Numer. Math. 13, 4—13 (1969)

Computational Aspects of F. L. Bauer's Simultaneous Iteration Method

HEINZ RUTISHAUSER

Dedicated to Professor E. STIEFEL on His 60th Birthday, 21. 4. 1969

Received January 16, 1969

In 1956, BAUER [1] introduced an iteration method, called "Bi-Iteration" for solving eigenvalue problems $\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}$ with arbitrary matrices A. The basic idea was that two sets of iteration vectors, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ and $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p$ are iterated simultaneously with $\boldsymbol{A}$ and $\boldsymbol{A}^T$ respectively. The iteration is then combined with linear combinations such that at any time the systems $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are biorthogonal. It can be shown that under certain conditions the $\boldsymbol{x}_i$ converge to right eigenvectors, the $\boldsymbol{y}_j$ to left eigenvectors corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ respectively.

In this paper a specialisation of this method to symmetric and positive definite matrices $\boldsymbol{A}$ is discussed and the general lines of a computing process are established. The details of such a process will be described later in a contribution to the Handbook series.

1. Simultaneous Iteration for a Symmetric and Positive Definite Matrix

If $\boldsymbol{A}$ is symmetric, then the two sets of iteration vectors $\boldsymbol{x}_1 \ldots \boldsymbol{x}_p$, $\boldsymbol{y}_1 \ldots \boldsymbol{y}_p$ can be chosen identical; they then form a system of orthonormal vectors $\boldsymbol{x}_1 \ldots \boldsymbol{x}_p$ which are combined into an $n \times p$-matrix X with the property $\boldsymbol{X}^T\boldsymbol{X}=\boldsymbol{I}_p$ ($=p\times p$ unit matrix).

It is assumed $1 \leqq p \leqq n$; the cases $p=1$ and $p=n$ are permitted, though usually $1<p \ll n$.

Denoting the matrix $\boldsymbol{X}$, as it is after k iteration steps, by $\boldsymbol{X}_k$, the method of BAUER is described as

$$\begin{aligned} &\text{Choose } \boldsymbol{X}_0 \text{ such that } \boldsymbol{X}_0^T\boldsymbol{X}_0=\boldsymbol{I}_p. \\ (1)\qquad &\boldsymbol{Z}_k := \boldsymbol{A}\boldsymbol{X}_{k-1} \\ &\boldsymbol{X}_k := \boldsymbol{Z}_k\boldsymbol{R}_k^{-1} \qquad (k=1, 2, \ldots), \end{aligned}$$

where $\boldsymbol{R}_k$ is an upper triangular matrix with positive diagonal elements chosen such that $\boldsymbol{X}_k$ again has p orthogonal columns. This implies

$$\boldsymbol{R}_k^T\boldsymbol{R}_k=\boldsymbol{Z}_k^T\boldsymbol{Z}_k, \tag{2}$$

but $\boldsymbol{R}_k$ is computed rather by Schmidt orthogonalisation of the columns of $\boldsymbol{Z}_k$ than by Cholesky decomposition of $\boldsymbol{G}_k = \boldsymbol{Z}_k^{\mathsf{T}} \boldsymbol{Z}_k$. In fact, the former process yields at the same time also $\boldsymbol{X}_k$, while the latter would require an extra calculation for this.

Provided $\boldsymbol{A}$ is positive definite, it can be shown (cf. [4], p. 182ff.) that with rule (1):

$$\begin{aligned} &\lim_{k\to\infty} \boldsymbol{X}_k = \boldsymbol{V} \\ &\lim_{k\to\infty} \boldsymbol{R}_k = \boldsymbol{D} \quad (= p\times p \text{ diagonal matrix}) \end{aligned} \tag{3}$$

both exist, and

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{D}. \tag{4}$$

That is, the columns of $\boldsymbol{X}_k$ converge to eigenvectors and the diagonal elements of $\boldsymbol{R}_k$ to eigenvalues of the matrix $\boldsymbol{A}$.

