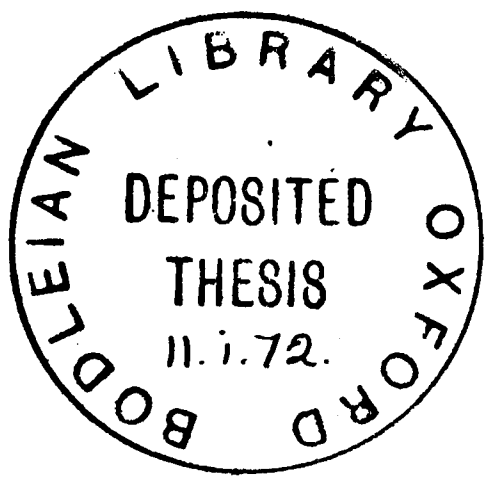
SOME NUMERICAL PROBLEMS IN QUANTUM THEORY

G. H. Redman

Magdalen College



A thesis submitted for the degree of

Doctor of Philosophy

in the University of Oxford

June 1971

ABSTRACT

This thesis presents the results of an exercise in practical numerical analysis whose aim is the estimation of realistic 'a posteriori' error bounds for the solutions of atomic structure calculations computed by finite-difference methods.

In performing these calculations the basic problem is to determine the eigenstates of a particle in a central force field by solving the one-dimensional radial form of the Schrodinger (non-relativistic) or Dirac (relativistic) equation, subject to homogeneous boundary conditions. The system represented in either case is essentially a second-order ordinary linear differential system of the type arising in many branches of atomic, molecular, and nuclear physics. In particular, the Hartree-Fock approach to the problem of calculating atomic structures involves the solution of one or more of these equations by an iterative scheme.

Chapter 1 describes the general aims of the thesis: these are to produce non-pessimistic error estimates for the solutions computed on any particular grid by obtaining first-order bounds on the solutions of the finite-difference systems satisfied by the errors themselves. The intention is to gain a realistic assessment of the order of magnitude of the error at every point over the range of tabulation of each of the computed solutions, thus indicating the number of figures worth quoting in the results. The sources of error taken into account include the errors arising from replacement of the derivatives by truncated finite-difference expansions ('truncation errors') and those arising from uncertainties in any initial data that may be used in performing the calculation ('physical errors'). Compared with these, the effect of rounding errors is assumed sufficiently small to be neglected.

Chapter 2 contains a description of a standard numerical method of determining the eigenstates of a particle in a central field in the non-relativistic approximation by solving the radial form of the Schrodinger

equation, employing for this purpose a combination of initial-value and boundary-value techniques. An extension of the method is described for dealing with the situation which arises for example in atomic structure theory in which an inhomogeneous term is present in the equation. The problem is not then strictly of eigenvalue type but a solution satisfying a specified normalization condition exists only for certain discrete values of the energy parameter.

The results which form the basis of the error analysis are presented in Chapter 3. The exact differential system is regarded as a perturbation of the finite-difference system actually solved, and the errors themselves are regarded as forming the solution of a finite-difference system. Taking into account each source of error and working to the first order in small quantities, this system is 'solved' by finding a first-order estimate of the error in the computed eigenvalue and a similar estimate of the error in each element of the vector representing the eigenfunction. Attention is paid to the problem of ensuring that the error estimate reflects as closely as possible the form of the actual error in the computed eigenfunction over the range of integration.

In Chapter 4 the formulae resulting from this analysis are rewritten in terms of the notation used in Chapter 2. Several sets of results are presented for the special case in which the central force field is a Coulomb field. In this case the analytic solutions of the exact differential system are known and the computed error estimates may be compared with the actual errors in the computed solutions. In making this comparison, separate consideration is given to the effect on the results of truncation error and simulated physical errors in the cases of both homogeneous and inhomogeneous systems.

Chapter 5 is concerned with the numerical calculation of the eigenstates of a particle in a central field by solving the radial form of the wave equation in Dirac's relativistic quantum theory. The radial wave functions

now have two components satisfying a pair of first-order differential equations, but the techniques employed for computing the solutions are similar to those described in Chapter 2. In presenting the method, the question of the stability of the techniques is discussed. An error analysis is developed along similar lines to the treatment presented in Chapter 3 and the results of the analysis are tested in the same way as in Chapter 4 by making use of the known solutions for the Coulomb field.

The subject of Chapter 6 is the application of the error analysis to the problem of calculating atomic structures in the non-relativistic approximation by solving the Hartree-Fock equations by the iterative self-consistent field technique. A general description of the technique is presented, and a method is proposed for determining 'a posteriori' error bounds on the solutions by incorporating the error analysis into an additional iteration of the self-consistent field procedure. The bounds are derived by taking into account the effect of the truncation errors associated with the replacement of the exact differential system by the approximating finite-difference system over a given grid, and these bounds are independent of the number of iterations performed to achieve self-consistency. Results are presented for the two simplest atomic systems, representing the ground states of atoms or ions of two electrons and four electrons respectively, and suggestions are made for modifying the methods of analysis so as to effect improvements in the results.

PREFACE

This thesis contains results of research undertaken at the Oxford University Computing Laboratory into methods of assessing the accuracy of atomic wave functions calculated by finite-difference techniques. In the context of atomic structure calculations there has been no previous attempt to apply methods of error estimation of the type described here. The numerical work was carried out on the KDF9 computer, using a version of the Algol language available in the Egdon system.

My warmest thanks are due to my supervisor, Dr. D. F. Mayers, for his guidance and encouragement throughout the course of this research.

I am grateful to the Science Research Council for financial support.

CONTENTS

Abstract

Preface

1 Introduction

1.1 Aims

1.2 Methods

2 The radial Schrodinger equation

2.1 The homogeneous equation

2.2 Numerical solution

2.3 The inhomogeneous equation

3 Development of error estimates

3.1 General analysis - homogeneous system

3.2 General analysis - inhomogeneous system

3.3 Errors in eigenvalue and scale factor

3.4 Error in eigenfunction

3.5 Further remarks

4 Application to the Schrodinger equation

4.1 Application of theory

4.2 Non-relativistic Coulomb functions

5 The radial Dirac equation

5.1 Numerical solution

5.2 Error analysis

5.3 Relativistic Coulomb functions

6 Radial wave functions of the self-consistent field

6.1 The Hartree-Fock equations

6.2 Error analysis

6.3 The radial equations

6.4 The potential and exchange terms

6.5 Configuration of one group .

6.6 Configuration of two groups

References

CHAPTER 1

INTRODUCTION

1.1 Aims

The advent of the elctronic computer has occasioned over the last fifteen years a renewed interest in the calculation of atomic structures and has enabled calculations to be performed on structures of ever increasing complexity. The mathematical system concerned forms a differential eigenvalue problem consisting of one or more coupled second-order ordinary differential equations with homogeneous boundary conditions. The problem of calculating the eigensolutions has been studied extensively, and efficient and stable numerical methods have been developed using finite differences. These methods, which employ a combination of initial-value and boundary-value techniques, are well known and of widespread application not only in atomic structure theory but also in many other branches of atomic, molecular, and nuclear physics and in physical chemistry and astrophysics. Many atomic properties such as transition probabilities and oscillator strengths can be evaluated as a result of these calculations, and approximate descriptions of molecular and solid-state phenomena can also be obtained in this way.

Thus the desirable situation has been reached in which methods originally developed by the numerical analyst are widely used by research workers of many disciplines. It is principally because of this widespread usage, however, that the numerical analyst has the responsibility of ensuring that at least some consideration is given to the problem of assessing the accuracy of the results obtained. This aspect of the problem has never received very much attention in spite of the obvious fact that an alleged 'eigensolution' without an error estimate is really worth nothing. A person experienced in the use of a particular technique may sometimes be able to form a good idea of the accuracy of his solutions from their degree of smoothness and consistency, but these tests are far from infallible and checks of a much more precise nature are needed for computer use.

An attempt will be made here to supply this need by performing an error analysis on the numerical methods of solution currently in use. This analysis will be aimed at the production of non-pessimistic error estimates which will have the effect of indicating the order of magnitude of the errors and hence the number of figures worth quoting in the results. The intention in estimating the error in a tabulated wave function, for example, will be to reflect as closely as possible the variations in magnitude of the actual error over the range of tabulation. This aim will be achieved by taking 'a posteriori' account of the various sources of error and determining the collective effect of these errors on the computed eigensolutions.

1.2 Methods

Consideration must first be given to the basic problem in this field of work, which is the problem of calculating the quantum states of a particle in a central force field - a field, that is, whose associated potential is spherically symmetric. Using spherical polar coordinates, the wave function for a particle in a central field is separable, and, whereas the angular part is simply a spherical harmonic, the radial component depends on the exact form of the potential and cannot be determined in closed form except in certain special cases. Analytic approximations to the radial functions can sometimes be constructed out of a suitable basis-set of functions, but this thesis is concerned only with the finite-difference methods that have been developed for calculating the eigenstates.

When the potential is of the general type occurring in the theory of atomic structures, the particle may exist in either of two distinct kinds of eigenstate: (i) states in which the particle remains bound to the force centre, corresponding to values of the energy less than the threshold required to free the particle from the influence of the force, and (ii) states in which the particle is unbound and free to move infinitely far from the force centre. The free-particle states form an infinite set of

'improper' eigenstates, there being an allowed state for any value of the energy above the threshold, whereas the bound states form a set of discrete energy levels whose determination constitutes a true eigenvalue problem. Only the bound states will be considered in this thesis.

The radial wave functions and the energy levels of a particle in a central field are obtained as the eigensolutions of the radial forms of either the Schrodinger equation or the Dirac equation (depending on whether or not relativistic effects are to be neglected) with homogeneous boundary conditions. The mode of approach to the problem of error analysis involves recognizing that the solution of the exact differential system, consisting of the radial equation (Schrodinger or Dirac) and the boundary conditions, is also the solution of a linear algebraic system which may be treated as a perturbation of the system obtained from it by finite-difference approximation. This latter system is the one which is actually solved, and the errors appearing in the solution can themselves be regarded as forming the solution of a similar algebraic system some of whose coefficients, however, are only known as bounds on the differences between the coefficients in the exact system and those in the system actually solved. The errors arise because of the replacement of each derivative by a truncated difference expansion (this will be referred to as 'truncation error') and because of uncertainties in any initial data that may be used in solving a particular problem (to be referred to as 'physical error').

In carrying out the numerical work for this thesis, the working accuracy available amounted to 11 or 12 significant decimal digits, an accuracy which is sufficiently high for the errors introduced by rounding to be ignored and all arithmetic to be assumed exact. Thus in every case the finite-difference interval used in solving the problem and the accuracy demanded of the solution are such that the effect of the rounding error remains insignificant compared with the effects of truncation and physical error. The methods of solution employed are all numerically stable in the

sense that they do not induce any relative growth in the effects of rounding error in performing the computation, and it is safe in all cases to ignore the propagated effect of this error on the computed solution.

As an illustration of the incidence of physical error can be cited the problem of determining the vibrational wave functions of a diatomic molecule when the potential function is known only approximately on the basis of inexact experimental data. A rather different situation arises in solving the radial equations in an atomic structure calculation, for now the potential term in each equation is a sum of integrals over the wave functions to be computed, and the error arises because the functions are not known exactly and because of the approximations made in evaluating the integrals by numerical methods.

The truncation error of the finite-difference solution on a particular grid is a function of the interval size (or grid spacing) h. If the form of this function, or at least its leading term, is known, it may be possible to obtain a more accurate solution by eliminating part of the error between solutions obtained at two different values of h. This is the well-known procedure called 'Richardson extrapolation', or 'the deferred approach to the limit'. However it does not give precise bounds for the solution and its use depends on the determination of an asymptotic form for the error as a function of h, in other words one which is valid when h is sufficiently small.

On a given fixed grid, a statement that the truncation error is of a certain order in h indicates nothing whatever about the accuracy of a particular solution, although it does indicate something about the method. The results are usually checked in practice by relying on consistency and comparing the solutions obtained at h and $\frac{1}{2}$h. This is not a rigorous check and can be unreliable if the function does not behave smoothly between grid points. In any case it is often impracticable to refine the grid until consistent approximate solutions are obtained, and it would be convenient

to be able to compute strict error bounds for the solution obtained on any particular grid.

Still considering the effect of truncation error, an improved solution can always be obtained without refining the grid by performing one or more stages of a deferred correction process. This is a well known procedure, of value as a means of correcting the first approximation to the solution, but again there is no provision for calculating precise error bounds for this solution. Furthermore it would not be possible to make a correction in this way to take into account the effects of any physical errors on the solution.

The idea behind the methods developed in this thesis will be to take first-order account of the various sources of error (both truncation and physical) and to continue working to the first order throughout the analysis; thus in a sense the error 'bounds' which will be produced would be more correctly described as first-order estimates. As regards the dependence on truncation error the procedure is equivalent in many respects to the performance of the first stage of a deferred correction process but with the difference that upper bounds are computed on the first-order perturbations instead of their being calculated exactly. Since the problems are of eigenvalue type, the procedure to be developed will involve first bounding the error in the computed eigenvalue before determining an error bound on the eigenfunction itself.

The justification for the use of the methods lies in the validity of the assumption that, for the types of problem and methods of solution to be considered, the algebraic system would converge to the exact differential system if h were reduced to zero, and that, for the particular values of h chosen, the deferred correction process would form a rapidly convergent procedure if continued beyond the first stage. These statements imply that the procedures developed will sufficiently meet the requirements stated in §1.1, yielding realistic order-of-magnitude estimates of the errors and

serving to indicate the number of decimal places that may reliably be quoted in the results.

It should be noted that these purposes would not be served in attempting to develop strict overall error bounds for the solutions by using the known theorems on 'a posteriori' bounds for the eigensolutions of normal matrices. Although the algebraic system could in some cases be cast in a form suitable for application of these theorems, much information about the nature of the system would be lost in the process and the bound computed for a particular wave function would at best be obtained as a bound on the Euclidean vector norm, which would be of little use for estimating the accuracy of the function. Further difficulties would attend the extension of this approach to deal with those cases in which inhomogeneous terms are present in the equations so that the problem is no longer strictly of eigenvalue type. By contrast, an advantage of the methods which will be developed in this thesis is that they are easily extended to apply in these cases.

In its application to problems of the type in which use is made of initial data whose possible uncertainty is of known maximum amount, the analysis to be developed has the merit of allowing a comparison to be made between the effects of the physical and the truncation error on the computed solution. Since the truncation error is a function of h whereas the physical error is not, the value of the analysis in this situation lies in its facilitating the selection of the most appropriate value of h with which to perform the calculation. The danger is then avoided of using a very fine grid and quoting results to an accuracy higher than is allowed by that of the initial data.

The ability of the analysis to provide some indication of the accuracy of the computed solution on a particular grid is again of value in its application to the calculation of atomic structures. The wave functions are calculated by an iterative process, continuing until a state of 'self-

consistency' is reached, in which the current estimate of each function agrees with the previous estimate to within a prescribed tolerance, and there is clearly little point in achieving self-consistency to a higher degree of accuracy than is allowed by the truncation error. Irrespective of the number of iterations performed, the computed solutions remain inaccurate to an unknown extent as a result of this error, and it is intended that the analysis should be capable of furnishing realistic bounds on the error so as to provide a criterion for stopping the iterative procedure.

It is in situations such as this that it would be extremely useful to have realistic estimates of the errors in the computed solutions so as to be able to state with some degree of certainty the number of figures that could be relied upon in tabulating the results. The work of the thesis is aimed at the development of practical methods of supplying such estimates.

CHAPTER 2

THE RADIAL SCHRODINGER EQUATION

This chapter contains a detailed description of a standard numerical method of solving the radial Schrodinger equation to determine the bound states of a particle in a central field.

2.1 The homogeneous equation

The three-dimensional time-independent Schrodinger equation for a particle in a central field may be divorced of its angular dependence by the technique of separation of variables and reduced to a one-dimensional equation for the radial wave function $R(r)$, where r is the distance from the force centre. The details may be found for example in §1 of Bethe and Salpeter (1957). The radial equation so formed may be written in atomic units as follows:

$$\frac{1}{2}\frac{d^2p(r)}{dr^2} - [f(r) + \lambda]p(r) = 0 \qquad (2.1)$$

where $p(r) = r\,R(r)$, λ is the negative of the energy of the system, and $f(r)$ is a potential of the type occurring in atomic or molecular structure whose typical form is shown in Figure 2.1.

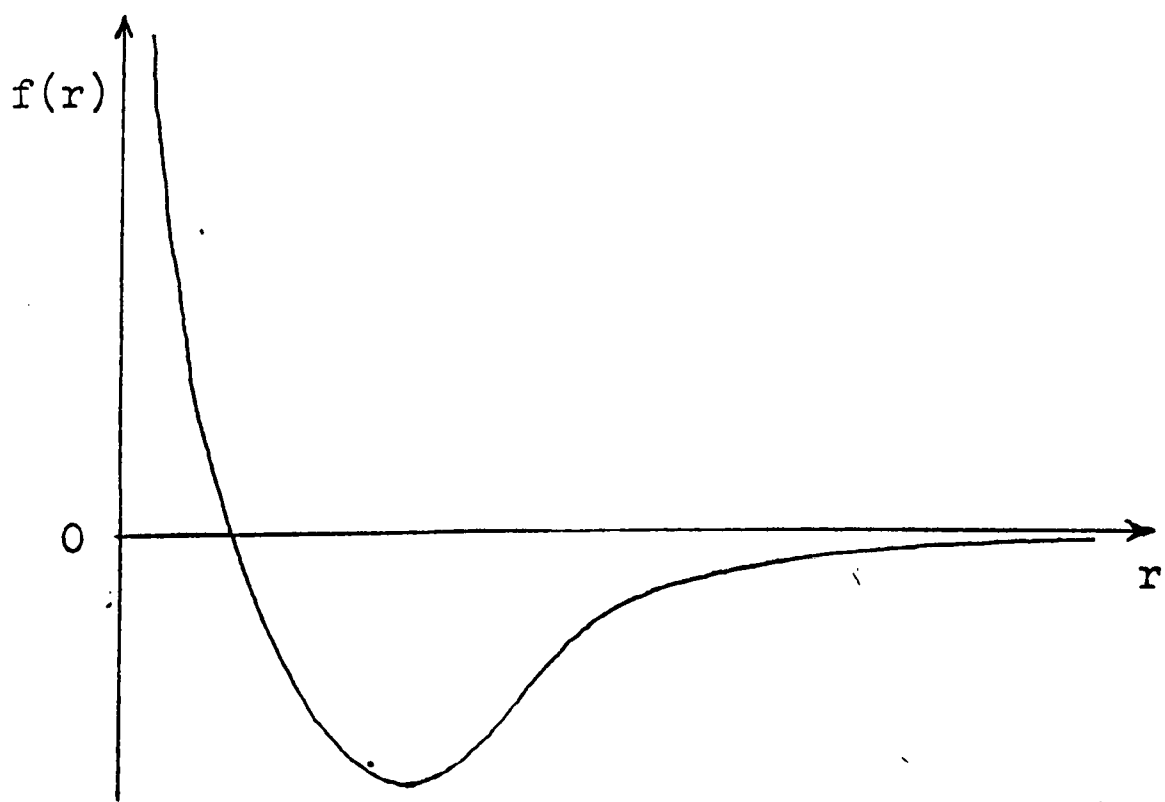


Fig. 2.1

The function $f(r)$ consists of two terms,

$$f(r) = \frac{l(l+1)}{2r^2} - v(r), \tag{2.2}$$

where l is the orbital quantum number (a non-negative integer) and where $v(r)$, the potential due to the central field, is positive in $(0, \infty)$, tends monotonically to zero as $r \to \infty$, and may be expanded about the origin as a convergent series in powers of r::

$$v(r) = \frac{1}{r}\left[v^{(0)} + v^{(1)}r + v^{(2)}r^2 + \ldots\right]. \tag{2.3}$$

The boundary conditions require the wave function $p(r)$ to vanish at $r = 0$ and to remain bounded as $r \to \infty$.

The self-adjoint system so defined belongs to the general class of Sturm-Liouville eigenproblems possessing an infinity of eigenvalues and a complete orthogonal set of eigenfunctions. In the present case the eigenspectrum consists of an infinite number of discrete positive eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_n > \ldots > 0$ belonging to a finite range with zero as limit point (these constitute the physically bound levels) and a continuous infinite spectrum of negative eigenvalues (the unbound levels). When arranged in this order, the n^{th} eigenvalue corresponds to an eigenfunction possessing exactly $n-1$ distinct nodes in $(0, \infty)$. [In the particular case of the Coulomb potential $v(r) = Z/r$ (representing for example a one-electron atom of central charge Z) the system, for a given quantum number l, has eigenvalues $\frac{1}{2}(Z/n)^2$ $(n = l+1, l+2, \ldots)$, the corresponding eigenfunctions having $n-l-1$ nodes.]

Interest will be centred on calculating the bound states, whose functions have exponentially-decreasing asymptotic behaviour and are square-integrable. The method of solution is developed so as to exclude the possibility of generating any of the unbound levels of the continuous spectrum, or any multiple of an increasing complementary function in regions where the required solution is decreasing. The functions so computed will

be normalized so as to satisfy

$$\int_0^\infty p^2(r)\,dr = 1. \tag{2.4}$$

2.2 Numerical solution

The method to be used employs a combination of initial-value and boundary-value techniques together with the process known as 'matching in the middle', and involves the solution of a variational system and subsequent iteration of the whole procedure, using Newton's method to gain improved solutions. The finite-difference approximation to equation (2.1) will be obtained by the method which takes advantage of the absence of a first derivative in the equation and gives rise to a three-term formula whose local truncation error is of sixth order in the step size. The process is usually attributed to Numerov [see, for example, Cooley(1961),§2] and reference will be made to it under that name. Although other methods will not be considered, it is only fair to point out, as in Sloan(1968a,b), that the order of error does not itself provide a sufficient criterion for comparing the Numerov method with, for instance, the fourth-order Runge-Kutta process since in both cases the cumulative error is of fourth-order in the step size. On the other hand the Runga-Kutta method suffers the limitation of requiring function evaluations at half-interval points and the disadvantage of possible induced instability when applied to certain differential systems [see for example: N.P.L.(1961),Ch.9; Fox(1962), Ch.4; Fox and Mayers(1968),Ch.10,§31]. Further comments on the comparison of the Numerov with other methods may be found in Froese(1963).

The form of potential depicted in Figure 2.1 is such that a step size sufficiently small for accurate representation of the eigenfunctions at small and intermediate values of r would prove unnecessarily small at large r. Blatt(1967) describes from a practical point of view a semi-automatic method of adjusting the step size during the course of the integration,

but this complication may be avoided by initially transforming on to a logarithmic scale. Such a transformation may lead to a grid which is rather too fine near the origin, but it retains the advantage of simplicity and is particularly useful in connection with the self-consistent field calculations considered in Chapter 6. Thus the transformation

$$t = \ln r \tag{2.5}$$

with equal intervals (h) in t produces a grid of points r_i ($= e^{t_i}$) in which Δr is proportional to r. Substitution of this transformation into equation(2.1) introduces a first derivative into the equation. This may be eliminated by a further transformation

$$\rho(r) = r^{-\frac{1}{2}} p(r). \tag{2.6}$$

[The general transformation is $\rho(r) = (\frac{dt}{dr})^{\frac{1}{2}} p(r)$ for any $t = t(r)$ satisfying $\frac{dt}{dr} > 0$ for all $r > 0$.] Substituting equations (2.2),(2.5), and (2.6) into equation (2.1), there results

$$\frac{d^2}{dt^2}\rho + \theta\rho = 0 \tag{2.7}$$

where

$$\begin{cases} \theta = \varphi - r^2\epsilon \\ \varphi = 2r^2 v - (l+\frac{1}{2})^2 \\ \epsilon = 2\lambda. \end{cases}$$

The Numerov method of approximation to this equation is based on the formula

$$\delta^2\rho_i - h^2(1 + \frac{1}{12}\delta^2)\rho_i'' + C_i(\rho) = 0 \tag{2.8}$$

where δ represents the central-difference operator and subscripted symbols denote values at grid points. The formula is simple to derive by combining Taylor expansions of ρ and ρ'' about the point r_i so as to eliminate ρ^{IV} and all odd derivatives, assuming that $\rho(r)$ has at least a continuous sixth derivative for all $r>0$.

The remainder term, represented in (2.8) by $C_i(\rho)$, can be written as $C_i(\rho) = -\frac{1}{240} h^6 \rho^{VI}(\xi)$ for some ξ in (r_{i-1}, r_{i+1}); see for example §6.1.3 of Henrici (1962). This term will be omitted in forming the finite-difference approximation so that it then constitutes the local truncation error which is subsequently to be included in the analysis leading to the estimation of error bounds for the computed eigensolutions. To this end, the term $C_i(\rho)$ may alternatively be represented as an expansion in terms of central differences

$$C_i(\rho) = \left(\frac{1}{240} \delta^6 - \frac{13}{15120} \delta^8 + \ldots \right) \rho_i \; ; \qquad (2.9)$$

see for example Fox (1957, Ch.4,§2).

Substituting for the second derivative in (2.8) from equation (2.7), the following discrete form of the equation is obtained:

$$(1+\tfrac{1}{12} h^2\theta_{i+1})\rho_{i+1} - (2-\tfrac{5}{6} h^2\theta_i)\rho_i + (1+\tfrac{1}{12} h^2\theta_{i-1})\rho_{i-1} + C_i(\rho) = 0.$$

It is convenient at this point to symmetrize the system by substituting

$$q = (1+\tfrac{1}{12} h^2\theta)\rho = (1+\tfrac{1}{12} h^2\theta)\, r^{-1/2} p \qquad (2.10)$$

to give

$$q_{i+1} - \left\{ \frac{2-\frac{5}{6} h^2\theta_i}{1+\frac{1}{12} h^2\theta_i} \right\} q_i + q_{i-1} + C_i(r^{-1/2} p) = 0. \qquad (2.11)$$

The finite-difference approximation to this equation is defined over the set of grid points as:

$$T_i(Q) = 0 \qquad (2.12)$$

where

$$\left.\begin{aligned} T_i(Q) &\equiv Q_{i+1} - a_i Q_i + Q_{i-1} \\ Q_i &= b_i r_i^{-1/2} P_i \\ a_i &= b_i^{-1}(2-\tfrac{5}{6} h^2\Theta_i) = \frac{12}{b_i} - 10 \\ b_i &= 1+\tfrac{1}{12} h^2\Theta_i \\ \Theta_i &= \Phi_i - r_i^2 E \; ; \quad \Phi_i = 2r_i^2 V_i - (l+\tfrac{1}{2})^2 ; \quad E = 2\Lambda. \end{aligned}\right\} \qquad (2.13)$$

[Capital letters denote grid-point approximations to the corresponding quantities appearing in the exact differential system.] The step size h must be chosen sufficiently small for $|2-a_i| \ll 2$ and $|1-b_i| \ll 1$ throughout the range of integration, which implies in particular that $h^2 r_n^2 E \ll 12$, where r_n represents the outermost grid point in the range.

Assuming estimates are known for the array of grid values V_i of the potential and the required eigenvalue Λ, the system of equations (2.12) (one equation based on each grid point) and boundary conditions can be solved for an approximate eigenfunction P. Each of the required eigenfunctions of the system is of a form which increases monotonically from zero as r increases from the origin to the inner 'classical turning point' [see Figure 2.2], undergoes bounded oscillation in the classical region, and decreases monotonically to zero as r increases beyond the outer turning point. However, the differential and approximating difference equations both possess complementary solutions which tend to infinity for small and large r, and the method of integration must be such as to avoid obtaining any multiples of these.

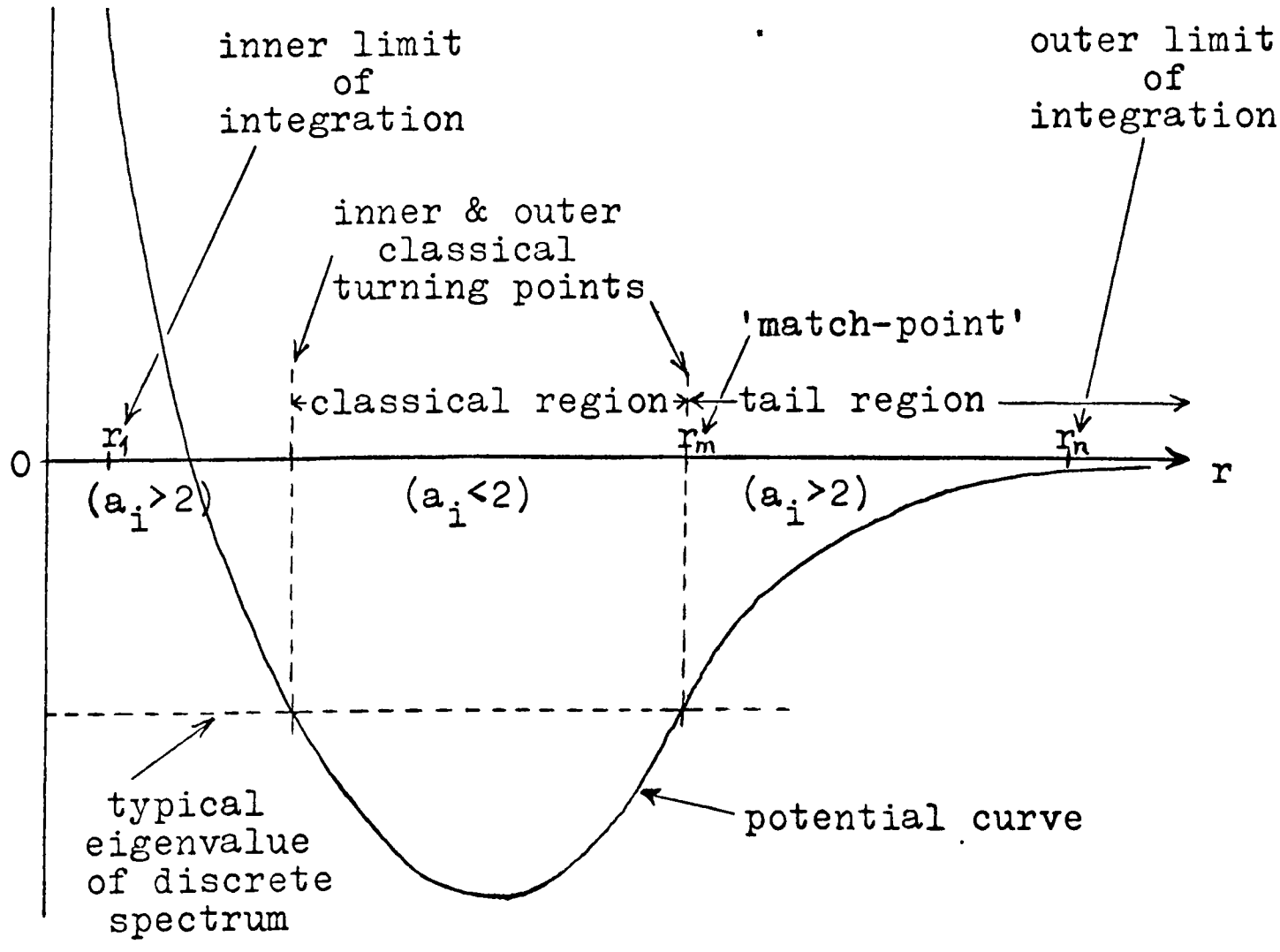


Fig. 2.2

§2.2

It is not possible using the logarithmic scale to begin integrating at the point r = 0, nor is it necessary since all wave functions become negligibly small near the force centre. The integration is started instead at a point r_1 sufficiently near the origin for the eigenfunction to be represented to the desired accuracy by the first few terms in its series expansion in powers of r. The exact position of r_1 is not critical but must certainly be less than the inner turning point for the level concerned. Equation (2.1) has a regular singularity at r = 0, and the function p(r) may be expanded in a convergent power series of the form:

$$p(r) = Ar^{\gamma} \sum_{j=0}^{\infty} \alpha_j r^j \qquad (2.14)$$

where A is an arbitrary multiplicative constant (subject only to the normalization condition).

Substituting equations (2.3) and (2.14) into (2.1) and equating to zero the coefficients of successive powers of r, the value $\gamma = l+1$ is derived from the indicial equation (applying the condition $p(0)=0$) and the following relation is obtained between the coefficients α_j for $j \geq 1$:

$$j(2l + 1 + j)\alpha_j = E\alpha_{j-2} - 2\sum_{k=0}^{j-1} v^{(k)}\alpha_{j-k-1}$$

(with $\alpha_0 \equiv 1$, $\alpha_{-1} \equiv 0$), where E has been substituted for the exact value ϵ.

Equation (2.12) can now be used to recur forwards from the starting point using the expansion

$$Q_i = Ab_i r_i^{l+\frac{1}{2}} \sum_{j=0,1,2,\ldots} \alpha_j r_i^j \qquad (2.15)$$

at the points r_1, r_2 and truncating the series when the contribution from subsequent terms becomes negligible.

The desired solution is monotonic increasing as far as the inner turning point, each coefficient a_i being greater than 2. Further, in the region between the turning points where $a_i < 2$, the complementary solutions of equation (2.12) possess the same bounded oscillatory behaviour as the

required solution. This means that throughout the region $r_1 < r < r_m$ (where r_m denotes the grid point nearest the 'outer turning point) the outward integration is numerically stable in the sense that it does not induce any growth of unwanted solutions or lead to any significant accumulation of rounding error.

Beyond this point the required solution is a decreasing function (the 'tail') and would always be swamped by the increasing complementary function if the outward integration were continued. This difficulty may be avoided by determining the solution at some large value of r from an asymptotic expansion, for example, and integrating inwards, matching the forward solution at r_m. It is avoided more simply by the procedure to be described, a method which was first demonstrated in this context by Froese (1962). The method is further detailed as a worked example in Fox and Mayers (1968, Ch.10,§§65-68). The system of equations in the tail region is treated as a boundary-value problem which is solved by a process of forward and backward integration, with the advantage that the outer limit need not be specified in advance (assuming all arrays have been declared sufficiently large to accommodate the range over which the integration is likely to proceed).

The relations (2.12) centred on the points $r_{m+1}, r_{m+2}, \ldots$ are considered as a set of simultaneous linear equations in the unknowns $Q_{m+1}, Q_{m+2}, \ldots$. This set has the following matrix representation:

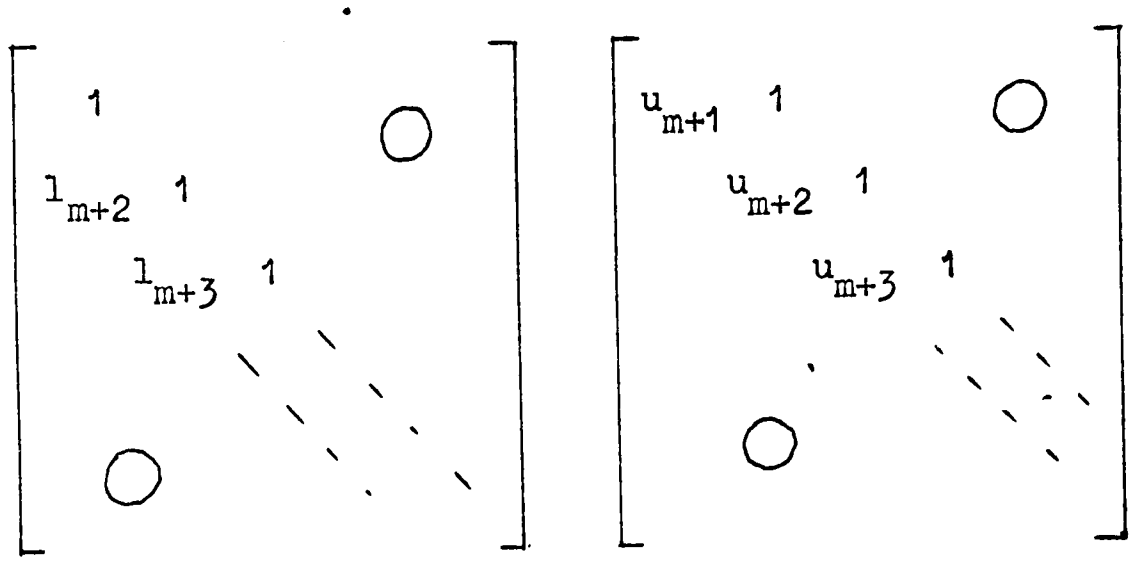
$$\begin{bmatrix} -a_{m+1} & 1 & & & \\ 1 & -a_{m+2} & 1 & & \\ & 1 & -a_{m+3} & 1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} Q_{m+1} \\ Q_{m+2} \\ Q_{m+3} \\ \vdots \end{bmatrix} = \begin{bmatrix} -Q_m \\ 0 \\ 0 \\ \vdots \end{bmatrix} \qquad (2.16)$$

where Q_m has been obtained as the final value in the outward integration.

In the equivalent matrix notation

$$A\,q = b$$

the matrix A may be decomposed into the following matrix product LU:



(2.17)

where

$$u_{m+1} = -a_{m+1} \ ;$$

$$\left.\begin{aligned} l_i &= 1/u_{i-1} \\ u_i &= -a_i - l_i \end{aligned}\right\} \quad i = m+2, m+3, \ldots \ .$$

Since $a_i > 2$ for all $i > m$, it follows that $|l_i| < 1$ and $|u_i| > 1$ for each value of i, so that the factorization of A in this manner forms a stable procedure. [It may be noted that A cannot be singular, since it is both irreducibly and strictly diagonally-dominant.]

The resulting matrix equation

$$L\,U\,q = b$$

can be solved in two stages. Introducing an auxiliary vector z (with components z_i ; $i = m+1, m+2, \ldots$), the system of equations represented by $Lz = b$ is first solved for z by forward recurrence and the vector q is then obtained as the solution of the system $Uq = z$ by backward recurrence. In the first stage the components of z are given by

$$z_{m+1} = -Q_m \ ; \quad z_i = -l_i z_{i-1} \ , \quad i = m+2, m+3, \ldots \ .$$

This forward recurrence is stopped when the magnitude of $(r_i^{1/2} z_i)/(b_i u_i)$ falls below a prescribed tolerance. Denoting this point as $r = r_n$, the components of q are now obtained by recurring backwards in the system $Uq = z$, to give

$$Q_n = z_n/u_n \; ; \quad Q_i = (z_i - Q_{i+1})/u_i \; , \quad i = n-1, n-2, \ldots, m+1 .$$

Since $|l_i| < 1$ and $|u_i| > 1$ for all $i > m$, this method of forward and backward integration will automatically produce the required decreasing solution, stopping where the wave function has become 'insignificant', since $P_n = r_n^{1/2} Q_n / b_n = (r_n^{1/2} z_n)/(b_n u_n)$.

Having so far ensured complete stability throughout the computation, it remains to match the solutions on either side of the outer turning point r_m. This point is the only one at which equation (2.12) is not yet satisfied, and substitution of the computed values will yield a residual non-zero quantity R satisfying

$$T_m(Q) = R. \tag{2.18}$$

This residue is a function of the value of E used as initial data in the computation and will not vanish until E is adjusted to be an exact eigenvalue of the algebraic system for the particular step size h. Newton's method gives $-R/(\frac{\partial R}{\partial E})$ as a first correction to E, where

$$\frac{\partial R}{\partial E} = T_m(\frac{\partial Q}{\partial E}) - (\frac{\partial a}{\partial E})_m Q_m$$

with $(\frac{\partial a}{\partial E})_i = (h r_i / b_i)^2$. The determination of $\frac{\partial R}{\partial E}$ therefore entails solution of the 'variational system' satisfied by the function $\frac{\partial Q}{\partial E}$. This system may be solved in the respective ranges by exactly the same methods as for the original system, using the recurrence

$$T_i(\frac{\partial Q}{\partial E}) = Q_i (h r_i / b_i)^2 . \tag{2.19}$$

Since equations (2.12) and (2.19) share the same coefficients, the solution of (2.19) involves the same matrix A in the tail region and the same decomposition (2.17). The starting values for the recurrence are obtained from equation (2.14) by neglecting the dependence of the coefficients α_j on E, so that for i = 1,2 :

$$\left(\frac{\partial Q}{\partial E}\right)_i = A\left(\frac{\partial b}{\partial E}\right)_i r_i^{l+1/2} \sum_j \alpha_i r_i^j$$

$$= -\frac{h^2 r_i^2 Q_i}{12 b_i} .$$

Using the newly corrected value of E, the entire computation is now repeated to yield a new solution-vector Q and a smaller value of the residual quantity R. The whole process is repeated as many times as necessary to reduce R below a specified value, and using Newton's method successive iterations would be expected to exhibit quadratic convergence. In practice, successive estimates are checked to ensure that the value of E lies within the finite range of the discrete eigenspectrum and that the eigenfunction possesses the requisite number of nodes, and adjustment made to the estimate of E as appropriate. Given a good initial estimate, additional adjustments of this sort are rarely necessary unless the problem corresponds to a very loosely bound physical state.

