Modification of the DIIS method for diagonalizing large matrices

1994 J. Phys. A: Math. Gen. 27 1199

(http://iopscience.iop.org/0305-4470/27/4/015)

J. Phys. A: Math. Gen. 27 (1994) 1199–1204. Printed in the UK

Modification of the DIIS method for diagonalizing large matrices

Hanchul Kim, Byung Deok Yu† and Jisoon Ihm

Department of Physics and Center for Theoretical Physics, Seoul National University, Seoul, 151-742, Korea

Received 14 May 1993

Abstract. The existing 'residual minimization/direct inversion in the iterative subspace' (DIIS) method for the iterative calculation of low-lying eigenstates of a large matrix is further developed and modified. The DIIS method, which uses the residual minimization criterion, may fail to provide correct low-lying eigenspectra in the case of ill-formed matrices, e.g. the momentum-space representation of Hamiltonian matrices of systems containing transition metal, rare earth, or first-row elements. We suggest the inclusion of another criterion—the vanishing of the overlap integral of an iterative eigenvector with already obtained low-lying eigenvectors in order to prevent the eigenvector from collapsing to lower states. Two numerical examples of the success of our modified DIIS method in contrast to the failure of the conventional DIIS method are presented.

In modern electronic structure calculations, the quantum mechanical Schrödinger equation has the form of the eigenvalue problem

$$H|a\rangle = \lambda|a\rangle \tag{1}$$

where H is an $N \times N$ Hermitian matrix. The present paper concerns an improvement of the iterative method of diagonalizing large Hermitian matrices proposed by Wood and Zunger [1]. We will restrict ourselves to the case of the plane-wave basis set [2] which is widely used because of its simplicity and versatility. The number of plane waves (N) required in the basis set is moderate when dealing with ordinary crystalline semiconductors or simple metals, but becomes huge when dealing with either a supercell structure (defects, superlattices, surfaces, lattice dynamics, etc) or a system where more or less localized states are involved (materials containing first-row atoms, transition metals, or rare-earth elements). In the former (supercell) case, the number of plane waves is proportional to the number of atoms in the large unit cell, and in the latter case, the deep non-local pseudopotentials necessitate short-wavelength plane waves. If one is to deal with a supercell geometry containing first-row atoms, transition metals or rare earth metals, the number of basis functions grows enormously. In the past, the eigenvalue problem was usually solved with use of the conventional procedure by Householder [3]. There are, however, two representative practical difficulties with this approach. First, the entire matrix H must be stored in the central memory. Second, the floating-point operation count required by this procedure scales as N^3 even if only a fraction of the eigenvalues and eigenvectors are desired in actual electronic structure calculations. As a result, this procedure is restricted

† Present address: Fundamental Research Laboratories, NEC Co., Tsukuba, Japan.

0305-4470/94/041199+06$19.50

to the order of 10^3 plane-wave basis functions. In response to these serious drawbacks in the usual Householder method, various alternative methods have been suggested and applied. They can be categorized into two groups: one is the direct minimization of the total energy functional in the Hilbert space of wavefunctions [4], and the other is the more conventional iterative diagonalization of the Hamiltonian matrix [5]. We concentrate on the latter category in this paper.

A number of iterative diagonalization methods have been developed over the last three decades, which have the twin virtues that only one row of the matrix H is required at a time and their floating-point operation count scales as N^2. (For pseudopotential plane-wave calculations, this can be reduced to $N \log_2 N$ if one uses fully separable pseudopotentials [6–9] and incorporates the fast Fourier transformation in the matrix-vector multiplication [10, 11].) Such iterative schemes have been used to compute the desired lower part of the eigenspectrum. These low-lying eigenstates are composed primarily of long-wavelength plane waves. Therefore, if the basis vectors are ordered by the wavenumber, an initial guess for the desired solution is obtained by diagonalizing an $N_0 \times N_0$ leading submatrix of H using the conventional Householder method. The Rayleigh–Ritz procedure [12] is commonly used in computing the optimal approximation to the eigensolutions of interest. This strategy has come to be known as the Davidson [13–15] and the Davidson–Liu (or block Davidson) [16] methods in the literature. On the other hand, Bendt and Zunger [17] and Wood and Zunger [1] have proposed an algorithm (the DIIS method—residual minimization/direct inversion in the iterative subspace [10, 11]) different from the previous ones. The main feature of the DIIS method lies in the use of the residual minimization criterion instead of the Ritz procedure in determining the optimal approximation to a desired solution. In the DIIS method, a correction vector $|\delta x^{(m)}\rangle$, where m represents the mth iteration, is generated with use of the modified Jacobi relaxation scheme [1, 10]. To compute the new approximate eigenvector $|x_{\text{new}}^{\text{DIIS}}\rangle = \sum_{k=0}^{m} c_k |\delta x^{(k)}\rangle$ of H, the expansion coefficients, c_k's, are chosen so that they can minimize the magnitude of the residual $\rho = \langle x_{\text{new}}^{\text{DIIS}} | (H - \lambda_{\text{old}} I) | x_{\text{new}}^{\text{DIIS}} \rangle / \langle x_{\text{new}}^{\text{DIIS}} | x_{\text{new}}^{\text{DIIS}} \rangle$. The requirement of minimal ρ leads to an expansion-space-projected $(m+1) \times (m+1)$ generalized Hermitian eigenvalue problem

