Conf-740110--2

ON THE EFFICIENT SOLUTION OF THE RELATIVISTIC
APW SECULAR EQUATION FOR BOTH EIGENVALUES AND EIGENVECTORS

by

D. D. Koelling

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

Presented at:

International Symposium on Atomic, Molecular, Solid State Theory,
and Quantum Statistics, Sanibel Island, Florida, January 20-26, 1974

J of C-AUA-USAEC

MASTER

ARGONNE NATIONAL LABORATORY, ARGONNE, ILLINOIS

fig

On the Efficient Solution of the Relativistic APW
Secular Equation for Both Eigenvalues and Eigenvectors*

D. D. Koelling

Argonne National Laboratory, Argonne, Illinois 60439

ABSTRACT

The special problems introduced into the solution of the secular equation by the inclusion of relativistic effects are discussed. A method based on the use of linear combination of RAPW's is then described for solving the secular equation. This method greatly reduces the computer time and core space required to solve the secular problem and is particularly advantageous in large systems.

*Work performed under the auspices of the U. S. Atomic Energy Commission.

Before describing an elaborate technique for dealing with the secular equation resulting from the relativistic augmented plane wave (RAPW) variational problem, it is appropriate to explain that there is, in fact, a problem worthy of consideration. After all, techniques for the solution of the energy-dependent non-relativistic secular equations have been in use for quite some time.¹ At first glance, it would appear that one need only convert from real matrices to complex ones (because of the spin-orbit coupling) and double the matrix size. This would result in an increase in computer core storage required by a factor of eight and computation time by a factor of 32. (Large matrix limit used.) This not only greatly increases the cost of the calculation but it adds greatly to the seriousness of the existing numerical instabilities. These instabilities do not greatly affect the eigenvalues but can cause substantial errors in the resulting eigenvectors--which are required in all applications of interest to us.

In this paper, the various techniques used to solve the non-relativistic secular equation are first reviewed and their sensitivities and limitations discussed. Then, it is shown how the self-consistent diagonalization approach can be effectively used for the solution of the relativistic secular equation by first solving a pseudo-nonrelativistic problem.

I. Solution of the Non-Relativistic Secular Equation

There are a number of techniques used to solve the non-relativistic secular equation to obtain eigenvectors. Two¹ involve first finding the eigenvalues by determinant plotting and the other two involve iteration with a diagonalization procedure.^{2,3} These procedures are used because the basis functions are energy dependent and thus the Hamiltonian H and overlap matrix S involved are a non-linear variational problem.

1. GAUSSIAN ELIMINATION--BACK SUBSTITUTION. In this procedure one finds the determinant by Gaussian elimination. This determinant is then plotted as a function of trial energy until a zero is found representing an eigenvalue. The triangular matrix remaining after the Gaussian elimination for this energy can then be solved as a set of linear equations by back substitution. This technique is very fast but suffers from considerable instability--especially when there is degeneracy or near-degeneracy. In the relativistic case where the crystal has inversion symmetry, the Kramers degeneracy implies that all eigenvalues will be doubly degenerate. Furthermore, as one includes more APW's to obtain the convergence required for good wavefunctions, the near linear dependence of the APW basis set will introduce more approximate roots⁴ (one for each atom in the unit cell is usually observed).

2. DETERMINANT PLOT--DIAGONALIZATION. The instability of the back substitution can be eliminated by replacing it with a diagonalization package used on the original secular equation.

This requires again setting up the secular equation after the eigenvalue is found, diagonalizing it, and then renormalizing the resultant eigenvector associated with the zero eigenvalue (since the APW's were not orthonormal). As originally implemented, this technique was used with a Jacobi diagonalization routine which is slow compared to other diagonalization techniques and required that one also get the irrelevant eigenvectors as well as the one desired. This has been the major criticism of the procedure.³ However, that difficulty is easily overcome by using a Householder-Givens diagonalization routine with the bounds on the Sturm sequence routine set quite close to zero. Then the extra diagonalization will cost roughly the same as one additional determinant evaluation.⁵