There remains only a simple normalization so as to satisfy the condition (2.4), where the integral is solved analytically in $[0, r_1]$ using the series expansion (2.14), and by numerical quadrature in $[r_1, r_n]$ using Simpson's rule. Ignoring the quadrature error, the defining relations $Q_i \equiv b_i r_i^{-1/2} P_i$ and $dr \equiv r\, dt$ transform (2.4) into

$$S_1 + \sum_{i=1}^{n} w_i (r_i Q_i / b_i)^2 = 1 \qquad (2.20)$$

where $\{w_i\}$ is the set of quadrature weights and

$$S_1 \equiv \int_0^{r_1} P^2\, dr$$

$$= A^2 r_1^{2l+3} \left\{ \frac{1}{2l+3} + \frac{2\alpha_1 r_1}{2l+4} + \frac{(\alpha_1^2 + 2\alpha_2) r_1^2}{2l+5} + \ldots \right\}. \qquad (2.21)$$

2.3 The inhomogeneous equation

In most cases of physical interest, the radial equations for the self-consistent field (Chapter 6) are not of the homogeneous type represented by equation (2.1) but take the form

$$\frac{d^2p(r)}{dr^2} - [2f(r) + \epsilon]p(r) = x(r) \qquad (2.22)$$

where $f(r)$ is defined by equation (2.2) and $x(r)$ is a given function belonging to the space of eigenfunctions of the discrete spectrum of the homogeneous system; it is implied that $x(r)$ vanishes as $r \to \infty$ and that its series expansion may be written in the form

$$x(r) = r^{l+1}[x^{(0)} + x^{(1)}r + x^{(2)}r^2 + \ldots]. \qquad (2.23)$$

A solution of equation (2.22) belonging to the same function-space exists for any value of the parameter ϵ within the range of the discrete eigenspectrum of the homogeneous system but excluding the point set of eigenvalues $\{\epsilon_i\}$. The inhomogeneous problem is therefore not strictly of eigenvalue type, but a solution satisfying a specified normalization condition will exist only for certain discrete values of ϵ. This may be shown by expanding the solution $p(r)$ and the function $x(r)$ in terms of the complete orthogonal set of discrete eigenfunctions $p^{(i)}(r)$ of the homogeneous system:

$$p(r) = \sum_{i=1}^{\infty} \eta_i p^{(i)}(r) \; ; \qquad x(r) = \sum_{i=1}^{\infty} \beta_i p^{(i)}(r) \; . \qquad (2.24)$$

Substitution into equation (2.22) and use of equation (2.1) gives

$$\sum_{i=1}^{\infty} \{ (\epsilon_i - \epsilon)\eta_i - \beta_i \} p^{(i)}(r) = 0 \; .$$

It follows from the linear independence of the functions $p^{(i)}(r)$ that $(\epsilon_i - \epsilon)\eta_i = \beta_i$ for all i. Hence

$$p(r) = \sum_{i=1}^{\infty} \frac{\beta_i p^{(i)}(r)}{(\epsilon_i - \epsilon)}$$

and, assuming the $p^{(i)}(r)$ to be separately normalized,

$$\int_0^\infty p^2(r)\,dr = \sum_{i=1}^{\infty} \frac{\beta_i^2}{(\epsilon_i - \epsilon)^2} \qquad (2.25)$$

$$= N(\epsilon)\,, \text{ say.}$$

For a given set of coefficients $\{\beta_i\}$, the integral $N(\epsilon)$ varies with ϵ in the manner shown typically in Figure 2.3.

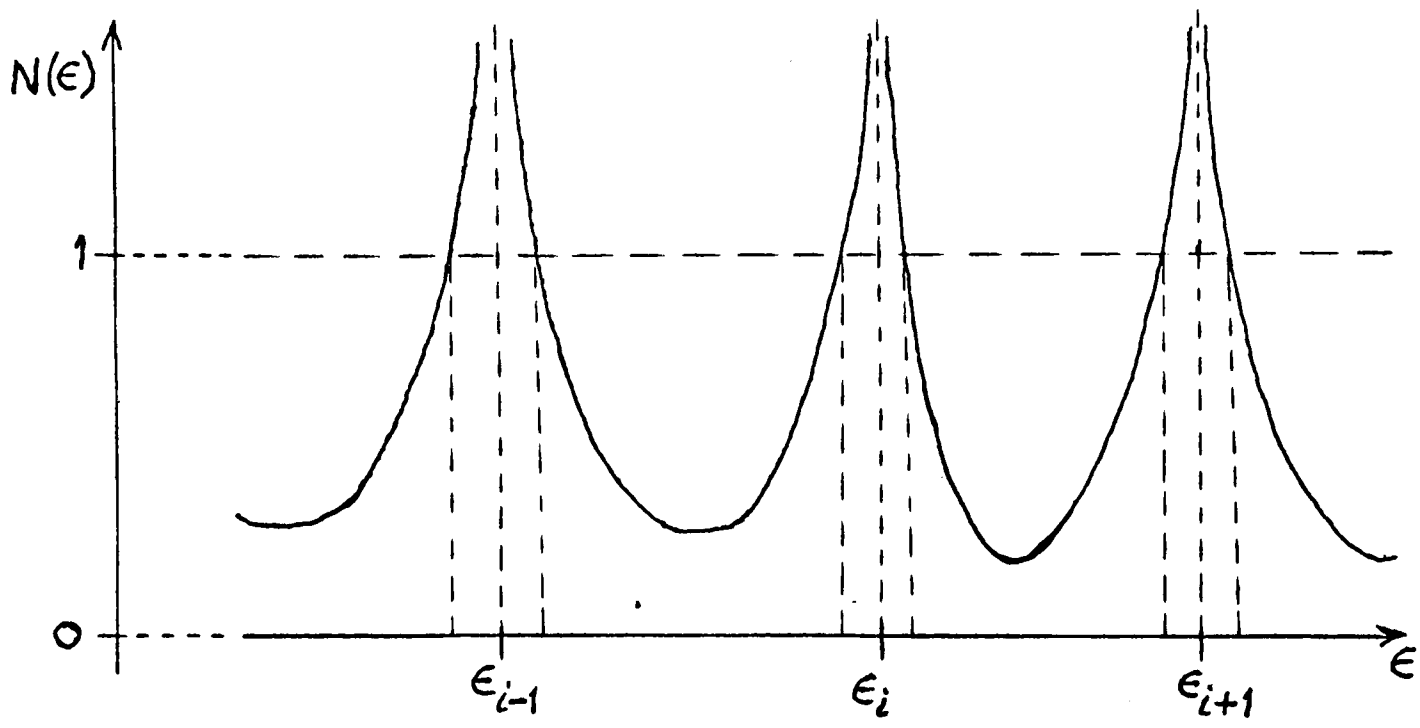


Fig. 2.3

If the function $x(r)$ is of very small magnitude throughout the range $r > 0$, the values of ϵ for which $N(\epsilon)$ has a prescribed value (for example, unity) will occur in pairs centred on each of the eigenvalues ϵ_i. In the limit as $x(r) \to 0$ for all $r > 0$ these values converge on to the point set $\{\epsilon_i\}$, each of the pair of corresponding solutions becoming identical except in sign and forming the set of eigenfunctions $\{\pm p^{(i)}(r)\}$ and $N(\epsilon)$ reducing to zero everywhere except over the set $\{\epsilon_i\}$ where it is undefined. As $|x(r)|$ increases, the values of ϵ satisfying $N(\epsilon) = 1$ will diverge from $\{\epsilon_i\}$ until eventually the situation is reached where no real solution exists satisfying $N(\epsilon) = 1$.

One way of solving equation (2.22) would be to obtain several solutions for different values of ϵ and then interpolate for the solution satisfying

the normalization condition. It is usually the case however that $x(r)$ is not very large, so that values of ϵ close to the final result will also be close to the corresponding eigenvalue of the homogeneous problem and small changes in ϵ will induce large fluctuations in the solution. For this reason it is better to apply the method already developed for the homogeneous system, starting with approximations to ϵ and to the scaling factor A of equation (2.14) and subsequently correcting these values by Newton's method after solving the variational systems.

In the notation of §2.2, equation (2.22) is transformed by the substitution of equations (2.2), (2.5), and (2.6) into

$$\frac{d^2}{dt^2}\rho + \theta\rho = r^{3/2}x .$$

Applying Numerov's method, equation (2.22) is further transformed into an equation of the form (2.11) with the addition of a function g_i on the right-hand side, where

$$g_i \equiv g(r_i) = \frac{h^2}{12}\left[r_{i+1}^{3/2}x_{i+1} + 10r_i^{3/2}x_i + r_{i-1}^{3/2}x_{i-1}\right]. \qquad (2.26)$$

The difference approximation, namely

$$T_i(Q) = G_i , \qquad (2.27)$$

where G_i approximates $g(r_i)$, is solved in similar manner to the homogeneous equation (2.12), starting as before with the first few terms of the series expansion (2.14). The coefficients in this expansion are found by substituting equations (2.3), (2.14), and (2.23) into (2.22), to give:

$$j(2l+1+j)\alpha_j = E\alpha_{j-2} - 2\sum_{k=0}^{j-1} v^{(k)}\alpha_{j-k-1} + A^{-1}x^{(j-2)}$$

for $j \geq 1$, where $\alpha_0 \equiv 1$, $\alpha_{-1} \equiv 0$, and $x^{(-1)} \equiv 0$.

In contrast to the homogeneous case the scaling factor A is not arbitrary for a specified normalization but must be adjusted, together with the estimate of E, to reduce to negligible values the magnitudes of R in the match-point equation

$$T_m(Q) = G_m + R \tag{2.28}$$

and of a residual quantity S in the normalization equation

$$S_1 + \sum_{i=1}^{n} w_i (r_i Q_i / b_i)^2 = 1 + S \tag{2.29}$$

[compare equation (2.20)].

Corrections $(\Delta E, \Delta A)$ are obtained to first order by solving the pair of equations

$$\begin{cases} 0 = R + \Delta E \frac{\partial R}{\partial E} + \Delta A \frac{\partial R}{\partial A} \\ 0 = S + \Delta E \frac{\partial S}{\partial E} + \Delta A \frac{\partial S}{\partial A} \end{cases}$$

where values are derived for the partial derivatives by computing the corresponding variational systems. Thus $\frac{\partial R}{\partial E}$ is obtained from the set of equations (2.19) as in the homogeneous case and $\frac{\partial R}{\partial A}$ from the system

$$T_i\left(\frac{\partial Q}{\partial A}\right) = \delta_{im} \frac{\partial R}{\partial A} \tag{2.30}$$

(where δ_{im} is the Kronecker delta) with starting values $\left(\frac{\partial Q}{\partial A}\right)_i = Q_i/A$ $(i = 1,2)$. The values of $\frac{\partial S}{\partial E}$ and $\frac{\partial S}{\partial A}$ are found by differentiation in equation (2.29), to give

$$2 \sum_{i=1}^{n} w_i (r_i/b_i)^2 Q_i \left\{ \left(\frac{\partial Q}{\partial E}\right)_i + \frac{h^2 r_i^2 Q_i}{12 b_i} \right\} = \frac{\partial S}{\partial E} \tag{2.31}$$

and

$$2 S_1/A + 2 \sum_{i=1}^{n} w_i (r_i/b_i)^2 Q_i \left(\frac{\partial Q}{\partial A}\right)_i = \frac{\partial S}{\partial A}, \tag{2.32}$$

where $\left(\frac{\partial b}{\partial E}\right)_i$ has been replaced by the equivalent expression $-h^2 r_i^2/12$, and it has been assumed that the coefficients α_j in equation (2.14) are only negligibly dependent on E and A so that $\partial S_1/\partial E$ may be neglected and $\partial S_1/\partial A$ replaced by $2 S_1/A$.

The correction process and solving procedure are iterated until further corrections to E and A become insignificant (in relative or absolute terms).

It is of interest to note that the gradient of the curve shown in Figure 2.3 may be obtained for the algebraic system from the final values of the partial derivatives after convergence to a solution. In fact it is evident from equation (2.29) that the gradient, which will be denoted by $\frac{dN}{dE}$, is the total derivative of S with respect to E, subject to the value of R remaining zero. That is,

$$\frac{dN}{dE} = \frac{\partial S}{\partial E} + \frac{\partial S}{\partial A}\frac{\partial A}{\partial E}$$

subject to the condition

$$0 = \frac{\partial R}{\partial E} + \frac{\partial R}{\partial A}\frac{\partial A}{\partial E} ,$$

and therefore

$$\frac{dN}{dE} = \frac{\partial S}{\partial E} - \left[\frac{\partial R}{\partial E} \Big/ \frac{\partial R}{\partial A} \right] \frac{\partial S}{\partial A} .$$

This gives the gradient at a point corresponding to the computed value of E.

CHAPTER 3

DEVELOPMENT OF ERROR ESTIMATES

By regarding the exact system as a perturbation of the system actually solved and the errors themselves as forming the solution of a finite-difference system, the methods described in this chapter lead to the determination of error bounds in accordance with the introductory remarks made in Chapter 1. The discussion is based on that of Chapter 2 but the treatment is in a slightly more general form in order to demonstrate more clearly the underlying ideas.

3.1 General analysis - homogeneous system

The finite-difference approximation to the homogeneous system, represented by the set of equations (2.12) and (2.18), may be written in matrix notation as follows (assuming homogeneous boundary conditions):

$$T(\Phi,E)\ Q = R\ e_m \tag{3.1}$$

where Q and Φ are column vectors of components Q_i and Φ_i, and $T(\Phi,E)$ is a symmetric tridiagonal $n \times n$ matrix whose i^{th} row has the non-zero elements $1, -a_i, 1,$ the coefficient a_i being a function of Φ and E through the relations (2.13). The vector e_m is that column of the unit matrix whose non-zero element lies in the row corresponding to the match-point equation (2.18), and R is the residual quantity which is assumed reduced to zero by the numerical solving process. That is,

$$T(\Phi,E)\ Q = 0 \tag{3.2}$$

so that the matrix of coefficients is then singular. This implies that the corresponding variational system [compare equation (2.19)]

$$T(\Phi,E)\ \frac{\partial Q}{\partial E} + \frac{\partial T}{\partial E}\ Q = \frac{\partial R}{\partial E}\ e_m \tag{3.3}$$

which is essentially inhomogeneous, cannot have a solution unless $\frac{\partial R}{\partial E}$ is non-zero. The vector $\frac{\partial Q}{\partial E}$ is therefore discontinuous in the sense that its

components cannot satisfy at the match-point the recurrence satisfied at all other grid points. It follows that convergence to a solution would not be obtained if the eigenvalue E were of multiplicity greater than 1, since in that case $\frac{\partial R}{\partial E}$ would vanish. As indicated in §2.1, none of the eigensolutions with which this thesis is concerned is degenerate, but the convergence (and ultimate accuracy attained) is related nevertheless to the nature of the eigenspectrum. Thus the values of $|\frac{\partial R}{\partial E}|$ at the eigensolutions [that is, at the nodes of R(E)] increase with the spacing between them, whereas in the case of near-degeneracy (when two or more nodes almost coincide) the values at these solutions become very small. The character of the spectrum is therefore reflected in these values and they in turn affect the rates of convergence to the eigensolutions.

It may be noted in passing that the derivation of strict 'a posteriori' error bounds on eigensolutions of normal matrices depends explicitly on at least some knowledge of the spacing between eigenvalues [compare Wilkinson (1965), Ch.3, §§52 et seq.; Isaacson and Keller (1966), Ch.4, §1.1]. By contrast, the order-of-magnitude error estimates which are the concern of the present analysis will be found to depend only in an implicit way on the nature of the spectrum.

It is evident from equation (2.11) that the solution of the exact differential system can be thought of as satisfying a set of linear difference equations which, for the purpose of comparison with the set of equations (3.2), may be written in the form

$$T(\varphi,\epsilon)\ q + C\ q = 0 \qquad (3.4)$$

where q represents a vector of components $q_i = q(r_i)$, and the matrix C, when acting on q, reproduces the local truncation error for each equation of the system.

In comparing the sets of equations (3.2) and (3.4), the differences between exact and calculated values of the quantities appearing in the

relations (2.13) will be denoted over the set of grid points as follows:

$$\left.\begin{aligned} \delta p_i &= p(r_i) - P_i \\ \delta q_i &= q(r_i) - Q_i \\ \delta \epsilon &= \epsilon - E \\ \delta v_i &= v(r_i) - V_i \\ \delta \varphi_i &= \varphi(r_i) - \Phi_i = 2r_i^2 \delta v_i \\ \delta \theta_i &= \theta(r_i) - \Theta_i = \delta \varphi_i - r_i^2 \delta \epsilon \end{aligned}\right\} . \qquad (3.5)$$

The symbols without their subscripts will be used for denoting the corresponding vectors of grid values. Note that δv, or $\delta \varphi$, represents a vector of errors in the given data ('physical' errors) whereas the other vectors represent the effects of truncation and physical error on the computed results.

Returning to the system of equations represented by (3.4), the matrix $T(\varphi,\epsilon)$ may be expanded in terms of $\delta\epsilon$ and $\delta\varphi$ to give

$$T(\varphi,\epsilon) = T(\Phi,E) + \delta\epsilon \frac{\partial T}{\partial E} + F + O(\delta\epsilon,\delta\varphi)^2 \qquad (3.6)$$

where F is a diagonal matrix whose i^{th} diagonal element is $-\left(\frac{\partial a}{\partial \Phi}\right)_i \delta\varphi_i$.

If the derivative in equation (2.7) had been approximated not by using (2.8) but by the less accurate formula which results simply from truncating its difference expansion after the first term, the matrix T would have been symmetric and its elements linear in E and in the elements of Φ. The linearity cannot be maintained if T is to remain symmetric when the more accurate Numerov approximation is used, and linearity has been sacrificed for symmetry in applying the transformation (2.10).

From here on, expansions of the type (3.6) will be limited to terms of first order in the quantities defined by (3.5). With this restriction, (3.6) transforms equation (3.4) into

$$T(\Phi,E)\, q = -\{C + \delta\epsilon \frac{\partial T}{\partial E} + F\} q \qquad (3.7)$$

which, by comparison with (3.2), yields the following system satisfied by the error vector δq:

$$T(\Phi,E)\,\delta q \;=\; -\{C + \delta\epsilon\frac{\partial T}{\partial E} + F\}q\ . \qquad (3.8)$$

An expression for $\delta\epsilon$ can be gained immediately by making use of the symmetry of T. In fact, premultiplying equation (3.8) by Q^t (where t denotes the transpose), it follows from equation (3.2) that

$$Q^t\{C + \delta\epsilon\frac{\partial T}{\partial E} + F\}q \;=\; 0$$

from which is obtained

$$\delta\epsilon \;=\; -\frac{Q^t\{C + F\}q}{Q^t\frac{\partial T}{\partial E}q}\ . \qquad (3.9)$$

This accords with the general result of perturbation theory for self-adjoint systems that the first-order correction to a non-degenerate eigenvalue is essentially the expectation value of the perturbation in the unperturbed state and is therefore not explicitly dependent on the first-order correction to the eigenfunction or on the proximity of other eigen-solutions.

In practice, the vector q in equation (3.9) may be replaced by Q so as to yield a calculable 'a posteriori' first-order estimate for $\delta\epsilon$. [As regards the dependence on truncation error, this is now equivalent to a one-step deferred correction.] In line with these approximations, the assumption of homogeneous boundary conditions made purely for convenience in writing equation (3.1) is seen not to invalidate equation (3.9), since the effect of any non-zero function values outside the limits of integration will not be to make any first-order contribution to the implied summations in (3.9).

The determination of error estimates from equations (3.8) and (3.9) will be considered further after corresponding results have been obtained for the inhomogeneous system.

3.2 General analysis - inhomogeneous system

The inhomogeneous finite-difference system, represented by the set of

equations (2.27) and (2.28) with homogeneous boundary conditions and normalization condition (2.29), may be represented as follows in matrix notation:

$$\left.\begin{aligned} T(\Phi,E)\,Q &= G + R e_m \\ Q^t W Q &= 1 + S \end{aligned}\right\} \qquad (3.10)$$

where G is the vector of elements G_i and W is a diagonal matrix whose elements are obtainable from equation (2.29). When the residual quantities R and S are reduced to zero, the system becomes

$$\left.\begin{aligned} T(\Phi,E)\,Q &= G \\ Q^t W Q &= 1 \end{aligned}\right\} . \qquad (3.11)$$

In the inhomogeneous case, two variational systems are defined [compare equations (2.19), (2.30), (2.31), and (2.32)]:

$$\left.\begin{aligned} T(\Phi,E)\,\frac{\partial Q}{\partial E} + \frac{\partial T}{\partial E} Q &= \frac{\partial R}{\partial E} e_m \\ 2 Q^t W \frac{\partial Q}{\partial E} &= \frac{\partial S}{\partial E} \end{aligned}\right\} \qquad (3.12)$$

$$\left.\begin{aligned} T(\Phi,E)\,\frac{\partial Q}{\partial A} &= \frac{\partial R}{\partial A} e_m \\ 2 Q^t W \frac{\partial Q}{\partial A} &= \frac{\partial S}{\partial A} \end{aligned}\right\} . \qquad (3.13)$$

[The dependence of the elements of W on E, which has been neglected here for the sake of simplicity, will be taken into account in the analysis of specific problems.]

For comparison with the system (3.10), the discretized version of the exact differential system may be written in the form:

$$\left.\begin{aligned} T(\varphi,\epsilon)\,q + Cq &= g \\ q^t W q + C_s &= 1 \end{aligned}\right\} \qquad (3.14)$$

where C is defined as in the homogeneous equation (3.4), g is the vector of compononts g_i defined in (2.26), and C_s represents the quadrature error which results from expressing the normalization condition (2.4) for the exact solution in the form (2.29). The expansion (3.6) transforms the first of equations (3.14) into

§3.2

$$T(\Phi,E)\,q \;=\; g - \{C + \delta\epsilon\frac{\partial T}{\partial E} + F\}q\,. \qquad (3.15)$$

Comparing this with the first of equations (3.11) and denoting $g(r_i) - G_i$ by δg_i, the following is obtained for the error vector:

$$T(\Phi,E)\,\delta q \;=\; \delta g - \{C + \delta\epsilon\frac{\partial T}{\partial E} + F\}q\,. \qquad (3.16)$$

Equations (3.15) and (3.16) correspond in the inhomogeneous case to the homogeneous equations (3.7) and (3.8).

Premultiplying equation (3.16) by Q^t and using the first of equations (3.11), it follows from the symmetry of T that

$$G^t\delta q \;=\; Q^t\delta g - Q^t\{C + \delta\epsilon\frac{\partial T}{\partial E} + F\}q$$

and therefore

$$G^t\delta q + \delta\epsilon\{Q^t\frac{\partial T}{\partial E}q\} \;=\; Q^t(\delta g - \{C + F\}q)\,. \qquad (3.17)$$

It is clear that in this case $\delta\epsilon$ cannot be estimated independently of δq, in contrast to the situation in the homogeneous case represented by equation (3.9).

3.3 Errors in eigenvalue and scale factor

Each iteration of the procedure described in Chapter 2 for obtaining an approximation to the eigenfunction by solving the set of equations represented by (3.1) or the first of (3.10) uses as parameters the currently held values of E and the scaling factor A. By analogy, the error vector δq defined by the set of equations (3.8) or (3.16) may be regarded as depending in part on the errors $\delta\epsilon$ and δA in E and A as if these errors arose independently, when in fact, of course, all the error quantities have their origin in the same truncation and physical errors. In furtherance of this point of view, $\delta\epsilon$ and δA may be considered to form independent variables in terms of which the vector δq may be expanded in a power series:

$$\delta q \;=\; \delta q^{(0)} \;+\; \delta\epsilon\frac{\partial Q}{\partial E} \;+\; \delta A\frac{\partial Q}{\partial A} \;+\; O(\delta\epsilon,\delta A)^2\,. \qquad (3.18)$$

Here, $\frac{\partial Q}{\partial E}$ satiafies (3.3) or (3.12), while in the inhomogeneous case $\frac{\partial Q}{\partial A}$ satisfies (3.13) whereas the homogeneous system has $\frac{\partial Q}{\partial A} = A^{-1}Q$. The zero-order component, $\delta q^{(0)}$, satisfies

$$T(\Phi,E)\,\delta q^{(0)} = \delta g - (C + F)\,Q - (\delta\epsilon \frac{\partial R}{\partial E} + \delta A \frac{\partial R}{\partial A})\,e_m \qquad (3.19)$$

where δg is absent and $\frac{\partial R}{\partial A}$ zero in the homogeneous case. This equation follows from (3.8) or (3.16) (with vector q replaced by Q) upon substituting expansion (3.18) and abstracting zero-order terms.

It is intended to use the series (3.18) for estimating bounds on the magnitudes of the components of δq by separately bounding $|\delta\epsilon|$, $|\delta A|$, and the components of $|\delta q^{(0)}|$ (where $|\delta q^{(0)}|$ denotes a vector of components $|\delta q_i^{(0)}|$).

In the homogeneous case, an estimate of $|\delta\epsilon|$ may be made directly from equation (3.9), but the corresponding inhomogeneous expression (3.17) for $\delta\epsilon$ involves δq as well. However, if expansion (3.18) is substituted the dependence is reduced to that of $\delta\epsilon$ on δA and $\delta q^{(0)}$:

$$\begin{aligned} 0 = {} & \{\, Q^t[\delta g - (C + F)\,q] - G^t \delta q^{(0)} \,\} \\ & - \delta\epsilon\,(Q^t \frac{\partial T}{\partial E}\,q + G^t \frac{\partial Q}{\partial E}) - \delta A\,(G^t \frac{\partial Q}{\partial A}). \end{aligned} \qquad (3.20)$$

In both homogeneous and inhomogeneous cases, a further relation between $\delta\epsilon$ and δA may be obtained to first order by comparing the conditions for normalization of the exact and computed solutions, represented respectively by the second of equations (3.14) and of (3.11). These combine to form

$$2\,Q^t W\,\delta q + C_s = 0$$

which, on substituting (3.18), becomes

$$0 = (Q^t W\,\delta q^{(0)} + \tfrac{1}{2} C_s) + \delta\epsilon\,(Q^t W \frac{\partial Q}{\partial E}) + \delta A\,(Q^t W \frac{\partial Q}{\partial A}). \qquad (3.21)$$

In the homogeneous case, where the coefficient of δA reduces to A^{-1}, estimates of $|\delta\epsilon|$ and $|\delta A|$ are obtainable from equations (3.9) and (3.21),

supposing bounds to be known on the magnitudes of C_s and of the components of Cq, $\delta\varphi$, and $\delta q^{(0)}$. In the inhomogeneous case, where the coefficients of $\delta\epsilon$ and δA in equation (3.21) are given by the second of equations (3.12) and of (3.13), the corresponding estimates are provided by solving equations (3.20) and (3.21) for $\delta\epsilon$ and δA.

3.4 Error in eigenfunction

The remaining task is the estimation of a realistic bound on the magnitude of each component of the vector $\delta q^{(0)}$ by considering the boundary-value system (3.19). Using the first of the defining relations (2.13), the i^{th} equation (for $i \neq m$) of the set (3.19) may be written in the form

$$T_i(y) = b_i \qquad (3.22)$$

where the vector y stands for $\delta q^{(0)}$. Given a strict bound B_i on $|b_i|$ for each value of i, the problem amounts to the estimation of a strict bound Y_i, as realistic as possible, on each component of $|y|$. It will be assumed that Y_i is known for $r \to 0$ (and in particular for the points $i = 1$ and $i = 2$) and for $r > r_n$ (so that Y_{n+1} is assumed known).

Equation (3.22) in the elements y_i forms a linear three-term recurrence whose complementary solutions are of oscillatory type in the classical region [see Figure 2.2] and of increasing or decreasing exponential type elsewhere. Now the method described in Chapter 2 for solving for the eigenfunction itself, involving forward recurrence in the region $r < r_m$ and matrix methods for $r > r_m$, is such that the actual error increases monotonically from the starting position to the inner turning point, then oscillates with gradually increasing amplitude through the classical region, and finally decreases monotonically to zero in the tail region. In order for the computed bound to reflect this behaviour as closely as possible, its estimation will be considered separately in each of these three regions.

§3.4

In the innermost region the function y satisfies an initial-value system comprising the recurrence (3.22) and known values of Y_1 and Y_2. An upper bound on y_i for any $i > 2$ can be found by a forward error analysis of the type described, for example, in Fox and Mayers (1968, Ch.3, §§18,19), in which y_i is expressed as a sum of multiples of y_1, y_2, and the quantities b_j $(j < i)$. The coefficients are readily built up according to the following table:

	y_2	(b_2-y_1)	b_3	b_4	. . .
y_3	a_2	1			
y_4	a_3a_2-1	a_3	1		
y_5	$\{a_4(a_3a_2-1)-a_2\}$	(a_4a_3-1)	a_4	1	
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Thus $y_i = \sum_{j=1}^{i} d_{ij} b'_j$ $(i = 3,4,\ldots)$ where $b'_1 = y_2$, $b'_2 = b_2-y_1$, $b'_j = b_j$ for $j > 2$. The coefficients d_{ij} satisfy the homogeneous recurrence

$$\left.\begin{aligned} d_{i+1,j} - a_i d_{ij} + d_{i-1,j} &= 0 \quad (i > j), \\ d_{j+1,j} = 1, \quad d_{jj} &= 0 \end{aligned}\right\} \quad (j > 0)$$

or alternatively

$$\left.\begin{aligned} d_{i,j-1} - a_j d_{ij} + d_{i,j+1} &= 0 \quad (0 < j < i), \\ d_{i,i-1} = 1, \quad d_{ii} &= 0 \end{aligned}\right\} \quad (i > 2).$$

The required bounds are therefore given by

$$|y_i| \leq Y_i = \sum_{j=1}^{i} |d_{ij}| B'_j \quad \text{for } i = 3,4,\ldots \qquad (3.23)$$

where $B'_1 = Y_2$, $B'_2 = B_2 + Y_1$, $B'_j = B_j$ for $j > 2$.

It should be noted that in this section of the range, where $a_i > 2$, the coefficients d_{ij} are all positive and each bound Y_i is theoretically attainable. Obviously this would not be the case if equation (3.22) were used simply to bound each $|y_i|$ in terms of Y_{i-1}, Y_{i-2}, and B_{i-1}.

In the classical region, where each a_i has a value less than 2, the actual error is an oscillatory function possessing one or more nodes whereas any error bound obtained by simple means would inevitably be monotonically increasing. However, the coefficients d_{ij} defined above satisfy the same recurrence as the solution itself, and the solution in this region is oscillatory in nature. For this reason the continued use of equation (3.23) through the region would be expected to yield bounds which, though not attainable, nevertheless increase less rapidly than in the non-classical region and possibly even decrease over short sections. It will become apparent subsequently that this hypothesis is borne out in practice.

If the forward error analysis is carried through as far as the match-point r_m, there remains to be solved a boundary-value system for the errors in the tail region. This system is comprised of the relation (3.22) for each of the values $i = m+1, m+2, \ldots, n$, together with bounds on the magnitudes of y_m and y_{n+1}. The system may be written in matrix form as

$$A\,\hat{y} = \hat{b}$$

where A is the non-singular matrix of equation (2.16) (terminated at the $(n-m)^{th}$ row and column), and

$$\hat{y} = \begin{bmatrix} y_{m+1} \\ y_{m+2} \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix}, \qquad \hat{b} = \begin{bmatrix} b_{m+1} - y_m \\ b_{m+2} \\ \vdots \\ b_{n-1} \\ b_n - y_{n+1} \end{bmatrix}.$$

A bound on $|y_m|$ has already been found as the final value, Y_m, in the forward analysis; by comparison, $|y_{n+1}|$ has an insignificant value since

the error in the tail region decreases monotonically by several orders of magnitude, in common with the solution itself, and is negligible at the outer limit of integration. It is essential that the computed bound, if it is to be at all realistic, should fall off at the same rate; this rules out its determination by simply bounding some norm of A^{-1}. Since knowledge of $\hat{b}$ is limited to bounds on its components, the best bound obtainable for $\hat{y}$ will be given by

$$|\hat{y}| \leq |A^{-1}||\hat{b}| \tag{3.24}$$

where $|A^{-1}|$ denotes the matrix of coefficients $|(A^{-1})_{ij}|$. The inversion of A is effected most easily by retaining the elements of the factored form L U from the original part of the calculation; then the j^{th} column, $c^{(j)}$, of A^{-1} satisfies $L U c^{(j)} = e_j$ where e_j is the j^{th} column of the unit matrix. The solution is obtained as before in two stages: $L z = e_j$ and $U c^{(j)} = z$.

In general the inverse is a full matrix and it might be thought that the storage requirement imposes an obvious limitation on the applicability of the method, even though it is only temporarily required while the bounds Y_i are obtained from (3.24). However, the size of A is determined by the number of grid points in the tail region and this may be relatively small compared with the total over the whole range, remembering that the density of points varies inversely as the value of r on account of the transformation (2.5) to the logarithmic scale. In the applications of practical interest to be considered, it proves quite feasible to choose a step size such that the storage requirement for using (3.24) is not prohibitively large.

If a situation were to arise in which this no longer remained a practical proposition, a rather poorer bound could always be gained by making use of the monotonicity property of the matrix A. Since each of the coefficients a_i appearing on the diagonal of A is of value greater than 2, it follows that $A \leq H$, where H is the tridiagonal matrix formed from A by

replacing each diagonal element by -2. [The notation implies $(A)_{ij} \leq (H)_{ij}$ for all i,j.] Both A and H are the negatives of monotone matrices and consequently $H^{-1} \leq A^{-1}$, with $A^{-1} \leq 0$. Thus $|A^{-1}| \leq |H^{-1}|$, and from (3.24) is obtained the result

$$|\hat{y}| \leq |H^{-1}||\hat{b}| \tag{3.25}$$

whose advantage over (3.24) is that the elements of H^{-1} may be written down explicitly [compare §7.2.2 of Henrici (1962)]. In fact the $(i,j)^{th}$ element, denoted by η_{ij}, is given by

$$-\eta_{ij} = \begin{cases} \dfrac{(j-m)(n+1-i)}{n+1-m}, & j \leq i, \\ \dfrac{(n+1-j)(i-m)}{n+1-m}, & j \geq i; \end{cases}$$

$$i,j = m+1, m+2, \ldots, n.$$

The bound (3.25) will necessarily be more pessimistic than (3.24) since information about the system, contained in the values of the coefficients a_i, has been discarded. The value of a_{m+1} is bounded away from 2 and successive coefficients satisfy $a_{i+1} > a_i$. Writing $\xi_n = a_n - 2$, it is clear that $H - \xi_n I \leq A$ (where I is the identity matrix). Thus

$$|(H - \xi_n I)^{-1}| \leq |A^{-1}| \leq |H^{-1}|,$$

so that the wider the spread of the coefficients a_i the more pessimistic the bound (3.25).

3.5 Further remarks

It has been shown in §3.4 how it is proposed, by using formulae (3.23) and (3.24), to obtain close bounds on the moduli of the components of the vector $\delta q^{(0)}$. This vector satisfies equation (3.19) and occurs as the first term in expansion (3.18). Truncating this expansion after the terms of first order in $\delta\epsilon$ and δA, a bound on the error in the computed eigenvector is given by

$$|\delta q| \leq |\delta q^{(0)}| + |\delta\epsilon|\left|\frac{\partial Q}{\partial E}\right| + |\delta A|\left|\frac{\partial Q}{\partial A}\right|$$

where bounds on $|\delta\epsilon|$ and $|\delta A|$ are obtainable by the methods described in §3.3. These bounds are all based on the combined effect of truncation error and physical error.

Considering first the effect of truncation, it is well known that the actual error in the computed solution for the type of equation and method of solution described in Chapter 2 is of order h^4. This is indicated for example in Fox (1962), Ch.8, §§17-21 in connection with a discussion of Richardson extrapolation, and is shown by Mayers (1964) to remain valid in spite of the regular singularity of equation (2.1) at the origin. To a good approximation, the computed error estimates should be of the same order in the step size, since their derivation is based on a process closely akin to the performance of one stage of a deferred correction; see Fox (1962), Ch.3, §§22 et seq. In fact the process bears a closer relation to the calculation of a first correction to the computed solution than to the direct calculation of a second approximation, with the advantage, as pointed out in N.P.L. (1961, Ch.9, §14), that fewer significant figures need be retained in performing the calculation. Osborne (1960) discusses the idea of extrapolation in the context of the Sturm-Liouville differential-equation eigenproblem and connects it with the deferred correction process. The latter is considered from a more general viewpoint by Stetter (1965) and Pereyra (1966, 1967, 1968), their aim being to establish under what conditions the repeated application of the process forms a convergent scheme involving the elimination of successive terms in the expansion of the error regarded as a function of h. However, as outlined in Chapter 1, the purpose of the present work is not to pursue these developments in the theory of iterative refinement but simply to establish useful error bounds on the first approximation to the solution for a given step size. To this extent, the exact form of the dependence of the error on h is not of primary importance.

By contrast, the effect of physical error on the computed solution would not automatically be diminished by refining the grid, and even for

completely stable methods of solution the presence of uncertainties in the initial data could impose definite restrictions on the ultimate accuracy attainable. It is in this type of situation that the error-bounding procedures so far developed could provide a useful indication of the propagated effect of such errors in the data. On the basis of this knoowledge could be selected a step size for which the estimated error arising from truncation is of the same magnitude as that resulting from the physical error. This would guard against the use of an unnecessarily small grid; it would ensure on the other hand that the accuracy achieved is limited only by the inherent inaccuracy in the data.

CHAPTER 4

APPLICATION TO THE SCHRODINGER EQUATION

The purpose of this chapter is twofold: to interpret the general ideas of the previous chapter in terms of the specific systems described in Chapter 2, and to test the error-estimating procedures in a few simple cases where it is possible to make a comparison with the actual errors induced in the computed solutions.

4.1 Application of theory

A few comments are called for in applying the results of Chapter 3 to the estimation of error bounds for the homogeneous and inhomogeneous equations of Chapter 2. In the first place, returning to the expansion (3.6), it may be noted that the coefficient $\frac{\partial T}{\partial E}$ is a diagonal matrix whose i^{th} diagonal element is

$$(\frac{\partial T}{\partial E})_{ii} = -(\frac{\partial a}{\partial E})_i = -(hr_i/b_i)^2.$$

The corresponding element of the diagonal matrix F is

$$F_{ii} = -(\frac{\partial a}{\partial \Phi})_i \delta\varphi_i = (h/b_i)^2 \delta\varphi_i .$$

The i^{th} component of the vector Cq in equation (3.4) represents the local truncation error $C_i(r^{-1/2}p)$ of equation (2.11). Replacing p by P in accordance with the comments at the end of §3.1, and using the second of the defining relations (2.13), the i^{th} element of Cq becomes

$$(Cq)_i = C_i(Q/b).$$

Then using the first of relations (2.13), the i^{th} equation in the set (3.8) or (3.16) may be written

$$T_i(\delta q) = K_i + (hr_i/b_i)^2 Q_i \delta\epsilon$$

where

$$K_i \equiv -C_i(Q/b) - (h/b_i)^2 Q_i \delta\varphi_i + \delta g_i , \qquad (4.1)$$

the term δg_i being absent in the homogeneous case.

