JOURNAL OF COMPUTATIONAL PHYSICS 17, 87–94 (1975)

# Note

# The Iterative Calculation of a Few of the Lowest Eigenvalues and Corresponding Eigenvectors of Large Real-Symmetric Matrices

## I. Introduction

Large-scale configuration interaction (CI) calculations of electronic wavefunctions require the construction of a few eigenvalues and eigenvectors of large, sparse, real-symmetric matrices. The root-shifting optimal-relaxation (MOR) procedure developed by Shavitt, Bender, Pipano, and Hosteny [1] is probably the most widely used algorithm. Their scheme has several disadvantages, however. It requires the calculation of $\sum_J A_{IJ}c_J$ for one value of $I$ at a time; it has convergence difficulties for nearly degenerate eigenvalues; it requires a large amount of central memory to find several eigenvectors; and it must always find all eigenvalues below the one desired.

There are several different viewpoints from which one can show that the correction to the eigenvector at each step in the MOR procedure is reasonable. If $A$ is the matrix and $c$ is some vector, then the Rayleigh quotient of the scalar products,

$$\rho(\mathbf{c}) = (\mathbf{c}, \mathbf{A}\mathbf{c})/(\mathbf{c}, \mathbf{c}), \tag{1}$$

is a minimum at the vector corresponding to the lowest eigenvalue. Further, it has a saddle point at every eigenvector of $\mathbf{A}$. If one component of the vector $\mathbf{c}$, say $c_I$, is varied by an amount $\delta_I$, holding all other components constant, the optimum choice for $\delta_I$ from

$$\partial\rho/\partial c_I\,|_{c_I+\delta_I} = 0 \tag{2}$$

is simply

$$\delta_I = (\rho - A_{II})^{-1}\, q_I, \tag{3}$$

where $\mathbf{q} = (\mathbf{A} - \rho\mathbf{1})\mathbf{c}$ and $\rho$ is evaluated at $\mathbf{c} + \delta_I\hat{\mathbf{e}}_I$ (where $\hat{\mathbf{e}}_I$ is a unit vector). Cooper [2], Nesbet [3], and Shavitt [4] have developed algorithms for the lowest eigenvalue using this formula for $\delta_I$ with $\rho$ approximated by $\rho(\mathbf{c})$. Fadeev and Fadeeva [5] have shown how $\delta_I$ and $\rho(\mathbf{c} + \delta_I\hat{\mathbf{e}}_I)$ can be found exactly. Bender and Davidson [6, 7] have generalized this further to allow a few of the $c_I$ to be varied simultaneously. Shavitt *et al.* [1] have further modified (3) to allow higher roots to be found by shifting the lower ones.

A second derivation of (3) can be obtained from an expansion of $\rho$ in a Taylor's series to second order in $\boldsymbol{\delta}$,

$$\rho(\mathbf{c} + \boldsymbol{\delta}) \approx \rho(\mathbf{c}) + (\boldsymbol{\delta}, \nabla\rho) + \tfrac{1}{2}(\boldsymbol{\delta}, \mathbf{K}\boldsymbol{\delta}), \tag{4}$$

where

$$\partial\rho/\partial c_I \,|_{\mathbf{c}} = (\nabla\rho)_I = 2q_I/(\mathbf{c}, \mathbf{c}), \tag{5}$$

and

$$K_{IJ} = \left.\frac{\partial^2\rho}{\partial c_I\,\partial c_J}\right|_{\mathbf{c}} = \{2[A_{IJ} - \rho\delta_{IJ}] - 4[q_I c_J + c_I q_J]/(\mathbf{c}, \mathbf{c})\}/(\mathbf{c}, \mathbf{c}). \tag{6}$$

If $\rho$ in (4) is minimized with respect to one $\delta_I$ holding all the rest fixed at zero, Eq. (3) is obtained. The choice of $\boldsymbol{\delta}$ which minimizes $\rho$ when all components are varied simultaneously is more difficult to derive because all derivatives of $\rho$ in the direction $\mathbf{c}$ are identically zero. If $\boldsymbol{\delta}$ is varied in directions orthogonal to $\mathbf{c}$, a modified Newton–Raphson equation is obtained;

$$\mathbf{c} + \boldsymbol{\delta} \approx (\rho\mathbf{1} - \mathbf{A})^{-1}\,\mathbf{c}/(\mathbf{c}, (\rho\mathbf{1} - \mathbf{A})^{-1}\,\mathbf{c}). \tag{7}$$

