An iterative method for calculating low lying eigenvalues of an Hermitian operator

1977 J. Phys. A: Math. Gen. 10 1089

(http://iopscience.iop.org/0305-4470/10/7/006)

J. Phys. A: Math. Gen., Vol. 10, No. 7, 1977. Printed in Great Britain.

An iterative method for calculating low lying eigenvalues of an Hermitian operator

W A Berger, H G Miller, K-G Kreuzer and R M Dreizler

Institut für Theoretische Physik, Universität Frankfurt, West Germany

Received 25 February 1977

Abstract. An iterative method which may under certain circumstances be an alternative to the Lanczos method has been investigated. The main properties of this scheme are discussed and a small numerical example is presented.

1. Introduction

In the last few years a number of groups (Sebe and Nachamkin 1969, Cole *et al* 1973, 1974a–c, 1975a–d, Toepfer 1975) have made extensive use of the Lanczos algorithm (Lanczos 1950) to perform diagonalisations of quantum mechanical operators. The algorithm itself provides a method of generating a matrix representation of a symmetric (or Hermitian) operator in which the operator has a tri-diagonal form and is therefore easier to diagonalise. However, as has been shown (Whitehead 1972) the extreme usefulness of the algorithm rests on the fact that one need not generate the full tri-diagonal matrix if one is only interested in some of the eigenvectors corresponding to either the lower (or higher) eigenvalues. In this case the algorithm defines an iterative scheme which exhibits rapid convergence at the extreme ends of the eigenvalue spectrum.

On the other hand certain numerical problems are inherent in the method because of the necessity of successively generating basis vectors of the desired representation which are orthogonal. Hence because of the numerical errors in generating the basis vectors it may be necessary in the nth step to re-orthogonalise the nth basis vector to the foregoing $n-1$ basis vectors. Needless to say this requires additional computer time and more storage space.

In the present work we have investigated another algorithm which under certain conditions may be considered as an useful alternative to the iterative Lanczos scheme as discussed by Whitehead (1972). In essence the algorithm may be stated as follows. Starting with an arbitrary start vector the first two iterations of the Lanczos algorithm are performed and the subsequent 2×2 matrix is diagonalised. The vector corresponding to the lowest eigenvalue of the 2×2 matrix is taken then as the new start vector and is used to generate a subsequent 2×2 matrix which is again diagonalised allowing as before the selection of a new start vector. We shall prove that iterations performed in this manner are convergent and converge ultimately to the eigenvector corresponding to the lowest eigenvalue which has a non-zero component in the original start vector. Hence assuming that the original start vector contains a non-zero component of the

eigenvector corresponding to the lowest eigenvalue this algorithm provides a means of obtaining exactly the ground state eigenvector of the operator considered.

In some respects this algorithm reduces part of the aforementioned numerical and computational problems prevalent in the original Lanczos algorithm. Since we are in essence iterating a 2×2 matrix only two vectors are required in any given iteration step, i.e. less storage space in a computer is required. Furthermore, in each iteration step we have only to construct one vector which is orthogonal to the start vector in that iteration and as is the case in the original Lanczos algorithm need not worry about its orthogonality with respect to all the previous vectors which have been generated. On the other hand the present algorithm although convergent, converges at a slower rate as compared to the original Lanczos algorithm and as previously stated yields only one eigenvector.

In the remaining sections of the present paper we describe the method in more detail and present a proof of its convergence. Furthermore, we show that this algorithm is closely related to a special form of the power method. A small numerical calculation has been performed to illustrate the method.

2. Description of the method

This section is divided into three parts. In § 2.1 we give a description of the iterative algorithm; in § 2.2 we present a proof of the convergence properties of the algorithm; and in § 2.3 we show its relation to other methods.