This result is established by considering the fact that (1) is equivalent to the LRCH-transformation[1], if the latter is applied to the initial matrix $\tilde{\boldsymbol{G}}_1 = \tilde{\boldsymbol{X}}_0^{\mathsf{T}} \boldsymbol{A}^2 \tilde{\boldsymbol{X}}_0$, where $\tilde{\boldsymbol{X}}_0$ is an $n\times n$ orthogonal matrix obtained by appending $n-p$ further columns to $\boldsymbol{X}_0$. Indeed, if (1) is applied with initial matrix $\tilde{\boldsymbol{X}}_0$ and $p=n$, then the k-th iterate $\tilde{\boldsymbol{X}}_k$ is related to the k-th LRCH transformate $\tilde{\boldsymbol{G}}_{k+1}$ of $\tilde{\boldsymbol{G}}_1$ through

$$\tilde{\boldsymbol{G}}_{k+1} = \tilde{\boldsymbol{X}}_k^{\mathsf{T}} \boldsymbol{A}^2 \tilde{\boldsymbol{X}}_k, \tag{5}$$

while $\boldsymbol{X}_k$ is contained in the p first columns of $\tilde{\boldsymbol{X}}_k$.

Provided (1) applied to the initial matrix $\tilde{\boldsymbol{X}}_0$ yields stable convergence (which is equivalent to saying that no "disorder of latent roots"[2] occurs if the LRCH-transformation is applied to $\tilde{\boldsymbol{G}}_1$), the above considerations also yield the convergence rate:

Theorem 1 [1]. Let $\boldsymbol{A}$ have the eigenvalues $\lambda_1 \geqq \lambda_2 \ldots \geqq \lambda_p > \lambda_{p+1} \geqq \cdots \geqq \lambda_n > 0$. If $\boldsymbol{v}_j$ is an appropriately chosen eigenvector to the eigenvalue λ_j, and if $\boldsymbol{x}_j^{(k)}$ denotes the j-th column of $\boldsymbol{X}_k$, then

$$\|\boldsymbol{v}_j - \boldsymbol{x}_j^{(k)}\| = O(q^k), \tag{6}$$

where $q = \operatorname{Max}(\lambda_{j+1}/\lambda_j,\ \lambda_j/\lambda_{j-1})$.

2. E-Iteration

The iteration rule (1) hides the fact that actually an iteration with p-dimensional spaces is performed. Indeed, if E_k denotes the linear space spanned by the columns of $\boldsymbol{X}_k$, then

$$E_k = \{\boldsymbol{x} \mid \boldsymbol{x} = \boldsymbol{A}\boldsymbol{y},\ y \in E_{k-1}\}. \tag{7}$$

We know, that if $\boldsymbol{A}$ is symmetric and positive definite, then $\lim_{k\to\infty} E_k$ exists and is an invariant subspace of $\boldsymbol{A}$, usually the one corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$. (In this latter case we speak of *stable convergence.*)

1. By LRCH-transformation we mean the LR-transformation with symmetric (Cholesky-) decomposition into two transposed factors (cf. [4], p. 149ff.).

2. See [3], § 4.3.

The $\boldsymbol{X}_k$ as produced by (1) just appear as a means for spanning E_k, but this is just one of many possible ways. As an example, also the trivial rule

$$\boldsymbol{X}_k := \boldsymbol{A}\boldsymbol{X}_{k-1} \tag{8}$$

defines the same sequence E_k, but in contrast to (1), the $\boldsymbol{X}_k$ produced by (8) are not well suited for numerical purposes. (8) can be used only with special care, while (1) is nearly foolproof.

