Revista Brasileira de Flsica, Vol. 7, N? 1, 1977

# Numerical Solution of Eigenvalue Systems of Second Order Differential Equations by the Matching Method

J. A CASTILHO ALCARÁS
*Instituto de Física Teórica*, São Paulo SP.*

Recebido em 10 de Fevereiro de 1977

A method for solving numerically eigenvalue systems of second order differential equations is presented. This method is a result of an interplay of the matching method and the linear multistep methods for the solution of the system as well as for the derivative of that solution. An analysis and control of the errors envolved in the method is also presented.

Apresenta-se um método para a solução numérica do problema a autovalores para um sistema de equações diferenciais de segunda ordem. Esse método resulta de uma combinação do método do *match* com os métodos lineares a k-passos e fornece tanto a solução do sistema como sua derivada. Apresenta-se também uma análise e um controle dos erros envolvidos no método.

## 1. INTRODUCTION

Methods for the numerical integration of differential equations and systems of differential equations, over a finite range of the defining domain of the independent variable, are well known in the literature [see, e.g. Refs. 1-3]. Nevertheless, in order to determine, in quantum mechanical problems, eigenvalues and eigenfunctions of the Schrodinger equation, one is lead to the integration of eigenvalue differential equations (or systemof equations), the solution of which must consist of square integrable functions, which one seeks to determine over an infinite range of the independent variable. This new feature of the problem requires the use of an appropriate method.

* Postal Address: Caixa Postal 5956, 01000 - São Paulo SP.

One such a method is the "matching method", whose basic ideas are the following:

a) fix both a "guessed" value, for the eigenvalue, and a value for the integration step, and integrate the differential equation (or system) forward in a finite region;

b) find a solution in an asymptotic region, and use its values at some points of this region as initial values, in order to integrate the differential equation (or system) backward up to a finite region;

c) check if the guessed value for the eigenvalue will provide a match, in some prescribed point, of the solution obtained by forward integration , with that obtained by backward integration. If it does not, another value for the eigenvalue is to be tried.

The matching method for solving one single eigenvalue differential' equation is a well established thecnique[4]

In order to solve the Hartree-Fock equations for atoms and ions, Froese[5] established a method to solve systems of differential equations of the form

$$y_i''(x) = \sum_{j=1}^{N} v_{ij}(x)\, y_j(x) + E_i y_i(x) \quad , \ i = 1,2,\ldots,N, \tag{1.1}$$

with boundary conditions $y_i(0) = y_i(\infty) = 0$. Denoting by $g_i(x)$ the coupling terms, Eq. (1.1) reads

$$y_i''(x) = \left[v_{ii}(x) + E_i\right] y_i(x) + g_i(x) \ . \tag{1.2}$$

In Froese's method, one starts finding a first estimate, for $y_i(x)$ $(i= 1,2,\ldots,N)$, which in turn determines $g_i(x)$. Considering the $g_i(x)$ as independent functions, the system (1.2) becomes a set of $N$ decoupled differential equations which are then solved by the matching method for <u>one</u> differential equation. With the values of $y_i(x)$ soobtained, one evaluates again $g_i(x)$ and repeats the process, until the results are "self-consistent".

For an eigenvalue differential system of the type

$$y_i''(x) = \sum_{i=1}^{N} v_{ij}(x)\, y_i(x) + E\, y_i(x), \quad (i = 1,2,\ldots,N), \tag{1.3}$$

$$y_i(0) = y_i(\infty) = 0 \ ,$$

which will concern us here, Froese's method does not apply since, once "self-consistency" is achieved, one obtains $N$ parameters $E_i$ instead of a single one.

Recently, the matching method has been applied to the solution of eigenvalue systems of second order differential equations (1.3) by Raynal[6]. However, the solution of (1.3) appeared in Raynal's paper as a part of a broader problem, and, probably due to that, it did not receive a thorough treatment. It is our aim to provide such a treatment in the present paper.

In order to fix the notation, and to collect most of the relevant formulas, a brief survey of linear multistep methods for the numerical integration of second order differential equations (and systems), based on Refs. 1-3, is given in Section 2. Section 3 deals with the development of the matching method, starting with one single differential equation, in which case one can get a better insight into the problem, and then going to systems of differential equations. The errors envolved in the method are analised in Section 4.

