A new method for the solution of eigenvalue problems

By M. R. Osborne*

One among the class of iterations proposed by Osborne and Michaelson (1964) is discussed. It is shown that for the ordinary eigenvalue problem this iteration is of third order, and that it may readily be modified to be of third order for the general eigenvalue problem provided the eigenvalue parameter occurs linearly. The analysis also suggests a further modification which gives a third-order iteration for the non-linear problem.

1. Introduction

In a recent paper (Osborne and Michaelson (1964)) the author has discussed the solution of eigenvalue problems having the form

$$\boldsymbol{M}(\lambda)\boldsymbol{v} = 0. \tag{1.1}$$

Here the eigenvalues are found as the zeros of a function $\beta(\lambda)$ defined by imposing a scaling condition on the solution vector $\boldsymbol{v}$ in the equation

$$\boldsymbol{M}(\lambda)\boldsymbol{v} = \beta(\lambda)\boldsymbol{x}. \tag{1.2}$$

By scaling $\boldsymbol{v}$ so that one component is fixed in value independent of λ, and by applying Newton's method to the resulting function $\beta(\lambda)$, Osborne and Michaelson derive the iteration

$$\left.\begin{aligned} \boldsymbol{M}(\lambda_i)\boldsymbol{v}_{i+1} &= \boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i}, \\ \boldsymbol{M}(\lambda_i)\boldsymbol{x}_{i+1} &= \frac{d\boldsymbol{M}}{d\lambda}(\lambda_i)\boldsymbol{v}_{i+1} \\ \lambda_{i+1} &= \lambda_i - \frac{(\boldsymbol{v}_{i+1})_{p_i+1}}{(\boldsymbol{x}_{i+1})_{p_i+1}} \end{aligned}\right\} \tag{1.3}$$

where p_i is the index of the component of maximum modulus in $\boldsymbol{x}_i$. In this paper iterations similar to (1.3) are discussed with a view to determining their order. (We say an iteration is of order p if δ_{j+1} is proportional to δ_j^p where δ_j is a parameter giving a measure of the error at the jth stage of the iteration.)

In the next section the ordinary eigenvalue problem is discussed $\left(\frac{d\boldsymbol{M}}{d\lambda} = -I\right)$. In this case the iteration of Osborne and Michaelson is of third order. Conditions sufficient to ensure the convergence of the iteration are also given. In Section 3 the case $\frac{d\boldsymbol{M}}{d\lambda} = -\boldsymbol{B}$, where $\boldsymbol{B}$ is a constant matrix is discussed. A modification of the basic iteration (1.3) is introduced and shown to be of third order. The iteration (1.3) is second order in this case. In Section 4, the case where M depends nonlinearly on λ is considered. Here both the iteration (1.3) and the modified iteration are of second order in general. However, the analysis suggests how the iteration might be modified to give third-order convergence, and it also suggests a possible second-order iteration which would require the solution of only one set of linear equations per step.

In the case in which $\boldsymbol{M}$ depends linearly on λ the iteration (1.3) amounts to two stages of inverse iteration followed by a shift in the origin of λ. It is of considerable interest that a third-order process can be obtained merely by carrying out two inverse iterations, keeping the same λ (requiring one triangular factorization and two forward and back substitutions), as inverse iteration is anyway recommended for the calculation of eigenvectors (Wilkinson (1958)).

It must be stressed that the iterations discussed in this paper apply only for simple eigenvalues, and must be modified to deal with repeated eigenvalues.

2. The ordinary eigenvalue problem

In this section it is shown that the iteration

$$\left.\begin{aligned} [\boldsymbol{A} - \lambda_i\boldsymbol{I}]\boldsymbol{v}_{i+1} &= \boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i}, \\ [\boldsymbol{A} - \lambda_i\boldsymbol{I}]\boldsymbol{x}_{i+1} &= -\boldsymbol{v}_{i+1}, \\ \lambda_{i+1} &= \lambda_i - \frac{(\boldsymbol{v}_{i+1})_{p_i+1}}{(\boldsymbol{x}_{i+1})_{p_i+1}} \end{aligned}\right\} \tag{2.1}$$

is of third order. It is assumed that the eigenvalue being determined is distinct. Let this eigenvalue be μ_1, then the precise assumption made is that

$$\min_j (|\mu_j - \mu_1|, |\mu_j - \lambda_i|) \geqslant a \tag{2.2}$$

where the index j runs over the other eigenvalues of $\boldsymbol{A}$. The need for the introduction of the parameter a, and its numerical significance, are discussed in Wilkinson (1961). Initially it is also assumed that the matrix $\boldsymbol{A}$ has a complete set of eigenvectors $\boldsymbol{u}_i$ scaled so that their maximum norm is 1. Whenever norms are used in this paper the maximum norm is assumed.