Equation (3.9) for $\delta\epsilon$ in the homogeneous case is transformed into

$$h^2 \delta\epsilon \sum (r_i Q_i / b_i)^2 + \sum K_i Q_i = 0 \qquad (4.2)$$

where in practice the summation is taken from $i = 1$ to $i = n$ and any slight end-effects ignored. The corresponding inhomogeneous equation (3.20) becomes

$$0 = \sigma_0 + \sigma_1 \delta\epsilon + \sigma_2 \delta A \qquad (4.3)$$

where

$$\sigma_0 = \sum K_i Q_i - \sum G_i \delta q_i^{(0)}$$

$$\sigma_1 = h^2 \sum (r_i Q_i / b_i)^2 - \sum G_i \left(\frac{\partial Q}{\partial E}\right)_i$$

$$\sigma_2 = - \sum G_i \left(\frac{\partial Q}{\partial A}\right)_i .$$

The derivation of equation (3.21) must be considered in slightly greater detail before application. The second of equations (3.11) represents equation (2.20) and the second of equations (3.14) represents the equation

$$S_1' + \sum_{i=1}^{n} w_i r_i p_i^2 + C_s = 1$$

where $S_1' = S_1 + 2S_1 A^{-1} \delta A + 0(\delta A)^2$ with S_1 defined by equation (2.21). Combining these and using the relations (2.10) and (3.5), the result is obtained:

$$0 = \sum w_i r_i^2 \left\{ \frac{(Q_i + \delta q_i)^2}{(b_i + \frac{1}{12} h^2 \delta\Theta_i)^2} - \frac{Q_i^2}{b_i^2} \right\} + 2S_1 A^{-1} \delta A + C_s$$

$$= 2 \sum w_i (r_i / b_i)^2 Q_i \left\{ \delta q_i - \frac{1}{12} h^2 \frac{Q_i}{b_i} (\delta\varphi_i - r_i^2 \delta\epsilon) \right\} + 2S_1 A^{-1} \delta A + C_s .$$

Substitution of (3.18) leads to the equivalent of equation (3.21):

$$0 = \sum w_i (r_i / b_i)^2 Q_i \left\{ \delta q_i^{(0)} - \frac{h^2}{12} \frac{Q_i}{b_i} \delta\varphi_i \right\} + \frac{1}{2} C_s$$

$$+ \delta\epsilon \sum w_i (r_i / b_i)^2 Q_i \left\{ \left(\frac{\partial Q}{\partial E}\right)_i + \frac{h^2}{12} \frac{Q_i}{b_i} r_i^2 \right\}$$

$$+ \delta A \left\{ \sum w_i (r_i / b_i)^2 Q_i \left(\frac{\partial Q}{\partial A}\right)_i + S_1 / A \right\} . \qquad (4.4)$$

In the homogeneous case, where $(\frac{\partial Q}{\partial A})_i = A^{-1}Q_i$, equations (2.20) and (4.2) [with (4.1)] transform this relation into

$$0 = \sum w_i (r_i/b_i)^2 Q_i \delta q_i^{(0)} + \frac{1}{2} C_s + \chi \sum C_i (Q/b) Q_i + h^2 \sum (Q_i/b_i)^2 \{\chi - \frac{1}{12} w_i r_i^2/b_i\} \delta\varphi_i + \frac{\delta A}{A} \qquad (4.5)$$

$$\text{where } \chi \equiv \frac{\sum w_i (r_i/b_i)^2 Q_i \{ (\frac{\partial Q}{\partial E})_i + \frac{h^2}{12} Q_i r_i^2/b_i \}}{h^2 \sum (r_i Q_i/b_i)^2} .$$

The estimated bounds on $|\delta\epsilon|$ and $|\delta A|$ follow from equations (4.2) and (4.5):

$$\left.\begin{aligned} |\delta\epsilon| &\leq \frac{\sum |C_i(Q/b)||Q_i| + h^2 \sum (Q_i/b_i)^2 |\delta\varphi_i|}{h^2 \sum (r_i Q_i/b_i)^2} \\ |\frac{\delta A}{A}| &\leq \sum w_i (r_i/b_i)^2 |Q_i||\delta q_i^{(0)}| + \frac{1}{2}|C_s| \\ &\quad + |\chi| \sum |C_i(Q/b)||Q_i| \\ &\quad + h^2 \sum (Q_i/b_i)^2 |\chi - \frac{1}{12} w_i r_i^2/b_i||\delta\varphi_i| \end{aligned}\right\} \qquad (4.6)$$

(homogeneous case).

By using the relations (2.31) and (2.32), equation (4.4) in the inhomogeneous case reduces to

$$0 = \tau_0 + \tau_1 \delta\epsilon + \tau_2 \delta A \qquad (4.7)$$

where

$$\tau_0 = \sum w_i (r_i/b_i)^2 Q_i \{ \delta q_i^{(0)} - \frac{h^2}{12} \frac{Q_i}{b_i} \delta\varphi_i \} + \frac{1}{2} C_s$$

$$\tau_1 = \frac{1}{2} \frac{\partial S}{\partial E}$$

$$\tau_2 = \frac{1}{2} \frac{\partial S}{\partial A} .$$

Solving equations (4.3) and (4.7) for $\delta\epsilon$ and δA, the estimated bounds are obtained as:

$$\left.\begin{aligned} |\delta\epsilon| &= \frac{N_2}{|\sigma_1\tau_2 - \sigma_2\tau_1|} \\ |\delta A| &= \frac{N_1}{|\sigma_1\tau_2 - \sigma_2\tau_1|} \end{aligned}\right\} \qquad (4.8)$$

(inhomogeneous case)

where, for $j = 1,2$,

$$
\begin{aligned}
N_j &= \left|\tau_j\Big\{\sum K_i Q_i - \sum G_i \delta q_i^{(0)}\Big\} - \sigma_j\Big\{\sum w_i \Big(\frac{r_i}{b_i}\Big)^2 Q_i\Big[\delta q_i^{(0)} - \frac{h^2}{12}\frac{Q_i}{b_i}\delta\varphi_i\Big] + \frac{1}{2}C_s\Big\}\right| \\
&= \Big|\tau_j \sum \{-C_i(Q/b) + \delta g_i\}Q_i - \frac{1}{2}\sigma_j C_s \\
&\quad - \sum \{(h/b_i)^2 Q_i^2 \tau_j - w_i (r_i/b_i)^2 \frac{h^2}{12}\frac{Q_i^2}{b_i}\sigma_j\}\delta\varphi_i \\
&\quad - \sum \{G_i\tau_j + w_i(r_i/b_i)^2 Q_i \sigma_j\}\delta q_i^{(0)}\Big| \\
&\le |\tau_j| \sum \{|C_i(Q/b)| + |\delta g_i|\}|Q_i| + \frac{1}{2}|\sigma_j||C_s| \\
&\quad + h^2 \sum (Q_i/b_i)^2 \Big|\tau_j - w_i \frac{r_i^2}{12 b_i}\sigma_j\Big||\delta\varphi_i| \\
&\quad + \sum |G_i\tau_j + w_i(r_i/b_i)^2 Q_i\sigma_j||\delta q_i^{(0)}| .
\end{aligned}
$$

where the defining relation (4.1) has been substituted for K_i.

Turning to the system of equations (3.19) for $\delta q^{(0)}$, the boundary conditions on $\delta q^{(0)}$ must be given some consideration. The condition at the outer limit of integration, r_n, may be approximated by setting $\delta q_i = 0$ for $i > n$, thereby assuming that the true function has decreased to such a small value that the effect of its being non-zero is negligible compared with other sources of error. The condition as $r \to 0$ is derived from expansion (2.15) by assuming that $\delta q_i^{(0)}$ is non-vanishing because of the possible effect of the physical error $\delta\varphi_i$ on the value of b_i. Thus $\delta q_i^{(0)} = \frac{\delta b_i}{b_i} Q_i = \frac{1}{12} h^2 \frac{Q_i}{b_i}\delta\varphi_i$. Using (4.1) and the first of relations (2.13), the system (3.19) may now be written

$$
\left.
\begin{aligned}
T_i(\delta q^{(0)}) &= K_i \quad \text{for all } i \neq m, \\
\delta q_i^{(0)} &= \frac{1}{12} h^2 \frac{Q_i}{b_i}\delta\varphi_i \quad \text{for } r \to 0, \\
\delta q_i^{(0)} &= 0 \quad \text{for } i > n
\end{aligned}
\right\} \qquad (4.9)
$$

from which, by using the method developed in §3.4, a bound on each component of $|\delta q^{(0)}|$ may be obtained from known bounds on the magnitudes of the functions K_i defined in equation (4.1).

The error estimate for the computed function $P(r_i)$ is finally obtained by comparing (2.10) with the equivalent relation in (2.13). Then, to first order in errors,

$$\delta p_i = \frac{r_i^{1/2}}{b_i}\{\delta q_i - \frac{1}{12}h^2\frac{Q_i}{b_i}\delta\theta_i\}$$

$$= \frac{r_i^{1/2}}{b_i^2}\{b_i\delta q_i - \frac{h^2}{12}Q_i\delta\varphi_i + \frac{h^2}{12}Q_i r_i^2\delta\epsilon\}$$

and, by using expansion (3.18),

$$\delta p_i = \frac{r_i^{1/2}}{b_i^2}\{b_i\delta q_i^{(0)} - \frac{h^2}{12}Q_i\delta\varphi_i + \delta\epsilon[b_i(\frac{\partial Q}{\partial E})_i + \frac{h^2}{12}Q_i r_i^2] + b_i(\frac{\partial Q}{\partial A})_i\delta A\}$$

so that

$$|\delta p_i| \le \frac{r_i^{1/2}}{b_i^2}\{b_i|\delta q_i^{(0)}| + \frac{h^2}{12}|Q_i||\delta\varphi_i| + |\delta\epsilon||b_i(\frac{\partial Q}{\partial E})_i + \frac{h^2}{12}Q_i r_i^2| + b_i|\delta A||(\frac{\partial Q}{\partial A})_i|\} \tag{4.10}$$

where $(\frac{\partial Q}{\partial A})_i = A^{-1}Q_i$ in the homogeneous case.

The use of these results requires that bounds be set on the magnitudes of the components of K_i [equation (4.1)]. Among these, the term $\delta\varphi_i$ defined by (3.5) gains its bound from an estimate of the likely error in the value V_i of the potential used as initial data for the problem. Similarly, for the inhomogeneous system, a bound on δg_i may be estimated through equation (2.26) by considering the accuracy of the given function $x(r)$. It has been assumed throughout that these bounds are sufficiently small to ensure the validity of retaining only first-order terms in expansions such as (3.6).

The remaining component of K_i represents an approximation to the local truncation error which is associated with equation (2.12) or (2.27) and was first defined in association with equation (2.8). Using the representation (2.9) involving sixth and higher differences of the computed function, it

is clearly essential that the step size be sufficiently small for the function to possess converging differences throughout the range. Indeed this assumption is implicit in the use of any formula such as (2.8) for approximating to the derivative by means of a truncated difference expansion, and is equivalent to assuming that the function could be represented to the desired accuracy over the whole range by a set of interlacing polynomials. In line with the approximations made so far, only the first term in expansion (2.9) will be included in the estimation of a bound on the local truncation error, the contribution from the higher differences being neglected.

For the justification and success of this procedure it is extremely important to examine the difference table of the function when tabulated to the required number of decimal places and if necessary to adjust the step size until the differences converge. Admittedly, however, it would be difficult to develop an automatic procedure which achieved a satisfactory investigation of this sort, and this is not attempted here. As regards the storage of differences, the requirement is only temporary and in any case not as great as that of the matrix inversion associated with the use of equation (3.24). No difficulty is encountered in calculating extra function values to obtain central differences at the ends of the range; at the inner limit the series expansion (2.15) is valid for all r in $(0, r_1)$ while the use of a logarithmic grid excludes the possibility of overshooting or even closely approaching the singular point $r = 0$; at the outer limit the forward and backward integration of equations (2.16) may be extended for a few points beyond r_n or alternatively the function values may simply be set to zero at these points.

The quadrature error C_s appearing in equation (4.4) arises from the use of Simpson's rule in approximating to the integral in the normalization condition (2.4); for a sufficiently smooth integrand the error may be expanded in terms of central differences: see Fox (1962), Ch.1, §7.

The first term is $-\frac{h}{90}\sum'\delta^4 z_i$, where $z_i = r_i P_i^2 = (r_i Q_i/b_i)^2$ and $\sum'$ denotes a summation over alternate values of i. Expanding the fourth difference, rearranging the summation, and ignoring end-effects, a first approximation to the quadrature error is given as

$$C_s \doteq \frac{4}{45} h \sum_{i=1}^{n} (-1)^i z_i$$

from which an estimate of its magnitude is obtained as

$$|C_s| \approx \frac{4}{45} h \left| \sum_{i=1}^{n} (-1)^i z_i \right| .$$

This error is extremely small compared with other sources of error. For instance, its contribution to the bound on $|\delta A|$ calculated from the second of equations (4.6) is never more than 0.1% in the cases for which results are listed in Table 4.1 (see next section).

4.2 Non-relativistic Coulomb functions

A convenient test of the analysis developed so far is provided by applying it to the calculation of the bound states of the single-electron atomic system (for example, the hydrogen atom) for which the exact solution is well known and easily determined analytically.

For a one-electron atom of nuclear charge Z, the central field may be ascribed to the Coulomb potential

$$v(r) = Z/r \qquad (4.11)$$

which assumes that the nuclear charge is localized at the point r = 0. By considering the asymptotic form of equation (2.1) for large values of r, the solution valid for all r is obtained in the form

$$p(r) = f(r)\, e^{-r\sqrt{2\lambda}}$$

where f(r) is a function which varies slowly at infinity. Expanding f(r) as a power series in r,

$$f(r) = r^{\gamma} \sum_{s=0}^{\infty} a_s r^s \qquad (a_0 \neq 0)$$

[compare equation (2.14)], substituting into equation (2.1), and equating coefficients of successive powers to zero, the value $\gamma = l+1$ is obtained, together with the set of relations:

$$a_{s+1}\{(s+l+2)(s+l+1) - l(l+1)\} = 2a_s\{\sqrt{2\lambda}\,(s+l+1) - Z\} \quad \text{for } s \geq 0 .$$

This series must terminate, otherwise $f(r)$ would behave asymptotically as a multiple of $e^{2r\sqrt{2\lambda}}$ and $p(r)$ would not then satisfy the boundary condition at infinity. The series terminates at $s = t$ if $a_{t+1} = 0$, and this implies $\sqrt{2\lambda}\,(t+l+1) = Z$. Writing $n = t+l+1$, where n is referred to as the principal quantum number, the eigenvalues are obtained in terms of n as

$$\lambda_n = \tfrac{1}{2}(Z/n)^2 \quad (n = l+1, l+2, \ldots). \tag{4.12}$$

Writing A for a_0, the corresponding eigenfunctions may be written

$$p_{n,l}(r) = Ar^{l+1}e^{-Zr/n}\,F\{-(n-l-1),\ 2l+2,\ 2Zr/n\} \tag{4.13}$$

where F is the confluent hypergeometric function:

$$F\{\alpha,\beta,x\} \equiv 1 + \frac{\alpha}{\beta}\frac{x}{1!} + \frac{\alpha(\alpha+1)}{\beta(\beta+1)}\frac{x^2}{2!} + \ldots .$$

Since $n-l-1$ is non-negative, F represents a polynomial of degree $n-l-1$ in r, and the function $p_{n,l}(r)$ has $n-l-1$ nodes in the range $r > 0$. The normalization condition (2.4) applied to these functions leads to the following value (dependent on n and l) for the scaling factor:

$$A_{n,l} = \frac{1}{n(2l+1)!}\left\{\frac{Z(n+l)!}{(n-l-1)!}\right\}^{1/2}\left\{\frac{2Z}{n}\right\}^{l+1} .$$

For the derivation of this value see, for example, Bethe and Salpeter (1957;§3).

For a given value of Z and quantum state (n,l), the computed error bounds for the solution of the homogeneous system defined in §2.1 can now be assessed against the actual difference between the exact and numerical solutions.

Considering first the effect of truncation error, assuming no uncertainty in initial data (which at present consists only of the scalar parameter

Z), the vector $\delta\varphi$ is set to zero in equations (4.1), (4,6), (4.9), and (4.10). Typical results for the specimen case Z = 10 and a fixed step size are set out in Table 4.1 for several of the lowest-lying one-electron states; the labelling in the first column accords with standard spectroscopic notation. The computed values of the energy λ_n and scale factor $A_{n,l}$ are tabulated to eight significant figures for each state; the actual errors and the computed error estimates are given to two significant figures. The actual value and estimated bound are also given for the maximum norm of the error in each of the computed wave functions: $||\delta p||_\infty \equiv \max_{(i)} |\delta p_i|$.

The results in Table 4.1 indicate that the calculated estimates for $|\delta\lambda|$ and $||\delta p||_\infty$ are mostly less than one order of magnitude greater than the actual errors, and that they change in a manner reflecting the actual error changes in passing from one quantum state to another. The estimate for $|\delta A|$ varies less predictably but generally improves with increasing principal quantum number and decreasing number of nodes.

For a given eigensolution, the maximum estimated and actual errors in the function, as recorded in the last two columns of Table 4.1, do not necessarily occur at the same point in the range. However, the similarity between the methods of determination of the eigenfunction and its error estimate ensures that the positions of these maxima are never far apart. The estimated $|\delta p_i|$ is computed in such a way that the maximum is always reached at or just before the match-point, whereas the actual maximum could occur beyond this point, though it would never be far away since the eigenfunction itself rapidly decreases in the tail region. An impression of the way in which the estimated bound on $|\delta p_i|$ compares with its actual value over all the grid points in the range is given in Figures 4.1-4.6, which correspond to a number of the states for which results are tabulated in Table 4.1. Each graph is plotted logarithmically on both axes; in each case the upper curve represents the estimated bound and the lower curve

the magnitude of the actual error in the computed function. The actual error may pass through several zeros as r increases through the classical region and into the tail region, and the error bound partially reflects this behaviour to the extent that its amplitude attains the same number of minima as there are nodes in the wave function itself.

From a knowledge of the radial wave functions many atomic properties can be evaluated by calculating the quantum-mechanical mean or 'expectation' values of various powers of the radial coordinate r. For a function normalized according to equation (2.4) the mean value of r^ν is defined as

$$\overline{r^\nu} = \int_0^\infty r^\nu p^2(r)\,dr . \qquad (4.14)$$

The differences between functions become particularly apparent when these values are formed, the major contributions to the integral arising from different parts of the range according to the magnitude and sign of the power of r involved. In analogy with equation (2.20), the numerical approximation to the mean value defined by (4.14) will be obtained as

$$S_1^{(\nu)} + \sum_{i=1}^{n} w_i r_i^{\nu+1} P_i^2$$

where $S_1^{(\nu)} = \int_0^{r_1} r^\nu P^2 dr$

$$= A^2 r_1^{2l+3+\nu} \left\{ \frac{1}{2l+3+\nu} + \frac{2\alpha_1 r_1}{2l+4+\nu} + \frac{(\alpha_1^2 + 2\alpha_2) r_1^2}{2l+5+\nu} + \ldots \right\}$$

and a first-order approximation to the error is given by

$$2S_1^{(\nu)} A^{-1} \delta A + 2 \sum_{i=1}^{n} w_i r_i^{\nu+1} P_i \delta p_i$$

(neglecting the quadrature error).

For the Coulomb functions, analytic expressions for the mean values are known (Bethe and Salpeter, 1957, §3) and the error estimates can be compared with the actual errors in the computed results. Such a comparison is made in Table 4.2 for the same states as in Table 4.1 and the powers $\pm 1, \pm 2$ of r. The calculated estimates are mostly between one and two orders

of magnitude greater than the actual errors; they become more pessimistic with increasing orbital quantum number (l), particularly for the inverse powers where the major contribution arises at small values of r for which the function behaves like r^{l+1}.

The effect on the errors of changing the nuclear charge Z is illustrated by the results for the specimen cases included in Table 4.3. The estimates for the errors in eigenvalue and eigenfunction remain constant relative to their actual values. The fact that the estimate for $|\delta A|$ does not decrease in proportion to its actual value as Z decreases has little effect on the estimate for $|\delta p|$ since the value and therefore influence of A itself also decreases with Z.

Table 4.3 includes further results illustrating the effect of altering the step size, verifying that the actual errors in eigenvalue and eigenfunction satisfy the expected h^4-extrapolation. This is also satisfied by the computed estimates, at least approximately and the more so as the grid is refined. It is also apparent that the smaller the step size the less pessimistic the estimate for $|\delta A|$.

Turning now to the effect of physical error on the computed estimates, a simple test can be arranged by ignoring the effect of truncation error [setting the appropriate terms to zero in equations (4.1) and (4.6)] and assuming some small variation δZ in the central charge. The error estimates resulting from this source are obtained from the analysis by inserting a non-zero value for $\delta\varphi$ in equations (4.1), (4.6), (4.9), and (4.10); this value is obtained from relations (3.5) and (4.11): namely $\delta\varphi = 2r^2\,\delta v$, where $\delta v = \delta Z/r$. The corresponding 'actual error' in this situation is represented by the actual change in the exact solution, induced by the change in central charge from Z to $Z+\delta Z$. This is readily calculated; for example, from equation (4.12), the eigenvalue λ_n is altered by an amount $Z\delta Z/n^2$.

Results are listed in Table 4.4 for the first six one-electron s-states, the value of δZ being chosen in each case so as to maintain a constant

value for $\delta\lambda$. Given this form of physical error, which is necessarily rather artificial, the computed estimate for $|\delta\lambda|$ turns out to be an exact bound; the estimate for $||\delta p||_\infty$ is greater than the actual value by one order of magnitude, and the estimate for $|\delta A|$ is greater to a similar extent though it improves as the principal quantum number (and number of nodes) decreases. The calculated estimates for the mean values of powers of r listed in Table 4.4 are again (as in the case of truncation error) one to two orders of magnitude greater than the actual changes induced by the alteration from Z to $Z+\delta Z$.

The discussion throughout this section has been restricted to a consideration of the homogeneous system; some thought must now be given to an assessment of the error analysis as it applies to the inhomogeneous system defined in §2.3. It is necessary in the first place to construct a system whose analytic solution is known; this involves a suitable choice of the function $x(r)$ appearing in equation (2.22). The solution $p(r)$ and the function $x(r)$ have already been represented [equations (2.24)] in terms of the set of discrete eigenfunctions $p^{(i)}(r)$ of the homogeneous system. Using the same notation and choosing $\beta_i = 0$ for all $i \neq j$, with $\beta_j \neq 0$ for given j, the function $x(r)$ is reduced to a multiple β_j of the eigenfunction $p^{(j)}(r)$ of the homogeneous system. In this rather simple case, equation (2.25) shows that the values of ϵ satisfying $N(\epsilon) = 1$ are given by $\epsilon_j - \epsilon = \pm\beta_j$ where $\epsilon_j = 2\lambda_j$. That is, writing $\lambda = \epsilon/2$, the solution is obtained as

$$\left.\begin{aligned} \lambda &= \lambda_j \mp \tfrac{1}{2}\beta_j \\ p(r) &= \pm p^{(j)}(r) \quad \text{(normalized)} . \end{aligned}\right\}$$

Using again the Coulomb potential (4.11) and writing for $x(r)$ some multiple β of a single Coulomb function $p_{n,l}(r)$ of the type (4.13), the simple inhomogeneous system so formed has the known solution

$$\left.\begin{aligned} \lambda &= \lambda_n \mp \tfrac{1}{2}\beta \\ p(r) &= \pm p_{n,l}(r) \end{aligned}\right\} . \qquad (4.15)$$

In solving the system numerically, the coefficients in the series expansion (2.23) for $x(r)$ are easily obtained by expanding the Coulomb function $p_{n,l}(r)$.

Choosing suitably small values for β (for example, 1% of λ_n), the effect of truncation error on the computed bounds can now be tested in exactly the same way as in the homogeneous case. The results obtained are basically the same as those of Table 4.1, the bounds differing at most only in the second significant figure. The same is true in the case of an assumed 'physical error' δZ: the results form virtually a duplication of those in Table 4.4.

In the inhomogeneous case, a further physical error can be simulated by setting $\delta\varphi$ and the local truncation error to zero in equations (4.1), (4.8), (4.9), and (4.10) and choosing a non-zero value for δg in equations (4.1) and (4.8). In the simple case under consideration, in which $x(r)$ is a multiple β of $p_{n,l}(r)$, it follows from the definition of g in equation (2.26) that a variation $\delta\beta$ will lead to the value $\delta g = \frac{\delta\beta}{\beta} g \doteq \frac{\delta\beta}{\beta} G$. The effect on the exact solution (4.15) of the variation $\delta\beta$ is to leave the solution $p(r)$ unaltered but to change the parameter λ by an amount $|\delta\lambda| = \frac{1}{2}|\delta\beta|$. Inserting the value $\delta g = \frac{\delta\beta}{\beta} G$ in equations (4.1) and (4.8), the resulting estimate for $|\delta\lambda|$ yields an exact bound, as was the case when the physical error originated from the term δZ. There is no variation in the exact solution $p(r)$ with which to compare the corresponding estimate for $|\delta p|$ but the bounds produced are again virtually identical to those which resulted from a variation δZ adjusted to produce the same change in λ. It seems therefore that for these specified physical errors the error-bounding procedures yield useful estimates of the maximum errors which could be induced in the solution.

Table 4.1

Effect of truncation error.

$Z = 10, \quad h = 0.1$

state label	n	l	no. of nodes	λ calculated	A calculated	$\lvert\delta\lambda\rvert$ estimate	$\lvert\delta\lambda\rvert$ actual	$\lvert\delta A\rvert$ estimate	$\lvert\delta A\rvert$ actual	$\lVert\delta p\rVert_\infty$ estimate	$\lVert\delta p\rVert_\infty$ actual
1s	1	0	0	50.000040	63.245552	$2.4_{10}-4$	$4.0_{10}-5$	$5.0_{10}-4$	$1.2_{10}-6$	$4.1_{10}-5$	$6.1_{10}-6$
2s	2	0	1	12.500059	22.360653	$2.0_{10}-4$	$5.9_{10}-5$	$1.3_{10}-3$	$2.7_{10}-5$	$1.9_{10}-4$	$2.0_{10}-5$
2p	2	1	0	12.500020	64.549560	$1.1_{10}-4$	$2.0_{10}-5$	$2.2_{10}-3$	$1.6_{10}-4$	$1.1_{10}-4$	$1.0_{10}-5$
3s	3	0	2	5.5556531	12.171535	$2.3_{10}-4$	$9.8_{10}-5$	$2.8_{10}-3$	$7.7_{10}-5$	$5.4_{10}-4$	$5.6_{10}-5$
3p	3	1	1	5.5556094	38.251401	$1.6_{10}-4$	$5.4_{10}-5$	$6.9_{10}-3$	$2.9_{10}-4$	$4.2_{10}-4$	$3.7_{10}-5$
3d	3	2	0	5.5555687	28.509406	$7.4_{10}-5$	$1.3_{10}-5$	$4.2_{10}-3$	$1.7_{10}-3$	$2.7_{10}-4$	$1.5_{10}-5$
4s	4	0	3	3.1251526	7.9055342	$2.8_{10}-4$	$1.5_{10}-4$	$5.6_{10}-3$	$1.6_{10}-4$	$1.3_{10}-3$	$1.3_{10}-4$
4p	4	1	2	3.1251058	25.514966	$2.2_{10}-4$	$1.1_{10}-4$	$1.5_{10}-2$	$5.5_{10}-4$	$1.1_{10}-3$	$1.0_{10}-4$
4d	4	2	1	3.1250468	22.095485	$1.3_{10}-4$	$4.7_{10}-5$	$9.1_{10}-3$	$1.6_{10}-3$	$7.1_{10}-4$	$5.5_{10}-5$
4f	4	3	0	3.1250098	6.9569587	$5.5_{10}-5$	$9.8_{10}-6$	$4.1_{10}-3$	$3.0_{10}-3$	$9.4_{10}-4$	$2.0_{10}-5$
5s	5	0	4	2.0002241	5.4316596	$3.5_{10}-4$	$2.2_{10}-4$	$1.1_{10}-2$	$2.8_{10}-4$	$3.2_{10}-3$	$3.3_{10}-4$
5p	5	1	3	2.0001753	18.474254	$2.9_{10}-4$	$1.8_{10}-4$	$2.5_{10}-2$	$9.5_{10}-4$	$2.2_{10}-3$	$2.5_{10}-4$
5d	5	2	2	2.0001032	16.931066	$2.0_{10}-4$	$1.0_{10}-4$	$1.8_{10}-2$	$1.7_{10}-3$	$1.7_{10}-3$	$1.5_{10}-4$
5f	5	3	1	2.0000407	6.4476971	$1.1_{10}-4$	$4.1_{10}-5$	$6.5_{10}-3$	$2.9_{10}-3$	$1.3_{10}-3$	$7.2_{10}-5$
6s	6	0	5	1.3892010	4.3028666	$4.2_{10}-4$	$3.1_{10}-4$	$1.5_{10}-2$	$4.5_{10}-4$	$5.0_{10}-3$	$6.6_{10}-4$

Table 4.2

Effect of truncation error.

$Z = 10, \quad h = 0.1$

state label	n	l	no. of nodes	$\lvert\delta(\overline{r^2})\rvert$ estimate	$\lvert\delta(\overline{r^2})\rvert$ actual	$\lvert\delta(\overline{r})\rvert$ estimate	$\lvert\delta(\overline{r})\rvert$ actual	$\lvert\delta(\overline{r^{-1}})\rvert$ estimate	$\lvert\delta(\overline{r^{-1}})\rvert$ actual	$\lvert\delta(\overline{r^{-2}})\rvert$ estimate	$\lvert\delta(\overline{r^{-2}})\rvert$ actual
1s	1	0	0	$1.4_{10}-6$	$1.5_{10}-7$	$5.9_{10}-6$	$3.2_{10}-7$	$2.4_{10}-4$	$7.9_{10}-6$	$3.8_{10}-3$	$7.9_{10}-5$
2s	2	0	1	$1.3_{10}-4$	$8.4_{10}-6$	$1.6_{10}-4$	$5.6_{10}-6$	$4.6_{10}-4$	$1.2_{10}-5$	$3.2_{10}-3$	$2.2_{10}-7$
2p	2	1	0	$5.4_{10}-5$	$2.9_{10}-6$	$7.9_{10}-5$	$2.0_{10}-6$	$3.0_{10}-4$	$3.9_{10}-6$	$8.4_{10}-4$	$6.6_{10}-6$
3s	3	0	2	$2.4_{10}-3$	$1.3_{10}-4$	$1.4_{10}-3$	$4.2_{10}-5$	$8.2_{10}-4$	$2.0_{10}-5$	$3.6_{10}-3$	$3.8_{10}-5$
3p	3	1	1	$1.7_{10}-3$	$7.4_{10}-5$	$1.1_{10}-3$	$2.4_{10}-5$	$6.6_{10}-4$	$1.1_{10}-5$	$1.1_{10}-3$	$4.8_{10}-6$
3d	3	2	0	$8.6_{10}-4$	$1.8_{10}-5$	$6.7_{10}-4$	$6.3_{10}-6$	$6.0_{10}-4$	$2.6_{10}-6$	$7.5_{10}-4$	$1.7_{10}-6$
4s	4	0	3	$2.3_{10}-2$	$1.1_{10}-3$	$7.8_{10}-3$	$2.0_{10}-4$	$1.4_{10}-3$	$3.1_{10}-5$	$4.6_{10}-3$	$7.2_{10}-5$
4p	4	1	2	$1.8_{10}-2$	$7.5_{10}-4$	$6.3_{10}-3$	$1.4_{10}-4$	$1.2_{10}-3$	$2.1_{10}-5$	$1.4_{10}-3$	$5.4_{10}-7$
4d	4	2	1	$1.0_{10}-2$	$3.3_{10}-4$	$3.8_{10}-3$	$6.4_{10}-5$	$8.7_{10}-4$	$9.3_{10}-6$	$7.3_{10}-4$	$2.7_{10}-6$
4f	4	3	0	$9.0_{10}-3$	$7.1_{10}-5$	$4.4_{10}-3$	$1.4_{10}-5$	$1.4_{10}-3$	$2.0_{10}-6$	$1.0_{10}-3$	$7.0_{10}-7$
5s	5	0	4	$1.5_{10}-1$	$5.8_{10}-3$	$3.3_{10}-2$	$6.9_{10}-4$	$2.5_{10}-3$	$4.5_{10}-5$	$6.6_{10}-3$	$1.1_{10}-4$
5p	5	1	3	$1.0_{10}-1$	$4.5_{10}-3$	$2.3_{10}-2$	$5.5_{10}-4$	$1.8_{10}-3$	$3.5_{10}-5$	$1.7_{10}-3$	$8.1_{10}-6$
5d	5	2	2	$7.4_{10}-2$	$2.7_{10}-3$	$1.7_{10}-2$	$3.3_{10}-4$	$1.5_{10}-3$	$2.1_{10}-5$	$9.0_{10}-4$	$2.0_{10}-6$
5f	5	3	1	$5.2_{10}-2$	$1.0_{10}-3$	$1.4_{10}-2$	$1.3_{10}-4$	$1.5_{10}-3$	$8.1_{10}-6$	$7.7_{10}-4$	$1.5_{10}-6$
6s	6	0	5	$5.7_{10}-1$	$2.3_{10}-2$	$8.9_{10}-2$	$2.0_{10}-3$	$3.3_{10}-3$	$6.2_{10}-5$	$6.8_{10}-3$	$1.4_{10}-4$

Table 4.3

Effect of truncation error.

1s state.

	Z	h	λ calculated	A calculated	|δλ| estimate	|δλ| actual	|δA| estimate	|δA| actual	$\|\|\delta p\|\|_\infty$ estimate	$\|\|\delta p\|\|_\infty$ actual
Variation of central charge	20	0.1	200.00016	178.88543	$9.7_{10}-4$	$1.6_{10}-4$	$1.4_{10}-3$	$1.1_{10}-5$	$5.8_{10}-5$	$8.6_{10}-6$
	15	0.1	112.50009	116.18950	$5.5_{10}-4$	$8.9_{10}-5$	$9.2_{10}-4$	$4.6_{10}-6$	$5.1_{10}-5$	$7.4_{10}-6$
	10	0.1	50.000040	63.245552	$2.4_{10}-4$	$4.0_{10}-5$	$5.0_{10}-4$	$1.2_{10}-6$	$4.1_{10}-5$	$6.1_{10}-6$
	5	0.1	12.500010	22.360680	$6.1_{10}-5$	$9.9_{10}-6$	$1.8_{10}-4$	$2.0_{10}-8$	$2.9_{10}-5$	$4.3_{10}-6$
	2	0.1	2.0000016	5.6568543	$9.7_{10}-6$	$1.6_{10}-6$	$4.4_{10}-5$	$5.6_{10}-8$	$1.8_{10}-5$	$2.7_{10}-6$
	1	0.1	0.5000004	2.0000000	$2.4_{10}-6$	$4.0_{10}-7$	$1.6_{10}-5$	$2.1_{10}-8$	$1.3_{10}-5$	$1.9_{10}-6$
Variation of step size	10	0.2	50.000641	63.245531	$3.6_{10}-3$	$6.4_{10}-4$	$8.8_{10}-3$	$2.2_{10}-5$	$7.3_{10}-4$	$9.6_{10}-5$
	10	0.1	50.000040	63.245552	$2.4_{10}-4$	$4.0_{10}-5$	$5.0_{10}-4$	$1.2_{10}-6$	$4.1_{10}-5$	$6.1_{10}-6$
	10	0.05	50.000002	63.245553	$1.5_{10}-5$	$2.5_{10}-6$	$2.9_{10}-5$	$1.0_{10}-7$	$2.4_{10}-6$	$3.8_{10}-7$
	10	0.025	50.000000	63.245553	$9.7_{10}-7$	$1.5_{10}-7$	$2.0_{10}-6$	$4.8_{10}-8$	$1.5_{10}-7$	$2.2_{10}-8$

Table 4.4

Effect of physical error.