## 2. A BRIEF SURVEY OF LINEAR MULTISTEP METHODS

### 2.1 ONE SINGLE DIFFERENTIAL EQUATION

Consider the problem of solving numerically the differential equation,

$$y''(x) = f(x,y(x)), \tag{2.1}$$

for $x$ in the range $-\infty<a\leq x\leq b<\infty$, with the boundary conditions

$$y(a) = \eta;\ y'(a) = \eta' \quad . \tag{2.2}$$

There is a great variety of methods for obtaining approximate solutions of (2.1), once some starting values for $y(x)$ are given. We shall focus our atention only on the socalled linear multistep methods.

Let $\{x_n\}$ be a sequence of equally spaced points $x_n = a+nh$, in the range $a\leq x\leq b$, where one seeks a solution of (2.1). Let $y_n$ be an approximation to the exact solution, at $x_n$, that is, to $y(x_n)$ and $f_n \equiv f(x_n, y_n)$. A linear multistep method (LMM), of stepnumber k, is a computational procedure for deterrnining the sequence $\{y_n\}$ by the following linear relationship between $y_{n+j}$ and $f_{n+j}$ $(j= 0,1,\dots,k)$:

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h^2 \sum_{j=0}^{k} \beta_j f_{n+j} \ , \tag{2.3}$$

where $\alpha_j$ and $\beta_j$ are constants of the method. It is assumed, without loss of generality, that $\alpha_k = 1$ and $\alpha_0\beta_0 \neq 0$.

To the linear multistep relation (2.3) one associates the linear difference operator

$$L\left[g(x);h\right] = \sum_{j=0}^{k} \left[\alpha_j g(x+jh) - h^2\beta_j g''(x+jh)\right], \tag{2.4}$$

in which $g(x)$ is any arbitrary function whose second derivative does exist. If $g(x)$ possesses higher continuous derivatives, one uses Taylor's expansion

$$L\left[g(x),h\right] = C_0+C_1 h g^{(1)}(x)+\dots C_{p+2} h^{p+2} g^{(p+2)}(x)+\dots, \tag{2.5}$$

where the C's are well defined functions of the $\alpha$'s and $\beta$'s (Ref.3). The LMM, (2.3), is said to be of order $p$ if[1]

$$C_0 = C_1 = \ldots\ldots = C_p = C_{p+1} = 0 \ , \ C_{p+2} \neq 0 \ . \tag{2.6}$$

For an LMM of order $p$, the remainder in (2.5) is bounded by[3]

$$|L\ [g(x);h]\,| \leq h^{p+2}\ GY \ , \tag{2.7}$$

with

$$Y = \max_{\xi\,\varepsilon\,[x,x+kh]} \left| g^{(p+2)}(\xi) \right| , \tag{2.8}$$

$$G = \frac{1}{p\,!} \int_0^k \left| \sum_{j=0}^{k} \left[ \alpha_j (j-s)_+^p - p\beta_j (j-s)_+^{p-1} \right] \right| ds, \tag{2.9}$$

and $z_+ = z$ if $z \geq 0$, and zero otherwise. (2.10)

Ref.3 quotes all convergent LMM, for $k$ up to 4. The simplest ones are those with $k$=2, namely,

i) $y_{n+2} - 2y_{n+1} + y_n = h^2 f_{n+1}$ ; $p$=2 ; $C_{p+2} = C_4$=1/12 ; (2.11)

ii) $y_{n+2} - 2y_{n+1} + y_n = \frac{h^2}{12} [f_{n+2} + 10 f_{n+1} + f_n]$ ; $p$=4 , $C_{p+2} = C_6 = -\frac{1}{240}$ . (2.12)

LMM (2.12), known as Numerov's method, is the most popular of them all.

## 2.2 SYSTEM OF DIFFERENTIAL EQUATIONS

Consider now the problem of solving numerically the system of differential equations

$$y_i''(x) = \sum_{j=1}^{N} f_j(x, y_1(x), \ldots, y_N(x)); \quad i = 1,2,\ldots,N, \tag{2.13}$$

for x in the range $-\infty \leq a \leq x \leq b < \infty$, with boundary conditions

$$y_i(a) = \eta_i; \ y_i'(a) = \eta_i' \quad ; \ i = 1,2,\ldots,N. \tag{2.14}$$

Introducing $N$-component column vectors $\vec{y}(x)=[y_1(x), y_2(x),\ldots,y_N(x)]^T$, $\vec{f}(x,\vec{y}(x))= [f_1(x,y_1(x),\ldots,y_N(x)),\ldots,f_N(x,y_1(x),\ldots,y_N(x))]^T$, $\vec{\eta}=[\eta_1,\ldots,\eta_N]^T$, and $\vec{\eta}'=[\eta_1',\ldots,\eta_N']^T$, Eqs. (2.13) and (2.14) become

