Application of Lanczos and conjugate gradient methods to a class of computational problems in physics

Kashyap V. Vasavada
Department of Physics, Indiana–Purdue University, Indianapolis, Indiana 46223
Jack H. Freed
Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853

(Received 25 August 1988; accepted 10 March 1989)

It is shown that the equivalence of the Lanczos and the conjugate gradient algorithms can be used to give a very powerful method to study linear systems in which complex symmetric matrices arise. This method is illustrated for electron spin resonance calculations, but is applicable to a wide class of problems in physics and engineering.

A number of problems in science and engineering can be reduced to the solution of an equation of the type

$$A'|u\rangle = |v\rangle, \tag{1}$$

where A' is a large but sparse $N \times N$ matrix that can be calculated from the basic theoretical models, and $|v\rangle$ is a known N-dimensional vector. Calculation of the unknown vector $|u\rangle$ enables one to find various properties of the system. The sparsity of A' in many applications arises from the fact that, in some approximation, a given state is coupled to very few other states (due to selection rules). Our own experience with such equations has been in connection with electron spin resonance (ESR) calculations and solutions of Fokker–Planck equations, which in general yield *complex* matrices A'. However, matrix equations such as Eq. (1) arise in different areas when the relevant quantities are expanded in terms of eigenfunctions of some differential operator [e.g., the Wigner rotation functions $D^{L}_{MK}(\Omega)$ for angular variables[1]]. They also arise when finite difference[2] or finite element[3] methods are used. Frequently A' can be expressed as

$$A' = i\,\Delta\omega\,\mathbb{1} + A, \tag{2}$$

where (in the ESR case) $\Delta\omega = \omega - \omega_0$, ω_0 being the Larmor frequency at the center of the spectrum and ω the angular frequency of the applied radiation field. Here $\mathbb{1}$ is the identity matrix and A is independent of $\Delta\omega$. Then only one diagonalization for the entire range in $\Delta\omega$ is required to solve Eq. (1) instead of inverting this equation for many values of $\Delta\omega$. The reason, of course, is that the diagonalization by a similarity or orthogonal transformation leaves the identity matrix unchanged. This saves a very large amount of computer time in many practical cases where one needs to know the spectrum at hundreds of values of the frequency ω. In cases such as the finite element method, where nonorthogonal basis sets are used, one has

$$A' = i\,\Delta\omega\,C + A, \tag{3}$$

where C is not a unit matrix. Equation (3) can be recast into the form of Eq. (2) by first taking the Cholesky decomposition[4,5] of $C = LL^T$. Then

$$\tilde{A}' = i\,\Delta\omega\,\mathbb{1} + \tilde{A} \tag{4}$$

with

$$\tilde{A} = L^{-1}AL^{-T}, \tag{5}$$

and Eq. (1) becomes

$$\tilde{A}'|\tilde{u}\rangle = |\tilde{v}\rangle, \tag{6}$$

where

$$|\tilde{u}\rangle = L^T|u\rangle \tag{7}$$

and

$$|\tilde{v}\rangle = L^{-1}|v\rangle. \tag{8}$$

The spectral lineshape in ESR is given by[6,7]

$$I(\Delta\omega) = (1/\pi)\mathrm{Re}\langle v|u(\Delta\omega)\rangle. \tag{9a}$$

Here the usual Dirac notation for scalar product is used. It will be clear later that evaluation of Eq. (9a) can be done by using continued fractions and does not require eigenvalues of A. Note that even after Cholesky decomposition we have

$$\langle\tilde{v}|\tilde{u}\rangle = \langle v|u\rangle. \tag{9b}$$

Equation (9a) is a prototype for spectra and spectral densities in general, which are Fourier transforms of correlation functions. The methods described below yield powerful algorithms for computing correlation functions and spectral densities.

