On the Simultaneous Determination of Several Eigensolutions of a Self-Adjoint System of Differential Equations

By P. Laasonen

1. Abstract. Any finite number m of eigensolutions of a definitely self-adjoint system of ordinary differential equations may be approximated simultaneously to any desired accuracy by an iterative procedure and the solution of an $(m \times m)$-matrix eigenvalue equation. If the method is used with rounded numerical values, intermediate purification steps are found to be necessary at times; a rule for the estimation of roundoff errors is established. A simple example illustrates the theoretical argument.

2. Definitely self-adjoint problems. Let the nth order system of ordinary differential equations have the form

$$\frac{d}{dx}u(x) = (F(x) + \lambda G(x))u(x),$$

where u is an n-vector, F and $G(n \times n)$-matrices of continuous functions on the interval (a, b), and λ the eigenvalue parameter. The boundary conditions considered have the form

$$Au(a) + Bu(b) = 0,$$

where (A, B) is a constant $(n \times 2n)$-matrix of rank n. The eigenvalue problem thus formulated is called definitely self-adjoint by Bliss [1], [2], if the following conditions are fulfilled:

1°. There exists a non-singular matrix $T(x)$ such that

$$(1) \qquad \begin{cases} \dfrac{d}{dx}T + TF + F'T = 0 \\ TG + G'T = 0 \\ AT^{-1}(a)A' = BT^{-1}(b)B' \end{cases}$$

2°. The matrix $S(x) = T'(x)G(x)$ is symmetric and positive definite or semi-definite.

3°. If $\lambda = 0$ is an eigenvalue and $u_0(x)$ a corresponding eigenvector, then

$$u_0'(x)S(x)u_0(x) \not\equiv 0.$$

We will replace the last condition by the stronger one:

3'°. $\lambda = 0$ is not an eigenvalue.

Definitely self-adjoint problems thus defined have at most countably many eigenvalues $\lambda_i (i = 1, 2, \cdots)$, all real. In the following we assume $|\lambda_i| \leqq |\lambda_j|$, if $i < j$.

Received July 9, 1958. The preparation of this paper was sponsored by the Office of Naval Research.

The corresponding eigensolutions, denoted by $u_i(x)$, may be chosen to satisfy the orthonormalization condition

$$\int_a^b u_i'(x)S(x)u_j(x)\,dx = \delta_{ij}\,.$$

3. An iteration procedure. Let $V_0(x)$ be an $(n \times m)$-matrix whose columns are continuous and linearly independent on (a, b). Use $V_0(x)$ as the initial matrix for the sequence $V_i(x)$ determined successively as solutions of problems

$$\begin{cases} \dfrac{d}{dx}V_\alpha(x) = F(x)V_\alpha(x) + G(x)V_{\alpha-1}(x), \\ AV_\alpha(a) + BV_\alpha(b) = 0, \end{cases} \qquad (\alpha = 1, 2, \cdots)$$

and define matrices Q_β, for $0 \leqq \alpha \leqq \beta$, by

$$Q_\beta = \int_a^b V_\alpha'(x)S(x)V_{\beta-\alpha}(x)\,dx \qquad (\beta = 0, 1, 2, \cdots).$$

These generalizations of Schwarz's constants are symmetric and independent of the choice of the index α. In this analogy the eigenvalues of the matrix equations

$$(Q_{2\alpha-1} - \kappa Q_{2\alpha})t = 0 \tag{2}$$

correspond to the Schwarz's quotient. Denote the diagonal matrix of its eigenvalues by K_α and a matrix of corresponding eigenvectors by T_α ; hence

$$Q_{2\alpha-1}T_\alpha - Q_{2\alpha}T_\alpha K_\alpha = 0.$$

4. Theorem on convergence and its rate. Assume first the following conditions: (a) $V_0(x)$ is arbitrary in the sense that no linear combination of its columns is orthogonal to the first m eigensolutions $u_i(x)$ $(i = 1, 2, \cdots, m)$. (b) The first m eigenvalues are smaller in absolute value than all others, i.e. $|\lambda_m| < |\lambda_{m+1}|$. Then the sequence of the solutions of (2) provides successive approximations tending to the first m eigensolutions of the original problem:

If the eigenvalues of K_α are ordered and the vectors T_α normed properly, then the former ones converge to λ_i $(i = 1, 2, \cdots, m)$ and the column vectors of

$$V_\alpha(x)T_\alpha \tag{3}$$

to u_i $(i = 1, 2, \cdots, m)$ as $\alpha \to \infty$. *The rate of convergence is such that the* ith *eigenvalue as well as all components of the corresponding vector differ from their respective limits by amounts* $0(|\lambda_i/\lambda_{m+1}|^{2\alpha})$.

The convergence theorem has been proved under more general conditions in the author's papers [3] and [4]. There it was found for instance that if the condition about the generality of $V_0(x)$, expressed in (a), is not satisfied but some among the m first eigensolutions of the original problem are orthogonal to linear combinations of columns of $V_0(x)$, then convergence still prevails, the limiting eigenvalues and eigensolutions of the process are just replaced by some others, associated with the next smallest eigenvalues. Again, if condition (b) above is not fulfilled, then the convergence still occurs with respect to eigenvalues with smaller absolute value than $|\lambda_{m+1}|$ and to the corresponding eigensolutions.

The rate of convergence is not discussed in the previous papers; the error estimate is, however, obtainable from the proof in the latter paper by quite straightforward considerations.

5. Rounding off errors. In order to have an estimate of the effect of rounding off errors of the Q's, unavoidable in numerical computation with truncated values, differentiate the eigenvalue equation (2),

$$(dQ_{2\alpha-1} - \kappa dQ_{2\alpha})t - d\kappa Q_{2\alpha}t + (Q_{2\alpha-1} - \kappa Q_{2\alpha})\,dt = 0.$$

Multiplication from left by t' and division by

$$t'Q_{2\alpha-1}\,t = \kappa t'Q_{2\alpha}\,t \tag{4}$$

gives, observing the transposed form of (2), the final formula

$$\frac{d\kappa}{\kappa} = \frac{t'dQ_{2\alpha-1}\,t}{t'Q_{2\alpha-1}\,t} - \frac{t'dQ_{2\alpha}\,t}{t'Q_{2\alpha}\,t}. \tag{5}$$

Let Λ and $U(x)$ be the diagonal matrix of eigenvalues λ_i and the matrix of corresponding eigensolutions $u_i(x)$, respectively. Moreover, let C, defined by

$$C = \int_a^b U'(x)S(x)V_0(x)\,dx,$$

be the matrix of coefficients in the expansion of $V_0(x)$. Then any $V_\alpha(x)$ ($\alpha = 1, 2, \cdots$) has an absolutely and uniformly convergent expansion

$$V_\alpha(x) = U(x)\Lambda^{-\alpha}C.$$

The condition (a) above is equivalent with the statement that the top square submatrix C_1 of C, formed by its m first rows, is non-singular. Partition C and Λ as follows,

$$C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix},$$

where C_1 and Λ_1 are $(m \times m)$-matrices and use the abbreviation

$$p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} \Lambda_1^{-2\alpha}C_1\,t \\ \Lambda_2^{-2\alpha}C_2\,t \end{pmatrix}.$$

for the coefficient vector in the expansion of an approximation

$$V_{2\alpha}t = U\Lambda^{-2\alpha}Ct = Up.$$

In the proof, previously mentioned, it has been shown that when α increases indefinitely the coefficient vector p, provided the eigenvector t of (2) is chosen and normed properly, tends to a limit vector all of whose elements vanish with just one exception; this element, the ith, is among the first m and may be assumed to be 1. Moreover, the convergence is such that the deviations of the elements of p from their respective limits are of the magnitude $0(|\lambda_i/\lambda_{m+1}|^{2\alpha})$. Of course, the finite vector p_1 has the same properties. Hence, if $e_m^{(i)}$ is the notation of the m-vec-

tor whose ith element is 1, all other elements 0, then the vector p_1 at the αth stage has the expression

$$p_1 = e_m^{(i)} + |\lambda_i/\lambda_{m+1}|^{2\alpha} a_\alpha ,$$

where a_α is an m-vector bounded for $\alpha \to \infty$.

