{aligned} \tag{19.20}$$

and

$$\begin{aligned} \mathbf{G}'_{++} &= \mathbf{S} + \mathbf{X}^{\dagger}\mathbf{T}\mathbf{X}/2mc^2 \equiv \tilde{\mathbf{G}}, \\ \mathbf{G}'_{--} &= \mathbf{T}/2mc^2 + \mathbf{T}\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\dagger}\mathbf{T}/4m^2c^4. \end{aligned} \tag{19.21}$$

The NESC Hamiltonian is just $\mathbf{H}'_{++}$, and the NESC metric is $\mathbf{G}'_{++}$. These arise naturally from the transformation of the Hamiltonian.

We now make use of the transformed Hamiltonian to develop a perturbation theory for both NESC and UESC. The NESC partitioning is

$$\mathbf{H}_0^{\text{NESC}} = \begin{pmatrix} \mathbf{H}'_{++} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}'_{--} \end{pmatrix} \qquad \mathbf{H}_1^{\text{NESC}} = \begin{pmatrix} \mathbf{0} & \mathbf{H}'_{+-} \\ \mathbf{H}'_{-+} & \mathbf{0} \end{pmatrix} \tag{19.22}$$

$$\mathbf{G}_0^{\text{NESC}} = \begin{pmatrix} \mathbf{G}'_{++} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}'_{--} \end{pmatrix} \qquad \mathbf{G}_1^{\text{NESC}} = \mathbf{0} \tag{19.23}$$

and the UESC partitioning is

$$\mathbf{H}_0^{\text{UESC}} = \begin{pmatrix} \mathbf{H}'_{++} - \mathbf{Y}_{++} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}'_{--} - \mathbf{Y}_{--} \end{pmatrix} \qquad \mathbf{H}_1^{\text{UESC}} = \begin{pmatrix} \mathbf{Y}_{++} & \mathbf{H}'_{+-} \\ \mathbf{H}'_{-+} & \mathbf{Y}_{--} \end{pmatrix} \tag{19.24}$$

$$\mathbf{G}_0^{\text{UESC}} = \begin{pmatrix} \mathbf{G}'_{++} - \mathbf{Q}_{++} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}'_{--} - \mathbf{Q}_{--} \end{pmatrix} \qquad \mathbf{G}_1^{\text{UESC}} = \begin{pmatrix} \mathbf{Q}_{++} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{--} \end{pmatrix} \tag{19.25}$$

where

$$\mathbf{Y}_{++} = \mathbf{X}^{\dagger}\mathbf{V}^{p}\mathbf{X}/4m^2c^2 + \mathbf{X}^{\dagger}\mathbf{T}(\mathbf{I} - \mathbf{X}) \tag{19.26}$$

$$\mathbf{Q}_{++} = \mathbf{X}^{\dagger}\mathbf{T}\mathbf{X}/2mc^2 \tag{19.27}$$

and for the present purposes the choice of $\mathbf{Y}_{--}$ and $\mathbf{Q}_{--}$ is arbitrary. Note that both $\mathbf{Y}_{++}$ and $\mathbf{Q}_{++}$ are $\mathcal{O}(c^{-2})$. In both cases we have a block-diagonal zeroth-order Hamiltonian, and the solutions have nonzero components in either the positive- or the negative-energy space,

$$\mathbf{c}'_{+} = \begin{pmatrix} \mathbf{a}^{L} \\ \mathbf{0} \end{pmatrix}, \qquad \mathbf{c}'_{-} = \begin{pmatrix} \mathbf{0} \\ \mathbf{a}^{P} \end{pmatrix}. \tag{19.28}$$

The reference state is one of the positive-energy states, which we will denote $\mathbf{c}'_0$. The remaining states will be indexed with $k+$ or $k-$ for positive- and negative-energy states, respectively.

The energies at the first and second order are given by

$$E^{(1)} = \mathbf{c}_0'^{\dagger}(\mathbf{H}_1 - \mathbf{G}_1 E_0)\mathbf{c}_0' \tag{19.29}$$

$$\begin{aligned} E^{(2)} &= \sum_{k+} \frac{\mathbf{c}_0'^{\dagger}(\mathbf{H}_1 - \mathbf{G}_1 E_0)\mathbf{c}_{k+}'\mathbf{c}_{k+}'^{\dagger}(\mathbf{H}_1 - \mathbf{G}_1 E_0)\mathbf{c}_0'}{E_0 - E_{k+}} \\ &+ \sum_{k-} \frac{\mathbf{c}_0'^{\dagger}(\mathbf{H}_1 - \mathbf{G}_1 E_0)\mathbf{c}_{k-}'\mathbf{c}_{k-}'^{\dagger}(\mathbf{H}_1 - \mathbf{G}_1 E_0)\mathbf{c}_0'}{E_0 - E_{k-}}. \end{aligned} \tag{19.30}$$

The first-order energy is zero for NESC because the perturbation is purely off-diagonal and hence only connects positive- and negative-energy states. For UESC, there is a first-order energy correction, which is the expectation of

$$\mathbf{X}^{\dagger}\left[\mathbf{V}^{p}/4m^2c^2 - \mathbf{T}(1 + E_0/2mc^2)\right]\mathbf{X} + \mathbf{X}^{\dagger}\mathbf{T} \tag{19.31}$$

which is zero only if $\mathbf{X}$ is defined by (19.5) with $E = E_0$. Otherwise, it is nonzero.

In the second-order energy, UESC has contributions from both the positive- and the negative-energy states. The contributions from the positive-energy states are $\mathcal{O}(c^{-2})$, and the contributions from the negative-energy states are $\mathcal{O}(c^{-4})$. NESC only has contributions from the negative-energy states, of $\mathcal{O}(c^{-4})$, and these are positive because $E_0 > E_{k-}$. This shows that the NESC energy is below the Dirac energy and in an iterative procedure would converge on it from below. It does *not* necessarily mean that there is variational collapse.

What we learn from this analysis is that the traditional elimination of the small components (i.e., UESC) presents some issues that make its application to many-electron systems difficult. The center of the problem is the neglect of the normalization terms along with the corresponding terms in the Hamiltonian. The consequence is that either a separate Hamiltonian has to be set up for each one-particle state, or we must be content with a set of states that are in error by $\mathcal{O}(c^{-2})$. It is much better to start from a theory in which the normalization is properly incorporated, such as NESC, for which all one-particle states can be obtained from the same Hamiltonian and the error is $\mathcal{O}(c^{-4})$. NESC is not without its drawbacks, the main one being that the metric is energy-dependent,[2] and it would therefore be necessary to solve the generalized eigenvalue problem for each iteration. This need only be done once per iteration, compared with multiple operations to obtain separate solutions for UESC. We will therefore base the developments of this chapter on NESC or the Foldy–Wouthuysen transformation, and will not be concerned any further with UESC. For this reason, we simplify the notation by dropping the superscript NESC such that for any operator $\hat{\Omega}$ we set $\tilde{\boldsymbol{\Omega}} = \tilde{\boldsymbol{\Omega}}^{\mathrm{NESC}}$.

There is a parallel between UESC and NESC on the one hand and ZORA and IORA on the other hand. If we set the energy in $\mathbf{X}$ to zero, we get the matrix versions of

---

[2] Kutzelnigg and Liu (2005) have shown how $\mathbf{X}$ can be obtained iteratively without using the energy.

the ZORA and IORA equations. Like UESC and NESC, ZORA has errors of $\mathcal{O}(c^{-2})$, while the errors in IORA are of $\mathcal{O}(c^{-4})$. The reasons are the same: ZORA neglects the normalization term that makes the wave function orthogonal on the chosen metric, whereas IORA includes it.

Having shown that NESC may be formulated as an unnormalized Foldy–Wouthuysen transformation, it appears natural to proceed to a normalized Foldy–Wouthuysen transformation. The normalization must be defined on the appropriate metric, here denoted $\mathbf{G}''$, for which the positive–positive block must be $\mathbf{S}$. As we are no longer concerned with the negative-energy states, we can choose the negative–negative block arbitrarily. The normalized transformation can then be expressed in terms of the unnormalized transformation by

$$\mathbf{U}'' = (\mathbf{G}'')^{1/2}(\mathbf{G}')^{-1/2}\mathbf{U}'. \tag{19.32}$$

To obtain the one-electron matrices for the normalized modified Foldy–Wouthuysen (NMFW) transformation we must therefore premultiply the NESC matrices by $(\mathbf{G}'')^{1/2}_{++}(\mathbf{G}')^{-1/2}_{++} = \mathbf{S}^{1/2}\tilde{\mathbf{S}}^{-1/2}$ and postmultiply by the transpose, $\tilde{\mathbf{S}}^{-1/2}\mathbf{S}^{1/2}$:

$$\begin{aligned}
\tilde{\mathbf{V}}^{\text{NMFW}} &= \mathbf{S}^{1/2}\tilde{\mathbf{S}}^{-1/2}(\mathbf{V} + \mathbf{X}^{\dagger}\mathbf{V}^{p}\mathbf{X}/4m^2c^2)\tilde{\mathbf{S}}^{-1/2}\mathbf{S}^{1/2} \\
\tilde{\mathbf{T}}^{\text{NMFW}} &= \mathbf{S}^{1/2}\tilde{\mathbf{S}}^{-1/2}(\mathbf{TX} + \mathbf{X}^{\dagger}\mathbf{T} - \mathbf{X}^{\dagger}\mathbf{TX})\tilde{\mathbf{S}}^{-1/2}\mathbf{S}^{1/2} \\
\tilde{\mathbf{S}}^{\text{NMFW}} &= \mathbf{S}^{1/2}\tilde{\mathbf{S}}^{-1/2}(\mathbf{S} + \mathbf{X}^{\dagger}\mathbf{TX}/2mc^2)\tilde{\mathbf{S}}^{-1/2}\mathbf{S}^{1/2} = \mathbf{S}.
\end{aligned} \tag{19.33}$$

Likewise, the NMFW eigenvectors $\mathbf{A}^{N}$ are related to the eigenvectors of the unnormalized Foldy–Wouthuysen transformation, $\mathbf{A}^{L}$, by

$$\mathbf{A}^{N} = \mathbf{S}^{-1/2}\tilde{\mathbf{S}}^{1/2}\mathbf{A}^{L}. \tag{19.34}$$

These transformations are in fact nothing more than the application of a symmetric orthonormalization followed by a transformation to the new metric.

In practical terms, these transformations require two diagonalizations and back-transformations, so because the unnormalized expressions are more straightforward and entirely equivalent, these are to be preferred in general. Later, we will present an approximation that requires the normalized coefficients.

19.3 Inclusion of the Two-Electron Terms

For many-electron systems we wish to incorporate the two-electron terms of the Hamiltonian into the model. As for the one-electron terms, we may perform an unnormalized Foldy–Wouthuysen transformation on the two-electron integrals to obtain a transformed set of integrals,

$$\begin{aligned}
(\tilde{\mu}\tilde{\nu}|\tilde{\tau}\tilde{\rho}) = (\mu\nu|\tau\rho) &+ X^{*}_{\kappa\mu}(\hat{\Pi}\kappa\hat{\Pi}\lambda|\tau\rho)X_{\lambda\nu}/4m^2c^2 \\
&+ X^{*}_{\kappa\tau}(\mu\nu|\hat{\Pi}\kappa\hat{\Pi}\lambda)X_{\lambda\rho}/4m^2c^2 + X^{*}_{\kappa\mu}X^{*}_{\zeta\tau}(\hat{\Pi}\kappa\hat{\Pi}\lambda|\hat{\Pi}\zeta\hat{\Pi}\eta)X_{\lambda\nu}X_{\eta\rho},
\end{aligned} \tag{19.35}$$

where the Einstein convention of summation over repeated indices has been used, and will continue to be used in this section. $\hat{\Pi}$ is $\mathbf{p}$ or $(\boldsymbol{\sigma} \cdot \mathbf{p})$ depending on whether the spin-free or spin-dependent Hamiltonian is used. The spin-free modified two-electron integrals can be used in an otherwise nonrelativistic code, just as the modified one-electron integrals could replace the nonrelativistic integrals.

Naturally, we also need a new expression for $\mathbf{X}$ that includes the two-electron terms, and this we get from the Dirac–Hartree–Fock equations, which we represent here in unrestricted form,

$$\left[\mathbf{V} + \mathbf{J}^{LL} - \mathbf{K}^{LL}\right]\mathbf{A}^L + \left[\mathbf{T} - \mathbf{K}^{LP}\right]\mathbf{A}^P = \mathbf{S}\mathbf{A}^L\mathbf{E} \tag{19.36a}$$

$$\left[\mathbf{T} - \mathbf{K}^{PL}\right]\mathbf{A}^L + \left[(\mathbf{V}^p + \mathbf{J}^{PP} - K^{PP})/4m^2c^2 - \mathbf{T}\right]\mathbf{A}^P = \mathbf{T}/2mc^2\mathbf{A}^P\mathbf{E}. \tag{19.36b}$$

The direct and exchange matrices are defined by

$$\begin{aligned} J^{LL}_{\mu\nu} &= (\mu\nu|\kappa\lambda)D^{LL}_{\lambda\kappa} + (\mu\nu|\hat{\Pi}\kappa\hat{\Pi}\lambda)D^{PP}_{\lambda\kappa} \\ J^{PP}_{\mu\nu} &= (\hat{\Pi}\mu\hat{\Pi}\nu|\kappa\lambda)D^{LL}_{\lambda\kappa} + (\hat{\Pi}\mu\hat{\Pi}\nu|\hat{\Pi}\kappa\hat{\Pi}\lambda)D^{PP}_{\lambda\kappa} \end{aligned} \tag{19.37}$$

$$\begin{aligned} K^{LL}_{\mu\nu} &= (\mu\lambda|\kappa\nu)D^{LL}_{\lambda\kappa} \\ K^{LP}_{\mu\nu} &= (\mu\lambda|\hat{\Pi}\kappa\hat{\Pi}\nu)D^{LP}_{\lambda\kappa} \\ K^{PL}_{\mu\nu} &= (\hat{\Pi}\mu\hat{\Pi}\lambda|\kappa\nu)D^{PL}_{\lambda\kappa} \\ K^{PP}_{\mu\nu} &= (\hat{\Pi}\mu\hat{\Pi}\lambda|\hat{\Pi}\kappa\hat{\Pi}\nu)D^{PP}_{\lambda\kappa} \end{aligned} \tag{19.38}$$

and the density matrices by

$$\begin{aligned} \mathbf{D}^{LL} &= \mathbf{a}_i^L\mathbf{a}_i^{L\dagger} \equiv \tilde{\mathbf{D}} \\ \mathbf{D}^{LP} &= \mathbf{a}_i^L\mathbf{a}_i^{P\dagger} = \mathbf{D}^{LL}\mathbf{X}^\dagger \\ \mathbf{D}^{PL} &= \mathbf{a}_i^P\mathbf{a}_i^{L\dagger} = \mathbf{X}\mathbf{D}^{LL} \\ \mathbf{D}^{PP} &= \mathbf{a}_i^P\mathbf{a}_i^{P\dagger} = \mathbf{X}\mathbf{D}^{LL}\mathbf{X}^\dagger. \end{aligned} \tag{19.39}$$

To obtain an expression for $\mathbf{X}$, we insert $\mathbf{A}^P = \mathbf{X}\mathbf{A}^L$ into (19.36a) and invert it:

$$\mathbf{X} = [\mathbf{T} - \mathbf{K}^{LP}]^{-1}[\mathbf{S}\mathbf{A}^L\mathbf{E}\mathbf{A}^{L\dagger} - (\mathbf{V} + \mathbf{J}^{LL} - \mathbf{K}^{LL})] \tag{19.40}$$

The definition we end up with is recursive because both $\mathbf{J}^{LL}$ and $\mathbf{K}^{LP}$ contain $\mathbf{X}$ if we replace $\mathbf{A}^P$ with $\mathbf{X}\mathbf{A}^L$ throughout. In an iterative scheme this would not be a serious problem because we could use the value of $\mathbf{X}$ from the previous iteration to construct the $\mathbf{J}$ and $\mathbf{K}$ matrices, and at convergence the input and output $\mathbf{X}$ matrices would be the same to the required numerical precision.

With the expressions for the density, the direct and the exchange matrices, and the definition of the transformed integrals, we can write down the NESC SCF equation as

$$\left(\tilde{\mathbf{T}} + \tilde{\mathbf{V}} + \tilde{\mathbf{J}} - \tilde{\mathbf{K}}\right) \mathbf{A}^L = \tilde{\mathbf{S}}\mathbf{A}^L \mathbf{E} \tag{19.41}$$

which is just like a nonrelativistic SCF equation but with modified matrices. The NESC direct and exchange matrices are defined by

$$\tilde{J}_{\mu\nu} = (\tilde{\mu}\tilde{\nu}|\tilde{\kappa}\tilde{\lambda})\tilde{D}_{\lambda\kappa}, \qquad \tilde{K}_{\mu\nu} = (\tilde{\mu}\tilde{\lambda}|\tilde{\kappa}\tilde{\nu})\tilde{D}_{\lambda\kappa}. \tag{19.42}$$

We have so far only considered the Coulomb interaction, but it is straightforward to incorporate the Gaunt or Breit interactions.

The modified two-electron terms contain all the relativistic integrals, which means that the integral work is no different from that in the full solution of the Dirac–Hartree–Fock equations. It would save a lot of work if we could approximate the integrals, in the same way as we did for the Douglas–Kroll–Hess approximation. To do so, we must use the normalized Foldy–Wouthuysen transformation. The DKH approximation neglects the commutator of the transformation with the two-electron Coulomb operator, and in so doing removes all the spin-dependent terms. We must therefore also use a spin-free one-electron Hamiltonian. The approximate Hamiltonian (in terms of operators rather than matrices) is

$$\mathcal{H} \simeq \sum_i \hat{\mathcal{O}}_i^{-1} \left[V_i + \hat{\mathcal{X}}_i^\dagger \mathrm{p}V \cdot \mathrm{p}\hat{\mathcal{X}}_i / 4m^2c^2 + T_i\hat{\mathcal{X}}_i + \hat{\mathcal{X}}_i^\dagger T_i - \hat{\mathcal{X}}_i^\dagger T_i \hat{\mathcal{X}}_i \right] \hat{\mathcal{O}}_i^{-1} + \sum_{i>j} \frac{1}{r_{ij}} \tag{19.43}$$

where

$$\hat{\mathcal{O}}_i = \left[1 + \hat{\mathcal{X}}_i^\dagger T_i \hat{\mathcal{X}}_i / 4m^2c^2 \right]^{1/2} \tag{19.44}$$

The parallels between the one-electron operator and the NMFW one-electron matrices in (19.33) are obvious. It turns out that insertion of the resolution of the identity in a nonorthogonal basis does not strictly lead to the NMFW matrices. These matrices are rather to be regarded as the result of a symmetric reorthonormalization. However, given the approximation to the two-electron operator, the difference betwen the two is not likely to be serious. Armed with the NMFW one-electron matrices, we can write down the NMFW SCF equation,

$$\left(\tilde{\mathbf{T}}^{\mathrm{NMFW}} + \tilde{\mathbf{V}}^{\mathrm{NMFW}} + \tilde{\mathbf{J}}^{\mathrm{NMFW}} - \tilde{\mathbf{K}}^{\mathrm{NMFW}}\right) \mathbf{A}^N = \mathbf{S}\mathbf{A}^N \mathbf{E} \tag{19.45}$$

which again is just like a nonrelativistic SCF equation, but with modified matrices. The NMFW direct and exchange matrices are defined by

$$\tilde{J}_{\mu\nu}^{\mathrm{NMFW}} = (\mu\nu|\kappa\lambda)\tilde{D}_{\lambda\kappa}^{\mathrm{NMFW}}, \quad \tilde{K}_{\mu\nu}^{\mathrm{NMFW}} = (\mu\lambda|\kappa\nu)\tilde{D}_{\lambda\kappa}^{\mathrm{NMFW}} \tag{19.46}$$

where

$$\tilde{D}^{\mathrm{NMFW}} = \mathrm{a}_i^N \mathrm{a}_i^{N\dagger} \tag{19.47}$$

19.4 Atom-Centered Approximations

The elimination of the small component as a method for solving the Dirac equation gives very little advantage over the original. It merely reduces the size of the one-electron matrices that must be diagonalized, and unless the two-electron integrals are approximated there is no reduction of work in the most time-consuming part of the procedure. Part of the reason at least is the presence of the energies, or spinor eigenvalues, in the expression for **X**. What would be really useful is some approximations in which the energies were not present, or at least not varying.

Neglecting the energy in the definition of **X** in (19.5), we have

$$\mathbf{X}_0 = \left[\mathbf{T} - \mathbf{V}^p/4m^2c^2\right]^{-1}\mathbf{T}. \tag{19.48}$$

Substitution of $\mathbf{X}_0$ for **X** in (19.14) gives the ZORA equation in the modified Dirac representation

$$\left(\mathbf{V} + \mathbf{T}\left[\mathbf{T} - \mathbf{V}^p/4m^2c^2\right]^{-1}\mathbf{T}\right)\mathbf{A}^L = \mathbf{S}\mathbf{A}^L\mathbf{E}. \tag{19.49}$$

The expression in square brackets is the matrix of $2mc^2 - V$ evaluated over the set of basis functions $\{((\boldsymbol{\sigma}\cdot\mathbf{p})/2mc)\chi_\mu\}$, which is essentially the small-component basis set. Likewise, **T** can be regarded as the integral over $(\boldsymbol{\sigma}\cdot\mathbf{p})$ between the small- and large-component basis set. This makes it clear that the denominator in ZORA must be evaluated over the small-component basis functions.

Similarly, if we substitute $\mathbf{X}_0$ for **X** in (19.10), we get the IORA equation in the modified Dirac representation,

$$\begin{aligned}&\left(\mathbf{V} + \mathbf{T}\left[\mathbf{T} - \mathbf{V}^p/4m^2c^2\right]^{-1}\mathbf{T}\right)\mathbf{A}^L\\ &= \left(\mathbf{S} + \mathbf{T}\left[\mathbf{T} - \mathbf{V}^p/4m^2c^2\right]^{-1}\mathbf{T}\left[\mathbf{T} - \mathbf{V}^p/4m^2c^2\right]^{-1}\mathbf{T}\right)\mathbf{A}^L\mathbf{E}.\end{aligned} \tag{19.50}$$

We have discussed ZORA and IORA at length in chapter 18, and will not pursue these approximations here.

It is even possible to define a free-particle transformation from which we could continue to develop a matrix Douglas–Kroll–Hess or Barysz–Sadlej–Snijders approximation. But with the matrix formalism we now have available, there is the opportunity to take a different approach, one that centers around the properties of the matrix **X**.

For the purpose of the development, we will use (19.5) and expand **X** into the unit matrix and a residue,

$$\mathbf{X} = \mathbf{I} + \frac{1}{4m^2c^2}\mathbf{T}^{-1}(\mathbf{V}^p - 2mE\mathbf{T})\left[\mathbf{I} - \mathbf{T}^{-1}(\mathbf{V}^p - 2mE\mathbf{T})/4m^2c^2\right]^{-1}. \tag{19.51}$$

The size of the off-diagonal matrix elements of $\mathbf{X}$ will be determined by the size of the corresponding matrix elements of $\mathbf{V}^p$ and $\mathbf{T}$, so we need some way to estimate the size of these matrix elements. The expression in parentheses is the matrix of $(\boldsymbol{\sigma}\cdot\mathbf{p})(V-E)(\boldsymbol{\sigma}\cdot\mathbf{p})$. If we approximate E with the nonrelativistic energy, this is the operator from direct perturbation theory that gives the lowest-order relativistic correction to the energy, given in (17.76). Replacing E with the operators from the nonrelativistic energy expression $E = T + V$ we find that

$$\mathbf{V}^p - 2mE\mathbf{T} \simeq 2m\mathbf{T}^2. \tag{19.52}$$

The size of the off-diagonal matrix elements of $\mathbf{X}$ will therefore approximately scale as the off-diagonal elements of $\mathbf{T}$. These matrix elements will be largest for pairs of basis functions that describe the core of the same atom, and will be smallest for basis functions on different centers, due to the overlap. It is then a reasonable approximation to set the elements of $\mathbf{X}$ that connect basis functions on different centers to zero.

In this approximation, $\mathbf{X}$ will be block-diagonal, with each block corresponding to an atom in the molecule. We can now do a full four-component calculation for each atom A in the molecule and use the solutions to define the atomic matrices $\mathbf{X}_A$. It is not necessary for this purpose to use the second part of (19.8): we may simply invert the $\mathbf{A}_A^L$ and postmultiply the $\mathbf{A}_A^P$ by the inverses to obtain the $\mathbf{X}_A$. This we do once and for all in a given basis set and store $\mathbf{X}_A$. The practical advantage of this procedure is that there is no energy dependence in the molecular $\mathbf{X}$: it is all in the atomic $\mathbf{X}_A$, and the molecular $\mathbf{X}$ can be set up once and used to transform the integrals at the beginning of the calculation.

What has actually happened in this approximation? $\mathbf{X}$ is the matrix that connects the large and small components. In the approximation, we have fixed the relation between the large and small components so that this no longer varies in the molecule. We have therefore projected the molecular solutions onto the positive-energy states of the isolated atoms. The missing term is the change of the small/large ratio when the molecule is formed, which we saw in chapter 5 corresponds to the polarization of the vacuum. The energy involved in molecular formation is very small compared with $2mc^2$, so the neglected term will be very small. Tests for molecules containing heavy atoms show that it is of the order of a few microhartrees. Note that the neglected term itself is $\mathcal{O}(c^{-2})$, but the coefficient multiplying c^{-2} is small.

The development so far has been in terms of a primitive, or uncontracted, basis set. In molecular calculations it is often the case that we wish to use contracted basis sets, which are defined for each atom type and are derived from some atomic calculation. When contracting the basis set for NESC, we must use a general contraction based on the positive-energy atomic solutions, because we have made a projection onto the positive-energy states. If we use any other kind of contraction, we could be mixing in negative-energy states, which might create problems if these states were occupied. Of course, we can make further approximations with the contractions, for example to obtain a set of SCF functions with uncontracted primitives, but we must ensure that the basis does not contain significant contamination from negative-energy states.

The contraction is in fact a transformation of the basis, which can be represented by a rectangular matrix. The transformation matrix for the contraction is taken from $\mathbf{A}^L$

by choosing the vectors corresponding to the occupied spinors and a selection of the virtual spinors. If we apply the atomic **X** to these, we generate the vectors from $\mathbf{A}^P$. All the integrals can then be represented in terms of a transformation of the primitive integrals using a selection from $\mathbf{A}^L$ and $\mathbf{A}^P$, and there is no need to generate **X** at all. The contracted one-electron integral matrices can be written

$$\begin{aligned}
\tilde{\mathbf{V}}^{\text{NESC,cnt}} &= \mathbf{A}^{L\dagger}\mathbf{V}\mathbf{A}^L + \mathbf{A}^{P\dagger}\mathbf{V}^p\mathbf{A}^P/4m^2c^2 \\
\tilde{\mathbf{T}}^{\text{NESC,cnt}} &= \mathbf{A}^{L\dagger}\mathbf{T}\mathbf{A}^P + \mathbf{A}^{P\dagger}\mathbf{T}\mathbf{A}^L - \mathbf{A}^{P\dagger}\mathbf{T}\mathbf{A}^P = \mathbf{A}^{L\dagger}\mathbf{T}\mathbf{A}^L - (\mathbf{A}^L - \mathbf{A}^P)^\dagger\mathbf{T}(\mathbf{A}^L - \mathbf{A}^P) \\
\tilde{\mathbf{S}}^{\text{NESC,cnt}} &= \mathbf{A}^{L\dagger}\mathbf{S}\mathbf{A}^L + \mathbf{A}^{P\dagger}\mathbf{T}\mathbf{A}^P/2mc^2
\end{aligned} \tag{19.53}$$

where we now take $\mathbf{A}^L$ and $\mathbf{A}^P$ to be the rectangular matrices of contraction coefficients. What we do in practice then is to store two sets of contraction coefficients, $\mathbf{A}^L$ and $\mathbf{A}^P$.

If we make the two-electron approximation given above, we will need to obtain $\mathbf{A}^N$, and for this we will need **X** for each atom. $\mathbf{A}^N$ will be used to contract the primitive two-electron integrals, but we can still use $\mathbf{A}^L$ and $\mathbf{A}^P$ for the one-electron integrals because the contracted one-electron integrals are independent of the representation we choose for the calculation. This can be seen by performing the contraction, for example with the potential energy matrix:

$$\begin{aligned}
\tilde{\mathbf{V}}^{\text{NMFW,cnt}} &= \mathbf{A}^{N\dagger}\mathbf{S}^{1/2}\tilde{\mathbf{S}}^{-1/2}(\mathbf{V} + \mathbf{X}^\dagger\mathbf{V}^p\mathbf{X}/4m^2c^2)\tilde{\mathbf{S}}^{-1/2}\mathbf{S}^{1/2}\mathbf{A}^N \\
&= \mathbf{A}^{L\dagger}\mathbf{V}\mathbf{A}^L + \mathbf{A}^{P\dagger}\mathbf{V}^p\mathbf{A}^P/4m^2c^2 = \tilde{\mathbf{V}}^{\text{NESC,cnt}}
\end{aligned} \tag{19.54}$$

where we have used (19.34) to back-transform $\mathbf{A}^N$. In this approximation, we have three sets of contraction coefficients, $\mathbf{A}^L$, $\mathbf{A}^P$, and $\mathbf{A}^N$.

As we have already discussed, the largest amount of work will be expended in the evaluation of the two-electron relativistic integrals. One method for reducing the work was presented in the previous section, but we cannot use this approximation if we are interested in retaining the spin–orbit effect or if the spin-free two-electron contribution to the energy is not adequately represented. If the molecules we want to perform calculations on are composed entirely of heavy atoms we probably have no alternatives. If, on the other hand, we are interested in coordination chemistry, in which we have a few heavy atoms surrounded by large organic ligands, it would be convenient to include the relativistic effects only for the heavy atoms and not for the ligands. We do not expect relativistic effects to be large for organic molecules.

One possible scheme for doing this is to classify the atoms into two groups: "relativistic" and "nonrelativistic" (Dyall and Enevoldsen 1999). Relativistic effects are included only for the relativistic atoms and interactions between the relativistic and nonrelativistic atoms. We have already made a block-diagonal atomic approximation for **X**, and we now make a further approximation. What distinguishes relativistic from nonrelativistic is the speed of light. If we classify atoms into these two groups, we are selecting a value of the speed of light for each atom: for relativistic atoms it has the physical value, and for nonrelativistic atoms it is infinite, or alternatively the inverse is zero.

The effect of setting $1/c = 0$ on the one-electron integrals, (19.16), is to zero out the relativistic correction. For $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{S}}$ this follows immediately; for $\tilde{\mathbf{T}}$ it follows from the fact that $\mathbf{X} \to \mathbf{I}$ and therefore the correction vanishes. This is just the nonrelativistic limit, which we have already discussed. When the basis functions in the integral come from one nonrelativistic and one relativistic center, we must consider that one factor of $1/2mc$ belongs to each basis function, because it came from the transformation from the small component to the pseudo-large component. Therefore, the relativistic corrections will also be zero in this case. The integral will just be the nonrelativistic integral over the large components, which in the nonrelativistic case is simply the nonrelativistic wave function. The same applies to the two-electron integrals. The general principle is that the relativistic correction survives only when all the pseudo-large components come from relativistic centers. So, for example, if we have a single relativistic center, the only relativistic corrections to the two-electron integrals that will survive are the one-center integrals and those with two pseudo-large components on the relativistic center.

When combined with a frozen core, this approximation is equivalent to pseudopotential or ab initio model potential methods, which are developed in chapter 20. These methods incorporate relativistic effects into a one-electron operator for heavy atoms, and the rest of the molecule is treated nonrelativistically.

The remaining issue is the dividing line between “relativistic” and “nonrelativistic.” Where the line is drawn will depend on the purpose and the desired accuracy of the calculations. For “chemical accuracy” of 4 kJ/mol, atoms to the end of the third period, that is, up to Ar, can probably be treated nonrelativistically. These include all the common elements used for ligands. But a caution must be sounded also, because polarization of sulfur- and phosphorus-containing ligands by a metal can result in relativistic corrections of a few kilojoules per mole. In the long run, where the line can be drawn will have to be determined by careful testing. Ultimately, this approximation will be useful for calculations in which high accuracy is not demanded.

19.5 Properties in the Matrix Approximations

The calculation of first-order properties in the atom-centered matrix approximations is almost absurdly simple, particularly for contracted basis sets. What we have done in the approximations is to expand the molecular wave function in a set of atomic positive-energy one-particle functions. So all we need do for the first-order properties is to evaluate the primitive property integrals and transform them. In an uncontracted basis, the electric and magnetic property matrices are given by

$$\begin{aligned} \tilde{\mathbf{W}} &= \mathbf{W} + \mathbf{X}^\dagger \mathbf{W}^p \mathbf{X}/4m^2c^2 \\ \tilde{\mathbf{Y}} &= \mathbf{Y}\mathbf{X} + \mathbf{X}^\dagger \mathbf{Y}^\dagger \end{aligned} \tag{19.55}$$

where $\mathbf{W}$ is the matrix of the electric operator W, $\mathbf{W}^p$ is the matrix of the operator $(\boldsymbol{\sigma} \cdot \mathbf{p}) W (\boldsymbol{\sigma} \cdot \mathbf{p})$ or $\mathbf{p}W \cdot \mathbf{p}$, and $\mathbf{Y}$ is the matrix of the operator $(\boldsymbol{\sigma} \cdot \mathbf{A})(\boldsymbol{\sigma} \cdot \mathbf{p})/2m$, all in the primitive basis. In the contracted basis we can express the property matrices in terms of $\mathbf{A}^L$ and $\mathbf{A}^P$:

$$\begin{aligned}\tilde{\mathbf{W}}^{\mathrm{cnt}} &= \mathbf{A}^{L\dagger}\mathbf{W}\mathbf{A}^L + \mathbf{A}^{P\dagger}\mathbf{W}^p\mathbf{A}^P/4m^2c^2 \\ \tilde{\mathbf{Y}}^{\mathrm{cnt}} &= \mathbf{A}^{L\dagger}\mathbf{Y}\mathbf{A}^P + \mathbf{A}^{P\dagger}\mathbf{Y}^\dagger\mathbf{A}^L.\end{aligned} \tag{19.56}$$

For second-order properties there is a further approximation we must make, because the sum for the first-order wave function extends over both positive- and negative-energy states and we have discarded the negative-energy states. This will not matter very much for electric properties because the contribution from the negative-energy states is $\mathcal{O}(c^{-4})$. But this will not be the case for the magnetic properties, where the contributions from the positive-energy states and the negative-energy states are both $\mathcal{O}(c^0)$, as we saw in chapter 13. We must therefore find a strategy for second-order magnetic properties.

The problem arises because the magnetic property operator is an odd (off-diagonal) operator, that is, it connects large and small components. One way to solve this problem is to convert it to an even operator using the approach outlined in section 13.7. The magnetic operator correct to $\mathcal{O}(c^0)$ is

$$\hat{M} = \frac{e}{2m}\beta(2\mathbf{A}\cdot\mathbf{p} + \hbar\boldsymbol{\Sigma}\cdot\mathbf{B}) + \frac{e^2}{2m}\beta\mathbf{A}^2. \tag{19.57}$$

This is essentially the nonrelativistic operator multiplied by β. We must transform this operator to the modified Dirac representation, in which we have

$$\tilde{Y}' = \begin{pmatrix} \hat{M}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & -(\boldsymbol{\sigma}\cdot\mathbf{p})\hat{M}_2(\boldsymbol{\sigma}\cdot\mathbf{p})/4m^2c^2 \end{pmatrix} \tag{19.58}$$

where $\hat{M}_2$ is the nonrelativistic two-component magnetic property operator,

$$\hat{M}_2 = \frac{e}{2m}(2\mathbf{A}\cdot\mathbf{p} + \hbar\boldsymbol{\sigma}\cdot\mathbf{B}) + \frac{e^2}{2m}\mathbf{A}^2. \tag{19.59}$$

The magnetic perturbation Hamiltonian $\tilde{Y}'$ now behaves like an electric perturbation, which is what we want for the approximations of this chapter. We can evaluate the matrix elements of this new operator in the same way as we did for the electric perturbations. There is a relativistic correction to the magnetic operator $\hat{M}$, but it requires a second transformation of the perturbed Hamiltonian, and is discussed in section 13.7. As stated for the approximations to the energy, these approximations are useful where high accuracy is not required.

20

Core Approximations

When we think of chemical bonding, we usually think only of the valence orbitals. It is these orbitals that form bonds, and the core orbitals are not involved in the chemistry. To be sure, this is only a qualitative picture, but it raises the question of whether we really need to consider the core orbitals in our calculations.

For heavy elements, the number of core orbitals is not small. An element such as platinum, from the third transition series, has 30 orbitals in the first four shells that could be classed as core orbitals. If these remain essentially atomic over some region of the molecular potential energy surface, they might as well be fixed in their atomic form. That is, we would make a *frozen-core* approximation, and all that the core orbitals are doing is supplying a nonlocal static potential that could be evaluated once and used for the remainder of the calculation. As relativistic effects are to a large extent localized in the core region, they could be included in the frozen-core potential. We could then treat the valence orbitals and the orbitals on the light atoms nonrelativistically, as we did in the previous chapter. This would save all the work of calculating the relativistic integrals, and the calculation would be as cheap as a nonrelativistic calculation.

There is one main difficulty with this idea, and that is the orthogonality of the rest of the orbitals to the frozen core. The basis sets we use in molecular calculations are not automatically orthogonal to the core of any one atom: we must make them so by some procedure, such as Schmidt orthogonalization. But this involves taking linear combinations of the core and valence orbitals, and then we not only have to calculate all the integrals involving the core, we also have to transform them to the orthogonal basis. The reintroduction of the core integrals means that we have to calculate all the relativistic contributions that we had previously put into the frozen-core potential. Obviously, this is not a satisfactory state of affairs.

Two solutions to this problem are in common use. One is to create a pseudopotential (PP) or effective core potential (ECP) that incorporates the core orthogonality terms along with the potential terms. The result is usually expressed as a combination of a local and a semilocal one-electron pseudopotential,

$$V^{\mathrm{PP}} = U(r) + \sum_{\ell} U_{\ell}(r) \sum_{m} |\ell m\rangle\langle \ell m|. \tag{20.1}$$

We will use the term pseudopotential for this kind of potential. In the second, the core orbitals are removed to a very high energy by a level-shifting procedure. As a result, any tendency of the valence orbitals to gain core character is energetically unfavorable. This second method was developed under the name ab initio model potentials (AIMP), though in some quarters this method is also referred to as an effective core potential or pseudopotential method. We will refer to these potentials as *model potentials*. The form of the potential is

$$V^{\mathrm{MP}} = \hat{J}^c - \hat{K}^c + \sum_{k} B_k |k\rangle\langle k| \tag{20.2}$$

where $\hat{J}^c$ and $\hat{K}^c$ are the core direct and exchange operators, and B_k is the level shift for core orbital k. In both approaches, the core orbitals are left out of the calculations, and the core tails of the valence orbitals can be approximated to remove the need for core-like basis functions.

Where do we draw the line between the core and the valence orbitals? If we take the valence orbitals to be those with the highest principal quantum number and assign the rest to the core, the further down the periodic table the element is, the more polarizable the core is. Thus, for those elements where the use of a core approximation is most desirable, the core orbitals are also most easily influenced by changes in valence shell electronic structure. Again, there are two main options: include more orbitals in the valence space or introduce a potential to describe the core polarization.

To formalize these concepts, we first develop the frozen-core approximation, and then examine pseudoorbitals or pseudospinors and pseudopotentials. From this foundation we develop the theory for effective core potentials and ab initio model potentials. The fundamental theory is the same whether the Hamiltonian is relativistic or nonrelativistic: the form of the one- and two-electron operators is not critical. Likewise with the orbitals: the development here will be given in terms of spinors, which may be either nonrelativistic spin–orbitals or relativistic 2-spinors derived from a Foldy–Wouthuysen transformation. We will use the terminology “pseudospinors” because we wish to be as general as possible, but wherever this term is used, the term “pseudoorbitals” can be substituted. As in previous chapters, the indices p, q, r, and s will be used for any spinor, t, u, v, and w will be used for valence spinors, and a, b, c, and d will be used for virtual spinors. For core spinors we will use k, l, m, and n, and reserve i and j for electron indices and other summation indices.

In our exposition of effective core potentials, we follow one particular line of development, represented by the work of Kahn, Baybutt, and Truhlar (1976) and the developments that followed from it. We recognize, however, that there have been many lines of developments, some earlier, such as that of Durand and Barthelat (1975), some in parallel with this one, and many procedures for obtaining pseudopotentials. Fortunately, there have been many reviews of pseudopotentials, which can be located either in Pyykkö's bibliographies (Pyykkö 1986, 1993, 2000) or in a review by Pyykkö and Stoll (2000). An excellent overview of both pseudopotentials and model potentials is provided by Seijo and Barandiarán (1999).

20.1 The Frozen-Core Approximation

From a formal point of view, we can represent the spinors of any system of interest in terms of a complete orthonormal set of spinors located on a single center. These spinors include both bound state and continuum spinors and range over all angular momenta. In an application, the particular choice of spinors will be determined by the state of the atom whose core we wish to freeze. But for the purpose of analysis, it does not matter which spinors we choose provided the set is complete and orthonormal.

The n_c core spinors selected from this set are occupied in all possible configurations. The many-electron wave function is partitioned into a core and a valence part,

$$\Psi(1,\ldots,n) = \hat{\mathcal{A}}\Psi^c(1,\ldots,n_c)\;\Psi^v(n_c+1,\ldots,n), \tag{20.3}$$

where $\hat{\mathcal{A}}$ is the antisymmetrizer and the core wave function is

$$\Psi^c(1,\ldots,n_c) = \hat{\mathcal{A}}\prod_k^{n_c}\psi_k. \tag{20.4}$$

Next, we integrate the contribution due to the core electrons out of the Hamiltonian, and work with a valence Hamiltonian and a valence wave function. The valence Hamiltonian (from which we exclude the core energy) is

$$\hat{\mathcal{H}}^v = \sum_i^{n_v}\hat{f}^c(i) + \frac{1}{2}\sum_{i\neq j}^{n_v}\frac{1}{r_{ij}}. \tag{20.5}$$

The sums now range over the n_v valence electrons. The one-electron operator has been replaced by the core Fock operator,

$$\hat{f}^c = \hat{h} + \sum_k^{n_c}(\hat{J}_k - \hat{K}_k), \tag{20.6}$$

with $\hat{J}_k$ and $\hat{K}_k$ the operators of the direct ("Coulomb") and exchange potentials for spinor k. This Hamiltonian is precisely the Hamiltonian we would use in a correlated calculation with the same frozen core. If we maintain orthogonality to the core at all times, this is as far as we need to go. The direct potential is a local potential, which could be accurately fitted, and the exchange potential is at most a three-index quantity. But, orthogonality to the core introduces the core spinors into the valence through the orthogonalization procedure, so we need to go further.

In order to be able to relax the constraints of orthogonality to the core, we must now introduce projection operators. The core projection operator is

$$\hat{\mathcal{P}}^c = \sum_k^{n_c} |\,\psi_k\,\rangle\langle\,\psi_k\,| \tag{20.7}$$

and the valence projection operator is obtained from the closure requirement,

$$\hat{\mathcal{P}}^v = 1 - \hat{\mathcal{P}}^c. \tag{20.8}$$

Every operator must now be bracketed by valence projection operators, including the metric operator, so we must insert the valence projection operator in every place where we would normally compute an overlap integral. This converts every one-electron operator into an N-electron operator,

$$\hat{h}(i) \rightarrow \hat{\mathcal{P}}^v(i)\hat{h}(i)\hat{\mathcal{P}}^v(i)\prod_{j\neq i}\hat{\mathcal{P}}^v(j). \tag{20.9}$$

and every two-electron operator similarly into an N-electron operator. In the present case, $N = n_v$.