Concerning the convergence rate of the E-iteration, it occurs that Theorem 1 does not apply here, but rather

Theorem 2. In case of stable convergence the angle $\phi_j^{(k)}$ between the j-th eigenvector $\boldsymbol{v}_j$ and E_k is asymptotically for $k \to \infty$:

$$O(q_j^k), \quad \text{with} \quad q_j = \lambda_{p+1}/\lambda_j. \tag{9}$$

Proof. First, the iteration rules (1) and (8) are both orthogonally invariant, i.e. replacing $\boldsymbol{A}$ by $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{U}$ (with $\boldsymbol{U}=n\times n$ orthogonal) and $\boldsymbol{X}_0$ by $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{X}_0$ has the effect that all $\boldsymbol{X}_k$ are replaced by $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{X}_k$, while the $\boldsymbol{R}_k$ are not changed at all. Therefore we can assume without loss of generality, that

$$\boldsymbol{A} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n).$$

Second, in case of stable convergence E_0 can be spanned by p vectors

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \\ x_{p+1,1} & x_{p+1,2} & x_{p+1,p} \\ \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,p} \end{pmatrix}. \tag{10}$$

According to (7), E_k is therefore spanned by the p vectors

$$\begin{pmatrix} \lambda_1^k & 0 & 0 \\ 0 & \lambda_2^k & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \lambda_p^k \\ \lambda_{p+1}^k x_{p+1,1} & \lambda_{p+1}^k x_{p+1,2} & \lambda_{p+1}^k x_{p+1,p} \\ \vdots & \vdots & \vdots \\ \lambda_n^k x_{n,1} & \lambda_n^k x_{n,2} & \lambda_n^k x_{n,p} \end{pmatrix}. \tag{11}$$

From (11) immediately follows, that the angle $\phi_j^{(k)}$ is at most $O(\lambda_{p+1}^k/\lambda_j^k)$, q.e.d.

3. Ritz-Iteration

Theorem 2 above reveals the interesting and important fact, that there are directions in E_k which are closer to the eigenvectors $\boldsymbol{v}_j$ than the columns of the matrices $\boldsymbol{X}_k$ as generated by (1).

Our present aim is to bring this improved convergence to light by a modified iteration rule, for which the convergence rate of the columns of $\boldsymbol{X}_k$ is the same as in (9). The key to such a modification are the eigendirections of the operator $\boldsymbol{A}$ projected onto E_k; these are not difficult to compute and for $k \to \infty$ approach the "optimal directions" (11). However, for convenience of computation it is preferable to use instead the eigendirections of the projection of $\boldsymbol{A}^{-2}$. The $n \times p$ matrix $\boldsymbol{Y}$ whose columns are such eigendirections, is given by the "projected eigenvalue equation":

$$\boldsymbol{Y}^{\mathsf{T}} \boldsymbol{A}^{-2} \boldsymbol{Y} = \boldsymbol{D}_k^{-2} \quad (= p \times p \text{ diagonal}), \tag{12}$$

where $\boldsymbol{Y}^{\mathsf{T}} \boldsymbol{Y} = \boldsymbol{I}_p$. The solution of (12) is given by

$$\boldsymbol{Y} = \boldsymbol{A} \boldsymbol{X}_{k-1} \boldsymbol{Q}_k \boldsymbol{D}_k^{-1}, \tag{13}$$

where $\boldsymbol{Q}_k$ is a $p \times p$ orthogonal matrix which transforms

$$\boldsymbol{G}_k = \boldsymbol{X}_{k-1}^{\mathsf{T}} \boldsymbol{A}^2 \boldsymbol{X}_{k-1} \quad \text{into} \quad \boldsymbol{Q}_k^{\mathsf{T}} \boldsymbol{G}_k \boldsymbol{Q}_k = \boldsymbol{D}_k^2, \tag{14}$$

and $\boldsymbol{X}_{k-1}$ is the previous iteration matrix, whose columns are assumed orthonormal. In the sequel, it is assumed throughout, that the diagonal elements of $\boldsymbol{D}_k$ are in decreasing order:

$$d_{11} \geqq d_{22} \geqq \cdots \geqq d_{pp} > 0.$$

With (13), the step from $\boldsymbol{X}_{k-1}$ to $\boldsymbol{X}_k$ is described by[3,4]:

(15)
a) $\boldsymbol{Z}_k := \boldsymbol{A} \boldsymbol{X}_{k-1}$.
b) $\boldsymbol{G}_k := \boldsymbol{Z}_k^{\mathsf{T}} \boldsymbol{Z}_k$.
c) Solve the eigenvalue problem for the $p \times p$ matrix $\boldsymbol{G}_k$, i.e. compute $\boldsymbol{Q}_k$ and $\boldsymbol{D}_k$ as indicated by (14).
d) $\boldsymbol{X}_k := \boldsymbol{Z}_k \boldsymbol{Q}_k \boldsymbol{D}_k^{-1}$.