It will be noticed that, apart from normalization, this is the inverse-iteration prescription for computing the eigenvector. Now (7) may be written as

$$(\rho\mathbf{1} - \mathbf{A})(\mathbf{c} + \boldsymbol{\delta}) \approx \epsilon\mathbf{c}, \tag{8}$$

where

$$\epsilon \approx \rho - \lambda, \tag{9}$$

and $\lambda$ is the exact eigenvalue. This may be rewritten as

$$(\rho - A_{II})\,\delta_I \approx q_I + \sum_{J \neq I} A_{IJ}\delta_J + \epsilon c_I\,. \tag{10}$$

Neglect of $\delta$ and $\epsilon$ on the right-hand side of (10) leads to (3) again. In most cases, however, even though the $\epsilon c_I$ are negligible, the $A_{IJ}\delta_J$ are comparable to the $q_I$. Consequently, unless the $A_{IJ}/(\rho - A_{II})$ are unusually small, $\delta_I = q_I/(\rho - A_{II})$ is not a very good approximation to the solution to (8). A Gauss–Seidel iterative solution to (10) in the form

$$(\rho - A_{II})(\delta_I^{(n+1)} - \delta_I^{(n)}) = [(\mathbf{A} - \rho\mathbf{1})(\mathbf{c} + \boldsymbol{\delta}^{(n)})]_I + \epsilon c_I \tag{11}$$

will in fact diverge unless the norm of the matrix $\mathbf{D}^{-1}\mathbf{F}$ satisfies

$$\| \mathbf{D}^{-1}\mathbf{F} \| < 1, \tag{12}$$

where

$$D_{IJ} = \delta_{IJ}(\rho - A_{II}), \tag{13}$$

and

$$F_{IJ} = A_{IJ}(1 - \delta_{IJ}). \tag{14}$$

Condition (12) does not seem to be obeyed by most matrices encountered in CI calculations. Even though use of Eq. (3) to change the $\delta_I$ sequentially is a stable method for finding the lowest eigenvalue, simultaneous change of all the $c_I$ by $q_I/(\rho - A_{II})$ usually does not converge. Thus, the seemingly small distinction between simultaneous or sequential relaxation of the $c_I$ often makes the difference between divergence or convergence. Even in the sequential case, large $A_{IJ}$ will slow convergence, and direct convergence on higher eigenvalues generally is not possible. In the next section a new method, based on simultaneous variation of the $c_I$ by $q_I/(\rho - A_{II})$, will be presented which leads to stable monotonic convergence for higher eigenvalues.

The principal alternatives to the method of optimal relaxation are the gradient and power methods. Since the gradient of $\rho$ is

$$\nabla\rho = 2(\mathbf{Ac} - \rho\mathbf{c})/(\mathbf{c}, \mathbf{c}), \tag{15}$$

moving along the line $\mathbf{c} + \alpha\nabla\rho$ is equivalent to moving along $\mathbf{c} + \alpha\mathbf{Ac}$. Hence, the gradient method of Hestenes and Karush [8, 9] is equivalent to iteratively choosing $\mathbf{c}_{(k+1)} = \mathbf{c}_{(k)} + \alpha_k\mathbf{Ac}_{(k)}$. Because $\nabla\rho$ and $\boldsymbol{\delta}$ in (3) are quite different (unless $\mathbf{K}$ is proportional to a unit matrix), the gradient method can be expected to give poor convergence in general. This problem is closely related to the well-known difficulties of the gradient method under changes of the scale of the independent variables. The direct power method based on the fact that $\mathbf{A}^k\mathbf{c}$ converges to the dominant eigenvector is likewise slowly convergent for matrices arising in CI calculations because the dominant eigenvalues tend to be close together in magnitude.

