JOURNAL OF COMPUTATIONAL PHYSICS 3, 17–28 (1968)

Round-off Errors in Variational Calculations

L. M. DELVES

School of Mathematical and Physical Sciences, University of Sussex, Falmer, Brighton, Sussex, England

Received March 15, 1968

ABSTRACT

Rigorous bounds are derived for the effect of round-off errors in variational calculations for eigenvalues of linear operators. These bounds are simple to compute. They are used to derive an alternative variation principle which minimizes the effect of round-off errors. A numerical example of the use of the techniques is given.

I. INTRODUCTION

$\mathscr{H}$, $\mathscr{N}$ are Hermitian operators in and ψ an element of a Hilbert space R, and $\mathscr{N}$ is positive-definite.

We consider in this paper the variational calculation of the eigenvalues λ of $\mathscr{H}$ with respect to $\mathscr{N}$:

$$[\mathscr{H} - \lambda\mathscr{N}]\psi = 0 \tag{1}$$

The Rayleigh–Ritz variational principle for this eigenvalue problem is

$$\lambda \approx \lambda_V \underset{x}{\overset{=}{\text{stat}}} (x, \mathscr{H}x)/(x\mathscr{N}x) \tag{2}$$

For a trial function x containing only linear parameters a,

$$x = \sum_{i=1}^{M} a_i\phi_i\,, \tag{3}$$

it is well known that (2) reduces to the $M \times M$ eigenvalue problem

$$[H - \lambda_u N]\mathbf{a} = 0, \tag{4}$$

where H, N are $M \times M$ Hermitian matrices with elements

$$(H)_{ij} = (\phi_i, \mathscr{H}\phi_j), \qquad (5)$$

$$(N)_{ij} = (\phi_i, \mathscr{N}\phi_j),$$

and $\mathbf{a}$ is the M-vector of amplitudes a_i. Further, it is well known that if $\mathscr{H}$ has a smallest eigenvalue λ_0, then the eigenvalues of (4) are upper bounds to the eigenvalues of $\mathscr{H}$,

$$\lambda_i \leqslant \lambda_{\mu i}, \qquad (6)$$

where (6) supposes both sets of eigenvalues to be in algebraically increasing order.

It is also a familiar phenomenon that this Rayleigh–Ritz procedure, and similar variational procedures for other linear operator equations such as the Kohn and related principles for scattering states, are subject in practice to apparently severe round-off errors. The result (4) and the bound (6) both assume that the arithmetic involved is carried out exactly; in particular, that the inner products (5) are known exactly. In most practical calculations this is not so; the matrix elements of H and N are calculated to finite precision, which is limited either by the word length of the computer, or by the numerical integration techniques available. Moreover, for large M, the observed errors in λ_μ may be much greater than the individual errors in the inner products, and the bound (6) may be violated. This loss of accuracy is disastrous if it is not recognized; when it is recognized, it is either countered by double or multiple precision working, or by increasing the accuracy of the numerical integrations if such are involved. Both of these palliatives, and especially the latter, are time-consuming, and it is often the round-off errors, rather than any upper limit in the number of terms M that can be handled, which determines the final accuracy of a calculation.

In this paper we analyze these round-off errors in detail. We first derive a rigorous bound on the magnitude of the error in λ_μ due to the finite accuracy of H and N. This bound involves only readily available quantities, and is easily computed at the same time as λ_μ; its existence gives a simple method of recognizing and quantizing possible error buildup. The error bound is then combined with the inequality (6) to form a rigorous bound on the eigenvalues of $\mathscr{H}$, valid even in the presence of roundoff errors.

In Section III we look further at these bounds. In the presence of round-off errors, the inequality (6) is largely academic, and the modified bounds of Section II are the only relevant ones. But the procedure embodied in (4) is designed to minimize the bound (6). We therefore discuss in this section a procedure designed to minimise the modified bound, and hence to get the best possible final result from a calculation of H, N of given accuracy. The resulting algorithm, and the simpler bounds of Section II, are illustrated by an example in Section IV. Finally, an extension of the method to scattering states given in Section V.