Theorem 3. Under the same conditions as in Theorem 1, we have for the columns of the matrices $\boldsymbol{X}_k$ as generated by (15):

$$\| \boldsymbol{v}_j - \boldsymbol{x}_j^{(k)} \| = O(q_j^k), \quad \text{with} \quad q_j = \lambda_{p+1} / \lambda_j. \tag{16}$$

Proof. Taking the p vectors (11), each one divided by λ_j^k, as coordinate vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p$ in E_k, the eigendirections of the projected operator $\boldsymbol{A}^{-2}$ are $\sum_1^p t_j \boldsymbol{w}_j$, where the t_i are solutions of the eigenvalue problem

$$\sum_{j=1}^{p} b_{ij} t_j - \mu^{-2} \sum_{j=1}^{p} c_{ij} t_j = 0 \quad (i = 1, \ldots, p), \tag{17}$$

3. A different method to find such optimal directions in E_k was proposed in [2]. It uses first order approximations to the eigenvectors of the matrix $\boldsymbol{G}_k$, and thus saves some computing time compared with (15). However, if $p \ll n$, this saving is less important than that (15) immediately yields optimal directions. The latter is a must for the method described in § 5.

4. Cf. [4], p. 185/186, where the modification (15) is described without giving information about the convergence rate.

with

$$b_{ij} = \delta_{ij}/\lambda_i^2 + \sum_{r=1}^{p} y_{ri}\, y_{rj} \left(\lambda_r^2/(\lambda_i \lambda_j)\right)^k,$$
$$c_{ij} = \delta_{ij} + \sum_{r=1}^{p} x_{ri}\, x_{rj} \left(\lambda_r^2/(\lambda_i \lambda_j)\right)^k, \tag{18}$$

where $y_{ri} = x_{ri}/\lambda_r$. Eqs. (17) can therefore be rewritten as

$$[\mathrm{diag}\,(\lambda_\gamma^{-2} - \mu^{-2}) + \boldsymbol{E}]\, \boldsymbol{t} = 0, \tag{19}$$

the elements of $\boldsymbol{E}$ being of magnitude $e_{ij} = O(q_i^k\, q_j^k)$.

Assume now $\lambda_i = \lambda_{i+1} = \cdots = \lambda_h$ being a $h - i + 1$-fold eigenvalue of $\boldsymbol{A}$; then, as $k \to \infty$, $h - i + 1$ independent eigensolutions of (17) with $\mu \to \lambda_i$ exist. Every one of these is described by p values $t_1, \ldots, t_p$, which we assume normalized such that $t_i^2 + \cdots + t_h^2 = 1$. Then the t_j with $j \neq i, i+1, \ldots, h$ are of the order $O(q_i^k\, q_j^k)$. This means that the angle between $\sum_i^h t_j \boldsymbol{w}_j$ and $\sum_1^p t_j \boldsymbol{w}_j$ is of the order $O(q_p^k\, q_i^k) = o(q_i^k)$, while according to (11) the angle between $\sum_i^h t_j \boldsymbol{w}_j$ and the eigenspace to $\lambda_i, \lambda_{i+1}, \ldots, \lambda_h$ of $\boldsymbol{A}$ is of the order $O(q_i^k)$. This establishes Theorem 3.