The more complicated two-dimensional ESR spectra can be calculated once the eigenvalues of A are known. For example, the signal in two-dimensional electron spin echo spectroscopy[6,7] is given by

$$S(\omega,\omega') \propto \sum_j c_j^2 \frac{T_{2j}}{1+\omega^2 T_{2,j}^2} \exp\left(-\frac{(\omega'-\omega_j)^2}{\Delta^2}\right), \tag{10a}$$

where, for the jth dynamic spin packet (i.e., the jth normal mode solution $|\psi_j\rangle$ to A corresponding to eigenvalue a_j), we have $T_{2j}^{-1} = \mathrm{Re}(a_j)$ as its Lorentzian width and $\omega_j = \mathrm{Im}(a_j)$ as its resonant frequency. The spectrum is inhomogeneously broadened (with respect to ω' sweep variable) by convolution with a Gaussian distribution of half width Δ. The weight factor is given by

$$c_j^2 = \langle \psi_j | v \rangle^2. \tag{10b}$$

In practical applications very often the dimension N of the matrix A becomes very large. This is the case in ESR for slow motion in oriented fluids (e.g., liquid crystals and model membranes). The usual diagonalization and inversion methods given in software packages such as EISPACK,[8] LINPACK,[9] and IMSL[10] become impractical once N becomes greater than about 200–500. The computer memory and time required become prohibitive. To overcome such difficulties the Lanczos algorithm (LA) has been used.[4–7,11–15] The LA produces a tridiagonal matrix [cf. Eq. (24) below] of dimension n_S, and this can be used to find eigenvalues of the original matrix, or quantities such as $I(\Delta\omega)$ can be obtained as a continued fraction [cf. Eq. (28) below] by using the elements of the tridiagonal matrix without even having to find the eigenvalues. Advantages of the LA in such problems are that (1) it is almost always the case that $n_S \ll N$, i.e., the LA projects out a reduced subspace of dimension n_S sufficient to represent the solution, and (2) the sparse matrix A is not modified by the algorithm, so only the nonzero elements need to be stored and utilized. The computer time required in the previous methods[8–10] usually increases as N^3 whereas in the LA (and the conjugate gradient method discussed below) it increases as[12] $n_S N n_E$, where n_E is the average number of nonzero matrix elements in a row of A.

The conjugate gradient method (CGM) of Hestenes and Stiefel[16] has been also used principally as a linear equation solver. It is known to mathematicians that the LA and CGM are in fact equivalent[4,5,16,17] (for a real symmetric positive definite matrix). In our recent ESR studies we have found that the CGM can be readily applied to complex symmetric matrices in general and that the equivalence between the LA and the CGM can be successfully exploited to use the CGM as a very powerful iterative technique to tridiagonalize matrices, which has significant advantages over the conventional LA. We like to refer to the approach we have taken of combining the advantages of these two algorithms as one of "turbo-charging" the Lanczos algorithm. Although our experience has been primarily with the ESR spectral calculations, we believe that this technique can have much wider application in various areas of science and engineering, e.g., for dissipative systems in general. It has been shown that many such cases can be formulated in a manner to produce complex symmetric matrices.[7,12] The purpose of this paper is to make the physical scientist aware of this technique. Greater details are given in Refs. 6 and 7.

There are various forms of the CG algorithm. The one we have used is given in the following. We first consider the CGM as a method of solving Eq. (1) directly.

One starts with a (complex) residual vector

$$|r_1\rangle = |v\rangle - A\,|u_1\rangle \tag{11}$$

and a conjugate vector

$$|p_1\rangle = |r_1\rangle, \tag{12}$$

with $|u_1\rangle$ being an initial guess for the solution (see, however, discussion below).

Successive approximants for $|r_k\rangle$ and $|p_k\rangle$ are obtained by

$$|r_{k+1}\rangle = |r_k\rangle - a_k A\,|p_k\rangle \tag{13}$$

and

$$|p_{k+1}\rangle = |r_{k+1}\rangle + b_k\,|p_k\rangle, \tag{14}$$

where the a_k and b_k are given by

$$a_k = \langle r_k | r_k \rangle / \langle p_k | A\,| p_k \rangle, \tag{15}$$

$$b_k = \langle r_{k+1} | r_{k+1} \rangle / \langle r_k | r_k \rangle. \tag{16}$$

Then

$$|u_{k+1}\rangle = |u_k\rangle + a_k\,|p_k\rangle \tag{17}$$

gives the $(k+1)$th approximant to the solution vector $|u\rangle$. At each step, the norm of $|r_k\rangle$ (defined in the following) gives a measure of the extent of convergence to the final solution $|u\rangle$.