3. SELF-CONSISTENT DIAGONALIZATION. In this procedure, the secular equation is set up in an actual linear eigenvalue problem format:

$$H(E_0) C(E_0) = E S(E_0) C(E_0)$$

where $H(E_0)$ and $S(E_0)$ are the Hamiltonian and overlap matrices. The dependence on energy results from using APW's defined using an energy E_0 in the solution of the radial equation within the muffin-tin spheres.⁶ This secular equation can be solved by conventional techniques consisting of a Choleski decomposition of S and transformation of H followed by a Householder-Givens diagonalization of the transformed Hamiltonian. The eigenvectors of the transformed matrix are then back transformed to get the eigenvectors of the original secular equation. The eigenvalue is determined by the self consistency requirement.

$$E(E_0) = E_0$$

Because E_0 is a variational parameter (albeit a non-linear one), we know that $E(E_0)$ is greater than the actual eigenvalue and that $dE(E_0)/dE_0 = 0$ at $E = E_0$. Plots of $E(E_0)$ are shown in Ref. 2 and one such plot is reproduced in Fig. 1 (The AAPW method² has been chosen for the construction of these plots as a matter of convenience. Such a plot would be almost identical for the standard APW method except for the encircled region. The spurious roots in the encircled region of Fig. 1 result from a numerical instability problem in the AAPW which allows the beginning of an orthogonalization catastrophe to occur unless extremely high precision is used.) It is useful to note that $E(E_0)$ is relatively insensitive to the eigenvector. Thus, a better energy to use for the next iteration is

$$\tilde{E}_n = E(E_{n-1})/N(E_{n-1})$$

$$E(E_{n-1}) \equiv C^\dagger(E_{n-1})H(E(E_{n-1}))C(E_{n-1})$$

$$N(E_{n-1}) \equiv C^\dagger(E_{n-1})S(E(E_{n-1}))C(E_{n-1})$$

This quantity is shown in Fig. 2. One notes that $\tilde{E}(E_0)$ is a much more slowly varying function of E_0 . In fact, the plane-wave states (the highest 4 eigenvalues) are extremely flat once one has gotten beyond the point where the principal maximum of the p-wave is pulled inside the muffin-tin sphere. The d-states (the lowest 5 eigenvalues) do show more variation. This can be improved with a repeated application of the process.¹⁰

3b. SELF CONSISTENT DIAGONALIZATION WITHOUT ORTHONORMALIZATION.

The question arises as to why one bothers to first orthonormalize (i.e., perform the Choleski decomposition). One could, instead, use a procedure very similar to the second type of approach: One would diagonalize $(H(E_0) - E_0 S(E_0))$ and then renormalize. The resulting eigenvalues after renormalization would be the energy correction required. Several problems do arise, however, when implementing such a procedure because of the near linear dependent of the APW basis set. The eigenvalue of $(H - E_0 S)$ are $(E - E_0)\langle S \rangle$ and not $(E - E_0)$. Thus there will be one or more small eigenvalues resulting from $\langle S \rangle$ being small. In practice, we find that there is usually one such eigenvalue for fcc and bcc crystals and two for hcp crystals. If one diagonalizes S for an APW set large enough to reach acceptable convergence, one finds that its eigenvalues range over 3-7 orders of magnitude with the one or two eigenvalues representing the near linear dependence another 1-2 orders of magnitude smaller still. Unfortunately, these eigenvalues resulting from the near linear dependence must be found at each iteration and it is in this manipulation of a the tridiagonal matrix that much of the computer time is expended. Then the normalization integral must be calculated before one can identify and eliminate these unwanted solutions. Further, even if one wants only to find one particular band, one must renormalize all bands to obtain proper band orderings.

II. Solution of the Relativistic Secular Equation

Historically, our first attempt to solve the secular equation resulting from the use of RAPW's was a straightforward applications of method number 3 above. But, when one attempts to apply the same techniques to the equations resulting from the relativistic APW technique, one finds that it is almost impossible to make it work. The major difficulty arises in attempting to perform the Choleski decomposition. This decomposition is only applicable to positive definite matrices--which the overlap matrix S most certainly is. However, if one has the very small near-linear-dependence eigenvalues which now appear twice because of the Kramers degeneracy (we assume that the crystal has inversion symmetry so that the Kramers degeneracy is equivalent to a spin-flip degeneracy), it is possible for one to lose the positive definite property of S . This is caused by the round-off errors resulting from the use of a finite precision computer. The situation is made worse by the doubling of the matrix size and by making it complex. Thus it has proven to be virtually impossible to do a direct Choleski decomposition. One can, of course, orthonormalize by first diagonalizing S but this is far too expensive as it requires obtaining all eigenvalues and eigenvectors of S .