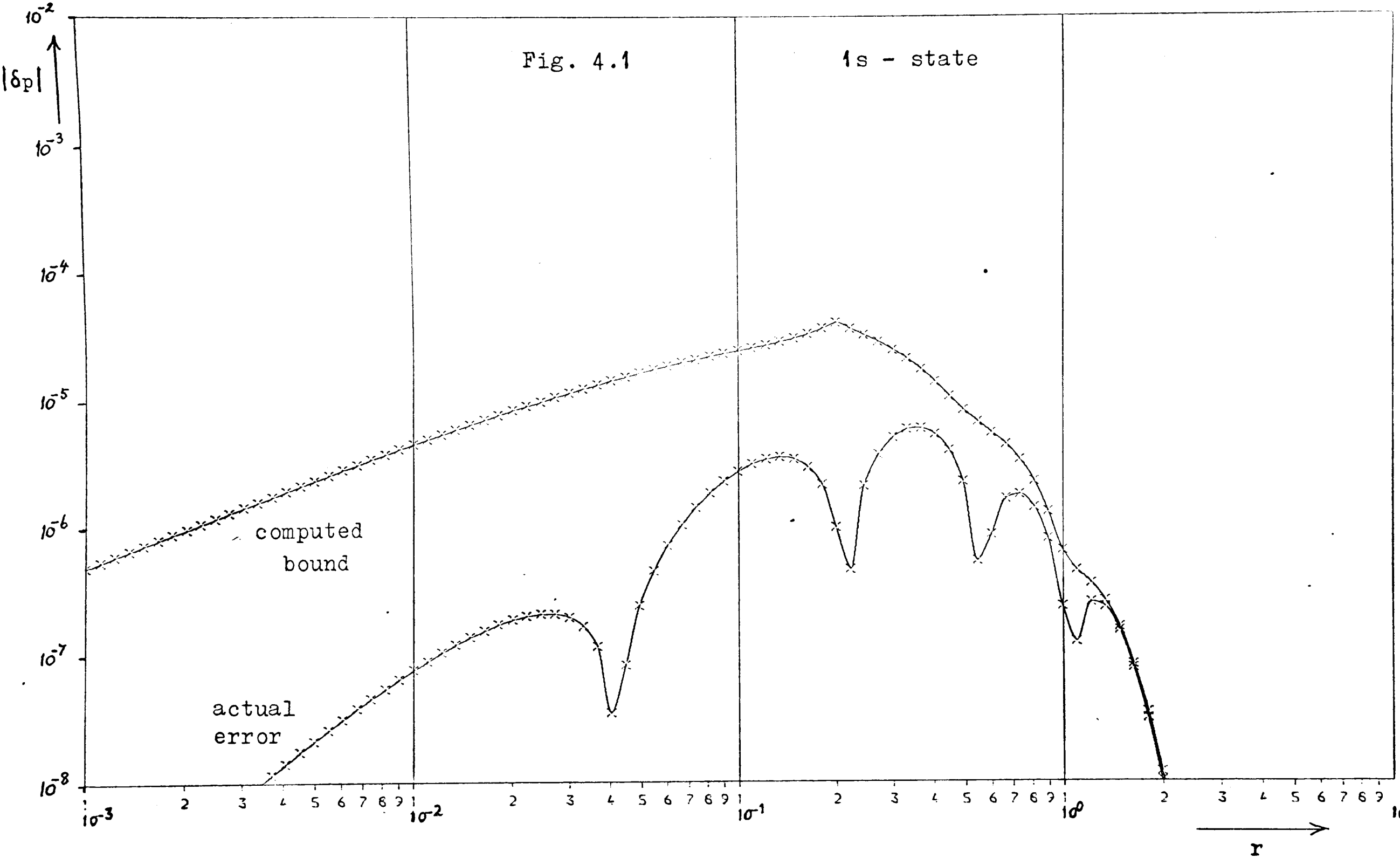
$Z = 10, \quad h = 0.1$

state label	δZ	$\lvert\delta\lambda\rvert$ estimate	$\lvert\delta\lambda\rvert$ actual	$\lvert\delta A\rvert$ estimate	$\lvert\delta A\rvert$ actual	$\lVert\delta p\rVert_\infty$ estimate	$\lVert\delta p\rVert_\infty$ actual
1s	$1_{10}-5$	$1.0_{10}-4$	$1.0_{10}-4$	$2.6_{10}-4$	$9.5_{10}-5$	$2.2_{10}-5$	$1.9_{10}-6$
2s	$4_{10}-5$	$1.0_{10}-4$	$1.0_{10}-4$	$7.0_{10}-4$	$1.3_{10}-4$	$1.0_{10}-4$	$9.0_{10}-6$
3s	$9_{10}-5$	$1.0_{10}-4$	$1.0_{10}-4$	$1.2_{10}-3$	$1.6_{10}-4$	$2.4_{10}-4$	$2.2_{10}-5$
4s	$16_{10}-5$	$1.0_{10}-4$	$1.0_{10}-4$	$1.9_{10}-3$	$1.9_{10}-4$	$4.8_{10}-4$	$4.2_{10}-5$
5s	$25_{10}-5$	$1.0_{10}-4$	$1.0_{10}-4$	$3.1_{10}-3$	$2.1_{10}-4$	$9.2_{10}-4$	$6.9_{10}-5$
6s	$36_{10}-5$	$1.0_{10}-4$	$1.0_{10}-4$	$3.3_{10}-3$	$2.3_{10}-4$	$1.2_{10}-3$	$1.0_{10}-4$

state label	δZ	$\lvert\delta(\overline{r^2})\rvert$ estimate	$\lvert\delta(\overline{r^2})\rvert$ actual	$\lvert\delta(\overline{r})\rvert$ estimate	$\lvert\delta(\overline{r})\rvert$ actual	$\lvert\delta(\overline{r^{-1}})\rvert$ estimate	$\lvert\delta(\overline{r^{-1}})\rvert$ actual	$\lvert\delta(\overline{r^{-2}})\rvert$ estimate	$\lvert\delta(\overline{r^{-2}})\rvert$ actual
1s	$1_{10}-5$	$6.3_{10}-7$	$6.0_{10}-8$	$2.9_{10}-6$	$1.5_{10}-7$	$1.3_{10}-4$	$1.0_{10}-5$	$2.0_{10}-3$	$4.0_{10}-4$
2s	$4_{10}-5$	$6.6_{10}-5$	$3.4_{10}-6$	$8.6_{10}-5$	$2.4_{10}-6$	$2.5_{10}-4$	$1.0_{10}-5$	$1.8_{10}-3$	$2.0_{10}-4$
3s	$9_{10}-5$	$1.0_{10}-3$	$3.7_{10}-5$	$6.1_{10}-4$	$1.2_{10}-5$	$3.6_{10}-4$	$1.0_{10}-5$	$1.6_{10}-3$	$1.3_{10}-4$
4s	$16_{10}-5$	$7.9_{10}-3$	$2.1_{10}-4$	$2.7_{10}-3$	$3.8_{10}-5$	$5.0_{10}-4$	$1.0_{10}-5$	$1.7_{10}-3$	$1.0_{10}-4$
5s	$25_{10}-5$	$4.3_{10}-2$	$7.9_{10}-4$	$9.2_{10}-3$	$9.4_{10}-5$	$7.0_{10}-4$	$1.0_{10}-5$	$1.9_{10}-3$	$8.0_{10}-5$
6s	$36_{10}-5$	$1.3_{10}-1$	$2.3_{10}-3$	$2.0_{10}-2$	$1.9_{10}-4$	$7.3_{10}-4$	$1.0_{10}-5$	$1.5_{10}-3$	$6.7_{10}-5$

Fig. 4.1

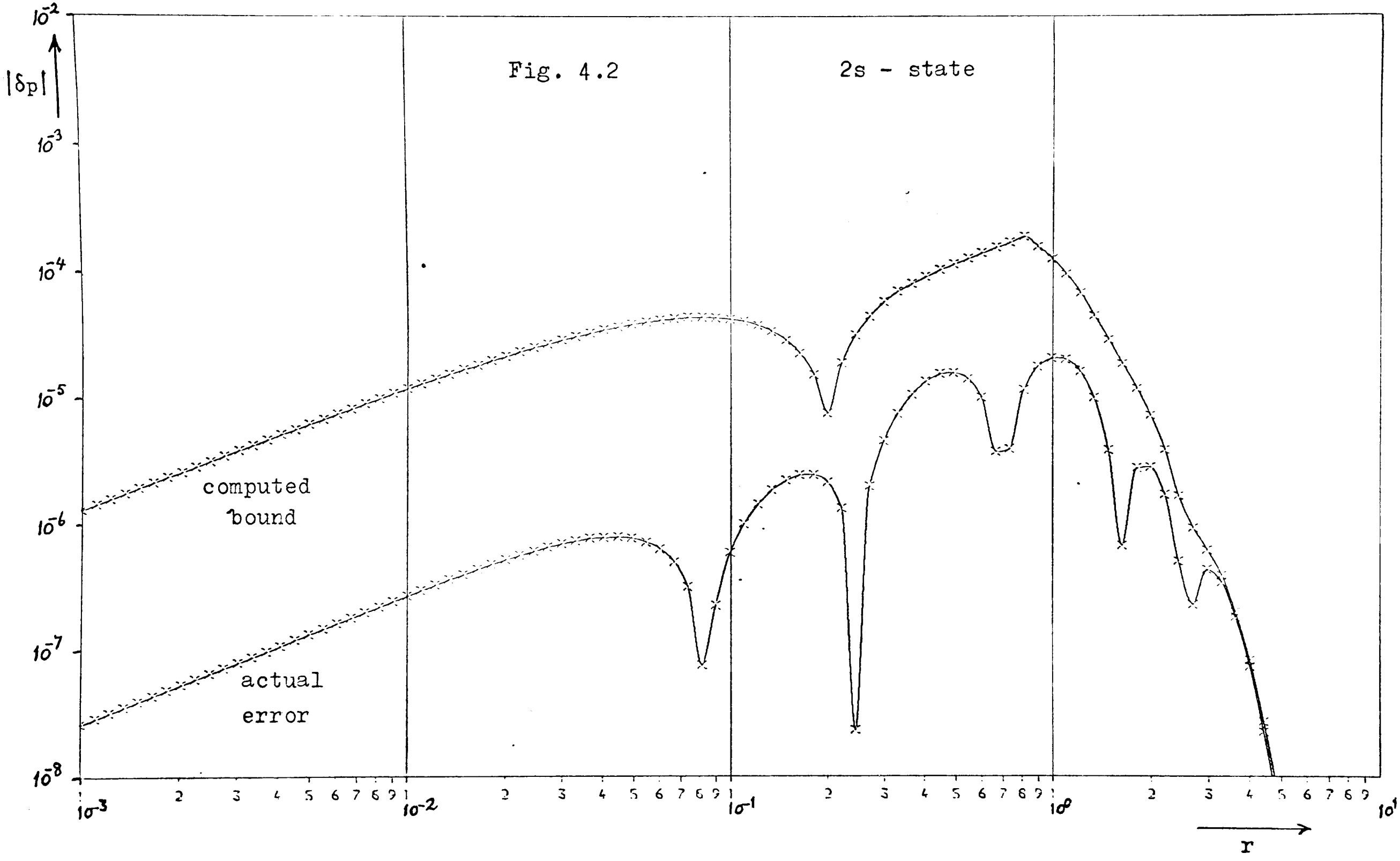
1s - state



§4.2

Fig. 4.2

2s - state



§4.2

Fig. 4.3

3s - state

$|\delta_p|$

10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}

computed bound

actual error

10^{-3} 2 3 4 5 6 7 8 9 10^{-2} 2 3 4 5 6 7 8 9 10^{-1} 2 3 4 5 6 7 8 9 10^{0} 2 3 4 5 6 7 8 9 10^{1}

r

§4.2

Fig. 4.4

4s - state

$|\delta_p|$

10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}

computed bound

actual error

10^{-3} 2 3 4 5 6 7 8 9 1 10^{-2} 2 3 4 5 6 7 8 9 1 10^{-1} 2 3 4 5 6 7 8 9 1 10^{0} 2 3 4 5 6 7 8 9 1 10^{1} 2 3 4 5 6 7 8 9 10^{2}

r

§4.2

Fig. 4.5

5s - state

$|\delta_p|$

10^{-2}

10^{-3}

10^{-4}

10^{-5}

10^{-6}

10^{-7}

10^{-8}

computed bound

actual error

10^{-3} 2 3 4 5 6 7 8 9 10^{-2} 2 3 4 5 6 7 8 9 10^{-1} 2 3 4 5 6 7 8 9 10^{0} 2 3 4 5 6 7 8 9 10^{1} 2 3 4 5 6 7 8 9 10^{2}

r

§4.2

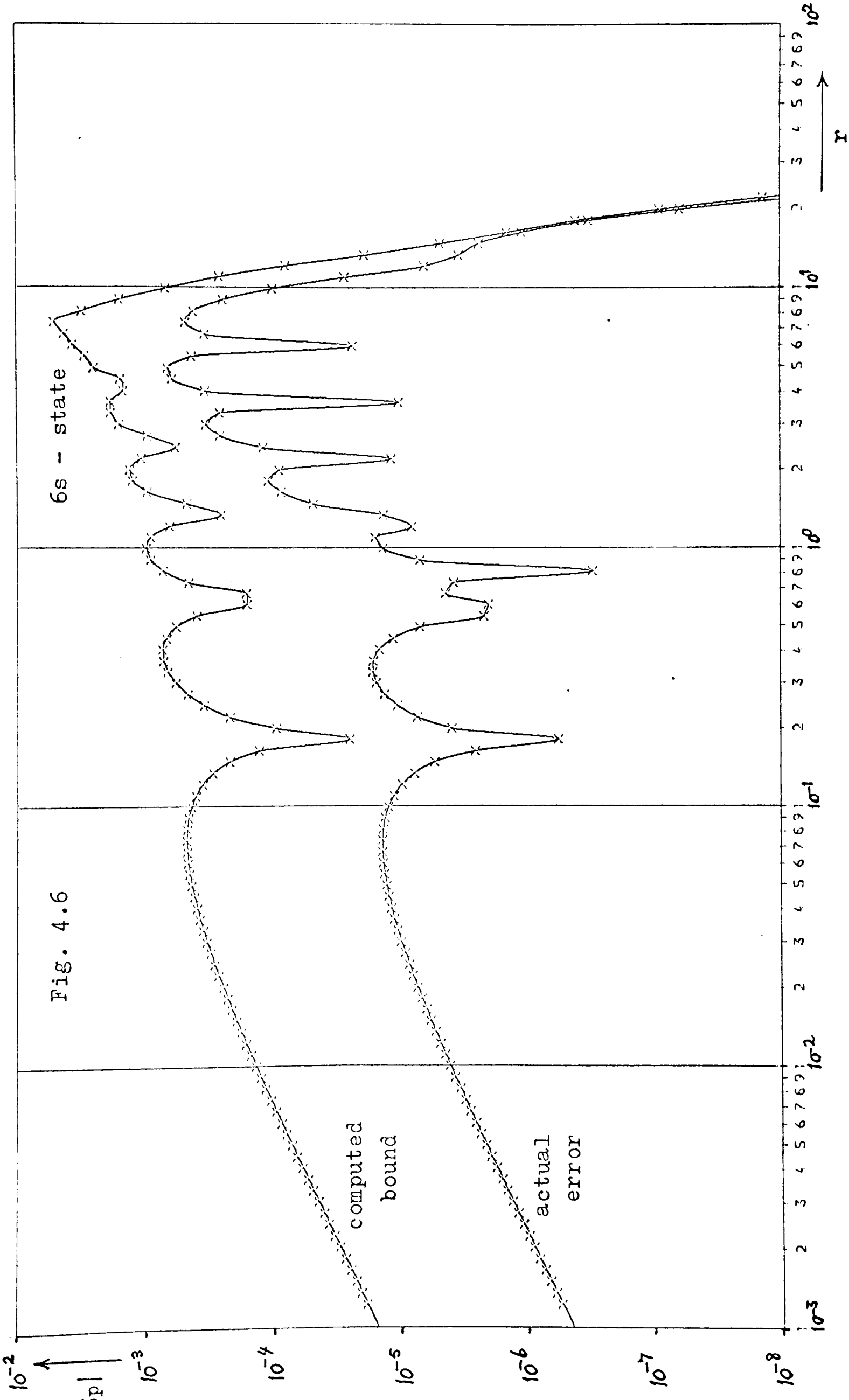


Fig. 4.6

CHAPTER 5

THE RADIAL DIRAC EQUATION

The preceding chapters have been concerned with the solution of equations derived from the Schrodinger theory of quantum mechanics, in which relativistic effects have been ignored. In this chapter the corresponding equations from Dirac's relativistic theory are considered from the point of view of their numerical solution. An analysis of the errors is developed along similar lines to the foregoing treatment, and the practicality of the results is assessed by methods similar to those of §4.2.

5.1 Numerical solution

In this section Dirac's equation (Dirac 1958) is solved for the radial wave functions by a method which follows the same general pattern as that adopted in Chapter 2 for the non-relativistic functions. The method provides an important means of solving the equations for the relativistic self-consistent field [Swirles (1935), Mayers (1957), Grant (1961, 1965), Desclaux, Mayers, and O'Brien (1971)].

The method will be described under the following headings:

(i) The radial equations

(ii) Starting procedure

(iii) Finite-difference system

(iv) Forward recurrence

(v) Tail region

(vi) Normalization and eigenvalue adjustment.

(i) The radial equations

For an electron in a central field, the Dirac theory gives rise to a radial wave function of two components which satisfy a pair of homogeneous first-order equations. Denoting the radial coordinate by r, the potential function by $v(r)$, the energy parameter by ϵ, and the components of the wave function by $p(r)$ and $q(r)$, and working in atomic units, the equations may

be written:

$$\left.\begin{aligned} \frac{dp(r)}{dr} + \frac{k}{r}\,p(r) + \{2c + \frac{v(r) - \epsilon}{c}\}q(r) &= 0 \\ \frac{dq(r)}{dr} - \frac{k}{r}\,q(r) - \{\frac{v(r) - \epsilon}{c}\}\,p(r) &= 0 \end{aligned}\right\} \qquad (5.1)$$

where c, the inverse of the fine-structure constant, has the value 137.037 in atomic units, correct to 6 significant figures.

The constant k has its value determined by the orbital and total angular momentum quantum numbers of the electronic state (referred to by l and j respectively). The orbital quantum number is a non-negative integer: for $l = 0$ the values of j and k are $\frac{1}{2}$ and -1, and for each $l > 0$ there are two states corresponding to the values $j = l + \frac{1}{2}$, $k = -(l+1)$ and $j = l - \frac{1}{2}$, $k = +l$. Thus k is a positive or negative integer, and $|k| = j + \frac{1}{2}$.

In the non-relativistic case the parameter ϵ is used as a substitution for 2λ, where $-\lambda$ represents the energy of the electronic state, but its occurrence in the Dirac equations (5.1) is as a replacement for λ itself, where $-\lambda$ again represents the energy of the state (discounting the rest-mass energy). It should also be noted that the function labelled as q(r) does not represent, as in the non-relativistic case, a substitution for p(r) but rather a separate component of the total wave function.

As in the non-relativistic case, only the discrete spectrum of physically bound states will be considered, the restriction being effected by imposing the usual homogeneous boundary conditions on each component of the wave function:

$$p(0) = q(0) = 0\,; \quad p(r),\, q(r) \to 0 \text{ as } r \to \infty\,. \qquad (5.2)$$

The energy parameter ϵ, as defined in equations (5.1), then assumes only positive values and the two-component wave functions are square-integrable, the normalization condition appropriate to the relativistic problem being expressible in the form:

$$\int_0^\infty \{p^2(r) + q^2(r)\}\,dr = 1\,. \qquad (5.3)$$

Strictly speaking, two of the conditions (5.2) are redundant, since the pair of conditions on either one of the components is sufficient for specifying the eigensolutions uniquely, as can be seen by transforming the first-order equations (5.1) into a pair of second-order equations for p and q respectively. If either of these equations, with the two boundary conditions, were to be solved by the methods of Chapter 2, the value of the other component would then be specified uniquely by one or other of the equations (5.1) and would necessarily satisfy the stated conditions. (It will not be necessary in practice to force the system into second-order form in this way with the consequent possibility of a loss of precision since the first-order equations can conveniently be solved directly.)

An inhomogeneous system may be defined by imposing the same boundary and normalization conditions (5.2) and (5.3) but modifying the radial equations (5.1) by the addition of functions which it will be convenient to write as $x(r)/r$ and $y(r)/r$. The equations in their inhomogeneous form will therefore be represented as follows:

$$\left.\begin{aligned} \frac{dp(r)}{dr} + \frac{k}{r}p(r) + \{2c + \frac{v(r) - \epsilon}{c}\}q(r) &= x(r)/r \\ \frac{dq(r)}{dr} - \frac{k}{r}q(r) - \{\frac{v(r) - \epsilon}{c}\}p(r) &= y(r)/r \end{aligned}\right\}. \qquad (5.4)$$

Much of the analysis will apply equally to the homogeneous and inhomogeneous cases, the former being regarded as a particular case of the latter and obtainable from it by setting the functions x and y to zero.

It is of interest to note that the non-relativistic Schrodinger equation may be recovered from the pair of equations (5.1) by omitting the term in c^{-1} from the first of these, differentiating, and substituting for $\frac{dq}{dr}$ from the second equation. Then, noting that $k(k+1) = l(l+1)$, the equation for $p(r)$ becomes

$$\frac{d^2p}{dr^2} - \frac{l(l+1)}{r^2}p + 2(v - \epsilon)p = 0$$

which is essentially the same as equation (2.1). The relativistic effect

can therefore be considered as arising from the term involving the ratio v/c: the effect is small on the outermost electronic orbits but is particularly marked on those which penetrate close to the nucleus, the more so the higher the nuclear charge. As the effect diminishes the function p(r) is transformed into the non-relativistic radial wave function whereas, as is evident from the first of equations (5.1), the function q(r) assumes the form

$$q = -\frac{1}{2c}\left(\frac{dp}{dr} + \frac{k}{r}p\right)$$

and this component vanishes entirely when the relativistic effect is neglected. For this reason it is usual to designate p as the 'large' component and q as the 'small' component, though the difference in size diminishes as the importance of relativistic effects increases. A description is given by Burke and Grant (1967) of certain aspects of the modifications which atomic wave functions are liable to undergo in consequence of relativistic effects.

(ii) Starting procedure

The required solutions have the same general behaviour as those of the non-relativistic Schrodinger theory, and comparable methods of solution can therefore be adopted. Anticipating the transformation to a logarithmic scale and the consequent need of function values from which to start the forward integration, the first requirement will be to determine the solution in series for small values of r. Assuming a point nuclear charge Z, the potential may be expanded in the form

$$v(r) = \frac{1}{r}\sum v^{(j)} r^j \quad \text{where } v^{(0)} = Z .$$

It is also assumed in the first instance that the functions x(r) and y(r) belong to the space of eigenfunctions of the discrete spectrum of the homogeneous system. This being so, the following expansions may be written:

$$p(r) = r^{\gamma} \sum p^{(j)} r^j \; ; \qquad q(r) = r^{\gamma} \sum q^{(j)} r^j \tag{5.5}$$

and

$$x(r)/r = r^{\gamma} \sum x^{(j)} r^j \; ; \qquad y(r)/r = r^{\gamma} \sum y^{(j)} r^j \; .$$

Each summation extends over the values 0, 1, 2, ... of the index j (not to be confused with the previous use of j as a quantum number).

Substituting these expansions into the pair of equations (5.4) and equating coefficients of successive powers of r to zero, the following relations are obtained:

$$\left.\begin{aligned} (\gamma+k)p^{(0)} + \frac{Z}{c} q^{(0)} &= 0 \\ (\gamma-k)q^{(0)} - \frac{Z}{c} p^{(0)} &= 0 \end{aligned}\right\} \text{ (indicial equations)}$$

and, for $j = 1, 2, \ldots,$

$$\left.\begin{aligned} (\gamma+j+k)p^{(j)} + \frac{Z}{c} q^{(j)} &= -\sum_{i=1}^{j} \frac{v^{(i)}}{c} q^{(j-i)} - (2c - \frac{\epsilon}{c}) q^{(j-1)} + x^{(j-1)} \\ (\gamma+j-k)q^{(j)} - \frac{Z}{c} p^{(j)} &= +\sum_{i=1}^{j} \frac{v^{(i)}}{c} p^{(j-i)} - \frac{\epsilon}{c} p^{(j-1)} + y^{(j-1)} \end{aligned}\right\} .$$

The indicial equations give

$$p^{(0)}/q^{(0)} = -\frac{Z}{c}(\gamma+k)^{-1} = +(\frac{Z}{c})^{-1}(\gamma-k)$$

so that

$$\gamma = +\sqrt{k^2 - Z^2/c^2}$$

where only the positive root is admissible on account of the boundary condition at $r = 0$.

Thus for a given system represented by specified values of $v(r)$, $x(r)$, $y(r)$, and k, the value of γ is fixed, and each of the coefficients $p^{(j)}$ and $q^{(j)}$ is uniquely specified (given an estimate for ϵ) by assigning an arbitrary non-zero value (α) to one of them, for example to $p^{(0)}$. Then the value of $q^{(0)}$ follows from the indicial equations, and, for $j \geq 1$, the values of $p^{(j)}$ and $q^{(j)}$ are determined as follows:

§5.1

$$\begin{bmatrix} (\gamma+j+k) & \frac{Z}{c} \\ -\frac{Z}{c} & (\gamma+j-k) \end{bmatrix}\begin{bmatrix} p^{(j)} \\ q^{(j)} \end{bmatrix} = \sum_{i=1}^{j}\begin{bmatrix} 0 & -v^{(i)}/c \\ +v^{(i)}/c & 0 \end{bmatrix}\begin{bmatrix} p^{(j-i)} \\ q^{(j-i)} \end{bmatrix}$$

$$-\begin{bmatrix} 0 & (2c-\frac{\epsilon}{c}) \\ \epsilon/c & 0 \end{bmatrix}\begin{bmatrix} p^{(j-1)} \\ q^{(j-1)} \end{bmatrix} + I\begin{bmatrix} x^{(j-1)} \\ y^{(j-1)} \end{bmatrix}$$

where I represents the 2×2 identity matrix.

Now

$$\begin{bmatrix} (\gamma+j+k) & \frac{Z}{c} \\ -\frac{Z}{c} & (\gamma+j-k) \end{bmatrix}^{-1} = \{(\gamma+j)^2 - k^2 + Z^2/c^2\}^{-1}\begin{bmatrix} (\gamma+j-k) & -\frac{Z}{c} \\ \frac{Z}{c} & (\gamma+j+k) \end{bmatrix}$$

where $\{(\gamma+j)^2 - k^2 + Z^2/c^2\} = (\gamma+j)^2 - \gamma^2 = j(2\gamma+j)$.

Therefore

$$j(2\gamma+j)\begin{bmatrix} p^{(j)} \\ q^{(j)} \end{bmatrix} = \sum_{i=1}^{j}\begin{bmatrix} -\frac{Z}{c}v^{(i)}/c & -(\gamma+j-k)v^{(i)}/c \\ (\gamma+j+k)v^{(i)}/c & -\frac{Z}{c}v^{(i)}/c \end{bmatrix}\begin{bmatrix} p^{(j-i)} \\ q^{(j-i)} \end{bmatrix}$$

$$-\begin{bmatrix} -\frac{Z}{c}\frac{\epsilon}{c} & (\gamma+j-k)(2c-\frac{\epsilon}{c}) \\ (\gamma+j+k)\frac{\epsilon}{c} & \frac{Z}{c}(2c-\frac{\epsilon}{c}) \end{bmatrix}\begin{bmatrix} p^{(j-1)} \\ q^{(j-1)} \end{bmatrix}$$

$$+\begin{bmatrix} (\gamma+j-k) & -\frac{Z}{c} \\ \frac{Z}{c} & (\gamma+j+k) \end{bmatrix}\begin{bmatrix} x^{(j-1)} \\ y^{(j-1)} \end{bmatrix} .$$

The last vector in this expression for $p^{(j)}$ and $q^{(j)}$ is absent in the homogeneous case, and the first j matrices would be absent from the expression if the potential were purely Coulomb. Each of the series (5.5) is truncated at the point when the next term in each expansion is of negligible value.

In the case in which the inhomogeneous equations (5.4) are derived as the equations of the relativistic self-consistent field, the function x(r) does not necessarily belong to the space of eigenfunctions of the homogeneous system since it may include terms which are multiples of functions corresponding to other values of k and therefore of γ. Since γ is not an integer this means that the powers of r in the expansion of x may differ from those in the expansions of p and q. However, for the leading term the discrepancy

is small whenever the ratio Z/c is small, and in any case the coefficients in the expansion of x are themselves of order c^{-2} so that any slight discrepancy would be of no practical significance. If the integration is started at a sufficiently small value of r, only the first two terms need be retained in the expansions for p and q and this demands a knowledge of only the first term in the expansion of x.

For atoms of small and intermediate atomic number Z this starting procedure is found to be quite adequate in practice, but trouble arises when Z approaches the order of 100. In this case the approximation of a point nuclear charge also becomes unsatisfactory since many of the electronic orbits may penetrate the nuclear region to a significant extent. The nucleus is then more realistically modelled by assuming a uniform spherical charge distribution of radius r_N and total charge Z. For $0 \leq r \leq r_N$ the term Z/r in the expansion of v(r) is replaced by $\frac{1}{2}\frac{Z}{r_N}(3 - r^2/r_N^2)$, so that in this region the potential may be written

$$v(r) = \sum_{j=1,2,\ldots} \tilde{v}^{(j)} r^{j-1}$$

where

$$\begin{aligned}
\tilde{v}^{(1)} &= v^{(1)} + \frac{3}{2}\frac{Z}{r_N}, \\
\tilde{v}^{(2)} &= v^{(2)}, \\
\tilde{v}^{(3)} &= v^{(3)} - \frac{1}{2}\frac{Z}{r_N^3}, \\
\tilde{v}^{(j)} &= v^{(j)} \quad \text{for } j \geq 4.
\end{aligned}$$

This expansion is of the same form as the previous one except for the omission of the term in 1/r, so that the values of γ and $\{p^{(j)}, q^{(j)}\}$ are found by setting $v^{(0)}$ $(\equiv Z)$ to zero in the formulae and replacing $v^{(j)}$ by $\tilde{v}^{(j)}$ $(j \geq 1)$; it follows in particular that $\gamma = |k|$ and that either $p^{(0)}$ or $q^{(0)}$ is zero depending on the sign of k. Thus γ is now an integer and there can no longer be any discrepancy between the powers of r in the expansion of x and those in the expansions of p and q. The disadvantages associated with the concept

of a point nuclear charge are therefore overcome by using this model, although its use requires the integration to be started at a value of r no greater than r_N so that in practice the point r_N itself would be chosen for this purpose.

As regards the finding of solutions by numerical calculation it should be emphasized that the exact determination of the series solutions is not of critical importance.

(iii) Finite-difference system

Taking as independent variable

$$t = \ln r$$

as in the non-relativistic case (and for the same reasons), the radial equations (5.4) become

$$\left.\begin{aligned} \frac{dp}{dt} + kp + r(2c + \frac{v-\epsilon}{c})q &= x \\ \frac{dq}{dt} - kq - r(\frac{v-\epsilon}{c})p &= y \end{aligned}\right\} . \qquad (5.6)$$

Each derivative at a given point r_i could now be replaced by a mean first-difference, that is by the first term in its central-difference expansion corresponding to a grid of equal intervals h in t:

$$h \frac{d}{dt} p_i \equiv (\mu\delta - \tfrac{1}{6}\mu\delta^3 + \tfrac{1}{30}\mu\delta^5 - \ldots)p_i$$

where p_i stands for $p(r_i)$ and μ is the 'averaging' operator (Fox 1957, Ch.2, §§3,6). However, the forward integration of the resulting system would give rise to so-called weak induced instability associated with the replacement of first derivatives by second-order difference schemes (see Fox and Mayers 1968, Ch.10, §§36-38). The instability is avoided by using the 'trapezium rule' formula which connects the mean of the derivatives p'_i, p'_{i+1} with the differences centred on the half-way point and is given by

$$\tfrac{1}{2} h \frac{d}{dt} (p_{i+1} + p_i) \equiv (\delta + \tfrac{1}{12}\delta^3 - \tfrac{1}{120}\delta^5 + \tfrac{1}{840}\delta^7 - \ldots)p_{i+\frac{1}{2}} ,$$

and similarly for $\frac{dq}{dt}$. [The derivation of the formula is given for example

in Fox 1957 (Ch.2, §14) where mention is also made of its application to the case of simultaneous first-order equations (Ch.10, §3).]

Truncating the series after the first term, replacing this term by pivotal values (that is, $\delta p_{i+1/2} = p_{i+1} - p_i$), and substituting into the radial equations, the resulting algebraic equations are still of first order, as desired, but form only a crude approximation to the true system. Hence the second term must also be taken into account, and to avoid inducing instability this can only be done by incorporating it as a deferred correction during each iteration of the overall solving process. The remaining terms will be used in the production of 'a posteriori' error bounds. It is convenient therefore to express the complete formula in the following way:

$$\begin{aligned} \tfrac{1}{2}h\,\frac{d}{dt}\,(p_{i+1}+p_i) &= (p_{i+1}-p_i) + C_{i+1/2}(p) \\ &= (p_{i+1}-p_i) + C^{(1)}_{i+1/2}(p) + C^{(2)}_{i+1/2}(p) \end{aligned}$$

where

$$\begin{aligned} C^{(1)}_{i+1/2}(p) &= \tfrac{1}{12}\,\delta^3 p_{i+1/2} \\ &= \tfrac{1}{12}(p_{i+2}-p_{i-1}) - \tfrac{1}{4}(p_{i+1}-p_i) \end{aligned}$$

and

$$C^{(2)}_{i+1/2}(p) = (-\tfrac{1}{120}\delta^5 + \tfrac{1}{840}\delta^7 - \ldots)\,p_{i+1/2}$$

and

$$C = C^{(1)} + C^{(2)},$$

with similar expressions for the function $q(r)$.

Substitution of these in the differential system (5.6) yields the following pair of equations:

$$\left.\begin{aligned} (1+\tfrac{1}{2}hk)\,p_{i+1} - (1-\tfrac{1}{2}hk)\,p_i + \tfrac{1}{2}hr_{i+1}(2c - \frac{\epsilon - v_{i+1}}{c})\,q_{i+1} + \tfrac{1}{2}hr_i(2c - \frac{\epsilon - v_i}{c})\,q_i \\ = \tfrac{1}{2}h(x_{i+1}+x_i) - C_{i+1/2}(p)\,; \\ (1-\tfrac{1}{2}hk)\,q_{i+1} - (1+\tfrac{1}{2}hk)\,q_i + \tfrac{1}{2}hr_{i+1}\{(\epsilon - v_{i+1})/c\}p_{i+1} + \tfrac{1}{2}hr_i\{(\epsilon - v_i)/c\}p_i \\ = \tfrac{1}{2}h(y_{i+1}+y_i) - C_{i+1/2}(q) \end{aligned}\right\}. \qquad (5.7)$$

The approximating algebraic system (that is, the system whose solution is the one actually computed) is obtained from these equations by truncating the operator C after its first term $C^{(1)}$ so that the remaining terms, represented by $C^{(2)}$, form the local truncation error. At the same time, the values at the grid points of the functions v, x, and y (which represent the 'correct' data for the problem) are replaced by the corresponding values of the mesh functions $\{V_i\}$, $\{X_i\}$, $\{Y_i\}$ to represent the actual data, so that the discrepancy (if any) between the two represents the assumed 'physical' error associated with the data. Denoting the computed approximations to the energy parameter and the two-component wave function by E and $\{P_i,Q_i\}$, the algebraic system may now be written in the form:

$$\left.\begin{aligned} w_+P_{i+1} - w_-P_i + \theta_{i+1}Q_{i+1} + \theta_iQ_i &= \xi_i \\ w_-Q_{i+1} - w_+Q_i + \varphi_{i+1}P_{i+1} + \varphi_iP_i &= \eta_i \end{aligned}\right\} \qquad (5.8)$$

where

$$\begin{aligned} w_\pm &= 1 \pm \tfrac{1}{2}hk \\ \theta_i &= \tfrac{1}{2}hr_i(2c - \mu_i/c) = hcr_i - \varphi_i \\ \varphi_i &= \tfrac{1}{2}hr_i\mu_i/c = \tfrac{1}{2}\tfrac{h}{c}r_i(E - V_i) \\ \mu_i &= E - V_i \\ \xi_i &= \tfrac{1}{2}h(X_{i+1} + X_i) - c^{(1)}_{i+1/2}(P) \\ \eta_i &= \tfrac{1}{2}h(Y_{i+1} + Y_i) - c^{(1)}_{i+1/2}(Q) . \end{aligned}$$

(iv) Forward recurrence

Choosing a starting point $r = r_1$ as in the non-relativistic case (§2.2) and estimates E and α for the energy parameter ϵ and the coefficient $p^{(0)}$ of the starting series (5.5), values may be found for P_1 and Q_1 from which to begin recurring forward using the relations (5.8). In these relations, the terms ξ_i, η_i are given by

$$\left\{\begin{aligned} \xi_i &= \tfrac{1}{2}h(X_{i+1} + X_i) - \tfrac{1}{12}(P_{i+2} - P_{i-1}) + \tfrac{1}{4}(P_{i+1} - P_i) \\ \eta_i &= \tfrac{1}{2}h(Y_{i+1} + Y_i) - \tfrac{1}{12}(Q_{i+2} - Q_{i-1}) + \tfrac{1}{4}(Q_{i+1} - Q_i) \end{aligned}\right.$$

where the functions X,Y are absent in the homogeneous case. The values of

P and Q required to determine ξ and η are obviously not known initially and are therefore set to zero before the initial integration, but thereafter, in the manner of a deferred correction, the values used are the ones computed during the previous iteration of the total solving procedure.

The recurrence (5.8) may be expressed in the form:

$$\begin{bmatrix} w_+ & \theta_{i+1} \\ \varphi_{i+1} & w_- \end{bmatrix}\begin{bmatrix} P_{i+1} \\ Q_{i+1} \end{bmatrix} = \begin{bmatrix} w_- & -\theta_i \\ -\varphi_i & w_+ \end{bmatrix}\begin{bmatrix} P_i \\ Q_i \end{bmatrix} + \begin{bmatrix} \xi_i \\ \eta_i \end{bmatrix} .$$

Now

$$\begin{bmatrix} w_+ & \theta_{i+1} \\ \varphi_{i+1} & w_- \end{bmatrix}^{-1} = \Delta_{i+1}^{-1}\begin{bmatrix} w_- & -\theta_{i+1} \\ -\varphi_{i+1} & w_+ \end{bmatrix}$$

where

$$\begin{aligned} \Delta_i &= w_+w_- - \theta_i\varphi_i \\ &= 1 - \tfrac{1}{4}h^2k^2 - \tfrac{1}{4}h^2r_i^2\frac{\mu_i}{c}(2c - \frac{\mu_i}{c}) \\ &= 1 - \tfrac{1}{4}h^2k^2 - \tfrac{1}{2}h^2r_i\mu_i + \tfrac{1}{4}h^2r_i\mu_i^2/c^2 \\ &= 1 - \tfrac{1}{4}h^2(k^2 + 2r_i^2\mu_i) + O(\tfrac{h}{c})^2 . \end{aligned}$$

Clearly the value of h must be sufficiently small to ensure that throughout the region of forward integration the values of the Δ_i are bounded away from zero. This being so, the recurrence assumes the final form:

$$\begin{bmatrix} P_{i+1} \\ Q_{i+1} \end{bmatrix} = G_i\begin{bmatrix} P_i \\ Q_i \end{bmatrix} + \Delta_{i+1}^{-1}\begin{bmatrix} w_-\xi_i - \theta_{i+1}\eta_i \\ w_+\eta_i - \varphi_{i+1}\xi_i \end{bmatrix} \qquad (5.9)$$

where

$$G_i = \Delta_{i+1}^{-1}\begin{bmatrix} w_-^2 + \theta_{i+1}\varphi_i & -w_-\theta_i - w_+\theta_{i+1} \\ -w_+\varphi_i - w_-\varphi_{i+1} & w_+^2 + \varphi_{i+1}\theta_i \end{bmatrix} .$$

As in the non-relativistic case, the forward integration is continued as far as the outer classical turning point [see Figure 2.2]. A sufficiently close approximation to this point is provided by the grid point r_m for which $V_m > E > V_{m+1}$, and so the recurrence is allowed to proceed as far as the calculation of P_m and Q_m. One of these is required as a boundary value for computing the solution in the region $r > r_m$ (the tail region).

If P_m is chosen for this purpose, the tail-region solution will yield a second estimate for Q_m, as the method described in §5.1(v) will show. The matching process appropriate to this problem, to be described in §5.1(vi), will be concerned with adjusting the estimate of E so as to reduce to zero the residual quantity R defined as

$$R = Q_m^{(f)} - Q_m^{(t)} \tag{5.10}$$

where $Q_m^{(f)}$ represents the value obtained for Q_m from the forward integration and $Q_m^{(t)}$ the value from the tail-region solution. In the inhomogeneous case, the estimated scaling parameter α will also be adjusted so as simultaneously to obtain the correct normalization.

(v) Tail region

In the region $r > r_m$, each component of the required wave functions exhibits the same type of exponentially-decreasing behaviour as that shown by the non-relativistic functions, and a similar method of forward and backward integration may be used. It will be assumed for the moment that the method to be described is stable and yields solutions whose behaviour is of the correct type. Then, given a value for P_m from the previous stage of the calculation and anticipating that the integration will be stopped at some point r_n where the values $Q_n, P_{n+1}, Q_{n+1}, \ldots$ have become negligibly small, the complete system of equations (5.8) may be written in matrix form as:

$$A \begin{bmatrix} P \\ Q \end{bmatrix} = \begin{bmatrix} \xi \\ \eta \end{bmatrix} \tag{5.11}$$

where

$$A \equiv \begin{bmatrix} -w_+ & \varphi_{m+1} & w_- & & & & \\ \theta_m & w_+ & \theta_{m+1} & & & & \\ & \varphi_{m+1} & -w_+ & \varphi_{m+2} & w_- & & \\ & -w_- & \theta_{m+1} & w_+ & \theta_{m+2} & & \\ & & & & & \ddots & \\ & & & & & \varphi_{n-1} & -w_+ & \varphi_n \\ & & & & & -w_- & \theta_{n-1} & w_+ \end{bmatrix}$$

with $\begin{bmatrix} P \\ Q \end{bmatrix} \equiv \begin{bmatrix} Q_m \\ P_{m+1} \\ Q_{m+1} \\ P_{m+2} \\ \vdots \\ Q_{n-1} \\ P_n \end{bmatrix}$ and $\begin{bmatrix} \xi \\ \eta \end{bmatrix} \equiv \begin{bmatrix} \eta_m - \varphi_m P_m \\ \xi_m + w_- P_m \\ \eta_{m+1} \\ \xi_{m+1} \\ \vdots \\ \eta_{n-1} \\ \xi_{n-1} \end{bmatrix}$.