As we have already mentioned, the original mathematical results and applications of both the Lanczos and the conjugate gradient algorithms involved real symmetric positive definite (or Hermitian) matrices. In our applications, the matrices are complex symmetric or else they can be transformed into complex symmetric forms by choosing an appropriate basis. For the general non-Hermitian (nonsymmetric) case, the algorithms can be justified by introducing a biorthonormal set of vectors x, x' such that

$$(x^i)^\dagger x_j = \delta_{ij}. \tag{18}$$

Then one can write down both Lanczos and conjugate gradient algorithms for a general complex matrix. Moro and Freed[12] have, however, shown that for the case of nondefective complex symmetric matrices it is possible to let

$$x^j = x_j^*. \tag{19}$$

With this, Eq. (18) becomes

$$x_j^{\mathrm{tr}} x_j = \delta_{ij}, \tag{20}$$

where tr stands for transposition. Thus both LA and CGM remain applicable by redefining the norms and the scalar products. One redefines the bra vectors without the usual complex conjugation in the Hilbert space. Now the norm (actually Euclidean pseudonorm)

$$\|r_k\|^2 = r_k^T r_k \tag{21}$$

becomes a complex quantity. All scalar products in the algorithm are defined in the same manner. Then the algorithm [Eqs. (11)–(17)] (and also the corresponding LA) can be used with complex quantities and it leads to convergence, as will be explained later. More mathematical details can be found in Refs. 6, 7, and 12.

Since Eq. (21) gives a complex quantity, we found it convenient to consider other norms for checking the convergence numerically:

$$r_{k,ps}^2 \equiv \left| \sum_j y_{k,j} \right|^2, \tag{22a}$$

$$r_{k,\mathrm{H}}^2 \equiv \sum_j |y_{k,j}|^2, \tag{22b}$$

$$r_{k,\mathrm{true}}^2 \equiv \sum_j |y_{\mathrm{true},k,j}|^2, \tag{22c}$$

where the $y_{k,j}$ are the components of $|r_k\rangle$ in the original basis set, and

$$|r_{k,\mathrm{true}}\rangle \equiv |v\rangle - A\,|u_k\rangle. \tag{23}$$

Any of the three norms can be used to estimate numerical convergence. The second and third norms are

equal in exact arithmetic and also were found to be equal in finite precision arithmetic to just about the limit of the double precision accuracy while Eq. (22a) was always smaller. The last result can be understood as a consequence of Holder's identity in complex analysis. Once the double precision limit is reached, any further attempt to improve the calculation by iteration is, of course, unsuccessful. Equations (22a) and (22b) are readily available during each iteration. Equation (22c) requires an extra matrix–vector multiplication. We have used $r^2 = r_{k\mathrm{H}}^2$ as a good criterion for convergence.

Availability of r^2 during iterations is a major advantage of CGM over the ordinary LA. Although r^2 can be obtained from the LA also, it requires, however, many more computations in addition to the basic algorithm. So in the LA one normally checks the spectrum with different values of n_S in an effort to verify convergence. Sometimes such comparisons are misleading, and one may be led to terminate the calculation before convergence is truly achieved. In the CGM one can readily terminate the program as soon as r^2 falls below some tolerance level, and this is readily automated. In fact, by monitoring r^2, the process can be seen to converge before one's eyes on the computer terminal.

