SIAM REVIEW
Vol. 21, No. 3, July 1979

0036-1445/79/2103-0003$01.00/0

INVERSE ITERATION, ILL-CONDITIONED EQUATIONS AND NEWTON'S METHOD*

G. PETERS† AND J. H. WILKINSON‡

Abstract. Inverse iteration is one of the most widely used algorithms in practical linear algebra but an understanding of its main numerical properties has developed piecemeal over the last thirty years: a major source of misunderstanding is that it requires the solution of very ill-conditioned systems of equations. We describe closely related algorithms which avoid these ill-conditioned systems and explain why the standard inverse iteration algorithm may nevertheless be preferable. The discussion covers the generalized problems $Ax-\lambda Bx$ and $(A_r\lambda^r+\cdots+A_1\lambda+A_0)x=0$ in addition to the standard problem. The case when $A-\lambda I$ is almost singular to the working accuracy has only recently been understood, mainly through the work of Varah. The final sections give a detailed account of the current state of this work, concluding with an analysis based on the singular value decomposition.

1. Introduction. Inverse iteration is now the most widely used method for computing eigenvectors corresponding to selected eigenvalues which have already been computed more or less accurately. The original motivation for this algorithm was the following. Let A be a matrix with eigenvalues λ_i and a complete set of eigenvectors u_i which we shall assume for the moment to be normalized so that

$$\|u_i\|_2=1. \tag{1.1}$$

If λ is a good approximation to some eigenvalue λ_s, consider the sequence of vectors x_i derived from the relations

$$(A-\lambda I)x_p=k_p x_{p-1}, \tag{1.2}$$

where x_0 may be almost arbitrary and k_p is the scale factor chosen so that x_p is normalized. If

$$x_0=\sum_1^n \alpha_i u_i \tag{1.3}$$

then

$$x_p=(k_1k_2\cdots k_p)\sum_1^n \alpha_i u_i/(\lambda_i-\lambda)^p \tag{1.4}$$

and if $|\lambda_s-\lambda|\ll|\lambda_i-\lambda|(i\neq s)$ and $\alpha_s\neq 0$ the component of u_s becomes increasingly dominant. In fact if λ is *very* accurate, correct, say, to working accuracy, x_1 will probably already be an accurate eigenvector, possibly correct almost to working accuracy. Inverse "iteration" is widely used in this form, i.e. with only one iteration.

Inverse iteration is usually attributed to Wielandt (1944) though a number of people seem to have had the idea independently. Although basically a simple concept its numerical properties have not been widely understood. If λ really is very close to an eigenvalue, the matrix $(A-\lambda I)$ is almost singular and hence a typical step in the iteration involves the solution of a very ill-conditioned set of equations. Indeed if λ is an exact eigenvalue $A-\lambda I$ is exactly singular, though in numerical work involving rounding errors the distinction between "exact" singularity and "near" singularity is

* Received by the editors July 8, 1976, and in final revised form September 6, 1978. This work was supported in part by National Aeronautics and Space Administration Grant NSG 1443.

† Division of Numerical Analysis and Computer Science, National Physical Laboratory, Teddington, Middlesex, England.

‡ Division of Numerical Analysis and Computer Science, National Physical Laboratory, Teddington, Middlesex, England and Computer Science Department, Stanford University, Stanford, California 94305.

scarcely significant. The period when inverse iteration was first considered was notable for exaggerated fears concerning the instability of direct methods for solving linear systems and *ill-conditioned* systems were a source of particular anxiety. For this reason it was widely held to be inadvisable to use a λ which was too accurate; it was thought that such an eigenvalue should be debased a little so that the resulting matrix $(A-\lambda I)$ would not be too ill-conditioned. Although it is now generally recognized that this is not necessary, and indeed is not to be recommended, few numerical analysts discuss inverse iteration with any confidence.

In this paper we expose the relationship between inverse iteration and two other algorithms in a way which we hope will remove some of the mystery.

2. Ill-conditioned systems of equations. We consider first the solution of an $n \times n$ system of linear equations

$$Bx = b. \tag{2.1}$$

It is now generally accepted that if one solves such a system by a direct method on a computer, then provided the matrix B is not so special that rounding errors do not occur, one cannot, in general, do better than to obtain the exact solution x_c of some perturbed system

$$(B+E)x_c = b, \tag{2.2}$$

where

$$\|E\| \leqq f(n)\|B\| \text{ macheps}, \tag{2.3}$$

macheps being the computer precision, and $f(n)$ a modest function of n. For large systems, using Gaussian elimination with partial pivoting, or a QR factorization of B it is rare for $f(n)$ to exceed n in practice. Hence if

$$Bx_e = b, \tag{2.4}$$

i.e. x_e is the exact solution of the given system,

$$x_c = (I+B^{-1}E)^{-1}x_e, \tag{2.5}$$

giving

$$\begin{aligned} \|x_c - x_e\|/\|x_e\| &\leqq \|B^{-1}E\|/(1-\|B^{-1}E\|) \\ &\leqq \alpha/(1-\alpha), \quad (\text{assuming } \alpha < 1), \end{aligned} \tag{2.6}$$

where

$$\alpha = f(n)\|B\|\|B^{-1}\| \text{ macheps}. \tag{2.7}$$

If $\alpha \doteq 2^{-m}$ then x_c will have roughly m binary figures correct in the norm. As $\alpha \to 1$ the accuracy gradually deteriorates until when $\alpha \geqq 1$ we can expect no correct significant figures. Although this last statement is perfectly true it conceals a very useful property of the solution. This property persists even when B is singular provided any zero diagonal elements in the triangular factor of B that is being employed (one may obtain such a zero even when rounding errors occur) are replaced by quantities of the order of $\|B\|$ macheps.

Consider now the solution of the related system

$$Cz \equiv (B+bp^T)z = b \tag{2.8}$$

where p is an arbitrary vector. When B is ill-conditioned, indeed even when it is singular

with a null space of dimension one, the matrix C will, in general, be quite well-conditioned. Suppose for the moment that C is *very* well-conditioned. Then the computed z_c will be of high accuracy. However z is very closely related to x, the solution of the original system. In fact

$$Bz = (1 - p^T z)b \tag{2.9}$$

and $p^T z$ is a scalar. Hence

$$x = z/(1 - p^T z), \tag{2.10}$$

i.e. z is merely a scalar multiple of the required x.

At first sight this looks rather spectacular; it appears that we have an accurate method of solving an ill-conditioned system. Obviously this cannot be true. Indeed the general error analysis mentioned above applies equally to the system (2.8) and implies that the computed z_c is the exact solution of some

$$(C + F)z_c = b, \quad \text{where } \|F\| \leqq f(n)\|C\| \text{ macheps}, \tag{2.11}$$

giving $(B + F + bp^T)z_c = b$, and hence showing that z_c is directly related to the solution of $(B + F)x = b$ and not of $Bx = b$. It is inconceivable that the rounding errors which give rise to F should be correlated in such a way as to leave the solution comparatively unaffected.

The answer to this apparent paradox is that $1 - p^T z$ must be given to low accuracy. We now show heuristically how this comes about. For convenience in the discussion we assume that $\|B\| = \|b\| = \|p\| = 1$. If B is ill-conditioned this means that $\|B^{-1}\|$ must be large. Indeed if B is singular to working accuracy on a t digit base β machine, $\|B^{-1}\|$ is of the order of magnitude of β^t. Hence there must be right-hand sides b for which the correct solution is of order β^t. (Indeed this will be true of almost all right-hand sides). Now if C i.e. $B + bp^T$ is well-conditioned this implies that C^{-1} is of order unity. Hence the exact solution of $Cz = b$ is of order unity and since C is, by assumption, well-conditioned, z_c will be very close to z and therefore itself of order unity. Now $z/(1 - p^T z)$ is the exact solution of $Bx = b$ and is, by hypothesis very large; hence $1 - p^T z$ must be very small. In practice when we compute $1 - p^T z_c$ we find that almost complete cancellation takes place and the computed value is largely dependent on the end figures of z_c. Note that nothing is gained by accumulating the inner product $p^T z_c$ in double precision; to achieve a useful value of $1 - p^T z_c$, z_c is required to higher precision.

This may pose a second paradox to the reader. Suppose b is a right-hand side for which the corresponding solution is *not* large. Such right-hand sides exist, however ill-conditioned B may be; indeed we have only to take a random x with $\|x\|$ of order unity and then $b = Bx$ is, of course, just such a right-hand side. Since x is not now large, and z cannot be small, $1 - p^T z$ is now not small and no cancellation will take place when we compute it. Hence it really does look as though we shall get an accurate solution to an ill-conditioned system.