$$\vec{y}''(x) = \vec{f}(x,\vec{y}(x)), \tag{2.15}$$

$$\vec{y}(0) = \vec{\eta}\,, \quad \vec{y}'(0) = \vec{\eta}'\,. \tag{2.16}$$

Similarly to the case of one single differential equation, one seeks for a sequence $\{\vec{y}_n\}$ of vectors that represent an approximation to the vector function solution, $\vec{y}$, in a given range $a\leq x\leq b$. A linear multistep of stepnumber $k$ is then defined by

$$\sum_{j=0}^{k} \alpha_j \vec{y}_{n+j} = h^2 \sum_{j=0}^{k} \beta_j \vec{f}(x_{n+j}, \vec{y}_{n+j})\,, \tag{2.17}$$

and the definition of order remains formally the same, being enough to replace $g(x)$ by $\vec{g}(x)$, in (2.4) and (2.5). The bound for the remainder is now[2]

$$\| L[\vec{g}(x);h] \| \leq h^{p+2}\, GY\,, \tag{2.18}$$

where

$$Y = \max_{\xi\in[x,x+kh]} \| g^{(p+2)}(\xi) \|\,, \tag{2.19}$$

$$\|\vec{g}(x)\| = |g_1(x)|+\ldots+|g_N(x)| \tag{2.20}$$

## 3. THE MATCHING METHOD

Consider now the problem of finding a numerical solution to:

i) the <u>eigenvalue</u> differential equation,

$$y''(x)= [v'(x)+E]\,y(x)\equiv v(x)y(x); \quad 0\leq x\leq\infty, \tag{3.1}$$

with boundary conditions

$$y(0) = y(\infty) = 0; \tag{3.2}$$

**ii) the** system of <u>eigenvalue</u> differential equations,

$$y_i''(x)= \sum_{j=1}^{N} v'_{ij}(x)\, y_j(x)+Ey_i(x)= \sum_{j=1}^{N} v_{ij}(x)y_j(x); \; i=1,2,\ldots,N, \tag{3.3}$$

with boundary conditions

$$y_i(0) = y_i(0) = 0 \quad ; i= 1,2,\ldots,N. \tag{3.4}$$

In applying the theory of the previous section to (3.1)-(3.2) and (3.3)-(3.4), two difficulties arise, namely, 1) the eigenvalue E is an unknown constant to be determined; 2) the range of x is infinite.

The rnatching method, which we shall describe in the remaining of this section, was introduced to solve such difficulties.

### 3.1 THE MATCHING METHOD FOR A SINGLE DIFFERENTIAL EQUATION

Let $x_{max}$ be a point such that for $x \geq x_{max}$, $v(x)$ can be approximated, within a good accuracy, by a simpler expression $v_{aspt}(x)$ which allows one to find analytical solutions to (3.1). Call $\phi_c(x;E)$ a solution of (3.1) with $v_{aspt}(x)$ in place of $v(x)$ which goes to zero as x goes to infinity. Then one takes a "guessed" value for E, and test whether (3.1), with the true $v(x)$, admits a solution $y(x)$ such that $y(0)=0$ and $y(x)=\phi_c(x,E)$, for $x \geq x_{max}$. If it does not, one tries another value, and so on.

The test consists in the following:

1) One chooses a set of m equally spaced points, $x_0=0$, $x_1=h,\ldots,x_m=x_{max}$, in the range $[0, x_{max}]$, and integrates (3.1) backward from $x_{ma}$ to some $x_\ell=x_{match}$ (Ref. 7), using a convergent k-step LMM (2.3), with starting values

$$y_{m+1}=\phi_c(x_{max}+h,E),\ y_{m+2}=\phi_c(x_{max}+2h,E),\ldots,y_{m+k}=\phi_c(x_{max}+kh,E). \tag{3.5}$$

In this way, one obtains a set of numbers $\{y^b_\ell(E),\ y^b_{\ell+1}(E),\ldots,y^b_m(E)\}$, as a numerical approximation to the values of the solution $y(x)$ of (3.1), at the points $x_R=x_{match}$, $x_{\ell+1}=x_R+h,\ldots,$ $x_m=x_{max}$.