Continuing the analogy with the non-relativistic case, the matrix A is factorized into a product of lower and upper triangular matrices LU:

$$\begin{bmatrix} 1 & & & & & \\ l_m & 1 & & & & \\ & j_{m+1} & 1 & & & \\ & k_{m+1} & l_{m+1} & 1 & & \\ & & & & \ddots & \\ & & & & j_{n-1} & 1 \\ & & & & k_{n-1} & l_{n-1} & 1 \end{bmatrix} \begin{bmatrix} t_m & \varphi_{m+1} & w_- & & & & \\ & u_m & v_m & & & & \\ & & t_{m+1} & \varphi_{m+2} & w_- & & \\ & & & u_{m+1} & v_{m+1} & & \\ & & & & & \ddots & \\ & & & & & t_{n-1} & \varphi_n \\ & & & & & & u_{n-1} \end{bmatrix}$$

whose elements are obtained from the relations:

$$
\begin{aligned}
t_m &= -w_+ \\
l_m &= \theta_m / t_m \\
u_m &= w_+ - l_m \varphi_{m+1} \\
v_{i-1} &= \theta_i - l_{i-1} w_- \\
j_i &= \varphi_i / u_{i-1} \\
k_i &= -w_- / u_{i-1} \\
t_i &= -w_+ - j_i v_{i-1} \\
l_i &= (\theta_i - k_i v_{i-1}) / t_i \\
u_i &= w_+ - l_i \varphi_{i+1}
\end{aligned}
\qquad (i = m+1,\ m+2,\ \ldots,\ n-1).
$$

The resulting system, namely $LU\begin{bmatrix} P \\ Q \end{bmatrix} = \begin{bmatrix} \xi \\ \eta \end{bmatrix}$, is solved by forward and backward recurrence, introducing in the process an auxiliary vector which will be denoted by

$$
\begin{bmatrix} \psi \\ \gamma \end{bmatrix} \equiv \begin{bmatrix} \psi_m \\ \gamma_m \\ \hline \psi_{m+1} \\ \gamma_{m+1} \\ \hline \vdots \\ \hline \psi_{n-1} \\ \gamma_{n-1} \end{bmatrix}.
$$

The elements of $\begin{bmatrix} \psi \\ \gamma \end{bmatrix}$ are determined by forward recurrence in the system of equations $L\begin{bmatrix} \psi \\ \gamma \end{bmatrix} = \begin{bmatrix} \xi \\ \eta \end{bmatrix}$, to give

$$
\begin{aligned}
\psi_m &= \eta_m - \varphi_m P_m \\
\gamma_m &= \xi_m + w_- P_m - l_m \psi_m \\
\psi_i &= \eta_i - j_i \gamma_{i-1} \\
\gamma_i &= \xi_i - k_i \gamma_{i-1} - l_i \psi_i
\end{aligned}
\qquad (i = m+1,\ m+2,\ \ldots,\ n-1).
$$

This assumes that a suitable criterion is employed for stopping the forward recurrence: for example by demanding that the ratio $|y_i/u_i|$ be less than a specified value. The solution vector $\begin{bmatrix} P \\ Q \end{bmatrix}$ is then found by recurring back through the system $U\begin{bmatrix} P \\ Q \end{bmatrix} = \begin{bmatrix} \psi \\ \gamma \end{bmatrix}$, with the result that

$$\left.\begin{aligned} P_n &= \gamma_{n-1}/u_{n-1} \\ Q_{n-1} &= (\psi_{n-1} - \varphi_n P_n)/t_{n-1} \\ P_i &= (\gamma_{i-1} - v_{i-1}Q_i)/u_{i-1} \\ Q_{i-1} &= (\psi_{i-1} - \varphi_i P_i - w_{-}Q_i)/t_{i-1} \end{aligned}\right\} \quad (i = n-1,\ n-2,\ \ldots,\ m+1).$$

The last element to be computed in this scheme is Q_m, and this is the value which is referred to as $Q_m^{(t)}$ in equation (5.10).

The stopping criterion for determining the outer limit of integration ensures that P_n and therefore all P_i, Q_i for $i \geq n$ are of negligible value in relation to the required accuracy.

It is not immediately obvious that the method just described for solving the boundary-value system (5.11) represents a stable procedure, automatically yielding a solution that decreases monotonically through the tail region in the correct manner. The corresponding system in the non-relativistic case, written as equation (2.16), gave rise to a symmetric tridiagonal matrix whose strict diagonal dominance ensured that the magnitudes of the elements l_i and u_i of the factorized matrix were bounded above and below respectively. It followed from this that the system was inherently well-conditioned and the method of solution free from instabilities, with no danger of the matrix becoming singular or of any of the elements u_i vanishing. As pointed out by Fox (1957, Ch.8, §45), these desirable properties are closely related to the nature of the complementary solutions of the recurrence (2.12). Thus if the mesh function $\zeta = \{\zeta_i\}$ satisfies the homogeneous form of the recurrence, with 'initial' values $\zeta_m = 0$, $\zeta_{m+1} \neq 0$, it follows that $A\zeta = 0$, where A is the infinite matrix of equation (2.16). The factorization (2.17) implies from this that $U\zeta = 0$, since L is non-singular, and this indicates that

$u_i\zeta_i + \zeta_{i+1} = 0$. Hence $-u_i = \zeta_{i+1}/\zeta_i$ and therefore any of the u_i could be very small or zero if ζ were oscillatory in nature but not, as in the tail region, when its behaviour is monotonic increasing or decreasing. In fact in the latter case, in which $\zeta_{i+1}/\zeta_i > 1$, it follows immediately that $|u_i| > 1$. A further relation can be obtained by noting that the determinant of A (truncated after the $(n-m)^{th}$ row and column), being equal to the product of all the elements u_i, is proportional to ζ_{n+1}/ζ_1, thus verifying the connection between the non-singularity of A and the exponential-type behaviour of the solutions.

Similarly, for there to be solutions of the required form in the relativistic case, the matrix A of equations (5.11) must also be non-singular, but this is not now so simply deduced from the form of the solutions or from the values of the matrix elements. The matrix, which is unsymmetric, is a band matrix of width five and the non-zero elements are defined in association with equations (5.8). The form of the potential in the tail region is such that the mesh function μ changes sign near the point r_m and then increases monotonically through the region towards a limiting value given by E. The function φ is therefore positive (bounded above and bounded away from zero) throughout the region, as also is ϑ (assuming $E < 2c^2$), and the values of each function are directly proportional to the step size. The elements w_+ and w_- are constants for a given step size, while as h decreases their values tend towards unity from above and below respectively: or vice versa, depending on the sign of k. Thus irrespective of the step size the matrix A is not necessarily diagonally dominant even with the equations ordered so as to bring the elements w_+ on to the diagonal. Nor is it block diagonally dominant [as defined according to Feingold and Varga (1962)] relative to any of the simple vector norms for any obvious partitioning of the matrix such as the one indicated by broken lines in the expression for A following equation (5.11).

However, it will be sufficient for the present purpose to show that the factored form LU has diagonal elements (t_i, u_i) bounded away from zero (which immediately implies non-singularity) and that upper bounds exist on the magnitudes of the other elements (j_i, k_i, l_i, v_i), with the implication that the factorization and the forward and backward recurrence form a set of procedures free of possible instability.

Now

$$u_i = w_+ - l_i \varphi_{i+1} \qquad (i \geq m);$$

and

$$\begin{aligned} -t_i &= w_+ + j_i v_{i-1} \qquad (i > m) \\ &= w_+ + \varphi_i v_{i-1} / u_{i-1} \\ &= w_+ + \varphi_i \left\{ \frac{\theta_i - l_{i-1} w_-}{w_+ - l_{i-1} \varphi_i} \right\}, \qquad \text{with } -t_m = w_+ \, . \end{aligned}$$

Assuming $h < |1/k|$ (so that w_+ and w_- are positive quantities) and remembering that $\theta_i > 0$ and $\varphi_i > 0$ for all $i > m$, these relations would show that $u_i > w_+$ and $-t_i \geq w_+$ for all $i \geq m$ if it could be proved that each $l_i < 0$. It would also follow that $0 < j_i < \varphi_i / w_+$ and $-w_- / w_+ < k_i < 0$ for all $i > m$.

Now

$$\begin{aligned} -l_{i+1} &= -(\theta_{i+1} - k_{i+1} v_i)/t_{i+1} \qquad (i \geq m) \\ &= \frac{\theta_{i+1} + k_{i+1} v_i}{w_+ + j_{i+1} v_i} \\ &= \frac{\theta_{i+1} + w_- v_i / u_i}{w_+ + \varphi_{i+1} v_i / u_i} \\ &= \frac{\theta_{i+1} + w_- (\theta_{i+1} - l_i w_-)/(w_+ - l_i \varphi_{i+1})}{w_+ + \varphi_{i+1} (\theta_{i+1} - l_i w_-)/(w_+ - l_i \varphi_{i+1})} \\ &= \frac{(w_+ + w_-)\theta_{i+1} - l_i (w_-^2 + \theta_{i+1} \varphi_{i+1})}{-(w_+ + w_-) l_i \varphi_{i+1} + (w_+^2 + \theta_{i+1} \varphi_{i+1})} \\ &= \frac{2\theta_{i+1} + (-l_i)(w_-^2 + \theta_{i+1} \varphi_{i+1})}{2\varphi_{i+1}(-l_i) + (w_+^2 + \theta_{i+1} \varphi_{i+1})} \, ; \end{aligned}$$

and

$$-l_m = -\theta_m / t_m = \theta_m / w_+ \, .$$

It therefore follows by induction that $l_i < 0$ for all $i \geq m$, as required. The bounds thus determined for $|k_i|$ and $|t_i|$ can be used to establish limits on the growth in magnitude of the elements l_i and v_i. Thus, for $i > m$,

$$\begin{aligned} |l_i| &= (\theta_i - k_i v_{i-1})/|t_i| \\ &= [\theta_i + |k_i|(\theta_i + |l_{i-1}|w_-)]/|t_i| \\ &< [\theta_i + \frac{w_-}{w_+}(\theta_i + |l_{i-1}|w_-)]/w_+ \\ &< (w_-/w_+)^2|l_{i-1}| + 2\theta_i/w_+^2 \\ &< (w_-/w_+)^{2(i-m)}|l_m| + \frac{2}{w_+^2}\sum_{j=m+1}^{i}(w_-/w_+)^{2(i-j)}\theta_j \end{aligned}$$

with

$$|l_m| = \theta_m/w_+ .$$

A corresponding limit is obtained for v_i by noting that

$$v_i = \theta_{i+1} + |l_i|w_- \quad \text{for all } i \geq m .$$

This analysis has shown that the elements j_i, l_i, v_i have only a limited growth as the index i increases, that every element k_i is bounded in magnitude by w_-/w_+, and that the elements t_i, u_i are bounded away from zero by w_+. This is only true for $h < |1/k|$, and only of real value as a condition of stability when $h \ll |1/k|$. Moreover it applies only for the particular ordering of the equations that was used in setting up equation (5.11) for the tail region, but this is not to say that the alternative ordering in which the quantities w_+ (or w_-) would not lie along the diagonal of A would be any less stable in practice. Certainly in the application of these methods to the solution of atomic systems no case is known of an instability having arisen whichever ordering was used.

(vi) Normalization and eigenvalue adjustment

(a) Homogeneous case

As indicated at the end of §5.1(iv), the solutions computed in the two regions $r < r_m$ and $r > r_m$ are matched at the point r_m by constructing an iterative scheme for reducing $|R|$ below a specified value, where R is the residual quantity defined in equation (5.10) as $Q_m^{(f)} - Q_m^{(t)}$. In the homogeneous case R is a function only of the estimated value E assigned to the energy parameter during the integration; a correction (ΔE) to this value may be obtained along the same lines as in the non-relativistic case by solving the variational system for $\frac{\partial}{\partial E}(P,Q)$ and finding a value for $\frac{\partial R}{\partial E}$, from which $\Delta E = -R\Big/(\frac{\partial R}{\partial E})$. However there is no need actually to solve the variational system (except after the iterative process has converged to a solution and the error analysis of §5.2 is applied); this will be demonstrated by performing an analysis on the original differential system.

The homogeneous equations were written in their original form (5.1) as:

$$\left\{\begin{array}{l} \frac{dp}{dr} + \frac{k}{r}p + (2c + \frac{v-\epsilon}{c})q = 0 \\ \\ \frac{dq}{dr} - \frac{k}{r}q - (\frac{v-\epsilon}{c})p = 0 \end{array}\right. .$$

and, differentiated with respect to ϵ, they become:

$$\left\{\begin{array}{l} \frac{\partial^2 p}{\partial\epsilon\partial r} + \frac{k}{r}\frac{\partial p}{\partial\epsilon} + (2c + \frac{v-\epsilon}{c})\frac{\partial q}{\partial\epsilon} - \frac{1}{c}q = 0 \\ \\ \frac{\partial^2 q}{\partial\epsilon\partial r} - \frac{k}{r}\frac{\partial q}{\partial\epsilon} - (\frac{v-\epsilon}{c})\frac{\partial p}{\partial\epsilon} + \frac{1}{c}p = 0 . \end{array}\right.$$

Multiplying these four equations by $\frac{\partial q}{\partial\epsilon}$, $-\frac{\partial p}{\partial\epsilon}$, $-q$, p respectively, and adding, the following result is obtained:

$$\frac{\partial p}{\partial r}\frac{\partial q}{\partial\epsilon} - \frac{\partial q}{\partial r}\frac{\partial p}{\partial\epsilon} - q\frac{\partial^2 p}{\partial\epsilon\partial r} + p\frac{\partial^2 q}{\partial\epsilon\partial r} + \frac{1}{c}(p^2+q^2) = 0 .$$

That is,

$$\frac{\partial}{\partial r}(p\frac{\partial q}{\partial\epsilon} - q\frac{\partial p}{\partial\epsilon}) + \frac{1}{c}(p^2+q^2) = 0$$

and therefore

$$\left[\, p\,\frac{\partial q}{\partial \epsilon} \;-\; q\,\frac{\partial p}{\partial \epsilon}\,\right]_{r_i}^{r_j} \;=\; -\frac{1}{c}\int_{r_i}^{r_j}(p^2+q^2)\,dr\,.$$

Assuming that for a specified value of ϵ the solution satisfies the boundary conditions (5.2) and is such that $p^{(f)}(r_m) = p^{(t)}(r_m) = p_m$ and $q^{(f)}(r_m) \neq q^{(t)}(r_m)$ [where the superscripts f,t have the same meaning as in equation (5.10)], then:

$$p_m\,\frac{\partial}{\partial\epsilon}q_m^{(f)} \;-\; q_m^{(f)}\,\frac{\partial}{\partial\epsilon}p_m \;=\; -\frac{1}{c}\int_0^{r_m}(p^2+q^2)\,dr$$

and

$$p_m\,\frac{\partial}{\partial\epsilon}q_m^{(t)} \;-\; q_m^{(t)}\,\frac{\partial}{\partial\epsilon}p_m \;=\; +\frac{1}{c}\int_{r_m}^{\infty}(p^2+q^2)\,dr\,.$$

That is,

$$\frac{\partial}{\partial\epsilon}(q_m^{(f)}/p_m) \;=\; -\frac{1}{cp_m^2}\int_0^{r_m}(p^2+q^2)\,dr$$

and

$$\frac{\partial}{\partial\epsilon}(q_m^{(t)}/p_m) \;=\; +\frac{1}{cp_m^2}\int_{r_m}^{\infty}(p^2+q^2)\,dr\,.$$

Now to match the functions $q^{(f)}$ and $q^{(t)}$ at the point r_m the required first-order correction to ϵ may be expressed as

$$\Delta\epsilon \;=\; -\,\frac{\{q_m^{(f)}/p_m \;-\; q_m^{(t)}/p_m\}}{\frac{\partial}{\partial\epsilon}\{q_m^{(f)}/p_m \;-\; q_m^{(t)}/p_m\}}$$

so that

$$\Delta\epsilon \;=\; \frac{cp_m(q_m^{(f)} \;-\; q_m^{(t)})}{\{\int_0^{r_m}+\int_{r_m}^{\infty}\}(p^2+q^2)\,dr}$$

$$=\; cp_m(q_m^{(f)} \;-\; q_m^{(t)})\,,$$

assuming the normalization (5.3). Thus, to the first order, the required correction ΔE to the estimated value E may be expressed directly in terms of the computed solution as

$$\Delta E \;=\; cP_m(Q_m^{(f)} - Q_m^{(t)}) \;=\; cP_mR\,,$$

where it should be noted that P_m is necessarily non-zero.

Each complete iteration is accompanied by a normalization of the computed function so as to satisfy the following algebraic approximation to equation (5.3):

$$S_1 + \sum_{i=1}^{n} w_i r_i (P_i^2 + Q_i^2) = 1 \quad . \tag{5.12}$$

where the w_i denote quadrature weights corresponding to Simpson's rule, and where

$$S_1 \equiv \int_0^{r_1} (p^2 + q^2)\, dr$$

so that, after substituting the series expansions (5.5),

$$S_1 = \int_0^{r_1} r^{2\gamma} \{ (p^{(0)^2} + q^{(0)^2}) + 2(p^{(0)}p^{(1)} + q^{(0)}q^{(1)})\, r + \dots \}\, dr$$

$$= \frac{(p^{(0)^2} + q^{(0)^2})}{2\gamma + 1} r_1^{2\gamma+1} + \frac{2(p^{(0)}p^{(1)} + q^{(0)}q^{(1)})}{2\gamma + 2} r_1^{2\gamma+2} + \dots .$$

(b) Inhomogeneous case

The matching process in the inhomogeneous case is concerned with the simultaneous adjustment of the estimated values of the energy parameter E and the scaling factor α so as to reduce to negligible proportions the quantities R and S defined as:

$$\begin{cases} R = Q_m^{(f)} - Q_m^{(t)} \\ S = S_1 + \sum_{i=1}^{n} w_i r_i (P_i^2 + Q_i^2) - 1 \end{cases}$$

where the notation is that of equations (5.10) and (5.12). In contrast to the situation in the homogeneous case it is now necessary to resort to a calculation of the variational systems satisfied by $\frac{\partial}{\partial E}(P,Q)$ and $\frac{\partial}{\partial \alpha}(P,Q)$. It will be assumed for this purpose that the correction terms included in the definitions of ξ and η in equations (5.8) are independent of E and α, being regarded as known functions in relation to a given iteration of the solving process. Then, by differentiating equations (5.8) with respect to E and using the fact that $(\frac{\partial \varphi}{\partial E})_i = -(\frac{\partial \theta}{\partial E})_i = \frac{1}{2}\frac{h}{c} r_i$, the first variational

system is obtained as:

$$\begin{cases} w_+(\frac{\partial P}{\partial E})_{i+1} - w_-(\frac{\partial P}{\partial E})_i + \theta_{i+1}(\frac{\partial Q}{\partial E})_{i+1} + \theta_i(\frac{\partial Q}{\partial E})_i = \frac{1}{2}\frac{h}{c}(r_{i+1}Q_{i+1} + r_iQ_i) \\ w_-(\frac{\partial Q}{\partial E})_{i+1} - w_+(\frac{\partial Q}{\partial E})_i + \varphi_{i+1}(\frac{\partial P}{\partial E})_{i+1} + \varphi_i(\frac{\partial P}{\partial E})_i = -\frac{1}{2}\frac{h}{c}(r_{i+1}P_{i+1} + r_iP_i) \end{cases}.$$

Assuming that the coefficients in the initial series expansions (5.5) have no significant dependence on E, the starting conditions for this system are obtained simply as

$$(\frac{\partial P}{\partial E})_1 = (\frac{\partial Q}{\partial E})_1 = 0,$$

and the system is solved by repeating the entire procedure for solving equations (5.8), the only difference being that ξ and η are replaced by the new right-hand sides. The result of the calculation is to obtain values for $\frac{\partial R}{\partial E}, \frac{\partial S}{\partial E}$ from their defining relations:

$$\begin{cases} \frac{\partial R}{\partial E} = (\frac{\partial Q}{\partial E})_m^{(f)} - (\frac{\partial Q}{\partial E})_m^{(t)} \\ \frac{\partial S}{\partial E} = 2\sum_{i=1}^{n} w_i r_i \{P_i(\frac{\partial P}{\partial E})_i + Q_i(\frac{\partial Q}{\partial E})_i\} \end{cases}$$

neglecting any dependence of S_1 on E.

The second variational system is obtained likewise as:

$$\begin{cases} w_+(\frac{\partial P}{\partial \alpha})_{i+1} - w_-(\frac{\partial P}{\partial \alpha})_i + \theta_{i+1}(\frac{\partial Q}{\partial \alpha})_{i+1} + \theta_i(\frac{\partial Q}{\partial \alpha})_i = 0 \\ w_-(\frac{\partial Q}{\partial \alpha})_{i+1} - w_+(\frac{\partial Q}{\partial \alpha})_i + \varphi_{i+1}(\frac{\partial P}{\partial \alpha})_{i+1} + \varphi_i(\frac{\partial P}{\partial \alpha})_i = 0 \end{cases}$$

the starting values being determined from the series expansions (5.5).

Writing these in the form $P = \alpha r^\gamma \sum_{j\geq 0} \alpha^{-1} p^{(j)} r^j$, $Q = q^{(0)} r^\gamma \sum_{j\geq 0} [q^{(0)}]^{-1} q^{(j)} r^j$,

and noting that $\frac{\partial q^{(0)}}{\partial \alpha} = q^{(0)}/\alpha$, the following expressions are obtained:

$$\begin{cases} \frac{\partial P}{\partial \alpha} = \frac{P}{\alpha} + \alpha r^\gamma \frac{\partial}{\partial \alpha}\{\sum_{j>0} \alpha^{-1} p^{(j)} r^j\} \\ \frac{\partial Q}{\partial \alpha} = \frac{Q}{\alpha} + q^{(0)} r^\gamma \frac{\partial}{\partial \alpha}\{\sum_{j>0} (q^{(0)})^{-1} q^{(j)} r^j\} \end{cases}$$

and these give $(\frac{\partial P}{\partial \alpha}) = P_1/\alpha$, $(\frac{\partial Q}{\partial \alpha}) = Q_1/\alpha$ when terms in $\frac{\partial}{\partial \alpha}(x^{(j)}/\alpha)$,

$\frac{\partial}{\partial\alpha}(y^{(j)}/\alpha)$ are neglected. The solution of the second variational system enables values for $\frac{\partial R}{\partial\alpha}$, $\frac{\partial S}{\partial\alpha}$ to be found from the relations

$$\left\{\begin{array}{l} \frac{\partial R}{\partial\alpha} = (\frac{\partial Q}{\partial\alpha})_m^{(f)} - (\frac{\partial Q}{\partial\alpha})_m^{(t)} \\ \frac{\partial S}{\partial\alpha} = \frac{2}{\alpha} S_1 + 2\sum_{i=1}^{n} w_i r_i \{P_i(\frac{\partial P}{\partial\alpha})_i + Q_i(\frac{\partial Q}{\partial\alpha})_i\} \end{array}\right.$$

the replacement of $\partial S_1/\partial\alpha$ by $\frac{2}{\alpha} S_1$ being made to the same degree of approximation as for the starting values.

The matching procedure for each iteration is concluded by calculating the corrections $(\Delta E, \Delta\alpha)$ to the current estimates of E and α, using the first-order expressions:

$$\left\{\begin{array}{l} 0 = R + \Delta E \frac{\partial R}{\partial E} + \Delta\alpha \frac{\partial R}{\partial\alpha} \\ 0 = S + \Delta E \frac{\partial S}{\partial E} + \Delta\alpha \frac{\partial S}{\partial\alpha} \; . \end{array}\right.$$

5.2 Error analysis

The method of solution adopted for the relativistic problem bears sufficient resemblance to that of the non-relativistic to allow the development of the error analysis to proceed in a similar way. The first objective therefore will be to establish a finite-difference system for the errors by comparing the discrete form of the true differential system (5.7) with the corresponding algebraic system (5.8), so taking into account the effects of truncation and physical error. The errors associated with the functions v, x, and y may be defined in the usual way:

$$\left\{\begin{array}{l} \delta v_i = v(r_i) - V_i \\ \delta x_i = x(r_i) - X_i \\ \delta y_i = y(r_i) - Y_i \; ; \end{array}\right.$$

likewise the errors associated with the computed solution:

$$\left\{\begin{array}{l} \delta p_i = p(r_i) - P_i \\ \delta q_i = q(r_i) - Q_i \\ \delta\epsilon = \epsilon - E \; . \end{array}\right.$$

The system (5.7) satisfied by the true solution may then be written:

$$\left\{\begin{aligned} &w_+p_{i+1} - w_-p_i + \{\theta_{i+1} - \tfrac{1}{2}\tfrac{h}{c}r_{i+1}(\delta\epsilon - \delta v_{i+1})\}q_{i+1} + \{\theta_i - \tfrac{1}{2}\tfrac{h}{c}r_i(\delta\epsilon - \delta v_i)\}q_i \\ &\qquad = \xi_i + \tfrac{1}{2}h(\delta x_{i+1} + \delta x_i) - C_{i+1/2}(p) + C^{(1)}_{i+1/2}(P); \\ &w_-q_{i+1} - w_+q_i + \{\varphi_{i+1} + \tfrac{1}{2}\tfrac{h}{c}r_{i+1}(\delta\epsilon - \delta v_{i+1})\}p_{i+1} + \{\varphi_i + \tfrac{1}{2}\tfrac{h}{c}r_i(\delta\epsilon - \delta v_i)\}p_i \\ &\qquad = \eta_i + \tfrac{1}{2}h(\delta y_{i+1} + \delta y_i) - C_{i+1/2}(q) + C^{(1)}_{i+1/2}(Q). \end{aligned}\right.$$

Therefore, introducing the mesh functions $\delta\xi_i$, $\delta\eta_i$ defined as

$$\left\{\begin{aligned} \delta\xi_i &= -\tfrac{1}{2}\tfrac{h}{c}(r_{i+1}q_{i+1}\delta v_{i+1} + r_iq_i\delta v_i) + \tfrac{1}{2}h(\delta x_{i+1} + \delta x_i) - C_{i+1/2}(p) + C^{(1)}_{i+1/2}(P); \\ \delta\eta_i &= +\tfrac{1}{2}\tfrac{h}{c}(r_{i+1}p_{i+1}\delta v_{i+1} + r_ip_i\delta v_i) + \tfrac{1}{2}h(\delta y_{i+1} + \delta y_i) - C_{i+1/2}(q) + C^{(1)}_{i+1/2}(Q) \end{aligned}\right.$$

(where δx and δy are absent in the homogeneous case), the system (5.7) becomes:

$$\left\{\begin{aligned} &w_+p_{i+1} - w_-p_i + \theta_{i+1}q_{i+1} + \theta_iq_i = \xi_i + \delta\xi_i + \tfrac{1}{2}\tfrac{h}{c}(r_{i+1}q_{i+1} + r_iq_i)\delta\epsilon \\ &w_-q_{i+1} - w_+q_i + \varphi_{i+1}p_{i+1} + \varphi_ip_i = \eta_i + \delta\eta_i - \tfrac{1}{2}\tfrac{h}{c}(r_{i+1}p_{i+1} + r_ip_i)\delta\epsilon . \end{aligned}\right.$$

Comparing this with the system (5.8) satisfied by the computed solution, the errors associated with this solution are seen to satisfy the following finite-difference system:

$$\left.\begin{aligned} &w_+\delta p_{i+1} - w_-\delta p_i + \theta_{i+1}\delta q_{i+1} + \theta_i\delta q_i = \delta\xi_i + \tfrac{1}{2}\tfrac{h}{c}(r_{i+1}q_{i+1} + r_iq_i)\delta\epsilon \\ &w_-\delta q_{i+1} - w_+\delta q_i + \varphi_{i+1}\delta p_{i+1} + \varphi_i\delta p_i = \delta\eta_i - \tfrac{1}{2}\tfrac{h}{c}(r_{i+1}p_{i+1} + r_ip_i)\delta\epsilon \end{aligned}\right\} . \qquad (5.13)$$

Working as usual to the first order in small quantities, the functions p, q occurring on the right-hand sides of these equations (including those appearing in the definitions of $\delta\xi, \delta\eta$) may be replaced by P, Q; in particular the expression $-C_{i+1/2}(p) + C^{(1)}_{i+1/2}(P)$ becomes $-C^{(2)}_{i+1/2}(P)$, representing the central-difference expansion $(\frac{1}{120}\delta^5 - \frac{1}{840}\delta^7 + \ldots)P_{i+1/2}$, and similarly for Q. On the same assumptions as made in establishing the starting values in the calculation of the variational systems $\frac{\partial}{\partial E}(P,Q)$ and $\frac{\partial}{\partial \alpha}(P,Q)$, the error functions may be given the following starting values:

$$\delta p_1 = \frac{\delta\alpha}{\alpha} P_1 , \quad \delta q_1 = \frac{\delta\alpha}{\alpha} Q_1 ,$$

where $\delta\alpha$ represents the error in the computed value of α.

The next step in the analysis involves expanding the functions δp, δq in terms of $\delta\epsilon$ and $\delta\alpha$ and retaining only first-order terms:

$$\left.\begin{aligned} \delta p_i &= dp_i + \left(\frac{\partial P}{\partial E}\right)_i \delta\epsilon + \left(\frac{\partial P}{\partial \alpha}\right)_i \delta\alpha \\ \delta q_i &= dq_i + \left(\frac{\partial Q}{\partial E}\right)_i \delta\epsilon + \left(\frac{\partial Q}{\partial \alpha}\right)_i \delta\alpha \end{aligned}\right\} . \qquad (5.14)$$

These expressions, in which the notation dp, dq is merely a convenient representation for the terms of zero order, are based on essentially the same assumptions as were justified at the corresponding stage in the development of the analysis for the non-relativistic problem. The expressions will be used in the same way for estimating bounds on $|\delta p|$, $|\delta q|$:

$$\left.\begin{aligned} |\delta p_i| &\le |dp_i| + \left|\left(\frac{\partial P}{\partial E}\right)_i\right| |\delta\epsilon| + \left|\left(\frac{\partial P}{\partial \alpha}\right)_i\right| |\delta\alpha| \\ |\delta q_i| &\le |dq_i| + \left|\left(\frac{\partial Q}{\partial E}\right)_i\right| |\delta\epsilon| + \left|\left(\frac{\partial Q}{\partial \alpha}\right)_i\right| |\delta\alpha| \end{aligned}\right\} . \qquad (5.15)$$

In the inhomogeneous case the functions forming the coefficients of $\delta\epsilon$ and $\delta\alpha$ in these expansions have already been computed as the solutions of variational systems during the iterative solving process. In the homogeneous case, where $\frac{\partial P}{\partial \alpha} \equiv \frac{P}{\alpha}$ and $\frac{\partial Q}{\partial \alpha} \equiv \frac{Q}{\alpha}$, the calculation of $\frac{\partial}{\partial E}(P,Q)$ has not already been performed and must therefore be included as part of the error analysis, the method of calculation being identical to that described in the context of the matching procedure for the inhomogeneous problem.

The equations satisfied by the zero-order terms dp, dq are formed from equations (5.13) by omitting the terms involving $\delta\epsilon$, so as to give:

$$\left.\begin{aligned} w_+ dp_{i+1} - w_- dp_i + \theta_{i+1} dq_{i+1} + \theta_i dq_i &= \delta\xi_i \\ w_- dq_{i+1} - w_+ dq_i + \varphi_{i+1} dp_{i+1} + \varphi_i dp_i &= \delta\eta_i \end{aligned}\right\}$$

where the functions $\delta\xi$, $\delta\eta$ are subject to the known bounds:

$$\left.\begin{aligned} |\delta\xi_i| &\le \tfrac{1}{2}\tfrac{h}{c}\left(r_{i+1}|Q_{i+1}||\delta v_{i+1}| + r_i|Q_i||\delta v_i|\right) + \tfrac{1}{2}h\left(|\delta x_{i+1}|+|\delta x_i|\right) \\ &\qquad + |c^{(2)}_{i+1/2}(P)|\,; \\ |\delta\eta_i| &\le \tfrac{1}{2}\tfrac{h}{c}\left(r_{i+1}|P_{i+1}||\delta v_{i+1}| + r_i|P_i||\delta v_i|\right) + \tfrac{1}{2}h\left(|\delta y_{i+1}|+|\delta y_i|\right) \\ &\qquad + |c^{(2)}_{i+1/2}(Q)| \end{aligned}\right\}. \tag{5.16}$$

Continuing the analogy with the non-relativistic problem, bounds can now be obtained on the elements of the mesh functions dp, dq by performing a forward error analysis as far as the match-point r_m and then a matrix inversion in the tail region.

Comparing the equations satisfied by dp, dq with the basic system (5.8) for P,Q, it is clear that the forward recurrence may be expressed in the same form as (5.9), namely:

$$f_{i+1} = G_i f_i + b_i \quad (i = 1, 2, \ldots, m-1)$$

where G_i is the matrix used in equation (5.9), and

$$f_i \equiv \begin{bmatrix} dp_i \\ dq_i \end{bmatrix}, \quad b_i \equiv \Delta^{-1}_{i+1}\begin{bmatrix} w_-\delta\xi_i - \theta_{i+1}\delta\eta_i \\ w_+\delta\eta_i - \varphi_{i+1}\delta\xi_i \end{bmatrix},$$

with $\Delta_i \equiv w_+w_- - \theta_i\varphi_i$. The recurrence has the starting values $dp_1 = dq_1 = 0$, and so the required bounds may be derived from the relations:

$$|f_i| \le \sum_{j=1}^{i-1} |D_{ij}||b_j| \quad (i = 2, 3, \ldots, m)$$

where

$$\left.\begin{aligned} D_{2,1} &= I \quad \text{(the } 2\times 2 \text{ identity matrix)}, \\ D_{ij} &= G_{i-1}D_{i-1,j} \quad (j = 1, 2, \ldots, i-2) \\ D_{i,i-1} &= I \end{aligned}\right\} \quad (i = 3, 4, \ldots, m),$$

and

$$|b_j| \le \Delta^{-1}_{j+1}\begin{bmatrix} w_-|\delta\xi_j| + |\theta_{j+1}||\delta\eta_j| \\ w_+|\delta\eta_j| + |\varphi_{j+1}||\delta\xi_j| \end{bmatrix}.$$

§5.2

The bounds corresponding to the tail region are found from the matrix system:

$$\begin{bmatrix} |dq_m| \\ |dp_{m+1}| \\ \hline |dq_{m+1}| \\ |dp_{m+2}| \\ \hline \\ \\ \hline |dq_{n-1}| \\ |dp_n| \end{bmatrix} \leq |A^{-1}| \cdot \begin{bmatrix} |\delta\eta_m| + |\varphi_m||dp_m| \\ |\delta\xi_m| + w_- |dp_m| \\ \hline |\delta\eta_{m+1}| \\ |\delta\xi_{m+1}| \\ \hline \\ \\ \hline |\delta\eta_{n-1}| \\ |\delta\xi_{n-1}| \end{bmatrix}$$

This has been written in a form corresponding to equation (5.11) and assumes that the value of $|dp_m|$ is taken from the final value $|f_m|$ computed in the forward recurrence, and that the values of $|dq_n|$, $|dp_{n+1}|$, $|dq_{n+1}|$, ... are negligibly small. As in the non-relativistic case (§3.4), the inverse of the matrix A is computed column by column using the relation $LUA^{-1} = I$; for a given number of grid points in the tail region the storage requirement for this operation is now four times as great and it is possible that a more economical formulation of A^{-1} (for example in factored form) would be called for in certain cases. As mentioned in §3.4, the storage requirement is determined by the number of grid points in the tail region: this depends on the step size and the form of the potential. The examples quoted in the next section certainly do not necessitate the implementation of a more compact formulation of A^{-1} and it is predicted that no difficulty would arise in the ultimate application of this work, which is to the calculation of atomic wave functions in the relativistic self-consistent field approximation.

The remaining analysis concerns the estimation of bounds on $|\delta\epsilon|$ and $|\delta\alpha|$ by deriving a pair of simultaneous equations for $\delta\epsilon$ and $\delta\alpha$; these will be equivalent to equations (3.20) and (3.21) developed for the inhomogeneous

form of the non-relativistic problem. Equation (3.20) was obtained from equation (3.17) by applying the series expansion (3.18); equation (3.17) was derived by multiplying each member of the set of finite-difference equations for the error vector by the corresponding element of the solution vector and summing over all such equations, then utilizing the symmetry of the system to simplify the result. The same general method will now be applied to the equations of the relativistic system in such a way as to minimize the effect of the slight asymmetry in the system.

It will be convenient to rewrite the pair of finite-difference equations (5.13) for the errors as follows:

$$\begin{bmatrix} -w_- & \theta_i \\ -\varphi_i & w_+ \end{bmatrix}\begin{bmatrix} \delta p_i \\ \delta q_i \end{bmatrix} + \begin{bmatrix} w_+ & \theta_{i+1} \\ -\varphi_{i+1} & -w_- \end{bmatrix}\begin{bmatrix} \delta p_{i+1} \\ \delta q_{i+1} \end{bmatrix} = \begin{bmatrix} \delta\xi_i \\ -\delta\eta_i \end{bmatrix} + \frac{1}{2}\frac{h}{c}\delta\epsilon\begin{bmatrix} r_i Q_i + r_{i+1}Q_{i+1} \\ r_i P_i + r_{i+1}P_{i+1} \end{bmatrix}.$$

Premultiplying by $[Q_i \; P_i]$, transposing the left-hand side, and summing over all values of i, the following result is obtained:

$$\sum [\delta p_i \; \delta q_i]\begin{bmatrix} -w_- & -\varphi_i \\ \theta_i & w_+ \end{bmatrix}\begin{bmatrix} Q_i \\ P_i \end{bmatrix} + \sum [\delta p_{i+1} \; \delta q_{i+1}]\begin{bmatrix} w_+ & -\varphi_{i+1} \\ \theta_{i+1} & -w_- \end{bmatrix}\begin{bmatrix} Q_i \\ P_i \end{bmatrix}$$

$$= \sum (Q_i\delta\xi_i - P_i\delta\eta_i) + \frac{1}{2}\frac{h}{c}\delta\epsilon \sum \{r_i(Q_i^2+P_i^2) + r_{i+1}(Q_iQ_{i+1} + P_iP_{i+1})\}. \quad (5.17)$$

In practice the terms in each summation arising from values of i outside the range $i = 1, 2, \ldots, n$ will be considered to contribute only insignificantly and be disregarded. To the same degree of approximation the first summation on the left-hand side may be replaced by

$$\sum [\delta p_{i+1} \; \delta q_{i+1}]\begin{bmatrix} -w_- & -\varphi_{i+1} \\ \theta_{i+1} & w_+ \end{bmatrix}\begin{bmatrix} Q_{i+1} \\ P_{i+1} \end{bmatrix}$$

where the range of summation is still $i = 1, 2, \ldots, n$. The two summations on the left-hand side may now be combined to give

$$\sum [\delta p_{i+1} \; \delta q_{i+1}]\left\{\begin{bmatrix} -w_- & -\varphi_{i+1} \\ \theta_{i+1} & w_+ \end{bmatrix}\begin{bmatrix} Q_{i+1} \\ P_{i+1} \end{bmatrix} + \begin{bmatrix} w_+ & -\varphi_{i+1} \\ \theta_{i+1} & -w_- \end{bmatrix}\begin{bmatrix} Q_i \\ P_i \end{bmatrix}\right\}.$$

§5.2

According to the finite-difference equations (5.8) for P, Q, this is equivalent to

$$\sum [\delta p_{i+1} \;\; \delta q_{i+1}] \begin{bmatrix} -(\varphi_{i+1} - \varphi_i)P_i - \eta_i \\ (\theta_{i+1} - \theta_i)Q_i + \xi_i \end{bmatrix} .$$

Taking this as the new left-hand side of equation (5.17) and inserting the expanded forms of δp, δq from equations (5.14), the result obtained may be written as follows:

$$0 = M + \sigma_1 \delta\epsilon + \sigma_2 \delta\alpha \qquad (5.18)$$

where

$$M = \sum \left\{ (Q_i \delta\xi_i - P_i \delta\eta_i) + [(\varphi_{i+1} - \varphi_i)P_i + \eta_i] dp_{i+1} - [(\theta_{i+1} - \theta_i)Q_i + \xi_i] dq_{i+1} \right\}$$

$$\sigma_1 = \sum \left\{ \frac{1}{2} \frac{h}{c} [r_i(Q_i^2 + P_i^2) + r_{i+1}(Q_i Q_{i+1} + P_i P_{i+1})] \right.$$

$$\left. + [(\varphi_{i+1} - \varphi_i)P_i + \eta_i]\left(\frac{\partial P}{\delta E}\right)_{i+1} - [(\theta_{i+1} - \theta_i)Q_i + \xi_i]\left(\frac{\partial Q}{\delta E}\right)_{i+1} \right\}$$

$$\sigma_2 = \sum \left\{ [(\varphi_{i+1} - \varphi_i)P_i + \eta_i]\left(\frac{\partial P}{\partial \alpha}\right)_{i+1} - [(\theta_{i+1} - \theta_i)Q_i + \xi_i]\left(\frac{\partial Q}{\partial \alpha}\right)_{i+1} \right\}$$

with $\frac{\partial P}{\partial \alpha}$ and $\frac{\partial Q}{\partial \alpha}$ replaced by P/α and Q/α in the homogeneous case.