The fallacy in this argument is the following. We have assumed that although B is ill-conditioned, $B + bp^T$ is well-conditioned. In the case when b is such that the solution x is of order unity, $B + bp^T$ will be ill-conditioned for all choices of p. We can see how this comes about by considering the extreme case when B is singular. There then exists a nonzero y such that $y^T B = 0$. Now if we take a random right-hand side b the corresponding solution is infinite; to have a finite solution x, b must be such that $b = Bx$.

Then

$$y^T C = y^T(B + bp^T) = y^T(B + Bxp^T) = 0 \tag{2.12}$$

and hence for such a right-hand side C is also singular for all choices of p.

To carry over this argument to the case when B is ill-conditioned rather than singular it is perhaps most instructive to work with the singular value decomposition (SVD)

$$B = U \operatorname{diag}(\sigma_i) V^H \tag{2.13}$$

of B. (See Golub and Kahan (1965)). Here U and V are unitary and the σ_i, satisfying

$$\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_n \geqq 0, \tag{2.14}$$

are the *singular values* of B. If B is ill-conditioned then at least one singular value σ_n is small and we have

$$u_n^H B = \sigma_n v_n^H, \tag{2.15}$$

where u_n^H, v_n^H are the nth rows of U^H and V^H. Hence if

$$b = Bx \tag{2.16}$$

we have

$$u_n^H C = u_n^H(B + bp^T) = u_n^H(B + Bxp^T) = \sigma_n[v_n^H + (v_n^H x)p^T]. \tag{2.17}$$

Thus the right-hand side of (2.17) will be small for any unit p if x is of order unity, and since $\sigma_n(C) = \min_{\|u\|_2=1} \|u^H C\|_2 \leqq \|u_n^H C\|_2$, the matrix $C = B + bp^T$ will also have a small singular value.

On the whole the situation is very much better with respect to those right-hand sides having large solutions. With such b the related system $B + bp^T$ can usually be chosen to be well-conditioned (indeed will be so for almost all p^T); hence z_c is an accurate approximation to z and the corresponding x_c given by

$$x_c = z_c/(1 - p^T z_c) \tag{2.18}$$

will be a vector in almost exactly the correct direction. It will be wrong only by the scalar factor $(1 - p^T z)/(1 - p^T z_c)$. If we are not interested in a constant multiplier, as is the case when we are computing an eigenvector, the nature of the error in the computed x_c is of no consequence.

In inverse iteration our matrix B is $A - \lambda I$; it is an essential feature of the algorithm that $A - \lambda I$ be reasonably near to singularity and therefore ill-conditioned. Moreover it is hoped that x_0 will contain a reasonable sized component of u_s and hence that the solution is large. If this is not true of x_0 it is certainly to be expected that x_1 or some subsequent x_i will be sufficiently rich in u_s for the corresponding solution of

$$(A - \lambda I)x = x_i \tag{2.19}$$

to be large. Circumstances are therefore propitious for working with the modified system

$$(A - \lambda I + x_i p^T)z = x_i. \tag{2.20}$$

However, although the material we have presented gives some insight into the nature of inverse iteration it is natural to ask if there really is any justification in practice for working with $A - \lambda I + x_i p^T$; is one paying too high a price for the "comfort" of working

with a well-conditioned system? In my opinion one is. The disadvantages are the following.

(i) In practice one often uses inverse iteration with some reduced form of the original matrix, so that A is, for example, banded (including the case when it is tridiagonal) or of Hessenberg form. In this case $A-\lambda I$ is also of the special form but $A-\lambda I+x_i p^T$ will not in general be so.

(ii) The modified matrix $A-\lambda I+x_i p^T$ is a function of x_i and if there is more than one iteration one has the disadvantage of working with a different matrix each time. This disadvantage can be diminished by choosing p to be e_m, where the largest component of $|x_0|$ is the mth and in the triangularization of $A-\lambda I+x_i e_m^T$ treating the mth column last. The triangularization is then substantially the same each time. However this may not be particularly economic if A is of a very condensed form, tridiagonal for instance.

It might be felt that the solution of equations with the ill-conditioned matrix $A-\lambda I$ is so unsatisfactory that it is preferable to use the well-conditioned $A-\lambda I+x_i p^T$ but this is hardly true. When the solution of $(A-\lambda I)x=x_i$ is large, the nature of the error is precisely the same as that in the solution obtained via $A-\lambda I+x_i p^T$, namely that it is wrong by a constant multiplier and this does not concern us. Indeed we obtain the solution of some

$$(A+E-\lambda I)x=x_i \tag{2.21}$$

and we have seen that this is a multiple of the solution z of

$$(A+E-\lambda I+x_i p^T)z=x_i. \tag{2.22}$$

If $A-\lambda I+x_i p^T$ is well-conditioned, equation (2.22) has essentially the same solution for virtually all small E. The only disadvantages of working directly with $(A-\lambda I)$ is that the solution of (2.21) may be so large as to cause overflow; also we may obtain zero diagonal elements in the triangularization and these must be replaced by some suitably small quantity.

3. Numerical example. Consider the 2×2 linear system $Bx=b$ given by

$$\begin{aligned} .51273x_1+.62137x_2&=.14012\\ .41835x_1+.50701x_2&=.34827. \end{aligned} \tag{3.1}$$

This is extremely ill-conditioned; indeed for work of 5 significant decimals it may be considered to be singular. The first 9 figures of the exact solution vector are

$$x_1=-15977.7406\cdots, \qquad x_2=13184.4264\cdots. \tag{3.2}$$

The size of this solution fully reflects the ill-condition of B. Now consider the modified system $(B+be_2^T)z=b$, i.e.

$$\begin{aligned} .51273z_1+.76149z_2&=.14012\\ .41835z_1+.85528z_2&=.34827. \end{aligned} \tag{3.3}$$

This system is very well-conditioned and when solved by Gaussian elimination it leads to

$$\begin{aligned} .51273z_1+.76149z_2&=.14012\\ .23396z_2&=.23394 \end{aligned} \tag{3.4}$$

giving $z_1=-1.21177$, $z_2=.99992$. This is an accurate solution of the modified system

as is to be expected since $(B+be_2^T)$ is a well conditioned matrix. Now

$$x = z/(1-p^T z)$$

and $p^T z = .99992$. Hence $1-p^T z = .00008$ and almost complete cancellation has taken place. To obtain a value of $1-p^T z$ correct to 5 significant figures we would need 9 correct figures in z_2. However using the result $x = z/.00008$ gives

$$\tilde{x}_1 = -15147. \qquad \tilde{x}_2 = 12499, \tag{3.5}$$

to five significant decimals. $\tilde{x}$ does not agree well with x but

$$1.05484\tilde{x}_1 = -15978. \quad \text{and} \quad 1.05484\tilde{x}_2 = 13184,$$

to five figures and these agree to working accuracy with x. The computed $\tilde{x}$ is correct to within a scaling factor.

Consider now the same matrix B but with the right-hand side b given by

$$b^T = (.36873, .30086). \tag{3.6}$$

for which the exact solution is

$$x_1 = .48573,2814\cdots, \qquad x_2 = .19260,7003\cdots. \tag{3.7}$$

The solution corresponding to this right-hand side is not large i.e. it doesn't give any indication of the ill-condition of B. The modified system $(B+be_2^T)z = b$ is

$$\begin{aligned} .51273z_1 + .99010z_2 &= .36873 \\ .41835z_1 + .80787z_2 &= .30086. \end{aligned} \tag{3.8}$$

The system is as ill conditioned as the original. Indeed if we perform Gaussian elimination the reduced system is

$$\begin{aligned} .51273z_1 + .99010z_2 &= .36873 \\ .00002z_2 &= .00000. \end{aligned} \tag{3.9}$$

almost complete cancellation taking place. The computed solution is

$$z_1 = .71915, \qquad z_2 = 0 \tag{3.10}$$

and now bears no relation to the true solution. We now have $1-p^T z = 1-e_2^T z = 1$ so that no loss of accuracy takes place in determining this factor. It is easy to see that $B+bp^T$ is ill-conditioned for all choices of p such that $\|p\|_2 = 1$. The systems of linear equations arising in connexion with inverse iteration are peculiarly satisfactory from the point of view of the above discussion. In fact the more accurate the λ and hence the more ill-conditioned the matrix $B \equiv A - \lambda I$ the more "satisfactory" is the system.

However if we return to our original example and solve it by Gaussian elimination without making the preliminary transformation we find it, too, is satisfactory. In fact elimination of the system (3.1) leads to

$$\begin{aligned} .51273x_1 + .62137x_2 &= .14012 \\ .00002x_2 &= .23394 \end{aligned}$$

giving

$$\tilde{x}_1 = -14175, \qquad \tilde{x}_2 = 11697.$$

This is, of course, quite inaccurate but

$$1.12718\tilde{x}_1 = -15977.7\cdots, \qquad 1.12718\tilde{x}_2 = 13184.6\cdots,$$

and hence again $\tilde{x}$ has the correct direction to working accuracy and is wrong only by the scale factor 1.12718. Nothing has been gained by working with the modified system except the "comfort" of working with well-conditioned equations.