Further, it is assumed that $\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i}$ is an approximation to $\boldsymbol{u}_1$ in the sense that

$$\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i} = \alpha_1^i\boldsymbol{u}_1 + \boldsymbol{q}_i, \tag{2.3}$$

* *Edinburgh University Computer Unit, 7, Buccleuch Place, Edinburgh 8.*

where $$\boldsymbol{q}_i = \sum_{j=2}^{n} \alpha_j^i \boldsymbol{u}_j, \tag{2.4}$$

and $$||\boldsymbol{q}_i|| \leqslant \sum_{j=2}^{n} |\alpha_j^i| \leqslant \delta_i, \tag{2.5}$$

and that λ_i is an approximation to μ_1 in the sense that

$$\lambda_i = \mu_1 + \delta_i \eta_i. \tag{2.6}$$

It is assumed that the parameter δ is small, and that the parameter η is of the same order of magnitude as a.

Theorem. If $\delta_i < \frac{1}{3}$, and $|\eta_i| < 1{\cdot}5a$, then for any $j \geqslant i$

$$\delta_{j+1} < 5{\cdot}4\delta_j^3,$$
$$|\eta_{j+1}| < 1{\cdot}5a,$$

and the iteration is convergent.

Proof. From equation (2.1) it follows that

$$\boldsymbol{v}_{i+1} = -\frac{\alpha_1^i}{\delta_i \eta_i}\boldsymbol{u}_1 + \boldsymbol{r}_{i+1} \tag{2.7}$$

where $$\boldsymbol{r}_{i+1} = \sum_{j=2}^{n} \frac{\alpha_j^i}{\mu_j - \lambda_i}\boldsymbol{u}_j, \tag{2.8}$$

and $$||\boldsymbol{r}_{i+1}|| \leqslant \delta_i/a. \tag{2.9}$$

Also $$\boldsymbol{x}_{i+1} = -\frac{\alpha_1^i}{(\delta_i\eta_i)^2}\boldsymbol{u}_1 + \boldsymbol{s}_{i+1}, \tag{2.10}$$

where $$\boldsymbol{s}_{i+1} = -\sum_{j=2}^{n} \frac{\alpha_j^i}{(\mu_j - \lambda_i)^2}\boldsymbol{u}_j, \tag{2.11}$$

and $$||\boldsymbol{s}_{i+1}|| \leqslant \delta_i/a^2. \tag{2.12}$$

Now, by equations (2.1) and (2.6),

$$\mu_1 - \lambda_{i+1} = \mu_1 - \lambda_i + \frac{\frac{\alpha_1^i}{\delta_i\eta_i}(\boldsymbol{u}_1)_{p_{i+1}} - (\boldsymbol{r}_{i+1})_{p_{i+1}}}{\frac{\alpha_1^i}{(\delta_i\eta_i)^2}(\boldsymbol{u}_1)_{p_{i+1}} - (\boldsymbol{s}_{i+1})_{p_{i+1}}}$$

$$= \delta_i\eta_i \frac{(\delta_i\eta_i)^2(\boldsymbol{s}_{i+1})_{p_{i+1}} - (\delta_i\eta_i)(\boldsymbol{r}_{i+1})_{p_{i+1}}}{\alpha_1^i(\boldsymbol{u}_1)_{p_{i+1}} - (\delta_i\eta_i)^2(\boldsymbol{s}_{i+1})_{p_{i+1}}}. \tag{2.13}$$

Also, by equations (2.3) and (2.5)

$$1 - \delta_i \leqslant |\alpha_1^i| \leqslant 1 + \delta_i,$$

and by equations (2.10) and (2.12),

$$\left.\begin{aligned} |\alpha_1^i| - \frac{\delta_i^3\eta_i^2}{a^2} &\leqslant |\alpha_1^i(\boldsymbol{u}_1)_{p_{i+1}}| + \frac{\delta_i^3\eta_i^2}{a^2} \\ \text{so that } \quad |\alpha_1^i(\boldsymbol{u}_1)_{p_{i+1}}| &\geqslant 1 - \delta_i - 2\delta_i^3\eta_i^2/a^2 \end{aligned}\right\}. \tag{2.14}$$

(It is readily verified that the numerical values of δ_i and $|\eta_i/a|$ stated in the theorem permit these inequalities to be satisfied with positive values on both sides of the inequality).