II. Bounds on the effect of truncation errors in H and N

The usual Rayleigh–Ritz procedure leads to the calculation of the eigenvalues λ_μ of (4),

$$[H - \lambda_\mu N]\mathbf{a} = 0,$$

where H, N are the Hamiltonian and normalization matrices (5). In fact however we calculate the eigenvalues λ_c of the approximate equation

$$[H + h - \lambda_c(N + n)]\mathbf{b} = 0 \tag{7}$$

where h, n are the error matrices of the computed H, N. In any calculation the final calculated eigenvalue has two sources of error. First, equation (7) is not in practice solved exactly; and second, h and n are not null, so that $\lambda_c \neq \lambda_\mu$. The first error may always be made as small as is wished by *solving* (7) to multiple precision. The time taken to do this is usually negligible compared with the time required to improve the accuracy of the inner products (5); and hence, the second source of error usually dominates. If we ignore the first source, we can compare the exact eigenvalues λ_μ, λ_c as follows. We write (5) and (7) in the form

$$\begin{aligned}[][\bar{H} - \lambda_\mu I]\mathbf{a}^1 &= 0,\\ [\bar{H} + C - \lambda_c I]\mathbf{b}^1 &= 0,\end{aligned} \tag{8}$$

where

$$\begin{aligned}\bar{H} &= N^{-1/2}HN^{-1/2},\\ C &= N^{-1/2}(h - \lambda_c n)N^{-1/2},\\ \mathbf{a}^1 &= N^{1/2}\mathbf{a} \quad \mathbf{b}^1 = N^{1/2}\mathbf{b}.\end{aligned} \tag{8a}$$

These equations express λ_μ, λ_c as the eigenvalues of the Hermitian matrices $\bar{H}$, $\bar{H} + C$, respectively. We then have an immediate bound for the change $\lambda_{ci} - \lambda_{\mu i}$ in the i—the eigenvalue of $\bar{H}$ induced by the perturbing matrix C (Ref. [1])

$$|\lambda_{ci} - \lambda_{\mu i}| \leqslant \| C \| \leqslant \| N^{-1/2} \|^2 \{\| h \| + |\lambda_{ci}| \, \| n \|\}. \tag{9}$$

In Eq. (9), $\| A \|$ denotes a norm of the matrix A (Ref. [2]). For convenience, the norm used in the numerical work of this paper is defined in the appendix. Equation (9) yields in principle a bound on the round-off errors induced in the eigenvalues λ_μ. We consider first a special case.

A. *Orthogonal Expansion of the Trial Function*

If the expanding functions ϕ_i in (3) are orthonormal,[1] the normalisation matrix reduces to the unit matrix, and we have

$$\begin{gathered} N = I, \qquad n = 0, \\ |\lambda_c - \lambda_\mu| \leqslant \| h \|. \end{gathered} \tag{10}$$

This is a very satisfactory bound. First, it is simple to compute, since bounds on $\| h \|$ follow trivially if we know either the relative or the absolute accuracy of the inner products (5). Second, it is as small as one might reasonably hope to achieve. If the absolute error in any element of H is less than ϵ we have for an $M \times M$ matrix

$$| h_{ij} | < \epsilon; \qquad \| h \|_\infty < M\epsilon \tag{11}$$

and we can hardly hope to do better than this.

B. *The general case*

If the expansion used is not orthogonal, the situation is less satisfactory. The bound (9) now contains the (unknown) matrix N. We first remove this dependence by writing

$$\begin{gathered} P = N + n, \\ [H - \lambda_\mu(P - n)]a = 0, \\ C^1 = P^{-1/2}(h - \lambda_\mu n)P^{-1/2}, \end{gathered} \tag{12}$$

where, as before, we derive the bound

$$| \lambda_c - \lambda_\mu | \leqslant \| C^1 \| \leqslant \| P^{-1/2} \|^2 \{ \| h \| + | \lambda_\mu | \, \| n \| \}. \tag{13}$$

If we know the sign of λ_μ we may rewrite (13); for instance, if $\lambda_\mu < 0$, we have the bound

$$\lambda_\mu \leqslant \frac{\lambda_c + \| P^{-1/2} \|^2 \| h \|}{1 + \| P^{-1/2} \|^2 \| n \|} \tag{13a}$$