This is not as bad as it sounds. If we ensure that the spinors are orthonormal on the valence projection operator,

$$\langle\,\psi_p\,|\hat{\mathcal{P}}^v\,|\,\psi_q\,\rangle = \delta_{pq}, \tag{20.10}$$

then the orthogonality requirements are automatically satisfied. We need no longer explicitly consider the projection operators inserted for orthogonality, and we can write the operators as

$$\begin{aligned}
\hat{h}^v(i) &= \hat{\mathcal{P}}^v(i)\hat{h}(i)\hat{\mathcal{P}}^v(i)\\
\hat{g}^v(i,j) &= \hat{\mathcal{P}}^v(i)\hat{\mathcal{P}}^v(j)\hat{g}(i,j)\hat{\mathcal{P}}^v(j)\hat{\mathcal{P}}^v(i)\\
\hat{G}^v(i) &= \hat{\mathcal{P}}^v(i)
\end{aligned} \tag{20.11}$$

where we have also included the metric operator $\hat{G}^v$, which is unity in the full orthonormal set of spinors including the core but deviates from unity here because of the frozen core.

Once the projection operators have been introduced we may remove the requirement that the valence spinors should be orthogonal to the core spinors. From the properties of determinants, we know that we can always add a linear combination of the core spinors to the valence spinors without changing the total wave function. The resulting spinor we term a *pseudospinor*,

$$\phi_v = \psi_v + \sum_k a_{kv}\psi_k. \tag{20.12}$$

This pseudospinor is normalized on the metric $\hat{G}^v$ and is orthogonal to other pseudospinors on the same metric. The coefficients can be determined to minimize the density of the pseudospinor in the core region. For this reason, we will call the sum of core spinors the *core tail*,

$$\psi_v^{\mathrm{ct}} = \sum_k a_{kv}\psi_k. \tag{20.13}$$

At this point, we have a valence wave function given in terms of pseudospinors that have little or no core contribution, and a set of valence operators that only operate on the valence space. We do however have explicit use of the valence projection operators, which are composed of infinite sums. What we would like to do is to write the Hamiltonian in terms of the "normal" Hamiltonian and a correction, which is termed a *pseudopotential*. This we can do by replacing $\hat{\mathcal{P}}^v$ with $1 - \hat{\mathcal{P}}^c$ and extracting out the unprojected Hamiltonian,

$$\hat{\mathcal{H}}^v = \sum_i \hat{h}(i) + \tfrac{1}{2}\sum_{ij} g(i,j) + \hat{V}^{\mathrm{FCPP}}. \tag{20.14}$$

The sums are over the valence electrons only, and the frozen-core pseudopotential is

$$\begin{aligned}
\hat{V}^{\mathrm{FCPP}} = {} & \sum_i \Big[\hat{V}^c(i) - \hat{\mathcal{P}}^c(i)\hat{V}^c(i) - \hat{V}^c(i)\hat{\mathcal{P}}^c(i) + \hat{\mathcal{P}}^c(i)\hat{V}^c(i)\hat{\mathcal{P}}^c(i)\Big] \\
& + \frac{1}{2}\sum_{i\neq j}\Big[-\hat{\mathcal{P}}^c(i) - \hat{\mathcal{P}}^c(j) + \hat{\mathcal{P}}^c(i)\hat{\mathcal{P}}^c(j)\Big] g(i,j) \\
& + g(i,j)\Big[-\hat{\mathcal{P}}^c(i) - \hat{\mathcal{P}}^c(j) + \hat{\mathcal{P}}^c(i)\hat{\mathcal{P}}^c(j)\Big] \\
& + \Big[-\hat{\mathcal{P}}^c(i) - \hat{\mathcal{P}}^c(j) + \hat{\mathcal{P}}^c(i)\hat{\mathcal{P}}^c(j)\Big] g(i,j) \Big[-\hat{\mathcal{P}}^c(i) - \hat{\mathcal{P}}^c(j) + \hat{\mathcal{P}}^c(i)\hat{\mathcal{P}}^c(j)\Big]
\end{aligned} \tag{20.15}$$

where the core potential $\hat{V}^c$ is

$$\hat{V}^c = \sum_k^{n_c} (\hat{J}_k - \hat{K}_k). \tag{20.16}$$

and the sum is over all core spinors. This complicated frozen-core pseudopotential contains both one- and two-electron terms, and the core spinors remain in the projection operators both in the pseudopotential and in the metric operator. And in fact, applied just as it is, it does not solve any of the problems of the number of integrals or the inclusion of relativistic effects in a potential.

Before concluding this section, we should add that frozen-core approximations of various (often ad hoc) forms have a well-established place in quantum chemistry. These are, however, not the central theme of this chapter. Here, we have used the frozen-core approximation only as a necessary precursor to the discussion of pseudopotentials.

20.2 The Generalized Philips–Kleinman Pseudopotential

What we would like to do is to remove all explicit reference to core spinors and to core basis functions from the Hamiltonian, incorporating their effect into some kind of local or nonlocal potential. The foundation for all such approaches is the Philips–Kleinman (1959) procedure, which was generalized by Weeks and Rice (1968). It starts from the all-electron self-consistent field (SCF) equation, which we will assume here is converged,

$$\hat{f}\psi_p = \sum_{q=1}^{n} \epsilon_{qp}\psi_q, \tag{20.17}$$

where $\hat{f}$ is the Fock operator,

$$\hat{f} = \hat{h} + \sum_{q=1}^{n}(\hat{J}_q - \hat{K}_q), \tag{20.18}$$

and the ϵ_{qp} are Lagrange multipliers, introduced to ensure orthogonality. They are defined by

$$\langle\psi_q|\hat{f}|\psi_p\rangle = f_{qp} = \epsilon_{qp}. \tag{20.19}$$

If ψ_p is a valence spinor, we may insert the valence projection operator as defined by (20.8) into (20.17) to obtain a projected equation that is satisfied for any valence spinor,

$$(1 - \hat{\mathcal{P}}^c)\hat{f}(1 - \hat{\mathcal{P}}^c)\psi_v = \sum_{u=1}^{n_v} \epsilon_{uv}(1 - \hat{\mathcal{P}}^c)\psi_u. \tag{20.20}$$

The sum on the right is restricted to the valence spinors because the terms involving the core spinors vanish. Now that we have introduced the projection operators, the equation is satisfied by the valence pseudospinors ϕ_v of (20.12) as well. This is the SCF equation we would get from the projected frozen-core Hamiltonian.

If there is only one valence spinor, and we substitute the pseudospinor ϕ_v for the spinor ψ_v, (20.20) simplifies to

$$(1 - \hat{\mathcal{P}}^c)\hat{f}(1 - \hat{\mathcal{P}}^c)\phi_v = \epsilon_v(1 - \hat{\mathcal{P}}^c)\phi_v. \tag{20.21}$$

To write this equation in terms of the normal operator plus a correction, we move the projector from the right-hand side and rearrange,

$$\hat{f}\phi_v + \left[\epsilon_v \hat{\mathcal{P}}^c - \hat{\mathcal{P}}^c \hat{f} - \hat{f}\hat{\mathcal{P}}^c + \hat{\mathcal{P}}^c \hat{f} \hat{\mathcal{P}}^c\right]\phi_v = \epsilon_v \phi_v. \tag{20.22}$$

The term in the square brackets is the generalized Philips–Kleinman pseudopotential, $\hat{V}_v^{\mathrm{GPK}}$,

$$\hat{V}_v^{\mathrm{GPK}} = \epsilon_v \hat{\mathcal{P}}^c - \hat{\mathcal{P}}^c \hat{f} - \hat{f}\hat{\mathcal{P}}^c + \hat{\mathcal{P}}^c \hat{f} \hat{\mathcal{P}}^c \tag{20.23}$$

and we may write (20.22) as

$$(\hat{f} + \hat{V}_v^{\mathrm{GPK}})\phi_v = \epsilon_v \phi_v. \tag{20.24}$$

Since the core spinors are eigenfunctions of $\hat{f}$, the generalized Philips–Kleinman pseudopotential can be written

$$\hat{V}_v^{\mathrm{GPK}} = \sum_k (\epsilon_v - \epsilon_k) |\psi_k\rangle\langle\psi_k|. \tag{20.25}$$

We would also like to partition the Fock operator into a valence and a core part: this we will do later. In the rest of this section, we explore the development of the generalized Philips–Kleinman pseudopotential for many valence electrons and many valence spinors.

The generalized Philips–Kleinman pseudopotential depends on the eigenvalue of the spinor v, unlike the frozen-core pseudopotential, which depends only on the core spinors. The appearance of the ϵ_v term in the pseudopotential came about because we transferred a term from right-hand side of (20.21). This means that we have changed the metric, which has implications for orthogonality that we pursue later. The new operator makes the core spinors degenerate with the valence spinor:

$$(\hat{f} + \hat{V}_v^{\mathrm{GPK}})\psi_k = \epsilon_v \psi_k. \tag{20.26}$$

The formation of a pseudospinor follows naturally: we can always mix degenerate spinors. There is a danger here as well. If a reasonable representation of a core spinor could be made from the basis set, it too would have an eigenvalue equal to the valence spinor. Moreover, neither the overlap nor the one-electron matrix element between the core spinor and the valence pseudospinor would be zero. This would complicate matters in a real system, and it is therefore imperative to keep core-like functions out of the basis set.

An atom in which we can only have one spinor in the valence region is not very interesting. We need a theory in which there are a number of valence spinors, and perhaps more than one of any given symmetry. Furthermore, we are interested not only in the occupied spinors but also in the virtual spinors or some combination of them, since it is these that will be used to provide the needed basis set flexibility in a molecular calculation.

For spinors of different symmetries there is no problem: we simply construct the pseudospinors in the same way for each symmetry. For an atomic shell the radial factorization and spinor degeneracy enable us to construct a single radial pseudopotential for the entire shell. And because there is rarely more than one symmetry in what we consider to be the valence spinors, this permits us to construct valence pseudopotentials for most atoms in the periodic table.

If we want pseudopotentials for more than one spinor of the same symmetry, including the virtual spinors needed for flexibility in molecular calculations, we need to start from the more general equation, (20.20), with pseudospinors substituted for spinors,

$$(1 - \hat{\mathcal{P}}^c)\hat{f}(1 - \hat{\mathcal{P}}^c)\phi_v = \sum_{u=1}^{n_v} \epsilon_{uv}(1 - \hat{\mathcal{P}}^c)\phi_u. \tag{20.27}$$

To construct a pseudopotential, we must transfer the core projection operator with the Lagrange multipliers to the left side. The question is, how do we do this?

Since the off-diagonal Lagrange multipliers are zero at convergence, we could neglect them. Each spinor would satisfy an equation of the form (20.22) with a pseudopotential of the form (20.23). These pseudopotentials are different for each spinor because they contain the spinor eigenvalue. In addition, the pseudospinors are no longer orthogonal because the transfer of the core projection operators into the pseudopotential changes the metric: the overlap of two pseudospinors on the new, unit metric is

$$\langle \phi_v | \phi_u \rangle = \delta_{vu} + \langle \psi_v^{\mathrm{ct}} | \psi_u^{\mathrm{ct}} \rangle = \delta_{vu} + \sum_k a_{kv} a_{ku}. \tag{20.28}$$

To fix this problem, we could impose a new orthogonality constraint by introducing new Lagrange multipliers λ_{uv},

$$(\hat{f} + \hat{V}_v^{\mathrm{GPK}})\phi_v = \sum_{u=1}^{n_v} \lambda_{uv}\phi_v. \tag{20.29}$$

Multiplying from the left by $\phi_u \hat{\mathcal{P}}^c$ and integrating we get

$$\epsilon_v \langle \phi_u | \hat{\mathcal{P}}^c | \phi_v \rangle = \lambda_{uv} \langle \phi_u | \hat{\mathcal{P}}^c | \phi_v \rangle \tag{20.30}$$

If we interchange v and u in the above derivation and require that the matrix of Lagrange multipliers be Hermitian, we get the equation

$$\epsilon_v \langle \phi_u | \hat{\mathcal{P}}^c | \phi_v \rangle = \epsilon_u \langle \phi_v | \hat{\mathcal{P}}^c | \phi_u \rangle. \tag{20.31}$$

The integral is the overlap of the core tails, and we conclude that either this must be zero or $\epsilon_v = \epsilon_u$. The core tail overlap cannot in general be zero—the case of a single core spinor demonstrates this point—so the spinors must be degenerate. But because the spinors are of the same symmetry (and of the same row in a degenerate irrep), this is not possible. Moreover, the eigenvalues are the eigenvalues of the Hartree–Fock spinors, which are definitely not degenerate. We conclude that the pseudospinors cannot be made orthogonal on the new metric by a simple rotation. Not only that, we also have a pseudopotential for each spinor, which would be impractical. To implement such a scheme we would have to use projection operators onto each spinor, so the pseudopotential would be

$$V^{\mathrm{PP}} = \sum_{n\kappa m} |n\kappa m\rangle\langle n\kappa m|V^{\mathrm{PP}}_{n\kappa m}|n\kappa m\rangle\langle n\kappa m|, \tag{20.32}$$

and the only information in the pseudopotential is the matrix element for a given spinor. This is essentially a spectral representation of the pseudopotential.

An alternative is to transfer the original Lagrange multipliers ϵ_{uv} to the pseudopotential. This can be done by writing the Lagrange multipliers in terms of matrix elements of the pseudospinors,

$$\langle\phi_u|(1-\hat{\mathcal{P}}^c)\hat{f}(1-\hat{\mathcal{P}}^c)|\phi_v\rangle = \epsilon_{uv}. \tag{20.33}$$

Splitting the projector $1-\hat{\mathcal{P}}^c$ on the right-hand side of (20.27), and substituting for ϵ_{uv}, we get

$$\begin{aligned}(1-\hat{\mathcal{P}}^c)\hat{f}(1-\hat{\mathcal{P}}^c)\phi_v &= \sum_{u=1}^{n_v}\phi_u\epsilon_{uv} - \sum_{u=1}^{n_v}\hat{\mathcal{P}}^c|\phi_u\rangle\langle\phi_u|(1-\hat{\mathcal{P}}^c)\hat{f}(1-\hat{\mathcal{P}}^c)\phi_v \\ &= \sum_{u=1}^{n_v}\phi_u\epsilon_{uv} - \hat{\mathcal{P}}^c\hat{\mathcal{P}}^{\tilde{v}}(1-\hat{\mathcal{P}}^c)\hat{f}(1-\hat{\mathcal{P}}^c)\phi_v,\end{aligned} \tag{20.34}$$

where we have defined the projector onto the valence pseudospinors, $\hat{\mathcal{P}}^{\tilde{v}}$,

$$\hat{\mathcal{P}}^{\tilde{v}} = \sum_{u=1}^{n_v}|\phi_u\rangle\langle\phi_u|. \tag{20.35}$$

Transferring the last term on the right-hand side to the left-hand side, the pseudospinor equation becomes

$$(1+\hat{\mathcal{P}}^c\hat{\mathcal{P}}^{\tilde{v}})(1-\hat{\mathcal{P}}^c)\hat{f}(1-\hat{\mathcal{P}}^c)\phi_v = \sum_{j=1}^{n_v}\epsilon_{ji}\phi_u. \tag{20.36}$$

Now we can define a single pseudopotential, but it is no longer Hermitian. This is a drawback for practical purposes. The pseudospinors are still not orthogonal, and cannot be made orthogonal and eigenfunctions of the pseudooperator by a simple rotation.

We conclude that regardless of how we construct the pseudopotential, for several pseudospinors of the same symmetry we cannot construct a potential that makes the pseudospinors orthogonal. We do not wish to have a pseudopotential for every pseudospinor, virtual and occupied,[1] or to have a non-Hermitian pseudopotential. These constraints lead to a further approximation: the pseudopotential for all spinors of a given symmetry is chosen to be the pseudopotential for the lowest spinor. It is this approximation that ultimately permits us to write the pseudopotential in the semilocal form of (20.1).

For the main group elements, the lowest spinor is usually one of the valence spinors, which means that at least the valence pseudospinors for the atom will not be approximated, and these are the most important spinors in a molecular calculation. However, for the transition metals, it is often important to have both the ns and the $(n+1)s$ spinors in the valence space. The pseudopotential must then be chosen for the ns pseudospinor, and this choice will affect the $(n+1)s$ pseudospinor.

To examine how this choice affects the pseudospinors, we need to find the eigenvalues and eigenvectors of the pseudospinor equation

$$\left[(1-\hat{\mathcal{P}}^c)\hat{f}(1-\hat{\mathcal{P}}^c)+\epsilon_v\hat{\mathcal{P}}^c\right]\tilde{\phi}=\tilde{\epsilon}\tilde{\phi}. \tag{20.37}$$

This equation is the SCF equation for the lowest pseudospinor, ϕ_v, and so $\tilde{\phi}=\phi_v$ is an eigenfunction of this equation with eigenvalue ϵ_v. What we are interested in is the other eigenfunctions of this equation.

We expand the solutions of the above equation in terms of the nonorthogonal pseudospinors,

$$\tilde{\phi}=\sum_u \tilde{c}_u\phi_u, \tag{20.38}$$

and set up and diagonalize the Fock matrix with the chosen pseudopotential. The elements of the Fock matrix in this basis are

$$\tilde{f}_{uw}=\langle\phi_u|(1-\hat{\mathcal{P}}^c)\hat{f}(1-\hat{\mathcal{P}}^c)+\epsilon_v\hat{\mathcal{P}}^c|\phi_w\rangle=\epsilon_u\delta_{uw}+\epsilon_v S^{\text{ct}}_{uw}. \tag{20.39}$$

The overlap matrix is

$$\tilde{S}_{uw}=\langle\phi_u|\phi_w\rangle=\delta_{uw}+S^{\text{ct}}_{uw} \tag{20.40}$$

where the overlap of the core tails is

$$S^{\text{ct}}_{uw}=\sum_k a_{ku}a_{kw}. \tag{20.41}$$

We want to solve the generalized eigenvalue equation $\tilde{\mathbf{f}}\tilde{\mathbf{c}}=\tilde{\mathbf{S}}\tilde{\mathbf{c}}\tilde{\boldsymbol{\epsilon}}$. We already know that ϕ_v is the lowest eigenfunction with eigenvalue ϵ_v. We can shift the energy scale by subtracting ϵ_v from the eigenvalue matrix, $\tilde{\boldsymbol{\epsilon}}'=\tilde{\boldsymbol{\epsilon}}-\epsilon_v\mathbf{I}$, and we must therefore

---

[1] An approach that includes a few pseudopotentials for each symmetry has been developed by Titov and Mosyagin (1999).

subtract $\epsilon_v\tilde{\mathbf{S}}\tilde{\mathbf{c}}$ from the Fock matrix. The new Fock matrix $\tilde{\mathbf{f}}'$ is now diagonal, because the core tail term cancels:

$$\tilde{f}'_{uw} = \langle \phi_u \,|\, (1 - \hat{\mathcal{P}}^c)\hat{f}(1 - \hat{\mathcal{P}}^c) + \epsilon_v \hat{\mathcal{P}}^c - \epsilon_v \,|\, \phi_w \rangle = (\epsilon_u - \epsilon_v)\delta_{uw}. \tag{20.42}$$

Although the Fock matrix is now diagonal, the overlap matrix remains the same.

There are two important conclusions that can be drawn. First, the fact that the overlap matrix is nondiagonal means that the higher eigenfunctions must mix in some of the lowest pseudospinor to maintain orthogonality. Second, since the diagonal elements of the overlap matrix are greater than 1, the eigenvalues $\tilde{\epsilon}$ are lower than the all-electron eigenvalues ϵ. The amount of the lowering of the eigenvalues depends on the magnitude of the core tail overlaps. The compression of the spectrum has consequences for the calculation of the correlation energy, for example: smaller denominators in an MP2 calculation will result in an overestimation of the correlation energies.

20.3 Shape-Consistent Pseudospinors and Pseudopotentials

The removal of the core projectors from the metric to the potential is only the first step in the removal of all references to the core spinors. The next step is the partitioning of the Fock operator into core and valence parts.

Our goal is to write the pseudopotential in a form that looks like an all-electron operator plus a one-electron correction term. We would also like to ensure that the pseudopotential has no long-range terms. To do this we partition the nuclear attraction terms as well as the electron repulsion terms into a core and a valence part,

$$\hat{f} = \hat{T} \;-\; Z^c f^{\mathrm{nuc}}(r) \;-\; Z^v f^{\mathrm{nuc}}(r) \;+\; \sum_k (\hat{J}_k - \hat{K}_k) \;+\; \sum_v (\hat{J}_v - \hat{K}_v) \tag{20.43}$$

where Z^c is equal to the number of core electrons, Z^v is equal to the number of valence electrons, and $f^{\mathrm{nuc}}(r)$ is the potential for a single charge. For a point nucleus it is $1/r$, while for a Gaussian nucleus it is $\mathrm{erf}(\eta^{1/2}r)/r$. The part of the nuclear potential that cancels the Coulomb potential of the core electrons is put in the pseudopotential. The one-electron operator is redefined with the remainder of the nuclear potential,

$$\hat{h}' = \hat{T} \;-\; Z^v f^{\mathrm{nuc}}(r). \tag{20.44}$$

The valence repulsion terms also need to be partitioned because these are expressed in terms of the spinors, not the pseudospinors. Expanding the spinors using (20.12) and (20.13), the direct potential is

$$\begin{aligned}
\hat{J}_v(i) &= \int \mathrm{d}^3 r_j \frac{1}{r_{ij}} \left[\phi_v(j)^\dagger \phi_v(j) - \phi_v(j)^\dagger \psi_v^{\mathrm{ct}}(j) - \psi_v^{\mathrm{ct}}(j)^\dagger \phi_v(j) + \psi_v^{\mathrm{ct}}(j)^\dagger \psi_v^{\mathrm{ct}}(j) \right] \\
&= \hat{J}_{\tilde{v}}(i) + \hat{J}_v^R(i).
\end{aligned} \tag{20.45}$$

The first term is the potential for the pseudospinor; the remainder, which includes all the core tail terms, goes into the pseudopotential. We can make a similar partitioning for the exchange potential.

So far, we have not normalized the pseudospinors, but we must do so now because we are using the unit metric with the pseudopotential. As we noted at the end of the previous section, the diagonal elements of the overlap matrix, which are the norms of the pseudospinors, are greater than 1. From (20.40), the normalized pseudospinor is given by

$$\chi_v = \frac{1}{\sqrt{1 + S_{vv}^{\mathrm{ct}}}} \phi_v. \tag{20.46}$$

The density for the normalized pseudospinor is

$$\rho_v = \frac{1}{1 + S_{vv}^{\mathrm{ct}}} {\phi_v}^{\dagger} \phi_v, \tag{20.47}$$

and it is this density that will be used to construct the potential for the normalized valence pseudospinors. We must therefore partition the direct potential further:

$$\hat{J}_v(i) = \frac{1}{1 + S_{vv}^{\mathrm{ct}}} \hat{J}_{\tilde{v}}(i) + \left[\frac{S_{vv}^{\mathrm{ct}}}{1 + S_{vv}^{\mathrm{ct}}} \hat{J}_{\tilde{v}}(i) + \hat{J}_v^R(i) \right]. \tag{20.48}$$

The exchange terms can be partitioned in an analogous fashion. The final form of the pseudopotential for the normalized pseudospinors is

$$\hat{V}^{\mathrm{PP}} = -Z^c f^{nuc}(r) + \sum_k (\hat{J}_k - \hat{K}_k) + \hat{V}^{\mathrm{GPK}} + \sum_v \left[\frac{S_{vv}^{\mathrm{ct}}}{1+S_{vv}^{\mathrm{ct}}} (\hat{J}_{\tilde{v}} - \hat{K}_{\tilde{v}}) + (\hat{J}_v^R - \hat{K}_v^R) \right]. \tag{20.49}$$

The first two terms are the core potential part, the third term is the orthogonality part, and the last term comes from the valence pseudospinors and their core tails.

The presence and form of this last term has some important consequences. The first is that the definition of the pseudopotential is determined to some extent by the definition of the pseudospinors. This means that the pseudospinors and the pseudopotentials are not independent. The only dependence on the valence spinors in $\hat{V}^{\mathrm{GPK}}$ comes from the spinor eigenvalue, but the remainder terms and the term that came from the renormalization contain the valence pseudospinors, and therefore the detailed shape of the potential is dependent on the shape of the valence pseudospinors.

The second consequence is that the pseudopotential has a long-range part that comes from the renormalization term. Beyond the range of the core spinors, the valence electron density from the normalized pseudospinors is less than the Hartree–Fock density because of the normalization. The reduced density results in a reduced Coulomb repulsion from the valence electrons, and the reduction in the Coulomb repulsion is compensated for in the pseudopotential. The decrease in electron density in the outer region has a corresponding increase in density within the range of the core spinors, and the pseudopotential must be attractive in this inner region.

When the number of valence electrons is the same as in the Hartree–Fock state for which the pseudopotential was derived, these shifts in density and corresponding attractive and repulsive regions of the pseudopotential have no net effect. However, in a molecule, the valence electron occupation is almost always different from the atomic Hartree–Fock occupation, and these attractive and repulsive regions of the pseudopotential cause problems such as bond lengths that are too short and potential wells that are too deep.

This effect presents some serious problems for the development of pseudopotentials. A pseudopotential that depends critically on the shape of the pseudospinor and for which the results are sensitive to the valence occupation is of no value. The problem was overcome (Christiansen et al. 1979) by the definition of the so-called *shape-consistent* pseudospinors and the corresponding pseudopotentials.

The shape-consistent pseudospinor is defined to be equal to the Hartree–Fock spinor outside a certain radius r_C, and inside this radius it falls smoothly to zero,

$$\begin{aligned} \phi_v^{\mathrm{SC}}(r) &= \psi_v(r), \qquad r > r_C \\ &= g_v(r), \qquad r \le r_C. \end{aligned} \tag{20.50}$$

Alternatively, the shape-consistent pseudospinor can be written in terms of the Hartree–Fock spinor and a function that is confined inside the chosen radius r_C, $f_v(r) = 0,\ r > r_C$,

$$\phi_v^{\mathrm{SC}}(r) = \psi_v(r) + f_v(r). \tag{20.51}$$

To ensure that there is no renormalization problem, this pseudospinor must itself be normalized,

$$\begin{aligned} 1 &= \int r^2 \mathrm{d}r\ |\phi_v^{\mathrm{SC}}(r)|^2 \\ &= \int r^2 \mathrm{d}r\ |\psi_v(r)|^2 + 2\,\mathrm{Re} \int r^2 \mathrm{d}r\ \psi_v(r)^\dagger f_v(r) + \int r^2 \mathrm{d}r\ |f_v(r)|^2. \end{aligned} \tag{20.52}$$

The spinor ψ_v is normalized, so we deduce that

$$2\,\mathrm{Re} \int r^2 \mathrm{d}r\ \psi_v(r) f_v(r) = - \int r^2 \mathrm{d}r\ |f_v(r)|^2 \tag{20.53}$$

and the overlap between the function f_v and the spinor ψ_v must be negative.

Because we have a complete orthonormal set of spinors, we can expand the function f_v in this set:

$$f_v = \sum_p d_{vp} \psi_p. \tag{20.54}$$

Although f_v is confined to the core region, it cannot be composed only of core spinors. If it were, the overlap with the spinor ψ_v would be zero, and that would make f_v zero.

In fact, the only way that the overlap with the spinor ψ_v can be nonzero is if f_v contains ψ_v, and then the overlap is

$$\int r^2 \mathrm{d}r\ \psi_v(r) f_v(r) = d_{vv} = -\tfrac{1}{2} \int r^2 \mathrm{d}r\ |f_v(r)|^2. \tag{20.55}$$

The negative sign for the coefficient of ψ_v implies a reduction in the amplitude of the pseudospinor in the valence region. So, if f_v only contains core and valence spinors, we are in the same situation as for the Philips–Kleinman pseudospinor, with a reduction in the valence density outside the chosen radius and the resultant problems with the pseudopotential. This means that f_v must contain virtual spinors as well as occupied spinors, and the sum of virtual spinors must compensate for the reduction in the valence spinor outside the radius of the core spinors (which is not necessarily the same as r_C).

In practice, it is not the function $f_v(r)$ but the function $g_v(r)$ that is used to determine the pseudospinor, by polynomial expansion in r. The coefficients are determined by equating the derivatives of the polynomial and the Hartree–Fock spinor at the matching point, r_C, with the constraints that the pseudospinor has the minimum number of turning points and is normalized. Further details on the methodology for generating pseudopotentials is given later.

The introduction of shape-consistent pseudospinors solved the problems in the pseudopotential that were caused by the use of the Philips–Kleinman pseudospinors. However, the other characteristics of pseudospinors that were discussed in the previous section still apply to shape-consistent pseudospinors. The inclusion of virtual spinors in the expansion of the core tail does not alter the conclusions drawn: all but the lowest pseudospinor mix, and the eigenvalue spectrum is compressed.

The inclusion of virtual spinors in the pseudospinors also represents a departure from the strict frozen-core approximation. Mixing in virtual spinors means that the valence wave function is no longer equivalent to the all-electron version, and that the valence energy is no longer the same. However, we can always project out the appropriate linear combinations of the virtual spinors, as well as the core spinors, to remove their effect, and replace the core projector with a projector that includes these virtual spinors.

20.4 Energetics of Pseudopotentials

We have so far discussed the theory and characteristics of pseudospinors and the corresponding one-particle pseudopotentials. The next stage is to examine the use of these pseudopotentials in a valence-only Hamiltonian and the properties of the energy and the SCF equations derived from this energy. The desired form of the pseudopotential Hamiltonian is one in which only the one-electron operator is modified,

$$\hat{\mathcal{H}}^{\mathrm{PP}} = \sum_i \left[\hat{h}'(i) + \hat{V}^{\mathrm{PP}}(i) \right] + \frac{1}{2} \sum_{ij} \frac{1}{r_{ij}}. \tag{20.56}$$

The sums run over valence electrons only.

The valence energy using the frozen-core Hamiltonian can be written as

$$\begin{aligned} E^{\mathrm{val}} &= \sum_u \langle u|\hat{f}^c|u\rangle + \tfrac{1}{2}\sum_{uv}(J_{uv} - K_{uv}) \\ &= \sum_u \epsilon_u - \tfrac{1}{2}\sum_{uv}(J_{uv} - K_{uv}). \end{aligned} \tag{20.57}$$

The direct and exchange integrals are given in Mulliken notation by

$$J_{uv} = (uu|vv), \quad K_{uv} = (uv|vu). \tag{20.58}$$

The valence energy using the pseudopotential Hamiltonian can be written as

$$\begin{aligned} E^{\mathrm{PP}} &= \sum_u \langle\tilde{u}|\hat{h}' + \hat{V}^{\mathrm{PP}}|\tilde{u}\rangle + \tfrac{1}{2}\sum_{uv}(J_{\tilde{u}\tilde{v}} - K_{\tilde{u}\tilde{v}}) \\ &= \sum_u \epsilon_u - \tfrac{1}{2}\sum_{uv}(J_{\tilde{u}\tilde{v}} - K_{\tilde{u}\tilde{v}}) \end{aligned} \tag{20.59}$$

where, as before, the tilde distinguishes pseudospinors from spinors. The valence spinor energies are the same whether or not the pseudopotential is used, provided that there is not more than one valence spinor in any symmetry. The difference between the two energy expressions is

$$\begin{aligned} \Delta E^{\mathrm{val}} &= E^{\mathrm{PP}} - E^{\mathrm{val}} \\ &= \tfrac{1}{2}\sum_{uv}\Big[(J_{uv} - J_{\tilde{u}\tilde{v}}) - (K_{uv} - K_{\tilde{u}\tilde{v}})\Big]. \end{aligned} \tag{20.60}$$

Obviously, this difference is not necessarily zero. It would be fairly straightforward to add the condition that this difference must be zero to the equations to determine the pseudospinors. Reproducing the valence energy is, however, only one consideration in obtaining a pseudopotential. More important is the ability of the pseudopotential to reproduce excitation energies, dissociation energies, and potential energy surfaces.

The SCF equations derived for the pseudopotential from the valence energy are exactly equivalent to the original SCF equations, provided that the spinor space is restricted to noncore spinors. In the atom, it is possible to exclude core spinors from the calculation, but in molecules, basis functions on one atom can provide a core-like component on the atom where the pseudopotential resides. In addition, we need to know what happens if the basis set contains a reasonable representation of a core spinor. We must therefore examine what happens when a core spinor is mixed in with a valence pseudospinor.

Consider a valence space containing two electrons in a single orbital (Kramers pair) outside a closed-shell core. We form pseudospinors for the valence space and a pseudopotential for the core. Then, we permit an arbitrary spinor ϕ_r to mix with the pseudospinor. When we have developed the theory, we will examine the case in which the function is a core spinor and the case in which the function is some other spinor with a small amount of a core spinor mixed in. The arbitrary spinor is not orthogonal to the pseudospinor, and we write the mixed pseudospinor as follows:

$$\phi'_v = (\phi_v + \lambda\phi_r)/\sqrt{1 + 2\lambda S + \lambda^2} \tag{20.61}$$

where

$$S = \langle\, \phi_v \,|\, \phi_r \,\rangle. \tag{20.62}$$

Note that we are using normalized pseudospinors.

The question we are asking is this: "Is the gradient of the energy with respect to the mixing parameter λ zero when $\lambda = 0$?" That is, is the energy a stationary point at $\lambda = 0$? And further, is it a minimum or a maximum? If it is not stationary, the SCF process will mix in the function. If it is a stationary point but a maximum, it will be unstable and mix in the function. If the function is a core spinor, mixing could have serious consequences if it can lower the energy significantly.

Making use of time-reversal symmetry, the energy expression for the two-electron state with the mixed pseudospinor can be written to order λ^2 as follows:

$$\begin{aligned} E = {} & 2\,\frac{h^{\mathrm{PP}}_{vv} + \lambda h^{\mathrm{PP}}_{vr} + \lambda h^{\mathrm{PP}}_{rv} + \lambda^2 h^{\mathrm{PP}}_{rr}}{1 + 2\lambda S + \lambda^2} \\ & + \frac{(vv|vv)}{(1 + 2\lambda S + \lambda^2)^2} + \lambda\frac{(vr|vv) + (rv|vv) + (vv|vr) + (vv|rv)}{(1 + 2\lambda S + \lambda^2)^2} \\ & + \lambda^2\frac{(rr|vv) + (vv|rr) + (vr|vr) + (vr|rv) + (rv|vr) + (rv|rv)}{(1 + 2\lambda S + \lambda^2)^2} \\ = {} & 2\,\frac{h^{\mathrm{PP}}_{vv} + 2\lambda h^{\mathrm{PP}}_{vr} + \lambda^2 h^{\mathrm{PP}}_{rr}}{1 + 2\lambda S + \lambda^2} \\ & + \frac{(vv|vv) + 4\lambda(vr|vv) + 2\lambda^2(rr|vv) + 4\lambda^2(vr|rv)}{(1 + 2\lambda S + \lambda^2)^2}. \end{aligned} \tag{20.63}$$

Here, and subsequently, we drop the tilde on the pseudospinor index in the matrix elements and integrals. Since we are dealing with an atom, it is no restriction to make λ and the integrals real.

We now differentiate with respect to λ and set λ to zero, with the following result:

$$\left.\frac{\mathrm{d}E}{\mathrm{d}\lambda}\right|_{\lambda=0} = 4[h^{\mathrm{PP}}_{rv} + (rv|vv)] - 4S[h^{\mathrm{PP}}_{vv} + (vv|vv)]. \tag{20.64}$$

Because the pseudopotential and the all-electron SCF equations are equivalent, we can write

$$\hat{h}^{\mathrm{PP}} + \hat{J}_v = (1 - \hat{\mathcal{P}}^c)\hat{f}(1 - \hat{\mathcal{P}}^c) + \epsilon_v\hat{\mathcal{P}}^c, \tag{20.65}$$

and the combinations of integrals can be expressed as

$$\begin{aligned} h_{vv}^{\mathrm{PP}} + (vv|vv) &= f_{vv} = \epsilon_v \\ h_{rv}^{\mathrm{PP}} + (rv|vv) &= \epsilon_v S. \end{aligned} \tag{20.66}$$

Note that we would not normally write the Fock operator as $\hat{h}^{\mathrm{PP}} + \hat{J}_v$ but as $\hat{h}^{\mathrm{PP}} + 2\hat{J}_v - K_v$. For the gradient, it does not matter because the matrix elements are the same.

Substituting these expressions into the derivative, we find that the gradient at $\lambda = 0$ is indeed zero, and therefore this point is a stationary point. But is it a minimum or a maximum, and under what conditions?

The second derivative expression is

$$\begin{aligned} \left.\frac{\mathrm{d}^2 E}{\mathrm{d}\lambda^2}\right|_{\lambda=0} &= 4h_{rr}^{\mathrm{PP}} + 4(rr|vv) + 8(rv|vr) - 4h_{vv}^{\mathrm{PP}} - 4(vv|vv) \\ &\quad - 16S[h_{rv}^{\mathrm{PP}} + 2(rv|vv)] + 8S^2[2h_{vv}^{\mathrm{PP}} + 3(vv|vv)]. \end{aligned} \tag{20.67}$$

The largest terms are those that are of zeroth order in S, which can be written as

$$4[f_{rr}^{\mathrm{PP}} - \epsilon_v + 3(rv|vr) - (rr|vv)], \tag{20.68}$$

or as

$$4[f_{rr}^{\mathrm{PP}} - \epsilon_v + 2(rv|vr)], \tag{20.69}$$

depending on how we define the Fock operator.

In a molecular environment we now need to consider two different cases, depending on whether ϕ_r has contributions from a neighboring atom or not. If ϕ_r is a core spinor, ϕ_k say, then

$$f_{rr}^{\mathrm{PP}} \equiv f_{kk}^{\mathrm{PP}} = \epsilon_v \tag{20.70}$$

and the first two terms of zeroth order in S cancel, leaving some core–valence direct and exchange integrals. The second derivative at $\lambda = 0$ is

$$\left.\frac{\mathrm{d}^2 E}{\mathrm{d}\lambda^2}\right|_{\lambda=0} \simeq 4[3(kv|vk) - (kk|vv)] \tag{20.71}$$

or

$$\left.\frac{\mathrm{d}^2 E}{\mathrm{d}\lambda^2}\right|_{\lambda=0} \simeq 8(kv|vk) \tag{20.72}$$

depending on how we define the Fock operator. It is usually the case that the direct integral is more than three times larger than the exchange integral for core–valence interactions, so the second derivative term is negative in the first case. The exchange integral is positive, so in the second case the derivative term is positive. The details of the two-electron contributions therefore determine whether this point is a minimum or a maximum. Either way, the surface is not flat, as it is in the frozen-core approximation (section 20.1), and it is advisable to keep core basis functions out of the basis set.

We next consider the case of a neighboring atom providing a representation of the core. For this purpose, we can expand the function ϕ_r in a set of eigenfunctions of the pseudopotential Fock operator plus a core contribution, and for simplicity of notation we limit the expansion to one member, ϕ_w:

$$\phi_r = \omega\phi_w + \gamma\psi_k \tag{20.73}$$

from which we get

$$f_{rr}^{\text{PP}} = \omega^2\epsilon'_w + (1-\omega^2)\epsilon_v \tag{20.74}$$

where ϵ'_w is the eigenvalue of the pseudopotential Fock operator for pseudospinor w. The lowest-order terms in the second derivative expression are

$$\left.\frac{\mathrm{d}^2E}{\mathrm{d}\lambda^2}\right|_{\lambda=0} \simeq 4[\omega^2(\epsilon'_w - \epsilon_v) + 3(rv|vr) - (rr|vv)]. \tag{20.75}$$

Now, $\epsilon'_w > \epsilon_v$, and the one-electron contribution is positive. The same issue of the definition of the Fock operator exists for the two-electron terms, but it is most likely that the one-electron contribution is larger that the two-electron contribution. The Hessian is therefore positive, the stationary point is a minimum, and there is no danger of collapse from basis functions on a neighboring atom providing a representation of a core spinor on the atom with the pseudopotential.

Some other issues arise for the use of pseudopotential methods in calculations that go beyond the Hartree–Fock approximation. The lack of nodes changes the radial overlap of the pseudospinors, and hence changes the two-electron integrals. If the SCF spinors have nodes that reduce the radial density and these nodes are not present in the pseudospinors, it is possible to overestimate the off-diagonal elements of the Hamiltonian matrix. Combined with the closer spacing of spinor eigenvalues, the correlation energy could be overestimated. This problem has in fact been observed for the Pt atom, where the node in the 6*s* occurs near the radial maximum of the 5*d* and the correlation energies are overestimated (Rohlfing et al. 1986). One solution to this problem is to make the core smaller so that the 6*s* orbital has the radial node in the right place. Other solutions are discussed in the next section.

20.5 Generation of Pseudopotentials

The generation of a pseudopotential involves several choices: the number of spinors to be included in the valence space and the method for determining the pseudospinors and pseudopotentials. The partitioning between the core and the valence space determines to some extent the accuracy of the approximation. Too small a valence space can yield inaccurate results because of the magnitude of the core polarization. For the heavy elements, it is usually necessary to include at least the next outermost shell or shells below the valence shells: for the p block this is usually the $(n-1)d$ shell, and for the transition metals, the shells with the same n as the d shell must be included for results of reasonable accuracy. Some groups have generated both small-core and large-core pseudopotentials so that a faster result can be obtained with a larger core where higher accuracy is not needed.

The two principal approaches for generating pseudopotentials are inversion methods and fitting methods. The inversion methods rely on having a radial function without nodes. Then, the SCF equation for the pseudospinor can be inverted to obtain a pseudopotential:

$$V_v^{\mathrm{PP}} = \epsilon_v - \frac{(\hat{h}' + J_v - K_v)\phi_v}{\phi_v}. \tag{20.76}$$

This inversion is only valid if the pseudospinor is nodeless; otherwise singularities would arise in the pseudopotential. The inversion is usually done on a radial grid with tabulated potentials and pseudospinors. The resulting pseudopotential is then fitted to a form involving Gaussian functions,

$$r^2 V_v^{\mathrm{PP}} = \sum_p \sum_i c_{pi} r^p e^{\gamma_{pi} r^2}. \tag{20.77}$$

The sum over p usually includes only the values 0, 1, and 2. This form of the pseudopotential contains a pseudo-angular-momentum term that prevents the penetration of the pseudospinors to the nucleus, which also helps prevent core spinors from mixing into the pseudospinors.

In some of the early studies, the pseudospinors were obtained by minimizing a function involving the kinetic energy (Kahn et al. 1976). The shape-consistent pseudopotentials of Hay and Wadt (Hay and Wadt 1985, Wadt and Hay 1985) and Christiansen, Ermler, and coworkers (Pacios and Christiansen 1985, Hurley et al. 1986, La John et al. 1987, Ross et al. 1990, 1994, Ermler et al. 1991) are obtained by fitting a polynomial function to the core tail with the requirements that it have no nodes and the minimum number of inflection points and must match the derivatives to the order of the polynomial at the join point. While this procedure guarantees the smoothness of the function, especially after the pseudospinor is expanded in a Gaussian basis set, the choice of the join point must be made with care. If it is too far out, the results can be unsatisfactory, as was found for the $6p$ elements (Wildman et al. 1997).

The fitting methods start directly with the form of (20.77), and fit the pseudospinors to minimize some measure of the error in the pseudopotential, the pseudospinor, or the properties of the atomic system. The method of Durand and Barthelat (1975) is one such method. If the goal is to be able to reproduce the energetics of the unfrozen atomic system, a natural approach would be to fit the parameters to minimize the discrepancy between the all-electron energies and the pseudopotential energies for a wide variety of atomic states (see Dolg 2002). Pseudopotentials of this kind are known as *energy-consistent*. One advantage of fitting to atomic energies is that the effect of the compression of the eigenvalue spectrum inherent in the approximation of a single pseudopotential for a given angular symmetry is minimized.