By taking moduli in equation (2.13), and using equations (2.9), (2.12), and (2.14) there is obtained

$$\delta_{i+1}|\eta_{i+1}| \leqslant \frac{\delta_i^3\eta_i^2}{a}\,\frac{1 + \delta_i|\eta_i/a|}{1 - \delta_i - 3\delta_i^3\eta_i^2/a^2}. \tag{2.15}$$

Writing γ_i for $|\eta_i/a|$, and dividing both sides of equation (2.16) by a gives

$$\delta_{i+1}\gamma_{i+1} \leqslant \delta_i^3\gamma_i^2\frac{1 + \delta_i\gamma_i}{1 - \delta_i(1 + 3\gamma_i^2\delta_i^2)}. \tag{2.16}$$

Further $$\boldsymbol{x}_{i+1}/(\boldsymbol{x}_{i+1})_{p_{i+1}} = \alpha_1^{i+1}\boldsymbol{u}_1 + \boldsymbol{q}_{i+1}$$

where, by equations (2.10), (2.12), and (2.14)

$$||\boldsymbol{q}_{i+1}|| \leqslant \delta_{i+1} = \frac{\delta_i^3\gamma_i^2}{1 - \delta_i(1 + 3\gamma_i^2\delta_i^2)}. \tag{2.17}$$

(Clearly, by the manner of its construction, the right-hand side of equation (2.17) provides a bound for $\sum_{j=2}^{n} |\alpha_j^{i+1}|$ so that it can be used to define δ_{i+1}). From equations (2.16) and (2.17) it is seen that

$$\delta_{i+1}\gamma_{i+1} \leqslant \delta_{i+1}(1 + \delta_i\gamma_i)$$

whence $$\gamma_{i+1} \leqslant 1 + \delta_i\gamma_i. \tag{2.18}$$

Substituting the numerical values of bounds for δ_i and γ_i (noting that $\gamma_i \leqslant 1{\cdot}5$ for all values of a) shows that

(*a*) from equation (2.18), $\gamma_{i+1} \leqslant 1{\cdot}5$

(*b*) from equation (2.17), $\delta_{i+1} \leqslant 1/5 < 1/3$.

Therefore the conditions of the theorem are satisfied for $j = i + 1$ (and hence for all $j > i$) if they hold for $j = i$. The first part of the theorem is now obtained by inserting bounds into the coefficient of δ_i^3 in equation (2.17), while the second part follows immediately from (*a*) above. A direct computation using equation (2.15) gives

$$\delta_{i+1}|\eta_{i+1}| \leqslant 0{\cdot}6\delta_i|\eta_i| \tag{2.19}$$

and as the bounds used in obtaining this inequality are valid for all $j \geqslant i$, the inequality is also valid for all $j \geqslant i$. This, plus the result that $\delta_{j+1} < 5{\cdot}4\delta_j^3$, shows that the iteration converges.

In the case where the matrix $\boldsymbol{A}$ has principal vectors of grades > 1 the argument given above requires some modification. The crucial steps in the argument involve the bounding of $||\boldsymbol{r}_{i+1}||$ and $||\boldsymbol{s}_{i+1}||$ (equations (2.9) and (2.12)). Assume that $\boldsymbol{u}_p$ is a principal vector of grade two associated with the eigenvalue μ_r. Then

$$(\boldsymbol{A} - \mu_r\boldsymbol{I})^2\boldsymbol{u}_p = 0,$$

whence

$$(\boldsymbol{A} - \lambda_i\boldsymbol{I})\boldsymbol{u}_p + 2(\mu_r - \lambda_i)\boldsymbol{u}_p + (\mu_r - \lambda_i)^2(\boldsymbol{A} - \lambda_i\boldsymbol{I})^{-1}\boldsymbol{u}_p = 0. \tag{2.20}$$

Taking norms in this equation gives

$$||(\boldsymbol{A} - \lambda_i\boldsymbol{I})^{-1}\boldsymbol{u}_p|| \leqslant \frac{2||\boldsymbol{u}_p||}{a} + \frac{||\boldsymbol{A} - \lambda_i\boldsymbol{I}||\,||\boldsymbol{u}_p||}{a^2} \tag{2.21}$$

and

$$||(\boldsymbol{A} - \lambda_i \boldsymbol{I})^{-2}\boldsymbol{u}_p|| \leqslant \frac{2}{a}||(\boldsymbol{A} - \lambda_i \boldsymbol{I})^{-1}\boldsymbol{u}_p|| + \frac{||\boldsymbol{u}_p||}{a^2}. \quad (2.22)$$

The most significant feature of equations (2.21) and (2.22) is the occurrence of terms in $1/a^2$ and $1/a^3$ respectively. It is readily verified that the corresponding inequalities for a principal vector of grade m involve inverse powers of a of orders up to the mth and $(m+1)$th respectively. A possible corollary of this is that the numerical difficulties usually experienced when a is necessarily small are intensified if the close eigenvalue has associated with it principal vectors of grades higher than one.