Equation (5.18) is equivalent to equation (3.20) of the non-relativistic theory; a further relation between $\delta\epsilon$ and $\delta\alpha$, equivalent to equation (3.21), is gained from the normalization condition. Thus in both homogeneous and inhomogeneous cases the computed solution has been made to satisfy equation (5.12); expressed in the same form as this equation, the condition (5.3) for the exact solution becomes:

$$S_1' + \sum w_i r_i (p_i^2 + q_i^2) + C_s = 1$$

where C_s represents the quadrature error and $S_1' = S_1(1 + 2\frac{\delta\alpha}{\alpha})$. Combining the two equations and working to the first order in small quantities, there follows the relation

$$0 = 2 \sum w_i r_i (P_i \delta p_i + Q_i \delta q_i) + 2 S_1 \alpha^{-1} \delta\alpha + C_s .$$

Substituting the expansions (5.14), this relation assumes the form:

$$0 = N + \tau_1 \delta\epsilon + \tau_2 \delta\alpha \qquad (5.19)$$

where

$$N = \sum w_i r_i (P_i dp_i + Q_i dq_i) + \frac{1}{2} C_s$$

$$\tau_1 = \sum w_i r_i \{P_i (\frac{\partial P}{\partial E})_i + Q_i (\frac{\partial Q}{\partial E})_i\} = \frac{1}{2} \frac{\partial S}{\partial E}$$

$$\tau_2 = \frac{1}{\alpha} S_1 + \sum w_i r_i \{P_i (\frac{\partial P}{\partial \alpha})_i + Q_i (\frac{\partial Q}{\partial \alpha})_i\} = \frac{1}{2} \frac{\partial S}{\partial \alpha} .$$

In the homogeneous case, where $\frac{\partial P}{\partial \alpha} \equiv P/\alpha$ and $\frac{\partial Q}{\partial \alpha} \equiv Q/\alpha$, the condition (5.12) shows that $\tau_2 = 1/\alpha$.

The required bounds can now be obtained from the two relations (5.18) and (5.19):

$$|\delta\epsilon| = \frac{|M\tau_2 - N\sigma_2|}{|\sigma_1\tau_2 - \sigma_2\tau_1|}$$

$$|\delta\alpha| = \frac{|M\tau_1 - N\sigma_1|}{|\sigma_1\tau_2 - \sigma_2\tau_1|} .$$

The denominator can be computed exactly; if N is rewritten as

$$\sum w_{i+1} r_{i+1} (P_{i+1} dp_{i+1} + Q_{i+1} dq_{i+1}) + \frac{1}{2} C_s$$

the numerator $M\tau_j - N\sigma_j$ $(j = 1, 2)$ can be expressed in the form:

$$\tau_j \sum (Q_i \delta\xi_i - P_i \delta\eta_i) - \frac{1}{2} C_s \sigma_j$$
$$+ \sum \{[(\varphi_{i+1} - \varphi_i) P_i + \eta_i] \tau_j - w_{i+1} r_{i+1} P_{i+1} \sigma_j\} dp_{i+1}$$
$$- \sum \{[(\vartheta_{i+1} - \vartheta_i) Q_i + \xi_i] \tau_j + w_{i+1} r_{i+1} Q_{i+1} \sigma_j\} dq_{i+1} .$$

Then

$$|M\tau_j - N\sigma_j| \leq |\tau_j| \sum \{|Q_i||\delta\xi_i| + |P_i||\delta\eta_i|\} + \frac{1}{2}|C_s||\sigma_j|$$
$$+ \sum |dp_{i+1}| \left|\{(\varphi_{i+1} - \varphi_i) P_i + \eta_i\} \tau_j - w_{i+1} r_{i+1} P_{i+1} \sigma_j\right|$$
$$+ \sum |dq_{i+1}| \left|\{(\vartheta_{i+1} - \vartheta_i) Q_i + \xi_i\} \tau_j + w_{i+1} r_{i+1} Q_{i+1} \sigma_j\right| .$$

A method for computing bounds on $|dp|, |dq|$ has already been described and these bounds can be used again here. Bounds on $|\delta\xi|, |\delta\eta|$ may be computed

from the relations (5.16); the functions $\delta\xi$, $\delta\eta$ are equivalent to the function K used in the analysis of the non-relativistic system and defined by equation (4.1): the remarks made in the last three paragraphs of §4.1 relating to the estimation of bounds remain equally relevant. As in the non-relativistic case, the effect of the quadrature error C_s will be negligible compared with that of the other terms.

It is now possible to use the expressions (5.15) for the estimation of bounds on the components of the complete error vectors δp, δq.

5.3 Relativistic Coulomb functions

To test the analysis to see whether it does in fact yield realistic bounds, the example of the Coulomb field will be cited and used in the same way as it was in §4.2 in the non-relativistic case. Substituting the Coulomb potential $v(r) = Z/r$ into the homogeneous form (5.1) of the radial equations, the analytic solutions for the bound states are determined by a method which follows the same principles as the one outlined in §4.2 and is described in, for example, Bethe and Salpeter (1957, §14).

The eigenvalues (denoted as in §4.2 by the symbol λ) are obtained as

$$\lambda_{n,j} = c^2\{1 - \sqrt{1 - Z^2/(c^2N^2)}\}$$

where n, j, and N are respectively the principal, the total angular momentum, and the apparent principal quantum numbers. The subscripts indicate that the eigenvalues depend only on the quantum numbers n and j. The apparent principal quantum number has the value

$$N = \{(t+\gamma)^2 + Z^2/c^2\}^{1/2}$$

where

$$t = n - |k| = n - j - \tfrac{1}{2} \quad (\geq 0)$$

and, as in the series expansions (5.5),

$$\gamma = \sqrt{k^2 - Z^2/c^2}\ .$$

The relations between j, k, and the orbital quantum number l have already been stated in §5.1(i).

The eigenfunctions, normalized according to equation (5.3), are obtained in terms of confluent hypergeometric functions as follows:

$$p(r) = A(N+t+\gamma)^{1/2} r^{\gamma} e^{-Zr/N} \{(N-k)F_1 - tF_2\}$$

$$q(r) = A(N-t-\gamma)^{1/2} r^{\gamma} e^{-Zr/N} \{(N-k)F_1 + tF_2\}$$

where

$$F_1 = F\{-t,\ 2\gamma+1,\ 2Zr/N\}$$

$$F_2 = F\{-t+1,\ 2\gamma+1,\ 2Zr/N\}\ .$$

Since t is non-negative, the definition of the hypergeometric function [see equation (4.13)] indicates that the expressions $(N-k)F_1 \mp tF_2$ represent polynomials of degree t in r.

The scale factor α ($= p^{(0)}$) appearing in the series expansion (5.5) is obtained as the value given by $\alpha = A(N+t+\gamma)^{1/2}(N-k-t)$ where

$$A = \pm (2Z/N)^{\gamma} \left\{ \frac{Z}{2N^3(N-k)\,t!} \right\}^{1/2} \frac{\{\Gamma(2\gamma+t+1)\}^{1/2}}{\Gamma(2\gamma+1)}\ .$$

Knowing the exact solutions for a Coulomb potential corresponding to a given nuclear charge Z, it is now possible to make a comparison between the error estimates and the actual errors associated with the numerically-computed solutions of the homogeneous system (5.1). As in §4.2, separate consideration will be given to the effect on the results of truncation and physical error.

In the absence of any physical errors, only the final term will be non-zero in each of the expressions (5.16) and the error bounds which result from the analysis will arise solely from the effects of truncation error. Proceeding in the same way as in §4.2, results are quoted in Table 5.1 for the case Z = 10 using the same step size for the numerical calculation of each of several quantum states; the results are listed in a form comparable with that of Table 4.1. In contrast to the non-relativistic case, there are now two states corresponding to each of the allowed pairs of quantum numbers (n,l) for $l > 0$. Each function p(r) has n-l-1 nodes, as

it has in the non-relativistic limit; the other component q(r) has the same number of nodes for negative values of k (when $j = l + \frac{1}{2}$) and one additional node for positive k ($j = l - \frac{1}{2}$).

The results indicate a pattern of behaviour broadly similar to the findings of Table 4.1, although the calculated estimates for $||\delta p||_\infty$ and $||\delta q||_\infty$ appear to depend only on the principal quantum number and not to follow the variation shown by the actual values as the quantum number l is changed. Thus for example for n = 4 the extent to which the computed bounds provide an over-estimate of the actual errors changes from one order of magnitude to two as l increases through the allowed range. For fixed n and j, the actual values and the estimates are both slightly smaller for the state corresponding to the higher of the two allowed values of l; it should be noted that the value of the scale factor α is several orders of magnitude smaller for this state than for the other. The general trends here described are observed also in the estimated and actual values of $|\delta\lambda|$, though it should be pointed out that the actual error is particularly small in each of the states corresponding to the maximum values of l and j for each value of n, whereas the computed estimate does not reflect this characteristic. As in the non-relativistic case, the estimate for the error ($\delta\alpha$) in the scale factor becomes less pessimistic as the number of nodes decreases and n increases. Again it should be noted that the maximum estimated and actual errors in either component of the wave function only rarely occur at the same point in the range. In fact the computed bound turns out to be rather more pessimistic in some regions and rather less so in others, but taken over the whole range its general form bears a relation to that of the actual error similar to the one observed in the non-relativistic case.

The extent to which the analysis yields realistic estimates of the errors which might arise from inaccuracies in the initial data may be tested by a procedure similar to the one adopted in §4.2. Ignoring the truncation error and assuming some small uncertainty δZ in the chosen value

of Z, only the first term will now be non-zero in each of the expressions (5.16), the values $\{\delta Z/r_i\}$ being substituted for $\{\delta v_i\}$. The resulting error bounds are to be compared with the changes in the exact solutions corresponding to the replacement of Z by $Z+\delta Z$; these changes represent the maximum errors which could be induced in the solutions by an uncertainty of $\pm\delta Z$ in Z. In Table 5.2 are listed the results for the first six s-states; in each case the same value of δZ is used as in Table 4.4 and the results are entirely comparable: in fact the estimates for $||\delta p||_\infty$ show a slight improvement over their non-relativistic counterparts while those for $||\delta q||_\infty$ are even less pessimistic in relative terms.

Applied to the non-relativistic problem, slight differences in detail exist between the forms of error analysis developed for the homogeneous and inhomogeneous systems, and in §4.2 the inhomogeneous form was subjected to separate testing by constructing a simple inhomogeneous system whose analytic solution was known. By contrast, the analysis developed in §5.2 is of essentially the same form whether applied to the homogeneous or to the inhomogeneous system and no significant information will be gained by applying further tests of this sort. There is already sufficient reason to suppose that the analysis is generally capable of yielding error bounds with the same degree of realism as in the non-relativistic problem.

Table 5.1

Effect of truncation error.

Z = 10, h = 0.1

state label	n	l	j	no. of nodes in p(r)	λ calculated	α calculated	$\lvert\delta\lambda\rvert$ estimate	$\lvert\delta\lambda\rvert$ actual	$\lvert\delta\alpha\rvert$ estimate	$\lvert\delta\alpha\rvert$ actual	$\lVert\delta_p\rVert_\infty$ estimate	$\lVert\delta_p\rVert_\infty$ actual	$\lVert\delta_q\rVert_\infty$ estimate	$\lVert\delta_q\rVert_\infty$ actual
1s	1	0	1/2	0	50.066741	62.854851	$7.7_{10}-4$	$1.4_{10}-8$	$1.6_{10}-3$	$1.7_{10}-5$	$1.2_{10}-4$	$1.2_{10}-5$	$5.9_{10}-6$	$4.5_{10}-7$
2s	2	0	1/2	1	12.520771	22.274835	$7.1_{10}-4$	$8.9_{10}-5$	$3.8_{10}-3$	$8.2_{10}-5$	$5.4_{10}-4$	$5.4_{10}-5$	$7.5_{10}-6$	$8.7_{10}-7$
$2\bar{p}$	2	1	1/2	0	12.520771	0.025752	$4.1_{10}-4$	$8.9_{10}-5$	$3.0_{10}-6$	$1.3_{10}-7$	$3.9_{10}-4$	$4.5_{10}-5$	$5.0_{10}-6$	$6.3_{10}-7$
2p	2	1	3/2	0	12.504163	64.474494	$5.4_{10}-4$	$1.3_{10}-8$	$1.3_{10}-2$	$4.0_{10}-3$	$6.0_{10}-4$	$1.9_{10}-5$	$1.0_{10}-5$	$3.6_{10}-7$
3s	3	0	1/2	2	5.562731	12.126779	$9.1_{10}-4$	$2.4_{10}-4$	$8.2_{10}-3$	$2.9_{10}-4$	$1.6_{10}-3$	$1.7_{10}-4$	$1.5_{10}-5$	$1.8_{10}-6$
$3\bar{p}$	3	1	1/2	1	5.562731	0.015262	$6.6_{10}-4$	$2.4_{10}-4$	$8.2_{10}-6$	$4.0_{10}-7$	$1.4_{10}-3$	$1.6_{10}-4$	$1.2_{10}-5$	$1.3_{10}-6$
3p	3	1	3/2	1	5.557950	38.236022	$7.3_{10}-4$	$7.3_{10}-5$	$2.4_{10}-2$	$2.9_{10}-3$	$1.5_{10}-3$	$9.0_{10}-5$	$1.3_{10}-5$	$9.4_{10}-7$
$3\bar{d}$	3	2	3/2	0	5.557950	0.009492	$3.7_{10}-4$	$7.3_{10}-5$	$4.0_{10}-6$	$7.6_{10}-7$	$1.0_{10}-3$	$7.4_{10}-5$	$9.9_{10}-6$	$5.8_{10}-7$
3d	3	2	5/2	0	5.556378	28.529142	$7.9_{10}-4$	$2.9_{10}-9$	$3.5_{10}-2$	$2.1_{10}-2$	$2.5_{10}-3$	$2.7_{10}-5$	$2.7_{10}-5$	$3.3_{10}-7$
4s	4	0	1/2	3	3.127931	7.876595	$1.2_{10}-3$	$4.6_{10}-4$	$1.7_{10}-2$	$6.2_{10}-4$	$4.4_{10}-3$	$4.6_{10}-4$	$3.0_{10}-5$	$3.4_{10}-6$
$4\bar{p}$	4	1	1/2	2	3.127931	0.010181	$9.9_{10}-4$	$4.6_{10}-4$	$1.9_{10}-5$	$8.5_{10}-7$	$3.8_{10}-3$	$4.2_{10}-4$	$2.6_{10}-5$	$2.9_{10}-6$
4p	4	1	3/2	2	3.126062	25.511641	$1.0_{10}-3$	$2.4_{10}-4$	$5.0_{10}-2$	$3.3_{10}-3$	$4.0_{10}-3$	$2.8_{10}-4$	$2.3_{10}-5$	$2.3_{10}-6$
$4\bar{d}$	4	2	3/2	1	3.126062	0.007358	$6.7_{10}-4$	$2.4_{10}-4$	$1.0_{10}-5$	$1.0_{10}-6$	$3.0_{10}-3$	$2.6_{10}-4$	$1.8_{10}-5$	$1.6_{10}-6$
4d	4	2	5/2	1	3.125547	22.119592	$9.1_{10}-4$	$6.0_{10}-5$	$4.9_{10}-2$	$1.7_{10}-2$	$4.3_{10}-3$	$1.3_{10}-4$	$3.0_{10}-5$	$1.0_{10}-6$
$4\bar{f}$	4	3	5/2	0	3.125547	0.002165	$5.4_{10}-4$	$6.0_{10}-5$	$3.6_{10}-6$	$1.7_{10}-6$	$3.1_{10}-3$	$1.0_{10}-4$	$2.7_{10}-5$	$6.5_{10}-7$
4f	4	3	7/2	0	3.125260	6.988355	$1.6_{10}-3$	$1.8_{10}-8$	$3.9_{10}-2$	$2.6_{10}-2$	$9.8_{10}-3$	$3.5_{10}-5$	$8.2_{10}-5$	$3.2_{10}-7$
5s	5	0	1/2	4	2.001072	5.636310	$1.6_{10}-3$	$7.4_{10}-4$	$3.5_{10}-2$	$1.1_{10}-3$	$1.1_{10}-2$	$1.1_{10}-3$	$6.2_{10}-5$	$5.8_{10}-6$
6s	6	0	1/2	5	1.338376	4.288345	$2.1_{10}-3$	$1.1_{10}-3$	$5.1_{10}-2$	$1.8_{10}-3$	$1.7_{10}-2$	$2.4_{10}-3$	$9.0_{10}-5$	$1.0_{10}-5$

Table 5.2

Effect of physical error.

$Z = 10, \quad h = 0.1$

state label	δZ	$\lvert\delta\lambda\rvert$		$\lvert\delta\alpha\rvert$		$\lVert\delta p\rVert_\infty$		$\lVert\delta q\rVert_\infty$	
		estimate	actual	estimate	actual	estimate	actual	estimate	actual
1s	$1_{10}-5$	$1.1_{10}-4$	$1.0_{10}-4$	$2.2_{10}-4$	$9.3_{10}-5$	$2.0_{10}-5$	$1.9_{10}-6$	$7.7_{10}-7$	$1.5_{10}-7$
2s	$4_{10}-5$	$1.1_{10}-4$	$1.0_{10}-4$	$5.7_{10}-4$	$1.3_{10}-4$	$8.8_{10}-5$	$9.0_{10}-6$	$1.2_{10}-6$	$2.6_{10}-7$
3s	$9_{10}-5$	$1.2_{10}-4$	$1.0_{10}-4$	$9.6_{10}-4$	$1.6_{10}-4$	$2.0_{10}-4$	$2.3_{10}-5$	$1.9_{10}-6$	$3.4_{10}-7$
4s	$16_{10}-5$	$1.2_{10}-4$	$1.0_{10}-4$	$1.5_{10}-3$	$1.9_{10}-4$	$4.0_{10}-4$	$4.2_{10}-5$	$2.9_{10}-6$	$4.3_{10}-7$
5s	$25_{10}-5$	$1.2_{10}-4$	$1.0_{10}-4$	$2.4_{10}-3$	$2.1_{10}-4$	$7.5_{10}-4$	$6.9_{10}-5$	$4.5_{10}-6$	$5.4_{10}-7$
6s	$36_{10}-5$	$1.3_{10}-4$	$1.0_{10}-4$	$2.5_{10}-3$	$2.3_{10}-4$	$9.0_{10}-4$	$1.0_{10}-4$	$4.8_{10}-6$	$6.3_{10}-7$

CHAPTER 6

RADIAL WAVE FUNCTIONS OF THE SELF-CONSISTENT FIELD

In this chapter the error analysis of Chapter 3 is used as a basis upon which to develop a method of estimating bounds on the errors which arise in solving the equations for the structure of a many-electron atom by the numerical 'self-consistent' field technique. The discussion is restricted to the non-relativistic case.

6.1 The Hartree-Fock equations

The allowed energy states of a particle in a central field are derived in the Schrodinger theory of quantum mechanics by solving the radial wave equation (2.1). For a system which consists of a single electron in the field of a nucleus of charge Z, the potential function for the Coulomb field in which the electron moves is $v(r) = Z/r$ and the equation admits of an easily-determined analytic solution whose derivation is outlined in §4.2. Such a solution is not forthcoming in the case of an atom consisting of more than one electron in the nuclear field, for now each electron is influenced by repulsive Coulomb forces exerted by all the others. These forces are of the same order of magnitude as the nuclear field itself and their effect is too large for them to be treated in total by means of perturbation theory. However, owing to the approximately spherical symmetry of the complete electron cloud, the net field experienced by each electron is largely a function of its radial coordinate only and can be described in terms of a central force field. The residual interactions are then sufficiently small to be neglected in the first instance and applied as a perturbation at a later stage.

Each electron in this approximation is assumed to occupy an allowed quantum state in the combined fields of the nucleus and of the <u>average</u> charge distribution of the other electrons. This implies that the wave function for the complete system of electrons may be represented as a

simple product of one-electron orbitals, each itself the product of a radial function, a spherical harmonic, and a spin function, and that each of the radial functions satisfies an equation of the homogeneous form (2.1) in which the potential is a function only of the radial coordinate. However, each of these potentials depends for its derivation on the average form of the total electron distribution and therefore cannot be determined until the one-electron orbitals are known.

The method of solution adopted by Hartree (1957) involves the initial choice of a set of wave functions from which to derive a set of potentials; better approximations to the wave functions are then obtained by inserting these potentials into the radial equations and solving by the method of §2.2. The repeated application of this process, assuming it to converge, gives rise ultimately to a 'self-consistent field' in which the wave functions obtained by solving the equations are consistent with those used in deriving the potentials.

A complication arises when Pauli's exclusion principle is taken into account. This demands that the total wave function of the system be anti-symmetric with respect to interchange of any two electrons and in terms of the Hartree theory it implies that no more than one electron can occupy a given quantum state. To satisfy the exclusion principle the wave function formed in Hartree's theory as a single product of one-particle functions must be replaced by a linear combination of such products so as to guarantee complete antisymmetry; this may be arranged by choosing a determinant or a linear combination of determinants of one-electron functions each of central-field type.

With this modification the radial equations satisfied by the one-electron functions can no longer be written down directly and an appeal must be made to the variational principle to derive them. This involves finding an expression for the energy of the system in terms of these functions and requiring that it be minimized with respect to small

variations in each of the functions in turn, subject to any conditions of orthonormality which they may be required to satisfy. The procedure gives rise to a set of coupled integro-differential equations, one for each wave function; these are commonly referred to as the Hartree-Fock equations. To each one-electron function there corresponds a radial equation of the inhomogeneous type (2.22) in which the potential term $v(r)$ and the function $x(r)$ depend on the wave functions of all the other electrons. The function $x(r)$ is described as an 'exchange term': its presence results from modifying the Hartree theory to satisfy the exclusion principle and so it can be thought of as originating from the exchange of electrons between different one-particle states.

In the central-field approximation, the one-electron states may be classified according to their principal and orbital quantum numbers n,l: these are allowed the values $n = 1, 2, \ldots;\ l = 0, 1, \ldots, n-1$. A set of states with common n,l values is said to form an '(nl) group'; each group consists of $2(2l+1)$ different states and if these are all occupied by electrons the group is said to be 'complete'. Most atomic structures consist, in this approximation, of complete groups only, or of complete groups together with a few electrons in one or two incomplete groups: seldom more except perhaps in excited states. An (nl) group containing χ occupied states is written nl^{χ} and the values $0,1,2,3,4,5,\ldots$ of l are conventionally indicated in spectroscopic notation by the letters $s,p,d,f,g,h,\ldots$. The number of occupied states in each of the (nl) groups is said to specify the 'configuration' of the atom: for example the carbon atom ($Z = 6$) in its ground state has the configuration $1s^2 2s^2 2p^2$. It is implicitly assumed in the present approximation that the radial wave functions depend only on the quantum numbers n,l and do not vary between states of the same group: this is not an additional restriction except in the case of certain configurations involving incomplete groups [see for example Pratt (1956)].

§6.1

The physical and mathematical theory and the approximations involved in deriving the Hartree-Fock equations are considered in detail by Hartree (1957) and Slater (1960), and no further comment on the validity of the equations will be made here. The mathematical model which the equations represent will be regarded from the point of view of the error analysis as constituting the 'exact' system whose solution is to be determined by numerical methods.

Where necessary in the following analysis, (nl) groups will be differentiated one from another by the labels a,b,c,...; and when, for example, the letter a is being used in this way, the quantum numbers n,l will be represented by n_a, l_a and the radial wave function by $p(a;r)$.

Using this notation and writing the potential $v(a;r)$ in the form $y(a;r)/r$, the Hartree-Fock equations can be written in the general form

$$\frac{d^2p(a;r)}{dr^2} + \left\{ \frac{2}{r} y(a;r) - \epsilon_{aa} - \frac{l_a(l_a+1)}{r^2} \right\} p(a;r) = x(a;r) \quad (6.1)$$

for all a, where the potential and exchange terms are defined by relations of the form

$$y(a;r) = Z - \sum_k A^{(k)}_{aa}\, y^{(k)}(a,a;r) - \sum_{b\neq a} A^{(0)}_{ab}\, y^{(0)}(b,b;r) \quad (6.2)$$

$$x(a;r) = -\frac{2}{r} \sum_{b\neq a} \sum_k B^{(k)}_{ab}\, y^{(k)}(a,b;r)\, p(b;r) + \sum_{b\neq a} \epsilon_{ab} p(b;r) . \quad (6.3)$$

The function $y^{(k)}$, where k is a non-negative integer, represents the integral

$$y^{(k)}(a,b;r) = r \int_0^\infty \frac{r_<^k}{r_>^{k+1}}\, p(a;s)\, p(b;s)\, ds \quad (6.4)$$

where $r_<$ ($r_>$) is the lesser (greater) of r and s. The summations over k are restricted to those values which are not greater than, but have the same parity as, $l_a + l_b$. The problem of calculating the coafficients $A^{(k)}_{ab}$ and $B^{(k)}_{ab}$ is treated in detail by Slater (1960): though presented in a somewhat different form the coefficients are tabulated by Slater for a large

number of configurations.

The parameters ϵ_{ab} originate as Lagrange multipliers incorporated into the variational procedure to satisfy the orthonormality conditions; they form a matrix whose diagonal elements ϵ_{aa} are each equivalent to the energy parameter in the equation of inhomogeneous type (2.22). These diagonal elements originate from the imposition of the set of normalization conditions:

$$\int_0^\infty [p(a;r)]^2 \, dr = 1 \quad \text{for all } a. \tag{6.5}$$

The presence of the off-diagonal parameters arises from the requirement that the one-electron functions form an orthogonal set; however, most of the off-diagonal entries may usually be set to zero because functions belonging to groups with different quantum numbers l are automatically orthogonal from the properties of the spherical harmonics. In this case the conditions to be imposed are reduced to the following:

$$\left.\begin{aligned} &\int_0^\infty p(a;r)\, p(b;r)\, dr = 0 \\ &\text{for all } a,b \text{ for which } l_a = l_b,\ n_a \neq n_b. \end{aligned}\right\} \tag{6.6}$$

Again, however, this is not necessarily an additional restriction in cases in which the groups concerned are both complete, but in other cases involving incomplete groups it may be necessary to impose orthogonality constraints leading to the presence of non-zero off-diagonal parameters: expressions for these must then be derived in terms of the wave functions to provide estimates at each stage of the process of solving the Hartree-Fock equations. Further details and examples of such cases are provided by Froese (1965; 1966b; 1967a,b; 1968b).

It may be noted that the exchange term (6.3) and the terms involving the functions $y^{(k)}$ for $k>0$ in the expression for the potential (6.2) originate as a result of the use of a determinantal wave function to satisfy the exclusion principle and can be attributed to the 'exchange' of electrons between different one-particle states. If these terms are omitted, the

system of equations (6.1) reduces to that of the unmodified Hartree theory in which 'exchange effects' are neglected.

Mention has already been made of the way in which this system can be solved numerically, and the same technique can be adopted for solving the system of Hartree-Fock equations. Estimates of the potential and exchange terms are computed from estimates of the wave functions; these terms are then regarded as given functions for the purpose of solving each of the radial equations by the method of §2.3; the resulting wave functions are compared with those used in deriving the potential and exchange terms, and new estimates are formed if necessary. The process is continued until the self-consistency criterion is met - each computed wave function agreeing with its estimate to within a prescribed tolerance.

The general numerical method is described by Froese (1963) and a large number of results have been reported (1966a; 1968a). The work of this chapter is based on the modified version described by Mayers and O'Brien (1968).

6.2 Error analysis

Each iteration of the self-consistent field procedure involves recomputing the wave functions by solving in turn each radial equation (6.1). Using the same grid $\{r_i\}$ for all the functions and regarding the potential and exchange terms as having fixed tabulated values for each equation, the solution can be found by the method described in §2.3 with slight modification to allow for orthogonality conditions. The error analysis developed in Chapter 3, leading to the formulae in §4.1, is also applicable, since the error in each newly-computed wave function can be thought of as originating from errors in the potential and exchange terms and from the truncation error associated with the method of solving the equation. Indeed, it is possible to take the analysis a stage further by noting that, as a consequence of the restriction to terms of first order in the error analysis, the estimate of the error bound is linearly dependent on the various sources

of error. By this is meant that the bound on $|\delta p_i|$ computed from relation (4.10) can be expressed in the form

$$|\delta p_i| \leq \sum_j \{ \beta_{ij} |\delta y_j| + \alpha_{ij} |\delta x_j| \} + \gamma_i \qquad (6.7)$$

where the coefficients $\alpha_{ij}, \beta_{ij}, \gamma_i$ are each non-negative, the coefficient γ_i represents the effect of truncation error, the summation extends over all grid points in the range of integration, and the mesh functions $\delta y_i, \delta x_i$ are defined in the usual way:

$$\delta y_i = y(r_i) - Y_i$$
$$\delta x_i = x(r_i) - X_i$$

with Y_i and X_i representing the computed approximations to the exact potential and exchange terms. The validity of the expression becomes evident upon examining the defining relations (4.9), (4.8), and (4.1) for each of the terms appearing in (4.10); further details will be provided in §6.3.

In anticipation of the derivation of further inequalities linking the error vectors of the potentials, exchange terms, and wave functions, it is convenient at this point to rewrite the relation (6.7) in terms of a suitable vector norm and to express it in the following way:

$$||\delta p^{(\nu)}(a)|| \leq M_y(a) ||\delta y^{(\nu)}(a)|| + M_x(a) ||\delta x^{(\nu)}(a)|| + U(a) . \qquad (6.8)$$

The label a is introduced here to signify that the terms in the relation are associated with the wave functions of a particular group in the configuration. The superscript ν indicates that the errors involved are the ones associated with the values of the functions calculated from the ν^{th} iteration of the self-consistent field procedure; for the moment the coefficients M_y, M_x, and U are assumed not to vary from one iteration to the next, but this restriction will be relaxed later.

Each iteration also involves a recalculation of all the potential and exchange terms occurring in the radial equations (6.1), and these terms are computed by inserting the currently-held values of the wave functions

into the defining relations (6.2) and (6.3) - the function values concerned are the ones resulting from the previous iteration. It will become apparent when the method of calculation is considered in detail that the first-order estimates of bounds on the errors in the potential and exchange terms can each be expressed as a linear combination of errors in the tabulated functions, with an additional term representing the effect of the truncation error associated with the calculation. Applied to the γ^{th} iteration, these relationships can be written in generalized form as follows:

$$||\delta y^{(\gamma)}(a)|| \leq \sum_b N_y(a,b)||\delta p^{(\gamma-1)}(b)|| + V_y(a) \qquad (6.9)$$

$$||\delta x^{(\gamma)}(a)|| \leq \sum_b N_x(a,b)||\delta p^{(\gamma-1)}(b)|| + V_x(a) \qquad (6.10)$$

where here again the coefficients are regarded for the moment as being independent of γ.

Combining (6.8), (6.9), and (6.10), a relation is obtained between the error norms for a particular function at the γ^{th} stage and all the functions at the previous stage:

$$||\delta p^{(\gamma)}(a)|| \leq \sum_b D_{ab}||\delta p^{(\gamma-1)}(b)|| + T_a$$

where

$$\left.\begin{aligned} D_{ab} &\equiv M_y(a)N_y(a,b) + M_x(a)N_x(a,b) \\ T_a &\equiv M_y(a)V_y(a) + M_x(a)V_x(a) + U(a) \end{aligned}\right\}. \qquad (6.11)$$

The totality of such relations for all groups in the configuration can be expressed symbolically in the form

$$\underline{dp}^{(\gamma)} \leq D\underline{dp}^{(\gamma-1)} + \underline{t} \qquad (6.12)$$

where D is the square matrix $[D_{ab}]$, and $\underline{dp}^{(\gamma)}$ and $\underline{t}$ represent column vectors whose typical components are $||\delta p^{(\gamma)}(a)||$ and T_a; the size of the vectors depends on the number of occupied or partially occupied groups.

The error norms associated with the functions computed at the γ^{th} stage can now be related to the values they assume at the start of the

iterative procedure:

$$\underline{dp}^{(\gamma)} \leq D^{\gamma}\underline{dp}^{(0)} + (I + D + D^2 + \ldots + D^{\gamma-1})\underline{t}$$

where the 'zeroth iteration' ($\gamma=0$) represents the process of choosing initial estimates for the wave functions, and I represents the identity matrix. Clearly the result is of no practical value unless D is a convergent matrix, and this is the case if and only if its spectral radius is less than unity. Then $I-D$ is non-singular and the result can be written in the form

$$\underline{dp}^{(\gamma)} \leq D^{\gamma}\underline{dp}^{(0)} + (I-D)^{-1}(I-D^{\gamma})\underline{t}$$

where $D^{\gamma} \to 0$ as $\gamma \to \infty$.

In practice it is only demanded of the initial estimates of the wave functions that they be sufficiently realistic to guarantee the convergence of the process within a specified number of iterations; as the definition of self-consistency implies, the final solution is not influenced by the exact form of the starting functions. This means that if a solution is obtained at all its accuracy will not depend on that of the estimates. In terms of the theoretical error bound obtained above, the equivalent result is gained by assuming an infinite number of iterations:

$$\underline{dp} \leq (I-D)^{-1}\underline{t}\,. \qquad (6.13)$$

The bound now depends only on the values of the coefficients defined by the relations (6.11).

The practical realisation of the procedure involves terminating the iterative process when a specified condition is satisfied; for example, each newly-computed function might be required to agree with the previous estimate to, say, four decimal places at every grid point. Given suitable starting values, a criterion of this sort could easily be met after only a few (say, five or six) iterations: consequently the result (6.13) would appear to have very little validity in such a case. In practice, however, to the accuracy required, the computed values satisfying the criterion would not be affected by performing further iterations: this would simply

result in self-consistency being achieved to a greater number of figures - up to the limit imposed by the rounding errors. Thus it is reasonable to conceive of the bound (6.13) being derived directly from (6.12) on the assumption that, for all practical purposes, the computed solution is the same as if an infinite number of iterations had been performed. There is no longer any need then to regard the coefficients of relations (6.8), (6.9), and (6.10) as constants of the iterative procedure when in fact the values do change from one iteration to the next since they depend on the wave functions. To apply (6.13), only the final values will require calculation: these are values which correspond to the functions tabulated at the termination of the self-consistent field procedure.

Before the coefficients can be calculated, the vector norm introduced in the theory must be defined appropriately. Given that the outcome of the analysis for each wave function p(r) will be obtained from relation (6.13) as a bound on $||\delta p||$, the norm must be chosen so that this result can be used to determine as realistically as possible a bound on $|\delta p_i|$ at each point r_i in the range of integration. Of the simple vector norms, the only one which is readily applicable, when combined with a suitable weighting function, is the maximum norm. This is defined so that, for any mesh function $f = \{f_i\}$,

$$||f|| \equiv \max_{(i)} \frac{1}{W_i} |f_i| \qquad (6.14)$$

where $W \equiv \{W_i\}$ is a vector of positive weights, the index i ranging over all grid points within the limits of integration. In terms of this norm, the bound on $|\delta p_i|$ for all i is obtained as

$$|\delta p_i| \leq W_i ||\delta p|| . \qquad (6.15)$$

In view of the presence of the norm in each of the relations (6.8), (6.9), and (6.10), the choice of 'weights-vector' obviously plays a crucial role in determining the success or otherwise of the method. The choice particularly affects the values which are obtained for the coefficients

defined in (6.11). Since these values must be minimized as far as possible and certainly (in the cases of M_y, M_x, N_y, N_x) to the extent of ensuring that the spectral radius of the matrix D is less than unity, the first priority must be to examine the method of evaluation of the coefficients and the extent to which they depend on $\{W_i\}$. It should be noted at the outset that $\{W_i\}$ is arbitrary to the extent of a multiplicative constant, only the form and not the scale of the vector having any effect on the results.

The coefficients in relation (6.8) can be expressed in terms of those appearing in the basic relation (6.7) by deriving from (6.7) the following relation:

$$|\delta p_i| \leq \sum_j \left\{ \beta_{ij} W_j \|\delta y\| + \alpha_{ij} W_j \|\delta x\| \right\} + \gamma_i .$$

The definition $\|\delta p\| \equiv \max_{(i)} \frac{1}{W_i} |\delta p_i|$ now leads to the following expressions for the coefficients in (6.8) (with the group label omitted):

$$M_y = \max_{(i)} \left\{ \frac{1}{W_i} \sum_j \beta_{ij} W_j \right\}$$

$$M_x = \max_{(i)} \left\{ \frac{1}{W_i} \sum_j \alpha_{ij} W_j \right\}$$

$$U = \max_{(i)} \frac{1}{W_i} \gamma_i .$$

Thus M_y is calculated by performing the error analysis leading to the establishment of a bound on $|\delta p_i|$ $(i = 1, 2, \ldots, n)$ under the conditions that the vector δy is replaced by W, the vector δx set to zero, and the truncation error neglected; M_y is then evaluated as $\max_{(i)} \frac{1}{W_i} |\delta p_i|$. A value for M_x is obtained by repeating the analysis, this time setting δy to zero and substituting W for δx. Finally, U is calculated by performing a further repeat, taking into account the effect of truncation error but setting both δy and δx to zero.

Each performance of the analysis leads essentially to the determination of a bound on each component of the vector δp from which the relevant coefficient is derived as $\max_{(i)} \frac{1}{W_i} |\delta p_i|$. Clearly, therefore, the optimum situation would be one in which each performance of the analysis led to the computed bound on $|\delta p_i|$ being proportional to W_i for all i, for then the coefficients would assume the minimum possible values. This corresponds to the situation in which the weights-vector is chosen to have the same form as the vector representing the computed bound on $|\delta p|$ itself, and in such a case the norm (6.14) can be thought of as containing within its definition the maximum information about error bounds of this type. Of course, the bounds computed in connection with the evaluation of each of the three coefficients in (6.8) cannot all conform exactly to the same pattern since the conditions under which the analysis is carried out are slightly different in each case, but the vectors representing the bounds should be only marginally different and not such as to cause the values of any of the coefficients to increase significantly above the minimum values achieved in the optimum case.

Unfortunately the optimum situation cannot be achieved when the system comprises more than one occupied group. In common with the radial wave functions themselves, the vectors representing the computed bounds differ markedly between one group and another (as illustrated for example by the differences between the bounds depicted in Figures 4.1 - 4.6) and the bounds attain local maxima and minima at different points in the range. The differences are particularly apparent near the outer limit of integration, at points within the tail regions of some but not all of the functions, for then some of the computed bounds are in the final monotonic-decreasing phase while others are still in the oscillatory phase.