Express now t in terms of p_1 :

$$\begin{aligned} t &= C_1^{-1}\Lambda_1^{2\alpha} p_1 = C_1^{-1}[\lambda_i^{2\alpha} e_m^{(i)} + |\lambda_i/\lambda_{m+1}|^{2\alpha}\Lambda_1^{2\alpha} a_\alpha] \\ &= C_1^{-1}\lambda_i^{2\alpha}[e_m^{(i)} + |\lambda_m/\lambda_{m+1}|^{2\alpha} b_\alpha], \end{aligned} \tag{6}$$

where b_α is also an m-vector bounded for $\alpha \to \infty$. In order to estimate the right hand side terms in (5), i.e. quotients of the form

$$\frac{t' dQ_\beta t}{t' Q_\beta t}, \tag{7}$$

consider first their denominators. Q_β has an expression

$$Q_\beta = C'\Lambda^{-\beta} C = C_1'\Lambda_1^{-\beta} C_1 + C_2'\Lambda_2^{-\beta} C_2$$

and therefore, for the particular value $\beta = 2\alpha$,

$$\begin{aligned} t' Q_{2\alpha} t &= p_1'\Lambda_1^{2\alpha} C_1'^{-1}(C_1'\Lambda_1^{-2\alpha} C_1 + C_2'\Lambda_2^{-2\alpha} C_2) C_1^{-1}\Lambda_1^{2\alpha} p_1 \\ &= p_1'\Lambda_1^{2\alpha} p_1 + p_1'\Lambda_1^{2\alpha} C_1'^{-1}\Lambda_2^{-2\alpha} C_2 C_1^{-1}\Lambda_1^{2\alpha} p_1 \\ &= [e_m^{(i)'} + |\lambda_i/\lambda_{m+1}|^{2\alpha} a_\alpha']\Lambda_1^{2\alpha}[I + C_1'^{-1} C_2'\Lambda_2^{-2\alpha} C_2 C_1^{-1}\Lambda_1^{2\alpha}] \\ &\quad \times [e_m^{(i)} + |\lambda_i/\lambda_{m+1}|^{2\alpha} a_\alpha']. \end{aligned}$$

Since the second term in the second bracket decreases at least like $|\lambda_m/\lambda_{m+1}|^{2\alpha}$, the total product obviously has the asymptotic expression $\lambda_i^{2\alpha}$. By (4) one finds that the asymptotic expression for $t' Q_{2\alpha-1} t$ is $\lambda_i^{2\alpha+1}$.

Using (6), the numerator of (7), for $\beta = 2\alpha$, is clearly approximated by $\lambda_i^{4\alpha}$ multiplied by the ith diagonal term of $C_1'^{-1} dQ_{2\alpha} C_1$, as α increases. If C_1 is arbitrarily chosen, that is, if $V_0(x)$ is originally arbitrary, then the order of magnitude of this diagonal term is determined by the product of $\lambda_1^{-2\alpha}$, which determines the order of magnitude of all elements in $Q_{2\alpha}$, and a proper measure of the relative rounding off error in the elements of $Q_{2\alpha}$, say δ. Hence, for an arbitrarily chosen V_0 the quotient (7) has, for $\beta = 2\alpha$, the order of magnitude

$$\frac{\lambda_i^{4\alpha}\lambda_1^{-2\alpha}\delta}{\lambda_i^{2\alpha}} = \delta\left(\frac{\lambda_i}{\lambda_1}\right)^{2\alpha}.$$

The same estimate is, of course, also valid for $\beta = 2\alpha - 1$. For an increasing α the effect of the rounding off error in the approximations k_i for λ_i with an absolute value less than $|\lambda_1|$ will accordingly increase indefinitely even if the relative magnitude of the rounding off error should remain bounded.