In addition to giving control over the required number of recursive steps n_S, the equivalence between the CGM and the LA readily gives the Lanczos tridiagonal matrix, utilizing the quantities already generated during each iteration. The elements α_k, β_k of the tridiagonal Lanczos matrix

$$T_k = \begin{bmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \beta_3 & \\ & \beta_3 & \alpha_3 & \beta_4 \\ & & \beta_k & \alpha_k \end{bmatrix} \tag{24}$$

are given by

$$\alpha_k = \langle p_k | A | p_k \rangle / \rho_k^2 + (\rho_k^2/\rho_{k-1}^4)\langle p_{k-1} | A | p_{k-1} \rangle, \tag{25}$$

$$\beta_k = -(\rho_k/\rho_{k-1}^3)\langle p_{k-1} | A | p_{k-1} \rangle, \tag{26}$$

and

$$\rho_k \equiv \left(\sum_j y_{k,j}^2 \right)^{1/2}. \tag{27}$$

So after n_S iterations in the CGM, one has all the information one gets from the LA plus r^2 with virtually the same computation time. This is our "turbo-Lanczos" procedure. Although the algorithm will work with any random starting value of $|u_1\rangle$, one must start with $|r_1\rangle = |v\rangle$ corresponding to $|u_1\rangle = 0$, so that the first Lanczos vector $|\phi_1\rangle = |v\rangle$ (normalized to unity). This is required to establish the equivalence between the CGM and the LA. (The CGM with initial residual r_1 is equivalent to the LA with starting vector $r_1/\|r_1\|$. This choice is particularly useful, because we are interested in the projection of the final vector $|u\rangle$ on the starting vector $|v\rangle$ [cf. Eq. (9)]). The utilization of this physically relevant starting vector has the effect of biasing the Lanczos projections, generating Lanczos vectors, $|\phi_k\rangle \propto |r_k\rangle$ to approximate what we call the "optimal reduced space" to represent the physical problem. This guarantees very rapid convergence to the spectrum.

It should be emphasized that by convergence we mean numerical convergence for the physical problem at hand and not necessarily strict mathematical convergence. When the spectral quantity, which we calculate, does not change appreciably by increasing the number of iterative steps n_S or by decreasing the tolerance on the residual r^2 significantly, then the calculation is accepted as having converged. In the literature, mathematicians have discussed various problems that the Lanczos method runs into. These arise from round-off errors, which lead to loss of orthogonality of Lanczos vectors, spurious eigenvalues, and multiple copies of eigenvalues. There are also programs[5] available to extract "good" eigenvalues. The CGM would have similar problems. However, for the physical systems we have considered, such difficulties either do not arise or do not have a measurable effect on the computed physical observables such as the spectra given by Eq. (9a) or (10a). The spectra converge long before any problems arise. As we have already mentioned, this seems to us to be a consequence of our choosing the first Lanczos vector $|\phi_1\rangle = |v\rangle = |r_1\rangle$ (i.e., $|u_1\rangle = 0$). We recommend this choice for $|u_1\rangle$ which is based upon the physical nature of the problem instead of a random vector in spite of the fact that, in the strict mathematical sense, the algorithm should work for any random choice of $|u_1\rangle$. In fact, any spurious "eigenvectors" which may remain after spectral convergence has been achieved have negligible projections along $|v\rangle$ so [by Eq. (10b)] they will not influence the result.

We ran into only one problem with the CGM in comparison with the LA. If some diagonal elements of A are zero, a division by zero can occur in the CGM unlike the LA. A simple remedy is to add a small but finite real constant to the diagonal entries ("an intrinsic linewidth" in ESR calculations). Such an intrinsic linewidth is added anyway for physical reasons that have to do with inhomogenity of the magnetic field. But, in case it is not desired, it can simply be subtracted from all the eigenvalues after they are computed. Since we are calculating experimentally measurable quantities, we do not expect singular matrices! Any singularity must be a consequence of some unrealistic approximations. In any case we have checked that such a procedure is numerically stable and equivalence between the CGM and the LA is maintained.