4. Newton's method. We now turn to an algorithm which at first sight appears to be completely unrelated to inverse iteration. In finding an eigenvalue λ and a corresponding normalized eigenvector x we require a solution of the nonlinear system

$$Ax - \lambda x = 0, \qquad x^H x = 1, \tag{4.1}$$

if x is normalized according to the 2 norm. If $\lambda^{(0)}$ and x_0, $x_0^H x_0 = 1$ is an approximate eigenpair, the true pair $\lambda^{(0)} + \delta\lambda^{(0)}$, $x_0 + \delta x_0$ satisfies exactly the relations

$$A(x_0 + \delta x_0) - (\lambda^{(0)} + \delta\lambda^{(0)})(x_0 + \delta x_0) = 0, \qquad (x_0 + \delta x_0)^H (x_0 + \delta x_0) = 1, \tag{4.2}$$

and hence ignoring second order quantities

$$\begin{aligned} (Ax_0 - \lambda^{(0)} x_0) + (A - \lambda^{(0)} I)\,\delta x_0 - \delta\lambda^{(0)} x_0 &= 0, \\ x_0^H\,\delta x_0 &= 0. \end{aligned} \tag{4.3}$$

This gives the system

$$\begin{bmatrix} (A - \lambda^{(0)} I & x_0 \\ x_0^H & 0 \end{bmatrix} \begin{bmatrix} \delta x_0 \\ -\delta\lambda^{(0)} \end{bmatrix} = \begin{bmatrix} -r_0 \\ 0 \end{bmatrix} \quad \text{where } r_0 = Ax_0 - \lambda^{(0)} x_0 \tag{4.4}$$

which is a set of equations of order $n+1$ for the $n+1$ quantities $\delta\lambda^{(0)}$ and the n components of δx_0. Comparing (4.3) with (1.2) we see that apart from a scaling factor $x_0 + \delta x_0$ is the same as would be obtained by performing one step of inverse iteration with λ_0 and x_0. The relation between k_1 and $\delta\lambda^{(0)}$ is

$$k_1 = \delta\lambda^{(0)}/(1 + \|\delta x_0\|^2)^{1/2}. \tag{4.5}$$

When we are very close to an eigenpair one step of inverse iteration gives the same result to first order as one Newton step. The iterative scheme derived from (4.4) by replacing the index zero by a running index is not quite Newton's method for the system (4.1) since after the first step the vector x_p would not be normalized. We shall use the term "Newton iteration" in rather a loose sense in the rest of the paper. The point that we shall be making from time to time is that although inverse iteration has an entirely different motivation it will always turn out to be closely related to a method for solving a nonlinear system in which second order quantities have been ignored. It is interesting though, that inverse iteration is sometimes the exact equivalent of the classical Newton's method. (This is true, for example, for the system (4.8).)

We are particularly interested in the augmented matrix on the left-hand side of (4.4) when $\lambda^{(0)}$ and x_0 are an eigenvalue λ and the corresponding eigenvector x. At this stage $\lambda - \lambda I$ is singular. In general this not true of the augmented matrix. For if it is singular we have a nonnull vector $(y^T, k)^T$ such that

$$\begin{bmatrix} A - \lambda I & x \\ x^H & 0 \end{bmatrix} \begin{bmatrix} y \\ k \end{bmatrix} = 0 \tag{4.6}$$

giving

$$(A - \lambda I)y = -kx, \qquad x^H y = 0. \tag{4.7}$$

Obviously y cannot be null. If k is zero, y is an eigenvector and it cannot be a multiple of x since (4.7) would then require $x^H x = 0$. Hence in this case there are two eigenvectors corresponding to λ. If k is not zero equations (4.7) show that λ corresponds to a quadratic elementary divisor at least and y is a related principal vector of grade 2. Hence if λ is a simple eigenvalue the limiting augmented matrix cannot be singular. Note that when A is Hermitian the augmented matrix is Hermitian; when A is real and symmetric and λ and x are real, the augmented matrix is real and symmetric.

In practice it is usually simpler to normalize the successive $|x_p|$ so that the largest component (or one of the larger components) is unity. If the largest component of $|x_0|$ is the mth we may normalize it and subsequent x_p so that $e_m^T x_p = 1$. It may happen that in some subsequent x_i the mth component is not actually the largest, but if x_0 really is a good approximation to the eigenvector the mth component of the true x is bound to be one of its larger components. We shall show that the tie up between Newton iteration and inverse iteration is now complete.

For simplicity of discussion we assume that A and x_0 have been transformed, if necessary, by a permutation matrix so that it is the last component of $|x_0|$ which is the largest. We are then trying to solve the system

$$Ax - \lambda x = 0, \qquad e_n^T x = 1 \tag{4.8}$$

starting with $\lambda^{(0)}$ and x_0. The successive iterates satisfy the relations

$$A(x_{p-1} + \delta x_{p-1}) - (\lambda^{(p-1)} + \delta\lambda^{(p-1)})(x_{p-1} + \delta x_{p-1}) = 0, \qquad e_n^T(x_{p-1} + \delta x_{p-1}) = 0 \tag{4.9}$$

and ignoring second order quantities

$$(A - \lambda^{(p-1)} I)\,\delta x_{p-1} - \delta\lambda^{(p-1)} x_{p-1} = -r_{p-1} \quad \text{where } r_{p-1} = Ax_{p-1} - \lambda^{(p-1)} x_{p-1}, \tag{4.10}$$

$$e_n^T \delta x_{p-1} = 0. \tag{4.11}$$

Equation (4.11) shows that δx_{p-1}^T is of the form

$$(\delta x_{p-1,1}, \delta x_{p-1,2}, \cdots, \delta x_{p-1,n-1}, 0), \tag{4.12}$$

i.e. we have only to determine $(n-1)$ components, while equation (4.10) implies that

$$[\tilde{A} - \lambda^{(p-1)} \tilde{I} | x_{p-1}] \begin{bmatrix} \delta x_{p-1,1} \\ \vdots \\ \delta x_{p-1,n-1} \\ -\delta\lambda^{(p-1)} \end{bmatrix} = -r_{p-1} \tag{4.13}$$

where $\tilde{A} - \lambda^{(p-1)}\tilde{I}$ is the $n \times n-1$ matrix given by the first $n-1$ columns of $A - \lambda^{(p-1)} I$.

We now show that equations (1.2) and (4.13) are essentially the same equation. For if in (1.2) x_p and x_{p-1} are both normalized so that their last components are unity and we write

$$x_p = x_{p-1} + \delta x_{p-1} \tag{4.14}$$

equation (1.2) becomes

$$(A - \lambda I)(x_{p-1} + \delta x_{p-1}) = k_p x_{p-1} \tag{4.15}$$

i.e.

$$(A - \lambda I)\,\delta x_{p-1} - k_p x_{p-1} = -(A - \lambda I)x_{p-1} = -r_{p-1} \tag{4.16}$$

which may now be written as

$$[\tilde{A}-\lambda\tilde{I}|x_{p-1}]\begin{bmatrix}\delta x_{p-1,1}\\ \vdots\\ \delta x_{p-1,n-1}\\ -k_p\end{bmatrix}=-r_{p-1}. \tag{4.17}$$

This is precisely equation (4.13) and shows that k_p, the normalizing factor in inverse iteration is the appropriate correction to λ. In practice if the initial λ is quite accurate it may be inadvisable to correct λ since if one does this the matrix $\tilde{A}-\lambda\tilde{I}$ changes with each iteration. We can regard the original λ and the current x_p as the current approximate eigenpair. $\tilde{A}-\lambda\tilde{I}$ is then a fixed matrix. The column x_{p-1} changes with each iteration but in the triangularization this means we have only to reprocess the last column. Now $A-\lambda I$ is almost a singular matrix but in general $[\tilde{A}-\lambda\tilde{I}|x_{p-1}]$ will not be so. Indeed even if we improve λ each time by adding in the computed correction, so that λ tends to an eigenvalue and x_p to the corresponding eigenvector x, the limiting matrix $[\tilde{A}-\lambda\tilde{I}|x]$ will not, in general, be ill-conditioned even though $A-\lambda I$ is itself exactly singular. In fact we show that if λ and x are an exact eigenpair, $B=[\tilde{A}-\lambda\tilde{I}|x]$ can be singular only if λ is at least a double eigenvalue. For if B is singular there is a nonnull vector y which may be partitioned in the form

$$y^T=\underbrace{[\,\tilde{y}^t}_{n-1}\,\underbrace{|k]}_{1}, \tag{4.18}$$

such that $By=0$. If $k=0$ this implies that there are two independent eigenvectors x and y corresponding to λ. Otherwise we can take $k=1$ without loss of generality and we have

$$(A-\lambda I)\begin{bmatrix}\tilde{y}\\ 0\end{bmatrix}+x=0. \tag{4.19}$$

Since

$$(A-\lambda I)x=0 \tag{4.20}$$

this implies that $\begin{bmatrix}\tilde{y}\\ 0\end{bmatrix}$ is a vector of grade 2. In either case λ is an eigenvalue of multiplicity at least two.