The technique currently being used is based on a Linear Combination of RAPW's (LCRAPW) approach¹¹ where one first constructs a smaller set of basis functions using the RAPW's and then uses these in a variational calculation (with a smaller secular equation). The construction of the basis functions is based on the observation that the RAPW Hamiltonian and overlap matrices can be written^{12,13} as a sum of a "non-relativistic" matrix plus a spin-orbit matrix

$$H_{(\text{RAPW})} = H_{\text{NR}} + H_{\text{SO}}$$

$$S_{(\text{RAPW})} = S_{\text{NR}} + S_{\text{SO}}$$

where the "non-relativistic" matrix element would be exactly what remains in the non-relativistic limit. In actuality, the H_{NR} and S_{NR} terms contain the non-spin dependent relativistic effects, are identical for the "spin-up" and "spin-down", and do not mix the two spin states. (These spin labels are appropriate to the interstitial regions where no spin orbit coupling is included--but there is spin mixing in these basis states inside the muffin-tin spheres by the spin-orbit interactions.) Selecting one spin only, one solves by Choleski decomposition techniques the real secular equation of half the size

$$H_{\text{NR}} b_n = e_n S_{\text{NR}} b_c$$

retaining only those solutions for which $e_n \leq e_{\text{max}} \sim 3 \text{ Ry}$. These vectors are then used to form a set of spin-up and spin-down basis vectors which are used with the complete matrix elements.

This is merely the application of linear combinations of RAPW's as a basis set in a restricted variational calculation. The restriction in the variation is determined by how many eigenvectors are retained as a result of the energy cutoff at $e_{\max}$. The energy used has been found to be more than adequate and has not been optimized. It is motivated by the PbTe perturbation calculations.¹⁴

The resulting equations retain the same symmetry properties that are responsible for the Kramers degeneracy in the RAPW secular equation. Thus it is possible to exploit this symmetry much as the Soven¹⁵ technique is used in the evaluation of determinants. If the spin-up and spin-down states are ordered in the same manner but kept separate, the resulting secular equation has the block form

$$\begin{pmatrix} H_d & H_o \\ H_o^\dagger & H_d^* \end{pmatrix} \begin{pmatrix} C_{(up)} \\ C_{(down)} \end{pmatrix} = E \begin{pmatrix} S_d & S_o \\ S_o^\dagger & S_d^* \end{pmatrix} \begin{pmatrix} C_{(up)} \\ C_{(down)} \end{pmatrix}$$

where

$$\begin{aligned} H_d^\dagger &= H_d & S_d^\dagger &= S_d \\ H_o^\dagger &= -H_o & S_o^\dagger &= -S_o \end{aligned}$$

We now describe the series of transformations used to solve this secular equation. The first transformation is based on a Choleski-type decomposition of

$$S_d = T^\dagger T$$

where T is upper triangular. The inverse of T is easily constructed (in fact, is normally constructed directly). We first apply the transformation

$$\begin{pmatrix} T^{-1} & 0 \\ 0 & T^{-1*} \end{pmatrix}$$

to get

$$\begin{pmatrix} H'_d & H'_a \\ H'^{\dagger}_o & H'^{*}_d \end{pmatrix} \begin{pmatrix} C'_{(up)} \\ C'_{(down)} \end{pmatrix} = E \begin{pmatrix} 1 & S'_o & C'_{(up)} \\ S'^{\dagger}_o & 1 & C'_{(down)} \end{pmatrix}$$

$$H'_d = T^{-1\dagger} H_d T^{-1}$$

$$H'_o = T^{-1\dagger} H_o T^{-1*}$$

$$S'_o = T^{-1\dagger} S_o T^{-1*}$$

$$C'_{(up)} = T C_{(up)}$$

$$C'_{(down)} = T^* C_{(down)}$$

Note that one need only construct H'_d , H'_o , and S'_o with the other blocks determined by symmetry. The next transformation is of the form