Now it can be shown[7,12] that $I(\Delta\omega)$ [Eq. (9a)] can be obtained as a continued fraction from the elements of the tridiagonal matrix [Eq. (24)]:

$$I(\Delta\omega) = \frac{1}{\pi} \mathrm{Re} \left\{ \cfrac{1}{i\,\Delta\omega + \alpha_1 - \cfrac{\beta_2^2}{i\,\Delta\omega + \alpha_2 - \cfrac{\beta_3^2}{i\,\Delta\omega + \alpha_3 - \beta_4^2 \cdots}}} \right\}. \tag{28}$$

Thus to calculate $I(\Delta\omega)$, diagonalization of T_k is not necessary. For two-dimensional spectral quantities like $S(\omega,\omega')$ [Eq. (10a)], however, eigenvalues and eigenvectors (actually only components of the eigenvectors along $|\phi_1\rangle = |v\rangle$) are required. We use a version of the QR algorithm[4] which is useful for complex symmetric matrices and takes advantage of the bandedness.[18] Note that the dimensions of the matrix T_k are much smaller than those of the original matrix A.

Numerical details of our application of the CGM are given in Ref. 6. In one application we found that for a matrix with $N = 8196$ and 667965 nonzero elements, $r^2 = 10^{-2}$ and $r^2 = 10^{-10}$ required $n_S = 57$ and 143, respectively. The former value was entirely adequate for computing a continuous wave (CW) ESR spectrum $I(\Delta\omega)$. The 2-D ESR spectrum $S(\omega,\omega')$, which is more sensitive to the eigenvalues, required smaller values of r^2 and hence a larger n_S as expected. The computer times required depended on the computer used. However, both the LA and the CGM resulted in savings of orders of magnitudes in CPU times as compared to the previous methods,[8–10] even in the case when core memory was enough to use these latter methods. Very often (e.g., slow motional studies in ESR) core memory is generally not enough to employ the previous methods. In addition, in the direct methods[8–10] a large portion of the effort is expended in calculating irrelevant eigenvalues whose corresponding eigenvectors have negligible overlap with the starting vector $|v\rangle$ and hence they are very wasteful for our purpose.

We find that another important advantage of the CGM is that it very conveniently enables one to study schemes for basis set truncation. This is, in general, a difficult but important problem with very large basis sets. In ESR, when the motion becomes slow, a very large number of basis states (with "quantum numbers" such as L, K, M when using Wigner rotation functions) enter the calculations. However, not all the basis states contribute appreciably. This is true even if one can guess the maximum values of L, K, M needed for convergence, which is very often not the case. To estimate the contribution of each state, we found that one need only solve Eq. (1) for $|u(\Delta\omega)\rangle$ over the range of $\Delta\omega$ by utilizing Eqs. (11)–(17), since $\langle x_j|u\,\Delta\omega\rangle$ was shown[6] to be a measure of the importance of the jth basis vector $|x_j\rangle$. We could then eliminate the states which contributed less than some cutoff percentage (say 3% for CW ESR or 0.03% for 2-D ESR) to the value of $I(\Delta\omega)$ in the whole range. This "after the fact" elimination was found to be very useful for ESR, since in such studies one varies a number of parameters repeatedly to fit the experimental spectrum. Thus by establishing a minimal sufficient set one can save substantial time for subsequent calculations. Starting with an initial set, where one first makes an intelligent guess as to the maximum values of the quantum numbers, L, K, M, etc., we found a large reduction by a factor of at least 2 or 3 in the size of the basis set (sometimes even a factor of 10 for spectra in oriented media with lower symmetry). We suggest that such a truncation analysis may be of considerable use in other areas. In addition, we find that it is possible to first study cases where the convergence is more rapid (i.e., relatively small basis sets), in order to establish general truncation rules that can then be extrapolated to the more difficult cases involving slower convergence, hence enormous basis sets.

As we have mentioned before, our experience has been in connection with ESR. However, the mathematical structure of Eqs. (1)–(3) is very similar in different areas of physics. Therefore, we expect that the CGM (using the equivalence with Lanczos) will be found to be a surprisingly more powerful alternative to the usual application of Lanczos methods. Computer programs using the Lanczos and the conjugate gradient methods for simulating electron spin resonance spectra will be given in a forthcoming book[19] in the form of a diskette, which can be used on an IBM-PC. With some changes, these programs can be used on a mainframe computer or a supercomputer.