$$\frac{\delta \rho^2}{\delta c_k^*} = 0 \longrightarrow P|c\rangle = \varepsilon Q|c\rangle \tag{2}$$

where $P_{kl} = \langle \delta x^{(k)} | H | \delta x^{(l)} \rangle$, $Q_{kl} = \langle \delta x^{(k)} | \delta x^{(l)} \rangle$ and $\varepsilon = \lambda_{\text{old}} + \rho$.

Now we want to point out that the DIIS method using only the residual minimization criterion may break down in providing correct eigenvalues and eigenvectors for ill-formed matrices. As long as an initial guess is reasonably close to the exact eigensolution (i.e. the chosen submatrix preserves the level structure of the full matrix for the desired part of eigensolutions), it converges to the exact eigensolution. The minimum dimension of the submatrix $N_{0,\min}$ preserving the level structure for the low-lying eigensolutions depends on the degree of diagonal dominance of the Hamiltonian matrix. We emphasize that the minimum submatrix dimension guaranteeing convergence to the correct eigensolutions can be too large to be solved directly by use of the Householder method for the class of matrices lacking diagonal dominance. This is the case for the Hamiltonian matrices of materials including first-row atoms, transition-metal oxides, and rare-earth compounds. In such situations, the aforementioned benefits of using the iterative diagonalization schemes reduce drastically. A more serious practical difficulty is that one does not have a good criterion for determining the minimum submatrix dimension applicable to general matrices. Noting such a possibility of failure, Martins and Troullier [11] proposed to use the DIIS

method in combination with the Ritz procedure. However, their improvement was not on the DIIS method but on the Ritz procedure, i.e. they improved the Ritz procedure by employing the DIIS relaxation steps to obtain optimally chosen correction vectors.

In contrast to their approach, we make a modification to the DIIS algorithm itself. We pay attention to the fact that a state is fully identified when both the eigenvalue and the eigenvector are specified. When the residual minimization is used as a criterion for new approximate eigensolutions in the DIIS scheme, it is implicitly assumed that the characteristics of input guesses have enough resemblance to their exact eigenvectors so that one can mainly concentrate on the eigenvalues. As is pointed out above, such an assumption can be false in dealing with ill-formed matrices and cannot be used with confidence in general. Therefore, in determining the approximate eigensolutions at each iteration, we examine the overlap integral of the present iterative eigenvector with the already found eigenvectors for the lower-lying levels. Discarding an eigenvector having a large overlap integral with previously obtained eigenvectors is an efficient method of preventing an iterative eigenvector from collapsing into one of the lower eigenvectors.

The actual algorithm is implemented as follows. Changes are made in the part of the original DIIS algorithm where the new approximate eigenvector $|x_{\text{new}}\rangle = \sum_{k=0}^{m} c_k |\delta x^{(k)}\rangle$ is determined. In the DIIS method, the set of expansion coefficients $\{c_0(\varepsilon), c_1(\varepsilon), \ldots, c_m(\varepsilon)\}$ minimizing $|\varepsilon - \lambda_{\text{old}}|$ is chosen. In the modified DIIS (MDIIS) method, we make an additional test on the overlap integral of the eigenvectors. If the overlap integral of the new approximate eigenvector with a normalized eigenvector corresponding to a lower eigenvalue is close to unity, it means that the iterative eigenvector has collapsed into the lower eigenvector. In that case, we take the eigenvector corresponding to the next lowest $|\varepsilon - \lambda_{\text{old}}|$ as the iterative eigenvector. This criterion is applied to all of the lower eigenvectors. The rest of the procedure is the same as that of the original DIIS algorithm. At this point, we want to caution the reader about a remaining problem. Though the inclusion of the overlap integral criterion prevents an eigensolution from collapsing to already obtained lower ones, it is still possible that the eigensolution erroneously jumps to a higher one causing a hole in the eigenvalue spectrum. Martins and Troullier [11] invoked the combination of Ritz procedure and the DIIS method to avoid this problem. We suggest that taking the total number of states to be calculated as $3 \sim 4$ times the number of the desired low-lying eigenstates may be safe for certain materials, but it still does not always guarantee correct results. The following numerical exercises show that the iterative solutions do converge to the exact eigenspectrum even for an ill-formed (i.e. far from diagonal dominance) matrix with the suggested modification of the DIIS method.