The Lanczos method [10] is a slight improvement on these latter methods. The Lanczos method (although usually expressed differently) is equivalent to

$$\mathbf{c}_{(k)} = \mathbf{c}_{(0)} + \sum_{i=1}^{k} \alpha_i^{(k)}\nabla\rho(\mathbf{c}_{(i-1)}), \tag{16}$$

and the $\alpha_i^{(k)}$ are recomputed variationally after each $\nabla\rho$ is added. This is also equivalent to (with a different value of $\alpha_i^{(k)}$)

$$\mathbf{c}_{(k)} = \mathbf{c} + \sum_{i=1}^{k} \alpha_i^{(k)} \mathbf{A}^k \mathbf{c}. \tag{17}$$

Thus, the Lanczos method is essentially a bordering technique which iteratively expands in the sequence of vectors $\mathbf{A}^k\mathbf{c}$. Because of its relation to the gradient expansion (16), better convergence can be expected than with an iteratively bordered expansion in the set $\hat{\mathbf{e}}_k$. As can be seen from the previous discussion of Eq. (3), this particular choice of expansion vectors is still not the best for optimum convergence. Lanczos has shown, however, that this particular choice does lead to the elegant simplification that the matrix from which the $\alpha$'s are calculated is tridiagonal.

## II. Compromise Method

Equations (3, 16) are suggestive of a new method. Let

$$\mathbf{c}_{(k)} = \mathbf{c}_{(0)} + \sum_{i=1}^{k} \alpha_i^{(k)} \boldsymbol{\xi}_i , \tag{18}$$

where the components of the vector $\boldsymbol{\xi}_i$ are found from the components of the vector $\mathbf{q}_i$ by

$$\xi_{J,i+1} = (\rho(\mathbf{c}_{(i)}) - A_{JJ})^{-1} \, q_{J,i} , \tag{19}$$

$$\mathbf{q}_i = (\mathbf{A} - \rho(\mathbf{c}_{(i)})\, \mathbf{1})\, \mathbf{c}_{(i)} . \tag{20}$$

Since expansion in an orthonormal basis is simpler, the equivalent form

$$\mathbf{c}_{(k)} = \sum_{i=0}^{k} \alpha_i^{(k)} \mathbf{b}_i , \tag{21}$$

$$\mathbf{b}_0 = \mathbf{c}_0 , \tag{22}$$

$$\mathbf{b}_i = \mathbf{d}_i / \| \mathbf{d}_i \|, \tag{23}$$

$$\mathbf{d}_i = \left[ \prod_{j=0}^{i-1} (\mathbf{1} - \mathbf{b}_j \mathbf{b}_j^T) \right] \boldsymbol{\xi}_i , \tag{24}$$

$$\| \mathbf{d}_i \| = (\mathbf{d}_i , \mathbf{d}_i)^{1/2}$$

is preferable. While this method does not lead to the elegance of the Lanczos method, it offers a better chance for rapid convergence. Like the Lanczos method, it generates $k$ approximate eigenvalues at each step, so it offers a prospect for generating higher eigenvalues. In fact, if the $\mathbf{q}_i$ are based on approximate eigenvectors associated with higher eigenvalues, selective convergence on excited states is conceivable.

## III. Computational Details

A. If the $k$th eigenvalue is wanted, select a zeroth-order orthonormal subspace $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_l$ $(l \geqslant k)$ spanning the dominant components of the first $k$ eigenvalues. Form and save $\mathbf{Ab}_1, \mathbf{Ab}_2, \ldots, \mathbf{Ab}_l$ and $(\mathbf{b}_i, \mathbf{Ab}_j) = \tilde{A}_{ij}$, $1 \leqslant i \leqslant j \leqslant l$. Diagonalize $\tilde{\mathbf{A}}$ using a standard method for small matrices. Select the $k$th eigenvalue $\lambda_k^{(l)}$ and the corresponding eigenvector $\boldsymbol{\alpha}_k^{(l)}$.