3. A more general eigenvalue problem

In this section is considered the eigenvalue problem

$$[\boldsymbol{A} - \lambda \boldsymbol{B}]\boldsymbol{v} = 0. \quad (3.1)$$

If the matrix $\boldsymbol{B}$ has an inverse then the problem stated in equation (3.1) can be reduced to that considered in Section 2 by multiplying by $\boldsymbol{B}^{-1}$. The iteration (2.1) applied to the matrix $\boldsymbol{B}^{-1}\boldsymbol{A}$ is equivalent to the iteration

$$\left.\begin{aligned} [\boldsymbol{A} - \lambda_i \boldsymbol{B}]\boldsymbol{v}_{i+1} &= \boldsymbol{B}\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i} \\ [\boldsymbol{A} - \lambda_i \boldsymbol{B}]\boldsymbol{x}_{i+1} &= -\boldsymbol{B}\boldsymbol{v}_{i+1} \\ \lambda_{i+1} &= \lambda_i - \frac{(\boldsymbol{v}_{i+1})_{p_{i+1}}}{(\boldsymbol{x}_{i+1})_{p_{i+1}}} \end{aligned}\right\}. \quad (3.2)$$

This iteration differs significantly from that proposed by Osborne and Michaelson in the definition of $\boldsymbol{v}_{i+1}$. It will be shown to be of third order even when the matrix $\boldsymbol{B}$ does not have an inverse. The iteration proposed by Osborne and Michaelson gives only second order convergence for the problem (3.1).

First the solution of equation (3.3) is considered

$$[\boldsymbol{A} - \lambda_i \boldsymbol{B}]\boldsymbol{y} = \boldsymbol{x} \quad (3.3)$$

where $\lambda_i = \lambda + \delta_i$, and λ is an eigenvalue of equation (3.1), so that 0 is an eigenvalue of the matrix $[\boldsymbol{A} - \lambda \boldsymbol{B}]$. It is assumed that the other eigenvalues of the matrix $[\boldsymbol{A} - \lambda \boldsymbol{B}]$ are $\mu_2, \ldots, \mu_n$, that $\boldsymbol{u}_1$ is the eigenvector corresponding to the zero eigenvalue, and that the μ_i, $i = 2, \ldots, n$, satisfy the condition (2.2). The similarity normal form of $[\boldsymbol{A} - \lambda \boldsymbol{B}]$ is written $\boldsymbol{T}\boldsymbol{M}\boldsymbol{T}^{-1}$, where $\boldsymbol{M}_{11} = 0$. The vectors $\boldsymbol{y}$ and $\boldsymbol{x}$ have representations which are written

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{\alpha}, \; \boldsymbol{x} = \boldsymbol{T}\boldsymbol{\beta}. \quad (3.4)$$

In this notation equation (3.3) becomes

$$\begin{aligned} [\boldsymbol{A} - \lambda_i \boldsymbol{B}]\boldsymbol{y} &= [\boldsymbol{A} - \lambda \boldsymbol{B}]\boldsymbol{y} - \delta_i \boldsymbol{B}\boldsymbol{y} \\ &= \boldsymbol{T}\boldsymbol{M}\boldsymbol{\alpha} - \delta_i \boldsymbol{B}\boldsymbol{T}\boldsymbol{\alpha} \\ &= \boldsymbol{T}\boldsymbol{\beta} \end{aligned}$$

so that, defining $\boldsymbol{G}$ to be the matrix $-\boldsymbol{T}^{-1}\boldsymbol{B}\boldsymbol{T}$,

$$[\boldsymbol{M} + \delta_i \boldsymbol{G}]\boldsymbol{\alpha} = \boldsymbol{\beta}. \quad (3.5)$$

Equation (3.5) is represented in the partitioned form

$$\begin{bmatrix} \delta_i G_{11} & \delta_i \boldsymbol{g}^T \\ \delta_i \boldsymbol{f} & \boldsymbol{K} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \boldsymbol{\xi} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \boldsymbol{\eta} \end{bmatrix}. \quad (3.6)$$