$$\begin{pmatrix} 1 & -S'_o \\ 0 & 1 \end{pmatrix}$$

and is not a unitary transformation. It yields

$$\begin{pmatrix} H'_d & H''_o \\ H''^{\dagger}_o & H''_d \end{pmatrix} \begin{pmatrix} C_1 \\ C'_{down} \end{pmatrix} = E \begin{pmatrix} 1 & 0 & C'_1 \\ 0 & 1 - S'^{\dagger}_o S'_o & C'_{down} \end{pmatrix}$$

$$H''_O = H'_O - H'_d S'_O$$

$$H''_d = H'^*_d - H'^{\dagger}_O H'_O + S'^{\dagger}_O H'_d S'_O$$

$$C_1 = C'_{(\text{up})} + S'_O C'_{(\text{down})}$$

where we have used the fact that the inverse of the transformation matrix is

$$\begin{pmatrix} 1 & S'_O \\ 0 & 1 \end{pmatrix}$$

The final transformation is of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & U \end{pmatrix}$$

where U is another Choleski-type transformation

$$U^{\dagger}(1 - S'^{\dagger}_O S'_O)U = 1$$

yielding

$$\begin{pmatrix} H'_d & H''_O U \\ U^{\dagger} H''_O & U^{\dagger} H''_d U \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = E \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$$

$$C_2 = U^{-1} C'_{(\text{down})}$$

from which the eigenvalues can be obtained by diagonalization of the transformed Hamiltonian. The eigenvectors of this transformed

secular equation must then be back-transformed. This is easily done yielding

$$C_{(\text{up})} = T^{-1}C_1 - S_0'T^{-1*}C_2$$

$$C_{(\text{down})} = UT^*C_2$$

There are a number of advantages to this method of handling the secular equation as many of the transformations are done on the blocks which are half the size of the original matrix and these transformations are used elsewhere by the symmetry relations. In addition, the second transformation involves no decompositions or inverses at all. Unfortunately it does make necessary the third transformation. Were one to find the unitary transformation at this second step, one could retain the symmetry of the original matrix and thus also eliminate the calculation of the lower right hand block at this stage. Unfortunately, the construction of this unitary transformation involves far more effort than the procedure invoked.

The eigenvalues resulting from this technique are easily checked by comparison to the results of the more standard determinant plotting procedures. This has been done for bcc molybdenum and fcc thorium with agreement to better than 0.001 Rydberg (the tolerance normally specified in the determinant plotting method). Thus, the truncation energy $e_{\text{max}} = 3 \text{ Ry}$ is more than adequate. As mentioned earlier, no systematic tests have been made to attempt to find a lower bound. This energy cutoff results in keeping roughly 20 eigenvectors from the first "pseudo-non-relativistic" calculation. Thus the LCRAPW secular

equation is still fairly large. For a more complicated system such as a compound, it may well be necessary to investigate the quality of the solutions as a function of $e_{\max}$ in order to reduce the size of the secular equation.

The eigenvectors resulting from this technique are more difficult to test. The only tests we have made thus far are to calculate the bilinear products

$$E = C^\dagger H_{(\text{RAPW})} C$$

$$N = C^\dagger S_{(\text{RAPW})} C$$

and verify that these quantities give back the eigenvalue and the normalization correctly. This will be seen to be more than just a check on the computer programming as we are performing the full calculation here with the results from a truncated secular equation. In addition, we have found that the derivatives calculated by Hellman-Feynman techniques

$$\nabla_k E = C^\dagger (\nabla_k H_{(\text{RAPW})}) C$$

are very accurate. Unlike the multiple determinant scheme,¹⁶ this procedure is advantageous for tracing orbits in reciprocal space.