2) Since to start the integration of (3.1) by a k-step LMM, $k-1$ initial values $y_0(E),\ y_1(E),\ldots,y_{k-1}(E)$ are needed, and we have only $y_0(E)=0$ (Eq.3.2), we need to choose a starting procedure to obtain $y_1(E),\ldots,$ $y_{k-1}(E)$ as accurate app-oximations to $y(x_1),\ y(x_2),\ldots,\ y(x_{k-1})$. [One such a procedure is given in Appendix 1.] Getting hold of then, one makes useof ak-step LMM to integrate (3.1) forward, obtaining in this way, the values $y^f_k(E),\ y^f_{k+1}(E),\ldots,y^f_R(E)$, as an approximation to those of $y(x)$, at the points $x_k,\ x_{k+1},\ldots,x_\ell$.

3) Since (3.1) is linear in $y(x)$, then $y_1(E),\ldots,y_{k-1}(E),\ y^f_k(E),\ \ldots,$ $y^f_R(E)$ are determined up to an overall multiplicative constant which is chosen in such a way as to make the numerical approximation to $y(x)$, at $x=x_\ell$, obtained by forward integration, to coincide with that one arrived at by backward integration, i.e., $y^f_\ell(E)=y^b_\ell(E)$. In this way, one garantees the "continuity" of the numerical solution. The arbitrary value attributed to $E$ will be an eigenvalue of (3.1) if the same "continuity" holds for the numerical approximation to $y'(x)$. The matching condition, which decides whether an arbitrary value attributed to $E$ is an eigenvalue of (3.1) is then

$$\frac{y'^f_\ell(E)}{y^f_\ell(E)} \cong \frac{y'^b_\ell(E)}{y^b_\ell(E)}, \tag{3.6}$$

where $y'^f_R(E),\ y'^b_\ell(E)$ are the numerical approximations to $y'(x)$, obtained by forward and backward integration, respectively. Methods for obtaining numerical approximations $y'_0,\ y'_1,\ldots,\ y'_m$ to $y'(0),\ y'(x_1),\ \ldots,$ $y'(x_m)$, are given in Appendix 2.

In concluding this subsection, we observe that $\phi_c(x;E)$ is defined up to a multiplicative constant which will factor out in all subsequent calculations.. One can fix this constant, up to a phase, imposing that the set $\{y_1, y_2, \ldots, y_m\}$ be consistent with the normalization of the true solution of (3.1), i.e.,

$$\int_0^\infty |y(x)|^2 dx = 1 . \tag{3.7}$$

This can be done by writing (3.7) as

$$\int_0^{x_{max}} |y(x)|^2 dx + \int_{x_{max}}^\infty |\phi_c(x,E)|^2 dx = 1 \tag{3.8}$$

and then calculating thefirst integral in (3.8) by anumerical integration formula which uses a set of equally spaced ramples, $|y_i(E)|^2$ in this case.

## 3.2 THE MATCHING METHOD FOR SYSTEM OF EIGENVALUE DIFFERENTIAL EQUATIONS

In writing (3.3) in a matrix form, one defines a matrix, $V(x)$, by $[V(x)]_{ij} = v_{ij}(x)$. Then, in matrix form, equations (3.3) and (3.4) read

$$\vec{y}''(x) = V(x)\vec{y}(x), \tag{3.9}$$

$$\vec{y}(0) = \vec{y}(\infty) = 0. \tag{3.10}$$

For a system of eigenvalue differential equations, the essentials of the matching method are the same, if one allows for a few modifications which stem from the fact that one is now dealing with vector quantities instead of scalars.

The point $x_{max}$ is now chosen such as to make the off-diagonal elements of the matrix V vanish, and in a way that the diagonal elements can be approximated by simpler expressions with which one is able to solve analytically the resultant decoupled system. One takes, for each component, a solution, $d_i\phi_i(x,E)$, which goes to zero as x goes to infinity. Then, the boundary condition $\vec{y}(\infty)=0$ is fulfilled by putting

$$y_i(x)= d_i\phi_i(x,E), \quad i= 1,2...,N, \tag{3.11}$$

where the $d_i$'s are arbitrary constants. Defining the N × N diagonal matrix, $\Phi(x,E)$, by $[\Phi(x,E)]_{ij}= \phi_i(x,E)\delta_{ij}$ and the $N$ vestor, $\vec{d} =[d_1,d_2, ...,d_N]$, E . (3.11) can be writen as

$$\vec{y}(x) = \Phi(x,E)\vec{d} \text{ , for } x \geq x_{\max} . \tag{3.12}$$

Taking

$$\vec{y}_{m+i} = \Phi(x_m+ih,E)\vec{d} \text{ , } i= 1,2,...,k, \tag{3.13}$$

as initial values, one could use a k-step LMM (2.17) to integrate (3.9) backward, obtaining in this way $\vec{y}^{\,b}_m(E)$, $\vec{y}^{\,b}_{m-1}(E),...,$ $\vec{y}^{\,b}_{\ell}(E)$. One may observe that, due to the fact that (3.9) is linear in $\vec{y}(x)$, the $\vec{y}^{\,b}_i(E)$'s depend on $\vec{d}$ only through a right positioned factor. One then can write

$$\vec{y}^{\,b}_i(E) = Q^b_i(E)\vec{d} \text{ , } \quad i = \ell, \ell+1 ,...,m . \tag{3.14}$$