If δ_i is small enough then (by the condition (2.2)) $\boldsymbol{K}$ possesses an inverse, and equation (3.6) gives

$$\left.\begin{aligned} \boldsymbol{\xi} &= \boldsymbol{K}^{-1}(\boldsymbol{\eta} - \alpha_1 \delta_i \boldsymbol{f}) \\ \delta_i G_{11}\alpha_1 + \delta_i \boldsymbol{g}^T \boldsymbol{K}^{-1}(\boldsymbol{\eta} - \alpha_1 \delta_i \boldsymbol{f}) &= \beta_1 \end{aligned}\right\} \quad (3.7)$$

whence

$$\left.\begin{aligned} \alpha_1 &= \frac{\beta_1}{\delta_i G_{11}} + \left(\frac{\beta_1}{G_{11}^2}\boldsymbol{g}^T\boldsymbol{K}^{-1}\boldsymbol{f} - \frac{\boldsymbol{g}^T\boldsymbol{K}^{-1}\boldsymbol{\eta}}{G_{11}}\right) + \mathrm{O}(\delta_i), \\ \text{and} \\ \boldsymbol{\xi} &= \boldsymbol{K}^{-1}\boldsymbol{\eta} - \frac{\beta_1}{G_{11}}\boldsymbol{f} + \mathrm{O}(\delta_i). \end{aligned}\right\} \quad (3.8)$$

Notes.—(1) Because of the term $\frac{\beta_1}{G_{11}}\boldsymbol{f}$, $\boldsymbol{\xi}$ can be O(1) even when η is $\mathrm{O}(\delta_i)$ so that the projection of the vector $\boldsymbol{x}$ on vectors other than the appropriate eigenvector is small. It is this that causes the iteration of Osborne and Michaelson to be of second order only. In particular, for this iteration, equation (3.12) does not hold.

(2) It is the matrix $\boldsymbol{K}^{-1}$ that contains the terms $\mathrm{O}(1/a)$ when the spacing between the eigenvalues is small.

(3) If $G_{11} = 0$ then, by equation (3.7), $\alpha_1 = \mathrm{O}(1/\delta_i^2)$. In this case it is readily seen that the iteration (3.2) cannot be satisfactory if indeed it converges at all.

Now let $\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i}$ be an approximation to $\boldsymbol{u}_1$ in the sense that

$$\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i} = a_i \boldsymbol{u}_1 + \delta_i \boldsymbol{q}_i \quad (3.9)$$

where a_i and $||\boldsymbol{q}_i||$ are O(1). From equation (3.1)

$$[\boldsymbol{A} - (\lambda_i - \delta_i)\boldsymbol{B}](\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i} - \delta_i \boldsymbol{q}_i) = 0$$

whence

$$[\boldsymbol{A} - \lambda_i \boldsymbol{B}]a_i \boldsymbol{u}_1 = -\delta_i \boldsymbol{B}\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i} + \delta_i^2 \boldsymbol{B}\boldsymbol{q}_i. \quad (3.10)$$

Also, from equation (3.2),

$$[\boldsymbol{A} - \lambda_i \boldsymbol{B}]\delta_i \boldsymbol{v}_{i+1} = \delta_i \boldsymbol{B}\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i}$$

so that

$$[\boldsymbol{A} - \lambda_i \boldsymbol{B}](\delta_i \boldsymbol{v}_{i+1} + a_i \boldsymbol{u}_1) = \delta_i^2 \boldsymbol{B}\boldsymbol{q}_i. \quad (3.11)$$

Using equation (3.8) with $\boldsymbol{\beta} = \boldsymbol{T}^{-1}\boldsymbol{B}\boldsymbol{q}_i$ gives

$$\delta_i \boldsymbol{v}_{i+1} + a_1 \boldsymbol{u}_1 = \left[\delta_i \frac{\beta_1}{G_{11}} + \mathrm{O}(\delta_i^2)\right]\boldsymbol{u}_1 + \boldsymbol{r}_{i+1} \quad (3.12)$$

where $||\boldsymbol{r}_{i+1}|| = \mathrm{O}(\delta_i^2)$.

Thus $$\boldsymbol{v}_{i+1} = a_i^* \boldsymbol{u}_1 + \delta_i \boldsymbol{r}_{i+1}^* \quad (3.13)$$

where $a_i^* = \mathrm{O}(1/\delta_i)$, and $||\boldsymbol{r}_{i+1}^*|| = \mathrm{O}(1)$.