2.1. The algorithm

Let A be a linear Hermitian operator defined everywhere in a finite dimensional vector space v†. A is then a continuous and bounded operator. It could for example be the projection of some physical operator onto a subspace of the complete Hilbert space. Let $|k\rangle$ be an arbitrary normalised vector in the space v. Acting on it with A we obtain

$$A|k\rangle = e_k|k\rangle + v_k|k_\perp\rangle. \tag{2.1}$$

If we demand that

$$\langle k|k_\perp\rangle = 0 \qquad \text{and} \qquad \langle k_\perp|k_\perp\rangle = 1$$

the quantities e_k and v_k are given by

$$e_k = \langle k|A|k\rangle \tag{2.2}$$

$$v_k = \pm(\langle k|A^2|k\rangle - e_k^2)^{1/2}. \tag{2.3}$$

In equation (2.3) we shall choose the positive sign. The two states $|k\rangle$ and $|k_\perp\rangle$ now define a two-dimensional subspace and we can diagonalise A in this subspace which is equivalent to the diagonalisation of P_kAP_k where P_k is the projector on this two-dimensional subspace. The 2×2 matrix to be diagonalised has the form

$$\begin{pmatrix} e_k & v_k \\ v_k & \alpha_k \end{pmatrix} \tag{2.4}$$

† A need not necessarily be given in its matrix representation.

with

$$\alpha_k = \langle k_\perp | A | k_\perp \rangle \tag{2.5}$$

and we note that this is a real matrix.

The resulting two eigenvalues have the interesting property, that one always lies above e_k and the other always below e_k as long as $v_k \neq 0$. This is a consequence of the eigenvalue separation theorem (Householder 1964, Wilkinson 1965) but can easily be seen by looking at the explicit form for the eigenvalues. Upon diagonalisation of the 2×2 matrix given in equation (2.4) the normalised eigenvector $|k+1\rangle$ corresponding to the lower eigenvalue e_{k+1} is then used as new start vector in equation (2.1) and the process is repeated. This now defines the iterative scheme we want to consider.

The successive eigenvalues form a monotonically decreasing sequence, i.e.

$$e_{k+1} < e_k.$$

Since this is a sequence of expectation values of a bounded Hermitian operator, it is bounded by the lowest eigenvalue of this operator and therefore must be convergent.

2.2. *Convergence properties*

From the fact that the e_k form a convergent sequence it follows, as is shown in appendix 1, that v_k converges to zero. But v_k is just the square root of the variance of A computed with vectors $|k\rangle$. If we suppose that the sequence $|k\rangle$ converges to $|\phi\rangle \in \upsilon$ we then know, because A is continuous that the variance of $|\phi\rangle$ will vanish, and so $|\phi\rangle$ will be an eigenvector.

We shall now show that this limiting vector actually exists. Expanding $|k\rangle$ in terms of the eigenvectors of A we obtain

$$|k\rangle = \sum_\lambda b_k^\lambda |\lambda\rangle \tag{2.6}$$

where $A|\lambda\rangle = E_\lambda |\lambda\rangle$ and, assuming non-degeneracy for the moment

$$E_\lambda < E_{\lambda+1}. \tag{2.7}$$

The normalisation condition and variance may now be given by

$$\sum_\lambda |b_k^\lambda|^2 = 1 \tag{2.8}$$

$$\sum |b_k^\lambda|^2 (E_\lambda - e_k)^2 = v_k \xrightarrow[k\to\infty]{} 0. \tag{2.9}$$

Equation (2.9) implies, that for each λ either $|b_k^\lambda| \xrightarrow[k\to\infty]{} 0$ or $(E_\lambda - e_k) \xrightarrow[k\to\infty]{} 0$. Since we assume non-degeneracy and because of equation (2.8) one and only one μ must exist, such that

$$e_k \xrightarrow[k\to\infty]{} E_\mu \qquad \text{and} \qquad |b_k^\lambda| \xrightarrow[k\to\infty]{} \delta_{\lambda\mu}.$$

Let us now assume that $|\lambda_0\rangle$ is the non-zero component in the initial start vector $|k=0\rangle$ with the lowest eigenvalue, i.e.

$$|0\rangle = \sum_{\lambda \geqslant \lambda_0} b_0^\lambda |\lambda\rangle. \tag{2.10}$$

As is demonstrated in appendix 2 it follows from the explicit construction of the iterated vectors that $b_{k+1}^{\lambda o}$ has the same phase as $b_k^{\lambda o}$ and that we have

$$|b_{k+1}^{\lambda o}| \geqslant |b_k^{\lambda o}|$$

which proves that $b_k^{\lambda o}$ converges to 1 times the phase factor of $b_0^{\lambda o}$ or

$$|k\rangle \xrightarrow[k\to\infty]{} |\phi\rangle$$

$|\phi\rangle$ being proportional to $|\lambda_0\rangle$. This completes the proof in the non-degenerate case.