The consequences of Theorem 3 are obvious: E.g. for a matrix with $\lambda_1 = 100$, $\lambda_2 = 99$, $\lambda_3 = 50$, $\lambda_4 = 10$, and with $p = 3$ the asymptotic convergence quotients for the j-th column of $\boldsymbol{X}_k$ are (stable convergence assumed):

	Process (1)	Process (15)
$j = 1$	0.99	0.1
$j = 2$	0.99	0.101
$j = 3$	0.505	0.2

All we need therefore for good convergence is to choose p such that if r eigenvectors are wanted, λ_{p+1}/λ_r is not too near to one.

An example speaking for itself is the matrix $64 - \boldsymbol{B}^3$, where

$b_{ij} :=$ **if** $i = j$ **then** 2
else if $abs(i - j) = 1$ **then** 1
else 0.

With $n = 17$, the dominant eigenvalues are in this case $\lambda_1 = 63.999971948\ldots$, $\lambda_2 = 63.99824531\ldots$. Thus the process (1) would yield a convergence rate 0.999973 for the eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2$, but with the process (15) and $p = 8$ the convergence rate for $\boldsymbol{v}_1, \boldsymbol{v}_2$ reduces to about 0.875 which means that roughly 120 steps are needed for 6 digit precision. This is due to the fact that $\lambda_9 = 56$.

4. Intermediate Steps

As powerful as a set of m consecutive steps (15) seem to be, the final space E_m is the same as with rule (1). Even more so, if we start with the same matrix $\boldsymbol{X}_k$, then $m - 1$ steps (1), followed by just one step (15) produce (aside from roundoff-

errors) the same matrix $\boldsymbol{X}_{h+m}$ as m steps (15) would; in both cases the columns are the eigendirections of the operator $\boldsymbol{A}^{-2}$ projected onto E_{h+m}. In view of the higher complexity of (15) it is important to know that nothing is lost on the final results, if (15) is applied only from time to time. However, an even greater saving of computing time can be achieved through a combination of $m-1$ steps (8) with one step (15). Indeed, even (8) produces the same spaces E_k, but in this case two precautions are advisable: *First*, after $m-1$ steps (8) the columns of $\boldsymbol{X}_{h+m-1}$ should be orthonormalized, and *second*, the number m must be limited since (8) tends to make the columns of $\boldsymbol{X}_{h+k}$ more and more parallel as $k \to \infty$. The proper m can be found by the following considerations:

As long as $(\lambda_1/\lambda_p)^{m-1} < 10$, parallelisation of the columns of $\boldsymbol{X}_{h+m-1}$ will not have gone further than that at most one decimal digit is cancelled out when these columns are orthonormalized.

λ_1, λ_p are of course not known; therefore the diagonal elements d_{11}, d_{pp} of the matrix $\boldsymbol{D}_h$ computed in the most recent step (15) are taken instead. A certain difficulty arises in that at the beginning of the iteration the diagonal elements d_{11}, d_{pp} are either not known (for $h=0$) or still far from the eigenvalues λ_1, λ_p. The latter involves the danger of a too big m; to avoid this, we start with a prefixed value of m which is later gradually increased, but not over $\log(10)/\log(d_{11}/d_{pp})$. This finally leads to the following computing rule:

a) Start with $h=0$, $m=2$.

b) Choose $\boldsymbol{X}_0$ with orthonormal columns.

c) Perform $m-1$ steps (8) upon $\boldsymbol{X}_h$.

d) Orthonormalize the columns of $\boldsymbol{X}_{h+m-1}$.

(20) e) Perform one step (15) with $k=h+m$.

f) **if** $(d_{11}/d_{pp})^m < 10$ **then** $m := m+1$;

g) test for termination (cf. §6 below).

h) $h := h+m$;

i) **goto** c;

A further improvement is possible by accelerating convergence via Chebyshev iteration. Indeed, the $m-1$ steps (20, c) are equivalent to

$$\boldsymbol{X}_{h+m-1} := \boldsymbol{A}^{m-1}\boldsymbol{X}_h.$$

This can be considered as a special case of

$$\boldsymbol{X}_{h+m-1} := P(\boldsymbol{A})\,\boldsymbol{X}_h, \tag{21}$$

where $P(x)$ is an arbitrary polynomial of degree $m-1$. If $P(x)$ is taken as the Chebyshev polynomial of degree $m-1$ for the interval $0 \leqq x \leqq 2e$, where e meets the condition $\lambda_{p+1} \leqq 2e < \lambda_p$, then convergence of (20) — especially if it was slow — is increased considerably.