The argument used for deriving $\boldsymbol{v}_{i+1}$ can be used again to derive $\boldsymbol{x}_{i+1}$. Defining $\boldsymbol{\gamma}$ to be $\boldsymbol{T}^{-1}\boldsymbol{B}\boldsymbol{r}_{i+1}^*$, then

$$-\delta_i \boldsymbol{x}_{i+1} + a_i^* \boldsymbol{u}_1 = \left[\delta_i \frac{\gamma_1}{G_{11}} + \mathrm{O}(\delta_i^2)\right]\boldsymbol{u}_1 + \boldsymbol{s}_{i+1} \quad (3.14)$$

where $$||\boldsymbol{s}_{i+1}|| = O(\delta_i^2). \tag{3.15}$$

Define $\boldsymbol{s}^*_{i+1}$ to be the vector $\delta_i^{-2}\boldsymbol{s}_{i+1}$.

The error in the next approximation to the eigenvalue can now be estimated. From equation (3.2)

$$\begin{aligned}\lambda_{i+1} &= \lambda + \delta_{i+1}\\ &= \lambda + \delta_i - \\ &\quad - \frac{a_i^*(\boldsymbol{u}_1)_{p_{i+1}} + \delta_i(\boldsymbol{v}^*_{i+1})_{p_{i+1}}}{\frac{a_i^*}{\delta_i}\left[1 - \frac{\delta_i}{a_i^*}\frac{\gamma_1}{G_{11}} + O(\delta_i^2)\right](\boldsymbol{u}_1)_{p_{i+1}} - \delta_i(\boldsymbol{s}_{i+1})_{p_{i+1}}}\end{aligned}$$

where the fact that $a_i^* = O(1/\delta_i)$ has been used,

$$= \lambda + O(\delta_i^3). \tag{3.16}$$

It follows from equations (3.14), (3.15) and (3.16) that the iteration (3.2) is of third order.

4. The non-linear problem

The general form of the eigenvalue problem is

$$\boldsymbol{M}(\lambda)\boldsymbol{u}_1 = 0. \tag{4.1}$$

Iterative methods for the solution of this problem are considered in this section. The notation of Section 3 is followed here.

Expanding equation (4.1) by Taylor's theorem gives

$$\left\{\boldsymbol{M}(\lambda_i) - \delta_i\frac{d\boldsymbol{M}}{d\lambda}(\lambda_i) + \frac{\delta_i^2}{2}\frac{d^2\boldsymbol{M}}{d\lambda^2}(\check{\lambda})\right\}(\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i} - \delta_i\boldsymbol{q}_i) = 0,$$

where $\check{\lambda}$ is a mean value between λ and λ_i, so that

$$\begin{aligned}\boldsymbol{M}(\lambda_i)a_i\boldsymbol{u}_1 &= \delta_i\frac{d\boldsymbol{M}}{d\lambda}(\check{\lambda}_i)\boldsymbol{x}_i/(\boldsymbol{x}_i)_{pi}\\ &\quad - \delta_i^2\left\{\frac{d\boldsymbol{M}}{d\lambda}(\lambda_i)\boldsymbol{q}_i + \tfrac{1}{2}\frac{d^2\boldsymbol{M}}{d\lambda^2}(\check{\lambda})(\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i} - \delta_i\boldsymbol{q}_i)\right\}.\end{aligned} \tag{4.2}$$

In the first stage of the iteration a vector $\boldsymbol{v}_{i+1}$ is calculated from

$$\boldsymbol{M}(\lambda_i)\delta_i\boldsymbol{v}_{i+1} = \delta_i\frac{d\boldsymbol{M}}{d\lambda}(\lambda_i)\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i} \tag{4.3}$$

so that

$$\begin{aligned}\boldsymbol{M}(\lambda_i)(\delta_i\boldsymbol{v}_{i+1} - a_i\boldsymbol{u}_1) &= \frac{\delta_i^2}{2}\frac{d^2\boldsymbol{M}}{d\lambda^2}(\check{\lambda})\boldsymbol{x}_i/(\boldsymbol{x}_i)_{pi}\\ &\quad + \delta_i^2\left\{\frac{d\boldsymbol{M}}{d\lambda}(\lambda_i) - \frac{\delta}{2}\frac{d^2\boldsymbol{M}}{d\lambda^2}(\check{\lambda})\right\}\boldsymbol{q}_i\\ &= \delta_i^2\boldsymbol{y}_i.\end{aligned} \tag{4.4}$$