When λ is a simple eigenvalue, so that the limiting matrix is nonsingular, convergence of Newton's method will as usual be asymptotically quadratic. For the residual r_p we have

$$\begin{aligned} r_p &= (A-\lambda^{(p)}I)x_p\\ &= (A-\lambda^{(p-1)}I-\delta\lambda^{(p-1)}I)(x_{p-1}+\delta x_{p-1})\\ &= (A-\lambda^{(p-1)}I)x_{p-1}+(A-\lambda^{(p-1)}I)\,\delta x_{p-1}-\delta\lambda^{(p-1)}x_{p-1}-\delta\lambda^{(p-1)}\,\delta x_{p-1}\\ &= -\delta\lambda^{(p-1)}\,\delta x_{p-1} \end{aligned} \tag{4.21}$$

the final reduction being a consequence of (4.10). The residual therefore involves the product of the corrections $\delta\lambda^{(p-1)}$ and δx_{p-1} and when these are both small the new residual is of second order as we would expect.

If we compute r_{p-1} accurately by accumulating inner products in double-precision, rounding only on completion of the inner product, the iterative procedure embodied in

equation (4.13) can be made to give more than single precision accuracy in x and λ without ever performing any true double-precision arithmetic.

Working in terms of the increments $\delta\lambda^{(p-1)}$ and δx_{p-1} rather than computing the new x_p directly removes the need for dealing with almost singular matrices and it is illuminating to see inverse iteration and Newton's method in this common light. However, we emphasize again that if we correct λ each time in the inverse iteration algorithms by taking

$$\lambda^{(p)} = \lambda^{(p-1)} + k_p$$

it is every bit as satisfactory numerically as Newton's method with single-precision computation of r_{p-1} and one can take full advantage of any special form of $A - \lambda^{(p)}I$ while this form is to some extent lost in the matrix $[\tilde{A} - \lambda^{(p)}\tilde{I}|x_p]$.

5. The generalized eigenvalue problems. The algorithms we have described extend to the generalized problems

$$Ax = \lambda Bx \quad \text{and} \quad (A_r\lambda^r + \cdots + A_1\lambda + A_0)x = 0 \tag{5.1}$$

but there are some interesting points of detail.

If B is nonsingular then $Ax = \lambda Bx$ is equivalent to

$$B^{-1}Ax = \lambda x. \tag{5.2}$$

Hence inverse iteration gives

$$(B^{-1}A - \lambda I)x_p = k_p x_{p-1} \tag{5.3}$$

or

$$(A - \lambda B)x_p = k_p B x_{p-1}. \tag{5.4}$$

When we wish to correct λ each time we have

$$\lambda^{(p)} = \lambda^{(p-1)} + k_p. \tag{5.5}$$

If we assume that there is a complete set of vectors u_i such that

$$Au_i = \lambda_i Bu_i, \tag{5.6}$$

and write

$$x_0 = \sum \alpha_i u_i, \qquad x_1 = \sum \beta_i u_i \tag{5.7}$$

then

$$\sum \beta_i(\lambda_i - \lambda)Bu_i = k_1 \sum \alpha_i Bu_i. \tag{5.8}$$

Hence

$$\beta_i = k_1[\alpha_i/(\lambda_i - \lambda)] \tag{5.9}$$

and in general

$$x_p = k_1 k_2 \cdots k_p \sum [\alpha_i/(\lambda_i - \lambda)^p]u_i. \tag{5.10}$$

It is not widely appreciated that instead of (5.4) one can use the iteration

$$(A - \lambda B)x_p = k_p A x_{p-1}. \tag{5.11}$$

For again expressing x_0 and x_1 in terms of u_i as in (5.7) we have

$$\sum \beta_i(\lambda_i - \lambda)Bu_i = k_1 \sum \alpha_i Au_i = k_1 \sum \alpha_i \lambda_i Bu_i \tag{5.12}$$

giving

$$\beta_i = k_1[\alpha_i\lambda_i/(\lambda_i - \lambda)] \tag{5.13}$$

and

$$x_p = k_1k_2 \cdots k_p \sum \alpha_i[\lambda_i/(\lambda_i - \lambda)]^p u_i. \tag{5.14}$$

This second form of iteration gives better convergence to the eigenvalue of maximum modulus while the original form is better for convergence to that of smallest modulus. More generally one can perform the iteration

$$(A - \lambda B)x_p = k_p(A - \mu B)x_{p-1} \tag{5.15}$$

giving

$$x_p = k_1k_2 \cdots k_p \sum \alpha_i[(\lambda_i - \mu)/(\lambda_i - \lambda)]^p u_i \tag{5.16}$$

and if one has some information on all the eigenvalues it may be possible to choose μ so as to achieve significantly better convergence to a prescribed eigenvector.

When B is singular we cannot use the same motivation, but if we consider the solution of the nonlinear system

$$Ax - \lambda Bx = 0, \qquad e_n^T x = 1 \tag{5.17}$$

we have

$$A(x_{p-1} + \delta x_{p-1}) - (\lambda^{(p-1)} + \delta\lambda^{(p-1)})B(x_{p-1} + \delta x_{p-1}) = 0 \tag{5.18}$$

$$e_n^T \delta x_{p-1} = 0. \tag{5.19}$$

Ignoring second order quantities this gives the iterative procedure

$$\begin{aligned}(A - \lambda^{(p-1)}B)\,\delta x_{p-1} - \delta\lambda^{(p-1)}Bx_{p-1} &= -(Ax_{p-1} - \lambda^{(p-1)}Bx_{p-1}) \\ &= -r_{p-1}\end{aligned} \tag{5.20}$$

i.e.

$$[\tilde{A} - \lambda^{(p-1)}\tilde{B} | Bx_{p-1}] \begin{bmatrix} \delta x_{p-1,1} \\ \vdots \\ \delta x_{p-1,n-1} \\ -\delta\lambda^{(p-1)} \end{bmatrix} = -r_{p-1} \tag{5.21}$$

where $\tilde{A} - \lambda^{(p-1)}\tilde{B}$ consists of the first $n-1$ columns of $A - \lambda^{(p-1)}B$. This iterative procedure is of course directly equivalent to

$$(A - \lambda^{(p-1)}B)x_p = \delta\lambda^{(p-1)}Bx_{p-1} \tag{5.22}$$

which comes immediately from (5.18). This corresponds to the form of inverse iteration given in equation (5.4).

In this approach we are no longer dependent on the nonsingularity of B for our motivation. It can be shown that the limiting matrix of the system (5.21) is singular only if λ is an eigenvalue of multiplicity at least 2, provided we include in this category the cases when $\det(A - \mu B) \equiv 0$. The second form of inverse iteration, that of equation (5.11) can also be derived via Newton's method. For if x_{p-1} and $\lambda^{(p-1)}$ are good approximations we have to first order

$$Ax_{p-1} = \lambda^{(p-1)}Bx_{p-1} \tag{5.23}$$

and ignoring second order quantities (5.22) becomes

$$(A-\lambda^{(p-1)}B)x_p=\left(\frac{\delta\lambda^{(p-1)}}{\lambda^{(p-1)}}\right)Ax_{p-1}\equiv k_pAx_{p-1}. \tag{5.24}$$

This derivation, which reveals that the k_p of (5.11) is $\delta\lambda^{(p-1)}/\lambda^{(p-1)}$, underlines our remark that the second form of inverse iteration is appropriate for the determination of the eigenvalue of largest modulus. The general form of equation (5.15) is obtained in a similar way; we have

$$[A-\lambda^{(p-1)}B]x_p=k_p[A-\mu B]x_{p-1}, \tag{5.25}$$

where

$$k_p[\lambda^{(p-1)}-\mu]=\delta\lambda^{(p-1)}. \tag{5.26}$$

Turning finally to the eigenvalue problem

$$(A_r\lambda^r+\cdots+A_1\lambda+A_0)x=0 \tag{5.27}$$

our techniques will be adequately illustrated without notational inconvenience by the case $r=3$. The problem (5.27) is then equivalent to

$$\begin{bmatrix}0&I&0\\0&0&I\\A_0&A_1&A_2\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix}=\lambda\begin{bmatrix}I&&\\&I&\\&&-A_3\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix} \tag{5.28}$$

and hence inverse iteration gives

$$\begin{bmatrix}-\lambda I&I&0\\0&-\lambda I&I\\A_0&A_1&A_2+\lambda A_3\end{bmatrix}\begin{bmatrix}x_p\\y_p\\z_p\end{bmatrix}=k_p\begin{bmatrix}I&&\\&I&\\&&-A_3\end{bmatrix}\begin{bmatrix}x_{p-1}\\y_{p-1}\\z_{p-1}\end{bmatrix} \tag{5.29}$$

leading to

$$y_p=\lambda x_p+k_px_{p-1},\qquad z_p=\lambda y_p+k_py_{p-1}, \tag{5.30}$$

$$A_0x_p+A_1y_p+(A_2+\lambda A_3)z_p=-k_pA_3z_{p-1}. \tag{5.31}$$

Eliminating y_p and z_p these equations give

$$\begin{aligned}(A_0+A_1\lambda+A_2\lambda^2+A_3\lambda^3)x_p=-k_p[&(A_1+A_2\lambda+A_3\lambda^2)x_{p-1}\\&+(A_2+A_3\lambda)y_{p-1}+A_3z_{p-1}]\end{aligned} \tag{5.32}$$

and hence we can derive x_p by the solution of an $n\times n$ system (a system of order $nr\times nr$ is avoided) and k_p can be chosen to give the appropriate normalization; y_p and z_p may then be determined from (5.30).