Similar bounds can also be obtained in terms of the triangular decomposition $P = L^T L$ of P rather than $P^{-1/2}$. In principle these bounds can be computed. In practice, it is extremely inconvenient to form the matrix $P^{-1/2}$ (or the triangular

[1] We may always carry out a diagonal transformation to make the diagonal elements of N unity. The following discussion is therefore valid even when the normalization of the ϕ_i is not known exactly.

matrix L) merely to obtain this bound, and we derive below a much more convenient bound. We merely note here that the *magnitude* of this bound may be very much larger than that given by (10) for the orthogonal case, since we may well have $\| N^{-1/2} \| \gg 1$. Thus Eq. (13) makes specific the practical observation that truncation errors are minimized by using an orthogonal basis.

An alternative bound on the truncation errors. We now look at an alternative approach which leads to bounds on the eigenvalues λ of $\mathscr{H}$, but which are easier to compute than (13). We note that a valid variational estimate λ_v of λ is given for any M-vector $\mathbf{c}$ by

$$\lambda_V(c) = c^+Hc/c^+Nc. \tag{14}$$

The eigenvalue equation (4) is in fact relevant only as a way of choosing an optimum vector $\mathbf{c} = \mathbf{a}$. We therefore accept as the best available choice of $\mathbf{c}$ the vector $\mathbf{b}$ derived from the computed form (7) of (5), and define λ_v by

$$\lambda_v = b^+Hb/b^+Nb. \tag{15}$$

Then λ_v is a variational estimate of λ, and in fact represents an upper bound to the lowest eigenvalue of $\mathscr{H}$. We derive a *simple* bound for the difference $\lambda_v - \lambda_c$. We write

$$\lambda_c = b^+(H + h)b/b^+(N + n)b \tag{16}$$

and obtain

$$\lambda_c - \lambda_v = b^+(h - \lambda_v n)b/b^+Pb, \tag{17}$$

where, as before, P is the computed normalization matrix

$$P = N + n$$

We solve (17) for λ_v to obtain

$$\lambda_c - \lambda_v = (b^+hb - \lambda_c b^+nb)/(b^+Pb - b^+nb) \tag{18}$$

Equation (18) forms the basis of our bound. In general, b will be normalized so that

$$b^+Pb = 1 \tag{19}$$

and with this condition we have

$$| \lambda_c - \lambda_v | \leqslant \| b \|^2 \frac{\| h \| + | \lambda_c | \| n \|}{1 - \| b \|^2 \| n \|} \quad \text{provided that} \quad \| b \|^2 \| n \| < 1 \tag{20}$$

This bound on the round-off errors in λ_c is very simple to compute, and in general is much less pessimistic than (13); for instance, (20) may be used to derive a rigorous bound on the eigenvalues λ of $\mathscr{H}$ even if the errors of the computation are so large that the computed normalization matrix $P = N + n$ is not positive definite, so that $P^{-1/2}$ does not exist.

The bound can be improved somewhat if some information about λ_v is available. For example, we have, from (17) with normalization (19),

$$| \lambda_c - \lambda_v | \leqslant \| b \|^2 \{ \| h \| + | \lambda_v | \, \| n \| \}.$$

If we can be sure that λ_v is negative, we may write $| \lambda_v | = -\lambda_v$ and obtain

$$\lambda_v \leqslant (\lambda_c + \| b \|^2 \| h \|)/(1 + \| b \|^2 \| n \|) \text{ provided } \lambda_v < 0 \tag{21}$$

This result does *not* assume $\| b \|^2 \| n \| < 1$. We compare this with the weaker result obtained from (20) if no such assumption is made,

$$\lambda_v \leqslant \frac{\lambda_c + \| b \|^2 \{ \| h \| + (| \lambda_c | - \lambda_c) \| n \| \}}{1 - \| b \|^2 \| n \|} \tag{22}$$