ACKNOWLEDGMENTS

We acknowledge Grants No. DMR8604200 and No. CHE8703014 from NSF, Grants No. GM25862 (J. H. F.) and No. GM40132 (K. V. V.), both from NIH, and the donors of the Petroleum Research Fund administered by the American Chemical Society (K. V. V.).

REFERENCES

1. Application of this well-known technique to ESR problems is described by J. H. Freed, in *Spin Labeling: Theory and Applications*, edited by L. J. Berliner (Academic, New York, 1976), Vol. I, Chap. 3.
2. For application of the finite difference method to magnetic resonance problems see, e.g., (a) J. H. Freed and J. B. Peterson, Adv. Magn. Reson. **8**, 1 (1976); (b) J. I. Kaplan and K. V. Vasavada, J. Magn. Reson. **52**, 475 (1983).
3. The finite element method is applied to ESR problems in (a) A. E. Stillman, G. P. Zientara, and J. H. Freed, J. Chem. Phys. **71**, 113 (1979); (b) K. V. Vasavada and J. H. Freed (unpublished results).
4. G. H. Golub and C. F. Van Loan, *Matrix Computations* (John Hopkins Univ., Baltimore, MD, 1983).
5. J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations* (Birkhauser, Boston, 1985), Vols. I and II. Volume II gives programs for Cholesky decomposition and Lanczos methods.
6. K. V. Vasavada, D. J. Schneider, and J. H. Freed, J. Chem. Phys. **86**, 647 (1987).
7. D. J. Schneider and J. H. Freed, in *Lasers, Molecules, and Methods, Advances in Chemical Physics*, Vol. LXXIII, edited by J. O. Hirshfelder, R. E. Wyatt, and R. D. Coalson (Wiley, New York, 1989), Chap. 10.
8. B. T. Smith, F. M. Boyle, J. J. Dongara, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, *Matrix Eigensystem Routines—EISPACK Guide* (Springer, New York, 1976).
9. J. T. Dongara, J. R. Bunch, C. B. Moler, and G. W. Stewart, LINPACK User's Guide (SIAM, Philadelphia, 1979).
10. IMSL MATH/LIBRARY (IMSL, Houston, 1987).
11. C. C. Paige and M. A. Saunders, SIAM J. Numer. Anal. **12**, 617 (1975); C. C. Paige, J. Inst. Math. Its Appl. **18**, 341 (1976); C. C. Paige, Linear Algebra Appl. **34**, 235 (1980).
12. (a) G. Moro and J. H. Freed, J. Chem. Phys. **74**, 3757 (1981); (b) G. Moro and J. H. Freed, in *Large Scale Eigenvalue Problems, Mathematical Studies Series*, Vol. 127, edited by J. Cullum and R. A. Willoughby (Elsevier, New York, 1986).
13. *The Recursion Method and Its Applications*, Springer Series on Solid-State Science, Vol. 58, edited by D. G. Pettifor and D. L. Weaire (Springer, New York, 1985).
14. D. S. Scott, in *Sparse Matrices and Their Uses*, edited by I. Duff (Academic, New York, 1981), p. 139.
15. B. N. Parlett, *The Symmetric Eigenvalue Problem* (Prentice-Hall, Englewood Cliffs, NJ, 1980).
16. M. R. Hestenes and E. Steifel, J. Res. Natl. Bur. Stand. **49**, 409 (1952).
17. D. B. Szyld and O. B. Widlund, in *Proceedings of the Third IMACS International Symposium* (Lehigh Univ., Bethlehem, PA, 1979), p. 167.
18. H. Rutishauser, Proc. Amer. Math. Soc. Symp. Appl. Math. **15**, 219 (1963); R. G. Gordon and T. Messenger, in *Electron Spin Relaxation in Liquids*, edited by L. T. Muus and P. W. Atkins (Plenum, New York, 1972), Chap. 13. Programs to diagonalize a complex symmetric tridiagonal matrix based on QL decomposition are given in Ref. 5.
19. D. J. Schneider and J. H. Freed, in *Spin Labeling: Theory and Applications*, edited by L. J. Berliner (Plenum, New York, in press), Vol. III.