We now attempt to relate this to Newton's method for the system

$$(A_3\lambda^3+A_2\lambda^2+A_1\lambda+A_0)x=0,\qquad e_n^Tx=1. \tag{5.33}$$

If λ and x_{p-1} are an approximate pair (we omit the suffix from λ for convenience and more direct comparison) then $\lambda+\delta\lambda$, $x_{p-1}+\delta x_{p-1}$, the exact pair, satisfy

$$[A_3(\lambda+\delta\lambda)^3+A_2(\lambda+\delta\lambda)^2+A_1(\lambda+\delta\lambda)+A_0](x_{p-1}+\delta x_{p-1})=0 \tag{5.34}$$

$$e_n^T\delta x_{p-1}=0. \tag{5.35}$$

Ignoring second order quantities and working with $x_p = x_{p-1} + \delta x_{p-1}$ for the moment we have

$$(A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0)x_p + \delta\lambda(3A_3\lambda^2 + 2A_2\lambda + A_1)x_{p-1} = 0. \tag{5.36}$$

We do not have an immediate tie up with equation (5.32) but now observe that in inverse iteration the limiting x, y and λ satisfy

$$y = \lambda x, \qquad z = \lambda y. \tag{5.37}$$

Hence if we assume that x_{p-1}, y_{p-1}, z_{p-1} satisfy the relations (5.37) to first order we have, with second order errors (since k_p is of first order)

$$(A_0 + A_1\lambda + A_2\lambda^2 + A_3\lambda^3)x_p = -k_p(A_1 + 2A_2\lambda + 3A_3\lambda^2)x_{p-1} \tag{5.38}$$

which is precisely (5.36) with $\delta\lambda = k_p$. However, at this point we emphasize an advantage of inverse iteration both as regards the insight provided on convergence and potentially superior performance. If λ is not a particularly good approximation it will be better to use equations (5.30) and (5.32). We know that this will give a speed of convergence which is controlled by the relative sizes of $|(\lambda - \lambda_i)|$. We have no need to think in terms of asymptotic rates of convergence; the global rate of convergence is satisfactory from the start if λ is sufficiently close to the selected eigenvalue. The rate of convergence can be improved if we update λ each time but we then need a recalculation and re-triangularization of $A_0 + A_1\lambda + A_2\lambda^2 + A_3\lambda^3$ in each iteration.

If we express Newton's method in terms of a computation of δx_{p-1} as we have earlier, (5.36) gives

$$\begin{aligned}(A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0)\,\delta x_{p-1} + \delta\lambda(3A_3\lambda^2 + 2A_2\lambda + A_1)x_{p-1} \\ = -(A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0)x_{p-1}\end{aligned} \tag{5.39}$$

the right-hand side being the residual vector r_{p-1} corresponding to λ and x_{p-1}. Since we are assuming δx_{p-1} has a zero last component this gives

$$[(\tilde{A}_3\lambda^3 + \tilde{A}_2\lambda^2 + \tilde{A}_1\lambda + \tilde{A}_0)|(3A_3\lambda^2 + 2A_2\lambda + A_1)x_{p-1}]\begin{bmatrix}\delta x_{p-1,1}\\ \vdots \\ \delta x_{p-1,n-1}\\ \delta\lambda\end{bmatrix} = -r_{p-1} \tag{5.40}$$

where the first matrix in the square brackets consists of the first $n-1$ columns of $A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0$. Again provided only we calculate the residual to higher accuracy the algorithm is capable of attaining a δx_{p-1} and $\delta\lambda$ such that $x_{p-1} + \delta x_{p-1}$ and $\lambda + \delta\lambda$ are correct to more than single precision without working to more than single precision at any other stage.

6. Ill-conditioned eigensystems. In practice inverse iteration often behaves in a way which, at first sight, seems at variance with what has been discussed in earlier sections. Many of the most effective algorithms are based on a composite technique in which "very accurate" eigenvalues are computed first and then inverse iteration is used with each approximation in turn to give the corresponding eigenvector.

Suppose such a λ is used and, starting from an arbitrary x_0, a sequence of normalized vectors x_p is derived using inverse iteration. We have then

$$(A - \lambda I)x_p = k_p x_{p-1}, \tag{6.1}$$

where k_p is the scale-factor necessary to give a normalized x_p. As a check on the

performance it is natural to compute the residual r_p defined by

$$r_p = (A - \lambda I)x_p. \tag{6.2}$$

It is a common experience that after the first iteration the residual r_1 is very small but, rather surprisingly, in subsequent iterations it is *very* much larger. It is tempting to imagine that this is due to the fact that $A - \lambda I$ is nearly singular and hence rounding errors lead to poor solutions of the equation (6.1). However, this is not the explanation; the following analysis shows that this behavior is to be expected whenever λ belongs to an ill-conditioned eigenvalue, even if inverse iteration is performed without rounding errors.

For the moment then we consider exact inverse iteration and for convenience we assume that $A - \lambda I$ is scaled so that $\|A - \lambda I\|_2 = 1$. Starting with a unit vector x_0 we determine y_1 such that

$$(A - \lambda I)y_1 = x_0 \tag{6.3}$$

and we then normalize y_1 to give x_1. Hence

$$x_1 = y_1/\|y_1\|_2. \tag{6.4}$$

The residual r_1 is given by

$$r_1 = (A - \lambda I)x_1 = (A - \lambda I)y_1/\|y_1\|_2 = x_0/\|y_1\|_2 \tag{6.5}$$

so that

$$\|r_1\|_2 = \|x_0\|_2/\|y_1\|_2. \tag{6.6}$$

We therefore have the general result. If starting from any vector x (normalized or not) one exact step of inverse iteration gives y then the norm of the residual corresponding to the normalized y is $\|x\|_2/\|y\|_2$. Hence the greater the growth in norm produced by a step of inverse iteration, the smaller the residual.

Now we have referred to λ as "a very accurate" eigenvalue. In practice the best we can hope for is that λ is an exact eigenvalue of some $A + E$ where $\|E\|_2$ is small. Suppose this is true for an E such that

$$\|E\| \leqq \beta^{-t} \tag{6.7}$$

where the λ has been computed on a computer working in base β with t digits in the mantissa. We may refer to such a λ as being an "optimum" eigenvalue for t digit computation. It should be emphasized that such a λ need not be particularly "accurate" in the usual sense if A is ill-conditioned with respect to its eigenvalue problem. Thus if

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \tag{6.8}$$

$\lambda = 0.9999$ is an "optimum" eigenvalue for 12 digit decimal computation since it is an exact eigenvalue of

$$A + E = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ -10^{-12} & 0 & 1 \end{bmatrix}, \tag{6.9}$$

though since the eigenvalues of A are 1, 1, 1 this λ has an error of 10^{-4}.