Using equation (3.8) with $\boldsymbol{B} = \frac{d\boldsymbol{M}}{d\lambda}(\lambda_i) - \frac{\delta_i}{2}\frac{d^2\boldsymbol{M}}{d\lambda^2}(\check{\lambda})$ and $\boldsymbol{\beta} = \boldsymbol{T}^{-1}\boldsymbol{y}_i$ gives

$$\delta_i\boldsymbol{v}_{i+1} - a_i\boldsymbol{u}_1 = \left\{\frac{\delta_i\beta_1}{G_{11}} + O(\delta_i^2)\right\}\boldsymbol{u}_1 + \boldsymbol{r}_{i+1}, \tag{4.5}$$

where $$||\boldsymbol{r}_{i+1}|| = O(\delta_i^2). \tag{4.6}$$

Thus $$\boldsymbol{v}_{i+1} = a_i^*\boldsymbol{u}_1 + \delta_i\boldsymbol{r}^*_{i+1} \tag{4.7}$$

where $a_i^* = O(1/\delta_i)$, and $||\boldsymbol{r}^*_{i+1}|| = O(1)$.

The vector $\boldsymbol{x}_{i+1}$ is now calculated using equation (1.3). The previous argument gives

$$\begin{aligned}\boldsymbol{M}(\lambda_i)(\delta_i\boldsymbol{x}_{i+1} - a_i^*\boldsymbol{u}_1) &= \frac{\delta_i^2}{2}\frac{d^2\boldsymbol{M}}{d\lambda^2}(\check{\lambda})\boldsymbol{v}_{i+1}\\ &\quad + \delta_i^2\left\{\frac{d\boldsymbol{M}}{d\lambda}(\lambda_i) - \frac{\delta_i}{2}\frac{d^2\boldsymbol{M}}{d\lambda^2}(\check{\lambda})\right\}\boldsymbol{r}^*_{i+1}\\ &= \delta_i\boldsymbol{z}_i.\end{aligned} \tag{4.8}$$

The right-hand side of equation (4.8) is only $O(\delta_i)$ because of the occurrence of the term involving $\boldsymbol{v}_{i+1}$. Again applying equation (3.8) (with $\boldsymbol{\beta} = \boldsymbol{T}^{-1}\boldsymbol{z}_i$) gives

$$\delta_i\boldsymbol{x}_{i+1} - a_i^*\boldsymbol{u}_1 = \left\{\frac{\beta_1}{G_{11}} + O(\delta_i)\right\}\boldsymbol{u}_1 + \boldsymbol{s}_{i+1} \tag{4.9}$$

where $$||\boldsymbol{s}_{i+1}|| = O(\delta_i). \tag{4.10}$$

From equations (4.9) and (4.10) it is seen that the error in $\boldsymbol{x}_{i+1}/(\boldsymbol{x}_{i+1})_{p_{i+1}}$ is proportional to the square (not cube) of δ_i. Calculating the error in λ_{i+1} using equations (4.7) and (4.9) shows that this is also $O(\delta_i^2)$. Thus this iteration is only of second order.

To obtain a third-order iteration it is necessary to reduce the right-hand side of equation (4.8) to $O(\delta_i^2)$. One possibility follows from noting that, by equation (3.9),

$$a_i(\boldsymbol{u}_1)_{p_i} = 1 + O(\delta_i) \tag{4.11}$$

so that, by equation (4.5),

$$(\boldsymbol{v}_{i+1})_{p_i} = \frac{1}{\delta_i} + O(1). \tag{4.12}$$

If $\boldsymbol{x}_{i+1}$ is now defined to be the solution of

$$\boldsymbol{M}(\lambda_i)\boldsymbol{x}_{i+1} = \left[\frac{d\boldsymbol{M}}{d\lambda}(\lambda_i) - \frac{1}{2(\boldsymbol{v}_{i+1})_{p_i}}\frac{d^2\boldsymbol{M}}{d\lambda^2}(\lambda_i)\right]\boldsymbol{v}_{i+1} \tag{4.13}$$

then, using that $\lambda_i - \check{\lambda} = O(\delta_i)$, it is found that

$$\boldsymbol{M}(\lambda_i)(\delta_i\boldsymbol{x}_{i+1} - a_i^*\boldsymbol{u}_1) = O(\delta_i^2) \tag{4.14}$$

so that this modification gives a third-order iteration. Equations (4.5) and (4.12) also suggest that the iteration

$$\left.\begin{aligned}\boldsymbol{M}(\lambda_i)\boldsymbol{v}_{i+1} &= \frac{d\boldsymbol{M}}{d\lambda}(\lambda_i)\boldsymbol{v}_i/(\boldsymbol{v}_i)_{p_i},\\ \lambda_{i+1} &= \lambda_i - \frac{1}{(\boldsymbol{v}_{i+1})_{p_i}}\end{aligned}\right\} \tag{4.15}$$

is of second order. This iteration has the advantage of requiring the solution of only one set of linear equations per step. The author has not been able to derive this iteration by the methods used in Osborne and Michaelson.