III. Optimized bounds including roundoff errors

The bounds (21) and (22) on λ_v represent also bounds on the eigenvalues λ of $\mathscr{H}$, since we have for any vector **b**

$$\lambda \leqslant \lambda_v .$$

The Rayleigh–Ritz procedure as defined by (7) leads to a choice of the vector **b** which minimizes, not λ_v but the computed λ_c. If round-off errors are significant, the bounds (22) (say) for λ_v will differ significantly from λ_c, and in this case it is clear that the choice of **b** given by Eq. (7) may not be optimal. Rather, we should try to choose a vector **b** which minimizes the bounds (21) or (22). These bounds are not simple in structure in their detailed dependence on **b**, and a direct minimization in the M-dimensional space of the components of **b** is hardly a practical procedure. We give here an indirect approach. We note that the contribution $| \lambda_v - \lambda_v |$ of the round-off errors to the bounds on λ_v depends directly on the norm $\| b \|$ of the vector **b**. This norm is limited by the condition (19). For instance, if we choose the 2-norm $\| b \|^2 = b^t b$, we have the restriction

$$\lambda_{\max}^{-1} (P) \leqslant \| b \|^2 \leqslant \lambda_{\min}^{-1} (P), \tag{23}$$

where $\lambda_{\max}(P)$, $\lambda_{\min}(P)$ are the eigenvalues of maximum and minimum modulus of the matrix P. In many calculations the bounds (23) are very broad, and the eigenvector of (7) may have a very large norm. This suggests that improved bounds on λ_v might be obtained if we search for the minimum of the Rayleigh quotient (16) subject to a *restriction* on the norm of the vector **b**. We shall then obtain an inferior estimate of λ_c, but a better round-off estimate; and by suitably choosing the maximum norm allowed for **b**, we can optimize the final bound obtained on λ_v.

We give here a practical procedure for this process. We search for the minimum of (16) subject to the restriction on **b**

$$b^+b/b^+Nb = \alpha^2.$$

That is, for the minimum of the functional

$$\min_{b,\mu}\left[\frac{b^+Hb}{b^+Nb} - \mu\left(\frac{b^+b}{b^+Nb} - \alpha^2\right)\right]. \tag{25}$$

This minimum leads to Eq. (24), and the double eigenvalue equation

$$(B_\mu - \gamma N)b = 0, \tag{26}$$

where

$$B_\mu = H - \mu I + \mu\alpha^2 N \tag{27}$$

The suggested procedure is to solve Eqs. (24) and (26) for a given choice of α^2; the resulting vector **b** is then used in (16) and (21) to bound λ_v, and α^2 varied to yield the best such bound. In general this will yield an improvement over the Rayleigh–Ritz choice of **b**, since a value of α^2 always exists for which (25) yields the same solution as (16). In the numerical example given in the next section, the gain in accuracy is considerable.

Finally, we consider the solution of (27). We write the following iterative scheme:

$$\begin{gathered} B_n = B(\mu_n), \\ [B_n - \gamma_{n+1}N]b_{n+1} = 0, \\ \mu_{n+1} = \mu_n + \delta_n, \end{gathered} \tag{28}$$

$$[B_n - \gamma_n N]C_n = [\alpha^2 N - I]b_n\,, \tag{29}$$

$$\delta_n = \frac{b_n{}^+(\alpha^2 N - I)b_n}{2b_n{}^+(I - \alpha^2 N)C_n} \tag{30}$$

Equation (29) is derived by differentiating (26) with respect to μ and setting $\partial b/\partial\mu = \mathbf{C}$. The scheme (28)–(30) has been found to work well in practice. It has the advantage that the iterations involve only the same type of operations as the original Rayleigh–Ritz method. A simpler method of solution, which also works well in practice, is to use inverse interpolation of (24) in μ.

IV. A NUMERICAL EXAMPLE

To illustrate the use of the bounds of Section II and the technique of Section III, we give a numerical example in which round-off errors are indeed significant. The example involves the calculation of a lower bound on the energy of the positronium ion $e^+e^-e^-$. This is a loosely bound system with an energy (in atomic units).