On the other hand, if C_1 is almost diagonal, then the ith diagonal term of $C_1'^{-1} dQ_{2\alpha} C_1^{-1}$ has the order of magnitude $\lambda_i^{-2\alpha}\delta$ and the quotient (7) the order of magnitude

$$\frac{\lambda_i^{4\alpha}\lambda_i^{-2\alpha}\delta}{\lambda_i^{2\alpha}} = \delta.$$

In this case V_0 consists of columns already almost equal to the first m eigensolutions and the method gives all eigensolutions with an approximately equal accuracy.

If the condition (b) is not fulfilled, i.e. some of the first m eigenvalues are equal to some later eigenvalues in absolute value, then the above statements are still valid regarding those eigenvalues whose absolute values are smaller than $|\lambda_{m+1}|$.

6. Practical use of the method. After an initial matrix $V_0(x)$, i.e. any set of m linearly independent trial vectors $v_0^{(i)}(x)$, has been assigned, one has to solve m independent boundary value problems

$$(8) \qquad \begin{cases} \dfrac{d}{dx} v_1^{(i)}(x) = F(x)v_1^{(i)}(x) + G(x)v_0^{(i)}(x), \\ Av_1^{(i)}(a) + Bv_1^{(i)}(b) = 0; \end{cases} \qquad (i = 1, 2, \cdots, m).$$

The solution vectors $v_1^{(i)}(x)$, $i = 1, 2, \cdots, m$, form the matrix $V_1(x)$. Usually it is already advisable after this first step to improve the initial matrix for the next step. To this end the matrices Q_1 and Q_2 are formed from V_0 and V_1, the corresponding eigenvalue equation (2) solved and $V_1(x)$ replaced by the product $V_1(x)T_1$ from (3). It is true that the procedure gives theoretically completely equal results whether continued with $V_1(x)$ or $V_1(x)T_1$ as the next initial matrix. However, there may be a substantial difference in the practical results, due to the fact that the top square submatrix C_1 of the coefficient matrix C related to the latter is usually much closer to a diagonal matrix and the effect of rounding off errors accordingly decreased.

This effect may be estimated at any stage and for any eigenvalue by evaluating quotients (7) on the right hand side of (5). The denominator hereby may be computed directly after the eigenvector t has been determined; the order of magnitude of the numerator may be obtained by replacing dQ, for instance, by the positive diagonal matrix whose diagonal elements are the estimated positive rounding off errors of diagonal elements of Q.

If the fractions (7) are small, then, of course, several iteration steps like (8) may be performed without needing to interrupt the process for an intermediary purification step.

Finally it may be observed that if one wants to determine the first m eigensolutions, then it might be useful in some cases to carry out the computations with a larger number $m' > m$ and the excessive least accurate $m' - m$ solutions related to the largest eigenvalues may be omitted at the end.

7. Comparison with the customary method. If more than just the lowest eigenvalue is needed, then the customary iterative method computes the wanted eigenvalues successively in the order of increasing absolute value. Hence, the steps consisting of integrations of boundary value problems of type (8) are equal in the customary as well as in the described method. The orthogonalization steps which are necessary at times in the former method in connection with higher eigenvalues have their counterpart in the purification steps of the latter method. These intermediary steps, involving solution of auxiliary eigenvalue matrix equations, generate additional computation. The associated extra work is, however, mostly outweighted by following advantages.

If the approximate eigenvalues are interpreted as minimal values of functionals of Rayleigh-Ritz type, then the present method bases its trial functionals at each step on linear combinations of m vectors instead of one vector. Accordingly all results are obtained, after some number of iterations, with substantially greater accuracy than by the method based on the use of one vector for λ_1, two vectors for λ_2, etc.

Furthermore, separation of two or more eigensolutions associated with eigenvalues of almost equal absolute values is established automatically, whereas this case is always quite troublesome by methods which proceed with one vector at a time.

8. An example. To illustrate the described method consider the simplest possible problem:

$$F = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad G = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}, \qquad A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

The solution T of (1) and the corresponding S are easily found to be

$$T = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad S = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Moreover, the conditions 2° and 3'° are fulfilled. Hence the problem is definitely self-adjoint.