As for efficiency, we have only one other usable technique to compare with. Primarily for ferromagnetic systems (where the Kramers degeneracy is lost), we have implemented a generalized version of the non-relativistic method 3b above. Although the two procedures have not been run against each other in a controlled manner, the procedure described here runs roughly twice as fast--and no

optimizing has been performed on it (i.e., reduction of $e_{\max}$). Furthermore, this procedure is capable of running in such a manner as to use only slightly more computer core storage than the non-relativistic case. This can be accomplished either by creating the spin-orbit related terms as needed or by storing them on auxiliary devices until needed. The charging algorithm used at our installation gives no economic advantage to such procedures so we have not done so. However, the programs have been set up such that switching to the use of auxiliary devices is a minor modification and test runs have been made in this mode.¹⁷

The procedure described here becomes ever more advantageous as the number of basis functions required is increased. We are now implementing this technique for use with binary intermetallic compounds where the number of basis functions required is a serious problem. We are quite optimistic that this method for handling the secular equation will greatly reduce the limitations previously experienced in such systems.

ACKNOWLEDGMENTS

The author is greatly indebted to his coworker B. N. Harmon for innumerable and very fruitful discussions on this problem.

REFERENCES

1. L. F. Mattheiss, J. H. Wood, and A. C. Switendick, Methods in Computational Physics, Vol. 8, B. Adler, S. Fernbach, and M. Rotenberg, Editors (Academic Press, New York, 1968).
 2. D. D. Koelling, Phys. Rev. B 2, 290 (1970).
 3. B. N. Harmon, D. D. Koelling, and A. J. Freeman, J. Phys. C 6, 2294 (1973).
 4. This linear dependence is easily understood. A set of plane waves is a complete set over the space of the crystal. When used in forming APW's, however, this set is used only in the interstitial region and is thus overcomplete.
 5. In fact, additional manipulations could be eliminated if one were to use the Householder transformation to tridiagonal form at each step and obtain the determinants from the tridiagonal matrix. This would have quite comparable speed to the Gaussian elimination technique and is quite stable.
 6. Only the functional dependence of the various quantities will be shown here. The actual form of the matrices involved will not be written down as is not needed for this discussion and it will depend on the type of augmentation used. See Refs. 7-9.
 7. J. C. Slater, Phys. Rev. 51, 846 (1937); T. L. Loucks, The Augmented Plane Wave Method (W. A. Benjamin Inc., New York, 1967).
 8. P. M. Marcus, Intern. J. Quantum Chem. 1, 576 (1967).
 9. D. D. Koelling, J. Phys. Chem. Solids 33, 1335 (1972).
- In this paper, the possibility of constructing a basis

function by matching the derivative of the ℓ partial wave when a node lies too near the sphere boundary. The reader should be aware that when this is done, the term $(2\ell + 1)P_\ell(\hat{k}_n \cdot \hat{k}_m)d(j_\ell(k_n R)j_\ell(k_m R))/dr$ must be added to the matrix element in addition to using the modified term for that partial wave as described there.

10. B. N. Harmon and D. D. Koelling (to be published).
11. G. O. Arbman and D. D. Koelling, *Physica Scripta* 5, 273 (1972).
12. J. Dimmock, private communication.
13. D. D. Koelling, *Phys. Rev.* 188, 1049 (1969); O. K. Andersen, *Phys. Rev. B* 2, 883 (1970).
14. J. B. Conklin, Jr., L. E. Johnson, and T. W. Pratt, Jr., *Phys. Rev.* 137, A1292 (1965).
15. P. Soven, *Phys. Rev.* 137, 170 (1965).
16. J. H. Wood, Computational Methods in Band Theory, P. M. Marcus, J. Janak, and A. R. Williams, Editors (Plenum Press, New York, 1971), p. 59.
17. It is the only way the programs could be operated on the CDC 6400 at Northwestern University.

FIGURE CAPTIONS

Fig. 1 The Eigenvalue Spectrum as a Function of Trial Energy E_0 for the Point W of the Chodorow Potential in copper. The self-consistent solution occurs where the eigenvalues cross the 45 deg. line. Doubly degenerate eigenvalues have been shown by using a heavier line. The encircled region is a spurious root problem of the AAPW method as discussed in the text.

Fig. 2 The Refined Eigenvalue Spectrum as a Function of Trial Energy for the Point W of the Chodorow Potential in copper. Note the greater insensitivity of $\tilde{E}$ (defined in text) to the trial energy E_0 .