$$\lambda = -0.262000 \text{ a.u.}$$

We denote by H the three-body Hamiltonian including the Coulomb forces. Then if ζ is the energy of the first-excited state, which in this case consists of a free electron and a positronium atom,

$$\zeta = -0.25 \text{ a.u.},$$

we may derive a lower bound in the following way. We write

$$\gamma = 1/(\lambda - \zeta), \quad \mathscr{L} = H - \zeta \tag{31}$$

Then the wavefunction ψ for the system satisfies

$$(\mathscr{L} - \gamma\mathscr{L}^2)\psi = 0. \tag{32}$$

Equation (32) has the general form of (1). A lower bound λ_B for λ follows from the upper bound (2) for γ, and the procedures of this paper are applicable. We write a trial function ψ_{T} for ψ in the form

$$\begin{aligned} \psi_{\mathrm{T}} &= \sum_{L=1}^{m} a_i\phi_i \\ \phi_i(r_1, r_2, r_{12}) &= (1 + P_{12})\phi(l_i, m_i, n_i) \qquad (33) \\ \phi(l, m, n) &= \exp(-S\{Z^*(lr_1 + mr_2) + nr_{12}\}) \end{aligned}$$

The functional form (33), and the numerical results given here, derive from a general purpose program written for such calculations (3). The parameter S is an overall scale function which is varied to yield the best bound, while Z^* has here been assigned the value 2.

Figure 1 shows the raw results from a calculation with such a trial function, and plots the lower bound λ_B against the value of the scale factor S, for a 22 and a 50-term trial function. We see that the 22-term results are apparently weil-behaved, but that there are clearly large errors inherent in the 50-term calculation. These errors derive from the finite accuracy of the calculation of the matrix elements of L and of L^2, and illustrate the disastrous way in which these errors can build up.

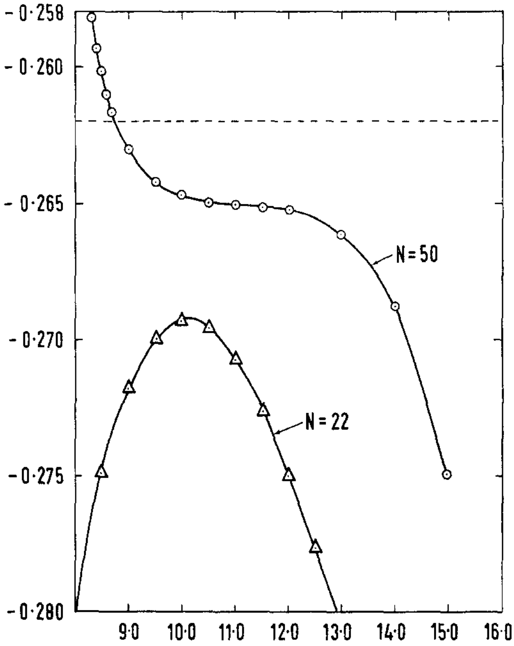

FIG. 1. Computed lower bounds on the positronium ion ground state. The ordinate gives the scale factor S defined in Eq. (33). The exact eigenvalue is shown as a dotted line. The results for a 50-term function show the presence of large round-off errors which cause the computed eigenvalue to rise above the exact result.

The estimated relative accuracy of these elements in this calculation is 10^{-10} for L, but only 10^{-6} for L^2. We thus have, for the error matrices h, n in Eq. (21),

$$\begin{aligned} \| h \| &\leqslant 10^{-10} \| L \|, \\ \| n \| &\leqslant 10^{-6} \| L^2 \|, \end{aligned} \qquad (34)$$

where L^2 is the $M \times M$ matrix of $\mathscr{L}^2$, and *not* $(L)^2$.

Curve II in Fig. 2 shows the modified bound for λ resulting from the inclusion of the round-off errors by the use of Eq. (21), for the $M = 50$ result. The results for $M = 22$ are not plotted, but show that for this case the round-off errors are indeed very small. We see that, although the round-off errors are very large for some values of the scale factor $\mathscr{L}$, we *can* derive a useful and reliable bound provided that their effect is included in this bound.