As a numerical test, we use a matrix of randomly generated elements in which diagonal elements are sorted in order of magnitude and off-diagonal elements are multiplied by a factor of r. The factor r is introduced as a control parameter of the degree of the diagonal dominance. In figure 1, we present the DIIS iterative evolution of the first five eigensolutions of a 70×70 matrix ($r = 0.1$) with a variable submatrix size N_0. When $N_0 = 15$, some eigenstates erroneously converge to wrong ones. Although correct convergence is achieved with use of a larger submatrix size ($N_0 = 45$), it is too large to claim merit in using an iterative diagonalization scheme. The result of the same calculation for $N_0 = 15$ with use of the suggested MDIIS scheme in figure 2 shows the correct convergence of eigenvalues. Although the addition of the overlap integral criterion slows down the convergence, such a cost is more than compensated for by the use of a much smaller submatrix. We have found that the minimum submatrix size $N_{0,\min}$ which gives the correct convergence in the above case is 42 and 11 for the DIIS and the MDIIS method, respectively.

We consider next the 3178×3178 Hamiltonian matrix (corresponding to the kinetic

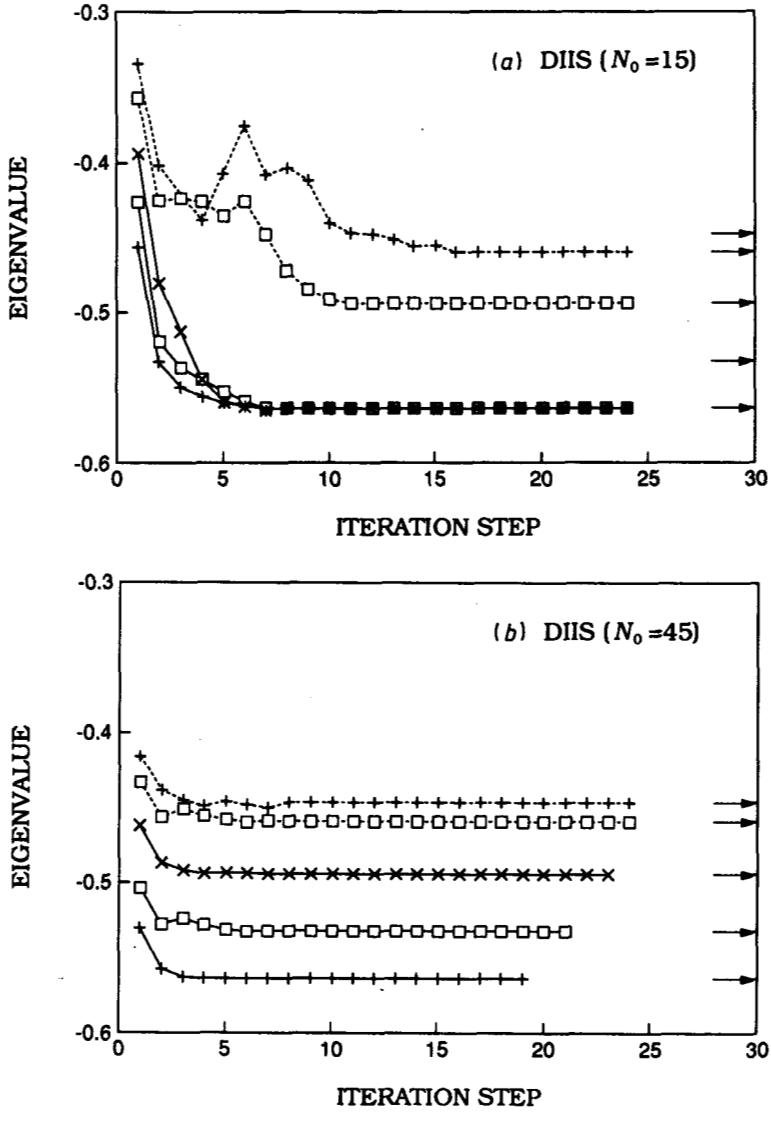

Figure 1. The DIIS iterative evolution. The arrows at the right-hand side represent the exact eigenvalues. (*a*) $N_0 = 15$, the second and third levels collapse to the first level, the fourth to the third, and the fifth to the fourth. (*b*) $N_0 = 45$, all the five lowest levels converge correctly.