To each group there corresponds one expression of the type (6.8) in which each coefficient depends for its computed value on the relation

between the weights-vector W and a vector representing a bound (computed under the conditions mentioned previously) on $|\delta p|$ for the appropriate group. Since W must be chosen so as to minimize the values of all the coefficients, its form must reflect as closely as possible the essential characteristics of each of the vectors of error bounds from which the coefficients are derived. It is reasonable to suppose that this might be achieved by choosing a suitable linear combination of these vectors, or by scaling them appropriately and selecting at each point the largest of the corresponding components of the scaled vectors. In choosing W in this way, the intention would be to prevent any of its components from falling in value far below the corresponding components of any of the vectors of error bounds; this in turn would prevent any significant rise in value of any of the coefficients. A further reason for exercising care in the choice of W is that the chosen form imprints itself through relation (6.15) on all the final error bounds, the value of $||\delta p||$ computed from (6.13) acting merely as a scaling factor in each case. In a sense, then, the weights-vector should resemble an envelope function: Figure 6.1 illustrates the general idea by showing how W might be constructed from the types of error bound typically computed in calculating the coefficients corresponding to each of two groups (a and b) in a system.

The selection of an appropriate weights-vector involves computing, for each group, at least one vector of error bounds resembling the ones which are ultimately computed in deriving the values of the coefficients in (6.8). One of these error vectors depends for its derivation on the truncation error alone and is the only one whose calculation does not involve a prior knowledge of the weights-vector itself. The set of such vectors (one for each group) can therefore be used in the first instance for selecting an appropriate weights-vector in the manner described.

§6.2

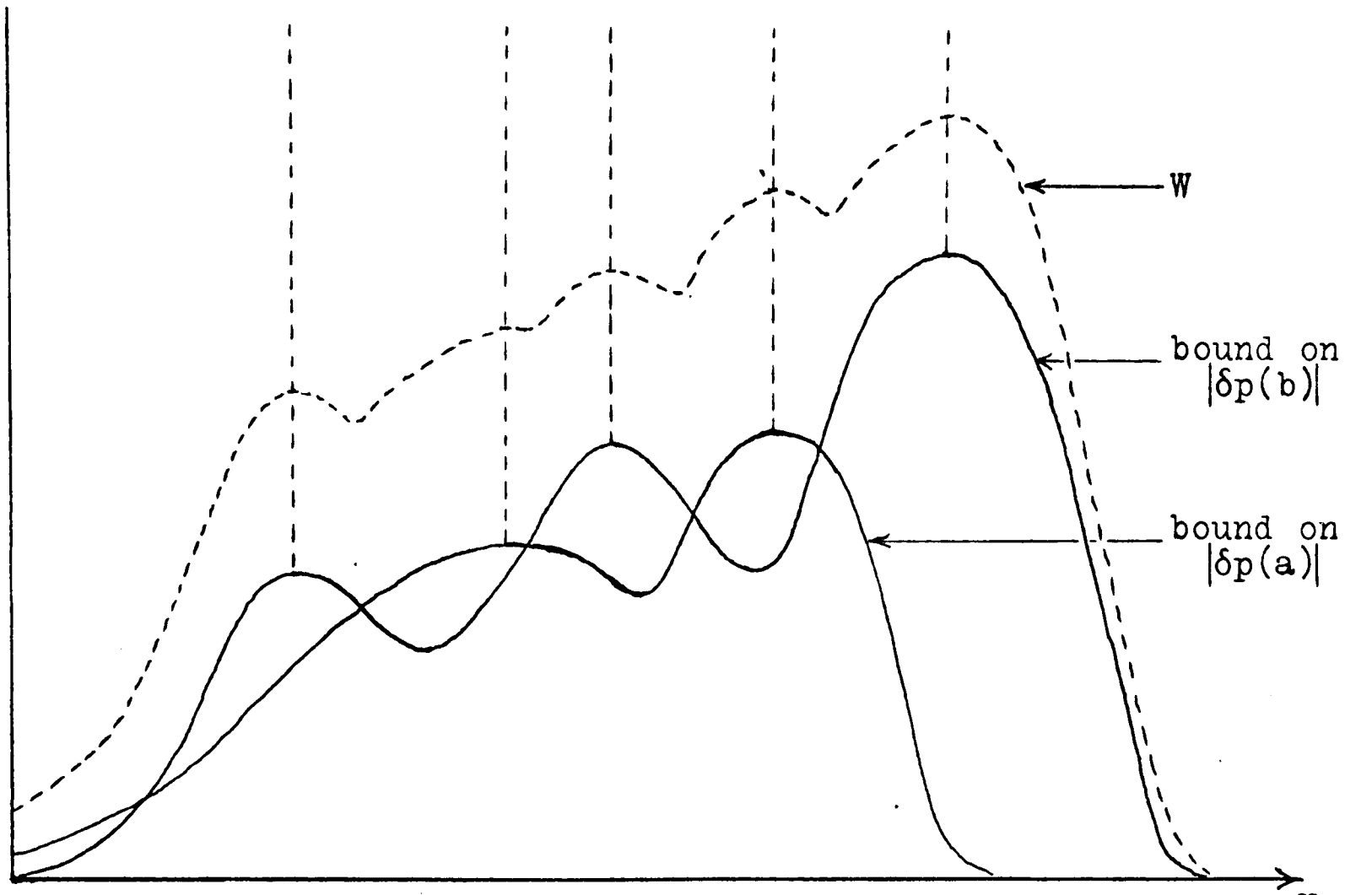


Fig. 6.1

In a particular group, the error vectors from which the coefficients M_y, M_x are calculated might be found to be rather different in pattern from the vector arising from truncation error and it might be possible to effect a reduction in the computed values of the coefficients by using a weights-vector corresponding more closely to these vectors. A knowledge of W is itself required for calculating the vectors and this could be provided by regarding the vector arising from truncation error as a first estimate of W and using it in performing the analysis leading to the second estimate.

It is evident that the selection of a weights-vector giving rise to optimum values for the coefficients in a particular case is largely a matter of experimentation, the primary aim being to ensure that the matrix D is convergent. To this end, the values of the coefficients N_y, N_x appearing

in relations (6.9) and (6.10) must also be minimized to the fullest possible extent. A detailed description of the method of calculating the coefficients N_y, N_x, V_y, V_x will be presented in §6.4 and it is sufficient to state here that values for the coefficients are derived in a similar way to those for M_y, M_x, U by computing certain error vectors and in each case taking the maximum over the set of vector elements divided by the corresponding elements of W. It is imperative, then, to ensure that these vectors of error bounds are also computed in such a way as to resemble the chosen weights-vector as closely as possible; this means for instance that towards each end of the range of integration the computed bounds must fall off to zero at least as fast as W itself.

Before proceeding to the discussion of specific examples to show how the procedure might work in practice, the two following sections are devoted to a more detailed consideration of the numerical methods involved in solving the Hartree-Fock equations and to the development of the related error analysis. This will lead to the derivation of procedures for computing all the coefficients appearing in (6.11).

It will be recalled that, in solving a particular problem, the general idea is to apply these procedures after self-consistency has been achieved to within the prescribed tolerance. This will involve performing a further iteration incorporating sufficient of the analysis to facilitate the estimation of an appropriate weights-vector; then, using this vector, the full analysis will be included in a final iteration leading to the estimation of error bounds on the computed solutions.

6.3 The radial equations

As mentioned at the beginning of the previous section, each of the radial equations (6.1) is solved by regarding the potential and exchange terms as fixed tabulated functions and applying the method described in §2.3 for solving the inhomogeneous Schrodinger equation. It is assumed that

values are also tabulated for the first few coefficients in the series expansions [equations (2.3), (2.23)] of the potential and exchange terms. In practice these expansions need not be taken further than terms of order r^2:

$$\left.\begin{aligned} y(r) &= Z + y^{(1)} r + y^{(2)} r^2 + O(r^3) \\ x(r) &= r^{l+1}[x^{(0)} + x^{(1)} r + O(r^2)] \end{aligned}\right\} \qquad (6.16)$$

and this allows values to be derived for the coefficients of terms up to r^{l+4} in the series expansion (2.14) of $p(r)$.

The method of §2.3 must be modified to allow for the possibility of orthogonality conditions being imposed on the wave function to be computed. Thus in addition to the normalization condition (6.5), the function may be required to satisfy one or more orthogonality conditions of the form (6.6). These may be written

$$\int_0^\infty p(a;r)\, p(b;r)\, dr = 0 \qquad (b \neq a)$$

where $p(a;r)$ represents the wave function to be computed and $p(b;r)$ represents each of the other functions to which $p(a;r)$ is to be made orthogonal. To each condition there corresponds, as explained in §6.1, a non-zero off-diagonal parameter ϵ_{ab}, and an initial estimate of each of these is required in addition to the estimates of the scaling factor and energy parameter ordinarily required for obtaining an approximate solution.

Writing E_b, E, and A (or A_a) for the estimates of ϵ_{ab}, ϵ_{aa}, and the scaling factor, with associated errors $\delta\epsilon_b \equiv \epsilon_{ab} - E_b$, $\delta\epsilon \equiv \epsilon_{aa} - E$, and δA, the values of E_b as well as of E and A must be adjusted so as to satisfy the orthogonality and normalization conditions and reduce to zero the residual quantity R in the match-point equation (2.28). Satisfying the normalization condition involves reducing to zero the residual quantity S in the normalization equation (2.29). A similar equation, involving a further residual quantity S_b, can be established for each of the orthogonality conditions:

$$S_{1,b} + \sum_{i=1}^{n} w_i r_i (r_i^{1/2} Q_i / b_i) P_i(b) = S_b \qquad (b \neq a)$$

where $P_i(b)$ represents the currently-tabulated approximation to $p(b;r_i)$ and, in analogy with the definition (2.21) of S_1, the quantity $S_{1,b}$ is defined as follows:

$$S_{1,b} = A_a A_b r_1^{l_a+l_b+3} \left\{ \frac{1}{l_a+l_b+3} + \dots \right\}.$$

The procedure for finding improved estimates of E, A, and the E_b therefore involves reducing R, S, and the S_b to zero. First-order corrections (ΔE, ΔA, ΔE_b) are found by solving the set of equations

$$\left\{ \begin{array}{l} 0 = R + \Delta E \frac{\partial R}{\partial E} + \Delta A \frac{\partial R}{\partial A} + \sum_{b \neq a} \Delta E_b \frac{\partial R}{\partial E_b} \\ 0 = S + \Delta E \frac{\partial S}{\partial E} + \Delta A \frac{\partial S}{\partial A} + \sum_{b \neq a} \Delta E_b \frac{\partial S}{\partial E_b} \\ 0 = S_c + \Delta E \frac{\partial S_c}{\partial E} + \Delta A \frac{\partial S_c}{\partial A} + \sum_{b \neq a} \Delta E_b \frac{\partial S_c}{\partial E_b} \qquad (c \neq a) \end{array} \right.$$

where the label c, as b, ranges over all groups whose functions are required to be orthogonal to $p(a;r)$.

Values for the partial derivatives are found from a calculation of the following variational systems (the notation is that of §2.3):

$$\left\{ \begin{array}{l} T_i\left(\frac{\partial Q}{\partial E}\right) = Q_i (h r_i / b_i)^2 + \delta_{im} \frac{\partial R}{\partial E} \\ \qquad \text{with starting values } \left(\frac{\partial Q}{\partial E}\right)_i = -\frac{1}{12} h^2 r_i^2 Q_i / b_i \quad (i = 1, 2), \\ 2 \sum w_i (r_i/b_i)^2 Q_i \left\{ \left(\frac{\partial Q}{\partial E}\right)_i + \frac{1}{12} h^2 r_i^2 Q_i / b_i \right\} = \frac{\partial S}{\partial E}, \\ \sum w_i r_i^{3/2} \left\{ \left(\frac{\partial Q}{\partial E}\right)_i + \frac{1}{12} h^2 r_i^2 Q_i / b_i \right\} P_i(c)/b_i = \frac{\partial S_c}{\partial E} \qquad (c \neq a); \end{array} \right.$$

$$T_i\left(\frac{\partial Q}{\partial A}\right) = \delta_{im}\frac{\partial R}{\partial A}$$

with starting values $\left(\frac{\partial Q}{\partial A}\right)_i = Q_i/A \quad (i = 1,2)$,

$$2S_1/A + 2\sum w_i(r_i/b_i)^2 Q_i\left(\frac{\partial Q}{\partial A}\right)_i = \frac{\partial S}{\partial A},$$

$$S_{1,c}/A + \sum w_i r_i^{3/2}\left(\frac{\partial Q}{\partial A}\right)_i P_i(c)/b_i = \frac{\partial S_c}{\partial A} \quad (c \neq a);$$

$$T_i\left(\frac{\partial Q}{\partial E_b}\right) = \left(\frac{\partial G}{\partial E_b}\right)_i + \delta_{im}\frac{\partial R}{\partial E_b}$$

with starting values $\left(\frac{\partial Q}{\partial E_b}\right)_i = 0$,

$(b \neq a)$

$$2\sum w_i(r_i/b_i)^2 Q_i\left(\frac{\partial Q}{\partial E_b}\right)_i = \frac{\partial S}{\partial E_b},$$

$$\sum w_i r_i^{3/2}\left(\frac{\partial Q}{\partial E_b}\right)_i P_i(c)/b_i = \frac{\partial S_c}{\partial E_b} \quad (c \neq a),$$

where $\left(\frac{\partial G}{\partial E_b}\right)_i$ is found from the definition (2.26) to take the form

$$\left(\frac{\partial G}{\partial E_b}\right)_i = \frac{1}{12}h^2\left\{ r_{i+1}^{3/2}\left(\frac{\partial X}{\partial E_b}\right)_{i+1} + 10 r_i^{3/2}\left(\frac{\partial X}{\partial E_b}\right)_i + r_{i-1}^{3/2}\left(\frac{\partial X}{\partial E_b}\right)_{i-1} \right\} \qquad (6.17)$$

with $\left(\frac{\partial X}{\partial E_b}\right)_i = P_i(b)$ from relation (6.3).

Having obtained corrections to the parameters, the whole procedure is repeated as many times as necessary to reduce each of the residual quantities to an insignificant value. Thus each of the radial equations is solved by performing an 'inner' iteration at each stage in the overall iterative procedure leading to self-consistency.

In applying the error analysis developed in Chapter 3, the assumed series expansion (3.18) for the first-order error vector must now be extended to include terms involving the off-diagonal parameters:

$$\delta q = \delta q^{(0)} + \delta\epsilon\frac{\partial Q}{\partial E} + \delta A\frac{\partial Q}{\partial A} + \sum_{b\neq a}\delta\epsilon_b\frac{\partial Q}{\partial E_b}.$$

The vector δg, in the form in which it appears in the matrix equation (3.16) for δq, is also dependent on the errors in the off-diagonal parameters:

it is convenient to separate out this dependence and to replace δg by

$$\delta g + \sum_{b \neq a} \delta \epsilon_b \frac{\partial G}{\partial E_b}$$

where the elements of the vector $\frac{\partial G}{\partial E_b}$ are defined by (6.17) and those of the newly-defined δg by

$$\delta g_i = \frac{1}{12} h^2 (r_{i+1}^{3/2} \delta x_{i+1} + 10 r_i^{3/2} \delta x_i + r_{i-1}^{3/2} \delta x_{i-1}) \qquad (6.18)$$

where $\{\delta x_i\}$ represents the error in the exchange term (6.3) arising from the errors in the wave functions but not from the errors in the off-diagonal parameters. The definition (6.18) applies to the term δg_i which appears in expression (4.1) for the function K_i, and the quantity $\delta \varphi_i$ appearing in (4.1) is defined in terms of the error in the potential term (6.2) as

$$\delta \varphi_i = 2 r_i \delta y_i \, . \qquad (6.19)$$

The two equations (4.3) and (4.7) which were used in §4.1 for deriving bounds on $|\delta \epsilon|$ and $|\delta A|$ are now extended into the following:

$$0 = \sigma_0 + \sigma_1 \delta \epsilon + \sigma_2 \delta A + \sum_{b \neq a} \sigma_b \delta \epsilon_b \, . \qquad (6.20)$$

$$0 = \tau_0 + \tau_1 \delta \epsilon + \tau_2 \delta A + \sum_{b \neq a} \tau_b \delta \epsilon_b \qquad (6.21)$$

where

$$\sigma_b = \sum_{i=1}^{n} \{ Q_i (\frac{\partial G}{\partial E_b})_i - G_i (\frac{\partial Q}{\partial E_b})_i \}$$

and

$$\tau_b = \frac{1}{2} \frac{\partial S}{\partial E_b} \, .$$

Equation (4.7) was derived by comparing the conditions for normalization of the exact and computed solutions, and similar equations can be established from the orthogonality conditions: their derivation is exactly comparable and leads to the following result (disregarding quadrature errors):

$$0 = \eta_c + \eta_{c,1} \delta \epsilon + \eta_{c,2} \delta A + \sum_{b \neq a} \eta_{c,b} \delta \epsilon_b \quad (c \neq a) \qquad (6.22)$$

where

$$\eta_0 = \sum_{i=1}^{n} w_i r_i^{3/2} \{\delta q_i^{(0)} - \frac{h^2}{12} \frac{Q_i}{b_i} \cdot \delta\varphi_i\} P_i(0)/b_i$$

$$\eta_{c,1} = \frac{\partial S_c}{\partial E}$$

$$\eta_{c,2} = \frac{\partial S_c}{\partial A}$$

$$\eta_{c,b} = \frac{\partial S_c}{\partial E_b} \quad .$$

Bounds on the magnitudes of $\delta\epsilon$, δA, and the $\delta\epsilon_b$ can now be obtained by solving the set of equations (6.20), (6.21), and (6.22). Denoting the inverse of the matrix of coefficients by $[H_{jk}]$, the estimated bounds can be represented as follows:

$$|\delta\epsilon| = L_1$$

$$|\delta A| = L_2$$

$$|\delta\epsilon_b| = L_j \quad (j > 2)$$

with

$$L_j \equiv |H_{j,1}\sigma_0 + H_{j,2}\tau_0 + \sum_{k>2} H_{jk}\eta_c|$$

where the values 3,4,... of the indices j and k correspond to each of the groups whose functions are to be orthogonal to the function $p(a;r)$. Inserting the expressions previously derived for σ_0, τ_0, and the η_c, and rearranging terms so as to distinguish between the various sources of error, the bounds can be expressed in the following way, comparable with the bounds (4.8) of §4.1:

$$L_j \leq |H_{j,1}| \sum_{i=1}^{n} |Q_i||C_i(Q/b)| + |H_{j,1}| \sum_{i=1}^{n} |Q_i||\delta g_i|$$

$$+ h^2 \sum_{i=1}^{n} \frac{|Q_i|}{b_i^2} \left| Q_i H_{j,1} + \frac{w_i}{12}\left\{\frac{r_i^2 Q_i}{b_i} H_{j,2} + r_i^{3/2} \sum_{k>2} P_i(c) H_{jk}\right\}\right| |\delta\varphi_i|$$

$$+ \sum_{i=1}^{n} \left| -G_i H_{j,1} + \frac{w_i}{b_i}\left\{\frac{r_i^2 Q_i}{b_i} H_{j,2} + r_i^{3/2} \sum_{k>2} P_i(c) H_{jk}\right\}\right| |\delta q_i^{(0)}| \; .$$

§6.3

As in §4.1, bounds are computed for the elements $|\delta q_i^{(0)}|$ by solving the system of equations (4.9) by the method of §3.4. The bound finally obtained for the error in the wave function is represented by an extension of expression (4.10):

$$|\delta p_i| \leq \frac{r_i^{1/2}}{b_i^2} \left\{ b_i |\delta q_i^{(0)}| + \frac{h^2}{12} |Q_i||\delta \varphi_i| + |\delta \epsilon| \left| b_i \left(\frac{\partial Q}{\partial E}\right)_i + \frac{h^2}{12} r_i^2 Q_i \right| + b_i |\delta A| \left|\left(\frac{\partial Q}{\partial A}\right)_i\right| + \sum_{b \neq a} b_i |\delta \epsilon_b| \left|\left(\frac{\partial Q}{\partial E_b}\right)_i\right| \right\}.$$

Essentially this is equivalent to expression (6.7) of the previous section, and values may be derived for the coefficients in the alternative expression (6.8) by noting from (6.19) and (6.18) that

$$|\delta \varphi_i| \leq 2 r_i W_i ||\delta y||$$

and

$$|\delta g_i| \leq \frac{h^2}{12} (r_{i+1}^{3/2} W_{i+1} + 10 r_i^{3/2} W_i + r_{i-1}^{3/2} W_{i-1}) ||\delta x|| .$$

Using these expressions wherever the terms $\delta \varphi_i$ and δg_i occur, and setting to zero the terms $C_i(Q/b)$ due to truncation error, the replacement of $||\delta y||$ and $||\delta x||$ by unity and zero (or vice versa) leads to the coefficient M_y (or M_x) being determined as $\max\limits_{(i)} \frac{1}{W_i} |\delta p_i|$ on the assumption that a weights-vector has been chosen.

The coefficient U is calculated by setting the terms $\delta \varphi_i$ and δg_i to zero wherever they occur, leaving truncation error as the only source of error; then $U = \max\limits_{(i)} \frac{1}{W_i} |\delta p_i|$.

The same analysis may be used for finding a suitable weights-vector, and in this case the vector is obtained in the first instance as $W_i = |\delta p_i|$ for all i, computed under the same conditions as specified for calculating the coefficient U. As mentioned in §6.2, it might be advantageous in some cases to 'refine' W by repeating the analysis, in which case the conditions assumed are those specified for calculating M_y (or M_x), and the first estimate of W is fed back into the analysis to derive the second.

6.4 The potential and exchange terms

As indicated earlier, the potential and exchange terms occurring in each Hartree-Fock equation are computed afresh at each stage of the self-consistent field procedure, making use of the function values calculated from the previous iteration. The method of computing these terms must now be considered in detail and an error analysis developed in such a way as to minimize the calculated values of the coefficients appearing in relations (6.9) and (6.10).

As seen from the definitions (6.2) and (6.3), the problem of computing the terms is essentially that of tabulating the functions $y^{(k)}(a,b;r)$ defined by (6.4). Each of these functions may be expressed as the sum of two integrals:

$$y^{(k)}(a,b;r) = \int_0^r (\frac{s}{r})^k \, p(a;s)\, p(b;s)\, ds + \int_r^\infty (\frac{r}{s})^{k+1} \, p(a;s)\, p(b;s)\, ds .$$

However, the function is not evaluated directly from the integrals but rather as the solution of a pair of differential equations each of which has a simple integrating factor. This leads to a fast and accurate algorithm for computing $y^{(k)}$, which is an advantage in view of the rate at which the number of such tabulations to be performed increases with the size of the atomic configuration.

It is convenient to express each function $y^{(k)}(a,b;r)$ in terms of two auxiliary functions

$$y^{(k)} = z^{(k)} + j^{(k)}$$

where

$$z^{(k)}(a,b;r) \equiv \int_0^r (\frac{s}{r})^k \, p(a;s)\, p(b;s)\, ds$$

$$j^{(k)}(a,b;r) \equiv \int_r^\infty (\frac{r}{s})^{k+1} \, p(a;s)\, p(b;s)\, ds .$$

Differentiating with respect to r, and performing the same transformation $t = \ln r$ [equation (2.5)] as used in solving the radial equations,

a set of first-order differential equations with constant coefficients is derived for evaluating $z^{(k)}$ or $j^{(k)}$ and then $y^{(k)}$:

$$\left.\begin{aligned} \frac{dz^{(k)}}{dt} + k\,z^{(k)} &= r\,p(a;r)\,p(b;r) \\ \frac{dy^{(k)}}{dt} - (k+1)\,y^{(k)} &= -(2k+1)\,z^{(k)} \end{aligned}\right\} \qquad (6.23)$$

$$\left.\begin{aligned} \frac{dj^{(k)}}{dt} - (k+1)\,j^{(k)} &= -r\,p(a;r)\,p(b;r) \\ \frac{dy^{(k)}}{dt} + k\,y^{(k)} &= (2k+1)\,j^{(k)} \end{aligned}\right\} \qquad (6.24)$$

The boundary conditions appropriate to each set of equations are

$$z^{(k)}(a,b;0) = 0, \quad y^{(k)}(a,b;r) \to z^{(k)}(a,b;r) \text{ as } r \to \infty$$

and

$$j^{(k)}(a,b;r) \to 0 \text{ as } r \to \infty, \quad y^{(k)}(a,b;0) = j^{(k)}(a,b;0) .$$

Since k is a non-negative integer, the step-by-step integration of equations (6.23) will be stable if the equation for $z^{(k)}$ is integrated in the outward direction, starting from the boundary condition at $r=0$, and the equation for $y^{(k)}$ integrated in the direction of decreasing r, starting at some sufficiently large upper limit $r = r_u$ where $y^{(k)}$ can be set equal to $z^{(k)}$. Alternatively, $y^{(k)}$ may be computed by solving the pair of equations (6.24), integrating inwards for $j^{(k)}$ from the boundary value $j^{(k)}(a,b;r_u) = 0$ and then outwards for $y^{(k)}$ from the value $y^{(k)} = j^{(k)}$ at $r=0$. In either case the upper limit r_u is chosen at least as far out as the limit of integration r_n of each of the radial equations: then the function values are negligibly small for $r \geq r_u$ and the boundary conditions for $r \to \infty$ may be assumed to hold with equal validity at $r = r_u$.

Each equation is of the general form

$$\frac{df}{dt} + c\,f = g(t)$$

where c is a constant and the function g(t) is tabulated over the logarithmic grid of points $\{t_i\}$. The solution satisfies

$$e^{ct_i} f(t_i) = \int^{t_i} e^{ct} g(t)\,dt + \text{constant}$$

so that

$$e^{ct_{i+1}} f(t_{i+1}) - e^{ct_i} f(t_i) = \int_{t_i}^{t_{i+1}} \varphi(t)\,dt \qquad (6.25)$$

where $t_{i+1} = t_i + h$ and $\varphi(t) = e^{ct} g(t)$. Mayers and O'Brien (1968) approximate to the integral by using the trapezium rule and the first term in the central-difference expansion of the associated error but neglecting the other terms. Including all the terms, the complete formula may be written

$$\int_{t_i}^{t_{i+1}} \varphi(t)\,dt = \tfrac{1}{2} h(\varphi_i + \varphi_{i+1}) + h\{-\tfrac{1}{12}\mu\delta^2_{i+1/2} + C_{i+1/2}\}\varphi$$

where

$$\mu\delta^2_{i+1/2}\varphi \equiv \tfrac{1}{2}(\varphi_{i+2} - \varphi_{i+1} - \varphi_i + \varphi_{i-1})$$

and

$$C_i \equiv \tfrac{11}{720}\mu\delta_i^4 - \tfrac{191}{60480}\mu\delta_i^6 + \ldots \,. \qquad (6.26)$$

Substitution of this expression into (6.25) gives rise to the following one-step recurrence relation for the solution:

$$e^{ch} f_{i+1} - f_i = -\tfrac{h}{24}\{e^{2ch} g_{i+2} - 13e^{ch} g_{i+1} - 13g_i + e^{-ch} g_{i-1}\}$$

$$+ he^{-ct_i} C_{i+1/2}(e^{ct} g). \qquad (6.27)$$

In applying this formula, the correction term (6.26) is omitted: it therefore constitutes the local truncation error whose effect will be taken into account in performing the error analysis.

When solving for $y^{(k)}$ from the pair of equations (6.23), the outward integration for $Z^{(k)}$ (where capital letters again denote the quantities actually computed) proceeds from the starting value

$$Z_1^{(k)} = \int_0^{r_1} \left(\frac{r}{r_1}\right)^k P(a;r)\,P(b;r)\,dr$$

where, from the series expansion (2.14),

$$P(a;r)\,P(b;r) = A_a A_b r^{l_a+l_b+2}\{1 + (\alpha_a^{(1)}+\alpha_b^{(1)})\,r + \ldots\}$$

so that

$$Z_1^{(k)} = A_a A_b r_1^{l_a+l_b+3}\left\{\frac{1}{l_a+l_b+3+k} + \frac{\alpha_a^{(1)}+\alpha_b^{(1)}}{l_a+l_b+4+k}\,r_1 + \ldots\right\}. \qquad (6.28)$$

Having calculated the grid values of $Z^{(k)}$, the inward integration for $Y^{(k)}$ takes $Y_u^{(k)} = Z_u^{(k)}$ as its starting value.

Using the alternative pair of equations (6.24), the inward integration for $J^{(k)}$ starts from the value $J_u^{(k)} = 0$ and the outward integration for $Y^{(k)}$ from the value $Y_1^{(k)} = Z_1^{(k)} + J_1^{(k)}$ where $Z_1^{(k)}$ is obtained as above.

In using these recurrence relations, any wave function values required at grid points beyond the usual limits of integration may be obtained at the inner limit from the series expansions and at the outer limit by assuming that the functions have become vanishingly small.

To determine values for the coefficients appearing in the series expansions (6.16), a similar expansion needs to be established for each $Y^{(k)}$ function. This is easily achieved by writing the function in the following form:

$$Y^{(k)} = Z^{(k)} + r^{k+1} I^{(k+1)} - Z^{(-k-1)}$$

where

$$I^{(k)}(a,b) \equiv \int_0^\infty s^{-k} P(a;s)\,P(b;s)\,ds\,.$$

Using the expansion (6.28) for $Z^{(k)}$ and $Z^{(-k-1)}$, it follows that for small values of r

$$Y^{(k)} = r^{k+1} I^{(k+1)} + O(r^{l_a+l_b+3})\,.$$

Since terms of order r^3 and higher in this expansion do not affect those of order r^2 or less in the expansions (6.16), only one term needs to be computed if $k = 0$ or 1 - namely the coefficient $I^{(k+1)}$ - and all terms may be neglected if $k > 1$. The value of $I^{(k+1)}$ may be computed to sufficient accuracy for this purpose by using Simpson's rule.

The error analysis associated with the computation of the $y^{(k)}$ functions must be developed into a method which ensures that the calculated values for the coefficients in each relation of the form (6.9) and (6.10) are as small as possible. Defining error vectors $\delta y^{(k)}$, $\delta z^{(k)}$, and $\delta j^{(k)}$ in the usual way, the analysis begins with the derivation of the recurrence relations which the components of these vectors satisfy. Each is of the general form obtained from subtracting from the recurrence (6.27) the same recurrence but with f and g replaced by their computed values F and G and the correction term omitted. Writing δf and δg for f-F and g-G, the relation which results can be written as follows:

$$e^{ch}\delta f_{i+1} - \delta f_i = -\frac{h}{24}\{e^{2ch}\delta g_{i+2} - 13e^{ch}\delta g_{i+1} - 13\delta g_i + e^{-ch}\delta g_{i-1}\}$$

$$+ h e^{-ct_i} C_{i+1/2}(e^{ct} G)$$

where the function g appearing in the correction term in the relation (6.27) has been replaced by G, in line with the usual approximations made in developing the first-order theory.

It would appear that bounds on the magnitudes of the grid values of $\delta y^{(k)}$ can be computed by performing a forward error analysis, recurring outwards for $\delta z^{(k)}$ or inwards for $\delta j^{(k)}$ (starting from $\delta z_1^{(k)} = 0$ or $\delta j_u^{(k)} = 0$) and then recurring for $\delta y^{(k)}$ either inwards from the value $\delta y_u^{(k)} = \delta z_u^{(k)}$ or outwards from the value $\delta y_1^{(k)} = \delta j_1^{(k)}$. The recurrence for $\delta z^{(k)}$ or $\delta j^{(k)}$ involves a sum of multiples of terms such as $P_i(a)\delta p_i(b) + P_i(b)\delta p_i(a)$ (to the first order in errors); it also involves the correction term (6.26) operating on $e^{kt_r} P(a) P(b)$ or $e^{-(k+1)t_r} P(a) P(b)$ respectively, while the recurrence for $\delta y^{(k)}$ as a function of $\delta z^{(k)}$ or $\delta j^{(k)}$ involves the correction term operating on $e^{-(k+1)t} z^{(k)}$ or $e^{kt} J^{(k)}$. If the terms involving $\delta p(a)$ and $\delta p(b)$ are omitted, the resulting error bounds will be derived solely from the correction terms, these representing the effect of truncation error on the computed $Y^{(k)}$ functions. On the other hand, if the correction terms are omitted, the bounds will depend on the

error vectors $\delta p(a)$ and $\delta p(b)$. This indicates how values may be calculated for the coefficients $N^{(k)}$ and $V^{(k)}$ defined in a relation of the form

$$||\delta y^{(k)}(a,b)|| \leq N^{(k)}(b)||\delta p(a)|| + N^{(k)}(a)||\delta p(b)|| + V^{(k)}(a,b) \qquad (6.29)$$

where $V^{(k)}(a,b) = V^{(k)}(b,a)$.

Thus the coefficient $V^{(k)}$ may be evaluated by omitting the terms involving $\delta p(a)$ and $\delta p(b)$ from the recurrence relation for $\delta z^{(k)}$ or $\delta j^{(k)}$ and computing a vector of error bounds $|\delta y_i^{(k)}|$ $(i = 1, 2, \ldots, u)$; then $V^{(k)} = \max_{(i)} \frac{1}{W_i} |\delta y_i^{(k)}|$. In practice, each of the correction terms (6.26) is limited to the first term in the expansion so that for instance the recurrence for $\delta z^{(k)}$ appears as follows:

$$|\delta z_{i+1}^{(k)}| \leq \tau|\delta z_i^{(k)}| + \left|\tau^{-2}\alpha_{i+3} - 3\tau^{-1}\alpha_{i+2} + 2\alpha_{i+1} + 2\tau\alpha_i - 3\tau^2\alpha_{i-1} + \tau^3\alpha_{i-2}\right|$$

where $\tau = e^{-kh}$ and $\alpha_i = \frac{11}{1440} h r_i P_i(a) P_i(b)$. After computing a set of bounds on $\{|\delta z_i^{(k)}|\}$, the required bounds on $\{|\delta y_i^{(k)}|\}$ are determined from a similar recurrence:

$$|\delta y_{i-1}^{(k)}| \leq \hat{\tau}|\delta y_i^{(k)}| + \hat{\tau}^{-1}|\hat{\alpha}_{i-2}| + 13|\hat{\alpha}_{i-1}| + 13\hat{\tau}|\hat{\alpha}_i| + \hat{\tau}^2|\hat{\alpha}_{i+1}|$$
$$+ \left|\hat{\tau}^{-2}\tilde{\alpha}_{i-3} - 3\hat{\tau}^{-1}\tilde{\alpha}_{i-2} + 2\tilde{\alpha}_{i-1} + 2\hat{\tau}\tilde{\alpha}_i - 3\hat{\tau}^2\tilde{\alpha}_{i+1} + \hat{\tau}^3\tilde{\alpha}_{i+2}\right|$$

where

$$\hat{\tau} = e^{-(k+1)h}$$
$$\hat{\alpha} = \frac{1}{24}(2k+1) h \delta z_i^{(k)}$$
$$\tilde{\alpha} = \frac{11}{1440}(2k+1) h Z^{(k)}.$$

Similar relations can be written for calculating the bounds when $J^{(k)}$ is computed as the auxiliary vector rather than $Z^{(k)}$.

To evaluate $N^{(k)}(b)$ or $N^{(k)}(a)$, the correction terms are omitted and one of the two vectors $\{\delta p_i(b)\}$, $\{\delta p_i(a)\}$ is set to zero, the other being replaced by $\{W_i\}$. The bounds on $\{|\delta z_i^{(k)}|\}$ are then computed from the following relations by recurring outwards from the value $\delta z_1^{(k)} = 0$:

$$|\delta z_i^{(k)}| \leq |\gamma_i| + 25\tau|\beta_{i-1}| + 12|\beta_i| + \tau^{-1}|\beta_{i+1}|$$

with

$$|\gamma_i| \leq \tau|\gamma_{i-1}| + 24\tau^2|\beta_{i-2}| \quad (i = 2, 3, \ldots)$$

and

$$|\gamma_1| \leq \tau^3|\beta_{-2}| + 12\tau^2|\beta_{-1}| + \tau|\beta_0|$$

where

$$|\beta_i| \leq \frac{1}{24} h r_i W_i |\Psi_i| \quad [\Psi_i \equiv P_i(b) \text{ or } P_i(a)].$$

[The indices $i < 1$ represent grid points beyond the inner limit chosen for integration of the radial equations.] Assuming $\delta y_i^{(k)} = \delta z_i^{(k)}$ for $i \geq u$, the bounds on $\{|\delta y_i^{(k)}|\}$ are computed by inward recurrence ($i = u-1, u-2, \ldots$) from the following system:

$$|\delta y_i^{(k)}| \leq |\hat{\gamma}_i| + 25\hat{\tau}|\hat{\beta}_{i+1}| + 12|\hat{\beta}_i| + \hat{\tau}^{-1}|\hat{\beta}_{i-1}|$$

with

$$|\hat{\gamma}_i| \leq \hat{\tau}|\hat{\gamma}_{i+1}| + 24\hat{\tau}^2|\hat{\beta}_{i+2}| \quad (i = u-1, u-2, \ldots)$$

and

$$|\hat{\gamma}_u| \leq \hat{\tau}^2|\delta z_{u+2}^{(k)}| + \hat{\tau}^3|\hat{\beta}_{u+3}| + 12\hat{\tau}^2|\hat{\beta}_{u+2}| + \hat{\tau}|\hat{\beta}_{u+1}|$$

where

$$|\hat{\beta}_i| \leq \frac{1}{24}(2k+1) h |\delta z_i^{(k)}|.$$

The bounds on $\{|\delta y_i^{(k)}|\}$ can be said to arise from errors of unit norm in $P(a)$ or $P(b)$, and the value of $N^{(k)}(b)$ or $N^{(k)}(a)$ is given by $\max_{(i)} \frac{1}{W_i} |\delta y_i^{(k)}|$.

Again a similar set of relations can be established for the case in which $Y^{(k)}$ is tabulated by the method involving the calculation of $J^{(k)}$ rather than $Z^{(k)}$.

The vectors of error bounds computed by these means, from which values may be calculated for the coefficients appearing in expressions of the form (6.29), are related to the coefficients N_y, V_y, N_x, V_x of

expressions (6.9) and (6.10) through the definitions of the potential and exchange terms (6.2), (6.3). Thus the following relation may be written from (6.2):

$$||\delta y^{(a)}|| \leq \sum_k |A_{aa}^{(k)}| \, ||\delta y^{(k)}(a,a)|| + \sum_{b\neq a} |A_{ab}^{(0)}| \, ||\delta y^{(0)}(b,b)||$$

and this gives rise to the following expressions for the coefficients in (6.9) in terms of those in (6.29):

$$\left.\begin{aligned} N_y(a,a) &= 2\sum_k |A_{aa}^{(k)}| \, N^{(k)}(a) \\ N_y(a,b) &= 2\,|A_{ab}^{(0)}| \, N^{(0)}(b) \qquad (b\neq a) \\ V_y(a) &= \sum_k |A_{aa}^{(k)}| \, V^{(k)}(a,a) + \sum_{b\neq a} |A_{ab}^{(0)}| \, V^{(0)}(b,b) \end{aligned}\right\} . \tag{6.30}$$

The determination of the coefficients N_x, V_x is facilitated by deriving the following relation (written to first order in errors) from the exchange term (6.3):

$$|\delta x_i(a)| \leq \frac{2}{r_i} \sum_{b\neq a} |P_i(b)| \sum_k |B_{ab}^{(k)} \, \delta y_i^{(k)}(a,b)|$$

$$+ \sum_{b\neq a} |\delta p_i(b)| \left\{ \frac{2}{r_i} \sum_k |B_{ab}^{(k)} \, Y_i^{(k)}(a,b)| + |\epsilon_{ab}| \right\}$$

where $|\delta p_i(b)|$ is bounded by $W_i \, ||\delta p(b)||$. Now the error analysis associated with the tabulation of the $y^{(k)}$ functions involves the derivation of bounds on $\{|\delta y_i^{(k)}|\}$ arising in turn from the truncation error and from errors of unit norm in P(a) and P(b). Denoting these bounds by $\{v_i^{(k)}(a,b)\}$, $\{n_i^{(k)}(b)\}$, and $\{n_i^{(k)}(a)\}$ respectively, the above relation gives rise to the following expressions for the coefficients in (6.10):

$$\left.\begin{aligned} N_x(a,a) &= \max_{(i)} \left\{ \frac{2}{r_i W_i} \sum_{b\neq a} |P_i(b)| \sum_k |B_{ab}^{(k)}| \, n_i^{(k)}(b) \right\} \\ N_x(a,b) &= \max_{(i)} \left\{ \frac{2}{r_i W_i} |P_i(b)| \sum_k |B_{ab}^{(k)}| \, n_i^{(k)}(a) \right. \\ &\qquad \left. + \frac{2}{r_i} \sum_k |B_{ab}^{(k)} \, Y_i^{(k)}(a,b)| + |\epsilon_{ab}| \right\} \qquad (b\neq a) \\ V_x(a) &= \max_{(i)} \left\{ \frac{2}{r_i W_i} \sum_{b\neq a} |P_i(b)| \sum_k |B_{ab}^{(k)}| \, v_i^{(k)}(a,b) \right\} \end{aligned}\right\} . \tag{6.31}$$

To minimize the computed values of the coefficients, the error vectors from which they are derived need to be calculated in such a way as to induce in them as close a resemblance as possible to the chosen weights-vector. It was stated in §6.2 that the weights-vector should be constructed so as to form an envelope over the error bounds from which the coefficients M_y and M_x are calculated, and it is now clear that the bounds from which N_y and N_x are calculated should also be enveloped by W. In particular, if W is of the typical form shown in Figure 6.1, these bounds must fall off in magnitude at each end of the range at a rate at least as fast as W itself.