Using (3.14) in the LMM (2.17), one finds that, due to the arbitrariness of $\vec{d}$, the $Q^b_i(E)$'s are determined by the same LMM used in obtaining the $\vec{y}_i$'s, with initial values

$$Q^b_{m+i}(E) = \Phi(x_m + ih), \quad i = 1,2,..., k. \tag{3.15}$$

The same arguments apply in approximating the derivative, and then one writes

$$\vec{y}'^{\,b}_i(E) = Q'^{\,b}_i(E)\vec{d}, \quad i = \ell,\ell+1 ,...,m. \tag{3.16}$$

In this way, one can compute the $Q_i^b(E)$'s, and the $Q_i'^b(E)$'s, leaving $\vec{d}$ as unknown.

For the <u>forward</u> integration, one chooses a <u>linear</u> starting procedure to obtain $\vec{y}_1(E), \vec{y}_2(E), \ldots, \vec{y}_{k-1}(E)$. Due to the linearity of that procedure, those vectors are obtained up to a right positioned vector factor, $\vec{c}$. One then can write

$$\vec{y}_i(E) = Q_i(E)\vec{c} \quad , \; i = 0,1,\ldots,k-1 \; . \qquad (3.17)$$

The use of a LMM to integrate (3.9) forward, using (3.17) as initial values, would give $\vec{y}_k^f(E),\ldots,\vec{y}_R^f(E)$, whose only dependence. on $\vec{c}$, would also be as a right positioned factor. Here, one defines the matrices $Q_i^f(E)$ by

$$\vec{y}_i^f(E) = Q_i^f(E)\vec{c} \quad , \; i = k, \; k+1,\ldots,\ell . \qquad (3.18)$$

Using (3.18) in the LMM, one finds, due to the arbitrariness of $\vec{c}$, that the $Q_i^f(E)$'s are determined by the same LMM, with rtarting values given by the $Q(E)$'s of Eq.(3.17).

The same arguments apply to the approximations to the derivative . One introduces the matrices $Q_i'^f(E)$ by

$$\vec{y}_i'^f(E) = Q_i'^f(E)\vec{c} \; , \; i = k', \; k'+1,\ldots,\ell, \qquad (3.19)$$

in order to compute the $Q^f(E)$'s and $Q'^f(E)$'s, leaving $\vec{c}$ as unknown.

The matching condition is now

$$\vec{y}_\ell^f(E) = y_\ell^i(E) \quad \text{and} \quad \vec{y}_\ell'^f(E) = \vec{y}_\ell'^b(E), \qquad (3.20)$$

which, by use of (3.14), (3.16), (3.18) and (3.19), becomes

$$Q_\ell^f(E)\vec{c} = Q_\ell^b(E)\vec{d} \quad \text{and} \quad Q_\ell'^f\vec{c} = Q_\ell'^b(E)\vec{d} \; . \qquad (3.21)$$

The matching condition, as shown in Eqs.(3.21), is not in a form appropriate for numerical purpose. From each one of Eqs. (3.21), one obtains a determination of $\vec{c}$. Equating those determinations, one is left with a homogeneous system of linear equations,

$$\{[Q_\ell^f(E)]^{-1} Q_\ell^b(E) - [Q_\ell'^f(E)]^{-1} Q_\ell'^b(E)\}\ \vec{d} \equiv M\vec{d} = 0\ , \tag{3.22}$$

in the unknowns $d_i (i = 1,2,...,N)$. As is well known, a necessary and sufficient condition for (3.22) having a non-trivial solution $\vec{d}$ is that

$$det\ \{[Q_\ell^f(E)]^{-1} Q_\ell^b(E) - [Q_\ell'^f(E)]^{-1}\ Q_\ell'^b(E)\ \} = 0\ , \tag{3.23}$$