Suppose now in the start vector there are components belonging to an n-fold degenerate eigenvalue E_λ. We collect the vectors which belong to the same degenerate subspace υ_λ, i.e. for each λ we form

$$|\bar{\lambda}\rangle = \sum_\mu c_{\lambda\mu}|\lambda_\mu\rangle$$

with the $c_{\lambda\mu}$ given in the initial start vector. Now we can rewrite this as a constant b_0^λ times a normalised vector:

$$|\bar{\lambda}\rangle = b_0^\lambda|\lambda\rangle, \qquad \langle\lambda|\lambda\rangle = 1.$$

If we do this for all degenerate vectors occurring in the start vector (equation (2.10)) equation (2.6) and the additional condition (2.7) are still valid so that the remainder of the proof is unchanged.

2.3. Relation to other methods

From the previous section it should be obvious that the present algorithm is closely related to the full Lanczos algorithm. In both cases the iterated vectors are essentially special linear combinations of the start vector and vectors of the form $A^n|0\rangle$, which are used in the power method and its variants (Faddejew and Faddejewa 1964). In our case the improvement over the power method may be understood from the following.

It is well known that the convergence of the usual power method can in principle be improved by allowing for a spectral shift of the matrix A during the iterations such that the iterated vector has the form

$$|\widetilde{k+1}\rangle = \prod_{n=1}^{k+1}(A-\lambda_n)|0\rangle = (A-\lambda_{k+1})|\tilde{k}\rangle.$$

However in principle not much can be said about the choice of λ_n unless at least some of the eigenvalues of the matrix are known. On the other hand if one is only interested in either the lowest (or highest) eigenvalues one may attempt to determine them in such a way, that convergence in each iteration step is optimal. This can be done by finding the extrema of the following functional:

$$\tilde{e}_{k+1} = \frac{\langle\tilde{k}|(A-\lambda_{k+1})A(A-\lambda_{k+1})|\tilde{k}\rangle}{\langle\tilde{k}|(A-\lambda_{k+1})(A-\lambda_{k+1})|\tilde{k}\rangle}.$$

But this expression is just the expectation value of the vector

$$|\widetilde{k+1}\rangle = (A-\lambda_{k+1})|\tilde{k}\rangle$$

which can be written of course as

$$|\widetilde{k+1}\rangle = c|\tilde{k}\rangle + (A - e_k)|\tilde{k}\rangle \tag{2.11}$$

where e_k is $\langle\tilde{k}|A|\tilde{k}\rangle/\langle\tilde{k}|\tilde{k}\rangle$ and c is suitably chosen. However since the extrema of the expectation values of the vector $|\widetilde{k+1}\rangle$ as defined in equation (2.11) are the two eigenvalues of A in the two-dimensional subspace defined by $|\tilde{k}\rangle$ and $(A-e_k)|\tilde{k}\rangle$ the optimal λ_{k+1} has to yield exactly one of these eigenvalues. In other words, the requirement of optimisation of the spectral shift in each iteration is satisfied in the present method.

The improvement of this method over the usual power method should now be clear. In cases where one looks for the dominant eigenvalue (i.e. the eigenvalue with the largest modulus) one can improve convergence; in cases where one looks for the lowest or highest eigenvalue one can enforce convergence to the eigenvalue required irrespective of whether it is dominant or not.

3. Numerical results and discussion

In order to demonstrate the convergence of the present method as compared with other methods we have performed the following numerical calculations. A 100×100 symmetric random matrix was generated. By subtracting the largest eigenvalue from the diagonal matrix elements the spectrum was made negative definite. This was done to simplify calculations performed using the power method which we wanted to include for comparison and which does not essentially influence the results obtained from the other methods.