In both cases, the angular expansion in the representation of the pseudopotential, as given by (20.1), is truncated. The expansion must include all angular momenta represented in the core, but is usually truncated at that point, or possibly one or more units of angular momentum higher. For spinors that are not occupied in the core, the content of the pseudopotential is only the exchange interaction with the core and the screening of the nuclear charge by the Coulomb potential of the core. If the local term in (20.1) is zero, functions that have higher angular momentum than represented in the angular expansion are assumed not to penetrate the core—a quite reasonable assumption—and therefore have zero matrix elements. If the local term is not zero, the assumption is that the pseudopotential is the same for all angular momenta higher than explicitly included in the angular expansion.

20.6 Relativistic Effects in Pseudopotentials

So far we have made no mention of relativistic effects, apart from a development in terms of pseudospinors rather than pseudoorbitals. Since the intended use of most pseudopotentials is in code based on a nonrelativistic formulation of quantum chemistry, the separation of the spin-free and spin–orbit terms is essential. Two approaches are again possible.

One is to use for the development of the pseudopotential a spin-free relativistic equation, such as the Cowan–Griffin equation[2] (Cowan and Griffin 1976), in which the spin–orbit terms have been already neglected. This approach results directly in a spin-free relativistic pseudopotential. There are also other methods for including spin-free relativity directly, the easiest being a direct fit to (atomic) four-component results. However, once a pseudopotential is developed using a spin-free approach, it is not usually possible to add an appropriate spin–orbit operator. An exception is the Wood–Boring equation[3] (Wood and Boring 1978), for which the spin–orbit term is an addition to the Cowan–Griffin equation and the spin–orbit term can be inverted to obtain a spin–orbit pseudopotential.

What we must not do is to add an all-electron spin–orbit operator, such as the Breit–Pauli or Douglas–Kroll operator, because the effect of the core is not included in these operators, nor is the removal of the core tail. The all-electron spin–orbit operators behave as $1/r^3$, and since the pseudospinors have minimal core amplitude, the spin–orbit effect will be grossly underestimated.

The other approach is to use a spin-dependent equation, such as the Dirac–Hartree–Fock or Wood–Boring equation, to obtain spin-dependent pseudopotentials, and take the appropriate averages and differences to obtain a spin-free relativistic pseudopotential and a spin–orbit pseudopotential. The formalism for the latter approach is as follows.

---

[2] See appendix K for details of this equation.

[3] See appendix K for details of this equation.

A pseudopotential obtained from a spin-dependent equation may be written as an expansion in ℓ and j,

$$V^{\mathrm{PP}} = \sum_{\ell j} U_{\ell,j}(r) \sum_{m_j} |\,\ell j m_j\,\rangle\langle\,\ell j m_j\,|. \tag{20.78}$$

Here we are only concerned with the angular expansion, so we have dropped the local potential $U(r)$. The angular spinors can be expanded into orbital and spin parts:

$$|\,\ell j m_j\,\rangle = \sum_{m_\ell, m_s} |\,\ell m_\ell\,\rangle|\,s m_s\,\rangle\langle\,\ell m_\ell s m_s\,|\,j m_j\,\rangle \tag{20.79}$$

where $\langle\,\ell m_\ell s m_s\,|\,j m_j\,\rangle$ is a Clebsch–Gordan coefficient. The contribution to the pseudopotential for a given ℓ value is therefore

$$V_\ell^{\mathrm{PP}} = \sum_{m_\ell, m'_\ell} |\,\ell m_\ell\,\rangle\langle\,\ell m'_\ell\,| \sum_{m_s, m'_s} |\,s m_s\,\rangle\langle\,s m'_s\,| \sum_{j, m_j} U_{\ell,j} \langle\,\ell m_\ell s m_s\,|\,j m_j\,\rangle\langle\,j m_j\,|\,\ell m'_\ell s m'_s\,\rangle. \tag{20.80}$$

We now define spin-free and spin–orbit potentials,

$$U_\ell^{\mathrm{sf}} = \frac{\ell U_{\ell,\ell-1/2} + (\ell+1)U_{\ell,\ell+1/2}}{(2\ell+1)}; \qquad U_\ell^{\mathrm{so}} = \frac{2(U_{\ell,\ell+1/2} - U_{\ell,\ell-1/2})}{(2\ell+1)}. \tag{20.81}$$

Here we note that U_ℓ^{sf} appears as a weighted average of what may be regarded as the potentials for two spin–orbit split levels, and U_ℓ^{so} is related to the difference between these two potentials. The factor of 2 in the numerator of the spin–orbit potential is included for representation of the final expression in terms of the spin operator. Rearranging these expressions, we can write

$$U_{\ell,\ell-1/2} = U_\ell^{\mathrm{sf}} - \tfrac{1}{2}(\ell+1)U_\ell^{\mathrm{so}}; \qquad U_{\ell,\ell+1/2} = U_\ell^{\mathrm{sf}} + \tfrac{1}{2}\ell U_\ell^{\mathrm{so}}. \tag{20.82}$$

We substitute in (20.80) and collect the terms in the spin-free and spin–orbit potentials to get

$$\begin{aligned} V_\ell^{\mathrm{PP}} = \sum_{m_\ell, m'_\ell} |\,\ell m_\ell\,\rangle\langle\,\ell m'_\ell\,| \sum_{m_s, m'_s} |\,s m_s\,\rangle\langle\,s m'_s\,| \Big[& \sum_{j, m_j} U_\ell^{\mathrm{sf}} \langle\,\ell m_\ell s m_s\,|\,j m_j\,\rangle\langle\,j m_j\,|\,\ell m'_\ell s m'_s\,\rangle \\ & + \sum_{m_j} U_\ell^{\mathrm{so}} \frac{\ell}{2} \langle\,\ell m_\ell s m_s\,|\,(\ell+\tfrac{1}{2}) m_j\,\rangle\langle\,(\ell+\tfrac{1}{2}) m_j\,|\,\ell m'_\ell s m'_s\,\rangle \\ & - \sum_{m_j} U_\ell^{\mathrm{so}} \frac{\ell+1}{2} \langle\,\ell m_\ell s m_s\,|\,(\ell-\tfrac{1}{2}) m_j\,\rangle\langle\,(\ell-\tfrac{1}{2}) m_j\,|\,\ell m'_\ell s m'_s\,\rangle \Big], \end{aligned} \tag{20.83}$$

exploiting the fact that $\ell+\frac{1}{2}$ and $\ell-\frac{1}{2}$ are the only possible values of j. The Clebsch–Gordan coefficients for the spin-free term reduce to a product of delta functions: this follows from the orthogonality properties of Clebsch–Gordan coefficients, or it may be demonstrated by reversing the expansion used above

$$\begin{aligned}\sum_{j,m_j}\langle \ell m_\ell s m_s \,|\, j m_j\rangle\langle j m_j \,|\, \ell m'_\ell s m'_s\rangle &= \langle \ell m_\ell \,|\langle s m_s \,|\Big[\sum_{j,m_j} |\, j m_j\rangle\langle j m_j \,|\Big]|\,\ell m_\ell\rangle|\, s m_s\rangle \\ &= \delta_{m_\ell,m'_\ell}\delta_{m_s,m'_s}. \end{aligned} \tag{20.84}$$

The sum in brackets is just an identity operator. The spin operator $\sum_{m_s} |\, s m_s\rangle\langle s m_s \,|$ is also an identity operator, so it can be omitted, and the spin-free potential is

$$V^{\mathrm{PP,sf}} = \sum_{\ell m_\ell} U_\ell^{\mathrm{sf}} \,|\, \ell m_\ell\rangle\langle \ell m_\ell \,|. \tag{20.85}$$

The spin-free potential derived in this way is also known as an *averaged relativistic effective potential*, or AREP.

To obtain a similar simplification for the spin–orbit potential, we note that

$$\boldsymbol{\ell}\cdot\mathbf{s}|\,(\ell+\tfrac{1}{2})m_j\rangle = \frac{\ell}{2}|\,(\ell+\tfrac{1}{2})m_j\rangle; \qquad \boldsymbol{\ell}\cdot\mathbf{s}|\,(\ell-\tfrac{1}{2})m_j\rangle = -\frac{\ell+1}{2}|\,(\ell-\tfrac{1}{2})m_j\rangle. \tag{20.86}$$

This enables us to write the various terms involving the spin–orbit potential as

$$\begin{aligned} &\sum_{m_j} U_\ell^{\mathrm{so}}\langle \ell m_\ell s m_s \,|\, \boldsymbol{\ell}\cdot\mathbf{s} \,|\, (\ell+\tfrac{1}{2})m_j\rangle\langle (\ell+\tfrac{1}{2})m_j \,|\, \ell m'_\ell s m'_s\rangle \\ &\quad + \sum_{m_j} U_\ell^{\mathrm{so}}\langle \ell m_\ell s m_s \,|\, \boldsymbol{\ell}\cdot\mathbf{s} \,|\, (\ell-\tfrac{1}{2})m_j\rangle\langle (\ell-\tfrac{1}{2})m_j \,|\, \ell m'_\ell s m'_s\rangle \\ &\quad = U_\ell^{\mathrm{so}}\sum_{j m_j}\langle \ell m_\ell s m_s \,|\, \boldsymbol{\ell}\cdot\mathbf{s} \,|\, j m_j\rangle\langle j m_j \,|\, \ell m'_\ell s m'_s\rangle \\ &\quad = U_\ell^{\mathrm{so}}\langle \ell m_\ell s m_s \,|\boldsymbol{\ell}\cdot\mathbf{s}\Big[\sum_{j m_j} |\, j m_j\rangle\langle j m_j \,|\Big]|\, \ell m'_\ell s m'_s\rangle \\ &\quad = U_\ell^{\mathrm{so}}\langle \ell m_\ell s m_s \,|\, \boldsymbol{\ell}\cdot\mathbf{s} \,|\, \ell m'_\ell s m'_s\rangle. \end{aligned} \tag{20.87}$$

Using this expression, we can write the spin–orbit potential as

$$\begin{aligned} V^{\mathrm{PP,so}} &= \sum_\ell U_\ell^{\mathrm{so}} \sum_{m_\ell m'_\ell}\sum_{m_s m'_s} |\, \ell m_\ell s m_s\rangle\langle \ell m_\ell s m_s \,|\boldsymbol{\ell}\cdot\mathbf{s}|\, \ell m'_\ell s m'_s\rangle\langle \ell m'_\ell s m'_s \,| \\ &= \sum_\ell U_\ell^{\mathrm{so}} \sum_{m_\ell m'_\ell}\sum_{m_s m'_s} |\, s m_s\rangle\langle s m_s \,||\, \ell m_\ell\rangle\langle \ell m_\ell \,|\boldsymbol{\ell}\cdot\mathbf{s}|\, \ell m'_\ell\rangle\langle \ell m'_\ell \,||\, s m'_s\rangle\langle s m'_s \,| \\ &= \sum_\ell U_\ell^{\mathrm{so}} \sum_{m_\ell m'_\ell} |\, \ell m_\ell\rangle\langle \ell m_\ell \,|\boldsymbol{\ell}\cdot\mathbf{s}|\, \ell m'_\ell\rangle\langle \ell m'_\ell \,| \end{aligned} \tag{20.88}$$

where we have exploited the fact that the spin operator $\sum_{m_s} |sm_s\rangle\langle sm_s|$ is an identity operator and can be omitted. We can extract $\mathbf{s}$ to separate the spin and orbital parts, and express the spin–orbit potential as

$$V^{\mathrm{PP,so}} = \mathbf{s} \cdot \sum_{\ell} U_{\ell}^{\mathrm{so}} \sum_{m_\ell m'_\ell} |\ell m_\ell\rangle\langle \ell m_\ell|\boldsymbol{\ell}|\ell m'_\ell\rangle\langle \ell m'_\ell|. \tag{20.89}$$

This potential is also known as a *spin–orbit relativistic effective potential*, or SOREP.

Finally, we can write the entire spin-dependent pseudopotential as

$$\begin{aligned} V^{\mathrm{PP}} &= V^{\mathrm{PP,sf}} + V^{\mathrm{PP,so}} \\ &= U(r) + \sum_{\ell} U_{\ell}^{\mathrm{sf}} \sum_{m_\ell} |\ell m_\ell\rangle\langle \ell m_\ell| + \mathbf{s} \cdot \sum_{\ell} U_{\ell}^{\mathrm{so}} \sum_{m_\ell m'_\ell} |\ell m_\ell\rangle\langle \ell m_\ell|\boldsymbol{\ell}|\ell m'_\ell\rangle\langle \ell m'_\ell|, \end{aligned} \tag{20.90}$$

where we have reintroduced the local potential, which is spin free.

By far the majority of relativistic pseudopotential calculations include only the spin-free pseudopotential at the SCF stage, and add the spin–orbit potential afterwards if they add it at all. There are a few where both the spin-free and spin–orbit effects are included at the SCF stage and the calculations performed with double-group symmetry, just like four-component calculations.

20.7 Model Potentials

In parallel with the development of pseudopotentials, a second approach to the frozen-core model has been taken. In this approach, the core orthogonality is retained, and the model includes an explicit representation of the core spinors. The rationale for retaining the core tail in the valence functions is that the cost of the primitive integrals is less important than the cost of retaining all the core basis functions. Some approximations to the core tails are generally made because the integral cost is still significant.

If the core orthogonality is retained, there is no necessity to insert the projection operators around the core Fock operator in (20.5), but we must still ensure that core spinors are kept out of the valence space. In an atom it is easy to maintain the orthogonality, but in a molecule the basis functions on another center expand into a linear combination of functions on the frozen-core center, including core spinors.

The equation that is used to ensure that the core spinors are projected out is the Huzinaga–Cantu equation (Huzinaga and Cantu 1971, Huzinaga et al. 1973), which we will now derive. Instead of writing the Fock equations for the valence spinors in terms of Lagrange multipliers and imposing the criterion that the overlaps must be zero at convergence, we can introduce projection operators into the Fock operator:

$$\left[\hat{f} - \hat{\mathcal{P}}^c \hat{f} - \hat{f}\hat{\mathcal{P}}^c\right] \psi_v = \epsilon_v \psi_v. \tag{20.91}$$

Premultiplying by ψ_k and integrating yields

$$-f_{kk}S_{kv} = \epsilon_v S_{kv} \tag{20.92}$$

which holds only if $S_{kv} = 0$. Premultiplying by ψ_v and integrating yields

$$f_{vv} - 2\sum_k f_{kv}S_{kv} = \epsilon_v \tag{20.93}$$

which gives the desired result, $f_{vv} = \epsilon_v$ if $S_{kv} = 0$. It is also easy to show that for any other valence spinor, $f_{uv} = 0$, provided that $S_{uv} = 0$.

If the core spinors are chosen to be eigenfunctions of the Fock operator, we can replace the Fock operators multiplying the projectors by the core eigenvalues,

$$\left[\hat{f} - 2\sum_k \epsilon_k |\psi_k\rangle\langle\psi_k|\right]\psi_v = \hat{f}^{\mathrm{MP}}\psi_v = \epsilon_v\psi_v. \tag{20.94}$$

This is the Huzinaga–Cantu equation. It has similar properties to (20.91), but instead of f_{kv} being indeterminate it is zero. The projector term in the Fock operator functions as a level shift: core spinors are moved up by $2|\epsilon_k|$. The advantage of this is that there is an energetic cost for any core spinors that come from other centers to mix into the valence spinors, and therefore the mixing is not likely. We will return to this point later.

The model potential is defined by partitioning the Fock operator into core and valence parts, just as in (20.43), and defining the modified one-electron operator as in (20.44). The model potential is

$$V^{\mathrm{MP}} = -Z^c f^{\mathrm{nuc}}(r) + \sum_k(\hat{J}_k - \hat{K}_k) + \sum_k B_k|\psi_k\rangle\langle\psi_k|. \tag{20.95}$$

The level shift has been generalized, but the usual value taken is $B_k = -2\epsilon_k$.

If no further approximations are made to the valence spinors, this potential should yield results exactly equivalent to the frozen-core approximation. Often, some approximation is made to the core tails because the number of basis functions used to describe the core tail can be large: typically, there are as many functions used to describe the first antinode of a noncore orbital as there are functions to describe all the remaining antinodes. Approximating the core has the potential to save up to half the number of basis functions, resulting in a saving of a factor of up to 16 in the primitive integrals. It is usually the inner core that is approximated, because a good description of the outer core, including the outermost nodes, is essential for accurate valence energetics.

One useful approximation to the core part of a valence spinor is to describe each antinode, with the possible exception of the subvalence antinode, by a single Gaussian function. The coefficients of these functions are determined by requiring the approximated valence spinor to be orthogonal to the core spinors. This approximation has the advantage of retaining strict core orthogonality. In terms of the analysis of pseudospinors made earlier in this chapter, the approximation corresponds to adding a linear combination of virtual spinors to the valence spinor. It is possible to make further approximations, which sacrifice the strict core orthogonality but retain the correct shape and nodal structure in the outer-core region.

20.8 Energetics of Model Potentials

Using the formalism developed in section 20.4, we can analyze the behavior of the approximations to the valence spinors in the AIMP method. The purpose of this exercise is to determine whether the approximations to the one-electron Hamiltonian and the pseudospinors are likely to cause any undesirable behavior, such as collapse when a core spinor is mixed in. As before, the model we use is the simplest: an atom with two valence electrons. We have three separate cases to consider.

In the first case, no approximation is made to the valence spinor. Then there is no approximation to the Fock matrix apart from the addition of the projector term. The gradient for mixing an arbitrary function into the valence spinor is deduced from (20.64)

$$\left.\frac{\mathrm{d}E}{\mathrm{d}\lambda}\right|_{\lambda=0} = 4\langle\,\psi_v\,|\,\hat{f}+\hat{B}\,|\,\psi_r\,\rangle - 4S\langle\,\psi_v\,|\,\hat{f}+\hat{B}\,|\,\psi_v\,\rangle = 4[f_{vr}+B_{vr}-S(f_{vv}+B_{vv})] \tag{20.96}$$

where

$$\hat{B} = \sum_k B_k|\,\psi_k\,\rangle\langle\,\psi_k\,|. \tag{20.97}$$

The matrix elements of $\hat{B}$ are zero, and $f_{vr} = S\epsilon_v$, so the gradient is zero and $\lambda = 0$ is a stationary point. The terms that are of zeroth order in S in the second derivative expression, equation (20.67), can be written

$$4[f_{rr}+B_{rr}-f_{vv}+3(rv|vr)-(rr|vv)]. \tag{20.98}$$

If ψ_r is a core spinor, ψ_k say, then

$$f_{rr} \equiv f_{kk} = \epsilon_k. \tag{20.99}$$

With $B_{kk} = -2\epsilon_k$, these terms come to

$$4[-\epsilon_k-\epsilon_v+3(kv|vk)-(kk|vv)] \tag{20.100}$$

which, since the spinor eigenvalues are negative, has a positive value. The remaining terms of first and second order in S cannot make the second derivative negative. Thus, the stationary point is a minimum and the valence spinor has no tendency to mix core spinors at all. If the arbitrary spinor is some other spinor with a small admixture of a core spinor, the projector term makes the second derivative a little more positive than it otherwise would have been.

In the second case, the core orthogonality is kept but approximated. The valence pseudospinor is no longer an eigenfunction of the Fock operator or the model potential. We can write the gradient as

$$\begin{aligned}\left.\frac{\mathrm{d}E}{\mathrm{d}\lambda}\right|_{\lambda=0} &= 4[h_{\tilde{v}r}^{\mathrm{MP}} + B_{\tilde{v}r} + (\tilde{v}\tilde{v}|\tilde{v}r)] - 4S[h_{\tilde{v}\tilde{v}}^{\mathrm{MP}} + B_{\tilde{v}\tilde{v}} + (\tilde{v}\tilde{v}|\tilde{v}\tilde{v})] \\ &= 4[f_{\tilde{v}r}^{\mathrm{MP}} + B_{\tilde{v}r}] - 4S[f_{\tilde{v}\tilde{v}}^{\mathrm{MP}} + B_{\tilde{v}\tilde{v}}]\end{aligned} \tag{20.101}$$

where the tilde indicates that the spinor is no longer the all-electron spinor but a pseudospinor. Because we did not change the model potential when we introduced the pseudospinor, the Fock operator for this pseudospinor is no longer the same as the all-electron Fock operator due to the two-electron terms. We can write the Fock operator as the all-electron Fock operator with a correction term:

$$\hat{f}^{\mathrm{MP}} = \hat{f} + \hat{f}^{R}. \tag{20.102}$$

The correction is essentially the difference in Coulomb potential between the valence spinor and the valence pseudospinor. Now the gradient can be written

$$\left.\frac{\mathrm{d}E}{\mathrm{d}\lambda}\right|_{\lambda=0} = 4[f_{\tilde{v}r} + f_{\tilde{v}r}^{R} + B_{\tilde{v}r} - S(f_{\tilde{v}\tilde{v}} + f_{\tilde{v}\tilde{v}}^{R} + B_{\tilde{v}\tilde{v}})]. \tag{20.103}$$

The matrix elements of $\hat{B}$ that involve the pseudospinor ϕ_v are zero because it is still orthogonal to the core, but the matrix elements of the Fock operator are not necessarily zero. The second derivative terms that are zeroth order in S are

$$\begin{aligned}&4[f_{rr}^{\mathrm{MP}} + B_{rr} - f_{\tilde{v}\tilde{v}}^{\mathrm{MP}} + 3(r\tilde{v}|\tilde{v}r) - (rr|\tilde{v}\tilde{v})] \\ &\quad = 4[f_{rr} + B_{rr} - f_{\tilde{v}\tilde{v}} + f_{rr}^{R} + f_{\tilde{v}\tilde{v}}^{R} + 3(r\tilde{v}|\tilde{v}r) - (rr|\tilde{v}\tilde{v})].\end{aligned} \tag{20.104}$$

For ψ_r a core spinor, ψ_k say, the off-diagonal Fock matrix element $f_{\tilde{v}k}$ is zero because the pseudospinor contains no components of the core spinors. S is zero for this case because the pseudospinor is orthogonal to the core. The gradient is therefore $4f_{\tilde{v}k}^{R}$, which is not necessarily zero, though it is likely to be small. The dominant terms in the second derivative are the same as in the previous case, $f_{kk} + B_{kk} = -2\epsilon_k$. The curvature is therefore positive, and strongly so, and the minimum, though not at $\lambda = 0$, is likely to be not far away. The interesting result is that even though the pseudospinor is orthogonal to the core, it mixes a small amount of core spinors to compensate for the change in potential due to the approximation to the core part of the pseudospinor.

For an arbitrary spinor r, we can expand the spinor into a contribution from the pseudospinor $\tilde{v}$ and an orthogonal term,

$$\phi_r = S\phi_{\tilde{v}} + \gamma\phi_{v_\perp} \tag{20.105}$$

where γ is a mixing factor, and should be close to 1. The gradient then becomes

$$\left.\frac{\mathrm{d}E}{\mathrm{d}\lambda}\right|_{\lambda=0} = 4[f^{\mathrm{MP}}_{\tilde{v}r} - Sf^{\mathrm{MP}}_{\tilde{v}\tilde{v}}] = 4\gamma\langle\phi_{v_\perp}|f^{\mathrm{MP}}|\phi_{\tilde{v}}\rangle \tag{20.106}$$

which is not necessarily zero, but is likely to be small. The contribution from any core component in ϕ_r reduces to $4f^R_{\tilde{v}k}$. The dominant second-derivative term is

$$f^{\mathrm{MP}}_{rr} + B_{rr} - f^{\mathrm{MP}}_{\tilde{v}\tilde{v}} = f_{rr} + B_{rr} - f_{\tilde{v}\tilde{v}} + f^R_{rr} - f^R_{\tilde{v}\tilde{v}}. \tag{20.107}$$

If we expand ϕ_r in the complete orthonormal set of spinors, the contributions to f_{rr} that come from core spinors are made positive by the terms from B_{rr}, and the remaining contributions sum to a value that is less negative than ϵ_v. It must also be true that $f_{\tilde{v}\tilde{v}} > \epsilon_v$. The terms from the residual Fock operator are small, so the second derivative is positive and the minimum, with a small gradient, must be nearby. In this case also, the deviation of the valence pseudospinor from the valence spinor causes mixing of other spinors to compensate for the change.

In the third and most general case, where the valence pseudospinor is not orthogonal to the core or to the trial function, the gradient is again given by (20.103). For ψ_r a core spinor ψ_k, we can use the expansion of the valence pseudospinor in terms of the complete set of spinors

$$\phi_{\tilde{v}} = \sum_p c_{vp}\psi_p \tag{20.108}$$

to show that $f_{\tilde{v}k} = S\epsilon_k$, and $B_{\tilde{v}k} = -2S\epsilon_k$. The gradient can then be written

$$\left.\frac{\mathrm{d}E}{\mathrm{d}\lambda}\right|_{\lambda=0} = 4[f^R_{\tilde{v}k} - S(\epsilon_k + f_{\tilde{v}\tilde{v}} + f^R_{\tilde{v}\tilde{v}} + B_{\tilde{v}\tilde{v}})]. \tag{20.109}$$

The second derivative is dominated by the core terms as before, and is positive. Again, there is a minimum, but how close it is to $\lambda = 0$ depends on the size of the core overlap S because ϵ_k is not small. The conclusion is that it is important to keep any representation of the core out of the basis set if core orthogonality is sacrificed and to keep the deviation from core orthogonality to a minimum.

For the case of an arbitrary trial function ψ_r we can use (20.105) to reduce the gradient to

$$\left.\frac{\mathrm{d}E}{\mathrm{d}\lambda}\right|_{\lambda=0} = 4\gamma[f^{\mathrm{MP}}_{\tilde{v}v_\perp} + B_{\tilde{v}v_\perp}]. \tag{20.110}$$

This expression contains a core projector matrix element in addition to the model potential Fock matrix element. However, because it is off-diagonal the core projector term could be of either sign, depending on the contribution of the orthogonal component to the trial spinor. Whatever contributions to the Fock matrix elements come from core spinors, their signs are changed by the projector matrix element. In any case, the gradient is not likely to be zero. The second derivative is dominated as before by the terms that do not contain S, and the curvature is positive, indicating that there is a minimum. How far it is from the point $\lambda = 0$ depends on the magnitude of the gradient.

The behavior of the SCF solutions is not the only issue of interest in the energetics of model potentials. Because B_k is finite, the core spinors are not moved an infinite distance from the valence space, but a finite distance. Any representation of the core spinors remains in the molecular basis set, and can contaminate a correlated calculation, if they are not removed (Klobukowski 1990). The result can be an overestimation of correlation effects. Diagonalizing $\hat{B}$ in the virtual spinor space should provide a means of recognizing and removing core-like spinors.

20.9 Model Potential Implementation

The form of the model potential is much simpler than that of the pseudopotential, and it is relatively easy to implement. The integrals over core projector terms are overlaps, which need no special coding. The core direct potential can be fitted to a linear combination of Gaussian functions. It is the core exchange potential that is the hardest to represent because of its nonlocality.

The simplest approach is to ignore the nonlocality and to represent both direct and exchange terms in a Gaussian series (Bonifacic and Huzinaga 1974, 1975, Sakai et al. 1997). This approach has the advantage that the integrals are overlaps and nuclear attraction type integrals. The disadvantage may be the loss of nonlocality. The most common form of the model potential in this implementation is

$$V^{\mathrm{MP}}(r) = -\frac{Z - n_c}{r}\left[\sum_i A_i e^{a_i r^2} + r\sum_i B_i e^{b_i r^2}\right]. \tag{20.111}$$

A local form was also used for pseudopotentials, and was abandoned for the same reasons: the lack of distinction between different symmetries.

A more sophisticated approach includes some form of angular projection. Because the integrand for each density in the exchange integrals involves a core spinor, the range of the core exchange term is short and projection onto a local (atomic) basis makes sense. However, the projection must contain an infinite sum of both angular and radial terms to be exact. Truncation of the projector to the atomic basis is one practical approach (Andzelm et al. 1985, Huzinaga et al. 1987). Then, it is only necessary to define the basis set and the core and fit the Coulomb potential to define the ab initio model potential. The potential can be expressed in the form

$$V^{\mathrm{MP}}(r) = \frac{1}{r}\sum_i A_i e^{a_i r^2} + \sum_{\ell m}\sum_{pq} |\, p\ell m\rangle X_{\ell;pq}\langle q\ell m\,| \tag{20.112}$$

where for a given ℓ the matrix $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{S}^{-1}\mathbf{K}\mathbf{S}^{-1} \tag{20.113}$$

and **K** is the matrix of the core exchange operator, given by

$$K_{pq} = \langle p | -\sum_k \hat{K}_k | q \rangle. \tag{20.114}$$

If the basis set used for this representation is the atomic basis set that is used in the molecule, the representation of the one-center exchange elements is exact. These are the most important elements because of the short range of the exchange potential.

20.10 Relativistic Effects in Model Potentials

Spin-free relativistic effects are readily incorporated into the ab initio model potential approximation by using a one-component spin-free relativistic method for the atom, such as the Cowan–Griffin method[4] or the Douglas–Kroll–Hess method.

In the case of the Cowan–Griffin method, the spin-free one-electron relativistic operators are incorporated into the spectral representation of the exchange potential:

$$\mathbf{X} = \mathbf{S}^{-1}\mathbf{M}\mathbf{S}^{-1}. \tag{20.115}$$

M is the matrix of the combined core exchange and spin-free relativistic operator, given by

$$M_{pq} = \langle p | -\sum_k \hat{K}_k | q \rangle + \langle p | V^{\mathrm{MVD}} | q \rangle. \tag{20.116}$$

V^{MVD} is an angular projection onto the Cowan–Griffin mass–velocity and Darwin potentials for the valence orbitals,

$$V^{\mathrm{CG-MVD}} = \sum_{\ell m} | \ell m \rangle \langle \ell m | V_{nl}^{\mathrm{CG-MVD}} | \ell m \rangle \langle \ell m | \tag{20.117}$$

where the sum extends over the valence symmetries, and

$$V_{nl}^{\mathrm{CG-MVD}} = -\frac{1}{4m^2c^2}(\epsilon_{n\ell} - V) - \frac{1}{2mc^2 + (\epsilon_{n\ell} - V)} \frac{\mathrm{d}V}{\mathrm{d}r} \left(\frac{1}{P_{n\ell}} \frac{\mathrm{d}P_{nl}}{\mathrm{d}r} - \frac{1}{r} \right). \tag{20.118}$$

The potential in this definition is the SCF potential, not merely the nuclear attraction potential. The model potential therefore includes both one- and two-electron relativistic corrections. The principal quantum number is that of the outermost valence orbital. Note that this makes the model potential dependent on the valence orbitals. For model

---

[4] See appendix K for details of this method.

potentials where there is more than one valence orbital of a given symmetry, the difference is apparently not large (Seijo and Barandiarán 1999).

In the case of the Douglas–Kroll–Hess method, no special treatment of relativistic effects is required apart from the use of the one-electron Douglas–Kroll–Hess operator for the valence electrons. The direct and exchange potentials for the core electrons are treated in exactly the same way as in the nonrelativistic case but using the atomic Douglas–Kroll–Hess orbitals. However, only the unmodified part of the nuclear attraction is partitioned into a core and a valence part; the core part is included with the core direct potential just as in the nonrelativistic case.

Spin–orbit effects can be incorporated using a two-component relativistic method for the atom such as the Wood–Boring method.[5] In this case, the usual representation is a semilocal form as is used in the pseudopotential method,

$$V^{\mathrm{so,MP}} = \sum_{\ell} V_{\ell}^{\mathrm{so}} \sum_{m,m'} |\ell m\rangle\langle \ell m|\boldsymbol{\ell}\cdot\mathbf{s}|\ell m'\rangle\langle \ell m'| \tag{20.119}$$

and the radial potential is expanded in a set of Gaussian functions,

$$V_{\ell}^{\mathrm{so}} = \sum_{i} D_{i,\ell}\, e^{d_i r^2}/r^2. \tag{20.120}$$

The coefficients and exponents are fitted to the radial components of the Wood–Boring spin–orbit operator, or some other suitable spin–orbit operator.

Just as for the Cowan–Griffin operator, the potential V_{ℓ}^{so} is the atomic SCF potential and so includes both one- and two-electron spin–orbit effects. In this respect the integrals over this potential resemble the atomic mean-field spin–orbit integrals of Hess et al. (1996).

20.11 Properties and Core Approximations

We turn finally to the calculation of molecular electric and magnetic properties in the pseudopotential and model potential approximations. For first-order properties and variational wave functions, we can write the property using the Hellmann–Feynman theorem as

$$\frac{\mathrm{d}E}{\mathrm{d}\lambda} = \langle \Psi|\frac{\partial\mathcal{H}}{\partial\lambda}|\Psi\rangle \tag{20.121}$$

where λ represents the perturbation strength. In an all-electron model this equation reduces to the sum of the expectations of the one-electron operator over the occupied spinors for the property. The same applies to the frozen-core approximation, and we may moreover ignore the contributions from the frozen-core electrons.

---

[5] See appendix K for details of this method.

For electric fields, this also means using the reduced nuclear charge Z^v for the nuclear contribution, due to the cancellation of the contribution from the core electrons with the rest of the nuclear contribution. For model potentials that are derived without further approximation, the calculation of properties follows straightforwardly from the frozen-core approximation.

However, if the valence spinors are modified to create pseudospinors—and this applies to both model potentials and pseudopotentials—the expectation of the valence wave function over the property operator is no longer the same as in the all-electron or the frozen-core case. For external fields, where the bulk of the contribution comes from the valence region of space, the deviation of the pseudospinor property integrals from their unmodified counterparts will be small. However, for nuclear fields, the approximation to the core part of the spinors will have serious consequences for the property. Pseudospinors used with pseudopotentials are designed to have vanishing amplitude at the nucleus, and therefore the property integrals will be much smaller than they should be. This means that properties such as NMR shielding constants calculated for the pseudopotential center will inevitably be erroneous.

The alternative is to derive pseudoproperty operators, which account for the fact that the core has been approximated, or to calculate the property integrals with the original spinors. The former has not to our knowledge been attempted; the latter has been done by Kozlov et al. (1987), for example, in the calculation of parity nonconservation effects.

21

Spin–Orbit Configuration Interaction Methods

In this chapter, we address the following problem: Assume that we have derived a method that gives a good approximation to the treatment of relativistic effects in molecular systems, but that leaves out any explicit references to the spin—that is, a spin-free relativistic approximation. How can we go about getting a reliable estimate of the spin–orbit interaction in the system?

There may be several reasons why we would want to take such an approach. One of them is that nonrelativistic calculations are much simpler to handle with respect to symmetry and logistics, and so if we could do the heavy computational work within the least demanding framework, it might entail considerable savings. Another reason might be that we are interested in light systems where the spin-free relativistic effects are small, but where the symmetry breaking induced by spin–orbit interaction may be of crucial importance for near-degenerate states and surface crossings.

To solve this problem, we have to answer two questions, "What?" and "How?" The first one is concerned with finding an operator $\hat{\mathcal{H}}^{\text{so}}$ that describes the spin–orbit interaction that has been left out of our zeroth-order Hamiltonian, $\hat{\mathcal{H}}_0$, making the total Hamiltonian

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{H}}^{\text{so}}. \tag{21.1}$$

The form of this operator may be a matter of choice, but we would preferably like to restore the terms dropped in developing $\hat{\mathcal{H}}_0$ from the fully relativistic Hamiltonian. This in turn leads to different $\hat{\mathcal{H}}^{\text{so}}$ operators depending on the approach taken towards $\hat{\mathcal{H}}_0$.

The other question is concerned with how to treat these $\hat{\mathcal{H}}^{\text{so}}$ operators in a computationally efficient manner in a configuration interaction calculation. Although the operators may differ, the problems inherent in their application are common for many of the choices of zeroth-order method. A further challenge is that spin–orbit energies are of comparable magnitude to the correlation energies for heavy elements. Thus, the two should be treated on a reasonably equal footing. Because most of these methods ultimately rely on diagonalizing $\hat{\mathcal{H}}^{\text{so}}$ in an N-particle basis, they fall within the family of CI methods or CI-related perturbation methods.

Below, we start by discussing some of the $\hat{\mathcal{H}}^{\text{so}}$ operators that may be used in the various settings. Some of these have already been derived in the preceding chapters on approximate methods. We also comment on the expectation values and integrals that they give rise to. We then give examples of efficient computational algorithms that have been developed to allow for manageable calculations using these operators. Our account is by no means exhaustive, but our aim is to illustrate some of the central features of spin–orbit methods and some of the more frequently employed strategies for exploiting these methods in practical computations.

21.1 Breit–Pauli Spin–Orbit Operators

While there are a number of possible pitfalls in the application of the Breit–Pauli Hamiltonian defined in (17.40), it does offer a convenient partitioning of the operator, where we may pick those contributions that are relevant for the applications at hand. For the study of spin–orbit contributions, a suitable starting point is obtained by retaining only those two-electron terms that involve direct coupling between spin and orbital motion. This entails the neglect of orbit–orbit, spin–spin, and Fermi contact terms, which also contribute to the electron–electron interaction energy at order c^{-2}, and is as such slightly inconsistent. However, the procedure may be defended on the grounds that the spin–orbit terms are the chief terms responsible for the breaking of the pure nonrelativistic symmetry of the system. For an atom[1] of nuclear charge Z, our Hamiltonian then becomes

$$\hat{\mathcal{H}}^{\text{BP}} = \sum_i \left[\frac{\mathbf{p}_i^2}{2m} - \frac{Z}{r_i} - \frac{\mathbf{p}_i^4}{8m^3c^2} + \frac{\pi Z}{2m^2c^2}\delta(\mathbf{r}_i) + \frac{Z}{2m^2c^2r_i^3}\mathbf{s}_i \cdot \boldsymbol{\ell}_i \right] + \frac{1}{2}\sum_{i \neq j}\left[\frac{1}{r_{ij}} - \frac{1}{m^2c^2r_{ij}^3}\left[(\mathbf{s}_i + 2\mathbf{s}_j) \cdot \mathbf{r}_{ij} \times \mathbf{p}_i\right]\right]. \tag{21.2}$$

We assume that we somehow know how to handle the spin-free terms of this equation, and that they form a suitable H_0. The spin–orbit part may be divided into a one-electron and a two-electron operator,

$$\hat{\mathcal{H}}^{\text{BPso}} = \hat{\mathcal{H}}^{\text{BPso1}} + \hat{\mathcal{H}}^{\text{BPso2}} = \sum_i \hat{h}_i^{\text{BPso}} + \frac{1}{2}\sum_{i \neq j} \hat{g}_{ij}^{\text{BPso}} \tag{21.3}$$

---

[1] Obviously, for molecules there must be a sum over nuclei in the Hamiltonian expression.

with the one-electron and two-electron operators defined by

$$\begin{aligned}\hat{h}_i^{\mathrm{BPso}} &= \frac{Z}{2m^2c^2r_i^3}\mathbf{s}_i \cdot \boldsymbol{\ell}_i, \\ \hat{g}_{ij}^{\mathrm{BPso}} &= -\left[\frac{1}{m^2c^2r_{ij}^3}(\mathbf{s}_i + 2\mathbf{s}_j) \cdot \mathbf{r}_{ij} \times \mathbf{p}_i\right].\end{aligned} \tag{21.4}$$

We may use these operators in various ways, to which we will return later. Regardless of which computational model we choose to follow, we end up evaluating basis function integrals over these operators. It is easy to see that this will require a somewhat different treatment from the ordinary nonrelativistic integral handling. Ignoring the constants, we can write the matrix element of the one-electron spin–orbit operator as

$$\begin{aligned}\langle p|\frac{Z}{r^3}\mathbf{s} \cdot \boldsymbol{\ell}|q\rangle &= \int \phi_p^*(\mathbf{r})\eta_p^\dagger(\tau)\frac{Z}{r^3}\mathbf{s} \cdot (\mathbf{r} \times \mathbf{p})\,\phi_q(\mathbf{r})\,\eta_q(\tau)\,\mathrm{d}\mathbf{r}\mathrm{d}\tau \\ &= \int \phi_p^*(\mathbf{r})\frac{Z}{r^3}(\mathbf{r} \times \mathbf{p})\phi_q(\mathbf{r})\,\mathrm{d}\mathbf{r} \cdot \int \eta_p^\dagger(\tau)\,\mathbf{s}\,\eta_q(\tau)\,\mathrm{d}\tau\end{aligned} \tag{21.5}$$

where ϕ_p is the spatial part of the one-electron function p, and η_p the corresponding spin function with τ the spin coordinate. For real ϕ, the spatial matrix element is pure imaginary because of the factor of $-i\hbar$ in the momentum operator. Taking out this factor, the spatial integrals can be written in terms of a real, but antisymmetric, matrix

$$V_{pq}^{\mathrm{BPso}} = \langle p|\frac{Z}{r^3}\mathbf{r} \times \nabla|q\rangle = -\langle q|\frac{Z}{r^3}\mathbf{r} \times \nabla|p\rangle = -V_{qp}^{\mathrm{BPso}}. \tag{21.6}$$

The antisymmetry can easily be established by integration by parts. This antisymmetry may be a little easier to see if we write the spin–orbit integral in terms of second derivatives of the regular nuclear attraction integrals:

$$V_{pq}^{\mathrm{BPso}} = \langle \nabla p|\frac{Z}{r} \times |\nabla q\rangle. \tag{21.7}$$

Since the basis functions depend on relative coordinates $\mathbf{r}_i - \mathbf{r}_A$, we can transfer the derivatives from the electron coordinates $\mathbf{r}_i$ to the nuclear coordinates $\mathbf{r}_A$, and demonstrate the antisymmetry:

$$V_{pq}^{\mathrm{BPso}} = \nabla_P \times \nabla_Q\ \langle p| - \frac{Z}{r}|q\rangle = -\nabla_Q \times \nabla_P\ \langle q| - \frac{Z}{r}|p\rangle = -V_{qp}^{\mathrm{BPso}}. \tag{21.8}$$

Here, P is the nucleus on which function p is centered, and likewise for Q and q. The nuclear attraction integral is symmetric to index interchange, but the vector product is antisymmetric.

Likewise, for the two-electron spin–orbit integrals in a real basis with real operators, the permutation properties are

$$\begin{aligned}\langle pr\,|\,\frac{1}{r_{ij}^3}\mathbf{r}_{ij}\times\nabla_i\,|\,qs\rangle &= -\langle qr\,|\,\frac{1}{r_{ij}^3}\mathbf{r}_{ij}\times\nabla_i\,|\,ps\rangle\\ &= \langle ps\,|\,\frac{1}{r_{ij}^3}\mathbf{r}_{ij}\times\nabla_i\,|\,qr\rangle.\end{aligned}\tag{21.9}$$

The difference between the first and second line is due to the fact that the gradient operator here only acts on electron i. Also, there is no longer any permutational symmetry between electrons i and j:

$$\langle pr|\hat{g}^{\mathrm{BPso}}|qs\rangle \neq \langle rp|\hat{g}^{\mathrm{BPso}}|sq\rangle.\tag{21.10}$$

As for the one-electron integrals, we can write the spatial integrals in terms of nuclear gradients of the two-electron integrals, given in Mulliken notation as

$$v_{pqrs}^{\mathrm{BPso}} = \nabla_P \times \nabla_Q (pq|rs),\tag{21.11}$$

and then the permutational symmetry can be expressed as

$$v_{pqrs}^{\mathrm{BPso}} = -v_{qprs}^{\mathrm{BPso}} = v_{pqsr}^{\mathrm{BPso}} \neq v_{rspq}^{\mathrm{BPso}}.\tag{21.12}$$

As we can see, the permutational symmetry of these integrals differs from that of the ordinary spin-free two-electron integrals, a symmetry that is extensively used in the formulation of computational algorithms. The loss of symmetry between electrons 1 and 2, combined with the fact that the operator is a vector operator, means that there are six times as many integrals as there are ordinary two-electron integrals. This factor of six makes up a considerable part of the difference between nonrelativistic calculations and four-component calculations.

The other noteworthy feature of these spin–orbit operators at this stage is in the spin part of the integrals. As we saw above, this may be extracted as

$$\int \eta_t(\tau)\,\mathbf{s}\,\eta_u(\tau)d\tau = \int \eta_t(\tau)\left[s_x\mathbf{i} + s_y\mathbf{j} + s_z\mathbf{k}\right]\eta_u(\tau)d\tau.\tag{21.13}$$

But we have shown in (6.25) that, for instance,

$$s_x = \frac{s_+ + s_-}{2}\tag{21.14}$$

and therefore the operators $\hat{h}^{\mathrm{so}}$ and $\hat{g}^{\mathrm{so}}$ connect α and β spin functions, which normally are orthogonal over spin-free operators.