5. Conclusion

In this paper the class of iterations discussed by Osborne and Michaelson has been examined critically with a view to determining the rate of convergence, and modifications have been suggested which ensure that the iteration is of third-order. The modified iteration (3.2) has already been tested and proved effective. The results obtained have been consistently better than those obtained using the iteration (1.3).

The iteration (4.15) has been tested on a family of matrices all of which depended non-linearly on the eigenvalue parameter. These eigenvalue problems were also solved using the iteration (1.3). In terms of the number of iterations required there was nothing to choose between the two methods (ten runs using equation (4.15) required a total of 44 iterations, the same problems with the same starting values required a total of 43 iterations using equation (1.3)). Thus equation (4.15) provided the more efficient procedure for the problems considered.

References

OSBORNE, M. R., and MICHAELSON, S. (1964). "The numerical solution of eigenvalue problems in which the eigenvalue parameter appears nonlinearly, with an application to differential equations," *The Computer Journal*, Vol. 7, pp. 58–65.

WILKINSON, J. H. (1961). "Rigorous Error Bounds for Computer Eigensystems," *The Computer Journal*, Vol. 4, pp. 230–241.

WILKINSON, J. H. (1958). "The Calculation of the Eigenvectors of Codiagonal Matrices," *The Computer Journal*, Vol. 1, pp. 90–96.

An error analysis of finite-difference methods for the numerical solution of ordinary differential equations

By M. R. Osborne*

A method is given for the calculation of strict, a-posteriori error bounds for the numerical solution by finite-difference methods of ordinary linear differential equations. The suggested procedure is illustrated by some numerical results for a particular differential equation.

1. Introduction

This paper is concerned with the derivation of strict, a-posteriori error bounds for the solution of ordinary differential equations by finite-difference methods. The basic idea of the method of error analysis is due to J. H. Wilkinson who has applied it to obtain strict error bounds for the solutions of sets of linear algebraic equations (see Wilkinson (1963) and the references quoted there). He argues as follows. Let the set of linear equations to be solved be

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{1.1}$$

In any process of calculation rounding errors almost always occur so that the process of numerical solution leads not to $\boldsymbol{x}$ but to a vector $\boldsymbol{z}$ satisfying the system of equations

$$(\boldsymbol{A} + \delta\boldsymbol{A})\boldsymbol{z} = \boldsymbol{b} + \delta\boldsymbol{b}. \tag{1.2}$$

By a careful analysis, bounds can be found for the magnitudes of the elements of $\delta\boldsymbol{A}$ and $\delta\boldsymbol{b}$. By suitably combining equations (1.1) and (1.2) and taking norms the result is obtained that

$$||\boldsymbol{x} - \boldsymbol{z}|| \leqslant \frac{||(\boldsymbol{A} + \delta\boldsymbol{A})^{-1}||}{1 - ||(\boldsymbol{A} + \delta\boldsymbol{A})^{-1}||\,||\delta\boldsymbol{A}||} \{||\delta\boldsymbol{b}|| + ||\delta\boldsymbol{A}||\,||\boldsymbol{z}||\}. \tag{1.3}$$

At this stage the quantities on the right-hand side of equation (1.3) can be estimated with the exception of $||(\boldsymbol{A} + \delta\boldsymbol{A})^{-1}||$ but, as equation (1.2) is the equation which is actually solved, there remains the possibility of obtaining at least an upper bound for the norm of the inverse of $(\boldsymbol{A} + \delta\boldsymbol{A})$.

An essential feature of the argument is the representation of the vector $\boldsymbol{z}$ as the solution of a set of linear equations. This permits the original problem to be treated as a perturbation of the one actually solved once bounds have been obtained for $\delta\boldsymbol{A}$ and $\delta\boldsymbol{b}$.

In general an error analysis of Wilkinson type does not seem to be applicable to the numerical solution of differential equations by finite-difference methods. This is because the process of solution requires the inversion of an operator of a different kind (the operator associated with a finite system of linear or non-linear algebraic equations) to the operator in the original problem.

This difficulty can be avoided in the case of finite-difference approximation to ordinary linear differential equations. Here it can be shown that the solution of the differential equation is also the solution of a linear difference equation, and bounds can be given for the difference between the coefficients in this equation and those in the equation produced by finite-difference approximation. The exact difference equation leads to

* *Edinburgh University Computer Unit, 7, Buccleuch Place, Edinburgh 8.*