(21) can be realized by using the three-term recurrence relation of the C-polynomials. In the present situation this leads to

$$\boldsymbol{X}_{h+1} := \frac{1}{e}\boldsymbol{A}\boldsymbol{X}_h - \boldsymbol{X}_h,$$

$$\boldsymbol{X}_{h+k} := \frac{2}{e}\boldsymbol{A}\boldsymbol{X}_{h+k-1} - 2\boldsymbol{X}_{h+k-1} - \boldsymbol{X}_{h+k-2}, \quad k = 2, 3, \ldots, m-1). \tag{22}$$

Of course, in this case m must be chosen different from (20, f), namely such that

$$m - 1 < \frac{\operatorname{ArCosh}(10)}{\operatorname{ArCosh}\left(\frac{d_{11}-e}{e}\right)}. \tag{23}$$

5. Countermeasures against Unstable Convergence

Unstable convergence means that the p columns of $\boldsymbol{X}_k$ generated by the process (20) do not converge to the eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ respectively, but to some other set of p eigenvectors. This is due to an unhappy choice of $\boldsymbol{X}_0$, namely such that the space E_0 is orthogonal to at least one of the said eigenvectors. In computing practice this is an illfated occurrence, but it is also very unlikely, since roundoff-errors usually turn unstable convergence into delayed stable convergence. However, also this is quite undesirable and should be prevented. The following considerations give a clue how this can be achieved:

After a step (15) has been performed, the last column of the matrix $\boldsymbol{X}_k$ has among all columns the biggest Rayleigh quotient with respect to the matrix $\boldsymbol{A}^{-2}$ and contains the most contributions from eigenvalues below λ_p (of $\boldsymbol{A}$). Little will be lost therefore, if this last column is replaced by a random vector which is subsequently orthonormalized against the first $p-1$ columns of $\boldsymbol{X}_k$. True, the last column is practically lost as an iteration vector in this way, but since by virtue of (16) convergence of this last column is the slowest of all columns, it would not have been very useful anyhow.

As a consequence, (20) is amended by an additional operation to be inserted after (20), e):

(24) Replace the p-th column of $\boldsymbol{X}_{h+m}$ by a random vector $\boldsymbol{y}$, and then orthonormalize $\boldsymbol{y}$ against the $p-1$ first columns of $\boldsymbol{X}_{h+m}$.

What is the effect of this? The columns of $\boldsymbol{X}_{h+m}$ are eigendirections of $\boldsymbol{A}^{-2}$ projected onto E_{h+m} with eigenvalues $\mu_1^{-2}, \mu_2^{-2}, \ldots, \mu_p^{-2}$. Replacing the last column of E_{h+m} by another vector creates a new space E'_{h+m} containing the $p-1$-dimensional subspace of E_{h+m} with projection eigenvalues $\mu_1^{-2}, \ldots, \mu_{p-1}^{-2}$. As a consequence of the minimax theorem, these latter values are interlaced between the eigenvalues of $\boldsymbol{A}^{-2}$ projected onto E'_{h+m}. Thus the $p-1$ smallest projection eigenvalues are at least not increased by the change from E_{h+m} to E'_{h+m} which means that convergence of the $p-1$ first columns of $\boldsymbol{X}_k$ is rather improved by the device (24).

A disadvantage is that not only λ_p and v_p can no longer be obtained by the amended process, but the last diagonal element d_{pp} computed in the next execution of (20, e) depends on the random vector $\boldsymbol{y}$ and is therefore itself random (but

never exceeding λ_p). Since this is undesirable, the maximum $\bar{d}_{pp}$ of all previously computed d_{pp}'s should be used in place of the current d_{pp}.