It is not surprising to find that the bounds computed by the procedures outlined in the foregoing treatment do not satisfy this requirement: in fact the computed bounds are found to increase monotonically from $r = r_1$ to $r = r_u$ although the rate of increase falls to zero as the outer limit is approached. Thus the bounds do diminish in size at a satisfactory rate towards the inner end of the range but unfortunately retain their maximum values on moving towards the outer end. The bounds are virtually the same whichever method is chosen for computing the $y^{(k)}$ functions - whether by solving the system of equations (6.23) or (6.24) - and it is difficult to see how the computed bounds could be reduced to any significant extent near the outer limit of integration without permitting the assumption that the values of the $y^{(k)}$ functions are determined exactly at the outer limit. This could only be possible in the particular case $k = 0$, since only then can an exact value be specified for $\lim_{r \to \infty} y^{(k)}(a,b;r)$: namely

$$y^{(0)}(a,b;\infty) = z^{(0)}(a,b;\infty) = \int_0^\infty p(a;s)\, p(b;s)\, ds = \delta_{ab} .$$

This can be used as an extra boundary condition at $r = r_u$ if the contribution to the integral from the region $r > r_u$ is neglected. The complete list of assumed boundary conditions for the computed functions and their errors now stands as follows:

$z_1^{(0)}$ obtained from equation (6.28);

$z_u^{(0)} = Y_u^{(0)} = \delta_{ab}$; $J_u^{(0)} = 0$;

$\delta z_1^{(0)} = 0$;

$\delta z_u^{(0)} = \delta y_u^{(0)} = \delta j_u^{(0)} = 0$.

In the case $k=0$, a wider choice may be exercised in selecting the most appropriate method by which to integrate equations (6.23) or (6.24). This is because the step-by-step integration of the equation for $z^{(0)}$ is stable in either direction of recurrence; the same is true of the equation for $y^{(0)}$ as a function of $j^{(0)}$, and the corresponding error vectors may also be computed from the respective recurrence relations by proceeding in either direction. It would appear, then, that the computed error bounds could be reduced in size in the region near the outer limit of integration by taking advantage of the extra boundary conditions and integrating all the equations in the direction of decreasing r. This would involve tabulating each $Y^{(0)}$ function by performing an inward integration in either $Z^{(0)}$ or $J^{(0)}$ and then an inward integration in $Y^{(0)}$, and calculating the error vectors correspondingly. On implementing this procedure it is found that although the computed bounds are indeed sufficiently reduced at points near the outer limit they now increase monotonically from r_u to r_1 albeit at a diminishing rate which drops to zero as r approaches r_1.

It is evident that the bounds will not be optimized throughout the range of integration unless two methods are employed for calculating the errror vectors so that at each grid point the bound can be taken as the smaller of two computed values. This may be achieved in practice by implementing either of two possible schemes:

(i) An outward recurrence is performed for $|\delta z^{(0)}|$ starting from $\delta z_1^{(0)} = 0$, and an inward recurrence for $|\delta j^{(0)}|$ starting from $\delta j_u^{(0)} = 0$; then an inward recurrence is begun for $|\delta y^{(0)}|$ starting from $\delta y_u^{(0)} = 0$, using the formula in which the bounds are specified in terms of the $|\delta j_i^{(0)}|$ but

testing at each step whether the computed bound on $|\delta y_i^{(0)}|$ is less than would be given by the formula which specifies the bounds in terms of the $|\delta z_i^{(0)}|$; a transfer is made to this formula as soon as the condition is not satisfied and the same formula is then used for the remainder of the inward recurrence for $|\delta y^{(0)}|$. Since the $|\delta z_i^{(0)}|$ are monotonic decreasing as the index i decreases whereas the $|\delta j_i^{(0)}|$ are monotonic increasing, it can safely be assumed that once having changed over from one recurrence relation to the other the bounds computed for the rest of the $|\delta y_i^{(0)}|$ would not be reduced by making any further changes.

(ii) An outward recurrence is performed for $|\delta z^{(0)}|$ starting from $\delta z_1^{(0)} = 0$, and a second bound on $|\delta z^{(0)}|$ is calculated by performing an inward recurrence starting from $\delta z_u^{(0)} = 0$; the smaller of the two bounds is taken at each point, to form a set of values $\{|\delta z_i^{(0)}|\}$ from which to calculate $|\delta y^{(0)}|$ by performing an inward recurrence starting from $\delta y_u^{(0)} = 0$.

A comparison of the practical effects of the two schemes in a few test cases shows that scheme (i) supplies the optimum bounds when the error vectors computed are those from which the coefficients $V^{(0)}(a,b)$ are derived whereas scheme (ii) is preferable for calculating the vectors which give rise to values for the coefficients $N^{(0)}(b)$, $N^{(0)}(a)$. The reason for the choice of scheme (i) in evaluating $V^{(0)}(a,b)$ is that the inward recurrence for $|\delta y^{(0)}|$ includes a term representing the first-order approximation to the local truncation error [the first term in the expansion (6.26)] and this vanishes, as desired, at the outer end of the range when the recurrence used is the one in which the $|\delta y_i^{(0)}|$ are computed from the $|\delta j_i^{(0)}|$ whereas this is not always true of the corresponding term when the $|\delta y_i^{(0)}|$ are computed, as in scheme (ii), from the $|\delta z_i^{(0)}|$. However, these considerations do not apply when it comes to evaluating the coefficients $N^{(0)}(b)$, $N^{(0)}(a)$; here it is found that scheme (ii) yields the better bounds because in avoiding the switch from one relation to the other during the inward recurrence for $|\delta y^{(0)}|$ full advantage is taken of the cancellation of terms

stemming from the use of the same relation throughout.

These methods for optimizing the error bounds and minimizing the values of the coefficients apply in the special case $k=0$, and it is only in this case that the analysis has been subjected to any practical testing. It is, however, the most important case: this is because, in solving the Hartree-Fock equations for any atomic configuration, a substantial number of the $y^{(k)}$ functions to be computed will in fact be $y^{(0)}$ functions. To deal with the others (those for which $k>0$) it would seem necessary to resort to the error analysis developed earlier in this section, leading as it does to the calculation of error bounds which increase monotonically from the inner to the outer end of the range. It should be noted, however, that the expressions (6.31) for the coefficients N_x and V_x involve products of elements of error vectors with corresponding grid values of wave functions and it seems plausible to suppose that, in consequence of the behaviour of the wave functions themselves, the functions of which these products form the grid values would fall off in magnitude at sufficiently fast rates towards the outer end of the range. The expressions (6.30) for N_y and V_y do not involve such products, but any $y^{(k)}$ functions for $k>0$ occurring in the potential term (6.2) could always be removed and transferred to the exchange term (6.3) where each would then be multiplied by one of the wave functions.

These ideas represent no more than suppositions at the present stage, and the problem of optimizing the error bounds in the cases in which k is non-zero requires further investigation.

6.5 Configuration of one group

The simplest system to which the analysis of §§6.2 - 6.4 may be applied is the one representing the ground state of an atom or ion of two electrons. In the central-field approximation the electrons occupy the two available states of the 1s group; the spin wave function of the system is antisymmetric

while the spatial function is symmetric and comprised of a single product of two identical one-particle functions. This implies the absence of any 'exchange effects', and equation (6.1) for either of the radial functions p(r) reduces to the homogeneous Schrodinger equation; the potential term is given by

$$y(r) = Z - y^{(0)}(r)$$

where

$$y^{(0)}(r) = \int_0^r p^2(s)\,ds + r\int_r^\infty s^{-1}p^2(s)\,ds .$$

Equations (6.8) and (6.9) for the vector norms of δp and δy are reduced to the simple forms

$$||\delta p|| \leq M||\delta y|| + U$$
$$||\delta y|| \leq N||\delta p|| + V .$$

Equation (6.29) for the norm of $\delta y^{(0)}$ becomes

$$||\delta y^{(0)}|| \leq 2N^{(0)}||\delta p|| + V^{(0)}$$

and relations (6.30) give $N = 2N^{(0)}$, $V = V^{(0)}$. The matrix D and the vector $\underline{t}$ whose elements are defined by (6.11) consist in this case of the single elements MN and MV+U respectively, and the final bound (6.13) assumes the form

$$||\delta p|| \leq \frac{MV + U}{1 - MN}$$

so that, from (6.15),

$$|\delta p_i| \leq w_i \frac{MV + U}{1 - MN} \quad \text{for all } i .$$

The condition that D be a convergent matrix now gives rise to the condition $MN < 1$; the weights-vector must therefore be adjusted so as to minimize the product of the coefficients M and N.

For an atom of central charge Z, an initial estimate of the function p(r) for either electron is conveniently provided by determining analytically (see §4.2) the radial function for the 1s state of a one-electron atom with a central charge of, say, $Z - \frac{1}{2}$. Starting with such a function, and

demanding self-consistency to four decimal places in the value of the function at each grid point, the number of iterations required is found to depend on the value of Z in the manner indicated:

Z	2	3	4	5 - 10	>10
number of iterations	8	6	5	4	3

This dependence on Z of the rate of convergence of the self-consistent field procedure is indicative of the variation in extent to which the computed function differs from the initial estimate; when Z is small the screening of the nuclear charge by each electron is relatively high and the wave function differs markedly from that of a one-electron atom whereas an increase in Z gives rise to a screening effect of relatively less importance and a function more closely resembling the initial estimate.

Choosing for W the vector $\{|\delta p_i|\}$ computed under the conditions specified for calculating the coefficient U (that is, arising solely from truncation error in the analysis associated with the solution of the radial equation), the values obtained for the coefficients M and N and for the error bound on the wave function are given (to two significant figures) in Table 6.1 for a range of values of Z. The step size used ($h = 0.1$) is the same as in most of the cases for which results are tabulated in previous chapters (§§ 4.2, 5.3).

The results indicate that, as Z decreases, the values of M and N undergo approximately proportional increases, with the result that the convergence criterion is not satisfied for $Z < 4$. For $Z \geq 4$ the computed error bound has the same form as the weights-vector over the range of integration: it increases monotonically from the starting point (r_1) to the match-point (r_m) and then falls off rapidly towards the outer limit; the value attained at $r = r_m$ is listed in the last column of Table 6.1.

Table 6.1

h = 0.1

Z	M	N	MN	$\|\|\delta p\|\|$	$\max_{(i)} \|\delta p_i\| \times 10^4$
20	0.69	0.21	0.15	1.6	0.86
18	0.72	0.22	0.16	1.7	0.84
16	0.74	0.24	0.18	1.8	0.81
14	0.87	0.26	0.22	2.0	0.92
12	0.88	0.28	0.24	2.1	0.84
10	0.98	0.30	0.30	2.5	0.90
8	1.1	0.34	0.36	3.1	0.92
6	1.2	0.39	0.49	4.5	1.1
4	1.7	0.49	0.84	19	4.1
3	2.0	0.58	1.1	-	-
2	2.3	0.73	1.7	-	-

The purpose of calculating the error bound is to provide some indication of the number of figures that may reliably be quoted in the tabulated value of the function at any point in the range. The results demonstrate that, for all but the lowest values of Z, the wave function is obtained to an accuracy of at least four decimal places at points near $r = r_m$, the same relative accuracy being maintained at all points $r_i < r_m$ (the wave function being nodeless). In the tail region ($r > r_m$) the computed bound falls off at a rate which is not sufficiently fast to indicate whether the actual error remains at the same relative value: this is a consequence of the chosen weights-vector not decreasing sufficiently rapidly as the outer limit of integration is approached. Taken overall, however, the size of the bound suggests that a relative accuracy of the order of 1 part in 10^5 is the most that can be guaranteed without reducing the step size and that it would be futile tp perform further iterations of the self-consistent field

§6.5

procedure to achieve any closer agreement between successive iterates.

To gain some idea of the extent to which the bound represents an over-estimate of the actual error, the results of repeating the calculation using different step sizes are given in Table 6.2 for the case $Z=4$. The final column in the table gives the maximum discrepancy between the tabulated values of the wave function computed using $h=0.1$ or 0.05 and the corresponding values computed using $h=0.025$. Assuming that halving the step size improves the accuracy by at least a factor of 10, this discrepancy can be regarded as a measure of the actual error in the function computed at the larger interval size. On this basis the results indicate that the computed error bound over-estimates the actual error by a factor of order 100 and a closer inspection reveals that this is the case throughout the range of integration.

Table 6.2

$Z=4$

h	M	N	MN	$\|\|\delta p\|\|$	$\max_{(i)} \|\delta p_i\|$	$\max_{(i)} \|\delta p_i\|$ observed
0.1	1.7	0.49	0.84	19	$4.1_{10}-4$	$3_{10}-6$
0.05	1.6	0.49	0.77	14	$1.7_{10}-5$	$1_{10}-7$
0.025	1.6	0.47	0.75	12	$9.4_{10}-7$	-

The bound may only be improved by altering the weights-vector in some way so as to effect a reduction in the computed values of the coefficients M and $N^{(0)}$. It will be recalled that each of these coefficients is evaluated by computing under certain conditions an error vector $\{|\delta p_i|\}$ or $\{|\delta y_i^{(0)}|\}$ and taking the maximum over the set of vector elements after division by the corresponding elements of the weights-vector. Thus $M = \max_{(i)} M_i$ where $M_i = \frac{1}{W_i}|\delta p_i|$, and $N^{(0)} = \max_{(i)} N_i^{(0)}$ where $N_i^{(0)} = \frac{1}{W_i}|\delta y_i^{(0)}|$.

A coefficient is regarded as being 'optimized' when the corresponding vector has all its elements equal in value.

An examination of the vectors from which are derived the coefficients listed in the results in Table 6.1 reveals that the values of M are not far from the optimum; for example, for $Z = 4$ the elements M_i satisfy $1.5 \lesssim M_i \lesssim 1.7$ for all values of i except those corresponding to points in the tail region: here the values fall monotonically (that is, $M_i > M_{i+1}$) because W has been chosen as the vector $\{|\delta p_i|\}$ computed under the conditions for calculating U rather than M and, dependent as it is on the truncation error, this decreases at a slower rate in the tail region than the error bound from which M is calculated. This means that if an attempt were made to use the latter to provide a second estimate of W in the manner described in §6.2 the first estimate would need to be retained for those elements W_i which correspond to points in the tail region - otherwise W would fall off too rapidly in this region and lead to an unacceptably high value for the coefficient U. However, the second estimate may certainly be used for all the other elements of W: when this is applied in the case mentioned above ($Z = 4$, $h = 0.1$) the result is that all the elements M_i except those corresponding to points in the tail region have the approximate value 1.6, and, since N remains unchanged, the values obtained for MN, $||\delta p||$, and $\max_{(i)} |\delta p_i|$ are 0.77, 11, and 3.9 respectively. Compared with the previous results (Table 6.1) the improvement is only marginal, but the achievement of a reduction in the value of the product MN indicates that the use of the second estimate of W might lead to a value less than unity for MN in the case $Z = 3$ or even $Z = 2$. Using a step size $h = 0.05$, the results of using this weights-vector in the cases $Z = 4, 3, 2$ are tabulated in Table 6.3; convergence is now achieved in the case $Z = 3$ but not when $Z = 2$.

Table 6.3

h = 0.05

Z	M	N	MN	$\|\|\delta_p\|\|$	$\max_{(i)} \|\delta_{p_i}\| \times 10^5$
4	1.4	0.49	0.70	8.0	1.6
3	1.7	0.57	0.94	48	8.5
2	2.1	0.73	1.5	-	-

Considering again the results presented in Table 6.1, it is found that the vectors which give rise to the values of the coefficients $N^{(0)}$ $(= \frac{1}{2}N)$ are not as well optimized, in the previously defined sense, as those from which the values of M are derived. In the case Z = 4, for instance, the elements $N_i^{(0)}$ satisfy $0.19 \lesssim N_i^{(0)} < N_{i+1}^{(0)} \lesssim 0.25$ for all i up to a value somewhat less than m (where i = m corresponds to the maximum of the elements M_i), and subsequent values are then monotonic decreasing to zero: $0.25 \gtrsim N_i^{(0)} > N_{i+1}^{(0)} > 0$. The latter represents an unavoidable decrease since W must be modelled on the error bound which falls to zero at the slowest rate and this is always a bound on $|\delta p|$ rather than $|\delta y^{(0)}|$. On the other hand it is legitimate to consider whether any reduction in value of the coefficients could be achieved by associating W to some extent with the error vector $\{|\delta y_i^{(0)}|\}$. For example, W could be chosen so that $W_i = \max(|\delta p_i|, |\delta y_i^{(0)}|)$ for all i, where $\{|\delta p_i|\}$ and $\{|\delta y_i^{(0)}|\}$ are the error vectors arising only from consideration of the truncation error associated with the solution of the radial equation and the tabulation of the $Y^{(0)}$ function respectively, these being the vectors computed under the conditions specified for calculating the coefficients U and $V^{(0)}$. In the case Z = 4 the elements $N_i^{(0)}$ now satisfy $0.21 \lesssim N_i^{(0)} < N_{i+1}^{(0)} \lesssim 0.24$ up to the point at which they begin to decrease towards the outer limit, and from this it follows that N = 0.48 (to be compared with the value 0.49 in Table 6.1).

However, the value of M suffers an attendant change from 1.7 to 2.0 so that the product MN increases from 0.84 to 0.94. It would seem, then, that the optimum state will not be achieved unless W is allowed to reflect as fully as possible the form of $\{|\delta p_i|\}$ only, and even then the analysis appears to be of little use when Z is less than 5. For higher values of Z, however, the results do give a useful indication of the accuracy likely to be achieved for a given step size in computing the self-consistent field wave functions.

It is worth mentioning also that in the simple case under discussion the possibility exists of finding a bound on the error in the computed value of ϵ at the termination of the self-consistent field procedure, where ϵ is the energy parameter appearing in the radial equation for either of the one-electron states. In analogy with expression (6.8) for the bound on $||\delta p||$, which in this case becomes $||\delta p|| \leq M||\delta y|| + U$, the bound on $|\delta\epsilon|$ may be written

$$|\delta\epsilon| \leq |\delta\epsilon^{(1)}|\,||\delta y|| + |\delta\epsilon^{(0)}|$$

where $|\delta\epsilon^{(1)}|$ and $|\delta\epsilon^{(0)}|$ are the values computed for $|\delta\epsilon|$ in performing the analysis leading to the calculation of the coefficients M and U respectively. A bound on $||\delta y||$ may be determined from the relation

$$||\delta y|| \leq \frac{NU + V}{1 - MN}$$

whose derivation involves the same assumptions as made in establishing the corresponding expression for $||\delta p||$.

For the same weights-vector, step size, and values of Z as were used in obtaining the results of Table 6.1, the computed values of the energy parameter and its error bound are tabulated in Table 6.4. According to Koopmans' theorem [see Slater (1960)] the parameter (which in atomic units is given by $\lambda = \frac{1}{2}\epsilon$) represents the single-ionization energy of the atom or ion on the assumption that the removal of one of the electrons does not affect the other orbital.

Table 6.4

h = 0.1

Z	ϵ	$\lvert\delta\epsilon\rvert$
20	375.3335	0.0028
18	301.8334	0.0024
16	236.3334	0.0020
14	178.8334	0.0016
12	129.3334	0.0013
10	87.8335	0.0010
8	54.3336	0.0008
6	28.8338	0.0006
4	11.3342	0.0010

In the particular case Z = 4, results are presented in Table 6.5 demonstrating the effect of halving the step size [compare Table 6.2].

Table 6.5

Z = 4

h	ϵ	$\lvert\delta\epsilon\rvert$
0.1	11.334228	0.00099
0.05	11.334231	0.000043
0.025	11.334231	0.0000025

6.6 Configuration of two groups

The next system to be considered is the one representing the ground state of an atom or ion of four electrons, in which the electrons occupy the available states of the 1s and 2s groups. Exchange effects are now present in the system but there is no necessity to impose orthogonality constraints leading to the presence of off-diagonal parameters since the

configuration consists of complete groups only - to this extent the treatment is simpler than it would be for the three electron system. In the central-field approximation, the radial functions p(1s;r) and p(2s;r) satisfy equations whose potential and exchange terms are given by the following expressions:

$$y(a;r) = Z - y^{(0)}(a,a;r) - 2\,y^{(0)}(b,b;r)$$

$$x(a;r) = -\frac{2}{r}\,y^{(0)}(a,b;r)\,p(b;r)$$

where a and b represent the groups 1s and 2s, or vice versa. Thus, as in the case of the two-electron atom, the only $y^{(k)}$ functions which arise are those for which $k=0$.

Equations (6.8), (6.9), and (6.10) for the bounds on the norms of δp, δy, and δx now assume the following forms:

$$||\delta p(a)|| \le M_y(a)||\delta y(a)|| + M_x(a)||\delta x(a)|| + U(a)$$

$$||\delta y(a)|| \le N_y(a,a)||\delta p(a)|| + N_y(a,b)||\delta p(b)|| + V_y(a)$$

$$||\delta x(a)|| \le N_x(a,a)||\delta p(a)|| + N_x(a,b)||\delta p(b)|| + V_x(a)$$

and expressions (6.30) and (6.31) for the coefficients in the last two of these relations become:

$$N_y(a,a) = 2\,N^{(0)}(a)$$

$$N_y(a,b) = 4\,N^{(0)}(b)$$

$$V_y(a) = V^{(0)}(a,a) + 2\,V^{(0)}(b,b)$$

$$N_x(a,a) = \max_{(i)}\left\{\frac{2}{r_i W_i}\,|P_i(b)|\,n_i^{(0)}(b)\right\}$$

$$N_x(a,b) = \max_{(i)}\left\{\frac{2}{r_i W_i}\,|P_i(b)|\,n_i^{(0)}(a) + \frac{2}{r_i}\,|Y_i^{(0)}(a,b)|\right\}$$

$$V_x(a) = \max_{(i)}\left\{\frac{2}{r_i W_i}\,|P_i(b)|\,v_i^{(0)}(a,b)\right\}.$$

The matrix D whose elements are given in terms of these coefficients by relation (6.11) is in this case a 2×2 matrix, and the weights-vector must be chosen so as to minimize its spectral radius.

As in the case of the single-group configuration, the self-consistent field procedure may be started by taking as an initial estimate of each wave function the analytically-known function corresponding to the appropriate state of a one-electron atom of suitably screened nuclear charge: convenient values for the central charge are found to be $Z-1$ for the 1s state and $Z-2.4$ for the 2s state. Starting the procedure with these Coulomb functions and imposing the condition for self-consistency that each new estimate should agree with the previous estimate to four decimal places at each grid point, the number of iterations required for this condition to be satisfied is found to vary with Z to the extent indicated:

Z	4	10	20	30
number of iterations	18	6	5	4

On the basis of the analysis of §6.2, it would appear that a suitable weights-vector could be provided in the first instance by choosing

$$W_i = \max(|\delta p_i(1s)|, |\delta p_i(2s)|)$$

where the two vectors of error bounds $|\delta p(1s)|$ and $|\delta p(2s)|$ are computed under the conditions specified for calculating the coefficients U(1s) and U(2s) (and arise, therefore, from the effects of truncation error alone). If W is chosen in this way after scaling one of the vectors so that $\max_{(i)} |\delta p_i(1s)| = \max_{(i)} |\delta p_i(2s)|$, its form will be of the general type depicted in Figure 6.1, with two maxima of equal height occurring at the match-points of the 1s and 2s functions. If the matrix D is convergent, this form will also be reflected in the final error bound obtained for each wave function.

Using a step size $h=0.1$, the following values (written to two significant figures) are obtained for the coefficients M_y, N_y, M_x, N_x in the case $Z=30$:

$M_y(1s) = 0.56$ $\quad$ $M_y(2s) = 0.90$

$N_y(1s,1s) = 0.19$ $\quad$ $N_y(1s,2s) = 0.72$

$N_y(2s,1s) = 0.38$ $\quad$ $N_y(2s,2s) = 0.36$

$M_x(1s) = 0.0038$ $\quad$ $M_x(2s) = 0.042$

$N_x(1s,1s) = 9.1$ $\quad$ $N_x(1s,2s) = 27$

$N_x(2s,1s) = 39$ $\quad$ $N_x(2s,2s) = 47$.

The resulting D matrix may be written

$$D = \begin{bmatrix} 0.11 + 0.035 & 0.40 + 0.11 \\ 0.34 + 1.6 & 0.32 + 1.9 \end{bmatrix}$$

where each of the two terms $M_y(a)N_y(a,b)$ and $M_x(a)N_x(a,b)$ whose sum according to (6.11) forms the element D_{ab} has been written separately. The spectral radius of this matrix is 2.7 - thus convergence has not been achieved even with a value of Z as high as 30. The spectral radius assumes even larger values when Z is reduced, increasing to 10 for Z = 10 and to 50 for Z = 4.

An examination of the vectors from which the coefficients listed above are derived reveals that the degree of optimization achieved (in the sense defined in §6.5) differs markedly from one coefficient to another. The elements of the vector whose maximum element is $M_y(1s)$ increase monotonically in value from 0.52 for those elements corresponding to points near the inner limit of integration (r_1) to 0.56 for those near the match-point of the 1s function (r_{m_1}); thereafter the values decrease monotonically towards zero as the outer limit is approached. Similarly the elements of the vector from which $M_x(1s)$ is derived vary between 0.0032 and 0.0038 before decreasing in value towards the outer limit. The vectors from which the values of $M_y(2s)$ and $M_x(2s)$ are derived have their maximum elements (0.90 and 0.042 respectively) at or near the match-point of the 2s function (r_{m_2}) (where $r_{m_2} > r_{m_1}$) and disregarding the elements corresponding to points in the tail region the vectors have their minimum elements (0.07 and 0.002 respectively) at r_{m_1}. This behaviour is of course a consequence of the

chosen form of the weights-vector with its two maxima at r_{m_1} and r_{m_2}, and the results will only be improved by changing W in some way with a view to reducing the values of some or all of the coefficients and rendering the matrix D convergent.

It should be noted at this stage that convergence could not be achieved simply by reducing the step size since the values obtained for the coefficients are affected hardly at all by so doing; for example the spectral radius of D is altered only from 2.7 to 2.4 when the calculation is performed with $h = 0.05$ rather than $h = 0.1$ in the case $Z = 30$. This lack of dependence on h can be observed also in the results presented in Table 6.2 for the single-group configuration.

Examining again the vectors from which the coefficients M_y, M_x are derived in the case $h = 0.1$, $Z = 30$, it would seem that those depending on error bounds for the 1s function - namely $M_y(1s), M_x(1s)$ - are optimized to a somewhat greater extent than those depending on bounds for 2s function and that perhaps W should be adjusted so as to reflect rather more of the latter and less of the former in order to effect an improvement in the results. The weights-vector has been chosen so that $W_i = \max(|\delta p_i(1s)|, |\delta p_i(2s)|)$ where the vectors are those arising from truncation error alone and W is formed after scaling one of the vectors so that $\max_{(i)} |\delta p_i(1s)| = \max_{(i)} |\delta p_i(2s)|$. This leads in practice to a vector whose form is that of $|\delta p(1s)|$ over the section of the range extending from r_1 to a point midway between r_{m_1} and r_{m_2} and $|\delta p(2s)|$ from this point to the outer limit. The balance can be altered in favour of the error bound for the 2s function by rescaling so that $\max_{(i)} |\delta p_i(1s)| = f \times \max_{(i)} |\delta p_i(2s)|$ for a value of f between 0 and 1 before choosing each element W_i as the larger of the two elements $|\delta p_i(1s)|$ and $|\delta p_i(2s)|$. When $f = \frac{1}{2}$, for instance, the weights-vector has the same form as $|\delta p(2s)|$ at all points except those within a short distance of r_{m_1}, but the effect on the results of altering

W in this way appears to be negligible: the values of some of the coefficients are marginally reduced but others are increased and the spectral radius of D remains virtually unchanged (for example, its value is 2.5 when $f = \frac{1}{2}$).

Again, the results are only slightly affected by replacing W (except for those of its elements corresponding to points near the outer limit) by a second estimate derived, in the same manner as described in previous sections, by basing it on error bounds computed under the conditions specified for calculating the coefficients M_y, M_x rather than U. Certainly it appears impossible by this means to promote substantial and simultaneous reductions in the values of all the coefficients.

The discussion so far has been concentrated on the problem of minimizing the computed values of the coefficients M_y, M_x, but of course it is equally important to ensure that the values of N_y, N_x are as small as possible. The coefficients N_y are multiples of $N^{(0)}(1s)$ and $N^{(0)}(2s)$ and are derived from error bounds which maximize in the region of the match-points r_{m_1} and r_{m_2} and decrease monotonically on moving away from this region towards either limit of integration. It could be said, then, that these bounds are qualitatively similar in behaviour to the bounds on $|\delta p(1s)|$ and $|\delta p(2s)|$ which have already been used as the preferred basis upon which to form the weights-vector. Thus it is unlikely that the results could be improved by basing W more directly on the bounds from which the coefficients N_y are derived, especially as any deviation of W from the forms of the bounds on $|\delta p(1s)|$ and $|\delta p(2s)|$ inevitably leads to increases in the values of the coefficients M_y, M_x in the same way as the coefficient M was observed to be affected in the case of the single-group configuration.

Unfortunately the same is not true of the bounds on which the coefficients N_x depend for their values. This is because each of the elements of the vectors forming the bounds includes $P(1s)/r$ or $P(2s)/r$ as an additional factor which results in the bounds maximizing at the inner limit

rather than at some point in the vicinity of the match-points. This explains the relatively large values obtained for the coefficients N_x in the results previously tabulated. The values will remain large unless W itself is allowed to possess the same general form as these bounds. In that case, however, the effect of any reductions would undoubtedly be nullified by the accompanying increases in the values of the other coefficients. The alternative policy would be to carry out a revue of the analysis leading to the derivation of the vectors $\{n_i^{(0)}(1s)\}$ and $\{n_i^{(0)}(2s)\}$ with the aim of reducing the values obtained for the elements of these vectors and particularly those which correspond to points near the inner limit of integration. Since $p(a;r)/r$ varies as r^{1a} for small values of r, the problem would be expected to be less acute when the functions appearing in expressions (6.31) for the coefficients N_x are other than s-functions, but this consideration does nothing to alleviate the difficulty in the case under discussion. It is difficult to see how the error analysis associated with the tabulation of the $y^{(k)}$ functions could be 'refined' further than has already been achieved, bearing in mind that the analysis presented in §6.4 represents the final outcome of many such attempts. Moreover, the analysis will not be of any significant use if it cannot easily be extended to apply in the case of more complicated atomic structures, where almost certainly the conditions for its application will be even more stringent.

As the number of occupied groups increases, so too does the number of different radial wave functions, and the chances of the analysis in its present form yielding a convergent matrix D would seem to become that much more remote. Beyond the minor improvements that might possibly be made by modifying particular sections of the analysis and by choosing the weights-vector more appropriately, any further improvement would appear hard to achieve without altering the whole basis of the method. An obvious limitation of the method as it now stands is that each error bound must be expressed in terms of the single vector norm defined in §6.2, and the

restriction which this represents becomes more untenable as the complexity of the atomic structure increases. The aims set out in Chapter 1 for bounding the errors as realistically as possible will never fully be realised unless a method can be developed in which a different norm may be defined for each group in the configuration. This might be achieved for example by defining each norm as in (6.14) but making use of a different weights-vector for each group. However, no way has been found of developing the analysis along these lines, and the methods that have been presented in this chapter constitute the only practicable scheme yet to have been devised.

Nevertheless, as already observed, even this scheme fails to yield the required error bounds when the configuration involves more than one occupied group. In the case of the four-electron system, some idea can be gained of the magnitudes of the error bounds likely to be obtained in the event of the matrix D converging by omitting the exchange terms from the Hartree-Fock equations, for then each element D_{ab} of the matrix consists only of the single term $M_y(a)N_y(a,b)$ and the matrix is more likely to be convergent. By neglecting the exchange effects the mathematical system is removed one further step from reality to correspond to Hartree's original formulation of the problem, and each of the radial equations is now homogeneous. Using the same form of weights-vector as led to the results tabulated earlier in this section and the same step size $h = 0.1$, the values obtained for the error bounds are presented (to two significant figures) in Table 6.6 for several values of Z. The bounds assume the same form (that of the weights-vector) for the 1s and 2s functions, and both attain their maximum values at each of the two match-points; these values are listed in the last two columns of Table 6.6.

Table 6.6

h = 0.1

Z	spectral radius of D	$\|\|\delta p(1s)\|\|$	$\|\|\delta p(2s)\|\|$	$\max_{(i)} \|\delta p_i(1s)\| \times 10^3$	$\max_{(i)} \|\delta p_i(2s)\| \times 10^3$
10	2.0	-	-	-	-
15	1.2	-	-	-	-
20	0.88	13	16	3.2	3.8
25	0.74	5.2	6.7	1.5	1.9
30	0.59	2.8	3.9	0.89	1.2

The results demonstrate that even with the exchange effects omitted the method does not converge until Z is increased to a value between 15 and 20, and even then the bounds are much more pessimistic than those obtained in the case of the single-group configuration considered in §6.5. Evidently there will be no further improvement until a means is found of lifting the restriction under which all the error bounds are constrained to have the same form.

In seeking to extend the analysis which has been developed in this and previous chapters, the aim must be to achieve a more effective application of these techniques to the problem of calculating the self-consistent field in cases of more complicated atomic structure. A similar extension of the analysis of Chapter 5 to the problem of solving the relativistic form of the Hartree-Fock equations would represent the ultimate achievement.

REFERENCES

Bethe, H. A. and Salpeter, E. E. (1957) Quantum mechanics of one- and two-electron systems. Hand. d. Phys. 35, 88 - 436.

Blatt, J. M. (1967) Practical points concerning the solution of the Schrodinger equation. J. Comp. Phys. 1, 382 - 396.

Burke, V. M. and Grant, I. P. (1967) The effect of relativity on atomic wave functions. Proc. Phys. Soc. 90, 297 - 314.

Cooley, J. W. (1961) An improved eigenvalue corrector formula for solving the Schrodinger equation for central fields. Math. Comp. 15, 363 - 374.

Desclaux, J. P., Mayers, D. F., and O'Brien, F. (1971) Relativistic atomic wave functions. J. Phys. B. (In press)

Dirac, P. A. M. (1958) The principles of quantum mechanics. Oxford Univ. Press; 4th ed.

Feingold, D. G. and Varga, R. S. (1962) Block diagonally dominant matrices and generalizations of the Gerschgorin circle theorem. Pacific J. Math. 12, 1241 - 1250.

Fox, L. (1957) The numerical solution of two-point boundary problems in ordinary differential equations. Oxford Univ. Press.

------- (Ed.) (1962) Numerical solution of ordinary and partial differential equations. Pergamon Press.

Fox, L. and Mayers, D. F. (1968) Computing methods for scientists and engineers. Oxford Univ. Press.

Froese, C. (1962) On solving $y'' = fy + g$ with a boundary condition at infinity. Math. Comp. 16, 492 - 494.

------- (1963) Numerical solution of the Hartree-Fock equations. Can. J. Phys. 41, 1895 - 1910.

------- (1965) Hartree-Fock results for some excited states of O^{++}, O^{+}, and Ne^{+}. Phys. Rev. 137A, 1644 - 1648.

------- (1966a) Hartree-Fock parameters for the atoms helium to radon. J. Chem. Phys. 45, 1417 - 1420.

------- (1966b) Hartree-Fock procedure for some nsn's 1S configurations. Phys. Rev. 150, 1 - 6.

------- (1967a) The orthogonality assumption in the Hartree-Fock approximation. Can. J. Phys. 45, 7 - 12.

------- (1967b) Some Hartree-Fock results for two- and four- electron atomic systems. J. Chem. Phys. 47, 4010 - 4015.

Froese Fischer, C. (1968a) Numerical Hartree-Fock results for the atoms helium to radon. Can. J. Phys. 46, 2336 - 2337.

------- (1968b) The orthogonality assumption in the Hartree-Fock approximation for $2p^n3p$ configurations. Can. J. Phys. 46, 2435 - 2440.

Grant, I. P. (1961) Relativistic self-consistent fields. Proc. Roy. Soc. A 262, 555 - 576.

------- (1965) Relativistic self-consistent fields. Proc. Phys. Soc. 86, 523 - 527.

Hartree, D. R. (1957) The calculation of atomic structures. John Wiley, N.Y.

Henrici, P. (1962) Discrete variable methods in ordinary differential equations. John Wiley, N.Y.

Isaacson, E. and Keller, H. B. (1966) Analysis of numerical methods. John Wiley, N.Y.

Mayers, D. F. (1957) Relativistic self-consistent field calculation for mercury. Proc. Roy. Soc. A 241, 93 - 109.

------- (1964) The deferred approach to the limit in ordinary differential equations. Computer J. 7, 54 - 57.

Mayers, D. F. and O'Brien, F. (1968) The calculation of atomic wave functions. J. Phys. B 1, 145 - 150.

N.P.L. (1961) Notes on applied science, 16 - Modern computing methods. HMSO, London; 2nd ed.

Osborne, M. R. (1960) h^2 - extrapolation in eigenvalue problems.
Quart. J. Mech. 13, 156 - 168.

Pereyra, V. (1966) On improving an approximate solution of a functional equation by deferred corrections. Num. Math. 8, 376 - 391.

------- (1967) Accelerating the convergence of discretization algorithms.
SIAM J. Num. Anal. 4, 508 - 533.

------- (1968) Iterated deferred corrections for nonlinear boundary value problems. Num. Math. 11, 111 - 125.

Pratt, G. W. (1956) Unrestricted Hartree-Fock method.
Phys. Rev. 102, 1303 - 1307.

Slater, J. C. (1960) Quantum theory of atomic structure, Vols I, II.
McGraw - Hill, N.Y.

Sloan, I. H. (1968a) Errors in the Numerov and Runga - Kutta methods.
J. Comp. Phys. 2, 414 - 416.

------- (1968b) Method for the numerical solution of linear second-order differential equations. J. Comp. Phys. 3, 40 - 45.

Stetter, H. J. (1965) Asymptotic expansions for the error of discretization algorithms for non-linear functional equations.
Numer. Math. 7, 18 - 31.

Swirles, B. (1935) The relativistic self-consistent field.
Proc. Roy. Soc. A 152, 625 - 649.

Wilkinson, J, H. (1965) The algebraic eigenvalue problem.
Clarendon Press, Oxford.