The present method, the Lanczos algorithm and the power method were used in order to obtain the lowest eigenvalue and eigenvector of the matrix (see figure 1). In order to obtain some feeling for the convergence rates of the respective methods we varied the amplitude of the lowest eigenvector in the original start vector. As can be seen from figure 1 the present method as expected converged at a slower rate than the Lanczos algorithm but appreciably faster than the power method. Hence bearing in mind that each iteration step in this method is somewhat simpler than that in the Lanczos algorithm (particularly if one must re-orthogonalise) the slower convergence rate in our opinion does not rule out the method as an alternative to the Lanczos algorithm.

On the other hand all three of the methods exhibit under certain conditions a disturbing tendency to converge to the wrong eigenvalue (see figure 1(*c*)). This problem arises when the original start vector contains large components (relative to that of the lowest lying eigenvector) of one or more low lying eigenvectors of the matrix. Unfortunately the only remedy at present for this problem is to perform additional iterations.

However in spite of this both methods provide a basis for realistic applications in the field of nuclear physics. One of the main advantages, in our opinion, being that quantum mechanical operators to be diagonalised need not be given in a matrix representation which is a feature of all direct iteration methods and may therefore reduce storage requirements and computing time (Whitehead 1972).

Finally we wish to point out that the present method may in principle be used to obtain other eigenvalues and eigenvectors in a simple, well known (Faddejew and

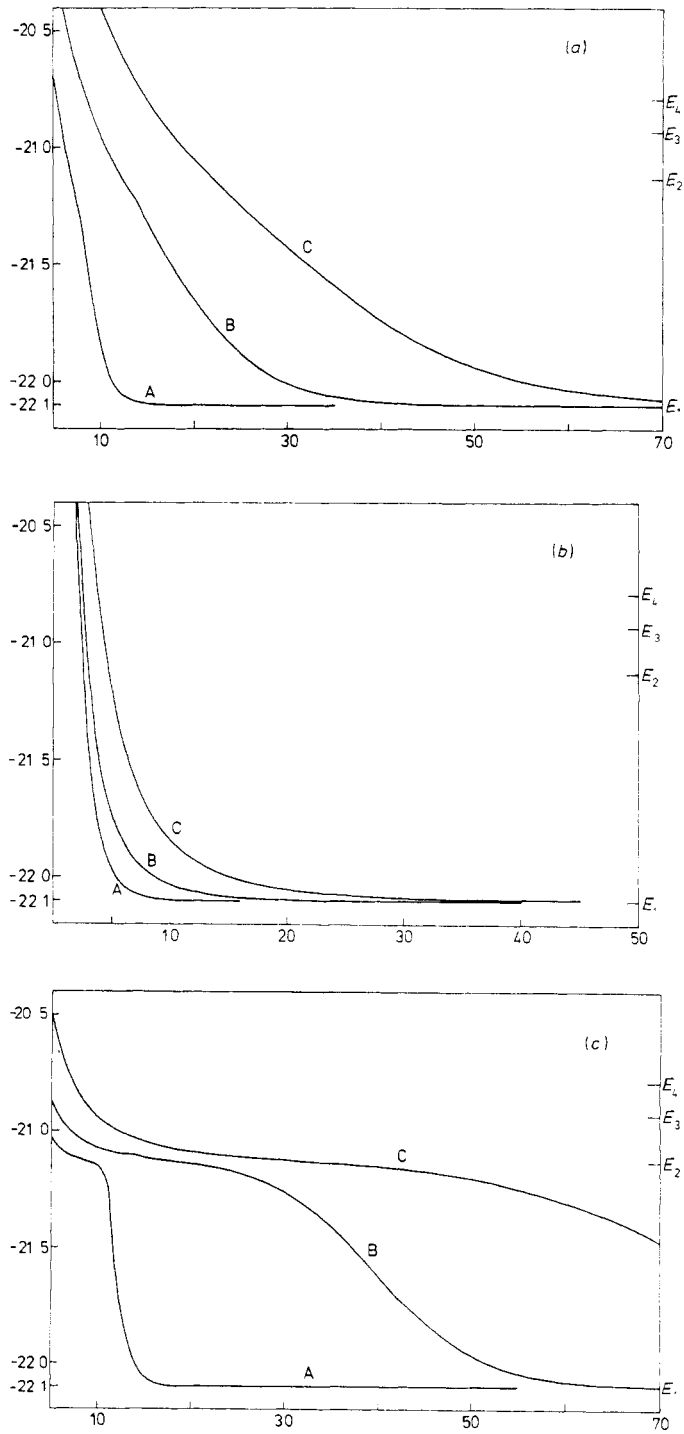

Figure 1. The convergence rate of the lowest-lying eigenvector in a random 100×100 symmetric matrix for: A, Lanczos method; B, the present method; and C, the power method. On the right-hand side of the figure the four lowest lying eigenvalues of the matrix are given. The amplitude of the lowest lying eigenvector in the original start vector is: (*a*) 0·1%, (*b*) 10%, and (*c*) 0·015%.