When using (24) in conjunction with Chebyshev iteration, it is best to use in (22), (23) the value $e=\bar{d}_{pp}/2$, since $\bar{d}_{pp}$ is the best lower bound to λ_p available in the computing process. The convergence rate for the j-th column of X_k is then approximately

$$\frac{d_{jj}-e}{e}-\sqrt{\left(\frac{d_{jj}-e}{e}\right)^2-1}. \tag{25}$$

For the example mentioned at the end of §3, we have $\lambda_1=63.99997\ldots$, $\lambda_8=59.48\ldots$. The convergence rate of the process (20), amended by (22), (23), (24), and with $p=8$, $e=29.74$ is therefore $q_1=0.58$ for the first column.

6. Termination of the Process

The most efficient computing process becomes doubtful, if it is not possible to determine the proper time for termination automatically.

Here, we have one clue: The first diagonal element of the matrix $\boldsymbol{D}_{k+m}$ which is computed in step (20, e), should theoretically increase while the iteration is proceeding. Thus, as soon as the first diagonal element stagnates (i.e. not increases from one step (20, e) to the next), the first eigenvalue is found within computer accuracy[5]. From then onwards the second diagonal element is tested, etc.

This device, however, is not sufficient for computing the eigenvectors, but can be used as a negative criterion: As long as the eigenvalues do not stagnate, it is obviously useless to test for eigenvectors. On the other hand, once λ_1 has been found, testing for $\boldsymbol{v}_1$ may be based on the following considerations:

Let $\boldsymbol{x}_1^{(k)}=\boldsymbol{x}_1$ be the first column of $\boldsymbol{X}_k$ and assume that the Rayleigh quotient of $\boldsymbol{x}_1$ already stagnates. Then

$$\boldsymbol{x}_1=c_0\boldsymbol{y}+\sum_{j=1}^{p}c_j\boldsymbol{v}_j, \tag{26}$$

where $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p$, $\boldsymbol{y}$ are an orthonormal system and $\sum_{j=0}^{p}c_j^2=1$. Now

$$\boldsymbol{A}\,x_1=\sum_{j=1}^{p}c_j\lambda_j\boldsymbol{v}_j+c_0\boldsymbol{A}\boldsymbol{y},\quad\text{with}\quad\|\boldsymbol{A}\boldsymbol{y}\|\leqq\lambda_{p+1},$$

hence

$$\boldsymbol{A}\boldsymbol{x}_1-\lambda_1\boldsymbol{x}_1=\sum_{j=1}^{p}c_j(\lambda_j-\lambda_1)\,\boldsymbol{v}_j+c_0(\boldsymbol{A}\boldsymbol{y}-\lambda_1\boldsymbol{y}).$$

This implies

$$\|\boldsymbol{A}\boldsymbol{x}_1-\lambda_1\boldsymbol{x}_1\|^2\geqq\sum_{j=2}^{p}c_j^2(\lambda_j-\lambda_1)^2+c_0^2(\lambda_1-\lambda_{p+1})^2,$$

or

$$|c_0|\leqq\frac{\|\boldsymbol{A}\boldsymbol{x}_1-\lambda_1\boldsymbol{x}_1\|}{\lambda_1-\lambda_{p+1}}. \tag{27}$$

5. This would be a fallacy in a slowly converging process, but in (20) m is finally so big that a full cycle of (20) should improve λ_1 by a factor 100. However, this is no longer true for the later eigenvalues, since d_{11} used in the criterion (20, f) is *not* replaced by d_{22}, etc. while testing for λ_2, etc.