showing then that (3.23) is a <u>necessary</u> condition for Eqs. (3.21) being satisfied. Conversely, if (3.23) is true, the homogeneous system of linear equations, (3.22), determines a non-trivial vector $\vec{d}$, which one uses to define a vector $\vec{c}$ by the relations

$$\vec{c} = [Q_\ell^f(E)]^{-1}\ Q_\ell^b(E)\vec{d} = [Q_\ell'^f(E)]^{-1} Q_\ell'^b(E)\vec{d}\ , \tag{3.24}$$

which, in turn, imply that Eqs.(3.21) do hold. Therefore, (3.23) is a necessary and sufficient condition for the matching equations (3.21) being satisfied. One, then, takes (3.23) as the <u>matching condition</u>. Such a condition can also be written in the form

$$det\ \{Q_\ell^b(E)[Q_\ell'^b(E)]^{-1} - Q_\ell^f(E)[Q_\ell'^f(E)]^{-1}\} = 0, \tag{3.25}$$

in which overall multiplicative factors, eventually present in the initial values, have cancelled out.

Strictly speaking, when (3.23) [or (3.24)] is satisfied, the system (3.22) yields a family of vectors $\vec{d}$. Assuming that the matrix M has rank $N-1$, those vectors differ among themselves only by a multiplicative constant. Again, this constant can be fixed, up to a phase, by imposing that the numerical approximation to $\vec{y}(x)$ be consistent with the normalization condition for $\vec{y}(x)$, the true solution of (3.9), i.e.,

$$\sum_{i=1}^{N} \int_0^\infty |y_i(x)|^2 dx = 1 \quad . \tag{3.26}$$

## 4. ERROR ANALYSIS AND CONTROL

There are, in the matching method, seven kinds of errors, which, to start with, are exhibited below for the case of a single differential equation:

i) the <u>local truncation error</u>: the error which arrises when

$\sum_{j=0}^{k} \alpha_j y(x_n+hj)$ is substituted by $h^2 \sum_{j=0}^{k} \beta_j f(x_n+jh, y(x_n+jh))$.

ii) the <u>discretization error</u>: the error committed when $y(x)$ is approximated by the sequence $\{y_n\}$, determined by the LMM (2.3), with initial values $y_i (i= 0, .., k-1)$.

iii) the <u>round-off error</u>: the error committed by the computer, when calculating $y_1, y_2, ..., y_m$ by (2.3), due to the finite representation of numbers in its memory.

iv) the errors in the starting values.

v) errors similar to i)-iv) but now related to the derivative.

vi) the error comitted when $y(x)$, for $x \geq x_{max}$ is substituted by $\phi_c(x, E)$.

vii) error in the eigenvalue $E$ obtained by the matching condition.

Errors i) to iv) are common to all integration methods, including the LMM employed here, and bounds for those errors can be found in Ref. 1, for the case of a single second order differential equation and convergent LMM. For a <u>system</u> of second order differential equations, similar results can be obtained following the same approach taken by Henrici [2].

As given in Section 2, the error i) is bounded by

$$h^{p+2} GY, \tag{4.1}$$

where G is defined by (2.9), and Y by (2.8) or (2.9) according to whether one is dialing with (3.1) or (3.3).

A bound for the discretization error, $e_n = y_n - y(x_n)$, is given by

<u>Theorem</u> 1. If $h^2 < L^{-1}|\alpha_k/\beta_k|$, the discretization error satisfies, for $a \leq x_n \leq b$,

$$|e_n| \leq \Gamma^* \left[ (x_n - a^*) k A \delta + (x_n - a^*)^2 G Y h^p / 2 \right] \exp \left[ (x_n - a^*)^2 \Gamma^* L B \right], \tag{4.2}$$

where $\delta$ is such that $|y_i - y(x_i)| \leq h\delta, \quad i = 0, 1, \ldots, k-1$ (4.3)

$$A = |\alpha_0| + \ldots + |\alpha_k|, \tag{4.4}$$

$$B = |\beta_0| + \ldots + |\beta_k|, \tag{4.5}$$

$$\Gamma^* = \frac{\Gamma}{1 - h^2 L |\beta_k/\alpha_k|} \tag{4.6}$$

$$a^* = a - h\gamma/\Gamma, \tag{4.7}$$

$$L = \text{Lipschitz constant of } f(x, y(x)),$$

$$Y = \max_{x \in [a,b]} |y^{(p+2)}(x)| , \tag{4.8}$$

$\Gamma$ and $\gamma$ being constants of the LMM such that

$$|\gamma_\ell| \leq \ell\Gamma + \gamma \quad , \ell = 0, 1, 2, \ldots, \tag{4.9}$$

while the sequence $\{\gamma_n\}$ is defined by

$$\frac{1}{\alpha_0 \xi^k + \alpha_1 \xi^{k-1} + \ldots + \alpha_k} = \gamma_0 + \gamma_1 \xi + \gamma_2 \xi^2 + \ldots \tag{4.10}$$