Both the reduced permutational symmetry and the need to account for nondiagonal matrix elements over spin functions will complicate any computational scheme. However, the difficulties introduced at this stage are certainly less serious than the integral handling introduced in the four-component calculations.

21.2 Douglas–Kroll-Transformed Spin–Orbit Operators

In chapter 16, we showed how approximate relativistic spin-free Hamiltonians could be derived using unitary transformations. For applications to molecular systems the most frequently used model is the Douglas–Kroll–Hess model. Having successfully carried out a spin-free Douglas–Kroll–Hess calculation, we may be faced with the challenge of trying to account for the spin–orbit effects. In order to do this, we have to identify the terms that have been neglected in order to arrive at a spin-free formalism.

The one-electron part of the second-order Douglas–Kroll transformed Hamiltonian is given in (16.61) and can be expressed in two-component form for the positive-energy states as

$$\hat{\mathcal{H}}^{\mathrm{DK2}} = E_p + \hat{\mathcal{A}}\left(V + \hat{\mathcal{R}}_2 V \hat{\mathcal{R}}_2\right)\hat{\mathcal{A}} + \left(\hat{\mathcal{W}}_1 E_p \hat{\mathcal{W}}_1 + \frac{1}{2}\left[\hat{\mathcal{W}}_1^2, E_p\right]_+\right). \tag{21.15}$$

The first parenthesis contains the first-order term from the free-particle Foldy–Wouthuysen transformation, and the second parenthesis contains the second-order Douglas–Kroll term. Both $\hat{\mathcal{A}}$ and E_p are spin-free, so the spin-dependence comes from the transformations involving $\hat{\mathcal{R}}_2$, which was defined as

$$\hat{\mathcal{R}}_2 = \frac{c}{E_p + mc^2}\boldsymbol{\sigma} \cdot \mathbf{p}. \tag{21.16}$$

Note that $\hat{\mathcal{W}}_1$ is linear in $\hat{\mathcal{R}}_2$, and thus both terms contain spin-dependence. In practice, we would want to include the spin–orbit operators from both terms: even though the second-order term is $\mathcal{O}(c^{-4})$, the spin–orbit interaction behaves as $1/r^3$ and therefore this term is likely to be important. For the purposes of this section, we need only be concerned with the lowest-order spin–orbit operator, which can be identified in (16.44) as

$$\hat{\mathcal{H}}^{\mathrm{DKso1}} = \frac{1}{m^2c^2}\sum_i \hat{\mathcal{Q}}_i \hbar \boldsymbol{\sigma}_i \cdot (\nabla V_i \times \mathbf{p}_i)\hat{\mathcal{Q}}_i \tag{21.17}$$

where $\hat{\mathcal{Q}}$ is defined by (16.45). We label this one-electron operator (and later the two-electron operator) with DK, even though it is only the lowest-order part of the operator. For a point nucleus, this operator can be written as

$$\hat{\mathcal{H}}^{\mathrm{DKso1}} = \frac{1}{m^2c^2}\sum_i \hat{\mathcal{Q}}_i \frac{Z}{r_i^3} \hbar \boldsymbol{\sigma}_i \cdot (\mathbf{r}_i \times \mathbf{p}_i)\hat{\mathcal{Q}}_i. \tag{21.18}$$

It is easy to check that this operator goes over to the Breit–Pauli operator in the limit $p \to 0$, where $\hat{\mathcal{Q}} \to 1/2$:

$$\hat{\mathcal{H}}^{\mathrm{DKso1}} \to \frac{1}{2m^2c^2}\sum_i \frac{Z}{r_i^3}\mathbf{s}_i \cdot \boldsymbol{\ell}_i = \hat{\mathcal{H}}^{\mathrm{BPso1}}, \tag{21.19}$$

where we have used the fact that $\hbar\boldsymbol{\sigma} = 2\mathbf{s}$.

For the two-electron terms we need to identify the spin–orbit operators in the transformed Hamiltonian. Again, we only consider the free-particle Foldy–Wouthuysen transformation: corrections from higher-order transformations are likely to be very small. The transformed Coulomb interaction was written in (16.63) as

$$\hat{g}_{ij}^{\mathrm{C,FPFW}} = \hat{\mathcal{A}}_i\hat{\mathcal{A}}_j\left[\frac{1}{r_{ij}} + \hat{\mathcal{R}}_i\frac{1}{r_{ij}}\hat{\mathcal{R}}_i + \hat{\mathcal{R}}_j\frac{1}{r_{ij}}\hat{\mathcal{R}}_j + \hat{\mathcal{R}}_i\hat{\mathcal{R}}_j\frac{1}{r_{ij}}\hat{\mathcal{R}}_j\hat{\mathcal{R}}_i\right]\hat{\mathcal{A}}_j\hat{\mathcal{A}}_i. \tag{21.20}$$

The first term in the brackets here cannot contribute any spin-containing terms. The last term, containing a product of four $\hat{\mathcal{R}}$ operators, can only contribute to higher order in $1/c$, and will be neglected for present purposes. The terms of interest are therefore the two middle terms of the type $\hat{\mathcal{R}}_i\frac{1}{r_{ij}}\hat{\mathcal{R}}_i$. The core of the operator expression to be evaluated here is

$$(\boldsymbol{\sigma}_i\cdot\mathbf{p}_i)\frac{1}{r_{ij}}(\boldsymbol{\sigma}_i\cdot\mathbf{p}_i). \tag{21.21}$$

This expression is completely analogous to the one-electron operator, and with two such terms and reindexing, the spin–orbit contribution from the Coulomb interaction is

$$\hat{g}^{\mathrm{DKso}} = -\frac{1}{m^2c^2}\sum_{i\neq j}\hat{\mathcal{Q}}_i\hat{\mathcal{A}}_j\frac{1}{r_{ij}^3}\boldsymbol{\sigma}_i\cdot(\mathbf{r}_{ij}\times\mathbf{p}_i)\hat{\mathcal{A}}_j\hat{\mathcal{Q}}_i. \tag{21.22}$$

This is the two-electron spin–orbit interaction operator, and reduces to the Breit–Pauli form when the limit $\mathbf{p}\to 0$ is applied in the kinematic factors.

When we evaluate the integrals, we apply the kinematic factors to the basis functions, using a resolution of the identity. Likewise, writing this operator in equivalent form as

$$\hat{g}^{\mathrm{DKso}} = \frac{i}{m^2c^2}\sum_{i\neq j}\hat{\mathcal{Q}}_i\hat{\mathcal{A}}_j\boldsymbol{\sigma}_i\cdot(\mathbf{p}_i\frac{1}{r_{ij}}\times\mathbf{p}_i)\hat{\mathcal{A}}_j\hat{\mathcal{Q}}_i \tag{21.23}$$

we can apply the momentum operators to the basis functions. Transferring the derivatives from the electron coordinates $\mathbf{r}_i$ to the nuclear coordinates $\mathbf{r}_A$, and placing a tilde over the indices of the orbitals to which the $\hat{\mathcal{Q}}$ kinematic factors apply, we see that the spatial integrals for this interaction can be written in Mulliken notation as

$$v_{pqrs}^{\mathrm{DKso}} = \nabla_P\times\nabla_Q\ (\tilde{p}\tilde{q}|rs). \tag{21.24}$$

As before, basis function p is centered on nucleus P, and so on. Note that there is an $\hat{\mathcal{A}}$ kinematic factor for orbitals r and s in this integral. The permutational symmetry of these integrals follows from the properties of the vector product and the symmetry of the two-electron integrals:

$$v_{pqrs}^{\mathrm{DKso}} = -v_{qprs}^{\mathrm{DKso}} = v_{pqsr}^{\mathrm{DKso}} \neq v_{rspq}^{\mathrm{DKso}}. \tag{21.25}$$

This symmetry is the same as for the Breit–Pauli integrals.

The spin–other–orbit interaction comes from the Gaunt interaction, as shown in chapter 15 and chapter 17. The derivation of these terms is somewhat more complicated because the kinematic factors $\hat{\mathcal{Q}}_i$ can appear on either side of $1/r_{ij}$ and they do not commute with it. The derivation gives us four terms,

$$\hat{g}^{\mathrm{DKsoo}} = \frac{1}{2m^2c^2}\sum_{i\neq j}\bigg(\hat{\mathcal{Q}}_i\hat{\mathcal{Q}}_j\frac{1}{r_{ij}^3}\boldsymbol{\sigma}_i\cdot(\mathbf{r}_{ij}\times\mathbf{p}_j)\hat{\mathcal{A}}_j\hat{\mathcal{A}}_i + \hat{\mathcal{Q}}_i\hat{\mathcal{A}}_j\frac{1}{r_{ij}^3}\boldsymbol{\sigma}_i\cdot(\mathbf{r}_{ij}\times\mathbf{p}_j)\hat{\mathcal{Q}}_j\hat{\mathcal{A}}_i$$
$$+\hat{\mathcal{A}}_i\hat{\mathcal{Q}}_j\frac{1}{r_{ij}^3}\boldsymbol{\sigma}_i\cdot(\mathbf{r}_{ij}\times\mathbf{p}_j)\hat{\mathcal{A}}_j\hat{\mathcal{Q}}_i + \hat{\mathcal{A}}_i\hat{\mathcal{A}}_j\frac{1}{r_{ij}^3}\boldsymbol{\sigma}_i\cdot(\mathbf{r}_{ij}\times\mathbf{p}_j)\hat{\mathcal{Q}}_j\hat{\mathcal{Q}}_i\bigg). \quad (21.26)$$

The reader can easily verify that in the limit $\mathbf{p}\to 0$ in the kinematic factors this yields the expression obtained in the Breit–Pauli Hamiltonian. Thus, for the Douglas–Kroll Hamiltonian, calculating the primitive integrals over basis functions for these operators will involve the same work as for the Breit–Pauli Hamiltonian, but at the same time the kinematic factors will have to be accounted for.

Because of these factors, however, the final spin–orbit and spin–other–orbit integrals are not the same, and cannot be made the same. To demonstrate this, we will use the equivalent form of the spin–other–orbit operators in which we have two momentum operators,

$$\hat{g}^{\mathrm{DKsoo}} = \frac{-i}{2m^2c^2}\sum_{i\neq j}\bigg(\hat{\mathcal{Q}}_i\hat{\mathcal{Q}}_j\boldsymbol{\sigma}\cdot\mathbf{p}_i\times\mathbf{p}_j\frac{1}{r_{ij}}\hat{\mathcal{A}}_j\hat{\mathcal{A}}_i + \hat{\mathcal{Q}}_i\hat{\mathcal{A}}_j\boldsymbol{\sigma}_i\cdot\mathbf{p}_i\frac{1}{r_{ij}}\times\mathbf{p}_j\hat{\mathcal{Q}}_j\hat{\mathcal{A}}_i$$
$$+\hat{\mathcal{A}}_i\hat{\mathcal{Q}}_j\boldsymbol{\sigma}_i\cdot\mathbf{p}_j\frac{1}{r_{ij}}\times\mathbf{p}_i\hat{\mathcal{A}}_j\hat{\mathcal{Q}}_i + \hat{\mathcal{A}}_i\hat{\mathcal{A}}_j\boldsymbol{\sigma}_i\cdot\frac{1}{r_{ij}}\mathbf{p}_i\times\mathbf{p}_j\hat{\mathcal{Q}}_j\hat{\mathcal{Q}}_i\bigg). \quad (21.27)$$

As for the spin–orbit interaction, we can transfer the derivatives from the electron coordinates to the nuclear coordinates and make the operators to the left of $1/r_{ij}$ act to the left. With the tilde notation, the real spatial integrals are

$$v_{pqrs}^{\mathrm{DKsoo}} = \nabla_P\times\nabla_R(\tilde{p}q|\tilde{r}s)+\nabla_P\times\nabla_S(\tilde{p}q|r\tilde{s})+\nabla_Q\times\nabla_R(p\tilde{q}|\tilde{r}s)+\nabla_Q\times\nabla_S(p\tilde{q}|r\tilde{s}). \quad (21.28)$$

The permutational symmetry is quite different from that of the spin–orbit integrals. Here we have

$$v_{pqrs}^{\mathrm{DKsoo}} = v_{qprs}^{\mathrm{DKsoo}} = v_{pqsr}^{\mathrm{DKsoo}} = -v_{rspq}^{\mathrm{DKsoo}}. \quad (21.29)$$

This symmetry is a result of the linear combination of the four contributions.

The consequence is that we must treat the spin–orbit and the spin–other–orbit interactions separately: we cannot combine them as in the Breit–Pauli Hamiltonian. The reason is that the functions on which the momentum operators operate are derived from the small component, and only in the nonrelativistic limit where the large and small components are related by kinetic balance can we rewrite the spatial part of the spin–other–orbit interaction in the same form as the spin–orbit interaction. The reader can verify (using integration by parts) that this can be done if we neglect the kinematic factors.

A similar analysis can be applied to the use of the spin-free modified Dirac method, whether in its full four-component form or in the reduced form of the matrix NESC method. In particular, the spin–orbit and spin–other–orbit integrals have the same properties, although different forms.

21.3 Spin–Orbit Operators for Model Potential and Pseudopotential Methods

The overwhelming majority of calculations that have tried to account for relativistic effects in molecular systems have used some sort of core approximation as described in chapter 20. Spin–orbit operators for the two approaches were presented in that chapter, but we repeat them here for completeness since they provide some of the answers to "What?"

In keeping with the general philosophy of the model potential method, it is desirable to express the spin–orbit operator as the spectral resolution of the operator over the valence orbitals. One possibility is to use a semilocal potential of the form

$$\hat{V}^{\mathrm{MPso}} = \sum_{\ell} V_{\ell}^{\mathrm{so}}(r) \sum_{mm'} |\,\ell m\,\rangle\langle\,\ell m\,|\boldsymbol{\ell}\cdot\boldsymbol{\sigma}|\,\ell m'\,\rangle\langle\,\ell m'\,| \tag{21.30}$$

where the radial potential $V_{\ell}^{\mathrm{so}}(r)$ is given by

$$V_{\ell}^{\mathrm{so}}(r) = \sum_{k} \frac{D_k^{\ell}\exp(-d_k^{\ell} r^2)}{r^2}. \tag{21.31}$$

The expansion coefficients and exponents D_k^{ℓ} and d_k^{ℓ} are obtained by least-squares fitting such that the weighted deviation between the original form of the radial part of the spin–orbit operator and the potential representation is minimized.

Another possibility is to represent the radial part in terms of the atomic matrix elements,

$$\hat{V}^{\mathrm{MPso}} = \sum_{lmm'tu} |t;\,\ell m\rangle V_{ltu}^{\mathrm{so}}\langle\,\ell m\,|\,\boldsymbol{\ell}\cdot\boldsymbol{\sigma}\,|\,\ell m\,\rangle\langle u;\,\ell m| \tag{21.32}$$

where V_{ltu}^{so} is the matrix element of the all-electron spin–orbit operator that corresponds to the relativistic Hamiltonian used for the development of the model potential. Given that spin–orbit effects are well-localized, this approach should prove quite accurate.

There is a variety of approaches to the development of pseudopotential spin–orbit operators. The spin–orbit potential is mostly expressed in the semi-local form

$$\hat{V}^{\mathrm{PPso}}(\mathbf{r}) = \sum_{\ell} \frac{\Delta V_{\ell}(r)}{2\ell+1} \sum_{mm'} |\,\ell m\,\rangle\langle\,\ell m\,|\boldsymbol{\ell}\cdot\boldsymbol{\sigma}|\,\ell m'\,\rangle\langle\,\ell m'\,|. \tag{21.33}$$

This is identical in content to the expression given by (21.30). In actual application of this approach, $\Delta V_\ell(r)$ is expressed as a suitable combination of atom-centered Gaussian functions, analogous to what was done for the model potential approach.

These operators represent the spin–orbit operators for the atom whose core is being replaced. In a molecule, the spin–orbit operator must be summed over all such atomic centers.

21.4 Mean-Field Approximations for Spin–Orbit Interaction

The two-electron spin–orbit term can amount to 20 to 50% of the total spin–orbit splitting, and has the opposite sign to the one-electron contribution. This two-electron interaction should therefore be included in any calculation that aims at a quantitative description of spin–orbit effects. However, the number of two-electron spin–orbit integrals is larger than the number of two-electron Coulomb integrals by almost an order of magnitude. It would therefore be useful if we could find an approximation in which we only needed to evaluate a set of one-electron spin–orbit integrals.

The spin–orbit operators for the model potential and pseudopotential approximations are one-electron operators. These operators include the effect of the two-electron spin–orbit interaction used in the mean-field approximation to derive the model potential or pseudopotential. Molecular calculations with these potentials therefore include, at least at the atomic level, the two-electron spin–orbit terms. This is just the kind of approximation we are looking for.

It would be very convenient if we could do the same for all-electron methods. One simple approach would be to just scale the nuclear charge of the one-electron term such that

$$\hat{\mathcal{H}}^{\mathrm{so}} = \hat{\mathcal{H}}^{\mathrm{so1}} + \hat{\mathcal{H}}^{\mathrm{so2}} \approx \sum_i \frac{Z^{\mathrm{eff}}}{2m^2c^2r_i^3}\mathbf{s}_i \cdot \boldsymbol{\ell}_i \tag{21.34}$$

where Z^{eff} is an effective nuclear charge (Blume and Watson 1962). Unfortunately, this approach does not work too well: a significant part of the two-electron contribution does not have the same functional dependence as this operator, and we may end up with state-specific values for Z^{eff} for the various atomic states, making the method rather impractical.

An alternative approach is based on the application of these spin–orbit terms in a perturbation theory (PT) or configuration interaction (CI) step following a scalar mean-field calculation. For such an application we will need to evaluate matrix elements over wave functions, expressed as linear combinations of Slater determinants. To do this, we apply the usual Slater–Condon rules. We start by looking at the one-electron term, where we may examine the elements of the operator

$$\sum_{i=1}^{n} \hat{h}^{\mathrm{so}}(\mathbf{x}_i) = \sum_{i=1}^{n} \frac{1}{r_i^3}\mathbf{s}_i \cdot (\mathbf{r}_i \times \mathbf{p}_i) \tag{21.35}$$

where the variable $\mathbf{x}_i$ includes spin as well as space coordinates. For a determinant Φ the matrix element over this operator is given by

$$\langle\Phi|\sum_{i=1}^{n}\hat{h}^{\mathrm{so}}(\mathbf{x}_i)|\Phi\rangle = \sum_{t} n(\Phi,t)\langle t|\hat{h}^{\mathrm{so}}(\mathbf{x}_1)|t\rangle \tag{21.36}$$

where the occupation number $n(\Phi, t)$ is 1 when the spin–orbital t is occupied in Φ, and 0 otherwise. But, the operator $\hat{h}^{\mathrm{so}1}(\mathbf{x}_1)$ is clearly antisymmetric with respect to spatial inversion, as may be verified by making the transformation $\mathbf{r}_1 \to -\mathbf{r}_1$, and thus this integral must vanish.

For a single excitation out of Φ from spin–orbital t to spin–orbital u, represented by Φ_t^u, we get the matrix element

$$\langle\Phi|\sum_{i=1}^{n}\hat{h}^{\mathrm{so}}(\mathbf{x}_i)|\Phi_t^u\rangle = \langle t|\hat{h}^{\mathrm{so}}(\mathbf{x}_1)|u\rangle. \tag{21.37}$$

Turning now to the two-electron operator, we see that the operator from (21.4)

$$\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2) = \frac{1}{r_{12}^3}(\mathbf{s}_1 + 2\mathbf{s}_2)\cdot\mathbf{r}_{12}\times\mathbf{p}_1 \tag{21.38}$$

is not symmetric with respect to index interchange. To use the Slater–Condon rules we therefore have to write the total two-electron operator as

$$\frac{1}{2}\sum_{i\neq j}\hat{g}^{\mathrm{so}}(\mathbf{x}_i,\mathbf{x}_j) = \frac{1}{2}\sum_{i<j}\left[\hat{g}^{\mathrm{so}}(\mathbf{x}_i,\mathbf{x}_j) + \hat{g}^{\mathrm{so}}(\mathbf{x}_j,\mathbf{x}_i)\right]. \tag{21.39}$$

Again, we can invoke the inversion antisymmetry of the spatial part of the operator to establish that the diagonal matrix element vanishes

$$\langle\Phi|\frac{1}{2}\sum_{i<j}\left[\hat{g}^{\mathrm{so}}(\mathbf{x}_i,\mathbf{x}_j) + \hat{g}^{\mathrm{so}}(\mathbf{x}_j,\mathbf{x}_i)\right]|\Phi\rangle = 0. \tag{21.40}$$

Matrix elements between Φ and the singly excited determinant Φ_t^u take the form

$$\begin{aligned}
&\langle\Phi|\frac{1}{2}\sum_{i<j}\left[\hat{g}^{\mathrm{so}}(\mathbf{x}_i,\mathbf{x}_j) + \hat{g}^{\mathrm{so}}(\mathbf{x}_j,\mathbf{x}_i)\right]|\Phi_t^u\rangle\\
&=\frac{1}{2}\sum_{m} n(\Phi,m)\Big[\langle tm|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|um\rangle + \langle tm|\hat{g}^{\mathrm{so}}(\mathbf{x}_2,\mathbf{x}_1)|um\rangle\\
&\qquad\qquad - \langle tm|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|mu\rangle - \langle tm|\hat{g}^{\mathrm{so}}(\mathbf{x}_2,\mathbf{x}_1)|mu\rangle\Big] \qquad (21.41)\\
&=\frac{1}{2}\sum_{m} n(\Phi,m)\Big[\langle tm|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|um\rangle - \langle tm|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|mu\rangle\\
&\qquad\qquad - \langle mt|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|um\rangle\Big]
\end{aligned}$$

where m runs over all occupied orbitals except t and u. The final step involves relabeling of indices and elimination of one of the Coulomb-like contributions by exploiting the inversion antisymmetry of the operator again. In similar fashion, matrix elements involving the doubly excited state Φ_{tv}^{uw} yield

$$\begin{aligned}\langle\Phi|\frac{1}{2}\sum_{i<j}\Big[\hat{g}^{\mathrm{so}}(\mathbf{x}_i,\mathbf{x}_j)+\hat{g}^{\mathrm{so}}(\mathbf{x}_j,\mathbf{x}_i)\Big]|\Phi_{tv}^{uw}\rangle \\ =\frac{1}{2}\Big[\langle tv|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|uw\rangle+\langle vt|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|wu\rangle \\ -\langle tv|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|wu\rangle-\langle vt|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|uw\rangle\Big].\end{aligned} \tag{21.42}$$

It is clear from the above expressions that the inclusion of the two-electron spin–orbit interaction may lead to the evaluation of an unpleasantly large number of two-electron integrals. We might intuitively argue that double excitations are likely to be of little importance for a term that is only indirectly dependent on Z, and $\mathcal{O}(c^{-2})$, and indeed practical experience shows that these four-index integrals may be neglected without appreciable loss of accuracy. This is not the case with the three-index integrals arising from the matrix element with singly excited determinants, which include the screening of the nuclear spin–orbit interaction by the core electrons. Even though these integrals are much less numerous than the four-index integrals, we would be very happy to replace the rather time-consuming evaluation of these integrals with something simpler.

Assume for the moment that we are doing calculations on an atom with one electron in a p orbital. Depending somewhat on our ground state, for the single excitation Φ_t^u, with neither t nor u in the open shell, we might get matrix elements involving in turn each of the six spin–orbitals. It is not unreasonable to assume that instead of calculating a new matrix element for each of these spin–orbitals, we could replace them with an average over all six orbitals. Thus, this particular contribution to the two-electron spin–orbit interaction could be cast in the form

$$\begin{aligned}&\frac{1}{2}\sum_{v\epsilon p}n(\Phi,v)\Big[\langle tv|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|uv\rangle-\langle tv|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|vu\rangle-\langle vt|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|uv\rangle\Big] \\ &=\frac{1}{2}\sum_{v\epsilon p}n(v)\Big[\langle t\tilde{v}|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|u\tilde{v}\rangle-\langle t\tilde{v}|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|\tilde{v}u\rangle-\langle\tilde{v}t|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|u\tilde{v}\rangle\Big]\end{aligned} \tag{21.43}$$

where $n(v)$ is an average occupation over the orbitals of an open shell (1/6 for the example chosen here), the index v runs over the orbitals of the open shell, and the notation $\tilde{v}$ indicates that the orbitals used have been suitably modified to describe a state that is averaged over these orbitals. The argument may be extended to any type of open shell and occupation. For closed shells, there is no problem because these appear in all matrix elements in the same form. This means that the final operator appears as an effective one-electron operator, analogous to the direct and exchange terms of the standard Fock operator. In matrix form, we can express this operator as

$$V_{pq}^{\mathrm{so}(2)} = \frac{1}{2}\sum_{i}\Big[\langle pi|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|qi\rangle - \langle pi|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|iq\rangle - \langle ip|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|qi\rangle\Big]$$
$$+\frac{1}{2}\sum_{v} n(v)\Big[\langle p\tilde{v}|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|q\tilde{v}\rangle - \langle p\tilde{v}|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|\tilde{v}q\rangle - \langle \tilde{v}p|\hat{g}^{\mathrm{so}}(\mathbf{x}_1,\mathbf{x}_2)|q\tilde{v}\rangle\Big] \tag{21.44}$$

where i runs over closed-shell orbitals and v over open-shell orbitals.

Returning to the general molecular case, we see that this form makes it easier to construct matrix elements between the determinants, but it has not really helped much in reducing the number of integrals to be calculated. We still have to calculate all the primitive integrals for each of the components of the average field. However, we may introduce a second approximation, based on the form of the spin–orbit operators. These have a radial dependence of the form r^{-3}, and we can reasonably expect that the main contributions to the integrals over these operators come from the region close to the atomic nucleus. Thus, we may argue that neglect of any integrals that have contributions from more than one atomic center in the molecule should not lead to serious errors in calculating the two-electron spin–orbit interaction. This idea has been confirmed in numerous calculations.

We now have an effective one-electron operator for the spin–orbit interaction, and this operator is mostly atomic in origin. The computational savings from this last approximation to the two-electron spin–orbit integrals are considerable: not only does it eliminate a large number of integrals from our calculations, but also those that are left are single-center and very easy to compute. The high symmetry of the atom may be exploited fully to speed up these calculations. For the one-electron contribution, two-center terms are easy to calculate, and we need introduce no further approximations. The replacement of the two-electron spin–orbit integrals with atomic mean-field integrals (AMFI) was introduced by Hess et al. (1996). A similar approach for AIMPs was taken by Fedorov and Klobukowski (2002).

We could further reduce this mean-field approximation by neglecting the exchange contributions and considering only the screening due to the spherically averaged electron distribution for each atom. We can write the effective one-electron operator, in the Douglas–Kroll approximation, for instance, as

$$\hat{h}^{\mathrm{eff,so}}(\mathbf{r}_i) = -2\hat{\mathcal{A}}_i\hat{\mathcal{D}}_i\boldsymbol{\sigma}_i\cdot\left(\mathbf{p}_i\left[\int \mathrm{d}\mathbf{r}_j\frac{1}{r_{ij}}\rho_j\right]\times\mathbf{p}_i\right)\hat{\mathcal{D}}_i\hat{\mathcal{A}}_i \tag{21.45}$$

where

$$\rho_j = \sum_m |\hat{\mathcal{A}}_j\phi_m(j)|^2 \tag{21.46}$$

and we have made use of the fact that we can make the kinematic factors operate on the basis functions. This density is essentially the electron density. With neglect of the overlap between atoms and with spherical averaging, the density is a sum of spherically symmetric atomic densities, which we can expand in a set of *s* Gaussians,

$$\rho_j = \sum_S \sum_k c_{kS} N_{kS} \exp(-\zeta_{kS}\, r_{kS}^2). \tag{21.47}$$

The evaluation of the spin–orbit integrals now reduces to the same form as the one-electron spin–orbit integrals with a finite Gaussian nuclear charge distribution,

$$V_{pq}^{\text{so2}} = \sum_{kS} c_{kS} \nabla_P \times \nabla_Q \langle p | \frac{\text{erf}\,(\zeta_{kS}^{1/2} r_S)}{r_S} | q \rangle. \tag{21.48}$$

The implementation of this approach requires atomic all-electron calculations for each atom of interest, which could be spin-free or full Dirac–Hartree–Fock calculations. The spherically averaged electron density is then fitted to a suitable series of *s* Gaussians. This procedure only needs to be done once and the fitted density stored. In the molecular calculation, the fitted densities are read and the integrals can be evaluated using code for the one-electron spin–orbit integrals for a finite nucleus. Such an approach has been suggested by van Wüllen (2004).

21.5 Strategies for Spin–Orbit Methods

Having arrived at a spin–orbit operator $\hat{\mathcal{H}}^{\text{so}}$ that is consistent with our computational approach, there are a number of paths we may explore to exploit this operator in actual calculations. For heavy elements, spin–orbit contributions are of the same order of magnitude as correlation effects and therefore need to be treated within a scheme that includes correlation. One example of the importance of spin–orbit interaction is in excitation energies, where the spin–orbit interaction results in splitting of energy levels. But with the changing electron occupation of various levels, these spectra are also very sensitive to correlation effects. Another example is in the calculation of reaction surfaces, where small shifts due to spin–orbit splitting may have large kinetic effects but where the breaking and formation of bonds clearly requires that correlation be accounted for. Thus, we are looking for schemes that incorporate correlation as well as spin–orbit interactions. Starting from a suitable lowest-order spin-free wavefunction, either SCF or MCSCF, the following three approaches could be taken:

- Include correlation, then use the correlated wave function in the spin–orbit treatment.
- Include spin–orbit interaction, then use the spin–orbit corrected wave function in the correlation treatment.
- Include spin–orbit interaction and correlation at the same time.

The first two strategies require two distinct steps to include both correlation and spin–orbit effects in the overall result whereas the third does both in one step. These strategies are therefore referred to as *two-step* and *one-step* methods.

As pointed out in the beginning of this chapter, when we talk about methods for describing correlation in this connection, we are mainly referring to configuration interaction (CI) or perturbation theory (PT) methods. Coupled-cluster methods have not been used in this connection, and density functional methods have been dealt with in a separate chapter. Great efforts have gone into developing computational methods for correlation, and it would be desirable to carry as much of this over to the spin–orbit case as possible. In particular, this means that we would like to avoid the use of the double-group symmetry and complex algebra that comes with the spin–orbit operator where possible. On this basis the second of the three approaches listed above appears least attractive, introducing these added computational complications in two steps instead of only in one step as with the other two approaches. And it is indeed the first and the third approaches that have mostly been exploited for spin–orbit effects in correlated calculations.

If the spin–orbit interaction is described by a one-electron operator, this operator will have nonvanishing elements between the reference space and singly excited configurations. The two-electron interaction responsible for the correlation, on the other hand, connects the reference space with singly and doubly excited states. For this latter reason, many CI schemes are formulated at the "CI with single and double excitations" (CISD) level. It follows that the number of matrix elements required for the correlation calculations is much larger than the number required for the spin–orbit calculation, given the same reference space.[2] On this basis, it appears reasonable to provide for correlation in a separate first step and then include spin–orbit interaction in a subsequent step, something that would permit us to take advantage of the highly developed techniques for nonrelativistic correlation calculations for the most time-intensive step of the calculation.

This strategy, however, leads to at least two problems. First, to treat the spin–orbit interaction properly, correlated wave functions must be obtained for all the states that interact with the reference via the spin–orbit interaction, and as a consequence, the number of states now involved in calculating the spin–orbit interaction matrix may be much larger than the reference space. Second, the reference space used for the correlation might be unrealistic or unsuitable for systems with large spin–orbit interaction.

As already mentioned, correlation methodology is a highly technical subject, and we do not intend to review this field here. Furthermore, even within the rather narrow scope of spin–orbit methods there is a fair number of approaches, differing to a large extent in their technical approaches to the problem. Here, we will try to concentrate on the typical features of some of these methods, and refer the reader to the rather extensive literature in the field, including a number of good review articles.

Below, we discuss specific examples of the one-step and two-step approaches. Although most of these have been applied with pseudopotentials or model potentials, they are by no means dependent on this, and they may also be adapted to all-electron calculations. The discussion is not exhaustive, and among the topics not covered below are the approaches pioneered by Yarkony (1992) (based on quasidegenerate perturbation theory) and Ågren et al. (1996) (based on response theory), both of which have found

---

[2] This is not true if we include the two-electron terms in the spin–orbit operator.

their main applications in the study of reaction surfaces. Fortunately, good reviews of both approaches have been provided by the respective research groups.

21.6 One-Particle and *N*-Particle Expansion Spaces

Although we may defer the introduction of double-group symmetry by performing the correlated calculations first, we still need to take the symmetry of the spin–orbit operator into account in the selection of the N-particle expansion space.

The symmetry of the spin–orbit operator was derived in section 10.3. The spatial part of this operator transforms as the vector of rotations $\mathbf{R} = (R_x, R_y, R_z)$. This means that the spin–orbit operator will connect states of different spatial symmetry. In C_{2v}, for example, states of all spatial symmetries are connected by the spin–orbit operator. In D_{2h}, the gerade states are all connected by the spin–orbit operator, and likewise the ungerade states, but there is no connection between gerade and ungerade states because the spin–orbit operator commutes with the inversion operator (it is an even operator). The spin operator transforms as a spherical tensor of rank 1; it is essentially a triplet operator. Therefore, it can connect states whose S and M_S values differ by 0 or ±1.

Both parts of the spin–orbit operator can lead to a chain of connected states. To illustrate with the spin operator, singlet states are connected to triplet states, which in turn are connected to quintet states, and so on. Thus, in principle, we may have to consider all possible states in the spin–orbit CI calculation. However, the product of the spin symmetry and the spatial symmetry must still transform as one of the irreducible representations of the double group, and we may also make use of time-reversal symmetry to reduce the size of the CI expansion.

The upshot is that, in a one-step spin–orbit CI calculation, the configuration expansion is the same length as the expansion in an equivalent four-component calculation. The same is true in the two-step approach: to include the same effects, the number of determinants in the total expansion is the same as in the four-component calculation. The set of determinants is partitioned between the single-group (boson) symmetries, and we must perform correlated calculations for each combination of spin and spatial symmetry that yields the required double-group symmetry.

In addition to choosing the range of N-particle states that is needed for the calculation, we would also like to make the best use of symmetry in the definition of the one-particle and N-particle states. In particular, we would like to use real arithmetic wherever possible, by making appropriate choices of phases.

Our basic expansion space is the set of determinants constructed from the molecular spin–orbitals. The spin–orbitals are defined as

$$\phi_{p\tau} = \phi_p(\mathbf{r})\eta(\tau) \qquad (21.49)$$

where $\eta(\tau)$ are the α and β spin functions. These spin–orbitals form a basis for the fermion irreps of the double group, but the matrix elements are in general complex. We can at least make the matrix elements for the real irreps real by an appropriate choice of phase. Noting that the imaginary parts of the spin–orbit operator are the x and z components, we can introduce a factor of i to each spin–orbital that transforms as R_x or R_z. The one-particle matrix elements over the spin–orbit operator will therefore become real, and since the two-particle matrix elements involve the same symmetries, they will also be real. If we make the factor for the β spin–orbitals $-i$, the spin–orbitals maintain the proper relations under time reversal and can be divided into Kramers pairs. This phase factor was introduced first by Pitzer and Winter (1991).

We now have, for the real groups, a set of real integrals. The phase factors have to be propagated into the N-particle matrix elements. This means that there will be some sign changes in the determinants, and therefore in the coefficients in the CI calculations. This adds a slight complication to the CI algorithm, which is offset by the ability to use real arithmetic for the spin–orbit step. In a two-step calculation, however, we can do a normal CI calculation, without the extra phases, then change the signs of the coefficients before we do the spin–orbit calculation.

While these phase factors are useful for the real groups, we would like to be able to make the Hamiltonian matrix real for any group, if possible. After all, for an even number of electrons, we can always construct a real basis, as demonstrated in section 9.3. In fact, we can transfer the principles from that section directly to the case of spin–orbit CI and construct linear combinations of determinants that are symmetric under time reversal.

21.7 One-Step Methods

One-step methods were introduced early in the development of relativistic methodology. In 1982, Pitzer and coworkers (Christiansen et al. 1982) applied a one-step method to the six lowest states of TlH. The one-electron basis was taken from an SCF calculation using an AREP. The Hamiltonian matrix was then constructed over the full operator, including both the AREP and a SOREP of the form described above, and treating spin–orbit and correlation on an equal footing. A reference space of seven determinants was used to generate a slightly reduced set of single and double excitations for a total space of approximately 1700 determinants. Such schemes, derived from a list of configurations from which the Hamiltonian matrix is constructed directly, are known as *list-driven* schemes.

As with most CI schemes of that period, the construction of the Hamiltonian matrix and its direct diagonalization effectively limited the size of calculations to a few thousand determinants. One possible strategy for extending the capability of this type of calculation is to introduce some sort of selection criterion for the N-particle functions, and to leave out those that do not contribute appreciably. Such methods had been developed within the framework of multireference CI (MR-CI) calculations, and Hess, Peyerimhoff, and coworkers (Hess et al. 1982) extended this to the case of spin–orbit interactions. Their procedure was based on performing a configuration-selected nonrelativistic MR-CI, followed by extrapolation to zero threshold. This technique may be applied in a one-step scheme, where selection criteria are introduced not only for the correlating many-particle states, but also for those that couple to the reference space via spin–orbit interaction. The size of the calculation that has to be performed in the double group may thereby be reduced. The errors introduced by these selection procedures appear to be small.

Ultimately, however, large-scale CI calculations, which are required to treat both spin–orbit and correlation effects, are simply not possible when the Hamiltonian is constructed directly and stored, and it is necessary to discard the list-driven schemes in favor of integral-driven, or direct, CI schemes. The Graphical Unitary Group Approach (GUGA) is a very efficient method for generating the many-particle basis. This scheme has been extended to cover spin-dependent operators (Gould and Battle 1993a, 1993b) and has been implemented by Yabushita et al. (1993), allowing calculations of more than 9×10^6 determinants to be carried out. However, rather than go into the more technical details of the GUGA approach, we turn to a different one-step method that has been developed as a direct CI scheme from the outset.

The theory of direct CI was reviewed in section 12.7. The central features of the nonrelativistic theory for the development of spin–orbit CI are as follows. We write the nonrelativistic Hamiltonian, $\hat{\mathcal{H}}^{\mathrm{NR}}$, in the form

$$\hat{\mathcal{H}}^{\mathrm{NR}} = \sum_{ij} h_{ij}\hat{E}_{ij} + \frac{1}{2}\sum_{ijkl}(ij|kl)\left[\hat{E}_{ij}\hat{E}_{kl} - \delta_{jk}\hat{E}_{il}\right] \tag{21.50}$$

where we have used the generators of the unitary group, $\hat{E}_{ij}$, from (9.8). Using an N-particle basis of the type discussed above, we get elements of the **a** vector in the form

$$\mathbf{a}(I_\alpha I_\beta) = \sum_{J_\alpha, J_\beta} \langle I_\alpha I_\beta|\hat{\mathcal{H}}^{\mathrm{NR}}|J_\alpha J_\beta\rangle C(J_\alpha J_\beta). \tag{21.51}$$

It can be shown that for maximum efficiency this vector should be separated into three parts

$$\mathbf{a} = \mathbf{a}_{\alpha\alpha} + \mathbf{a}_{\alpha\beta} + \mathbf{a}_{\beta\beta} \tag{21.52}$$

where

$$\begin{aligned}\mathbf{a}_{\alpha\alpha}(I_\alpha I_\beta) = \sum_{J_\alpha}\Bigg[&\sum_{ij}\langle I_\alpha|a_i^\dagger a_j\ |J_\alpha\rangle h_{ij}\\ &+ \sum_{i>k,l>j}\langle I_\alpha|a_i^\dagger a_k^\dagger a_l\ a_j\ |J_\alpha\rangle\left[(ij|kl) - (il|kj)\right]\Bigg]C(J_\alpha J_\beta)\end{aligned} \tag{21.53}$$

and similar definitions apply to $\mathbf{a}_{\alpha\beta}$ and $\mathbf{a}_{\beta\beta}$. We see that for $\mathbf{a}_{\alpha\alpha}$ there are no barred indices in the annihilation and creation operator products, and therefore excitations only occur out of the α strings, making it possible to vectorize $\mathbf{a}_{\alpha\alpha}$ construction over the β strings.

The challenge is to devise an algorithm for the spin–orbit case that incorporates as much as possible of the machinery that makes this nonrelativistic approach so efficient. We begin by writing the spin–orbit operator in second quantization. Looking at the one-electron contribution, we may write this as

$$\hat{\mathcal{H}}^{\mathrm{so1}} = \sum_i \mathbf{v}(i)\cdot\mathbf{s}(i) \tag{21.54}$$

where $\mathbf{v}(i)$ is an operator that could be derived from any of the spin–orbit Hamiltonians we have discussed in previous sections. In the second-quantized form, this operator becomes

$$\hat{\mathcal{H}}^{\text{so1}} = \sum_{ij} \left[\langle i|\mathbf{v}\cdot\mathbf{s}|j\rangle a_i^\dagger a_j \; + \langle i|\mathbf{v}\cdot\mathbf{s}|\bar{j}\rangle a_i^\dagger a_{\bar{j}} \; + \langle \bar{i}|\mathbf{v}\cdot\mathbf{s}|j\rangle a_{\bar{i}}^\dagger a_j \; + \langle \bar{i}|\mathbf{v}\cdot\mathbf{s}|\bar{j}\rangle a_{\bar{i}}^\dagger a_{\bar{j}} \; \right] \tag{21.55}$$

where spins are indicated explicitly, and the indices i, j run over spatial orbitals. The operator $\mathbf{v}\cdot\mathbf{s}$ may be expanded using the step operators, s_+ and s_-, from section 6.3

$$\mathbf{v}\cdot\mathbf{s} = v_x \frac{s_+ + s_-}{2} + v_y \frac{s_+ - s_-}{2i} + v_z s_z. \tag{21.56}$$

If we look only at the x component of $\hat{\mathcal{H}}^{\text{so1}}$, this may be written

$$\begin{aligned}\left(\hat{\mathcal{H}}^{\text{so1}}\right)_x = \sum_{ij} \langle i|v_x|j\rangle \Big[& a_i^\dagger a_j \; \langle\alpha|\frac{s_+ + s_-}{2}|\alpha\rangle + a_i^\dagger a_{\bar{j}} \; \langle\alpha|\frac{s_+ + s_-}{2}|\beta\rangle \\ & + a_{\bar{i}}^\dagger a_j \; \langle\beta|\frac{s_+ + s_-}{2}|\alpha\rangle + a_{\bar{i}}^\dagger a_{\bar{j}} \; \langle\beta|\frac{s_+ + s_-}{2}|\beta\rangle \Big]\end{aligned} \tag{21.57}$$

from which it is obvious that only the off-diagonal spin terms will contribute such that

$$\left(\hat{\mathcal{H}}^{\text{so1}}\right)_x = \sum_{ij} \langle i|v_x|j\rangle \frac{1}{2} \left[a_i^\dagger a_{\bar{j}} \; + a_{\bar{i}}^\dagger a_j \; \right]. \tag{21.58}$$

Similar expansion of the y and z components of $\hat{\mathcal{H}}^{\text{so1}}$ shows that we can write the second-quantized one-electron spin–orbit term as

$$\hat{\mathcal{H}}^{\text{so1}} = \sum_{i,j} \mathbf{v}_{ij} \cdot \hat{\mathbf{s}}_{ij} \tag{21.59}$$

in terms of the excitation operators

$$\begin{aligned}(\hat{\mathbf{s}}_{ij})_x &= \frac{1}{2} \left[a_i^\dagger a_{\bar{j}} \; + a_{\bar{i}}^\dagger a_j \; \right], \\ (\hat{\mathbf{s}}_{ij})_y &= \frac{-i}{2} \left[a_i^\dagger a_{\bar{j}} \; - a_{\bar{i}}^\dagger a_j \; \right], \\ (\hat{\mathbf{s}}_{ij})_z &= \frac{1}{2} \left[a_i^\dagger a_j \; + a_{\bar{i}}^\dagger a_{\bar{j}} \; \right].\end{aligned} \tag{21.60}$$

The two-electron spin–orbit term may be cast in the form

$$\hat{\mathcal{H}}^{\text{so2}} = \sum_{ijkl} \mathbf{w}_{ijkl} \cdot \left[\hat{\mathbf{s}}_{ij} \hat{E}_{kl} - \delta_{jk} \hat{\mathbf{s}}_{il} \right]. \tag{21.61}$$

The matrix elements $\mathbf{v}_{ij}$ and $\mathbf{w}_{ijkl}$ may be derived from Breit–Pauli or Douglas–Kroll spin–orbit operators, or other operators that conveniently lend themselves to this type of expression.