Since c_0 essentially describes the angle between E_k and $\boldsymbol{v}_1$, and the eigenvalue problem within E_k is taken care of by (15), the inequality (27) furnishes the basis for a termination criterion, however only under the hypothesis, that unstable convergence is offset. In fact, (27) allows — every time after a step (20, e) has been performed — to make a test

$$\|\boldsymbol{x}_1 - \boldsymbol{v}_1\| < \varepsilon. \tag{28}$$

There is, however, still a serious problem: If ε is preset in the computing process, (28) may never be fulfilled because of rounding errors. In order to prevent such a disaster, the notion of *discounted error* is introduced: Theoretically the quantity $\|\boldsymbol{x}_1 - \boldsymbol{v}_1\|$ should for $k \to \infty$ converge linearly to zero with a convergence quotient $q_1 = \lambda_{p+1}/\lambda_1$. In view of this we define the discounted error as follows:

Let f_k be the right side of (27) after step k. If for the first time the value d_{11} stagnates, set $t_k = f_k$. Later define

$$t_k = \operatorname{Min}\{(d_{pp}/d_{11}) \times t_{k-1}, f_k\}, \tag{29}$$

where d_{pp} is taken as a substitute for λ_{p+1}. (A different discounting is needed for Chebyshev iteration.)

As soon as the discounted error undershoots the tolerance ε, the current $\boldsymbol{x}_1$ is taken as $\boldsymbol{v}_1$. Of course, this $\boldsymbol{x}_1$ does not actually meet condition (28), but is somehow the best possible result.

Having thus computed $\boldsymbol{v}_1$, the iteration is — with obvious modifications of the testing process — continued in order to find $\boldsymbol{v}_2$, etc. It should be recognized, however, that once $\boldsymbol{v}_1$ has been found, the first column of $\boldsymbol{X}_k$ is frozen, i.e. no longer subject to the iteration and orthogonalisation, but is still used to orthogonalize the later columns.

7. Conclusion

Numerical tests with all sorts of test matrices have been performed. They have confirmed that the combination (20), (22), (23), (24) is an efficient algorithm for computing dominant eigenvalues and corresponding eigenvectors of symmetric matrices. The method is particularly interesting for matrices of high order or matrices not available as a two-dimensional array of values. Indeed, all the process (20) needs, is a rule for computing the vector $\boldsymbol{A}\boldsymbol{x}$ from the vector $\boldsymbol{x}$. This latter fact allows computing the eigenvalues at the lower end of the spectrum via the rule for computing $\boldsymbol{A}^{-1}\boldsymbol{x}$ from $\boldsymbol{x}$.

The highest efficiency is obtained if a small value of p is sufficient to yield a good convergence rate. On the other hand, if p must be chosen higher as $n/5$, say, methods for computing all eigenvalues — e.g. a combination of Householder transformation and QR-method — are usually more efficient.

A striking example for what simultaneous iteration can achieve, is the matrix (of order 30)

$$\frac{\pi}{2}\boldsymbol{I} + \boldsymbol{A}, \quad \text{where} \quad \boldsymbol{A} = \{a_{ij} = 1/(1 + 2n - 2i - 2j)\} \quad (i, j = 1, \ldots, 30).$$

The ten highest eigenvalues are within 10-digit precision equal to 3.141592654: furthermore $\lambda_{11} = 3.141592649$, $\lambda_{12} = 3.141591787$.

Despite of this hopeless situation, a total of 90 iteration steps ((8) and (15) together) with $p=5$ were sufficient to produce 2 eigenvectors corresponding to $\lambda=\pi$.

References

1. BAUER, F. L.: Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme. ZAMP **8**, 214–235 (1957).
2. JENNINGS, A.: A direct method for computing latent roots and vectors of a symmetric matrix. Proc. Camb. Phil. Soc. **63**, 755–765 (1967).
3. RUTISHAUSER, H.: Solution of eigenvalue problems with the LR-transformation. Nat. Bur. Standards Appl. Math. Ser. **49**, 47–81. Washington D.C.: US. Gov. Printing Office 1958.
4. SCHWARZ, H. R.: Numerik symmetrischer Matrizen. 244 S. Stuttgart: Teubner 1968.

Professor Dr. H. RUTISHAUSER
Eidgenössische Technische Hochschule
Fachgruppe Computer-Wissenschaften
8006 Zürich, Schweiz
Leonhardstr. 33