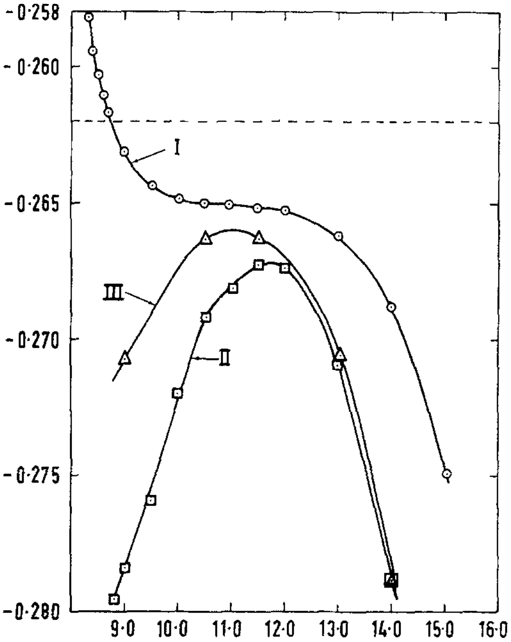

FIG. 2. Raw and corrected lower bounds for the positronium ion. Curve I: uncorrected for roundoff errors; curve II: corrected bound from Eq. (21); curve III: improved lower bounds derived by constraining the norm of the vector *a* using the iterative scheme (28)–(30).

Finally, curve III in Fig. 2 shows the improved bound which results from choosing the expansion coefficients **a** in the way suggested in paragraph 3. These values were obtained using the iterative scheme (28)–(30); convergence of the scheme was quite rapid, usually five or six iterations being sufficient to reduce the norm of **a** by the chosen factor α from its initial value for $\mu = 0$. We see that the procedure results in this instance in a modest but useful gain in the lower bound.

V. Extension to other variational schemes

We have restricted the discussion of this paper to the eigenvalue problem and the Rayleigh–Ritz variation principle; however, the method of Section III for bounding the round-off errors can be applied to a number of other variation principles for linear-operator equations in which expansions of the form (3) are made for the solutions. We illustrate these extensions by considering one, the Kohn variation principle for scattering states. The principle considers a trial function of the form

$$\psi_{\mathrm{T}} = \sum_{i=1}^{m} a_i \phi_i + \psi_s ,$$

where ψ_s is a normalized scattering part whose asymptotic form defines a trial phaseshift $\tan \delta_{\mathrm{T}}$. Then the variation principle for an energy $E = k^2$ (we set $2m/\hbar^2 = 1$),

$$\frac{\tan \delta_V}{k} = \frac{\tan \delta_{\mathrm{T}}}{k} - \int \psi_t (H - E) \psi_t \, dt,$$

takes the form

$$\frac{\tan \delta_V}{k} = \frac{\tan \delta_i}{k} - [a^+ L_a + a^+ L_s + L_s{}^+ a + L_{ss}] \tag{35}$$

where

$$\begin{aligned} L_{ss} &= \int \psi_s (H - E) \psi_s \, dt, \\ (L_s)_i &= \int \psi_s (H - E) \phi_i \, dt, \\ (L)_{ij} &= \int \psi_i (H - E) \psi_j \, dt. \end{aligned}$$

As before we assume that the computed L, L_s, L_{ss} have error matrices l, l_s, l_{ss}. Then we derive immediately a bound for the difference between the computed and the exact values of $\tan \delta_v$, for the given vector a,

$$| \tan \delta_V - \tan \delta_{\text{calc}} | \leqslant k[\| l \| \, \| a \|^2 + 2 \| a \| \, \| l_s \| + | l_{ss} |]. \tag{36}$$

Acknowledgments

The numerical calculations in this paper were carried out in part on the Atlas computer, Science Research Council, Chilton, and in part on the ICT 1905 computer of the University of Sussex. I am grateful to the directors and staff of both of these computing laboratories for the facilities provided.

APPENDIX. Matrix and Vector Norms

A matrix norm is a positive real number associated with a matrix satisfying the condition

$$\| \lambda A \| = | \lambda | \, \| A \|,$$
$$\| A \| + \| B \| \geqslant \| A + B \|,$$
$$\| A \| \, \| B \| \geqslant \| AB \|,$$
$$\| 0 \| = 0.$$

In this definition we regard a column vector as an ($n \times 1$) matrix. These properties lead directly to the results used in this paper. A number of specific norms are also discussed in (1). Of these, that leading to the sharpest bounds is the socalled infinity norm $\| \; \|_\infty$,

$$\| A \|_\infty = \max_i \sum_j | a_{ij} |.$$

For a vector, this yields

$$\| x \|_\infty = \max_i | x_i |.$$

References

1. J. H. Wilkinson, "The algebraic Eigenvalue Problem." p. 101. O.U.P. (1965).
2. Reference [1], p. 55.