Faddejewa 1964) way. Once the eigenvector of the lowest lying eigenvalue has been obtained it may be projected out of the original start vector yielding a vector which contains no component of the lowest lying eigenvector. This vector may then be used as a new start vector and the algorithm will yield ultimately the eigenvector corresponding to the next lowest lying eigenvalue whose amplitude in the start vector is non-zero.

Appendix 1

Proof that v_k (equation (2.3)) converges to zero. The explicit form of e_{k+1} after diagonalisation of the matrix (2.4) is:

$$e_{k+1} = \tfrac{1}{2}(\alpha_k + e_k) - \tfrac{1}{2}[(\alpha_k - e_k)^2 + v_k^2]^{1/2}.$$

From this follows (assuming of course $v_k \neq 0$)

$$|e_{k+1} - e_k| = |x_k - (x_k^2 + v_k^2)^{1/2}| = |v_k^2||x_k + (x_k^2 + v_k^2)^{1/2}|^{-1}$$

with $x = \frac{1}{2}(\alpha_{k+1} - e_k)$. Since v_k and x_k are bounded it follows from the convergence of e_k that $v_k \xrightarrow[k\to\infty]{} 0$.

Appendix 2

Proof of the convergence of $b_k^{\lambda_0}$. From the inequalities

$$e_{k+1} < e_k \qquad \text{and} \qquad e_{k+1} \geqslant E_{\lambda_0}$$

one obtains

$$|e_{k+1} - e_k| < |E_{\lambda_0} - e_k|$$

and therefore, constructing $|k+1\rangle$ explicitly from $|k\rangle$

$$\frac{v_{k+1}^{\lambda_0}}{b_k^{\lambda_0}} = \frac{1 + (e_{k+1} - e_k)(E_{\lambda_0} - e_k)v_k^{-4}}{[1 + (e_{k+1} - e_k)^2 v_k^{-4}]^{1/2}} > \frac{1 + (e_{k+1} - e_k)(E_{\lambda_0} - e_k)v_k^{-4}}{[1 + (e_{k+1} - e_k)(E_{\lambda_0} - e_k)v_k^{-4}]^{1/2}}$$

$$= [1 + (e_{k+1} - e_k)(E_{\lambda_0} - e_k)v_k^{-4}]^{1/2} > 1.$$

References

Cole B J, Watt A and Whitehead R R 1973 *Phys. Lett.* **45B** 429–32
—— 1974a *Phys. Lett.* **49B** 133–6
—— 1974b *J. Phys. A: Math., Nucl. Gen.* **7** 1374–98
—— 1974c *J. Phys. A: Math., Nucl. Gen.* **7** 1399–409
—— 1975a *J. Phys. G: Nucl. Phys.* **1** 17–34
—— 1975b *J. Phys. G: Nucl. Phys.* **1** 213–29
—— 1975c *J. Phys. G: Nucl. Phys.* **1** 303–23
—— 1975d *J. Phys. G: Nucl. Phys.* **1** 935–55
Faddejew D K and Faddejewa W N 1964 *Numerische Methoden der Linearen Algebra* (München: Oldenbourg)
Householder A S 1964 *The Theory of Matrices in Numerical Analysis* (New York: Blaschell)
Lanczos C 1950 *J. Res. NBS* **45** 255
Sebe T and Nachamkin J 1969 *Ann. Phys., NY* **51** 100
Toepfer A 1975 *Diplomarbeit* University of Frankfurt
Whitehead R R 1972 *Nucl. Phys.* A **182** 290
Wilkinson J H 1965 *The Algebraic Eigenvalue Problem* (Oxford: Clarendon)