From the final expression for $\hat{\mathcal{H}}^{\text{so}1}$ above, it is clear that operators of the type $a_i^\dagger a_{\bar{j}}$ that move electrons from β to α orbitals destroy the structure that made it possible to vectorize the construction of the **a** vectors for the nonrelativistic CI. However, the $(\hat{\mathbf{s}}_{ij})_z$ operator has the required structure, and so at least the calculations of the contributions to $\mathbf{a}_{\alpha\alpha}$ from the $(\hat{\mathcal{H}}^{\text{so}1})_z$ component should be vectorizable. To transform the expressions for the x and y components into a similar form to the z components, we note that if we introduce the overall spin step operators $\hat{S}_+$ and $\hat{S}_-$ defined by

$$\hat{S}_+ = \sum_k a_k^\dagger a_{\bar{k}} \ ; \quad \hat{S}_- = \sum_k a_{\bar{k}}^\dagger a_k \tag{21.62}$$

then the following two commutation relations hold

$$\begin{aligned} \left[\hat{S}_+, (\hat{\mathbf{s}}_{ij})_z\right] &= -a_i^\dagger a_{\bar{j}} \\ \left[\hat{S}_-, (\hat{\mathbf{s}}_{ij})_z\right] &= a_{\bar{i}}^\dagger a_j \end{aligned} \tag{21.63}$$

and the excitation operators $(\hat{\mathbf{s}}_{ij})_x$ and $(\hat{\mathbf{s}}_{ij})_y$ may be expressed as

$$\begin{aligned} (\hat{\mathbf{s}}_{ij})_x &= \frac{1}{2}\Big(\big[(\hat{\mathbf{s}}_{ij})_z, \hat{S}_+\big] + \big[\hat{S}_-, (\hat{\mathbf{s}}_{ij})_z\big]\Big) \\ (\hat{\mathbf{s}}_{ij})_y &= \frac{-i}{2}\Big(\big[(\hat{\mathbf{s}}_{ij})_z, \hat{S}_+\big] - \big[\hat{S}_-, (\hat{\mathbf{s}}_{ij})_z\big]\Big). \end{aligned} \tag{21.64}$$

Using these forms of the excitation operators, we may now express the various contributions to the **a** vector in terms of matrix elements over the $(\hat{\mathbf{s}}_{ij})_z$ excitation operator, and $\hat{S}_+$ and $\hat{S}_-$ operations performed on the vectors before or after accumulation of contributions from the integrals. For instance, the contribution from the x component to $\mathbf{a}_{\alpha\alpha}$ takes the form

$$\sum_{J_\alpha}\Big[\sum_{ij}\langle I_\alpha|a_i^\dagger a_j\ |J_\alpha\rangle(\mathbf{v}_{ij})_x + \sum_{ijkl}\langle I_\alpha|a_i^\dagger a_k^\dagger a_l\ a_j\ |J_\alpha\rangle(\mathbf{w}_{ijkl})_x\Big]C(J_\alpha, I_\beta). \tag{21.65}$$

By introducing a symmetrized two-electron matrix element $\mathbf{w}^s_{ijkl}$, defined as

$$\mathbf{w}^s_{ijkl} = \mathbf{w}_{ijkl} + \mathbf{w}_{klij}, \tag{21.66}$$

we can restrict the sum over four indices, and the x contribution to $\mathbf{a}_{\alpha\alpha}$ becomes

$$\begin{aligned} \sum_{J_\alpha}\Big[&\sum_{ij}\langle I_\alpha|a_i^\dagger a_j\ |J_\alpha\rangle(\mathbf{v}_{ij})_x \\ &+ \sum_{i>k,l>j}\langle I_\alpha|a_i^\dagger a_k^\dagger a_l\ a_j\ |J_\alpha\rangle\Big[(\mathbf{w}^s_{ijkl})_x - (\mathbf{w}^s_{ilkj})_x\Big]\Big]C(J_\alpha, I_\beta). \end{aligned} \tag{21.67}$$

This contribution is calculated twice, in fact: once for the original CI vector, after which the spin step operators are applied, and once for the CI vector after application of the spin step operators.

The action of the spin step operators on a vector **b** can be evaluated with the following expressions:

$$\begin{aligned} b_{+}(I_\alpha, I_\beta) &= \sum_{i J_\alpha J_\beta} (-1)^{n_\alpha(J_\alpha)} \langle I_\alpha | a_i^\dagger | J_\alpha \rangle \langle I_\beta | a_{\bar{\imath}} | J_\beta \rangle \, b(J_\alpha, J_\beta) \\ b_{-}(I_\alpha, I_\beta) &= \sum_{i J_\alpha J_\beta} (-1)^{n_\alpha(J_\alpha)+1} \langle I_\alpha | a_i | J_\alpha \rangle \langle I_\beta | a_{\bar{\imath}}^\dagger | J_\beta \rangle \, b(J_\alpha, J_\beta) \end{aligned} \tag{21.68}$$

where $\mathbf{b}_{\pm} = \hat{S}_{\pm}\mathbf{b}$, and $n_\alpha(J_\alpha)$ is the number of electrons in the α string J_α. We will not go into details of the algorithms involving $\hat{S}_{+}$ and $\hat{S}_{-}$ working on the various vectors. We only note that these operations are easily vectorized and that the operation count is much smaller than for the construction of the matrix elements over the $(\hat{\mathbf{s}}_{ij})_z$ excitation operator.

An alternative to this approach that does not exploit the Wigner–Eckart theorem is to factorize the products of α and β creation and annihilation operators. Such a scheme parallels the evaluation of contributions from the first off-diagonal block of the Hamiltonian in the four-component scheme, which is given in section 12.7.

The scheme outlined above (Sjøvoll et al. 1997) has been implemented in the program LUCIA. The program also exploits both double-group symmetry and time-reversal symmetry. The main computational costs over a nonrelativistic CI arise from the presence of vector operators, from the need to use complex arithmetic, and from the extended interaction space due to the fact that the spin–orbit operators connect determinants of different spin multiplicity.

21.8 Two-Step Methods

The one-step methods in principle provide the most accurate treatment of spin–orbit effects because they include both correlation and spin–orbit effects in the same computational step. The configuration expansions required to do this are large and place a practical limitation on the accuracy that can be achieved. Two-step methods promise comparable or superior accuracy based on a more rigorous treatment of electron correlation, with judicious approximations for the treatment of the spin–orbit interaction. This is especially so if we use a one-electron approximation to the spin–orbit operator. Two-step methods must still generate eigenvalues and eigenvectors of a Hamiltonian matrix that includes spin–orbit and correlation effects, but the matrix is generally much smaller than in one-step methods and the diagonalization time does not dominate the procedure.

There are a number of nonrelativistic correlation methods that are themselves two-step procedures. Those based on effective Hamiltonian theory provide a convenient starting point for the discussion of the two-step treatment of spin–orbit effects. As a framework, we present here the "configuration interaction with perturbation including spin–orbit coupling" (CIPSO) method developed in Toulouse by Teichteil and coworkers (Teichteil et al. 1983). There are also other approaches that conceptually may be cast in a similar formalism (Hess et al. 1982).

We wish to construct an appropriate effective Hamiltonian matrix $\mathbf{H}^{\text{eff}}$ for a set of M states of a system. We may make the observation that our space of N-particle functions contains functions that will have dominant contributions to these M states or will interact strongly with them. The remaining functions will be of less importance, and rather than include these in the CI they might be treated in a more approximate scheme. If we number the N-particle basis such that the first N_s functions are the strongly contributing functions, these make up a function space $S = \{|I\rangle; I \leq N_s\}$. Obviously, we must have $N_s > M$. As a first step, we diagonalize the Hamiltonian within the subspace S. This amounts to solving the eigenvalue problem

$$\hat{\mathcal{P}}_S \hat{\mathcal{H}} \hat{\mathcal{P}}_S |I\rangle = E_I |I\rangle \quad I \leq N_s \tag{21.69}$$

using the projector

$$\hat{\mathcal{P}}_S = \sum_{K=1}^{N_S} |K\rangle\langle K|. \tag{21.70}$$

The solutions obtained from this CI in the S subspace should provide us with M realistic zeroth-order approximations to the desired states.

In the second step, we build an $M \times M$ effective Hamiltonian matrix over the space of these M approximate solutions, which form a subspace S_M of S. We do this by now including all N-particle functions not in S—that is, in the external space X—through second-order perturbation theory:

$$\left[\mathbf{H}^{\text{eff}}\right]_{KL} = \langle K|\hat{\mathcal{H}}|L\rangle\delta_{KL} + \sum_{J>N_S} \frac{\langle K|\hat{\mathcal{H}}|J\rangle\langle J|\hat{\mathcal{H}}|L\rangle}{E_K - E_J}, \quad K, L \leq M. \tag{21.71}$$

Diagonalization of this matrix yields the required M states. One slight complication is that $\mathbf{H}^{\text{eff}}$ is not Hermitian, but this can be easily remedied by changing the denominator used in the perturbation expansion such that

$$E_K - E_J \ \rightarrow \ \frac{1}{2}(E_K + E_L) - E_J \equiv \Delta E_{KL,J}. \tag{21.72}$$

Up to this point, there is nothing particularly relativistic about this approach, and it is indeed a standard procedure for carrying out nonrelativistic correlated calculations. We would like to extend this scheme to include spin–orbit interaction by setting

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{H}}^{\text{so}} \tag{21.73}$$

where $\hat{\mathcal{H}}_0$ is the chosen spin-free relativistic Hamiltonian. This entire scheme now has to be cast in a relativistic format, carrying with it all the possible expenses of complex arithmetic and double-group symmetry. Furthermore, the space S has to accommodate the functions interacting over $\hat{\mathcal{H}}^{\text{so}}$, and consequently must either be much larger, or compromise on the degree to which correlation is included in the zeroth-order space.

Regardless of the sizes of the spaces, we can formulate a number of possible ways in which the effective Hamiltonian including spin–orbit interaction can be constructed. In all of these, the first step remains the solution of the eigenvalue problem in the space S for the spin-free Hamiltonian,

$$\hat{\mathcal{P}}_S \hat{\mathcal{H}}_0 \hat{\mathcal{P}}_S |I\rangle = E_I |I\rangle \quad I \leq N_s. \tag{21.74}$$

The first step therefore retains an entirely nonrelativistic symmetry treatment. Indeed, we can treat each nonrelativistic symmetry separately as they are only mixed (by $\hat{\mathcal{H}}^{\text{so}}$) in the next step.

The simplest scheme for the second step is to include spin–orbit interaction only as a first-order effect, in which case the matrix elements of the effective Hamiltonian would appear as

$$\left[\mathbf{H}^{\text{eff}}\right]_{KL} = \langle K | \hat{\mathcal{H}}_0 + \hat{\mathcal{H}}^{\text{so}} | L \rangle + \sum_{J > N_S} \frac{\langle K | \hat{\mathcal{H}}_0 | J \rangle \langle J | \hat{\mathcal{H}}_0 | L \rangle}{\Delta E_{KL,J}}, \qquad K, L \leq M. \tag{21.75}$$

This amounts to adding the spin–orbit matrix in the zeroth-order space to the spin-free effective Hamiltonian. Of course, the spin-free effective Hamiltonian now consists of blocks for each nonrelativistic symmetry, but it is not much more work than the nonrelativistic calculation for the same set of states.

For heavy elements, where the spin–orbit interaction would be expected to be large, the approach above may not be sufficiently accurate. After all, we have neglected second-order contributions from $\hat{\mathcal{H}}^{\text{so}}$ as well as any interaction over this operator between the M states selected from the diagonalization in step one and the other $N_s - M$ states. A more accurate approach would be to extend the number of states for which the spin–orbit interaction is included to the full S space,

$$\left[\mathbf{H}^{\text{eff}}\right]_{KL} = \langle K | \hat{\mathcal{H}}_0 + \hat{\mathcal{H}}^{\text{so}} | L \rangle + \sum_{J=N+1}^{N_S} \frac{\langle K | \hat{\mathcal{H}}^{\text{so}} | J \rangle \langle J | \hat{\mathcal{H}}^{\text{so}} | L \rangle}{\Delta E_{KL,J}}$$

$$+ \sum_{J > N_S} \frac{\langle K | \hat{\mathcal{H}}_0 | J \rangle \langle J | \hat{\mathcal{H}}_0 | L \rangle}{\Delta E_{KL,J}}; \qquad K, L \leq M \tag{21.76}$$

or even to the entire N-particle space including the external space,

$$\left[\mathbf{H}^{\text{eff}}\right]_{KL} = \langle K|\hat{\mathcal{H}}_0 + \hat{\mathcal{H}}^{\text{so}}|L\rangle + \sum_{J=N+1}^{N_S} \frac{\langle K|\hat{\mathcal{H}}^{\text{so}}|J\rangle\langle J|\hat{\mathcal{H}}^{\text{so}}|L\rangle}{\Delta E_{KL,J}} + \sum_{J>N_S} \frac{\langle K|\hat{\mathcal{H}}_0 + \hat{\mathcal{H}}^{\text{so}}|J\rangle\langle J|\hat{\mathcal{H}}_0 + \hat{\mathcal{H}}^{\text{so}}|L\rangle}{\Delta E_{KL,J}}; \qquad K, L \leq M. \tag{21.77}$$

This last approach involves calculating spin–orbit matrix elements for the entire set of N-particle states, and although this is probably less time-consuming than the one-step methods, would nevertheless be quite time-consuming. Which one of these schemes to use would depend on the actual system under consideration. For light elements, a sufficiently accurate description might be obtained using the first approach.

If our aim is to obtain very accurate energy differences between the spin–orbit split states, as for example when calculating an excitation spectrum, we might be concerned that the energies from the CI in subspace S are not sufficiently accurate. By its nature, the space S has to be limited, and the amount of correlation not accounted for here will give rise to errors in the calculated state energies. This in turn will affect the diagonal elements of the total Hamiltonian matrix $\mathbf{H}$. The second-order perturbation theory corrections from the external space might not be adequate to account for correlation accurately enough. We may try to improve the correlation treatment by getting better results for the spin-free calculation. For this we do not have to worry about the spin–orbit complications, and we can use a correlated calculation with a much larger N-particle basis. If we call this space R, and the energies obtained here E_I^R, we can shift the eigenvalue spectrum of the total Hamiltonian matrix by adding a term

$$\hat{\mathcal{H}}' = \hat{\mathcal{H}} + \sum_{J=1}^{N_S} |J\rangle\langle J| \left[\left(E_J^R - E_0^R\right) - \left(E_J^S - E_0^S\right)\right] \tag{21.78}$$

where E_J^S is the energy obtained in the calculation using the S subspace and E_0^R and E_0^S are the energies for a common reference state (normally the ground state) obtained in the respective function spaces. This correction, which was introduced by Llusar et al. (1996), has been shown to work quite well for molecular spectra. We are therefore constructing an effective Hamiltonian of the form

$$\left[\mathbf{H}^{\text{eff}}\right]_{KL} = \langle K|\hat{\mathcal{H}}_0 + \hat{\mathcal{H}}^{\text{so}}|L\rangle + \sum_{J=N+1}^{N_S} \frac{\langle K|\hat{\mathcal{H}}^{\text{so}}|J\rangle\langle J|\hat{\mathcal{H}}^{\text{so}}|L\rangle}{\Delta E_{KL,J}} + \sum_{J=1}^{N_S} |J\rangle\langle J| \left[\left(E_J^R - E_0^R\right) - \left(E_J^S - E_0^S\right)\right]; \qquad K, L \leq M. \tag{21.79}$$

The correlation described by second-order perturbation theory is now included variationally by means of the energy shift operator.

As an example, the space S might be derived from a CASSCF calculation, and we perform internally-contracted MR-CISD calculations for the states of interest, then evaluate the spin–orbit matrix elements over the CAS states. This reduces greatly the amount of work needed to evaluate the spin–orbit part of the total Hamiltonian, and is similar in cost at the spin–orbit CI stage to the second approach above.

There are several other schemes that treat correlation entirely variationally. In these schemes, CI calculations are performed for all M states, and the spin–orbit matrix elements are calculated between the M correlated states. The disadvantage is that MRCI calculations are required for each state in the total Hamiltonian matrix, and these states must span the space required for an adequate description of the spin–orbit coupling. The balance between spin–orbit and correlation effects in these methods is tipped towards correlation.

An alternative is to perform MRCI calculations for the key states, then extract the natural orbitals and construct a reduced set of states for the total Hamiltonian (Balasubramanian 1997a). In this approach, there is somewhat of a compromise on correlation effects, but spin–orbit effects are treated to a much higher level, and occupy a substantial part of the time taken. In any case, the diagonal can be corrected with the energies from the MRCI calculations, according to (21.78).

These are only a few of the schemes in use for computing spin–orbit interactions using correlated molecular wave functions. Whichever scheme is used, the balance between correlation and spin–orbit effects will have to be assessed on a case-by-case basis, and an appropriate method chosen.

Part V

The Nature of the Relativistic Chemical Bond

This page intentionally left blank

22

Relativistic Effects on Molecular Bonding and Structure

Having covered much of the theory relevant to relativistic quantum chemistry, the logical next step would be to turn to the application of this theory in the calculation of chemical phenomena and in the interpretation of experimental data. We shy away from this task for the reason that research and development in the field have increased greatly over the past decades, and with the present level of activity, even doubling the volume of this book might not do justice to much of the excellent work being done in this field. Also, the present publication rates and tempo of developments would probably make much of our analysis obsolete or incomplete within a short time. Fortunately, the specialist literature has been reviewed quite frequently, and with the theoretical background from the previous sections of this work this wealth of information should be accessible to the reader. A good overview can be found in the volumes edited by Schwerdtfeger (2002, 2004). The bibliographies of Pyykkö (1986, 1993, 2000) should also be consulted to locate the relevant literature sources.

There is, however, one area that we feel deserves some extra attention. This is the theory of bonding and structure in molecules. So far most of what we have presented could be regarded as pertaining as much to atomic physics as to quantum chemistry. Molecules have mainly appeared through the discussion of the molecular point groups and their treatment within the various schemes. Molecular orbital theory, valence bonding, and hybridization are concepts that are central to our understanding of molecular bonding and structure. In this final chapter, we therefore discuss how these concepts are influenced by relativity and what changes this might cause in the parameters we observe, such as valencies, bond lengths, vibrational frequencies, dissociation energies, and reaction barriers.

22.1 Relativistic Effects on Atomic Shell Structure

Our concepts of molecules are based on our concepts of atoms. Molecular orbitals are conveniently viewed in terms of linear combinations of atomic orbitals. Thus, our understanding of relativistic effects on bonding should start with an understanding of the relativistic effects on atomic spinors.

In chapter 7, we found that the mean radius of a hydrogenic spinor decreased due to relativistic effects. However, the radial extent of a spinor in a many-electron atom depends on the screening provided by the other electrons as well as the attraction of the nucleus. In a paper by Rose, Grant, and Pyper (1978), the relativistic effects on the properties of a single electron outside a closed shell were analyzed for a range of angular momentum values of the electron. Two sets of two cases were considered: a relativistic (Dirac) or a nonrelativistic (Schrödinger) Hamiltonian, which define the dynamics, combined with a relativistic or a nonrelativistic potential. These potentials were precalculated for the singly-charged ion.

The results obtained for the spinor eigenvalues (which relate to mean spinor radii through the virial theorem) are presented in table 22.1. For the s shell, the screening effects have little influence on the energy. The major influence is the change from nonrelativistic to relativistic dynamics. For the p shell, both the dynamics and the potential are important. The contributions are almost the same for the $p_{1/2}$ subshell, but for the $p_{3/2}$ subshell the effect of the dynamics is smaller by about a factor of three. For the d shell the dominant effect is the change in potential.

The effect of the dynamics is termed the *direct* relativistic effect, and the effect of the potential—the screening—is termed the *indirect* relativistic effect. The direct effects dominate for s and $p_{1/2}$ subshells, which both contract. Both have s-character, either in the large or the small component. One consequence of the contraction is that for the $6p$ elements, the $6s$ subshell stabilization contributes to the stability of the oxidation state $n-2$, where n is the valence occupation. This is often termed the "inert pair" effect. The $p_{1/2}$ shells contract less, but in the superheavy $7p$ elements both the $7s$ and the $7p_{1/2}$ form inert filled shells for the later members. For the $p_{3/2}$ subshell the dynamics and screening approximately cancel, and the behavior of the late p block elements shows little effect of relativity on many properties. For orbitals with $\ell > 1$, the screening effect dominates and the shells are destabilized. The $5d$ expands, for example, and contributes substantially to the bonding in third-row transition metals. Likewise, the $5f$ of the actinide series participates in bonding past the middle of the series, whereas in the lanthanides, the $4f$ forms a relatively inert shell inside the $5p$ shell early in the series.

Table 22.1 Effect of Dirac and Schrödinger dynamics and relativistic and nonrelativistic potentials on orbital eigenvalues (given in eV)

Dynamics:	*Dirac*	*Dirac*	*Schrödinger*	*Schrödinger*
Potential:	*Rel*	*Nonrel*	*Rel*	*Nonrel*
Au $6s$	−7.94	−7.97	−6.18	−6.01
Tl $6p_{1/2}$	−5.81	−6.79	−4.58	−5.24
Tl $6p_{3/2}$	−4.79	−5.63	−4.46	−5.24
Lu $5d_{3/2}$	−5.25	−7.32	−4.74	−6.63
Lu $5d_{5/2}$	−5.01	−6.90	−4.81	−6.63

Data taken from table 2 of Rose S. J., Grant I. P., and Pyper N. C. 1978. *J. Phys. B: At. Mol. Phys.* 11, 1171 (IOP Publishing Ltd).

Table 22.2 Relativistic shifts of outer orbitals for U

Spinor	r_{max}	$< r >$	ϵ_R	ϵ_{NR}
$6d_{5/2}$	2.503	3.290	−4.98	−7.07
$6d_{3/2}$	2.389	3.152	−5.25	−7.07
$7s$	3.657	4.341	−5.50	−4.49
$5f_{7/2}$	1.072	1.463	−8.71	−17.14
$5f_{5/2}$	1.052	1.423	−9.52	−17.14
$6p_{3/2}$	1.675	1.897	−26.67	−28.08
$6p_{1/2}$	1.488	1.664	−36.46	−28.08
$6s$	1.368	1.463	−57.96	−45.71

$r_{\max}$ – radial maximum of spinor, $< r >$ – expectation value of r over spinor, ϵ_R – spinor eigenvalue, ϵ_{NR} – orbital eigenvalue (nonrelativistic). Lengths in a.u., energies in eV.

As an example, we consider the case of U($7s^2 6d^1 5f^3$) discussed by Schwarz et al. (1989) in an extensive analysis of the origin of relativistic effects on atomic spinors. The data in table 22.2 nicely illustrate the interplay between the various effects.

The $6s$ and $7s$ spinors are both stabilized. The weighted average of the $6p$ spinors is slightly stabilized, but the $6p_{3/2}$ spinor is actually destabilized, demonstrating the poorer penetration of this subshell. The d spinors are destabilized because the indirect effect of the contracted inner shells is stronger than the direct dynamic effect. The really dramatic effect here is the strong destabilization of the $5f$ spinors. This comes about because, although the $5f$ spinors are much higher in energy that the stabilized $6s$ and $6p$ spinors, all three shells have their radial maxima in the same region. The $5f$ spinors are therefore much more sensitive to $6s$ and $6p$ contraction. The lesson to be drawn from this is that relative location in space is an important factor in the interplay of relativistic effects on different shells.

Differential screening effects are also important in the trend with Z across the periodic table. In the filling of the d shells, the screening of the $(n + 1)s$ shell becomes less and less effective, leading to a contraction. The contraction due to the decrease in screening is augmented by the direct relativistic effect, and there is a maximum in the relativistic s orbital contraction at Au (Pyykkö 1988). This effect is shown in figure 22.1, which displays the variation of the relative eigenvalue decrease and the relative mean radius decrease of the $6s$ spinor across row 6 of the periodic system. The close similarity between the two curves should be noted. At both ends of the row, there is the filling of $6s$ or $6p$ shells and the direct effect dominates. From La onwards, the $4f$ shell starts filling, and as the shell fills, it becomes less and less effective in screening the $6s$ spinor. This is even more pronounced as the $5d$ shell is completed from Hf to Hg. As the $6p$ shell starts filling, the $5d$ shell becomes more core-like and the additional stabilization from the expansion of the $5d$ shell diminishes.

Both figures show pronounced peaks for W, Pt, and Au, elements that all have ground states ($6s^1 5d^{n+1}$) with a singly occupied $6s$ spinor. This electron has to account for most of the direct relativistic effect normally shared by two electrons. The result is a further shrinkage of the $6s$ spinor and lowering of the eigenvalue. The figure of the relative spinor shrinkage also includes the results for the ($6s^2 5d^n$) configurations for these elements. These fall into the regular pattern and do not show the peaks of the ground state configurations. The "Au maximum" is therefore due to a combination of

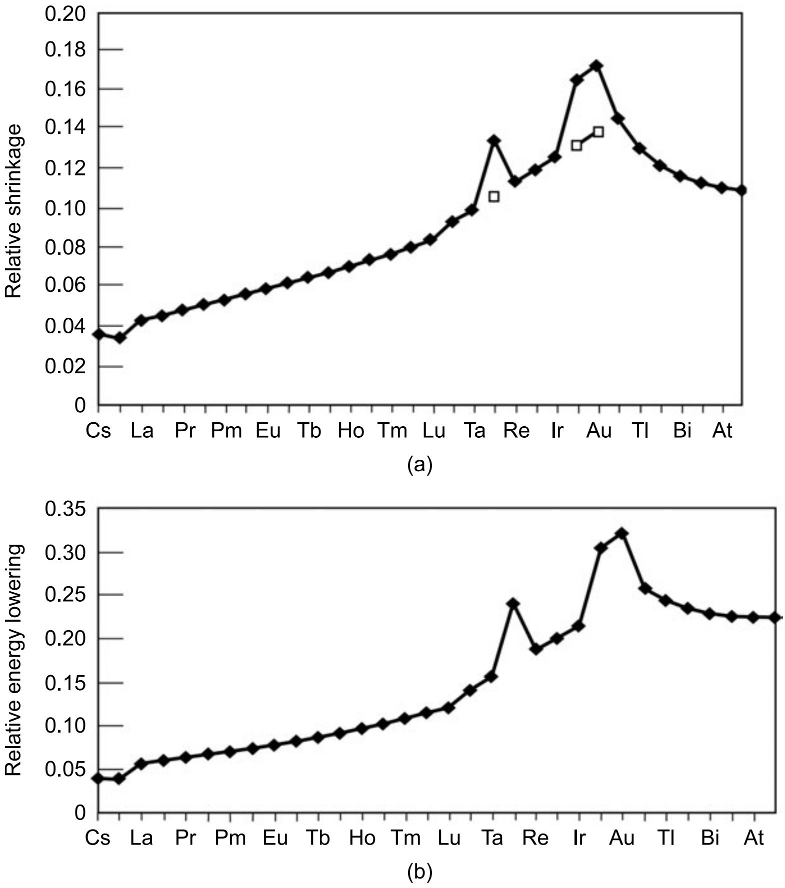


Figure 22.1 Changes in the 6*s* spinor due to relativity across the sixth period: (a) relative decrease in mean radius of the 6*s* orbital; (b) relative increase in magnitude of the 6*s* eigenvalue

the direct relativistic effect, a maximum in the stabilization provided by the relativistically expanding 5*d* shell, and the electronic structure, which favors a singly occupied 6*s* spinor. If not for the latter, we would have a “Hg maximum” instead.

22.2 Spin-Free Effects on Molecular Structure

The lowest-order effect of relativity on energetics of atoms and molecules—and hence usually the largest—is the spin-free relativistic effect (also called *scalar relativity*), which is dominated by the one-electron relativistic effect. For light atoms, this effect is relatively easily evaluated with the mass–velocity and Darwin operators of the Pauli Hamiltonian, or by direct perturbation theory. For heavier atoms, the Douglas–Kroll–Hess method or the NESC 1e method provide descriptions of the spin-independent relativistic effect that are satisfactory for all but the highest accuracy.

These two methods include terms beyond first order in perturbation theory that are important for heavy elements. Spin-free relativistic effects are also well described by relativistic effective core potentials and ab initio model potentials. In this section, we discuss the energetic and structural changes that result from these effects.

The origin of the bond-length contraction has been the subject of some debate. Is it due to the contraction of the atomic orbitals, or is it simply due to a change in dynamics? It seems to make sense that the contracted atomic orbitals should form shorter, stronger bonds. But the contraction of the orbitals involves the first-order correction to the wave function, and therefore should only influence the bond length in second order. The fact that the bond length contraction is already described in first-order perturbation theory suggests that the effect is primarily dynamic. This conclusion is in line with the discussion of atomic relativistic effects on *s* orbitals given above.

From the form of the mass–velocity and Darwin operators, it is clear that the largest contributions to the direct scalar relativistic effect come from *s* electrons. Consequently, $s\sigma$ bonds should be strengthened by relativity. The coinage metals have s^1d^{10} valence configurations and we would expect compounds of these elements to exhibit such effects in their bonding properties. Table 22.3 shows the bond lengths, harmonic

Table 22.3 Bond lengths in pm, dissociation energies in eV, and harmonic frequencies in cm^{-1} of CuH, AgH, and AuH

Molecule	*Method*	r_e^{SCF}	r_e^{MP2}	D_e^{SCF}	D_e^{MP2}	ω^{SCF}	ω^{MP2}
CuH							
	NR	156.9	145.4	1.416	2.585	1642	2024
	PT-MVD	153.9		1.485		1704	
	DKH	154.2	142.9	1.476	2.708	1698	2100
	RECP	154.3	142.9	1.465	2.696	1690	2095
	DC	154.1	142.8	1.477	2.711	1699	2101
	Exp		146.3		2.85		1941
AgH							
	NR	177.9	166.3	1.126	1.986	1473	1699
	PT-MVD	171.4		1.220		1588	
	DKH	170.1	158.7	1.229	2.190	1602	1870
	RECP	170.0	158.4	1.224	2.189	1607	1882
	DC	170.0	158.5	1.233	2.195	1605	1873
	Exp		161.8		2.39		1760
AuH							
	NR	183.1	171.1	1.084	1.901	1464	1169
	PT-MVD	162.2		1.530		1872	
	DKH	157.6	149.8	1.727	3.042	2045	2495
	RECP	157.1	149.5	1.751	3.075	2076	2512
	DC	157.0	149.7	1.778	3.114	2067	2496
	Exp		152.4		3.36		2305

NR – nonrelativistic, PT-MVD – perturbative treatment of mass–velocity and Darwin operators (only SCF), DKH – Douglas–Kroll–Hess, RECP – relativistic effective core potential, DC – four-component Dirac–Coulomb, Exp – experiment.

Data taken from Collins C. L., Dyall K. G., and Schaefer III H. F. 1995. *J. Chem. Phys.* 102, 2024 (AIP).

frequencies, and dissociation energies of the coinage metal hydrides, calculated by a number of methods (Collins et al. 1995).

Table 22.3 demonstrates clearly that relativistic effects are indeed important, even for the lightest member of the series, CuH. For the bond length, inclusion of relativistic effects leads to a reduction of around 2.5 pm, a change that must be accounted for in accurate calculations. For AuH, the nonrelativistic bond length is not even qualitatively correct. The relativistic bond reduction is 26 pm at the SCF level and 21 pm at the MP2 level. This difference of 5 pm between the two computational models for AuH also demonstrates that relativity and correlation are not simply independent, additive effects.

The simple perturbative treatment of spin-free terms from the Pauli Hamiltonian does reasonably well for the two light members of the series, but less so for AuH. As this approximation is extremely easy to program for SCF calculations and requires almost no extra computational effort, it appears as an attractive qualitative approach to relativity in medium heavy species. Only the Dirac–Coulomb (DC) results in table 22.3 account for spin–orbit interaction. The closeness of DKH and RECP results to the DC values indicate that spin–orbit effects are of minor importance, something we would expect in closed-shell molecules, where the bonding is dominated by *s* orbitals. Under these conditions, the two approximate spin-free methods can compete with the full DC operator. The agreement between the results from these three schemes also indicate that the discrepancy between the calculated and experimental values is due to insufficient description of the correlation. This observation is in line with the common experience that MP2 calculations on transition-metal compounds frequently yield somewhat short bond lengths.

If the dominant relativistic effect comes through the *s* populations, relativistic effects on reaction energies should be large when there is a significant change in *s* population. In the main group elements, such changes occur when there is a change in the s–p hybridization. For example, in a tetrahedral group 14 molecule, such as CF_4, the hybridization of C is sp^3, whereas the ground state configuration of the atom is s^2p^2. Therefore, we would expect to find a relativistic effect on the atomization energy equivalent to the effect on one $2s$ electron. Calculations show that the effect is around 4 kJ/mol, which is not negligible in accurate calculations (Bauschlicher 2000). The electronegativity of the fluorine atoms increases the effect because they withdraw more of the *s* density than, say, hydrogen would. Nevertheless, there is also a sizeable relativistic correction to the atomization energies of the group 14 hydrides.

Another example of a reaction that changes the *s* occupation is the reaction $XH_4 \rightarrow XH_2 + H_2$, where X is again from group 14. The XH_2 molecules have sp^2 hybridization, which leads to an effective population of $s^{1.5}p^{2.5}$ in the molecule. Energies for this reaction calculated at the SCF level within several approximations are given in table 22.4. Again, the importance of relativity increases as we move towards the heavier elements of the series. It becomes dramatic for Pb, which yields a solidly endothermic reaction from nonrelativistic data, but which is actually exothermic when relativity is taken into account. This is because the divalent species become increasingly more stable as Z increases, so that by Pb, the tetrahydride is thermodynamically unstable, though it is a local minimum on the potential energy surface. Note that the perturbative treatment does well for the lighter members of the series, and that the approximate RECP method competes with the full four-component DC

Table 22.4 Equilibrium SCF reaction energies ΔE_r in kJ/mol for the reaction $XH_4 = XH_2 + H_2$ with X = Si, Ge, Sn, Pb

Method	*Si*	*Ge*	*Sn*	*Pb*
NR	263	190	129	89
PT-MVD	262	174	111	30
RECP		195	102	−31
DC	261	177	97	−26

NR – nonrelativistic, PT-MVD – perturbative treatment of mass-velocity and Darwin operators, RECP – relativistic effective core potential, DC – four-component Dirac-Coulomb.

Data taken from Dyall K. G. 1992. *J. Chem. Phys.* 96, 1210 (AIP).

calculations. This is remarkable because we would expect spin–orbit effects to be of importance through the hybridization, and a fortuitous cancellation of errors cannot be excluded.

The stability of the divalent and tetravalent state for Pb also varies as a function of the electronegativity of the substituent. The inorganic chemistry of Pb is dominated by the divalent state, while the tetravalent state is dominant for organolead compounds (Kaupp and Schleyer 1993). This has traditionally been explained in terms of the *inert pair* effect: due to the low energy of the 6*s* spinor, it could only contribute poorly to the overall sp^3 hybridization required for the tetravalent state. From our insight into the interplay of relativistic effects, we are able to add nuances to this picture. Electronegative substituents withdraw electron density from the central atom and the result is a lowering of the 6*s* energy and a shrinkage of the spinor radius, in analogy to what we saw for the difference between the s^2 and s^1 configurations in the variation of the 6*s* spinor across row 6. The 6*p* spinors are also affected, but the effects of relativity act most strongly on *s* spinors. If we regard hybridization as a result of a first-order perturbative mixing of 6*s* and 6*p* spinors, the interaction will depend both on the relative location of the spinor maxima (numerator) and on the energy difference (denominator). When relativity increases the energy gap, and pushes the spinor maxima apart, the result is a decreased tendency towards sp^3 hybridization. In this context, the divalent compound is favored as there are only two substituents drawing charge away from the central atom. Electron-donating substituents would have the opposite effect, but whether the organolead compounds are thermodynamically stable or not is still an open question. Regardless, there is a large barrier to the elimination of a hydrocarbon from tetraalkyl lead compounds, so that organolead compounds at least have kinetic stability.

The stabilization of the *s* spinors as a function of *Z* thus plays an important role in hybridization. In figure 22.2, we present the variation of the valence *s* and *p* energy levels for the elements of group 14. Nonrelativistic calculations show the valence *s* and *p* orbital energies becoming closer as *Z* increases, although not evenly, due to the filling of the 3*d* and 4*f* shells. In scalar relativistic calculations, the *s* and *p* spinor energy gap increases with *Z*, making hybridization less energetically favorable (Saue et al. 1996).

In transition-metal chemistry, the energies of various states depend on the *s* population, and hence are altered by scalar relativistic effects. Even in the first transition

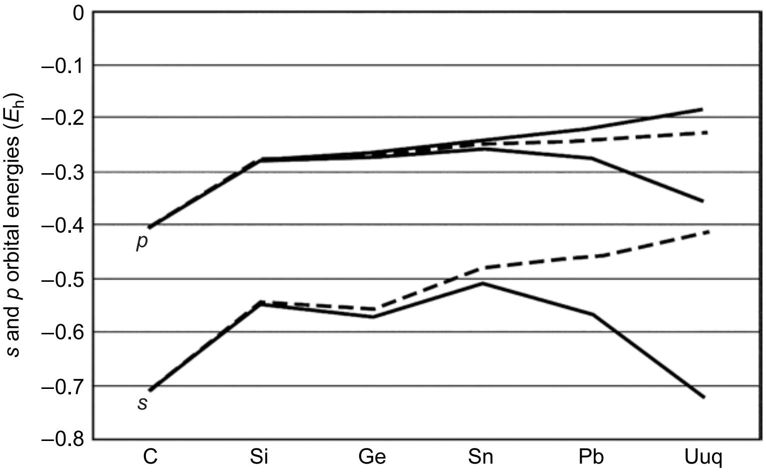


Figure 22.2 Variation of s and p orbital energies with n down group 14: solid lines – relativistic; dashed lines – nonrelativistic

series, the 4*s* spinor is relativistically lowered by several tenths of an electron volt. This series of transition metals is quite sensitive to shifts in the 4*s* orbital energies due to the spatial proximity of the 4*s* and the 3*d* orbitals (Bauschlicher et al. 1995).

In the third transition series, the 5*d* orbital expands and the 6*s* orbital contracts due to relativistic effects, as we have seen above, and many of the elements of this series display higher oxidation states than those in the second series. The greater participation of the 5*d* orbital in the late part of the series can change the qualitative bonding picture. An example comes from the activation of methane by Pt^+. In nonrelativistic calculations, the $PtCH_2^+$ product has only a single bond between the Pt^+ d^9 ion and the methylene, and the unpaired electron is located on the methylene carbon. When scalar relativistic effects are introduced, the d^8s state of Pt^+ is stabilized, and both electrons of the methylene can bond to the platinum ion, resulting in a double bond. The unpaired electron is now located on the platinum. The Pt–C bond energy more than doubles with scalar relativistic effects, from 200 to 450 kJ/mol (Heinemann et al. 1996).

Another effect of scalar relativity on transition metal chemistry (or coordination chemistry in general) is the change in binding energy of donor ligands due to the polarization of the ligand. A simple example is the binding of PH_3 to Au^+, in which the polarization of the phosphine by the metal ion reduces the *s* population on the phosphorus atom, and relativistic effects decrease the binding energy by about 1.5 kJ/mol. This is not a large effect for a single molecule, but the cumulative effect for several ligands could be significant, and the effect is larger for more highly charged ions.

A number of other examples of relativistic effects in transition metal compounds may be found in the calculations by Balasubramanian and coworkers (Balasubramanian 1997b). For the lanthanides and actinides, the situation becomes rather more complex. With a valence configuration of $s^2d^1f^n$ for the early actinides, these compounds are also strongly influenced by the spin-free relativistic effects. We refer the reader to the literature for a further discussion of these effects (Dolg 2002).

22.3 Spinor Bonds in Diatomic Molecules

Concepts such as σ and π bonds, and bonding and antibonding orbitals, are derived from the consideration of diatomic molecular orbitals. The ideas of directional bonds originate in the hybridization of atomic orbitals to form orbitals with the optimal spatial properties for bonding. More generally, the use of symmetry, molecular or local, helps us to classify orbitals into various bonding types.

With the effects of relativity on atomic orbitals that we have already discussed, it makes sense to see what pictures of bonding arise out of a consideration of the relativistic analogues of the nonrelativistic concepts. Spin-free relativistic effects change only the magnitude of bonding because there are no changes in the symmetry. Spin–orbit coupling, on the other hand, mixes states of different symmetry. In this section, we therefore consider the formation of diatomic molecular spinors from a linear combination of atomic spinors (LCAS), to gain insight into how the spin–orbit effect changes our picture of bonding.

In the following discussion, we consider only the large-component 2-spinors, because the small-component density is negligible for the valence spinors and would only complicate the picture. We analyze the spinors in terms of the α and β components, which can be represented in terms of nonrelativistic orbitals. The normalization is neglected because we are seeking a qualitative rather than a quantitative description. The molecular axis is taken to be the z axis. The two atomic spinors are considered to be on different centers, and the positive combination is given first, followed by the negative combination.

For homonuclear diatomic molecules, we can only make spinor combinations that preserve the inversion symmetry. The simplest combinations are those between spinors with $j = m_j$ for $j = \ell + \frac{1}{2}$, for which the first three are given here:

$$s_{1/2,1/2} \pm s_{1/2,1/2} = \begin{pmatrix} s \\ 0 \end{pmatrix} \pm \begin{pmatrix} s \\ 0 \end{pmatrix} \to \begin{pmatrix} s\sigma \\ 0 \end{pmatrix}, \begin{pmatrix} s\sigma^* \\ 0 \end{pmatrix} \tag{22.1}$$

$$p_{3/2,3/2} \pm p_{3/2,3/2} = \begin{pmatrix} p_1 \\ 0 \end{pmatrix} \pm \begin{pmatrix} p_1 \\ 0 \end{pmatrix} \to \begin{pmatrix} p\pi_1^* \\ 0 \end{pmatrix}, \begin{pmatrix} p\pi_1 \\ 0 \end{pmatrix} \tag{22.2}$$

$$d_{5/2,5/2} \pm d_{5/2,5/2} = \begin{pmatrix} d_2 \\ 0 \end{pmatrix} \pm \begin{pmatrix} d_2 \\ 0 \end{pmatrix} \to \begin{pmatrix} d\delta_2 \\ 0 \end{pmatrix}, \begin{pmatrix} d\delta_2^* \\ 0 \end{pmatrix}. \tag{22.3}$$

These combinations therefore give pure bonding and antibonding diatomic molecular spinors, which also have pure spin. For homonuclear diatomic molecules, the spinors have a definite inversion symmetry: $\sigma \equiv \sigma_g$, $\sigma^* \equiv \sigma_u$, $\pi \equiv \pi_u$, $\pi^* \equiv \pi_g$, and so on. The Kramers partners of these spinors are found by interchanging the spin parts. The first one is

$$s_{1/2,-1/2} \pm s_{1/2,-1/2} = \begin{pmatrix} 0 \\ s \end{pmatrix} \pm \begin{pmatrix} 0 \\ s \end{pmatrix} \to \begin{pmatrix} 0 \\ s\sigma \end{pmatrix}, \begin{pmatrix} 0 \\ s\sigma^* \end{pmatrix} \tag{22.4}$$

with similar results for the two others. With the exception of the s spinor, occupying the Kramers pair does not yield a nonrelativistic closed shell; instead, the Kramers pair is an equal linear combination of the $^3\Sigma^-$ and $^1\Sigma^+$ states.