We now show that with an "optimum" eigenvalue, whatever its error may be, there always exists an x_0 such that x_1 gives a residual r_1 with

$$\|r_1\| \leqq n^{1/2}\beta^{-t}. \tag{6.10}$$

Since λ is an optimum eigenvalue we know that $A+E-\lambda I$ is exactly singular for some E satisfying (6.7). Let y_s be the exact solution of

$$(A-\lambda I)y_s = e_s, \tag{6.11}$$

where e_s is the sth column of I. We show that at least one of the y_s satisfies the relation

$$\|y_s\| \geqq \beta^t/n^{1/2}. \tag{6.12}$$

Equation (6.11) gives

$$(A-\lambda I)Y = I$$

where Y is the matrix with columns y_s. Hence

$$(A+E-\lambda I)Y = I+EY. \tag{6.13}$$

But $A+E-\lambda I$ is singular and hence $I+EY$ is singular. Therefore

$$1 \leqq \|EY\|_2 \leqq \|E\|_2\|Y\|_2 \tag{6.14}$$

or

$$\|Y\|_2 \geqq 1/\|E\|_2 \geqq \beta^t \tag{6.15}$$

and the result follows. In general there will, of course, be a number of e_s for which it is true. If we were to take such an e_s as our initial x_0 then

$$x_1 = y_s/\|y_s\| \tag{6.16}$$

and

$$r_1 = (A-\lambda I)x_1 = (A-\lambda I)y_s/\|y_s\|_2 = e_s/\|y_s\|_2 \tag{6.17}$$

giving

$$\|r_1\|_2 \leqq n^{1/2}\beta^{-t}. \tag{6.18}$$

Hence with an "optimum" λ it will always be possible to choose an x_0 so that after one iteration the residual is "negligible to working accuracy." In fact if we take a random x_0 the probability that one interation will produce an x_1 having a pathologically small residual is very high indeed; this would only fail to be true if we happen to choose an x_0 which is almost completely deficient in all those e_s which give a y_s satisfying (6.12).

It might be imagined that in the second iteration conditions are even more favorable; surely an x_1 which gives a negligible residual is a better starting point than a random vector? We are encouraged in this view by our original motivation for inverse iteration. We might argue thus. When the first iteration is so successful, presumably x_1 is rich in the eigenvector corresponding to λ; x_2 will then be even more so and therefore give an even better residual. "Why is this so often untrue"?

The cause of the trouble lies in our use of the system of eigenvectors as the basis in which to express x_0. Assuming for the moment that A *has* a complete set of vectors, these vectors may nevertheless be very close to linear dependence. Indeed this will always be true when A is ill-conditioned with respect to its eigenvalue problem. The size of the individual components of a vector expressed in terms of a set of vectors which are nearly linear dependent can be very misleading.

Suppose there is a subset of almost linearly dependent unit eigenvectors $u_1, u_2, \cdots, u_r$. The matrix $U = [u_1, u_2, \cdots, u_r]$ will then have a small singular value ε (say) and we have

$$\sum_{i=1}^{r} \alpha_i u_i = \varepsilon v_1 \tag{6.19}$$

for some α_i and v_1 satisfying

$$\sum_{1}^{r} |\alpha_i^2| = 1, \qquad \|v_1\|_2 = 1. \tag{6.20}$$

Let v_i $(i = 1, \cdots, n)$ form a complete set of orthonormal vectors of which v_1 is the first. Corresponding to an arbitrary unit x_0 we may write

$$x_0 = \sum_{i=1}^{n} \beta_i v_i, \quad \text{where } \sum_{i=1}^{n} |\beta_i|^2 = 1. \tag{6.21}$$

Since the v_i are orthonormal only a small percentage of all possible x_0 will lead to a small β_1 (the same remark applies to any specific β_j, of course). In order to assess the effect of inverse iteration on x_0 we must now revert to the basis u_i. From (6.19) we have

$$v_1 = \sum_{1}^{r} \left(\frac{\alpha_i}{\varepsilon}\right) u_i. \tag{6.22}$$

The other relations we shall write as

$$v_j = \sum_{i=1}^{n} \gamma_{ji} u_i \qquad (j = 2, \cdots, n). \tag{6.23}$$

The coefficients in the expansion of v_1 are large while those in the expansion of the other v_j will not in general be comparably large. (We assume here that there is not another subset of almost linearly dependent u_i). Hence we have

$$x_0 = \sum_{i=1}^{r} \left[\frac{\beta_1 \alpha_i}{\varepsilon} + \sum_{j=2}^{n} \beta_j \gamma_{ji}\right] u_i + \sum_{i=r+1}^{n} \left[\sum_{j=2}^{n} \beta_j \gamma_{ji}\right] u_i. \tag{6.24}$$

Observe now that although x_0 is a unit vector the coefficients of u_i $(i = 1, \cdots, r)$ will all, in general, be large since they all involve ε^{-1}. Moreover although x_0 is to be regarded as a random vector the coefficients of $u_1, \cdots, u_r$ will be substantially in the ratio $\alpha_1, \cdots, \alpha_r$. The coefficients of the remaining u_i will not, in general be large or in any other way special. It is an interesting fact that a random vector has such special characteristics when expressed in terms of the poor basis formed by the u_i. The coefficients of $u_1, \cdots, u_r$ are large but are related in just such a way as to make the total contribution from them to x_0 of "normal size"; severe cancellation takes place in the addition.

Now consider the effect of inverse iteration starting with x_0. For convenience we shall not normalize after each step but write

$$x_p = (A - \lambda I)^{-p} x_0. \tag{6.25}$$

We know that the residual corresponding to the normalized version of x_p is $\|x_{p-1}\|_2 / \|x_p\|_2$; the explicit expressions for the norms are so involved that they confuse

the issue. We shall give only a rough assessment of their size. We have then

$$x_1 = \sum_{i=1}^{r} \frac{1}{(\lambda_i - \lambda)}\left[\frac{\beta_1\alpha_i}{\varepsilon} + \sum_{j=2}^{n} \beta_j\gamma_{ji}\right]u_i + \sum_{i=r+1}^{n} \frac{1}{(\lambda_i - \lambda)}\left[\sum_{j=2}^{n} \beta_j\gamma_{ji}\right]u_i. \tag{6.26}$$

If we assume that λ is close to $\lambda_1, \cdots, \lambda_r$ (these r values are likely to be moderately close) then the bulk of the contribution to x_1 will come from the components of $u_1, \cdots, u_r$. These components will be large for two reasons, the presence of the factor ε^{-1} and the presence of the factors $(\lambda_i - \lambda)^{-1}$. Moreover when adding these components together we can no longer expect cancellation since the factors $(\lambda_i - \lambda)^{-1}$ will destroy the special relationship between the coefficients. Assuming for convenience that λ is closest to λ_1, $\|x_1\|_2$ will be substantially of order $\varepsilon^{-1}(\lambda_1 - \lambda)^{-1}$ and since $\|x_0\|_2 = 1$ the residual corresponding to the normalized x_1 will be of order $\varepsilon(\lambda_1 - \lambda)$. From our previous discussion $\varepsilon(\lambda_1 - \lambda)$ must be of order β^{-t} if λ is an optimum eigenvalue.

The critical point now is the passage from x_1 to x_2. We have

$$x_2 = \sum_{i=1}^{r} \frac{1}{(\lambda_i - \lambda)^2}\left[\frac{\beta_1\alpha_i}{\varepsilon} + \sum_{j=2}^{n} \beta_j\gamma_{ji}\right]u_i + \sum_{i=r+1}^{n} \frac{1}{(\lambda_i - \lambda)^2}\left[\sum_{j=2}^{n} \beta_j\gamma_{ji}\right]u_i \tag{6.27}$$

and the coefficients of the $u_1, \cdots, u_r$ will not, in general, be of the critical ratio which leads to cancellation. Hence $\|x_2\|_2$ will be of the order of magnitude $\varepsilon^{-1}(\lambda_1 - \lambda)^{-2}$ and $\|x_2\|_2/\|x_1\|_2$ will be of the order of $(\lambda_1 - \lambda)^{-1}$, showing that in the second iteration ε^{-1} plays no role in the amplification factor. In general $\|x_p\|_2$ will be of the order of $\varepsilon^{-1}(\lambda_1 - \lambda)^{-p}$ and the amplification from $\|x_{p-1}\|_2$ to $\|x_p\|_2$ will be of the order of $(\lambda_1 - \lambda)^{-1}$, implying that the residual corresponding to the normalized x_p is of the order of $(\lambda_1 - \lambda)$. So far we have assumed that the presence of the additional factors $1/(\lambda_i - \lambda)^p$ in the coefficients of $u_1, \cdots, u_r$ will always ruin the special relationship. However if it happens that for some value of p

$$(\lambda_i - \lambda)^p = (\lambda_j - \lambda)^p \qquad (i, j \leqq r) \tag{6.28}$$

the relationship will be restored and cancellation will take place when these components are added. $\|x_p\|_2$ will then be of the order of $(\lambda_1 - \lambda)^{-p}$ rather than $\varepsilon^{-1}(\lambda_1 - \lambda)^{-p}$. The residual corresponding to the normalized x_p will therefore be particularly poor but the normalized x_{p+1} will behave as well as x_1! The example in § 7 illustrates this point very well.