B. Form $\mathbf{q}_M = \sum_{i=1}^{M} \alpha_{i,k}^{(M)}(\mathbf{Ab}_i) - \sum_{i=1}^{M} \alpha_{i,k}^{(M)} \lambda_k^{(M)} \mathbf{b}_i$. Here, $M$ is the dimension of $\tilde{A}$ used to find $\boldsymbol{\alpha}$ and $\lambda$.

C. Form $\| \mathbf{q}_M \|$ and check convergence by the Weinstein lower bound formula [11], $\lambda_k^{(M)} - \| q_M \| \leqslant \lambda_k \leqslant \lambda_k^{(M)}$. This is a pessimistic convergence test since $\| \mathbf{q}_M \| \to 0$ at about the same rate as $\| \mathbf{A} - \lambda_k \| \, \| \mathbf{c}_{(M)} - \mathbf{c}_{\text{exact}} \|$. The $\lambda_k^{(M)}$ converges to $\lambda_k$ much more rapidly. In cases where $\| \mathbf{A} - \lambda_k \|$ is very different from unity, a test on $\alpha_{M,k}^{(M)}$ is better than a test on $\| \mathbf{q}_M \|$.

D. Form $\xi_{I,(M+1)} = (\lambda_k^{(M)} - A_{II})^{-1} q_{I,M}$, $I = 1, \ldots, N$.

E. Form $\mathbf{d}_{(M+1)} = [\prod_{i=1}^{M} (\mathbf{1} - \mathbf{b}_i \mathbf{b}_i^T)] \boldsymbol{\xi}_{(M+1)}$.

F. Form $\mathbf{b}_{(M+1)} = \mathbf{d}_{(M+1)} / \| \mathbf{d}_{(M+1)} \|$.

G. Form $\mathbf{Ab}_{(M+1)}$.

H. Form $\tilde{a}_{i,M+1} = (\mathbf{b}_i, \mathbf{Ab}_{(M+1)})$, $i = 1, \ldots, M + 1$.

I. Diagonalize $\tilde{\mathbf{A}}$ and return to step $B$ with $\alpha_k^{(M+1)}$ and $\lambda_k^{(M+1)}$.

If several eigenvalues are wanted, the first $l$ of the $\sum_{i=1}^{M} \alpha_{ij} \mathbf{b}_i$ at the end of finding one root often provides a good starting set for the next root. The slowest step in this procedure for large matrices is the formation of **Ab**. All of the other steps together require negligible time. For this reason, this method requires the same time per iteration as the Lanczos method, the gradient method, and the relaxation method. Since the $\mathbf{b}_i$ and $\mathbf{Ab}_i$ are numerous and large in dimension, they must be kept in auxiliary storage. Only central memory storage space for two vectors is ever required. If $M$ becomes inconveniently large, the current set of $\sum_{i=1}^{M} \alpha_{ij} \mathbf{b}_i$, $j = 1, \ldots, l$, can be taken as a new initial set and the calculation restarted with step (A).

## IV. Test Results

This method has now been applied to a few matrices with the convergence criteria $\| \mathbf{q} \| \leqslant 10^{-6}$. Among the matrices used were some for the LiF molecule from the work of Kahn, *et al.* [12]. These matrices were of dimension ~1100 and had nearly degenerate third and fourth roots as well as fairly close first and second roots. In one example, the actual first and second roots were reversed in order from the results of the zeroth-order subspace. In several cases the MOR procedure had failed to converge for the third root after hundreds of iterations. In every case convergence was reached with the present method after only 10–20 iterations per eigenvalue (51 iterations total for 4 eigenvalues). Near degeneracy did not seem to affect the convergence rate. Since the time per iteration was the same as the MOR procedure, the overall speed was much faster because of the reduced number of iterations. For the lowest eigenvalue alone, the MOR algorithm remained superior requiring about three fewer iterations for convergence. One comparison with the Lanczos method was made using a matrix of dimension 372. This new method converged $\| \mathbf{q} \|$ to less than $10^{-6}$ in ten iterations, while the Lanczos method had only reached $\| \mathbf{q} \| = 2 \times 10^{-2}$ after 28 iterations.