The next example is the set of spinors with $\omega \equiv m_j = \frac{1}{2}$ for the p shell. Here, both spin–orbit components are represented:

$$p_{1/2,1/2} \pm p_{1/2,1/2} = \begin{pmatrix} -p_0 \\ \sqrt{2}p_1 \end{pmatrix} \pm \begin{pmatrix} -p_0 \\ \sqrt{2}p_1 \end{pmatrix} \rightarrow \begin{pmatrix} -p\sigma^* \\ \sqrt{2}p\pi_1 \end{pmatrix}, \begin{pmatrix} -p\sigma \\ \sqrt{2}p\pi_1^* \end{pmatrix}; \tag{22.5}$$

$$p_{3/2,1/2} \pm p_{3/2,1/2} = \begin{pmatrix} \sqrt{2}p_0 \\ p_1 \end{pmatrix} \pm \begin{pmatrix} \sqrt{2}p_0 \\ p_1 \end{pmatrix} \rightarrow \begin{pmatrix} \sqrt{2}p\sigma^* \\ p\pi_1 \end{pmatrix}, \begin{pmatrix} \sqrt{2}p\sigma \\ p\pi_1^* \end{pmatrix}. \tag{22.6}$$

For homonuclear diatomics, the positive combination has ungerade symmetry and the negative has gerade symmetry. Labeling the positive and negative combinations (for either j value) $e_{1/2}^+$ and $e_{1/2}^-$, we have $e_{1/2}^+ \equiv e_{1/2u}$, $e_{1/2}^- \equiv e_{1/2g}$.

Since the spinors formed from the two spin–orbit components have the same ω, we may take a linear combination of them. One way of combining them results in pure-spin spinors:

$$\sqrt{2}\begin{pmatrix} -p\sigma^* \\ \sqrt{2}p\pi_1 \end{pmatrix} + \begin{pmatrix} \sqrt{2}p\sigma^* \\ p\pi_1 \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ p\pi_1 \end{pmatrix}; \tag{22.7}$$

$$-\begin{pmatrix} -p\sigma^* \\ \sqrt{2}p\pi_1 \end{pmatrix} + \sqrt{2}\begin{pmatrix} \sqrt{2}p\sigma^* \\ p\pi_1 \end{pmatrix} \rightarrow \begin{pmatrix} p\sigma^* \\ 0 \end{pmatrix}. \tag{22.8}$$

This is not the only combination that can be formed, but the nature of the combination will be determined by the energetics, as we explain in what follows.

The inversion symmetry in homonuclear diatomics has an important effect on the p spinor bonding: σ bonding orbitals can mix only with π antibonding orbitals via spin–orbit coupling, and π bonding orbitals can only mix with σ antibonding orbitals. In the 2-spinor representation of these combinations this is denoted by using an asterisk to label antibonding components. In order to form a stable molecule, the bonding component must considerably outweigh the antibonding component. How feasible this is depends on the magnitude of the spin–orbit splitting. If the spin–orbit splitting is zero, the spinors derived from the two components can mix freely, and form pure-spin spinors that are either bonding or antibonding. As the spin–orbit splitting increases, the energetic cost of the mixing of the two components increases, and can exceed the spin-free dissociation energy. Consider the group 13 elements, where each has a single p valence electron. Because of the spin–orbit splitting, the electron is actually in a $p_{1/2}$ spinor, and the cost of promotion to a $p_{3/2}$ spinor increases down the periodic table. For Tl, it is about 1 eV. To form a pure σ bond requires 1/3 $p_{1/2}$ and 2/3 $p_{3/2}$, so the bond energy of Tl_2 is reduced by 2/3 of the spin–orbit splitting, or about 0.7 eV per atom. The result is a binding energy that is much smaller than expected. Experimentally, the bond dissociation energy of Tl_2 is not well known, but recent calculations and experiments place it around 0.4 eV (see Lee 2004). Similar considerations apply to the other homonuclear diatomics. The spin–orbit effect therefore reduces the bond strength of homonuclear diatomic molecules.

For heteronuclear diatomics, the inversion symmetry constraint is relaxed, and the bonding/antibonding mixtures are no longer required. Consider the following combinations:

$$p_{1/2,1/2} \pm p_{3/2,1/2} = \begin{pmatrix} -p_0 \\ \sqrt{2}p_1 \end{pmatrix} \pm \begin{pmatrix} \sqrt{2}p_0 \\ p_1 \end{pmatrix} \rightarrow \begin{pmatrix} -p\sigma \\ p\pi_1 \end{pmatrix}, \begin{pmatrix} -p\sigma^* \\ p\pi_1^* \end{pmatrix}; \tag{22.9}$$

$$p_{3/2,1/2} \pm p_{1/2,1/2} = \begin{pmatrix} \sqrt{2}p_0 \\ p_1 \end{pmatrix} \pm \begin{pmatrix} -p_0 \\ \sqrt{2}p_1 \end{pmatrix} \rightarrow \begin{pmatrix} p\sigma \\ p\pi_1 \end{pmatrix}, \begin{pmatrix} p\sigma^* \\ p\pi_1^* \end{pmatrix}. \tag{22.10}$$

These combinations are now either totally bonding or totally antibonding, thanks to the fact that one atomic spinor has two components of different sign while the other has the same sign for the two components. The molecular spinors are not symmetric: the σ part is skewed towards one atom and the π towards the other. Linear combinations of these spinors will return pure-spin spinors as before. The numerical factors for the σ and π orbitals have not been included in these spinors because they have no precise meaning here.

Because of the bonding or antibonding character of these molecular spinors, there is no promotion cost for the bonding of early p-block elements to late p-block elements to form a bond. Of course, the bond might not be optimal, and some promotion might be necessary to form the optimal bonds.

Experimental support for these bonding types can be seen in the dissociation energies of the diatomics of group 13 with group 13 and group 15 elements. In the one case, these are homonuclear diatomics in which spin–orbit effects should reduce the bond strength; in the other, they are heteronuclear diatomics in which spin–orbit effects should have little effect on the bond strength. The bonding in the second case is, to a first approximation, a bond between a $p_{1/2}$ spinor and a $p_{3/2}$ spinor. The dissociation energies D_0 are presented in table 22.5. The 13–15 diatomics show a reduction in D_0 as Z increases, a trend that can be explained by nonrelativistic calculations. Scaling the D_0 for the 13–13 diatomics by 1.5 yields a value that is close to the D_0 for the 13–15 diatomics—all except for Tl_2, whose bond energy is strongly reduced by the spin–orbit interaction.

Another example, this time theoretical, is the PbO molecule (Dyall 1993). The ground state of Pb, from DHF calculations, is approximately 92% $6p_{1/2}$ and 8% $6p_{3/2}$. The molecule has a positive charge on Pb and a negative charge on O, so the bonding is partly ionic and partly covalent. The Mulliken populations for the two highest $e_{1/2}$ spinors are given in table 22.6. The oxygen atom has a small spin–orbit splitting, but the

Table 22.5 Dissociation energies in eV for homonuclear group 13 diatomics and the heteronuclear group 13–group 15 diatomics

Molecule	D_0	$D_0 \times 1.5$	*Molecule*	D_0
Al_2	1.55	2.32	AlP	2.20
Ga_2	1.40	2.10	GaAs	2.18
In_2	1.01	1.51	InSb	1.54
Tl_2	~0.4	~0.6	TlBi	1.21

Data taken from Huber K. P. and Herzberg G., eds. 1979. *Molecular spectra and Molecular Structure. IV. Constants of Diatomic Molecules.* van Nostrand Reinhold: New York.

Table 22.6 Mulliken gross populations of PbO at the equilibrium distance r_e

	Pb		*O*	
	$p_{1/2,1/2}$	$p_{3/2,1/2}$	$p_{1/2,1/2}$	$p_{3/2,1/2}$
$25e_{1/2}$	0.00	0.25	1.45	0.01
$26e_{1/2}$	0.72	0.03	0.04	1.04

Data taken from Dyall K. G. 1993. *J. Chem. Phys.* 98, 2191 (AIP).

large spin–orbit splitting on Pb forces the bonding to conform to the $p_{1/2} - p_{3/2}$ pattern. Both spinors show little mixing of spin–orbit components on the individual atoms.

A third example is the bonding of group 13 and group 17 elements. The former have a single unpaired $p_{1/2}$ electron; the latter have a single unpaired $p_{3/2}$ electron. The Mulliken gross populations of the bonding spinors for TlAt and UutUus are given in table 22.7 (Fægri and Saue 2001). For both molecules, the nonrelativistic and spin-free relativistic results are similar: a small population on the group 13 element, statistically distributed between the $p_{1/2}$ and $p_{3/2}$ spinors. The group 17 element has the bulk of the population, indicating that the bond is fairly ionic. When the spin–orbit interaction is

Table 22.7 Mulliken gross populations for bonding atomic spinors in TlAt and UutUus

Molecule	*Method*	*Group 13*	*Group 13*	*Group 17*
		$p_{1/2,1/2}$	$p_{3/2,1/2}$	$p_{3/2,1/2}$
TlAt				
	NR	0.187	0.314	1.764
	SF	0.171	0.277	1.759
	DC	0.363	0.155	1.599
UutUus				
	NR	0.184	0.311	1.766
	SF	0.157	0.271	1.760
	DC	1.082	0.021	0.946

NR – nonrelativistic, SF – relativistic spin-free modified Dirac, DC – four-component Dirac-Coulomb. Data taken from Fægri, Jr. K. and Saue T. 2001. *J. Chem. Phys.* 115, 2456 (AIP).

added in the Dirac–Coulomb calculations, the population of the group 13 $p_{1/2}$ spinor increases at the expense of the other two spinors, making the bond less ionic. For UutUus the change is dramatic, and even leads to a change in the sign of the dipole moment. The bond is now almost a pure $p_{1/2} - p_{3/2}$ bond.

The analysis of bonding types can be extended to other angular momenta, and to linear polyatomic molecules. For a given ω value, we can formulate the following rules:

- The all-bonding or all-antibonding combinations occur for atomic spinors of opposite sign of the κ quantum number.
- The combination of bonding and antibonding character occurs for atomic spinors of the same sign of κ.

22.4 Hybridization and Bonding in Polyatomic Molecules

In organic chemistry, s–p hybridization is an important qualitative concept in the description of chemical bonds. Optimum orbitals for bonding in linear, planar trigonal, and tetrahedral configurations are formed by sp, sp^2, and sp^3 hybrids. We have previously discussed how hybridization may be influenced by spin-free relativistic effects. With the involvement of p spinors that may be highly spin–orbit split, hybridization will also be influenced by the spin–orbit coupling.

For linear molecules, we can form sp hybrids from the spinors:

$$s_{1/2} \pm p_{1/2,1/2} = \begin{pmatrix} s \\ 0 \end{pmatrix} \pm \begin{pmatrix} -p_0 \\ \sqrt{2}p_1 \end{pmatrix} \rightarrow \begin{pmatrix} sp\sigma \\ n_p \end{pmatrix}, \begin{pmatrix} sp\sigma^* \\ -n_p \end{pmatrix}; \tag{22.11}$$

$$s_{1/2} \pm p_{3/2,1/2} = \begin{pmatrix} s \\ 0 \end{pmatrix} \pm \begin{pmatrix} \sqrt{2}p_0 \\ p_1 \end{pmatrix} \rightarrow \begin{pmatrix} sp\sigma^* \\ n_p \end{pmatrix}, \begin{pmatrix} sp\sigma \\ -n_p \end{pmatrix}. \tag{22.12}$$

The spin–orbit coupling introduces lone-pair character into the bonding and antibonding orbitals. As with the p–p bonding orbitals, we may take linear combinations of these spinors. The bond with $p_{1/2}$ spinors is most affected because there is more density in the lone pair π part than in the σ part of the $p_{1/2}$ spinor. We would therefore expect that group 13 monohydrides would be more destabilized than group 17 monohydrides. To form a strong σ bond, there must be a promotion into the empty $p_{3/2}$ spinor: for group 13 the energetic cost is 2/3 the atomic spin–orbit splitting, but for group 17 it is 1/3. This cost is well demonstrated by the molecules TlH and HAt. For TlH, the cost is almost equal to the bond strength, so the molecule is weakly bound. For HAt the cost is less than the bond strength, but the bond is nevertheless weakened by the atomic spin–orbit splitting.

For molecules such as these, the spin–orbit effect is essentially quenched in the molecule, because to exist it must form a strong bond, and the spin–orbit splitting is not so large that the quenching cannot be achieved. This is obvious for light atoms such as C and Si, but even $Pb(C_2H_5)_4$ has four equivalent sp^3 hybridized bonds. We can regard the quenching of the spin–orbit interaction in both the s–p and the p–p bonding cases (and indeed in general) as a "relativistic hybridization." To form strong bonds, we must make hybrids of the $j = \ell - \frac{1}{2}$ and $j = \ell + \frac{1}{2}$ spinors, and there is an energetic cost to the hybridization, just as there is for nonrelativistic sp^n hybridization. Thus, a good approximation for many lighter elements is to reduce the bond strength calculated without spin–orbit effects by the appropriate multiple of the experimental atomic spin–orbit splitting.

The superheavy elements are beyond the quenching regime, however, and the spin–orbit effect is so large that it is not quenched. In HUus, for example, the $7p_{1/2}$ shell contracts significantly and forms an inert pair, leaving an unpaired $7p_{3/2}$ electron to form an s–$p_{3/2}$ bond, which is weaker than a regular σ bond and contains a significant lone-pair component (Saue et al. 1996).

The capability for relativistic hybridization is present for molecules of other than linear symmetry. The symmetry spinors for groups lower than cubic can be classified in terms of the m_j quantum numbers for spinors at the high symmetry point (or axis). Since the j quantum number is not part of the classification, relativistic hybridization can always take place. For example, in D_{3h} the fermion irreps are $e_{1/2}$, $e_{3/2}$, and $e_{5/2}$, for which the Kramers partners have $m_j \bmod 6 = \pm 1/2, \pm 3/2, \pm 5/2$, respectively. The $j = \ell - \frac{1}{2}$ and $j = \ell + \frac{1}{2}$ spinors for a given m_j both belong to the same irrep. In the cubic groups, the spin–orbit components of a shell can belong to different irreps, or be partially in the same irrep. For example, in T_d the $p_{1/2}$ and $p_{3/2}$ spinors at the center of symmetry belong to Γ_7 and Γ_8, respectively; the $d_{3/2}$ also belongs to Γ_8 and the $d_{5/2}$ is split between Γ_6 and Γ_8.

The simple bonding picture based on spinors and relativistic or spin–orbit hybridization is an appealing explanation of the effect of spin–orbit interaction on bonding, but it is by no means the only one. It is the natural explanation when spinors are the starting point, but it is also possible (and common) to start from a spin-free model and introduce spin–orbit effects at the post-SCF stage of the calculation. Now the bonding picture is a multi-electron picture that involves the interaction of states of different spin and symmetry.

Taking TlH as an example, we can expand the molecular spinor combination in terms of spin–orbitals, and collect together the two-electron determinants that form the ground state. Let us write the spinor and its Kramers partner as:

$$\begin{aligned} e_{1/2} &= \begin{pmatrix} c_1 sp\sigma \\ \\ c_2 p\pi_1 \end{pmatrix} = c_1(sp\sigma)\alpha + c_2(p\pi_1)\beta; \\ e_{-1/2} &= \begin{pmatrix} -c_2 p\pi_{-1} \\ \\ c_1 sp\sigma \end{pmatrix} = c_1(sp\sigma)\beta - c_2(p\pi_{-1})\alpha. \end{aligned} \tag{22.13}$$

The doubly occupied Kramers pair can then be written

$$\begin{aligned} e_{1/2}e_{-1/2} = {} & c_1^2(sp\sigma)^2\alpha\beta + c_2^2(p\pi_{-1})(p\pi_1)\alpha\beta \\ & - c_1c_2(sp\sigma)(p\pi_{-1})\alpha\alpha - c_1c_2(sp\sigma)(p\pi_1)\beta\beta. \end{aligned} \tag{22.14}$$

This Dirac–Hartree–Fock single-determinant state is a linear combination of the closed-shell σ-bonded $\sigma^2\ {}^1\Sigma^+$ state, an equal mixture of the $\pi^2\ {}^3\Sigma^-$ and $\pi^2\ {}^1\Sigma^+$ states and some of the $\sigma\pi\ {}^3\Pi$ states. As discussed in chapter 12, the DHF method must be considered a multiconfiguration method when the spinors are expanded into states based on spin–orbitals. To obtain a proper description, it is really necessary to perform a MCSCF calculation in the active space spanned by the two possible $e_{1/2}$ spinors.

The alternative model of the mixing of many-electron states is one that is probably better understood by quantum chemists. The change in dissociation energies due to spin–orbit interaction can often be viewed in terms of avoided crossings of states that have different spin. A good example is the dissociation of PtH_2. The ${}^1A_1$ ground state of this molecule dissociates to Pt $(d^9s^1)\ {}^1D$ and H_2, in the absence of spin–orbit coupling. However, the ground state of Pt is the $(d^9s^1)\ {}^3D$ state, and when spin–orbit effects are included, there is a spin–orbit avoided crossing at a fairly large bond distance that creates a barrier in the dissociation curve (Nakatsuji et al. 1998).

One important case to consider is that of Jahn–Teller distortion, which arises when the molecular state has spatial degeneracies. In the double group, all groups lower than cubic have only doubly degenerate fermion irreps, and this degeneracy represents the spin. Therefore, in molecules that have lower than cubic symmetry, there should in principle be no Jahn–Teller effect. The situation is not quite so simple, however. The spin–orbit effect mixes the components that are degenerate in the single group, and the effect is largest at the high symmetry point. The shape of the lowest state depends on the relative magnitudes of the spin–orbit splitting and the Jahn–Teller distortion. If the spin–orbit splitting is small, there will be a double-well potential with minima near the two Jahn–Teller minima. If the spin–orbit splitting is more than twice as large as the Jahn–Teller distortion energy, there is only a single well with the minimum at the symmetric point. An example is WF_5, which has a single electron in an e'' orbital at the D_{3h} symmetric point. Jahn–Teller distortion lowers the symmetry to C_{2v}, but the spin–orbit interaction quenches the Jahn–Teller effect to return the molecule to D_{3h} symmetry (Dyall 2000a). UF_5 is another example. Here, both the Jahn–Teller effect and the spin–orbit splitting are relatively small, and the molecule is fluxional with a C_{4v} minimum (Onoe et al. 1997). However, in a molecule like Au_3, which has an $(a_1')^2(e'')^1$ configuration in D_{3h} symmetry, there is no quenching because there is no spin–orbit interaction due to the fact that the bonding orbitals are composed of the Au 6*s* orbitals.

22.5 Relativistic Effects on Properties

While this chapter is devoted to relativistic effects on bonding and not to properties, there are some intrinsic properties that relate directly to the discussion of relativistic effects on bonding: the dissociation energy and the vibrational frequency (or the force constant). When it comes to these properties—and indeed any properties—the relativistic effect can be broken down into two parts. Since relativity changes the Hamiltonian, and consequently also changes the geometry of a molecule, the relativistic effect on a property has a part that is the change in the nonrelativistic property due to the relativistic change in the geometry, and a part that is due to the relativistic change in the property at the relativistic geometry.

We can illustrate this principle with an analysis of the force constant. To do so, we write the relativistic energy as a function of position in terms of the nonrelativistic energy and a relativistic correction:

$$E^{\text{rel}}(\mathbf{q}) = E^{\text{NR}}(\mathbf{q}) + \mathit{\Delta} E^{\text{rel}}(\mathbf{q}) \tag{22.15}$$

where $\mathbf{q}$ is the vector of nuclear coordinates. For the purpose of the analysis, we take the origin to be at the nonrelativistic geometry. The relativistic correction to a coordinate is written as $\mathit{\Delta} q_i^{\text{rel}}$ and the coordinate relative to the relativistic minimum is written as q_i^{rel}.

We expand the nonrelativistic energy in a Taylor series in the coordinates,

$$E^{\text{NR}}(\mathbf{q}) = E^{\text{NR}}(0) + E_i^{\text{NR}}(0)q_i + \tfrac{1}{2}E_{ij}^{\text{NR}}(0)q_i q_j + \tfrac{1}{6}E_{ijk}^{\text{NR}}(0)q_i q_j q_k + \cdots \tag{22.16}$$

where we have used the Einstein summation convention, and the subscripted quantities are the derivatives of the energy, that is

$$E_i^{\text{NR}} = \frac{\partial E^{\text{NR}}}{\partial q_i}, \qquad E_{ij}^{\text{NR}} = \frac{\partial^2 E^{\text{NR}}}{\partial q_i \partial q_j}, \qquad \cdots . \tag{22.17}$$

We can similarly expand the relativistic correction to the energy, but this we do around the relativistic geometry:

$$\mathit{\Delta} E^{\text{rel}}(\mathbf{q}) = \mathit{\Delta} E^{\text{rel}}(\mathbf{q}^{\text{rel}}) + \mathit{\Delta} E_i^{\text{rel}}(\mathbf{q}^{\text{rel}})q_i^{\text{rel}} + \tfrac{1}{2}\mathit{\Delta} E_{ij}^{\text{rel}}(\mathbf{q}^{\text{rel}})q_i^{\text{rel}} q_j^{\text{rel}} + \cdots \tag{22.18}$$

where the $\mathit{\Delta} E^{\text{rel}}$ terms are the relativistic corrections to the energy and its derivatives. We now combine these two expressions to obtain

$$\begin{aligned} E^{\text{rel}}(\mathbf{q}) = E^{\text{NR}}(0) &+ E_i^{\text{NR}}(0)q_i + \tfrac{1}{2}E_{ij}^{\text{NR}}(0)q_i q_j + \tfrac{1}{6}E_{ijk}^{\text{NR}}(0)q_i q_j q_k + \cdots \\ &+ \mathit{\Delta} E^{\text{rel}}(\mathbf{q}^{\text{rel}}) + \mathit{\Delta} E_i^{\text{rel}}(\mathbf{q}^{\text{rel}})q_i^{\text{rel}} + \tfrac{1}{2}\mathit{\Delta} E_{ij}^{\text{rel}}(\mathbf{q}^{\text{rel}})q_i^{\text{rel}} q_j^{\text{rel}} + \cdots . \end{aligned} \tag{22.19}$$

The relativistic quadratic force constant k_{ij}^{rel} is defined as the second derivative of the relativistic energy with respect to geometric displacements, evaluated at the minimum of the relativistic potential energy surface. We define the relativistic correction to a force constant as the difference between the relativistic force constant at the relativistic geometry and the nonrelativistic force constant at the nonrelativistic geometry. Differentiating twice, evaluating at the relativistic minimum and subtracting the nonrelativistic force constant, we get the relativistic correction to the quadratic force constant,

$$\mathit{\Delta} k_{ij}^{\text{rel}} = k_{ij}^{\text{rel}} - k_{ij}^{\text{NR}} = E_{ij}^{\text{rel}}(\mathbf{q}^{\text{rel}}) - E_{ij}^{\text{NR}}(\mathbf{q}^{\text{NR}}) = \mathit{\Delta} E_{ij}^{\text{rel}}(\mathbf{q}^{\text{rel}}) + E_{ijk}^{\text{NR}}(0)\mathit{\Delta} q_k^{\text{rel}} + \cdots \tag{22.20}$$

Thus, we can decompose the relativistic correction to the quadratic force constant into a correction to the force constant at the relativistic geometry—the change in curvature due to the relativistic Hamiltonian—and terms that depend on the magnitude and sign of the nonrelativistic cubic force constant (and higher terms) and the magnitude and sign of the relativistic correction to the bond length.

For stretching modes, the cubic force constant is usually negative. If the relativistic correction to the bond length is also negative, that is, the bond is shortened due to relativity, the product of these two terms yields a positive contribution to the relativistic correction to the quadratic force constant. If the increase in curvature is also positive, the force constant increases. This situation usually arises when relativity produces shorter bonds and deeper potential wells, such as for AuH. If, on the other hand, the bond length increases and the well becomes shallower, the force constant decreases, such as for PbO and Tl_2, where the spin–orbit interaction is primarily responsible for the changes. Another example is the lead hydrides, where the totally symmetric stretching frequency increases relativistically for PbH_4 but decreases for PbH_2, due to the different signs of the nonrelativistic cubic force constant (Dyall 1992).

For bending modes, the geometric effect is similar. The relativistic correction to the curvature usually increases if the lengths of the two bonds defining the bending mode shorten, so that shorter bond lengths produce an increase in the bending force constant. This is the case for PbH_2, for example.

This kind of analysis can be applied to other properties. For the correlation energy, the relativistic correction is simple:

$$\begin{aligned}\Delta E^{\mathrm{rel}} &= \left[E^{\mathrm{rel}}(\mathbf{q}^{\mathrm{rel}}) - E^{\mathrm{NR}}(\mathbf{q}^{\mathrm{rel}})\right] + \left[E^{\mathrm{NR}}(\mathbf{q}^{\mathrm{rel}}) - E^{\mathrm{NR}}(\mathbf{q}^{\mathrm{NR}})\right] \\ &= \Delta E^{\mathrm{Ham}}(\mathbf{q}^{\mathrm{rel}}) + \Delta E^{\mathrm{geom,NR}}\end{aligned} \tag{22.21}$$

where the energies here are the correlation energies.

Given that the lowest-order one-electron relativistic correlation contribution to the energy is zero for a closed shell, it is possible that the relativistic corrections to correlation energies are mostly geometric in origin. However, the perturbation is not small near the nucleus, and higher-order effects could be significant. An illustration is given in table 22.8 for the coinage metal hydrides, comparing DHF results with nonrelativistic results. For these systems, the spin–orbit contribution is small, and therefore the ambiguity over the multireference nature of the relativistic wave function is not an issue. The relativistic correction to the correlation energy increases with Z, as might be expected. The geometric effect, which increases almost linearly with the relativistic increase in bond length, is smaller than the effect of the Hamiltonian. We can conclude that the effect of changing the wave function, which contributes to the Hamiltonian effect, is significant even for Cu.

Table 22.8 Effect of geometry and Hamiltonian on the correlation energy of the coinage metal hydrides

System	ΔE^{rel}	$\Delta E^{Ham}(\boldsymbol{q}^{rel})$	$\Delta E^{geom,NR}$	$\Delta \boldsymbol{q}^{rel}$
CuH	−9.02	−7.50	−1.53	−2.53
AgH	−16.88	−12.98	−3.90	−7.58
AuH	−50.70	−38.01	−12.28	−21.26

Energies in millihartrees, relativistic correction to the bond length in pm.

22.6 A Final Warning

In this chapter, we have tried to provide an analysis of relativistic effects on bonding in molecules, pointing to some major trends that are easily recognizable in routine chemical research. However, real life only seldom reduces to simple model systems, and in most situations relativistic effects are accompanied by a number of other, more or less traditional, chemical effects. Added to this is the fact that relativistic effects are not really molecular properties, but rather they are defined as the difference between an incomplete (nonrelativistic) model and a more complete, but often less intuitive, (relativistic) model. This is of course also true of correlation effects. Trying to obtain qualitative insight by balancing all these effects against each other may be very difficult. In the long run, there may be no better way of obtaining the answers than doing calculations within a reliable model.

Appendices

This page intentionally left blank

Appendix A

Four-Vector Quantities

This appendix lists various four-vector quantities encountered in this book.

$$\mathrm{w} = (\mathbf{r}, ict) = (x, y, z, ict) \tag{A1}$$

$$\mathrm{p} = \left(\mathbf{p}, \frac{iE}{c}\right) = \left(p_x, p_y, p_z, \frac{iE}{c}\right) = (p_x, p_y, p_z, imc) \tag{A2}$$

$$\Box = \left(\nabla, \frac{-i}{c}\frac{\partial}{\partial t}\right) = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}, \frac{-i}{c}\frac{\partial}{\partial t}\right) \tag{A3}$$

$$\Box^2 = \left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) \tag{A4}$$

$$\mathrm{j} = (\mathbf{j}, ic\rho) = (j_x, j_y, j_z, ic\rho) \tag{A5}$$

$$\mathrm{A} = \left(\mathbf{A}, \frac{i\phi}{c}\right) = \left(A_x, A_y, A_z, \frac{i\phi}{c}\right) \tag{A6}$$

$$\alpha = (\boldsymbol{\alpha}, i\mathbf{I}_4) = (\alpha_x, \alpha_y, \alpha_z, i\mathbf{I}_4) \tag{A7}$$

Appendix B

Vector Relations

This appendix lists some of the more commonly used relations in vector algebra. Most of them are applied at least once in this book.

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) \tag{B1}$$

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}\,(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}\,(\mathbf{a} \cdot \mathbf{b}) \tag{B2}$$

$$\nabla\,(\mathbf{a} \cdot \mathbf{b}) = (\mathbf{a} \cdot \nabla)\,\mathbf{b} + (\mathbf{b} \cdot \nabla)\,\mathbf{a} + \mathbf{a} \times (\nabla \times \mathbf{b}) + \mathbf{b} \times (\nabla \times \mathbf{a}) \tag{B3}$$

$$\nabla \cdot (\phi \mathbf{a}) = (\nabla \phi) \cdot \mathbf{a} + \phi(\nabla \cdot \mathbf{a}) \tag{B4}$$

$$\nabla \times (\phi \mathbf{a}) = (\nabla \phi) \times \mathbf{a} + \phi(\nabla \times \mathbf{a}) = -\mathbf{a} \times (\nabla \phi) + \phi(\nabla \times \mathbf{a}) \tag{B5}$$

$$\nabla \times (\nabla \phi) = 0 \tag{B6}$$

$$\nabla \cdot (\nabla \times \mathbf{a}) = 0 \tag{B7}$$

$$\nabla \times (\nabla \times \mathbf{a}) = \nabla\,(\nabla \cdot \mathbf{a}) - \nabla^2 \mathbf{a} \tag{B8}$$

$$\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \;\Rightarrow\; \nabla \cdot \mathbf{r} = 3; \quad \nabla \times \mathbf{r} = 0 \tag{B9}$$

$$\mathbf{e}_r = \frac{\mathbf{r}}{r} \;\Rightarrow\; \nabla \times \mathbf{e}_r = 0; \quad (\mathbf{a} \cdot \nabla)\, \mathbf{e}_r = \frac{1}{r}\big[\mathbf{a} - \mathbf{e}_r\, (\mathbf{a} \cdot \mathbf{e}_r)\big] \tag{B10}$$

$$\nabla = \mathbf{e}_r \frac{\partial}{\partial r} + \mathbf{e}_\vartheta \frac{1}{r}\frac{\partial}{\partial \vartheta} + \mathbf{e}_\varphi \frac{1}{r \sin\vartheta}\frac{\partial}{\partial \varphi} \tag{B11}$$

$$(\boldsymbol{\sigma} \cdot \mathbf{a})(\boldsymbol{\sigma} \cdot \mathbf{b}) = (\mathbf{a} \cdot \mathbf{b})\mathbf{I}_2 + i\boldsymbol{\sigma} \cdot \mathbf{a} \times \mathbf{b} \tag{B12}$$

Appendix C

Elements of Group Theory

In this appendix, we briefly review some of those concepts in group theory that are used in our discussion of symmetry. Most of our readers will no doubt be familiar with these concepts already. We therefore provide no proofs or examples: these can be found in one of the many good introductory accounts of group theory (Atkins and Friedman 1997, Schonland 1965, Hamermesh 1962, Tinkham 1964).

We start by defining a group:

> Given a set G of elements $\{g, h, ...\}$ and a binary rule of combination (*group multiplication*) between elements, the set forms a group under the rule of combination if and only if
>
> — every product of two elements in the group is a member of the group, that is, $g, h \in G; k = gh \Rightarrow k \in G$
> — there exists an identity element $e \in G$ such that any element multiplied by e is the element itself, that is, $\exists\, e \in G$ such that $eg = ge = g$ for all $g \in G$
> — for any element g in the group there exists an inverse g^{-1} that is also an element of the group, that is, $g \in G \Rightarrow\ \exists\, g^{-1} \in G$ such that $gg^{-1} = g^{-1}g = e$
> — the associative law holds for group multiplication, that is, $g(hk) = (gh)k$ for all $g, h, k \in G$.

The elements of a group do not necessarily commute under the group multiplication, that is, we may in general have $gh \neq hg$ where $g, h \in G$. If all the elements of the group commute under group multiplication, then the group is called *Abelian*.

The *order* of the group is the number of elements in the set. This need not be finite, and the set need not even be countable.

A *subgroup* of G is a subset of G that itself fulfills the group requirements. G and e are considered *improper* subgroups of G, all other subgroups are *proper*.

A *homomorphism* is a mapping from one group *into* another that preserves products. If μ denotes the mapping from G to G', such that $g_1, g_2 \in G$ have the images $\mu(g_1), \mu(g_2) \in G'$, then homomorphism requires that $\mu(g_1 g_2) = \mu(g_1)\mu(g_2)$. If every element of G' is the image of an element of G, the mapping is *onto* the set G'. A homomorphism that is 1-to-1 and onto is called an *isomorphism*.

A *representation* of a group is any group of concrete mathematical entities that is homomorphic to the group. More commonly, representations refer to groups of matrices (linear transformations) that are homomorphic to the group. To be precise, a matrix representation of a finite group G is a homomorphism of G onto the linear group of the nonsingular $n \times n$ complex matrices, called the linear group $GL(n, C)$.

If a transformation can be found that reduces all the matrices of one representation to the same block diagonal structure (e.g., a set of 5×5 matrices are reduced to $2 \times 2 + 2 \times 2 + 1 \times 1$ diagonal blocks), the representation is *reducible*, otherwise it is *irreducible*. We will use the abbreviation *irrep* for irreducible representation, and *rep* for a representation in general.

Many of the properties of the representation matrices are carried by the trace of the matrix, that is, the sum of the diagonal elements of the matrix. For the matrix $\mathbf{D}(g)$, where $\mathbf{D}(g)$ is the image of $g \in G$, we write this as

$$Tr(\mathbf{D}(g)) = \sum_i D_{ii}(g). \tag{C1}$$

This trace is called the *character* of the operation in this representation.

Let the ith irrep of a group G of order n be a set of $d_i \times d_i$ matrices $\mathbf{D}^i(g)$. Then, the *great orthogonality theorem* says that

$$\sum_{g \in G} D^i_{\kappa\lambda}(g)^* D^j_{\mu\nu}(g) = \frac{n}{d_i} \delta_{ij} \delta_{\kappa\mu} \delta_{\lambda\nu}. \tag{C2}$$

Two elements g, h of the group G are *conjugate* if there is an element k of the group such that $g = khk^{-1}$. The group may be divided into disjoint sets of mutually conjugate elements. These sets are called *classes*. As the trace of a product of matrices is invariant to cyclic permutation of the matrices, that is

$$Tr(\mathbf{ABC}) = Tr(\mathbf{BCA}) = Tr(\mathbf{CAB}), \tag{C3}$$

all elements in the same class have the same characters

$$\begin{aligned} Tr\big(\mathbf{D}(k)\mathbf{D}(g)\mathbf{D}(k^{-1})\big) &= Tr\big(\mathbf{D}(g)\mathbf{D}(k^{-1})\mathbf{D}(k)\big) = Tr\big(\mathbf{D}(g)\mathbf{D}(k^{-1}k)\big) \\ &= Tr(\mathbf{D}(g)\mathbf{D}(e)) = Tr(\mathbf{D}(ge)) = Tr(\mathbf{D}(g)). \end{aligned} \tag{C4}$$

The identity element is clearly in a class by itself. It is also the only class that is a subgroup. All other classes lack the identity element. In an Abelian group, all elements commute, and every element is therefore in a class by itself.

For a finite group, the number of classes equals the number of irreps. Furthermore, the sum of the squares of the dimensions (d_i) of the irreps must equal the order (n) of the group

$$\sum_i d_i^2 = n. \tag{C5}$$

Appendix D

Group Tables

There is a great deal of material on group theory of both the single and double groups. The most complete is the treatise by Altmann and Herzig (1994). In this appendix, we present information on the basis functions and principles for determining the products of basis functions. Most of this information is available elsewhere, but for convenience we present a summary here.

In this appendix, we present information for groups with only one axis of order higher than two. Most of the information can be derived by considering finite rotations of angular momentum eigenfunctions about the principal axis, which is taken to be the z axis. The basis functions can be classified in terms of the azimuthal quantum number (m_ℓ or m_j) and the inversion symmetry. For this purpose, the notation for the doubly degenerate reps and irreps might differ from those presented in the literature. Here, we number them according to the value of m_ℓ mod n or m_j mod n, where n is the order of the principal axis. The labels for the singly degenerate irreps that form a rep include a sign: for example, in the group C_3 the two (boson) irreps that form the E rep are labeled E_1 and E_{-1}. These labels are also used to distinguish the rows of the doubly degenerate irreps, such as in C_{3v}. The subscript is unconventional but it is used for the purpose of analysis. The doubly degenerate irreps themselves are given a positive subscript. The slightly unfortunate side of this notation is that an irrep and the row of an irrep can have the same label, but the context should make it clear which is being referred to.

The boson basis functions are presented in table D1. The irreps for these groups consist of nondegenerate irreps, which are real; singly degenerate irreps occurring in complex conjugate pairs that form reps; and doubly degenerate irreps, which are complex. The basis function is presented for the first row of the degenerate irreps; the corresponding basis function for the second row is the complex conjugate of the function for the first row, that is the function for $-m_\ell$ (to within a phase). For the real irreps, linear combinations of the complex spherical harmonics must be made to form the real basis functions. For the C_n, C_{nh}, and S_n groups, the E reps occur as singly degenerate pairs.

Table D1 Boson basis functions for groups with only one principal axis of order greater than 2

Group	*(ir)rep*	*Basis functions*
C_{2n}, C_{2nv}, D_{2n}	A_*	$m_\ell \bmod 2n = 0$
	B_*	$m_\ell \bmod 2n = n$
	E_m	$m_\ell \bmod 2n = m$, $0 < m < n$
C_{2n+1},$C_{(2n+1)v}$, D_{2n+1}	A_*	$m_\ell \bmod (2n+1) = 0$
	E_m	$m_\ell \bmod (2n+1) = m$, $0 < m \leq n$
C_{2nh}, D_{2nh}	A_{*g}	$\ell \bmod 2 = 0$, $m_\ell \bmod 2n = 0$
	B_{*g}	$\ell \bmod 2 = 0$, $m_\ell \bmod 2n = n$
	E_{mg}	$\ell \bmod 2 = 0$, $m_\ell \bmod 2n = m$, $0 < m < n$
	A_{*u}	$\ell \bmod 2 = 1$, $m_\ell \bmod 2n = 0$
	B_{*u}	$\ell \bmod 2 = 1$, $m_\ell \bmod 2n = n$
	E_{mu}	$\ell \bmod 2 = 1$, $m_\ell \bmod 2n = m$, $0 < m < n$
$C_{(2n+1)h}$, $D_{(2n+1)h}$	A'_*	$\ell - m_\ell \bmod 2 = 0$, $m_\ell \bmod (2n+1) = 0$
	E'_m	$\ell - m_\ell \bmod 2 = 0$, $m_\ell \bmod (2n+1) = m$, $0 < m < n$
	A''_*	$\ell - m_\ell \bmod 2 = 1$, $m_\ell \bmod (2n+1) = 0$
	E''_m	$\ell - m_\ell \bmod 2 = 1$, $m_\ell \bmod (2n+1) = m$, $0 < m < n$
D_{2nd}	A_1, A_2	$\ell \bmod 2 = 0$, $m_\ell \bmod 2n = 0$
		$\ell \bmod 2 = 1$, $m_\ell \bmod 2n = 2n$
	B_1, B_2	$\ell \bmod 2 = 0$, $m_\ell \bmod 2n = 2n$
		$\ell \bmod 2 = 1$, $m_\ell \bmod 2n = 0$
	E_m	$m_\ell \bmod 2n = m$, $0 < m < 2n$
$D_{(2n+1)d}$	A_{1g}, A_{2g}	$\ell \bmod 2 = 0$, $m_\ell \bmod (2n+1) = 0$
	E_{mg}	$\ell \bmod 2 = 0$, $m_\ell \bmod (2n+1) = m$, $0 < m \leq n$
	A_{1u}, A_{2u}	$\ell \bmod 2 = 1$, $m_\ell \bmod (2n+1) = 0$
	E_{mu}	$\ell \bmod 2 = 1$, $m_\ell \bmod (2n+1) = m$, $0 < m \leq n$
S_{4n}	A	$\ell \bmod 2 = 0$, $m_\ell \bmod 4n = 0$
		$\ell \bmod 2 = 1$, $m_\ell \bmod 4n = 2n$
	B	$\ell \bmod 2 = 0$, $m_\ell \bmod 4n = 2n$
		$\ell \bmod 2 = 1$, $m_\ell \bmod 4n = 0$
	E_m	$m_\ell \bmod 4n = m$, $0 < m < 2n$
S_{4n+2}	A_g	$\ell \bmod 2 = 0$, $m_\ell \bmod (2n+1) = 0$
	E_{mg}	$\ell \bmod 2 = 0$, $m_\ell \bmod (2n+1) = m$, $0 < m \leq n$
	A_u	$\ell \bmod 2 = 1$, $m_\ell \bmod (2n+1) = 0$
	E_{mu}	$\ell \bmod 2 = 1$, $m_\ell \bmod (2n+1) = m$, $0 < m \leq n$

The fermion irreps for the binary groups (groups that contain only twofold operations) are presented in table D2, and for the general case in table D3. Although most of the information in table D2 follows from the table for the general case, the notation for D_2 and D_{2h} differs from that for the higher D_n and D_{nh} groups. The basis functions can be labeled with an m_j value because they transform as angular momentum eigenfunctions; the basis functions given in table D3 are listed for positive m_j in order of increasing m_j. Basis functions for negative m_j can be obtained by time reversal. The basis functions presented here do not contain phase information, merely boson symmetry information.