This analysis applies to exact computation. The inclusion of rounding errors in the inverse iteration process makes surprisingly little difference. However with exact inverse iteration the iterates will continue to converge to u_1, though generally rather slowly because under our hypothesis the ratio of $|\lambda_1 - \lambda|$ to $|\lambda_2 - \lambda|$ is not likely to be particularly small. However the residuals we discuss are not tending to zero since they are being computed always with the same λ; this suggests we should try to improve λ. When rounding errors are included this last point loses its force. When we attempt to solve

$$(A - \lambda I)y_i = x_i \tag{6.29}$$

we obtain at best the solution of

$$(A + F_i - \lambda I)y_i = x_i \tag{6.30}$$

where F_i satisfies some such relation as

$$\|F_i\|_2 \leqq f(n)\beta^{-t}\|A\|_2 \tag{6.31}$$

and $f(n)$ is some relatively innocuous function of n. The F_i changes with each iteration and since by hypothesis our initial λ is always exact for some $A+E$ we cannot hope to improve on it by using solutions of (6.30) subject to (6.31).

At this stage the reader may well ask if we have not attached too much importance to a small residual, bearing in mind that it does not imply a corresponding accuracy in the λ and x. Our attitude may be summed up as follows. If $\|x\|_2 = 1$ and

$$Ax - \lambda x = r \quad \text{with } \|r\|_2 = \varepsilon \quad \text{(say)} \tag{6.32}$$

then

$$(A - rx^H)x = \lambda x \quad \text{since } x^H x = \|x\|_2^2 = 1 \tag{6.33}$$

and $\|rx^H\|_2 = \|r\|_2 = \varepsilon$.

Hence if $\|r\| = \varepsilon$, λ and x are exact for the matrix $A - rx^H$ and $\|rx^H\|_2 = \varepsilon$. We have therefore shown that if λ is an optimum eigenvalue then it is always possible in one step of inverse iteration to obtain an eigenvector which is exact corresponding to that λ for a matrix within $n^{1/2}\beta^{-t}$ of A i.e. for a matrix "equal to A to working accuracy." There is much to be said for designing an algorithm which gives the exact solution to a "neighboring problem" and indeed one is hardly likely to design a general purpose algorithm which achieves a better result. Perhaps the strongest argument, though, for having an algorithm which provides eigenvectors giving negligible residuals is that computation of the residuals provides evidence that the algorithm has no flaws (as far as the example in hand is concerned) and that the computer has worked correctly. Consider the situation when A is very ill-conditioned with respect to its eigenvalue problem. Suppose we are working on a 12-digit decimal computer and A is such that relative perturbations of order 10^{-12} may make changes of order $10^{-2}\|A\|$ in some eigenvalues. Then however good the algorithm may be we are unlikely except by a happy accident, to obtain these eigenvalues with errors of less than $10^{-2}\|A\|$. If we use an algorithm which does not give correlated eigenvalues and eigenvectors the residuals corresponding to normalized vectors are likely to be of order $10^{-2}\|A\|$. We shall not know whether these residuals result from an error in our algorithm and/or from the malfunctioning of our computer or alternatively from the sensitivity of the eigensystem. The guarantee of a negligible residual eliminates the first two possibilities, or perhaps we should more correctly say, even if the algorithm *is* wrong and/or the computer *has* malfunctioned we have nevertheless solved a "neighboring" problem and that is all we can hope to do.

7. Numerical example. The points we have made can be well illustrated by a simple 2×2 example. Let

$$A = \begin{bmatrix} 1 & 1 \\ 10^{-10} & 1 \end{bmatrix}, \quad \lambda_1 = 1 - 10^{-5}, \quad \lambda_2 = 1 + 10^{-5}, \quad u_1 = \begin{bmatrix} 1 \\ -10^{-5} \end{bmatrix}, \quad u_2 \begin{bmatrix} 1 \\ +10^{-5} \end{bmatrix}. \tag{7.1}$$

The vectors u_1 and u_2 are close to linear dependence and

$$2^{-1/2}u_2 - 2^{-1/2}u_1 = 2^{1/2}(10^{-5})\begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{7.2}$$

showing that $\varepsilon = 2^{1/2}(10^{-5})$.

If we take $x_0^T = (\cos\theta, \sin\theta)$ then

$$x_0 = \tfrac{1}{2}(10^5 \sin\theta + \cos\theta)u_2 - \tfrac{1}{2}(10^{-5}\sin\theta - \cos\theta)u_1. \tag{7.3}$$

For most values of θ the coefficients of u_1 and u_2 are of order ε^{-1} and almost equal and opposite. Substantial cancellation takes place in the addition.

Suppose now we take $\lambda = 1-2(10^{-5})$, which is an exact eigenvalue of

$$\begin{bmatrix} 1 & 1 \\ 4.10^{-10} & 1 \end{bmatrix} \tag{7.4}$$

and therefore effectively an optimum eigenvalue for ten decimal digit computation, and let $x_0^T = (0, 1)$. The successive unnormalized y_i, the normalized x_i and the residuals r_i are as follows.

x_0	y_1	x_1	r_1	y_2	x_2	r_2	y_3	x_3	r_3
0	$-\frac{1}{3}(10^{10})$	1	0	$\frac{4}{3}(10^5)$	1	$\frac{3}{4}(10^{-5})$	$\frac{13}{12}(10^5)$	1	$\frac{12}{13}(10^{-5})$
1	$\frac{2}{3}(10^5)$	$-2(10^{-5})$	$-3(10^{-10})$	$-\frac{5}{3}$	$-\frac{5}{4}(10^{-5})$	$-\frac{3}{2}(10^{-10})$	$-\frac{7}{6}$	$-\frac{14}{13}(10^{-5})$	$-\frac{15}{13}(10^{-10})$

The vector y_1 is of order 10^{10} and r_1 is of order 10^{-10}; all subsequent y_i are of order 10^5 and the r_i of order 10^{-5}. It will be seen that x_i is tending slowly to u_1 since the λ is closer to λ_1 than to λ_2. This is because the computation is exact.

It is interesting to consider the behavior with the approximation $\lambda = 1$ since we have then

$$(\lambda_1-\lambda)^2 = (\lambda_2-\lambda)^2 \tag{7.5}$$

which is an example of the special case discussed above with $r = p = 2$. Starting with $x_0^T = (0, 1)$ the results are

x_0	y_1	x_1	r_1	y_2	x_2	r_2	y_3	x_3	r_3
0	10^{10}	1	0	0	0	1	10^{10}	1	0
1	0	0	10^{-10}	1	1	0	1	0	10^{-10}

The first residual is again of order 10^{-10} but now the second residual (i.e. the pth) is of order unity; again the third residual is of order 10^{-10} and the behavior repeats as was forecast by the general analysis.

This special behavior is not uncommon in practice. It tends to arise when a defective matrix is perturbed by rounding errors. Corresponding to an elementary divisor $(\lambda_1-\lambda)^p$ one tends to obtain eigenvalues $\lambda_1+\varepsilon^{1/p}$ where $\varepsilon^{1/p}$ takes the p possible values.

8. Inverse iteration and the singular value decomposition of $A-\lambda I$. In § 6 we analysed inverse iteration in terms of the system of eigenvectors in order to explain why the explicit expression for the successive iterates is so misleading. A more satisfactory analysis is provided by considering the SVD of $A-\lambda I$.

If we write

$$(A-\lambda I) = U \operatorname{diag}(\sigma_i) V^H, \tag{8.1}$$

then

$$(A-\lambda I)v_i = \sigma_i u_i, \qquad u_i^H(A-\lambda I) = \sigma_i v_i^H, \tag{8.2}$$

and in particular when λ is an exact eigenvalue of A the matrix $A-\lambda I$ is singular and σ_n,

the smallest singular value, is zero, so that

$$(A-\lambda I)v_n = 0, \qquad u_n^H(A-\lambda I) = 0. \tag{8.3}$$

Equation (8.3) implies that v_n and u_n are a right-hand eigenvector and a left-hand eigenvector respectively corresponding to λ. When λ is an "optimum" eigenvalue, $(A+E-\lambda I)$ is exactly singular and

$$\sigma_n \leqq \|E\|_2 \leqq \beta^{-t}. \tag{8.4}$$

It is well known that

$$\sigma_n = \min_{\|x\|_2=1} \|(A-\lambda I)x\|_2 \tag{8.5}$$

and the minimum is attained for $x = v_n$ since

$$(A-\lambda I)v_n = \sigma_n u_n. \tag{8.6}$$

The ideal starting vector for inverse iteration therefore is $x_0 = u_n$ and if y_1 is defined by $(A-\lambda I)y_1 = x_0$ we have $y_1 = v_n/\sigma_n$. For an arbitrary unit vector x_0 we may write

$$x_0 = \sum \alpha_i u_i, \qquad \sum |\alpha_i|^2 = 1 \tag{8.7}$$

$$y_1 = \sum (\alpha_i/\sigma_i)v_i \tag{8.8}$$

giving

$$\|r_1\|_2 = \|x_0\|_2/\|y_1\|_2 = 1/(\sum |\alpha_i/\sigma_i|^2)^{1/2}. \tag{8.9}$$

Clearly an essential feature of a good starting vector is that α_n should not be pathologically small; since the probability that a random vector will be pathologically deficient in u_n is very small, almost all starting vectors x_0 give a y_1 which, when normalized produces an x_1 having a small residual; indeed x_1 is dominated by its v_n component.