## V. Conclusions

In the opening paragraph four difficulties with the MOR procedure were mentioned. By direct calculation, it has been verified that the convergence difficulties, at least in these examples, do not occur in this new method. There is no reason why this method would not work even for exactly degenerate roots. Also, this method requires core storage for only two vectors at once, regardless of the number of roots found. In fact, the method could easily be adapted to block matrix operations which would require very little core for a few eigenvectors of matrices of dimension $10^6$ or larger. If desired, this new method can also be used to find higher eigenvalues directly without finding accurate values for lower eigenvalues.

Finally, the new method requires only the matrix operation $\mathbf{Ab}$ and not the sequential calculation of $(\mathbf{Ab})_I$. For very large CI calculations, this may offer some advantages. In CI calculations the eigenvalue problem for a $N \times N$ matrix,

$$\sum_{J=1}^{N} A_{IJ} x_J = \lambda x_I ,$$

can be rewritten as

$$\sum_{J=1}^{N} \sum_{ijkl}^{K} h_{ijkl} \Gamma_{ijklIJ} x_J = \lambda x_I ,$$

where the $h_{ijkl}$ are relatively expensive numbers to obtain and the $\Gamma_{ijklIJ}$, while very numerous, are relatively simple to compute. For very large matrices for which $N^2 \gg K^4 > 10^8$, it becomes impractical to actually form the $A_{IJ}$. In this case the problem can be reformulated as

$$\sum_{ijkl} h_{ijkl} \sum_{J} \Gamma_{ijklIJ} x_J = \lambda x_I . \tag{18}$$

The $\Gamma$ supermatrix is reconstructed during each iteration while forming **Ab**. In this event, a method which does not require the formation of the $(\mathbf{Ab})_I$ in any particular order has obvious advantages. Calculations based on (18) using Rayleigh–Schrödinger perturbation theory have recently been reported by Roos [13], and calculations using the Lanczos method have been reported by Hausman, Bender, and Bloom [14]. The method suggested here should improve the convergence of these calculations.

## Acknowledgment

The author is indebted to the Guggenheim Foundation for a fellowship which supported this work. Also, Battelle Memorial Institute provided computer resources and other support. The author also thanks I. Shavitt and L. Kahn for use of their matrices and several helpful conversations.

The author is on leave from the University of Washington, Seattle, Washington.

## References

1. I. Shavitt, C. F. Bender, A. Pipano, and R. P. Hosteny, *J. Computational Phys.* **11** (1973), 90.
2. J. L. B. Cooper, *Quart. Appl. Math.* **6** (1948), 179.
3. R. K. Nesbet, *J. Chem. Phys.* **43** (1965), 311.
4. I. Shavitt, *J. Computational Phys.* **6** (1970), 124.
5. D. K. Fadeev and V. N. Fadeeva, "Computational Methods of Linear Algebra" (English Translation), Section 61, Freeman, San Francisco, CA, 1963.
6. C. F. Bender, Ph.D. Thesis, University of Washington, Seattle, WA, 1968.
7. C. F. Bender and E. R. Davidson, *Phys. Rev.* **183** (1969), 23.
8. W. Karush, *Pacific J. Math.* **1** (1951), 233.
9. M. R. Hestenes, in "Simultaneous Linear Equations and the Determination of Eigenvalues" (L. J. Page and O. Taussky, Eds.), Chap. 12, National Bureau of Standards Applied Mathematics Series, No. 29, U. S. Government Printing Office, Washington, DC, 1953.
10. C. Lanczos, *J. Res. Nat. Bur. Stand.* **45** (1950), 255.
11. D. H. Weinstein, *Proc. Nat. Acad. Sci* **20** (1934), 529.

12. L. R. Kahn, P. J. Hay, and I. Shavitt, *J. Chem. Phys.* **61** (1974), 3530.
13. B. Roos, *Chem. Phys. Lett.* **15** (1972), 153.
14. R. F. Hausman, C. F. Bender, and S. D. Bloom, private communication.

Received: August 26, 1974

Ernest R. Davidson

*Battelle Memorial Institute*
*505 King Avenue*
*Columbus, Ohio 43220*