energy cut off of 81 Ry) of cubic ZnS at the L point of the Brillouin zone. Matrix elements are computed from the self-consistent charge density in the plane-wave basis set. We find that the MDIIS iteration initialized by diagonalizing a leading submatrix of $\sim$ 500 yields the lowest 40 eigensolutions correctly apart from the level ordering. (One can easily rearrange the converged eigensolutions in ascending order of eigenvalues.) On the other hand, the DIIS scheme requires quite a large submatrix size (more than 1000) to obtain the correct eigenspectrum due to the localized nature of Zn 3d originated bands. In figure 3, the iterative evolution of some selected low-lying eigenvalues is shown with use of the initial guesses obtained by diagonalizing the 568 $\times$ 568 leading submatrix. The levels of the two highest

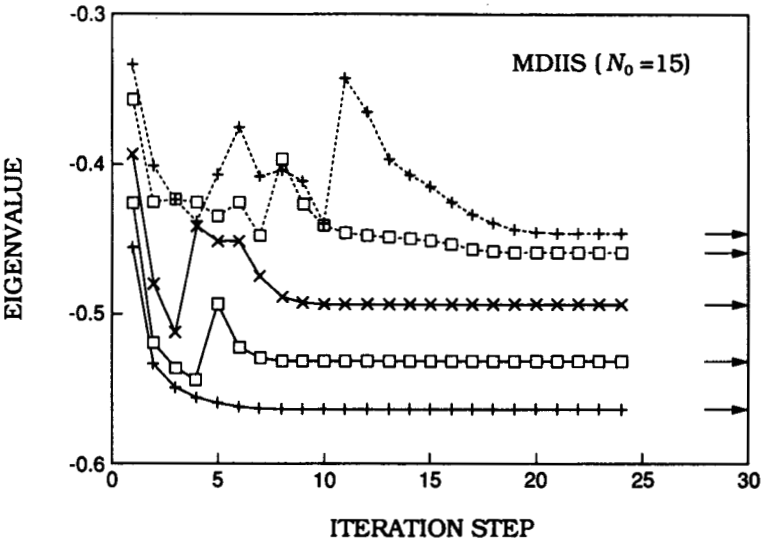

Figure 2. The MDIIS iterative evolution. For $N_0 = 15$, the levels converge correctly to the desired eigenspectrum.

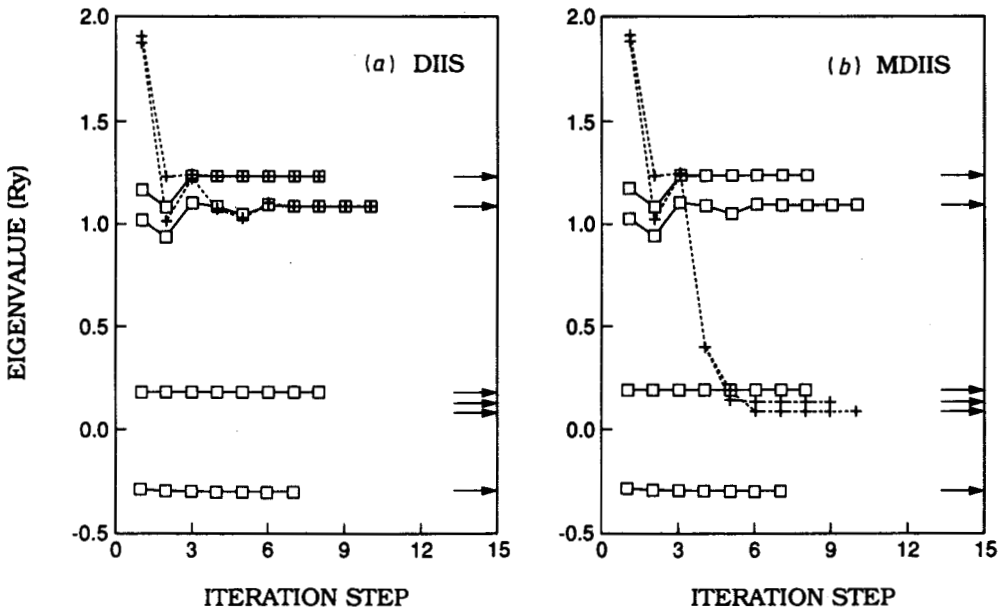

Figure 3. The iterative evolution of some representative low-lying levels of the cubic ZnS at the L point of the Brillouin zone. (*a*) The DIIS scheme; the erroneously evolving levels marked by '+' have the character of the Zn 3d orbitals. (*b*) the MDIIS scheme; two Zn 3d originated states converge correctly.