In order to find the first two eigensolutions take for instance $a = 0$, $b = 1$ and choose

$$V_0 = \begin{pmatrix} 1 & x \\ 0 & 0 \end{pmatrix}.$$

Solving (8) for the two corresponding column vectors results in

$$V_1 = \frac{1}{6}\begin{pmatrix} 6x - 3x^2 & 3x - x^3 \\ 6 - 6x & 3 - 3x^2 \end{pmatrix}$$

and the matrices

$$Q_1 = \frac{1}{120}\begin{pmatrix} 40 & 25 \\ 25 & 16 \end{pmatrix}, \qquad Q_2 = \frac{1}{5040}\begin{pmatrix} 672 & 427 \\ 427 & 272 \end{pmatrix}.$$

The eigenvalue equation (2) with $\alpha = 1$ gives now the eigenvalues,

$$\kappa = \frac{846 \mp 96\sqrt{51}}{65} = \begin{cases} 2.4680 \\ 23.563 \end{cases},$$

and the eigenvectors

$$T_1 = \begin{pmatrix} 2 & 2 \\ 4 + \sqrt{51} & 4 - \sqrt{51} \end{pmatrix}.$$

A second iteration step, based on the immediate use of V_1, gives the matrix

$$V_2 = \frac{1}{120}\begin{pmatrix} 40x - 20x^2 + 5x^4 & 25x - 10x^3 + x^5 \\ 40 - 60x^2 + 20x^3 & 25 - 30x^2 + 5x^4 \end{pmatrix}.$$

The corresponding matrices Q_3 and Q_4 are

$$Q_3 = \frac{1}{362880}\begin{pmatrix}19584 & 12465\\ 12465 & 7936\end{pmatrix}; \qquad Q_4 = \frac{1}{39916800}\begin{pmatrix}872960 & 555731\\ 555731 & 353792\end{pmatrix}$$

and the solutions of (2) with $\alpha = 2$:

$$\kappa = \frac{10496290 \mp 6400\sqrt{1725010}}{847269} = \begin{cases}2.46740208\\ 22.3093571\end{cases}$$

$$T_2 = \begin{pmatrix}2062 & 2062\\ 3328 + 5\sqrt{1725010} & 3328 - 5\sqrt{1725010}\end{pmatrix}$$

The relative errors of these approximations of λ_1 and λ_2 are $4 \cdot 10^{-7}$ and $4 \cdot 10^{-4}$, respectively. This accuracy is, however, possible only by using untruncated values of the elements of Q_3 and Q_4. This may be seen by computing the ratio (7) for instance with Q_4 in the denominator and $10^{-8} \cdot \mathrm{I}$ as the kernel dQ in the numerator. Inserting for t the first or the second column of T_2 one obtains $7 \cdot 10^{-7}$ or $5 \cdot 10^{-2}$, respectively. Since the elements of Q_4 are of the order of magnitude 10^{-2}, this result shows that truncation of the elements of Q_4 by 10^{-6} causes at the first eigenvalue an error of the same magnitude but at the second eigenvalue an error 10^{-1}. In fact, this result can also be obtained by a direct computation. If the elements of Q_3 and Q_4 are truncated to about 6-place values,

$$Q_3 = \begin{pmatrix}0.0539683 & 0.0343502\\ 0.0343502 & 0.0218695\end{pmatrix}, \qquad Q_4 = \begin{pmatrix}0.02186949 & 0.01392223\\ 0.01392223 & 0.00886324\end{pmatrix}$$

then the following eigenvalues are found for (2):

$$\kappa = \begin{cases}2.46740277\\ \\ 21.669\end{cases}$$

The error caused by this truncation into the first and second eigenvalue is hence at the eighth and at the second place, respectively.