Table D2 Fermion basis functions for the binary groups

Group	*Basis functions*
C_1	$\begin{pmatrix} a \\ a \end{pmatrix}$
C_i	$\begin{pmatrix} a_g \\ a_g \end{pmatrix}, \begin{pmatrix} a_u \\ a_u \end{pmatrix}$
C_s	$\begin{pmatrix} a' \\ a'' \end{pmatrix}$
C_2	$\begin{pmatrix} a \\ b \end{pmatrix}$
C_{2h}	$\begin{pmatrix} a_g \\ b_g \end{pmatrix}, \begin{pmatrix} a_u \\ b_u \end{pmatrix}$
C_{2v}	$\begin{pmatrix} a_1 + ia_2 \\ b_1 + ib_2 \end{pmatrix}$
D_2	$\begin{pmatrix} a + ib_1 \\ b_2 + ib_3 \end{pmatrix}$
D_{2h}	$\begin{pmatrix} a_g + ib_{1g} \\ b_{2g} + ib_{3g} \end{pmatrix}, \begin{pmatrix} a_u + ib_{1u} \\ b_{2u} + ib_{3u} \end{pmatrix}$

Table D3 Basis functions for fermion irreps in terms of boson irreps for groups that have only one axis of order greater than 2

Group	*Basis functions*
C_{2n}	$\begin{pmatrix} a \\ e_1 \end{pmatrix}, \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \ldots, \begin{pmatrix} e_{n-1} \\ b \end{pmatrix}$
C_{2n+1}	$\begin{pmatrix} a \\ e_1 \end{pmatrix}, \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \ldots, \begin{pmatrix} e_n \\ e_{-n} \end{pmatrix}$
C_{2nv}, D_{2n}	$\begin{pmatrix} a_1 + ia_2 \\ e_1 \end{pmatrix}, \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \ldots, \begin{pmatrix} e_{n-1} \\ b_1 + ib_2 \end{pmatrix}$
$C_{(2n+1)v}$, D_{2n+1}	$\begin{pmatrix} a_1 + ia_2 \\ e_1 \end{pmatrix}, \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \ldots, \begin{pmatrix} e_n \\ e_{-n} \end{pmatrix}$
C_{2nh}	$C_{2n} \otimes C_i$; add g and u subscripts to C_{2n}
$C_{(2n+1)h}$	$C_{2n+1} \otimes C_i$; add $'$ to α spin and $''$ to β spin in C_{2n+1}
D_{2nd}	$\begin{pmatrix} a_1 + ia_2 \\ e_1 \end{pmatrix}, \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \ldots, \begin{pmatrix} e_{2n-1} \\ b_1 + ib_2 \end{pmatrix}$
$D_{(2n+1)d}$	$D_{2n+1} \otimes C_i$; add g and u subscripts to D_{2n+1}
D_{2nh}	$D_{2n} \otimes C_i$; add g and u subscripts to D_{2n}

continued

Table D3 (Continued)

Group	*Basis functions*
$D_{(2n+1)h}$	$D_{2n+1} \otimes C_s$; add $'$ and $''$ to opposite rows in both possible ways, i.e. $\begin{pmatrix} a_1' + ia_2' \\ e_1'' \end{pmatrix}, \begin{pmatrix} a_1'' + ia_2'' \\ e_1' \end{pmatrix}, \begin{pmatrix} e_1'' \\ e_2' \end{pmatrix}, \begin{pmatrix} e_1' \\ e_2'' \end{pmatrix}, \ldots$
S_{4n}	$\begin{pmatrix} a \\ e_1 \end{pmatrix}, \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \ldots, \begin{pmatrix} e_{2n-1} \\ b \end{pmatrix}$
S_{4n+2}	$C_{2n+1} \otimes C_i$; add g and u subscripts to C_{2n+1}

For the groups with only one axis of order greater than 2, it is straightforward to derive the multiplication tables in terms of angular momentum eigenfunctions. Apart from the usual rules regarding g and u labels and single and double primes, the irrep of the product is determined by the sum of the magnetic quantum numbers for the basis functions. This applies equally to the boson and the fermion irreps. Just as for the basis functions, the resulting function for the nondegenerate irreps must be determined by taking the appropriate linear combination of products. For example, in C_{2v}, the product of $e_{1/2}$ and $e_{1/2}$ has $m = 1$, which transforms as one of the B irreps, whereas the product of $e_{1/2}$ and $e_{-1/2}$ has $m = 0$, which transforms as one of the A irreps.

Appendix E

Change of Metric for Modified Wave Functions

In this appendix, we describe the change of metric that occurs when a 4-spinor wave function is modified by operations on the small component. The unmodified spinor is expressed as

$$\psi = \begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix}. \tag{E1}$$

The metric operator $\hat{\mathcal{G}}$ is the operator that appears in the orthogonality condition,

$$\int \psi_p{}^\dagger \hat{\mathcal{G}} \psi_q \mathrm{d}\mathbf{r} = \delta_{pq}. \tag{E2}$$

Without modifications to the wave function, the metric is simply the unit matrix,

$$\hat{\mathcal{G}} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{I}_2 \end{pmatrix} \tag{E3}$$

and the orthogonality condition reduces to the usual expression,

$$\int \psi_p{}^\dagger \psi_q \mathrm{d}\mathbf{r} = \delta_{pq}. \tag{E4}$$

In the general case, the metric operator can be written as

$$\hat{\mathcal{G}} = \begin{pmatrix} \hat{\mathcal{G}}^{LL} & \hat{\mathcal{G}}^{LS} \\ \hat{\mathcal{G}}^{SL} & \hat{\mathcal{G}}^{SS} \end{pmatrix}. \tag{E5}$$

For a number of approaches, we want to modify the small component such that $\psi^S \rightarrow \phi$ where

$$\psi^S = \hat{a}\phi \tag{E6}$$

and $\hat{a}$ is an operator expression. In this modified representation we can establish a new spinor as

$$\psi' = \begin{pmatrix} \psi^L \\ \phi \end{pmatrix}. \tag{E7}$$

But in order to maintain the normalization for this new spinor, we have to change the metric, which we can derive by substituting for the small component in the normalization expression,

$$\begin{aligned} \psi^\dagger \hat{\mathcal{G}} \psi &= \begin{pmatrix} \psi^{L\dagger} & \psi^{S\dagger} \end{pmatrix} \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{I}_2 \end{pmatrix} \begin{pmatrix} \psi^L \\ \psi^S \end{pmatrix} \\ &= \begin{pmatrix} \psi^{L\dagger} & \phi^\dagger \hat{a}^\dagger \end{pmatrix} \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{I}_2 \end{pmatrix} \begin{pmatrix} \psi^L \\ \hat{a}\phi \end{pmatrix} \\ &= \begin{pmatrix} \psi^{L\dagger} & \phi^\dagger \end{pmatrix} \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \hat{a}^\dagger \hat{a} \end{pmatrix} \begin{pmatrix} \psi^L \\ \phi \end{pmatrix} = \psi'^\dagger \hat{\mathcal{G}}' \psi'. \end{aligned} \tag{E8}$$

The modified metric $\hat{\mathcal{G}}'$ is no longer a unit matrix, but has an operator for the $\hat{\mathcal{G}}^{SS}$ element,

$$\hat{\mathcal{G}}^{SS\prime} = \hat{a}^\dagger \hat{a} \tag{E9}$$

and we write the modified metric as

$$\hat{\mathcal{G}}' = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \hat{a}^\dagger \hat{a} \end{pmatrix}. \tag{E10}$$

Using the case of (15.12) as an example, the modified spinor is

$$\tilde{\psi} = \begin{pmatrix} \psi^L \\ \phi^L \end{pmatrix} \tag{E11}$$

where ϕ^L is defined through (15.4)

$$2mc\psi^S \equiv (\boldsymbol{\sigma} \cdot \mathbf{p})\phi^L. \tag{E12}$$

For this case we have

$$\hat{a}^2 = (\boldsymbol{\sigma} \cdot \mathbf{p})(\boldsymbol{\sigma} \cdot \mathbf{p})/4m^2c^2 \tag{E13}$$

and therefore

$$\hat{\mathcal{G}}^{SS\prime} = \hat{T}/2mc^2\mathbf{I}_2 \tag{E14}$$

and

$$\hat{\mathcal{G}}' = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \hat{T}/2mc^2\mathbf{I}_2 \end{pmatrix}. \tag{E15}$$

Appendix F

Two-Electron Gauge Terms for the Modified Dirac Operator

In this appendix, we derive the two-electron Breit terms for the modified Dirac operator of chapter 15.

The gauge term, which comprises the difference between the Gaunt and the Breit interactions (see (5.48) and (5.49)), is more complicated than the Gaunt term due to the scalar quadruple product involving the alpha matrices:

$$g^{\text{gauge}} = \frac{(\boldsymbol{\alpha}_i \times \mathbf{r}_{ij}) \cdot (\boldsymbol{\alpha}_j \times \mathbf{r}_{ij})}{2r_{ij}^3}. \tag{F1}$$

It may be written in the same form as the Gaunt interaction in (15.53),

$$\tilde{g}^{\text{gauge}} = \frac{1}{4m^2c^2}\left[\hat{\mathcal{Q}}_i^+\hat{\mathcal{Q}}_j^+\,(g_{ij}^{++}) + \hat{\mathcal{Q}}_i^+\hat{\mathcal{Q}}_j^-\,(g_{ij}^{+-}) + \hat{\mathcal{Q}}_i^-\hat{\mathcal{Q}}_j^+\,(g_{ij}^{-+}) + \hat{\mathcal{Q}}_i^-\hat{\mathcal{Q}}_j^-\,(g_{ij}^{--})\right], \tag{F2}$$

but the four g operators have a different structure. These are

$$\begin{aligned} g_{ij}^{++} &= (\boldsymbol{\sigma}_i \cdot \mathbf{p}_i)(\boldsymbol{\sigma}_j \cdot \mathbf{p}_j)\frac{1}{2r_{ij}^3}(\boldsymbol{\sigma}_i \times \mathbf{r}_{ij}) \cdot (\boldsymbol{\sigma}_j \times \mathbf{r}_{ij}) \\ &= \frac{1}{2r_{ij}^3}\big[(\mathbf{p}_i \times \mathbf{r}_{ij}) \cdot (\mathbf{p}_j \times \mathbf{r}_{ij}) + i(\boldsymbol{\sigma}_i \times \mathbf{p}_i) \times \mathbf{r}_{ij} \cdot (\mathbf{p}_j \times \mathbf{r}_{ij}) \\ &\quad + i(\boldsymbol{\sigma}_j \times \mathbf{p}_j) \times \mathbf{r}_{ij} \cdot (\mathbf{p}_i \times \mathbf{r}_{ij}) - (\boldsymbol{\sigma}_i \times \mathbf{p}_i) \times \mathbf{r}_{ij} \cdot (\boldsymbol{\sigma}_j \times \mathbf{p}_j) \times \mathbf{r}_{ij}\big], \end{aligned} \tag{F3a}$$

$$\begin{aligned}
g_{ij}^{+-} &= (\boldsymbol{\sigma}_i \cdot \mathbf{p}_i)\frac{1}{2r_{ij}^3}(\boldsymbol{\sigma}_i \times \mathbf{r}_{ij}) \cdot (\boldsymbol{\sigma}_j \times \mathbf{r}_{ij})(\boldsymbol{\sigma}_j \cdot \mathbf{p}_j) \\
&= \frac{1}{2r_{ij}^3}\big[(\mathbf{p}_i \times \mathbf{r}_{ij}) \cdot (\mathbf{p}_j \times \mathbf{r}_{ij}) + i(\boldsymbol{\sigma}_i \times \mathbf{p}_i) \times \mathbf{r}_{ij} \cdot (\mathbf{p}_j \times \mathbf{r}_{ij}) \qquad \text{(F3b)} \\
&\quad - i(\boldsymbol{\sigma}_j \times \mathbf{p}_j) \times \mathbf{r}_{ij} \cdot (\mathbf{p}_i \times \mathbf{r}_{ij}) + (\boldsymbol{\sigma}_i \times \mathbf{p}_i) \times \mathbf{r}_{ij} \cdot (\boldsymbol{\sigma}_j \times \mathbf{p}_j) \times \mathbf{r}_{ij}\big],
\end{aligned}$$

$$\begin{aligned}
g_{ij}^{-+} &= (\boldsymbol{\sigma}_j \cdot \mathbf{p}_j)\frac{1}{2r_{ij}^3}(\boldsymbol{\sigma}_i \times \mathbf{r}_{ij}) \cdot (\boldsymbol{\sigma}_j \times \mathbf{r}_{ij})(\boldsymbol{\sigma}_i \cdot \mathbf{p}_i) \\
&= \frac{1}{2r_{ij}^3}\big[(\mathbf{p}_i \times \mathbf{r}_{ij}) \cdot (\mathbf{p}_j \times \mathbf{r}_{ij}) - i(\boldsymbol{\sigma}_i \times \mathbf{p}_i) \times \mathbf{r}_{ij} \cdot (\mathbf{p}_j \times \mathbf{r}_{ij}) \qquad \text{(F3c)} \\
&\quad + i(\boldsymbol{\sigma}_j \times \mathbf{p}_j) \times \mathbf{r}_{ij} \cdot (\mathbf{p}_i \times \mathbf{r}_{ij}) + (\boldsymbol{\sigma}_i \times \mathbf{p}_i) \times \mathbf{r}_{ij} \cdot (\boldsymbol{\sigma}_j \times \mathbf{p}_j) \times \mathbf{r}_{ij}\big],
\end{aligned}$$

$$\begin{aligned}
g_{ij}^{--} &= \frac{1}{2r_{ij}^3}(\boldsymbol{\sigma}_i \times \mathbf{r}_{ij}) \cdot (\boldsymbol{\sigma}_j \times \mathbf{r}_{ij})(\boldsymbol{\sigma}_i \cdot \mathbf{p}_i)(\boldsymbol{\sigma}_j \cdot \mathbf{p}_j) \\
&= \frac{1}{2r_{ij}^3}\big[(\mathbf{p}_i \times \mathbf{r}_{ij}) \cdot (\mathbf{p}_j \times \mathbf{r}_{ij}) - i(\boldsymbol{\sigma}_i \times \mathbf{p}_i) \times \mathbf{r}_{ij} \cdot (\mathbf{p}_j \times \mathbf{r}_{ij}) \qquad \text{(F3d)} \\
&\quad - i(\boldsymbol{\sigma}_j \times \mathbf{p}_j) \times \mathbf{r}_{ij} \cdot (\mathbf{p}_i \times \mathbf{r}_{ij}) - (\boldsymbol{\sigma}_i \times \mathbf{p}_i) \times \mathbf{r}_{ij} \cdot (\boldsymbol{\sigma}_j \times \mathbf{p}_j) \times \mathbf{r}_{ij}\big].
\end{aligned}$$

It has been assumed in deriving these expressions that the momentum operators are always operating on the basis functions, and not on any of the operators involving r_{ij}, with which they will therefore commute. This is because the reduction using (15.2) requires the rearrangement of the order of the scalar and vector products in the intermediate stages. Nevertheless, since $\mathbf{p}_i \times \mathbf{r}_{ij}$ commutes with any scalar function of r_{ij}, the first term in each of the above expressions is in general correct even if $\mathbf{p}$ is not taken to operate on a basis function.

The first term in each of the above expressions admits a straightforward interpretation. The replacements $\boldsymbol{\ell}_{ij} = \mathbf{r}_{ij} \times \mathbf{p}_i$ and $\boldsymbol{\ell}_{ji} = -\mathbf{r}_{ij} \times \mathbf{p}_j$ give the scalar product $-\boldsymbol{\ell}_{ij} \cdot \boldsymbol{\ell}_{ji}$, which is an orbit–orbit interaction: the interaction of the orbital angular momentum of electron i around electron j with the orbital angular momentum of electron j around electron i. While it would seem reasonable that these two should be the same, the two operators $\boldsymbol{\ell}_{ij}$ and $\boldsymbol{\ell}_{ji}$ are not equivalent. The second and third terms also involve an orbital angular momentum, but these are interacting with a more complicated entity involving the spin of the other electron. The fourth term involves the spin of both electrons, and thus belongs to the spin–spin category.

Each of the operators in the gauge term consists of a scalar quadruple product, which may be rearranged as was the spin–spin term in the Gaunt interaction.

This rearrangement is equivalent to the rearrangement of the Breit interaction:

$$\begin{aligned} g^{\text{Breit}} &= -\frac{\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j}{r_{ij}} + \frac{(\boldsymbol{\alpha}_i \times \mathbf{r}_{ij}) \cdot (\boldsymbol{\alpha}_j \times \mathbf{r}_{ij})}{2r_{ij}^3} \\ &= -\frac{1}{2r_{ij}}\left[\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j + \frac{(\boldsymbol{\alpha}_i \cdot \mathbf{r}_{ij})(\boldsymbol{\alpha}_j \cdot \mathbf{r}_{ij})}{r_{ij}^2}\right] \end{aligned} \tag{F4}$$

and therefore the rearranged gauge terms will cancel half the value of terms in the Gaunt interaction.

Appendix G

The Second-Order Term of the Douglas–Kroll Expansion

In this appendix, we examine the content of the second-order term of the Douglas–Kroll expansion given by (16.61). For this purpose, we expand the products of the operators involving $\hat{\mathcal{W}}_1$. It is only necessary for this purpose to consider $\hat{\mathcal{W}}_1^2$, which expands to

$$\begin{aligned}\hat{\mathcal{W}}_1^2 = {} & \hat{\mathcal{A}}_i \hat{\mathcal{D}}_i c\boldsymbol{\alpha}_i \cdot \mathbf{p}_i \tilde{V}_{ij} \hat{\mathcal{A}}_j^2 \hat{\mathcal{D}}_j c\boldsymbol{\alpha}_j \cdot \mathbf{p}_j \tilde{V}_{jk} \hat{\mathcal{A}}_k \\ & - \hat{\mathcal{A}}_i \hat{\mathcal{D}}_i c\boldsymbol{\alpha}_i \cdot \mathbf{p}_i \tilde{V}_{ij} \hat{\mathcal{A}}_j^2 \tilde{V}_{jk} c\boldsymbol{\alpha}_k \cdot \mathbf{p}_k \hat{\mathcal{D}}_k \hat{\mathcal{A}}_k \\ & - \hat{\mathcal{A}}_i \tilde{V}_{ij} c\boldsymbol{\alpha}_j \cdot \mathbf{p}_j \hat{\mathcal{D}}_j \hat{\mathcal{A}}_j^2 \hat{\mathcal{D}}_j c\boldsymbol{\alpha}_j \cdot \mathbf{p}_j \tilde{V}_{jk} \hat{\mathcal{A}}_k \\ & + \hat{\mathcal{A}}_i \tilde{V}_{ij} c\boldsymbol{\alpha}_j \cdot \mathbf{p}_j \hat{\mathcal{D}}_j \hat{\mathcal{A}}_j^2 \tilde{V}_{jk} c\boldsymbol{\alpha}_k \cdot \mathbf{p}_k \hat{\mathcal{D}}_k \hat{\mathcal{A}}_k\end{aligned} \tag{G1}$$

where the modified potential is

$$\tilde{V}_{ij} = \frac{V(\mathbf{p}_i, \mathbf{p}_j)}{E_i + E_j}. \tag{G2}$$

In order to represent all potentials in terms of the two operators V and $\boldsymbol{\alpha} \cdot \mathbf{p} V \boldsymbol{\alpha} \cdot \mathbf{p}$, we insert $(c\boldsymbol{\alpha} \cdot \mathbf{p})^2/c^2 p^2$ into the second term in the expansion of $\hat{\mathcal{W}}_1^2$ and rearrange to get

$$\begin{aligned}\hat{\mathcal{W}}_1^2 = {} & \hat{\mathcal{A}}_i \hat{\mathcal{D}}_i c\boldsymbol{\alpha}_i \cdot \mathbf{p}_i \tilde{V}_{ij} c\boldsymbol{\alpha}_j \cdot \mathbf{p}_j \hat{\mathcal{A}}_j^2 \hat{\mathcal{D}}_j \tilde{V}_{jk} \hat{\mathcal{A}}_k \\ & - \hat{\mathcal{A}}_i \hat{\mathcal{D}}_i c\boldsymbol{\alpha}_i \cdot \mathbf{p}_i \tilde{V}_{ij} c\boldsymbol{\alpha}_j \cdot \mathbf{p}_j (\hat{\mathcal{A}}_j^2/p_j^2) c\boldsymbol{\alpha}_j \cdot \mathbf{p}_j \tilde{V}_{jk} c\boldsymbol{\alpha}_k \cdot \mathbf{p}_k \hat{\mathcal{D}}_k \hat{\mathcal{A}}_k \\ & - \hat{\mathcal{A}}_i \tilde{V}_{ij} \hat{\mathcal{D}}_j \hat{\mathcal{A}}_j^2 c^2 p_j^2 \hat{\mathcal{D}}_j \tilde{V}_{jk} \hat{\mathcal{A}}_k \\ & + \hat{\mathcal{A}}_i \tilde{V}_{ij} \hat{\mathcal{D}}_j \hat{\mathcal{A}}_j^2 c\boldsymbol{\alpha}_j \cdot \mathbf{p}_j \tilde{V}_{jk} c\boldsymbol{\alpha}_k \cdot \mathbf{p}_k \hat{\mathcal{D}}_k \hat{\mathcal{A}}_k.\end{aligned} \tag{G3}$$

Finally, we can factorize this expression and write it as

$$\hat{\mathcal{W}}_1^2 = -\hat{\mathcal{Y}}_{ij} c^2 p_j^2 \hat{\mathcal{Y}}_{kj}{}^{\dagger} \tag{G4}$$

where the operator $\hat{\mathcal{Y}}$ is defined by

$$\hat{\mathcal{Y}}_{ij} = \hat{\mathcal{A}}_i \left(\tilde{V}_{ij} \hat{\mathcal{D}}_j - \hat{\mathcal{D}}_i \boldsymbol{\alpha}_i \cdot \mathbf{p}_i \tilde{V}_{ij} \boldsymbol{\alpha}_j \cdot \mathbf{p}_j \frac{1}{p_j^2} \right) \hat{\mathcal{A}}_j . \tag{G5}$$

Pre- and postmultiplying by $\hat{E}_p$ is trivial, as is insertion of $\hat{E}_p$ between the two $\hat{\mathcal{W}}_1$ operators. The factorized form of the operator shows that it can be written as the negative of the product of an operator and its Hermitian conjugate, which must be negative. The factors of $\hat{E}_p$ do not change this conclusion because $\hat{E}_p$ is positive. Since all three contributions from $\hat{\mathcal{W}}_1^2$ appear with a negative sign in (16.61), the overall contribution to the energy is positive, redressing the too-negative energies from the first-order operator which came from the free-particle Foldy–Wouthuysen transformation.

As for the first-order term, the second-order term may be split into spin-free and spin-dependent operators. The spin-free version of $\hat{\mathcal{W}}_1^2$ is

$$\begin{aligned}
(\hat{\mathcal{W}}_1^2)^{sf} = & - \hat{\mathcal{A}}_i \left(\tilde{V}_{ij} \hat{\mathcal{D}}_j - \hat{\mathcal{D}}_i \mathbf{p}_i \tilde{V}_{ij} \cdot \mathbf{p}_j \frac{1}{p_j^2} \right) \hat{\mathcal{A}}_j^2 c^2 p_j^2 \left(\hat{\mathcal{D}}_j \tilde{V}_{jk} - \frac{1}{p_j^2} \mathbf{p}_j \tilde{V}_{jk} \cdot \mathbf{p}_k \hat{\mathcal{D}}_k \right) \hat{\mathcal{A}}_k \\
& + \hat{\mathcal{A}}_i \hat{\mathcal{D}}_i \left(\mathbf{p}_i \tilde{V}_{ij} \times \mathbf{p}_j \right) \frac{c^2 \hat{\mathcal{A}}_j^2}{p_j^2} \cdot \left(\mathbf{p}_j \tilde{V}_{jk} \times \mathbf{p}_k \right) \hat{\mathcal{D}}_k \hat{\mathcal{A}}_k .
\end{aligned} \tag{G6}$$

The last term is sometimes neglected.

Appendix H

Transformed Operators for Electric and Magnetic Properties

In this appendix, we provide the tools for deriving transformed property operators in which the momentum appears in complicated functions, and separating the terms of these operators (Dyall 2000b). The reason that we cannot simply differentiate the functions with respect to the property operators is that the property operators do not commute with the momentum operators, and therefore the commutation must be considered explicitly.

For this purpose, we consider the functions of momentum to be defined by their power series expansion. The expansion is not convergent for $p > mc$, but provided we do not truncate the expansion we may make transformations on the series outside its radius of convergence as a representation of the operator and re-sum the series in the final result.

The scalar functions of momentum that appear in the transformation operators are all functions of $\mathbf{p}^2/m^2c^2$ which we will denote x. The terms involving $(\boldsymbol{\sigma} \cdot \mathbf{p})$ must be considered separately. We may write the commutator of an operator W with powers of x as

$$\begin{aligned}\left[x^n, W\right] &= n\left[x, W\right]x^{n-1} + \binom{n}{2}\left[x, \left[x, W\right]\right]x^{n-2} + \cdots \\ &= \sum_{k=1}^{n} \binom{n}{k} \left[x, W\right]^{(k)} x^{n-k} \\ &= \sum_{k=1}^{n} \left[x, W\right]^{(k)} \frac{1}{k!}\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^{k} x^n\end{aligned} \tag{H1}$$

where we have defined the kth order multiple commutator

$$\left[x, W\right]^{(k)} = \left[x, \left[x, \ldots \left[x, W\right]\right]\right] \tag{H2}$$

with k occurrences of x. The commutator of any function of x can therefore be written

$$\big[f(x), W\big] = \sum_{k=1}^{\infty} \big[x, W\big]^{(k)} \frac{1}{k!} \left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^{k} f(x). \tag{H3}$$

The expansion of the commutator becomes essentially a power series in x, which means that it is an expansion whose terms can be defined by a leading power of $1/m^2c^2$. The nature of the multiple commutator and the derivative of $f(x)$ will determine whether the series is convergent or not. For external electric fields, the commutator series truncates and an exact expression may be obtained.

The magnetic terms for the free-particle Foldy–Wouthuysen transformation are more complicated because the operator is not linear in the momentum. To do so, we start with the field-dependent operator corresponding to $\hat{e}_p$, which we will call $\hat{\epsilon}_p$,

$$\hat{\epsilon}_p = \sqrt{\mathbf{I} + (\boldsymbol{\sigma} \cdot (\mathbf{p} + e\mathbf{A}))^2/m^2c^2}. \tag{H4}$$

We may substitute $\hat{\epsilon}_p$ for $\hat{e}_p$ in the kinematic factors in both the Douglas–Kroll and the Barysz–Sadlej–Snijders approximations. To separate out the zeroth-order Hamiltonian from the perturbation, we follow an analogous procedure to that used for the scalar potential. We will concern ourselves only with the terms linear in the vector potential, but extension to other terms is straightforward, if tedious.

We may write any even power of $\boldsymbol{\sigma} \cdot (\mathbf{p} + e\mathbf{A})$ in terms of the multiple commutators defined above as

$$\begin{aligned}
(\boldsymbol{\sigma} \cdot (\mathbf{p} + e\mathbf{A}))^{2n} &= \mathbf{p}^{2n} + e\sum_{k=1}^{n} \mathbf{p}^{2k-2}(2\mathbf{A}\cdot\mathbf{p} + \hbar\boldsymbol{\sigma}\cdot\mathbf{B})\mathbf{p}^{2n-2k} + \cdots \\
&= \mathbf{p}^{2n} + e\sum_{k=1}^{n} \binom{n}{k} \big[\mathbf{p}^2, (2\mathbf{A}\cdot\mathbf{p} + \hbar\boldsymbol{\sigma}\cdot\mathbf{B})\big]^{(k-1)} \mathbf{p}^{2n-2k} + \cdots \\
&= \mathbf{p}^{2n} + e\sum_{k=1}^{n} \big[\mathbf{p}^2, (2\mathbf{A}\cdot\mathbf{p} + \hbar\boldsymbol{\sigma}\cdot\mathbf{B})\big]^{(k-1)} \frac{1}{k!} \left(\frac{\mathrm{d}}{\mathrm{d}\mathbf{p}^2}\right)^{k} \mathbf{p}^{2n} + \cdots,
\end{aligned} \tag{H5}$$

where the zeroth-order commutator is defined to be $[x, y]^{(0)} = y$. The powers of $(\boldsymbol{\sigma} \cdot (\mathbf{p} + e\mathbf{A}))^2$ must be divided by m^2c^2, and writing $y = (\boldsymbol{\sigma} \cdot (\mathbf{p} + e\mathbf{A}))^2/m^2c^2$ we may deduce an expression for a general function that is very similar to the one for the electric perturbations,

$$f(y) = f(x) + \frac{1}{m^2c^2} \sum_{k=1}^{\infty} \big[x, (2\mathbf{A} \cdot \mathbf{p} + \hbar\boldsymbol{\sigma} \cdot \mathbf{B})\big]^{(k-1)} \frac{1}{k!} \left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^{k} f(x) + \cdots . \tag{H6}$$

The pure kinetic corrections come from $\hat{\epsilon}_p$, which we expand to $\mathcal{O}(c^{-4})$ to obtain

$$mc^2\hat{\epsilon}_p = mc^2\hat{e}_p + eY\frac{1}{4m\hat{e}_p} + \left[\nabla^2, Y\right]\frac{e\hbar^2}{8m^3c^2\hat{e}_p^3} + \left[\nabla^2, \left[\nabla^2, Y\right]\right]\frac{e\hbar^4}{16m^5c^4\hat{e}_p^5} + \cdots \tag{H7}$$

where $Y = 2\mathbf{A}\cdot\mathbf{p} + \boldsymbol{\sigma}\cdot\mathbf{B}$. For a constant magnetic field all the commutators vanish, leaving only the first term in Y. If we expand $\hat{e}_p$ in powers of $\mathbf{p}^2/m^2c^2$ we get the kinetic correction to the Pauli operator, given in (16.105). The expression given here is not symmetric but is easily symmetrized in the expansion by halving each term and commuting Y to the right and to the left, instead of only commuting Y to the left.

The cross terms between the magnetic field and the potential come from the field-dependent FPFW potential term,

$$\begin{aligned}
&\hat{\mathcal{A}}^{\mathrm{Y}} V \hat{\mathcal{A}}^{\mathrm{Y}} + \frac{1}{m^2c^2}\hat{\mathcal{Q}}^{\mathrm{Y}}\Big(\boldsymbol{\sigma}\cdot(\mathbf{p}+e\mathbf{A})V\boldsymbol{\sigma}\cdot(\mathbf{p}+e\mathbf{A})\Big)\hat{\mathcal{Q}}^{\mathrm{Y}}\\
&\quad = \hat{\mathcal{A}}^{\mathrm{Y}} V \hat{\mathcal{A}}^{\mathrm{Y}} + \frac{1}{m^2c^2}\hat{\mathcal{Q}}^{\mathrm{Y}}\Big(\boldsymbol{\sigma}\cdot\mathbf{p}V\boldsymbol{\sigma}\cdot\mathbf{p} + e^2\boldsymbol{\sigma}\cdot\mathbf{A}V\boldsymbol{\sigma}\cdot\mathbf{A} + \tfrac{1}{2}e(VY+YV)\\
&\qquad\qquad + e\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{A}\Big)\hat{\mathcal{Q}}^{\mathrm{Y}}
\end{aligned} \tag{H8}$$

where the superscripts indicate field-dependent operators. When expanding these operators, the field term Y is commuted to the left or the right, depending on where the primed operator is located, to give a symmetric expression. With similar manipulations of the terms of various orders, we arrive at an expression correct to $\mathcal{O}(c^{-4})$ for the cross terms,

$$\begin{aligned}
&\frac{e}{m^2c^2}\,\hat{\mathcal{Q}}\,\boldsymbol{\sigma}\cdot(\nabla V)\times\mathbf{A}\,\hat{\mathcal{Q}}\\
&\quad + \frac{e}{m^4c^4}\Bigg(Y\frac{1}{8\hat{e}_p^3}\,\hat{\mathcal{Q}}\left[V, \mathbf{p}^2\right]\hat{\mathcal{Q}} + \hat{\mathcal{Q}}\left[\mathbf{p}^2, V\right]\hat{\mathcal{Q}}\,\frac{1}{8\hat{e}_p^3}Y\\
&\qquad + Y\frac{2\hat{e}_p+1}{4\hat{e}_p^2(1+\hat{e}_p)}\,\hat{\mathcal{Q}}\,\boldsymbol{\sigma}\cdot\mathbf{p}\left[V, \boldsymbol{\sigma}\cdot\mathbf{p}\right]\hat{\mathcal{Q}}\\
&\qquad + \hat{\mathcal{Q}}\left[\boldsymbol{\sigma}\cdot\mathbf{p}, V\right]\boldsymbol{\sigma}\cdot\mathbf{p}\,\hat{\mathcal{Q}}\,\frac{2\hat{e}_p+1}{4\hat{e}_p^2(1+\hat{e}_p)}Y\\
&\qquad + \left[\mathbf{p}^2, Y\right]\frac{2\hat{e}_p^2+2\hat{e}_p+1}{16\hat{e}_p^3(1+\hat{e}_p)}\,\hat{\mathcal{Q}}V\hat{\mathcal{Q}}\\
&\qquad + \hat{\mathcal{Q}}V\hat{\mathcal{Q}}\,\frac{2\hat{e}_p^2+2\hat{e}_p+1}{16\hat{e}_p^3(1+\hat{e}_p)}\left[Y, \mathbf{p}^2\right]\Bigg).
\end{aligned} \tag{H9}$$

The scalar term of $\mathcal{O}(c^{-2})$ has vanished, leaving only a spin–orbit term, which agrees with the Pauli value in the limit $\hat{\mathcal{Q}} \to 2$.

It remains to consider the Douglas–Kroll and Barysz–Sadlej–Snijders transformations. Since these corrections have leading order c^{-4}, the magnetic terms of this order arise from the commutator $\left[V, \boldsymbol{\sigma} \cdot (\mathbf{p} + e\mathbf{A})\right] = \left[V, \boldsymbol{\sigma} \cdot \mathbf{p}\right]$, which is independent of the field. Hence, there is no contribution from either transformation to the field at $\mathcal{O}(c^{-4})$. Because this commutator is field-independent, all the corrections arise from the expansion of the kinematic factors, and are of leading order c^{-6}.

Appendix I

Gauge Term Contributions from the Breit Interaction to the Breit–Pauli Hamiltonian

In this appendix, we develop the two-electron terms for the Breit–Pauli Hamiltonian arising from the gauge term of the Breit interaction. We derive the expressions using the modified Dirac operators given in (F3). In the reduction of the spin-dependence in the derivation of (F3), no account was taken of the order of the momentum operators, assuming that they would all operate on basis functions. Now, however, we wish to write it in a form in which the momentum operators are all operating to the right. As for the contributions to the Gaunt interaction, we may split the contributions to the gauge term into three classes. The first class comprises the spin-free operators, for which we can readily insert the $1/2r_{ij}^3$ terms in the appropriate places, giving

$$\begin{aligned}&\frac{1}{2r_{ij}^3}(\mathbf{r}_{ij}\times\mathbf{p}_i)\cdot(\mathbf{r}_{ij}\times\mathbf{p}_j)-(\mathbf{p}_i\times\mathbf{r}_{ij})\frac{1}{2r_{ij}^3}\cdot(\mathbf{r}_{ij}\times\mathbf{p}_j)\\&\qquad-(\mathbf{p}_j\times\mathbf{r}_{ij})\frac{1}{2r_{ij}^3}\cdot(\mathbf{r}_{ij}\times\mathbf{p}_i)+(\mathbf{p}_i\times\mathbf{r}_{ij})\cdot(\mathbf{p}_j\times\mathbf{r}_{ij})\frac{1}{2r_{ij}^3}.\end{aligned}\tag{I1}$$

The first term needs no further consideration because $\mathbf{p}_i$ here is taken not to operate on $(\mathbf{r}_{ij}\times\mathbf{p}_j)$. In the second term, we rearrange the scalar quadruple product and differentiate (omitting the factor of $\hbar$ from $\mathbf{p}$) to give

$$\begin{aligned}(\mathbf{p}_i\times\mathbf{r}_{ij})\frac{1}{2r_{ij}^3}\cdot(\mathbf{r}_{ij}\times\mathbf{p}_j)&=\left(\mathbf{p}_i\cdot\frac{\mathbf{r}_{ij}}{2r_{ij}^3}\right)(\mathbf{r}_{ij}\cdot\mathbf{p}_j)-\left(\mathbf{p}_i\frac{1}{2r_{ij}}\cdot\mathbf{p}_j\right)\\&=-2\pi i\delta(\mathbf{r}_{ij})(\mathbf{r}_{ij}\cdot\mathbf{p}_j)+\frac{i}{r_{ij}^3}(\mathbf{r}_{ij}\cdot\mathbf{p}_j)+\frac{1}{2r_{ij}^3}(\mathbf{r}_{ij}(\mathbf{r}_{ij}\cdot\mathbf{p}_j)\cdot\mathbf{p}_i)-\frac{1}{2r_{ij}}(\mathbf{p}_i\cdot\mathbf{p}_j)\\&=\frac{i}{r_{ij}^3}(\mathbf{r}_{ij}\cdot\mathbf{p}_j)-\frac{1}{2r_{ij}^3}(\mathbf{r}_{ij}\times\mathbf{p}_i)\cdot(\mathbf{r}_{ij}\times\mathbf{p}_j).\end{aligned}\tag{I2}$$

The term involving the delta function vanishes because it multiplies a term involving $\mathbf{r}_{ij}$. Similarly, for the third term, we get

$$(\mathbf{p}_j \times \mathbf{r}_{ij})\frac{1}{2r_{ij}^3} \cdot (\mathbf{r}_{ij} \times \mathbf{p}_i) = -\frac{i}{r_{ij}^3}(\mathbf{r}_{ij} \cdot \mathbf{p}_i) - \frac{1}{2r_{ij}^3}(\mathbf{r}_{ij} \times \mathbf{p}_j) \cdot (\mathbf{r}_{ij} \times \mathbf{p}_i). \tag{I3}$$

For the fourth term, we have

$$(\mathbf{p}_i \times \mathbf{r}_{ij}) \cdot (\mathbf{p}_j \times \mathbf{r}_{ij})\frac{1}{2r_{ij}^3} = -(\nabla_i \cdot \nabla_j)\frac{1}{2r_{ij}} + \left(\nabla_i \left(\nabla_j \cdot \frac{\mathbf{r}_{ij}}{2r_{ij}^3}\right) \cdot \mathbf{r}_{ij}\right). \tag{I4}$$

The first part of this expression yields the result

$$-(\nabla_i \cdot \nabla_j)\frac{1}{2r_{ij}} = -2\pi\delta(\mathbf{r}_{ij}) + \frac{i}{2r_{ij}^3}(\mathbf{r}_{ij} \cdot \mathbf{p}_i) - \frac{i}{2r_{ij}^3}(\mathbf{r}_{ij} \cdot \mathbf{p}_j) + \frac{1}{2r_{ij}}(\mathbf{p}_i \cdot \mathbf{p}_j). \tag{I5}$$

In the second part, we first apply ∇_j to give

$$\left(\nabla_i \left(\nabla_j \cdot \frac{\mathbf{r}_{ij}}{2r_{ij}^3}\right) \cdot \mathbf{r}_{ij}\right) = 2\pi\nabla_i \cdot \left[\delta(\mathbf{r}_{ij})\mathbf{r}_{ij}\right] + \nabla_i \cdot \frac{\mathbf{r}_{ij}}{2r_{ij}^3} + \left(\nabla_i \cdot \frac{\mathbf{r}_{ij}}{2r_{ij}^3}\right)(\mathbf{r}_{ij} \cdot \nabla_j). \tag{I6}$$

The first term is zero, and the second can be straightforwardly differentiated. In the third term, the differentiation of $\mathbf{r}_{ij}/2r_{ij}^3$ to give a delta function produces a zero overall result because of the presence of $\mathbf{r}_{ij}$. The second part is therefore

$$\left(\nabla_i \left(\nabla_j \cdot \frac{\mathbf{r}_{ij}}{2r_{ij}^3}\right) \cdot \mathbf{r}_{ij}\right) = -2\pi\delta(\mathbf{r}_{ij}) + \frac{i}{2r_{ij}^3}(\mathbf{r}_{ij} \cdot \mathbf{p}_i) - \frac{i}{2r_{ij}^3}(\mathbf{r}_{ij} \cdot \mathbf{p}_j) - \frac{1}{2r_{ij}^3}(\mathbf{r}_{ij}(\mathbf{r}_{ij} \cdot \mathbf{p}_j) \cdot \mathbf{p}_i). \tag{I7}$$

Adding the two parts together, we get

$$\begin{aligned}(\mathbf{p}_i \times \mathbf{r}_{ij}) \cdot (\mathbf{p}_j \times \mathbf{r}_{ij})\frac{1}{2r_{ij}^3} &= -4\pi\delta(\mathbf{r}_{ij}) + \frac{i}{r_{ij}^3}\left[(\mathbf{r}_{ij} \cdot \mathbf{p}_i) - (\mathbf{r}_{ij} \cdot \mathbf{p}_j)\right] \\ &\quad + \frac{1}{2r_{ij}}(\mathbf{r}_{ij} \times \mathbf{p}_j) \cdot (\mathbf{r}_{ij} \times \mathbf{p}_i). \end{aligned} \tag{I8}$$

Summing all the spin-free contributions gives

$$\begin{aligned}\hat{g}^{\text{gauge,sf}} &= \frac{1}{2}\sum_{i \neq j} \frac{1}{m^2c^2}\left[\frac{1}{2r_{ij}}(\mathbf{p}_i \cdot \mathbf{p}_j) - \frac{1}{2r_{ij}^3}(\mathbf{r}_{ij}(\mathbf{r}_{ij} \cdot \mathbf{p}_j) \cdot \mathbf{p}_i)\right. \\ &\quad \left. + \frac{i}{2r_{ij}^3}(\mathbf{r}_{ij} \cdot \mathbf{p}_i) - \frac{i}{2r_{ij}^3}(\mathbf{r}_{ij} \cdot \mathbf{p}_j) - \pi\delta(\mathbf{r}_{ij})\right]. \end{aligned} \tag{I9}$$

The reduction of the remaining contributions from the gauge term is extremely tedious, and in fact gives a zero contribution to the operator to $\mathcal{O}(c^{-2})$. Thus, there is no spin-dependent contribution from the gauge term for the relativistic correction to the electron–electron interaction to the Breit–Pauli Hamiltonian.

Appendix J

Approximations in Relativistic Density Functional Theory

Prior to the wholesale adoption of DFT by quantum chemists, a variety of approximations were used to implement DFT methods in relativistic and nonrelativistic programs. A large portion of the literature on relativistic effects from DFT calculations is due to these programs, most of which are still in use today. In this appendix, we briefly describe some of these approximations.

The first approximation to make an impact on quantum chemistry was from the "multiple scattering Xα" (MSXα) method developed by Slater (1960, 1972, 1974), Johnson (1973), and others in the early 1970s. To deal with the absence of spherical symmetry in the general molecular problem, the *muffin-tin* approximation was adopted from solid-state physics. In this approximation, each atom is surrounded by a spherical domain within which the potential is assumed to be spherical. All the atomic spheres (which are made to touch) are then surounded by an overall molecular sphere, and a constant potential is assumed in the intersphere region. Using standard numerical techniques (partly adapted from scattering theory, hence the MS), the single-particle equations could be solved within each of these regions, and the solutions then joined at the boundaries. The MSXα method has provided a number of interesting and valuable results, in particular for properties directly related to electronic structure and energy levels. Due to the muffin-tin approximation, the method was less reliable in predicting molecular structures, and typically worked best for highly symmetric systems. The extension of the MSXα method to the relativistic case was done by Case and Yang (Case and Yang 1980, 1982; see also Cartling and Whitmore 1976). Programs that are descendants of the relativistic MSXα code are still successfully applied to the properties of large, (preferably) symmetric molecules (Alvarez-Thon et al. 2001).

A different approach was developed by Baerends, Ellis, and Ros (1973). In addition to adopting the Slater potential for the exchange, their approach had two distinct features. The first was an efficient numerical integration procedure, the "discrete variational method" (DVM), which permitted the use of any type of basis function for expansion, not only Slater-type orbitals or Gaussian-type orbitals, but also numerical atomic orbitals. The second feature was an evaluation of the Coulomb potential from an analytic fit of one-center functions, as is done in many modern programs. Thus, the density is approximated by the expansion

$$\rho(\mathbf{r}) = \sum_{\lambda} a_{\lambda} f_{\lambda}(\mathbf{r}) \tag{J1}$$

where $\{f_{\lambda}\}$ is a set of suitable atom-centered functions. The method has been called "discrete variational Xα" (DV-Xα), and was extended to the relativistic Dirac–Slater method by Rosén and Ellis (Rosén 1997, Rosén and Ellis 1975). The introduction of 4-spinors into the method was facilitated by the rather general nature of the DVM approach. The relativistic version also employed symmetry-adapted spinors generated by double-group projection operators. The relativistic calculations done with this method have mainly dealt with electronic structure, and photoelectron and optical spectra. The approach has been used extensively by Fricke, Pershina, and coworkers (Fricke et al. 1997, Pershina 2004) for the study of compounds involving superheavy elements.

Relativistic corrections have also been included in DFT calculations using perturbation theory, first by Herman and Skillman (1963) and later by Snijders and Baerends (Snijders 1979, Snijders and Baerends 1977, 1978). Following a nonrelativistic Hartree–Fock–Slater calculation, first-order perturbation theory was used to calculate the relativistic corrections from the Breit–Pauli terms of $\mathcal{O}(c^{-2})$. Herman and Skillman applied this approach to first order only for the energies.