For a system of second order differential equations, the discretization error, $\vec{e}_n = \vec{y}_n - \vec{y}(x_n)$, also satisfies (4.2), with $A, B, \Gamma, \Gamma^*, a^*$, defined as before, 6 being such that

$$\| \vec{y}_i - \vec{y}(x_i) \| \leq h\delta \ , \ i = 0,1,\ldots,k-1 \ , \qquad (4.11)$$

L standing for the Lipschitz constant of $\vec{f}(x,\vec{y}(x))$.

Let $\varepsilon_{n+k}$ be the <u>local round-off error</u> committed when the numerical value $\tilde{y}_{n+k}$ of $y_{n+k}$ is obtained by (2.3). Taking $\varepsilon$ such that $|\varepsilon_n| \leq \varepsilon$, then the accumulated round-off error, $r_n = \tilde{y}_n - y_n$, for $h^2 < L^{-1}|\alpha_k/\beta_k|$, satisfies

$$| r_n | \leq \frac{\varepsilon}{h^2} \Gamma^* (x_n - a^*)^2 \exp\left[(x_n - a^*)^2 \Gamma^* L B\right] \ . \qquad (4.12)$$

For a system of differential equations, $\vec{r}_n = \vec{y}_n - y_n$ also satisfies (4 . 12). .

Combining (4.2) and (4.12), one finds that, for $h^2 < L^{-1}|\alpha_k/\beta_k|$, the <u>resultant</u> numerical error, $\sigma_n = \tilde{y}_n - y(x_n)$ $\left[\text{or } \vec{\sigma}_n = \tilde{\vec{y}}_n - \vec{y}(x_n)\right]$ satisfies

$$|\sigma_n| \leq \Gamma^* \left[ (x_n - a^*) k A \delta + \frac{(x_n - a^*)^2}{2} \left( \frac{2\varepsilon}{h^2} + G Y h^p \right) \right] \exp\left[ (x_n - a^*)^2 \Gamma^* L B \right] . \qquad (4.13)$$

To assure a reasonable accuracy, one must control the relative error $|\sigma_n|/|y_n|$. Starting with a steplength $h \lesssim [L^{-1}|\alpha_k/\beta_k|]^{1/2}$, and using double precision, floating point, arithmetic, for which one assumes $\varepsilon \cong 0$, each time the relative error becomes larger than a given tolerance, one discards the last value $y_{n+1}$, and one proceeds with a steplength a half of the previous one. The additional approximations to $y(x)$ and $y'(x)$, obtained from (2.12) and (A2.9.4), are given by

$$\left[ 1 + \frac{5h^2}{24} v(x_n - h/2) \right] y_{n-1/2} = y_n + y_{n-1} - \frac{h^2}{48}\left[ f_n + f_{n-1} \right], \qquad (4.14)$$

$$y'_{n-1/2} = y'_{n-1} + \frac{h}{12}(f_{n-1/2} + 4 f_{n-3/4} + f_{n-1}), \qquad (4.15$$

where $f_{n-3/4} = v(x_n - 3h/4)y_{n-3/4}$, and $y_{n-3/4}$ given by

$$\left|2 + \frac{5}{96}v(x_n - 3h/4)\right| y_{n-3/4} = y_{n-1/2} + y_{n-1} - \frac{h^2}{192}(f_{n-1/2} + f_{n-1}). \tag{4.16}$$

[For systems of equations, one uses the vector version of (4.14)-(4.16).]

In the further evaluations of $\sigma_n$, one has to redefine $a^*$, $\delta$ and $\gamma$.

If $y_\ell$ and $y'_\ell$ [$\vec{y}_\ell$ and $\vec{y}'_\ell$] were known exactly, the more precise the matching condition (3.6) [(3.25)] were satisfied, more sharp the corresponding value of $E$ would be. However, since $y_\ell$ and $y'_\ell$ [$\vec{y}_\ell$ and $\vec{y}'_\ell$] are not exact, all values of $E$, that make the matching condition consistent with the errors of $y_\ell$ and $y'_\ell$ [$\vec{y}_\ell$ and $\vec{y}'_\ell$], are allowed, and hence $E$ is obtained within an imprecision $\Delta E$.