However when we turn to the second iteration the situation is entirely different. The difficulty is clearly exposed if we consider the most favorable choice of x_0 for the first iteration, namely $x_0 = u_n$. We then have $x_1 = v_n$. Now if we write

$$v_n = \sum \alpha_i u_i, \qquad \sum |\alpha_i|^2 = 1 \tag{8.10}$$

the all-important α_n is in fact $u_n^H v_n$. If λ is an ill-conditioned eigenvalue we can expect this to be small, since it is an approximation to s_n, the cosine of the angle between the left-hand eigenvector and the right-hand eigenvector. We have now

$$\|r\|_2 = \|v_n\|/\|(A-\lambda I)^{-1}v_n\| \doteq 1\Big/\Big(\sum_1^{n-1} |\alpha_i/\sigma_i|^2 + |s_n/\sigma_n|^2\Big)^{1/2}. \tag{8.11}$$

Suppose, for example, $\beta^{-t} = 10^{-12}$ and $s_n = 10^{-8}$; then the residual may be as large as 10^{-4} unless σ_{n-1} is appreciably smaller than 10^{-4}. Our very success in the first iteration handicaps us in the second and this handicap persists in subsequent iterations except in very special circumstances of the type discussed in § 6.

An advantage of this analysis is that it immediately suggests a remedy. The most "satisfactory" approximate eigenvector we can get is v_n; this is the *eigenvector* of the Hermitian matrix $(A-\lambda I)^H(A-\lambda I)$ corresponding to its smallest *eigenvalue* σ_n^2. If we perform iteration with the matrix $[(A-\lambda I)^H(A-\lambda I)]^{-1}$ we shall converge to v_n and we are now dealing with an Hermitian matrix which necessarily has an orthogonal set of eigenvectors. There can be no danger of the type we have just been discussing. When

one has a factorization of $(A-\lambda I)$ it can be used to solve systems of either of the forms

$$(A-\lambda I)y=x \quad \text{or} \quad (A-\lambda I)^H y=x. \tag{8.12}$$

One iteration in the modified process can be carried out via the steps

$$(A-\lambda I)^H z_{s+1}=x_s, \qquad (A-\lambda I)y_{s+1}=z_{s+1}, \qquad x_{s+1}=y_{s+1}/\|y_{s+1}\|_2. \tag{8.13}$$

If we now write

$$x_0=\sum \alpha_i v_i, \tag{8.14}$$

we have, ignoring normalization,

$$z_s=\sum (\alpha_i/\sigma_i^{2s-1})u_i, \qquad x_s=\sum (\alpha_i/\sigma_i^{2s})v_i, \tag{8.15}$$

so that z_s and x_s will tend rapidly to u_n and v_n respectively. With exact computation the progress of the x_s is now steady and indeed when λ is an "optimum" eigenvalue, x_1 will already have had "impurities" attenuated by a factor β^{-2t} relative to those in x_0. Only a fantastically unfortunate choice of x_0 could deny success in one complete iteration. When rounding errors intervene there can be no question of continuing to gain accuracy in the successive x_s; indeed the attainable accuracy is limited by the very representation of the vectors themselves and x_1 will, in general, already have the maximum representable accuracy.

The reader may have the uneasy feeling that this alternative form of iteration gives us something for nothing, reasoning as follows. "We are interested in it mainly in the case when λ is an ill-conditioned eigenvalue of A. It appears that in going over to $(A-\lambda I)^H(A-\lambda I)$ one removes all the ill-condition from the problem." This view may be reinforced by considering a simple example such as

$$A=\begin{bmatrix} 2 & 1 \\ -1 & 0 \end{bmatrix}, \qquad \lambda_1=\lambda_2=1. \tag{8.16}$$

This matrix has a quadratic elementary divisor and the eigenvector is $(1, -1)$; both the eigenvalue and eigenvector are ill-conditioned. On the other hand

$$(A-I)^T(A-I)=\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \tag{8.17}$$

with eigenvalues 0 and 4. These eigenvalues are well-conditioned, as are the corresponding eigenvectors.

The weakness in this argument is the following. When A has an ill-conditioned λ then; if rounding errors are involved, the corresponding computed value of λ may vary over a considerable range. Thus if the above trivial problem is solved on a *ten*-digit decimal computer and a single rounding error is made one obtains eigenvalues of the form $1+10^{-5}k$ where $|k|$ is of order unity. Corresponding to such a λ one has

$$C=(A-\lambda I)^T(A-\lambda I)=\begin{bmatrix} 2-2\varepsilon+\varepsilon^2 & 2 \\ 2 & 2+2\varepsilon+\varepsilon^2 \end{bmatrix} \tag{8.18}$$

where $\varepsilon \equiv 10^{-5}k$. As λ varies over the range of optimum computed values C shows variations of order 10^{-5} rather than 10^{-10}. For each acceptable λ the corresponding C has a well-conditioned eigenproblem and the solution can be computed with errors of order 10^{-10}. However the smallest eigenvalue of C is $\varepsilon^4/4$, ignoring higher powers of λ; this is to be expected since σ_n should be of the order of the machine precision 10^{-10}, and

the smallest eigenvalue of C is σ_n^2. The corresponding eigenvector is

$$[1, 1-\varepsilon+\varepsilon^2/2+O(\varepsilon^3)]. \tag{8.19}$$

We are likely still to have errors or order ε, i.e. 10^{-5}, in our computed eigenvector of A. Nevertheless there is a certain satisfaction in working with a well-conditioned C rather than ill-conditioned A since convergence behavior is much stronger.

It should not be concluded from our remarks that we recommend the use of the iteration defined by equations (8.13) when λ is an optimum eigenvalue. The ordinary process of inverse iteration will almost always succeed in one iteration; if it does not do so one has only to re-start with an initial vector orthogonal to the first. This process can be continued until one reaches an initial vector which gives success in one iteration. It is rare for the first vector to fail and the average number of iterations is unlikely to be as high as 1.2. This makes it more economical than our alternative scheme; one iteration with the latter requires about as much computation as two iterations with the standard process.

Acknowledgment. The authors wish to thank the referees for their criticisms and for numerous suggestions which proved very helpful.

REFERENCES

P. M. ANSELONE AND L. B. RALL (1968), *The solution of characteristic value vector problems by Newton's method*, Numerische Math., 11, pp. 38–45.

G. H. GOLUB AND W. KAHAN (1965, *Calculating the singular values and pseudo inverse of a matrix*, SIAM J. Numer. Anal., 2, pp. 205–224.

G. H. GOLUB AND C. REINSCH (1970), *Singular value decomposition and least squares*, Numerische Math. 14, pp. 403–20.

M. R. OSBORNE AND S. MICHAELSON (1964), *The numerical solutions of eigenvalue problems in which the eigenvalue parameter appears nonlinearly with an application to differential equations*, Comput. J., 7, pp. 66–71.

G. PETERS AND J. H. WILKINSON (1971), *The calculation of specified eigenvectors by inverse iteration*, Handbook for Automatic Computation, Vol. II, Linear Algebra, by J. H. Wilkinson and C. Reinsch, Springer-Verlag, Berlin, Heidelberg, New York, pp. 418–439.

L. B. RALL (1961), *Newton's method for the characteristic value problem AX = Lambda BX*, J. Soc. Indus. Appl. Math., 9, pp. 288–93.

——— (1966), *Convergence of the Newton process to multiple solutions*, Numerische Math., 9, pp. 23–37.

J. M. VARAH (1968a), *The calculation of the eigenvectors of a general complex matrix by inverse iteration*, Math. Comput., 22, pp. 785–791.

——— (1968b), *Rigorous machine bounds for the eigensystem of a general complex matrix*, Ibid., 22, pp. 793–801.

——— (1970), *Computing invariant subspaces of a general matrix when the Eigensystem is poorly conditioned*, Ibid., 24, pp. 137–49.

H. WIELANDT (1944), *Das Iterationsverfahren bei nicht selbstadjungierten linearen Eigenwertaufgaben*, Math. Z., 50, pp. 93–143.

J. H. WILKINSON (1965), *The Algebraic Eigenvalue Problem*, Oxford University Press, London.

——— (1972), *Inverse iteration in theory and in practice*, Symposia Mathematica Volume X of the Institute Nazionale di Alta Mathematica Monograf, Bologna, 19, p. 361.

——— (1974), *Note on inverse iteration and ill-conditioned eigensystems*, Acta Univ. Carolinae—Math. et Phys. No. 1–2, p. 173.