If on the other hand the second iteration step is made by an improved matrix, taking

$$V_1^* = V_1\, T_1$$

$$= \frac{1}{6}\begin{pmatrix}24x - 6x^2 - 4x^3 + \sqrt{51}(3x - x^3) & 24x - 6x^2 - 4x^3 - \sqrt{51}(3x - x^3)\\ 24 - 12x - 12x^2 + 3\sqrt{51}(1 - x^2) & 24 - 12x - 12x^2 - 3\sqrt{51}(1 - x^2)\end{pmatrix}$$

as initial matrix, then the corresponding new vector matrix is

$$V_2^* = \frac{1}{120}\begin{pmatrix}180x - 80x^3 + 10x^4 + 4x^5 + \sqrt{51}(25x - 10x^3 + x^5)\\ 180x - 80x^3 + 10x^4 + 4x^5 - \sqrt{51}(25x - 10x^3 + x^5)\\ 180 - 240x^2 + 40x^3 + 20x^4 + 5\sqrt{51}(5 - 6x^2 + x^4)\\ 180 - 240x + 40x^2 + 20x^4 - 5\sqrt{51}(5 - 6x^2 + x^4)\end{pmatrix}.$$

Related kernel matrices are

$$Q_3^* = \frac{1}{90720}\begin{pmatrix} 202372 + 28337\sqrt{51} & 4 \\ 4 & 202372 - 28337\sqrt{51} \end{pmatrix},$$

$$Q_4^* = \frac{1}{3326400}\begin{pmatrix} 3007300 + 421105\sqrt{51} & 68 \\ 68 & 3007300 - 421105\sqrt{51} \end{pmatrix},$$

and the corresponding solution of (2):

$$\kappa = \frac{10496\,290 \mp 6400\sqrt{1725010}}{847269} = \begin{cases} 2.46740208 \\ 22.309 \end{cases}$$

$$T_2^* = \begin{pmatrix} 1 & -0.0000115 \\ 0.1212 & 1 \end{pmatrix}.$$

In this case the truncation of the numerical values of the elements of the Q's does not produce essentially greater error in the second eigenvalue. This may be seen by computing the characteristic quotients (7): $t'Q_4^*t$, by inserting the second column of T_2^*; this gives $\sim 3 \cdot 10^{-6}$ and, if dQ_4^* is taken to be the diagonal matrix containing diagonal elements of Q_4^* multiplied by 10^{-6}, then $t'dQ_4^*t$ is $\sim 3 \cdot 10^{-12}$, hence the order of magnitude of the relative error of k is 10^{-6}. And actually, if the elements of Q_3^* and Q_4^* are truncated to about 6-place values,

$$Q_3^* = \begin{pmatrix} 4.46140 & 4.40917 \cdot 10^{-5} \\ 4.40917 \cdot 10^{-5} & 5.88913 \cdot 10^{-5} \end{pmatrix}, \qquad Q_4^* = \begin{pmatrix} 1.808138 & 2.044252 \cdot 10^{-5} \\ 2.044252 \cdot 10^{-5} & 2.639971 \cdot 10^{-6} \end{pmatrix}$$

then the following eigenvalues for (2) are found to be:

$$\kappa = \begin{cases} 2.4673997. \\ 22.3093475 \end{cases}$$

The relative errors due to the truncation of the elements of the matrices are hence 10^{-6} and $\frac{1}{2} \cdot 10^{-6}$, respectively.

University of California, Los Angeles

1. G. A. Bliss, "A boundary value problem for a system of ordinary differential equations of first order," *Transactions Amer. Math. Soc.* v. 28, 1926 p. 561–584.

2. G. A. Bliss, "Definitely self-adjoint boundary value problems," *Transactions Amer. Math. Soc.* v. 44, 1938 p. 413–428.

3. P. Laasonen, "Ein Problem bei der iterativen Bestimmung der Eigenwerte simultaner Differentialgleichungen," *Ann. Acad. Sci. Fenn.* A I 195, 1955 p. 1-13.

4. P. Laasonen, "Bemerkung zur iterativen Lösung der Eigenwertaufgabe einer Vektordifferentialgleichung," *Ann. Acad. Sci. Fenn.* A I 230, 1956 p. 1–8.