In the approach by Snijders and Baerends, the spinors ϕ_i^0 from the nonrelativistic equation satisfy the equation

$$\hat{f}_0 \phi_i^0 = \epsilon_i^0 \phi_i^0. \tag{J2}$$

The mass–velocity corrections to the orbital energy and the orbital are calculated from

$$\left(\hat{f}_0 - \epsilon_i^0\right)\phi_i^{mv} = \left(\langle \phi_i^0 | \hat{f}_{mv} | \phi_i^0 \rangle - \hat{f}_{mv}\right)\phi_i^0 \tag{J3}$$

where $\hat{f}_{mv}$ is the mass–velocity operator. Similar equations hold for the Darwin and spin–orbit terms. In addition, they also accounted for the change in potential due to the relativistic change of the density, given by the perturbation equation

$$\left(\hat{f}_0 - \epsilon_i^0\right)\phi_i^{\text{pot}} = \left(\langle \phi_i^0 | V^1(\rho') | \phi_i^0 \rangle - V^1(\rho')\right)\phi_i^0 \tag{J4}$$

with the first-order change of the density given by

$$\rho' = 2c^2 \sum_{i=1}^{N} n_i \phi_i^0 \phi_i^1. \tag{J5}$$

The total first-order correction to the orbital is then

$$\phi_i^1 = \phi_i^{mv} + \phi_i^D + \phi_i^{SO} + \phi_i^{\text{pot}}. \tag{J6}$$

As we see, this changes the potential $V^1(\rho')$ that determines ϕ_i^{pot}, and thus (J4) must be solved iteratively.

In actual applications of the method, Snijders and Baerends used a frozen core and incorporated core relativistic effects by using atomic all-electron relativistic orbitals. This is because the Pauli Hamiltonian is not bounded from below. The orthogonality requirement against the core prevents the orbitals from collapsing. With this approach they were able to reproduce quite accurately the valence shell orbital energies from fully relativistic all-electron Dirac–Slater calculations for atoms.

Appendix K

The Cowan–Griffin and Wood–Boring Equations

The Cowan–Griffin Hamiltonian was developed for spin-free relativistic atomic calculations (Cowan and Griffin 1976). However, it has also found some use as a starting point for developing spin-free relativistic Hamiltonians for molecular application. Here, we show the form of this operator, and the associated spin–orbit correction. For atoms, the Cowan–Griffin Hamiltonian follows directly from the radial form of the atomic Dirac equation (7.29), which may be given as

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}r^2} - \frac{\kappa(\kappa+1)}{r^2} + 2(E-V) + \frac{1}{c^2}(E-V)^2 \right.$$
$$\left. + \frac{1}{(2c^2+E-V)}\frac{\mathrm{d}V}{\mathrm{d}r}\left(\frac{\mathrm{d}}{\mathrm{d}r} - \frac{1}{r}\right) + \frac{1}{(2c^2+E-V)}\frac{\mathrm{d}V}{\mathrm{d}r}\left(\frac{\kappa+1}{r}\right)\right] P = 0. \qquad \text{(K1)}$$

The first three terms here are easily recognized as the nonrelativistic Hamiltonian. Terms four and five are related to the mass–velocity and Darwin operators, respectively. For the mass–velocity term, this is easily seen by making the approximate substitution $E - V \approx \hat{T}$, and inserting the usual nonrelativistic expression for the kinetic energy operator. For the Darwin term, integration by parts yields the second derivative of the potential, which is the Pauli Darwin operator. The last term of the operator above contains the one-electron spin–orbit interaction, which we can cast in a more familiar form by replacing $\kappa + 1$ with the corresponding operator expression $-\boldsymbol{\sigma} \cdot \boldsymbol{\ell}$. This equation is in fact the Wood–Boring equation. By dropping the last term, we obtain the spin-free Cowan–Griffin equation. We can now apply this in numerical calculations for many-electron atoms using the Coulomb interaction between the electrons. Note that the E of the expression above must be then replaced by the orbital energies ϵ_n to be consistent. Also, the equation is for the large component, which is neither normalized nor orthogonal to the large components for other spinors or orbitals. Orthogonalization, which would be required in an SCF method for many-electron atoms, is therefore an approximation because the metric that corresponds to the Cowan–Griffin equation is not unity.

If we want to incorporate spin–orbit effects, either subsequent to a spin-free relativistic calculation using this operator or directly as a first step using a modified operator, we must include the previously discarded last term of the large component of the Dirac equation above. For atoms, this was first done by Wood and Boring (1978), who used the operator perturbatively.

Appendix L

Supplementary Reading

Relativistic quantum chemistry is a subject that draws on many different fields of theory, such as special relativity, QED, and group theory, to mention only a few. For the practicing researcher, eager to apply the theory to the needs at hand, the prospects of sifting through volumes of information in order to come up with a workable platform of knowledge are not appealing. It is to help with this task that the present book has been written. However, in this process we have had to make a choice of what to include, and we realize that there are certainly those among our readers who wish for more or different information. We expect that these wishes go in two main directions—toward the pedagogical for fuller explanations of background and derivations, and toward the advanced treatment with more sophisticated physics and leading-edge applications.

For this reason, we provide here a list of supplementary reading. The list falls into three parts: textbooks, review volumes, and bibliographies. Some of the books referred to are out of print, or at least difficult to obtain, but a good science library should be able to provide most of these.

Textbooks

Special Relativity

Most introductory physics textbooks contain chapters on special relativity. The following two books are exclusively devoted to the subject.

French A. P. 1968. *Special Relativity.* W. W. Norton: New York.
Rindler W. 1987. *Introduction to Special Relativity.* Oxford University Press: New York.

Relativistic Quantum Mechanics

There are a large number of physics textbooks that deal with relativistic quantum mechanics. The first four books on the list below are general texts that include chapters on relativistic quantum mechanics, the others are exclusively devoted to the subject.

Merzbacher E. 1970. *Quantum Mechanics*, 2nd edn. John Wiley & Sons: New York.
Messiah A. 1961. *Quantum mechanics. Vol. II.* North-Holland: Amsterdam.
Sakurai J. J. 1967. *Advanced Quantum Mechanics.* Addison-Wesley: London.
Schiff L. I. 1968. *Quantum Mechanics*, 3rd edn. McGraw-Hill Kogakusha: New York.
Bjorken J. D. and Drell S. D. 1964. *Relativistic Quantum Mechanics.* McGraw-Hill: New York.
Das T. P. 1973. *Relativistic Quantum Mechanics of Electrons.* Harper & Row: New York.
Moss R. E. 1973. *Advanced Molecular Quantum Mechanics: An Introduction to Relativistic Quantum Mechanics and the Quantum Theory of Radiation.* Chapman and Hall: London.
Rose M. E. 1961. *Relativistic Electron Theory.* Wiley: New York.
Strange P. 1998. *Relativistic Quantum Mechanics: With Applications in Condensed Matter and Atomic Physics.* Cambridge University Press: Cambridge.
Thaller B. 1992. *The Dirac Equation.* Springer: Berlin.

QED

Again this is treated in many physics textbooks. The first two listed here are the classical references in the field, the third is oriented more towards molecular physics.

Berestetskii V. B., Lifshitz E. M., and Pitaevskii L. P. 1982. *Quantum Electrodynamics*, 2nd edn. Pergamon: Oxford.
Bjorken J. D. and Drell S. D. 1965. *Relativistic Quantum Fields.* McGraw-Hill: New York.
Craig D. P. and Tirunamachandran T. 1999. *Molecular Quantum Electrodynamics.* Dover: New York.

Double-group Symmetry

Treatments of double-group symmetry are usually hidden in general texts on group theory or relativistic physics. Some useful discussion may be found in the following texts.

Brink D. M. and Satchler G. R. 1993. *Angular Momentum*, 3rd edn. Oxford University Press: Oxford.
Jones H. F. 1998. *Groups, Representations and Physics.* Institute of Physics Publishing: Bristol.
Tinkham M. 1964. *Group Theory and Quantum Mechanics.* McGraw-Hill: New York.

Quaternions and Clifford Algebras

Adler S. L. 1995. *Quaternionic Quantum Mechanics and Quantum Fields.* Oxford University Press: Oxford.
Lounesto P. 2001. *Clifford Algebras and Spinors.* Cambridge University Press: Cambridge.

Review Volumes

Malli G. L. (ed.) 1983. *Relativistic Effects in Atoms, Molecules, and Solids.* Plenum: New York.

Grant I. P., Gyorffy B., and Wilson S. (eds.) 1990. *The Effects of Relativity in Atoms, Molecules and the Solid State.* Plenum: New York.

Malli G. L. (ed.) 1994. *Relativistic and Correlation Effects in Molecules and Solids.* Plenum: New York.

Balasubramanian K. 1997. *Relativistic Effects in Chemistry. Part A: Theory and Techniques.* Wiley: New York.

Balasubramanian K. 1997. *Relativistic Effects in Chemistry. Part B: Applications.* Wiley: New York.

Hess B. A. (ed.) 2002. *Relativistic Effects in Heavy Element Chemistry and Physics.* John Wiley & Sons: New York.

Schwerdtfeger P. (ed.) 2002. *Relativistic Electronic Structure Theory. Part 1: Fundamentals.* Elsevier: Amsterdam.

Schwerdtfeger P. (ed.) 2004. *Relativistic Electronic Structure Theory. Part 2: Applications.* Elsevier: Amsterdam.

Kaldor U. and Wilson. S. (eds.) 2003. *Theoretical Chemistry and Physics of Heavy and Superheavy Elements.* Kluwer Academic: Dordrecht.

Hirao K. and Ishikawa Y. (eds.) 2004. *Recent Advances in Relativistic Molecular Theory.* World Scientific: Singapore.

Bibliographies

The most comprehensive bibliography is that prepared by Pyykkö. The online version can be found at http://www.csc.fi/rtam/. The published books, which include some classification and commentary, are listed below.

Pyykkö P. 1986. *Relativistic Theory of Atoms and Molecules* (Lect. Notes. Chem. vol. 41). Springer-Verlag: Berlin.

Pyykkö P. 1993. *Relativistic Theory of Atoms and Molecules II* (Lect. Notes. Chem. vol. 60). Springer-Verlag: Berlin.

Pyykkö P. 2000. *Relativistic Theory of Atoms and Molecules III* (Lect. Notes. Chem. vol. 76). Springer-Verlag: Berlin.

This page intentionally left blank

Bibliography

Aaberge T. 1977. *Helvetica Physica Acta*, **50**, 917.

ADF — Amsterdam Density Functional code. Written by E. J. Baerends, J. A. Autschbach, A. Berces, C. Bo, P. M. Boerrigter, L. Cavallo, D. P. Chong, L. Deng, R. M. Dickson, D. E. Ellis, L. Fan, T. H. Fischer, C. Fonseca Guerra, S. J. A. van Gisbergen, J. A. Groeneveld, O. V. Gritsenko, M. Groning, F. E. Harris, P. van den Hoek, H. Jacobsen, G. van Kessel, F. Kootstra, E. van Lenthe, D. A. McCormack, V. P. Osinga, S. Patchkovskii, P. H. T. Philipsen, D. Post, C. C. Pye, W. Ravenek, P. Ros, P. R. T. Schipper, G. Schreckenbach, J. G. Snijders, M. Sola, M. Swart, D. Swerhone, G. te Velde, P. Vernooijs, L. Versluis, O. Visser, E. van Wezenbeek, G. Wiesenekker, S. K. Wolff, T. K. Woo, and T. Ziegler, 2005.

Ågren H., Vahtras O., and Minaev B. 1996. *Adv. Quantum Chem.*, **27**, 71.

Almlöf J., Fægri, Jr. K., and Grelland H. H. 1985. *Chem. Phys. Lett.*, **114**, 53.

Altmann S. L. and Herzig P. 1994. *Point-Group Theory Tables*. Clarendon Press: Oxford.

Alvarez-Thon L., Hernandez-Acevedo L., and Arratia-Perez R. 2001. *J. Chem. Phys.*, **115**, 726.

Andrae D. 2002. In *Relativistic Electronic Structure Theory. Part 1. Fundamentals*, Schwerdtfeger P. (ed.). Elsevier: Amsterdam, p. 203.

Andzelm J., Radzio E., Seijo L., and Barandiarán Z. 1985. *J. Chem. Phys.*, **83**, 4565.

Atkins P. W. and Friedman R. S. 1997. *Molecular Quantum Mechanics*, 3rd edn. Oxford University Press: Oxford.

Aucar G. A., Jensen H. J. Aa., and Oddershede J. 1995. *Chem. Phys. Lett.*, **232**, 47.

Baerends E. J., Ellis D. E., and Ros P. 1973. *Chem. Phys.*, **2**, 41.

Balasubramanian K. 1997a. *Relativistic Effects in Chemistry. Part A: Theory and Techniques*. Wiley: New York.

Balasubramanian K. 1997b. *Relativistic Effects in Chemistry. Part B: Applications*. Wiley: New York.

Barysz M., Sadlej A. J., and Snijders J. G. 1997. *Int. J. Quantum Chem.*, **65**, 225.

Bauschlicher C. W. 2000. *J. Phys. Chem. A*, **104**, 2281.

Bauschlicher C. W., Langhoff S. R., and Partridge H. 1995. In *Modern Electronic Structure Theory*, Yarkony D. R. (ed.). World Scientific: Singapore, p. 1280.

BDF—Beijing Density Functional code. Liu W., Hong G., Dai D., Li L., and Dolg M. 1997. *Theor. Chem. Acc.*, **96**, 75.

Berestetskii V. B., Lifshitz E. M., and Pitaevskii L. P. 1982. *Quantum Electrodynamics*, 2nd edn. Pergamon: Oxford.

BERTHA, a relativistic ab initio electronic structure program. Written by H. M. Quiney, H. Skaane, I. P. Grant, and T. Scott. Algorithms and implementation are described in Grant I. P. and Quiney H. M. 2002. In *Relativistic Electronic Structure Theory. Part 1. Fundamentals*, Schwerdtfeger P. (ed.). Elsevier: Amsterdam, p. 107.

Bethe H. A. and Salpeter E. E. 1957. *Quantum Mechanics of One- and Two-Electron Atoms*. Springer-Verlag: Berlin.

Bjorken J. D. and Drell S. D. 1964. *Relativistic Quantum Mechanics*. McGraw-Hill: New York.

Blume M. and Watson R. E. 1962. *Proc. Roy. Soc. Ser. A*, **270**, 127.

Bonifacic V. and Huzinaga S. 1974. *J. Chem. Phys.*, **60**, 2779.

Bonifacic V. and Huzinaga S. 1975. *J. Chem. Phys.*, **62**, 1507.

Brink D. M. and Satchler G. R. 1993. *Angular Momentum*, 3rd edn. Oxford University Press: Oxford.

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

Buenker R. J., Chandra P., and Hess B. A. 1984. *Chem. Phys.*, **84**, 1.

Cartling B. G. and Whitmore D. M. 1976. *Int. J. Quantum Chem.*, **10**, 393.

Case D. A. 1982. *Ann. Rev. Phys. Chem.*, **33**, 151.

Case D. A. and Yang C. Y. 1980. *J. Chem. Phys.*, **72**, 3443.

Chaix P. and Iracene D. 1989. *J. Phys. B: At. Mol. Phys.*, **22**, 3791.

Chang Ch., Pélissier M., and Durand P. 1986. *Physica Scripta*, **34**, 394.

Christiansen P. A., Balasubramanian K., and Pitzer K. S. 1982. *J. Chem. Phys.*, **76**, 5087.

Christiansen P. A., Lee Y. S., and Pitzer K. S. 1979. *J. Chem. Phys.*, **71**, 4445.

Collins C. L., Dyall K. G., and Schaefer III H. F. 1995. *J. Chem. Phys.*, **102**, 2024.

Cowan R. D. and Griffin D. C. 1976. *J. Opt. Soc. Am.*, **66**, 1010.

Dacre P. D. 1970. *Chem. Phys. Lett.*, **7**, 47.

Dalgarno A. and Stewart A. L. 1956. *Proc. Roy. Soc.*, **238**, 269, 276.

Darwin C. G. 1928. *Proc. Roy. Soc. (London) A*, **118**, 654.

Davidson E. R. 1975a. *J. Comput. Phys.*, **17**, 87.

Davidson E. R. 1975b. *J. Chem. Phys.*, **62**, 400.

DIRAC, a relativistic ab initio electronic structure program. Release DIRAC04.0 (2004), written by H. J. Aa. Jensen, T. Saue, and L. Visscher with contributions from V. Bakken, E. Eliav, T. Enevoldsen, T. Fleig, O. Fossgaard, T. Helgaker, J. Laerdahl, C. V. Larsen, P. Norman, J. Olsen, M. Pernpointner, J. K. Pedersen, K. Ruud, P. Salek, J. N. P. van Stralen, J. Thyssen, O. Visser, and T. Winther.

Dirac P. A. M. 1929. *Proc. Roy. Soc. (London) A*, **123**, 714.

Dirac P. A. M. 1949. *Rev. Mod. Phys.*, **21**, 392.

Dolg M. 2002. In *Relativistic Electronic Structure Theory. Part 1. Fundamentals*, Schwerdtfeger P. (ed.). Elsevier: Amsterdam, p. 793.

Douglas M. and Kroll N. M. 1974. *Ann. Phys. (N.Y.)*, **82**, 89.

Drake G. W. F. and Swainson R. A. 1990. *Phys. Rev. A*, **41**, 1243.

Dupuis M. and King H. F. 1977. *Int. J. Quantum Chem.*, **11**, 613.

Durand P. and Barthelat J. C. 1975. *Theor. Chim. Acta*, **38**, 283.

Dyall K. G. 1992. *J. Chem. Phys.*, **96**, 1210.

Dyall K. G. 1993. *J. Chem. Phys.*, **98**, 2191.

Dyall K. G. 1994. *J. Chem. Phys.*, **100**, 2118.

Dyall K. G. 1997. *J. Chem. Phys.*, **106**, 9618.

Dyall K. G. 2000a. *J. Phys. Chem. A*, **104**, 4077

Dyall K. G. 2000b. *Int. J. Quantum Chem.*, **78**, 412.

Dyall K. G. 2001. *J. Chem. Phys.*, **115**, 9136.

Dyall K. G. and Enevoldsen T. 1999. *J. Chem. Phys.*, **111**, 10000.

Dyall K. G., Grant I. P., and Wilson S. 1984. *J. Phys. B*: *At. Mol. Phys.*, **17**, 493.

Dyall K. G. and van Lenthe E. 1999. *J. Chem. Phys.*, **111**, 1366.

Elder M. 1973. *Int. J. Quantum Chem.*, **7**, 75.

Eliav E. and Kaldor U. 1996. *Chem. Phys. Lett.*, **248**, 405.

Eliav E., Kaldor U., and Hess B. A. 1998. *J. Chem. Phys.*, **108**, 3409.

Eliav E., Kaldor U., and Ishikawa Y. 1994. *Chem. Phys. Lett.*, **222**, 82.

Engel E. 2002. In *Relativistic Electronic Structure Theory. Part 1. Fundamentals*, Schwerdtfeger P. (ed.). Elsevier: Amsterdam, p. 523.

Engel E. and Dreizler R. M. 1996. *Topics in Current Chemistry*, **181**, 1.

Engel E., Müller H., Speicher C., and Dreizler R. M. 1995. In *Density Functional Theory*, Gross E. K. U. and Dreizler R. M. (eds.). Plenum: New York, p. 65.

Ermler W. C., Ross R. B., and Christiansen P. A. 1991. *Int. J. Quantum Chem.*, **40**, 829.

Eschrig H. 1996. *The Fundamentals of Density Functional Theory*. Teubner: Stuttgart.

Esser M. 1984a. *Int. J. Quantum Chem.*, **26**, 313.

Esser M. 1984b. *Chem. Phys. Lett.*, **111**, 58.

Esser M., Butscher W., and Schwarz W. H. E. 1981. *Chem. Phys. Lett.*, **77**, 359.

Faas S., Snijders J. G., van Lenthe J. H., van Lenthe E., and Baerends E. J. 1995. *Chem. Phys. Lett.*, **246**, 632.

Fægri, Jr. K. 2001. *Theor. Chem. Acc.*, **105**, 252.

Fægri, Jr. K. and Dyall K. G. 2002. In *Relativistic Electronic Structure Theory. Part 1. Fundamentals*, Schwerdtfeger P. (ed.). Elsevier: Amsterdam, p. 259.

Fægri, Jr. K. and Saue T. 2001. *J. Chem. Phys.*, **115**, 2456.

Farazdel A. and Smith, Jr. V. H. 1986. *Int. J. Quantum Chem.*, **29**, 311.

Fedorov D. and Klobukowski M. 2002. *Chem. Phys. Lett.*, **360**, 223.

Filatov M. 2002. *Chem. Phys. Lett.*, **365**, 222.

Filatov M. and Cremer D. 2003. *J. Chem. Phys.*, **119**, 11526.

Filatov M. and Cremer D. 2005. *J. Chem. Phys.*, **122**, 044104.

Foldy L. L. and Wouthuysen S. A. 1950. *Phys. Rev.*, **78**, 29.

Frank P. and Rothe H. 1911. *Ann. Physik*, **34**, 825.

French A. P. 1968. *Special Relativity*. W. W. Norton: New York.

Fricke B., Sepp W.-D. Bastug T., Varga S., Schultze K., Anton J., and Pershina V. 1997. *Adv. Quantum Chem.*, **29**, 109.

Goldstein H. 1950. *Classical Mechanics*. Addison-Wesley: Reading.

Gordon W. 1928. *Z. Phys.*, **50**, 630.

Gould M. D. and Battle J. S. 1993a. *J. Chem. Phys.*, **98**, 8843.

Gould M. D. and Battle J. S. 1993b. *J. Chem. Phys.*, **99**, 5983.

Grant I. P. 1982. *Phys. Rev. A*, **25**, 1230.

Grant I. P. and Quiney H. M. 1988. *Adv. At. Mol. Phys.*, **23**, 37.

Grant I. P. and Quiney H. M. 2002. In *Relativistic Electronic Structure Theory. Part 1. Fundamentals*, Schwerdtfeger P. (ed.). Elsevier: Amsterdam, p. 107.

Greiner W. 1997 *Relativistic Quantum Mechanics. Wave Equations* 2nd edn. Springer: Berlin.

Grelland H. H. 1980. *J. Phys. B*, **13**, L389.

Gross E. U. and Kurth S. 1994. In *Relativistic and Correlation Effects in Molecules and Solids*, Malli G. L. (ed.). Plenum: New York, p. 367.

Hafner P. 1980. *J. Phys. B*: *At. Mol. Phys.*, **13**, 3297.

Hamermesh M. 1962. *Group Theory and its Application to Physical Problems*. Addison-Wesley: Reading, MA.

Hay P. J. and Wadt W. R. 1985 *J. Chem. Phys.*, **82**, 270, 299.

Heinemann C., Koch W., Schwarz H., and Dyall K. G. 1996. *J. Chem. Phys.*, **104**, 4642.

Helgaker T. U., Jaszunski M., and Ruud K. 1999. *Chem. Rev.*, **99**, 293.

Helgaker T. U., Jørgensen P., and Olsen J. 2000. *Molecular Electronic Structure Theory*. Wiley: Chichester.

Helgaker T. U. and Taylor P. R. 1995. In *Modern Electronic Structure Theory, Part II*, Yarkony D. R. (ed.). World Scientific: Singapore, p. 725.

Herman F. and Skillman S. 1963. *Atomic Structure Calculations*. Prentice-Hall: Englewood Cliffs, NJ.

Hess B. A. 1985. *Phys. Rev. A*, **32**, 756.

Hess B. A. 1986. *Phys. Rev. A*, **33**, 3742.

Hess B. A., Buenker R. J., and Chandra P. 1986. *Int. J. Quantum Chem.*, **29**, 737.

Hess B. A., Buenker R. J., Marian C. M., and Peyerimhoff S. D. 1982. *Chem. Phys. Lett.*, **89**, 459.

Hess B. A., Marian C. M., Wahlgren U., and Gropen O. 1996. *Chem. Phys. Lett.*, **251**, 365.

Heully J.-L., Lindgren I., Lindroth E., Lundqvist S., and Mårtensson-Pendrill A.-M. 1986. *J. Phys. B: At. Mol. Phys.*, **19**, 2799.

Hinchliffe A. 1987. *Ab Initio Determination of Molecular Properties*. Adam Hilger: Bristol.

Hoffmann M. R. and Dyall K. G. (eds.) 2002. *Low-Lying Potential Energy Surfaces (ACS Symposium Series 828)*. American Chemical Society: Washington, DC.

Huber K. P. and Herzberg G. (eds.) 1979. *Molecular Spectra and Molecular Structure. IV. Constants of Diatomic Molecules*. van Nostrand Reinhold: New York.

Hurley M. M., Pacios L. F., Christiansen P. A., Ross R. B., and Ermler W. C. 1986. *J. Chem. Phys.*, **84**, 6840.

Huzinaga S. and Cantu A. A. 1971. *J. Chem. Phys.*, **55**, 5543.

Huzinaga S. and Klobukowski M. 1985. *Chem. Phys. Lett.*, **120**, 509.

Huzinaga S., Seijo L., Barandiarán Z., and Klobukowski M. 1987. *J. Chem. Phys.*, **86**, 2132.

Huzinaga S., McWilliams D., and Cantu A. A. 1973. *Adv. Quantum Chem.*, **7**, 187.

Ishikawa Y., Baretty R., and Binning, Jr. R. C. 1985. *Chem. Phys. Lett.*, **121**, 130.

Jackson J. D. 1975. *Classical Electrodynamics*, 2nd edn. Wiley: New York, p. 182.

Jamieson M. J., Drake G. W. F., and Dalgarno A. 1995. *Phys. Rev. A*, **51**, 3358.

Jensen H. J. Aa., Dyall K. G., Saue T., and Fægri, Jr. K. 1996. *J. Chem. Phys.*, **104**, 4083.

Jensen F. 1999. *Introduction to Computational Chemistry*. Wiley: Chichester.

Jeziorski B. and Monkhorst H. J. 1981. *Phys. Rev. A*, **24**, 1668.

Johnson K. H. 1973. *Adv. Quantum Chem.*, **7**, 143.

Jørgensen P. and Simons J. 1981. *Second Quantization-Based Methods in Quantum Chemistry*. Academic Press: New York.

Jørgensen P. and Simons J. (eds.) 1986. *Geometrical Derivatives of Energy Surfaces and Molecular Properties*. D. Reidel: Dordrecht.

Kahn L., Baybutt P., and Truhlar D. G. 1976. *J. Chem. Phys.*, **65**, 3826.

Kaupp M. and Schleyer P. von R. 1993. *J. Am. Chem. Soc.*, **115**, 1061.

Kellö V. and Sadlej A. J. 1990. *J. Chem. Phys.*, **93**, 8122.

Kellö V. and Sadlej A. J. 1998. *Int. J. Quantum Chem.*, **68**, 159.

Klobukowski M. 1990. *Chem. Phys. Lett.*, **172**, 361.

Knowles P. J. and Handy N. C. 1984. *Chem. Phys. Lett.*, **111**, 315.

Knowles P. J. and Werner H.-J. 1988. *Chem. Phys. Lett.*, **145**, 514.

Koch W. and Holthausen M. C. 2001. *A Chemist's Guide to Density Functional Theory*, 2nd edn. Wiley-VCH: Weinheim.

Kozlov M. G., Fomichev V. I., Dmitriev Y. Y., Labzowsky L. N., and Titov A. V. 1987. *J. Phys. B: At. Mol. Phys.*, **20**, 4939.

Kramers H. A. 1930. *Proc. Acad. Amsterdam*, **33**, 959.

Kullie O., Dusterhöft C., and Kolb D. 1999. *Chem. Phys. Lett.*, **314**, 307.

Kutzelnigg W. 1984. *Int. J. Quantum Chem.*, **25**, 107.

Kutzelnigg W. 1989. *Z. Phys. D*, **11**, 15.
Kutzelnigg W. 1990. *Z. Phys. D*, **15**, 27.
Kutzelnigg W. 1996. *Phys. Rev. A*, **54**, 1183.
Kutzelnigg W. 1999. *J. Chem. Phys.*, **110**, 8283.
Kutzelnigg W. 2003. *Phys. Rev. A*, **67**, 032109.
Kutzelnigg W. and Liu W. 2005. *J. Chem. Phys.*, **123**, 241102.
Kutzelnigg W., Ottschofski E., and Franke R. 1995. *J. Chem. Phys.*, **102**, 1740.
Laaksonen L. and Grant I. P. 1984a. *Chem. Phys. Lett.*, **109**, 485.
Laaksonen L. and Grant I. P. 1984b. *Chem. Phys. Lett.*, **112**, 157.
Lærdahl J. K., Wesendrup R., and Schwerdtfeger P. 2000. *Chem. Phys. Chem.*, **1**, 60.
LaJohn L. A., Christiansen P. A. Ross R. B., Atashroo T., and Ermler W. C. 1987. *J. Chem. Phys.*, **87**, 2812.
Lamb Jr. W. E. 1952. *Phys. Rev.*, **85**, 259.
Lee Y. S. 2004. In *Relativistic Electronic Structure Theory. Part 2. Applications*, Schwerdtfeger P. (ed.). Elsevier: Amsterdam, p. 352.
Lee Y. S. and McLean A. D. 1982a. In *Current Aspects of Quantum Chemistry* 1981, Carbó R. (ed.). Elsevier: Amsterdam, p. 219.
Lee Y. S. and McLean A. D. 1982b. *J. Chem. Phys.*, **76**, 735.
Lévy-Leblond J. M. 1967. *Comm. Math. Phys.*, **6**, 286.
Lévy-Leblond J. M. 1970. *Ann. Phys.*, **57**, 481.
Llusar R., Casarrubios M., Barandiarán Z., and Seijo L. 1996. *J. Chem. Phys.*, **105**, 5321.
London F. 1937. *Journal de Physique*, **8**, 397.
MAGIC code. Willetts A., Gagliardi L., Ioannou A. G., Simper A. M., Skylaris C.-K., Spencer S., and Handy N. C. 2000. *Int. Rev. Phys. Chem.*, **19**, 327.
Malli G. L. and Oreg J. 1975. *J. Chem. Phys.*, **63**, 830.
Manne R. 1972. *Mol. Phys.*, **24**, 935.
Mark F., Lischka H., and Rosicky F. 1980. *Chem. Phys. Lett.*, **71**, 507.
Mark F. and Rosicky F. 1980. *Chem. Phys. Lett.*, **74**, 562.
Matsuoka O. 1987. *Chem. Phys. Lett.*, **140**, 362.
Matsuoka O. and Okada S. 1989. *Chem. Phys. Lett.*, **155**, 547.
Merzbacher E. 1970. *Quantum Mechanics*, 2nd edn. John Wiley & Sons: New York.
Messiah A. 1961. *Quantum Mechanics*. North-Holland: Amsterdam.
Mittleman M. H. 1981. *Phys. Rev. A*, **24**, 1167.
Morrison J. D. and Moss R. E. 1980. *Mol. Phys.*, **41**, 491.
Moss R. E. 1973. *Advanced Molecular Quantum Mechanics: An Introduction to Relativistic Quantum Mechanics and the Quantum Theory of Radiation*. Chapman and Hall: London.
Nakajima N. and Hirao K. 2000. *J. Chem. Phys.*, **113**, 7786.
Nakatsuji H., Matsuzaki Y., and Yonezawa T. 1988. *J. Chem. Phys.*, **88**, 5759.
Onoe J., Nakamatsu H., Mukoyama T., Sekine R., Adachi H., and Takeuchi K. 1997. *Inorg. Chem.*, **36**, 1934.
Pacios L. F. and Christiansen P. A. 1985. *J. Chem. Phys.*, **82**, 2664.
Park C. and Almlöf J. 1994. *Chem. Phys. Lett.*, **231**, 269.
Parr R.G. and Yang W. 1989. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press: Oxford.
Pershina V. 2004. In *Relativistic Electronic Structure Theory. Part 2. Applications*, Schwerdtfeger P. (ed.). Elsevier: Amsterdam, p. 1.
Phillips J. C. and Kleinman L. 1959. *Phys. Rev.*, **116**, 287.
Pitzer K. S. 1979. *Acc. Chem. Res.*, **12**, 271.
Pitzer R. M. and Winter N. W. 1991. *Int. J. Quantum Chem.*, **40**, 773.
Primas H. 1980. *In Quantum Dynamics of Molecules*, Woolley R. G. (ed.). Plenum Press: New York, p. 39.

Pyykkö P. 1978. *Adv. Quantum Chem.*, **11**, 353.

Pyykkö P. 1986. *Relativistic Theory of Atoms and Molecules* (Lect. Notes Chem. vol. 41). Springer-Verlag: Berlin.

Pyykkö P. 1988. *Chem. Rev.*, **88**, 563.

Pyykkö P. 1993. *Relativistic Theory of Atoms and Molecules II* (Lect. Notes Chem. vol. 60). Springer-Verlag: Berlin.

Pyykkö P. 2000. *Relativistic Theory of Atoms and Molecules III* (Lect. Notes Chem. vol. 76). Springer-Verlag: Berlin.

Pyykkö P. and Desclaux J.-P. 1979. *Acc. Chem. Res.*, **12**, 276.

Pyykkö P., Dyall K. G., Császár A. G., Tarczay G., Polyansky O. L., and Tennyson J. 2001. *Phys. Rev. A*, **63**, 024502.

Pyykkö P. and Stoll H. 2000. RSC Specialist Periodical Reports: *Chemical Modelling, Applications and Theory, 1*, 239.

Quiney H. M. 2002. In *New Trends in Quantum Systems in Chemistry and Physics, Vol. 1*, Maruani J., Minot C., McWeeny R., Smeyers Y. G., and Wilson S. (eds.). Kluwer Academic: Dordrecht, p. 135.

Quiney H. M., Belanzoni P., and Sgamellotti A. 2002. *Theor. Chem. Acc.*, **108**, 113.

Quiney H. M., Lærdahl J. K., Fægri K., and Saue T. 1998. *Phys. Rev. A*, **57**, 920.

Quiney H. M., Skaane H., and Grant I. P. 1999. *Adv. Quantum Chem.*, **32**, 1.

Rajagopal A. K. 1978. *J. Phys. C*, **11**, L943.

Rajagopal A. K. and Callaway J. 1973. *Phys. Rev. B*, **7**, 1912.

Ramsey N. F. 1953. *Phys. Rev.*, **91**, 303.

Rao S., Stoicheff B. P., and Turner R. 1960. *Can. J. Phys.*, **38**, 1516.

Rohlfing C. M., Hay P. J., and Martin R. L. 1986. *J. Chem. Phys.*, **85**, 1447.

Roothaan C. C. J. and Bagus P. S. 1963. *Meth. Comp. Phys.*, **2**, 47.

Rösch N., Matveev A., Nasluzov A., Neyman K. M., Moskaleva L., and Krüger S. 2002. In *Relativistic Electronic Structure Theory. Part 2. Applications*, Schwerdtfeger P. (ed.). Elsevier: Amsterdam, p. 656.

Rose M. E. 1961. *Relativistic Electron Theory*. Wiley: New York.

Rose S. J., Grant I. P., and Pyper N. C. 1978. *J. Phys. B: At. Mol. Phys.*, **11**, 1171.

Rosén A. 1997. *Adv. Quant. Chem.*, **29**, 1.

Rosén A. and Ellis D. E. 1975. *J. Chem. Phys.*, **62**, 3039.

Rosicky F. and Mark F. 1979. *Theor. Chim. Acta*, **54**, 35.

Ross R. B., Gayen S., and Ermler W. C. 1994. *J. Chem. Phys.*, **100**, 8145.

Ross R. B., Powers J. M., Atashroo T., Ermler W. C., LaJohn L. A., and Christiansen P. A. 1990. *J. Chem. Phys.*, **93**, 6654.

Rutkowski A. 1986. *J. Phys. B: At. Mol. Phys.*, **19**, 149, 3431, 3443.

Rutkowski A., Jankowski K., and Mikielewicz B. 1988. *J. Phys. B: At. Mol. Phys.*, **21**, L147.

Rutkowski A., Rutkowska D., and Schwarz W. H. E. 1992. *Theor. Chim. Acta*, **84**, 105.

Ruud K., Helgaker T., Bak K. L., Jørgensen P., and Jensen H. J. Aa. 1993. *J. Chem. Phys.*, **99**, 3847.

Sadlej A. J. and Snijders J. G. 1994. *Chem. Phys. Lett.*, **229**, 435.

Sadlej A. J., Snijders J. G., van Lenthe E., and Baerends E. J. 1995. *J. Chem. Phys.*, **102**, 1758.

Sakai Y., Miyoshi E., Klobukowski M., and Huzinaga S. 1997. *J. Chem. Phys.*, **106**, 8084.

Sakurai J. J. 1967. *Advanced Quantum Mechanics*. Addison-Wesley: London.

Salahub D. R., Castro M., and Proynov E. I. 1994. In *Relativistic and Correlation Effects in Molecules and Solids*, Malli G. L. (ed.). Plenum: New York, p. 411.

Samzow R., Hess B. A., and Jansen G. 1992. *J. Chem. Phys.*, **96**, 1227.

Sandars P. G. H. 1967. *Phys. Rev. Lett.*, **19**, 1396.

Saue T., Fægri, Jr. K., and Gropen O. 1996. *Chem. Phys. Lett.*, **263**, 360.

Saue T. and Helgaker T. U. 2002. *J. Comp. Chem.*, **23**, 814.

Saue T. and Jensen H. J. Aa. 1999. *J. Chem. Phys.*, **111**, 6211.

Schiff L. I. 1968. *Quantum Mechanics*, 3rd edn. McGraw-Hill Kogakusha: New York.

Schonland D. S. 1965. *Molecular Symmetry*. D. van Nostrand: London.

Schwerdtfeger P. (ed). 2002. *Relativistic Electronic Structure Theory. Part 1. Fundamentals*. Elsevier: Amsterdam.

Schwerdtfeger P. (ed). 2004. *Relativistic Electronic Structure Theory. Part 2. Applications*. Elsevier: Amsterdam.

Schwarz W. H. E., van Wezenbeek E. M., Baerends E. J., and Snijders J. G. 1989. *J. Phys. B: At. Mol. Phys.*, **22**, 1515.

Schwerdtfeger P., Lærdahl J., and Chardonnet C. 2002. *Phys. Rev. A*, **65**, 042508.

Seijo L. and Barandiarán Z. 1999. In *Computational Chemistry: Reviews of Current Trends, Vol. 4*, Leszczynski X. (ed.). World Scientific: Singapore, p. 55.

Shavitt I. 1988. In *Mathematical Frontiers in Computational and Chemical Physics*, Truhlar D. G. (ed.). Springer-Verlag: New York, p. 300.

Siegbahn P. E. M. 1984. *Chem. Phys. Lett.*, **109**, 417.

Sjövoll M., Gropen O., and Olsen J. 1997. *Theor. Chem. Acc.*, **97**, 301.

Slater J. C. 1960. *Quantum Theory of Atomic Structure, Vol. 1*. McGraw-Hill: New York.

Slater J. C. 1972. *Adv. Quant. Chem.*, **6**, 1.

Slater J. C. 1974. *Quantum Theory of Molecules and Solids, Vol. 4*. McGraw-Hill: New York.

Snijders J. G. 1979. *Relativity and Pseudopotentials in the HFS Method*, Thesis, Vrije Universiteit te Amsterdam.

Snijders J. G. and Baerends E. J. 1977. *Mol. Phys.*, **33**, 1651.

Snijders J. G. and Baerends E. J. 1978. *Mol. Phys.*, **36**, 1789.

Stanton R. E. and Havriliak S. 1984. *J. Chem. Phys.*, **81**, 1910.

Sucher J. 1980. *Phys. Rev. A*, **22**, 348.

Sucher J. 1984. *Int. J. Quantum Chem.*, **25**, 3.

Sundholm D. 1988. *Chem. Phys. Lett.*, **149**, 251.

Sundholm D. 1994. *Chem. Phys. Lett.*, **223**, 469.

Sundholm D. and Ottschofski E. 1997. *Int. J. Quantum Chem.*, **65**, 151.

Sutcliffe B. T. 1980. In *Quantum Dynamics of Molecules.*, Woolley R. G. (ed.). Plenum Press: New York, p. 1.

Swanström P. and Hegelund F. 1975. In *Computational Techniques in Quantum Chemistry and Molecular Physics.*, Diercksen G. H. F., Sutcliffe B. T., and Veillard A. (eds.). D. Reidel: Dordrecht, p. 299.

Szabo A. and Ostlund N. S. 1989. *Modern Quantum Chemistry*. McGraw-Hill: New York.

Talman J. D. 1986. *Phys. Rev. Lett.*, **57**, 1091.

Teichteil C., Pélissier M., and Spiegelmann F. 1983. *Chem. Phys.*, **81**, 273.

Tinkham M. 1964. *Group Theory and Quantum Mechanics*. McGraw-Hill: New York.

Titov A. V. and Mosyagin N. S. 1999. *Int. J. Quantum Chem.*, **71**, 359.

Vahtras O., Ågren H., Jørgensen P., Jensen H. J. Aa., Helgaker T. U., and Olsen J. 1992. *J. Chem. Phys.*, **96**, 2118.

van Lenthe E. 1996. *The ZORA Equation*, doctoral thesis, Vrije Universiteit, Amsterdam.

van Lenthe E., Baerends E. J., and Snijders J. G. 1993. *J. Chem. Phys.*, **99**, 4597.

van Lenthe E., Baerends E. J., and Snijders J. G. 1994. *J. Chem. Phys.*, **101**, 9783.

van Wüllen C. 1995. *J. Chem. Phys.*, **103**, 3589.

van Wüllen C. 1998. *J. Chem. Phys.*, **109**, 392.

van Wüllen C. 2004. In *Relativistic Electronic Structure Theory. Part 2. Applications*, Schwerdtfeger P. (ed.). Elsevier: Amsterdam, p. 598.

Visscher L. 1997. *Theor. Chem. Acc.*, **98**, 68.

Visscher L., Aerts P. J. C., Visser O., and Nieuwpoort W. C. 1991. *Int. J. Quantum Chem. Symp.*, **25**, 131.

Visscher L. and Dyall K. G. 1997. *At. Data Nucl. Data Tables*, **67**, 207.

Visscher L., Dyall K. G., and Lee T. J. 1995. *Int. J. Quantum Chem. Symp.*, **29**, 411.

Visscher L., Dyall K. G., and Lee T. J. 1996. *J. Chem. Phys.*, **105**, 8769.

Visscher L., Eliav E., and Kaldor U. 2001. *J. Chem. Phys.*, **115**, 9720.

Visscher L. and Saue T. 2000. *J. Chem. Phys.*, **113**, 3996.

Visscher L. and van Lenthe E. 1999. *Chem. Phys. Lett.*, **306**, 357.

Visser O., Aerts P. J. C., Hegarty D., and Nieuwpoort W. C. 1987. *Chem. Phys. Lett.*, **134**, 34.

Wadt W. R. and Hay P. J. 1985. *J. Chem. Phys.*, **82**, 284.

Weeks J. D. and Rice S. A. 1968. *J. Chem. Phys.*, **49**, 2741.

Wigner E. P. 1959. *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra*. Academic Press: New York.

Wildman S. A., DiLabio G. A., and Christiansen P. A. 1997. *J. Chem. Phys.*, **107**, 9975.

Wolf A., Reiher M., and Hess B. A. 2002. *J. Chem. Phys.*, **117**, 9215.

Wood J. H. and Boring A. M. 1978. *Phys. Rev. B*, **18**, 2701.

Yabushita S., Zhang Z. Y., and Pitzer R. M. 1999. *J. Phys. Chem. A*, **103**, 5791.

Yanai T., Iikura H., Nakajima T., Ishikawa Y., and Hirao K. 2001a. *J. Chem. Phys.*, **115**, 8267.

Yanai T., Nakajima T., Ishikawa Y., and Hirao K. 2001b. *J. Chem. Phys.*, **114**, 6526.

Yanai T., Nakajima T., Ishikawa Y., and Hirao K. 2002. *J. Chem. Phys.*, **116**, 10122.

Yarkony D. R. 1992. *Int. Rev. Phys. Chem.*, **11**, 195.

Index