guessed eigenvalues originating from the Zn 3d orbitals converge incorrectly with use of the DIIS algorithm in figure 3(*a*), but these two levels start to deviate significantly from initial guesses in the first few MDIIS iterations and converge to exact levels in the end as

shown in figure 3(*b*). Such a substantial deviation from the initial guess confirms the fact that high-wavenumber plane waves are required to describe localized states correctly in the plane-wave basis set and, consequently, a small-sized submatrix is not capable of preserving the low-lying level structure of the full matrix. The added overlap integral criterion forces such unsatisfactorily provided initial guesses to converge to correct results.

In summary, we have pointed out that the DIIS method can fail to provide the correct eigensolutions from an initial guess obtained by directly diagonalizing an $N_0 \times N_0$ leading sub-block of H when H has no fully diagonal dominance. We suggest a modified DIIS scheme where the overlap of the present iterative eigenvector with the lower eigenvectors is to be examined to avoid the collapse into one of the lower eigenvectors. Inheriting all the merits of the DIIS method, our modified version has shown to be applicable to those matrices in which the DIIS method fails.

This work is supported by the Ministry of Science and Technology, the KOSEF-NSF Cooperative Research Program, and the Korea Science and Engineering Foundation through the SRC Program.

References

[1] Wood D M and Zunger A 1985 *J. Phys. A: Math. Gen.* **18** 1343
[2] Ihm J, Zunger A and Cohen M L 1979 *J. Phys. C: Solid State Phys.* **12** 4409
[3] Wilkinson J H 1965 *The Algebraic Eigenvalue Problem* (Oxford: Clarendon) ch 4, p 229; ch 5, p 290
[4] Teter M P, Payne M C and Allan D C 1989 *Phys. Rev.* B **40** 12255
Payne M C, Teter M P, Allan D C, Arias T A and Joannopoulos J D 1992 *Rev. Mod. Phys.* **64** 1045
[5] National Resource for Computation in Chemistry 1978 *Workshop Report, Numerical Algorithms in Chemistry: Algebraic Methods (National Resource for Computation in Chemistry, Lawrence Berkeley Laboratory)* LBL-8158, UC-32, CONF-780878
[6] Kleinman L and Bylander D M 1982 *Phys. Rev. Lett.* **48** 1425
Gonze X, Kräckell P and Scheffler M 1990 *Phys. Rev.* B **41** 12264
Gonze X, Stumpf R and Scheffler M 1991 *Phys. Rev.* B **44** 8503
[7] Blöchl P E 1990 *Phys. Rev.* B **41** 5414
Vanderbilt D 1990 *Phys. Rev.* B **41** 7892
Chou M Y 1992 *Phys. Rev.* B **45** 11465
[8] Saito M, Sugino O and Oshiyama A 1992 *Phys. Rev.* B **46** 2606
[9] Bylander D M and Kleinman L 1992 *Phys. Rev.* B **46** 9837
[10] Martins J L and Cohen M L 1988 *Phys. Rev.* B **37** 6134
[11] Martins J L and Troullier N 1991 *Phys. Rev.* B **43** 2213
[12] Wilkinson J W and Reinsch C 1971 *Linear Algebra* (Berlin: Springer) p 284
[13] Davidson E R 1975 *J. Comput. Phys.* **17** 87
[14] Davidson E R 1980 *J. Phys. A: Math. Gen.* **13** L179
[15] Davidson E R 1978 *Workshop Report, Numerical Algorithms in Chemistry: Algebraic Methods (National Resource for Computation in Chemistry, Lawrence Berkeley Laboratory)* LBL-8158, UC-32, CONF-780878, p 15
[16] Liu B 1978 *Workshop Report, Numerical Algorithms in Chemistry: Algebraic Methods (National Resource for Computation in Chemistry, Lawrence Berkeley Laboratory)* LBL-8158, UC-32, CONF-780878, p 49
[17] Bendt P B and Zunger A 1982 *Solar Energy Research Institute Technical Report* TP-212-1698, unpublished