## ACKNOWLEDGMENTS

The author thanks B.M. Pimentel Escobar for stimulating discussions. He is particularly indebted to J. Leal Ferreira for revising the manuscript.

## APPENDIX 1 – STARTING PROCEDURES

Procedures to obtain the starting values $y_1, y_2, \ldots, y_{k-1}$ needed to iniciate the numerical integration of (3.1) by a LMM will be presented now. It will be exploited the peculiarity of $f(x,y(x))$ of being linear in y, i.e., $f(x,y(x)) = v(x)y(x)$.

When $v(x)$ is <u>finite</u>, at $x=0$, approximate values to $y(x)$, satisfying (3.1), can be obtained for small values of $x$, in terms of $y(0)$ and $y'(0)$, by a Taylor expansion, since

$$y(h) = y(0) + hy^{(1)}(0) + \frac{h^2}{2!}y^{(2)}(0) + \frac{h^3}{3!}y^{(3)}(0) + \ldots$$

$$= \left[1 + \frac{h^2}{2!}v(0) + \frac{h^3}{3!}v^{(1)}(0) + \ldots\right] y(0) + \left[h + \frac{h^3}{3!}v(0) + \ldots\right] y'(0),$$

$$y'(h) = y^{(1)}(0) + hy^{(2)}(0) + \frac{h^2}{2!}y^{(3)}(0) + \ldots =$$

$$= \left[hv(0) + \frac{h^2}{2!}v^{(1)}(0) + \ldots\right] y(0) + \left[1 + \frac{h^2}{2!}v(0) + \ldots\right] y'(0). \qquad \text{(A1.1)}$$

It is now possible to compute $y(2h)$ from the expansion

$$y(2h) = y(h) + hy^{(1)}(h) + \frac{h^2}{2!}y^{(2)}(h) + \ldots$$

$$= \left[1 + \frac{h^2}{2!}v(h) + \frac{h^3}{3!}v^{(1)}(h) + \ldots\right] y(h) + \left[h + \frac{h^3}{3!}v(h) + \ldots\right] y'(h), \qquad \text{(A1.2)}$$

by substituting the approximations already obtained for $y(h)$ and $y'(h)$ into (~1.2). Continuing in this way, and truncating the expansions at appropriate points, additional starting values of any desired order of accuracy can be obtained. Since the boundary condition (3.2) gives $y(0) = 0$, one sees that $y(h), y(2h), \ldots, y((k-1)h)$ are determined up to a multiplicative constant, $c = y'(0)$, which is fixed by the matching method. In an analogous way, one obtains $y'(2h)$, $y'(3h), \ldots$ $y'((k'-1)h)$ up to the same multiplicative constant c.

For a system of second order differential equations with $v_{ij}(x)$ finite at x = 0, one has

$$y_i(h) = y_i(0) + hy_i^{(1)}(0) + \frac{h^2}{2!}y_i^{(2)}(0) + \frac{h^3}{3}y_i^{(3)}(0) + \ldots$$

$$= \sum_{j=1}^{N} \{\left[\delta_{ij} + \frac{h^2}{2!}v_{ij}(0) + \frac{h^3}{3!}v_{ij}^{(1)}(0) + \ldots\right] y_j(0) + \left[h\delta_{ij} + \frac{h^3}{3!}v_{ij}(0) + \ldots\right] y_j^{(1)}(0)$$

or, in matrix form,

$$\vec{y}(h) = \left[I + \frac{h^2}{2!}V(0) + \frac{h^3}{3!}V^{(1)}(0) + \ldots\right] \vec{y}(0) + \left[hI + \frac{h^3}{3!}V(0) + \ldots\right] \vec{y}^{(1)}(0),$$

(A1.3)

where $I$ is the N × N identity matrix. Analogously,

$$\vec{y}\,'(h)= \left[hV(0)+ \frac{h^2}{2!}\, V^{(1)}(0)+\ldots\right]\vec{y}(0)+ \left[I+ \frac{h^2}{2!}\, V(0)+\ldots\right] \vec{y}\,'(0) \ . \qquad \text{(A1.4)}$$

Proceeding in this way, one obtains approximations to $\vec{y}(2h),\vec{y}(3h),\ldots$ , $\vec{y}((k-1)h)$ as well as to $\vec{y}\,'(2h)$, $\vec{y}\,'(3h),\ldots,$ $\vec{y}\,'((k'-1)h)$. Since, by the boundary condition (3.10), $\vec{y}(0)=0$, all such starting values are obtained up to a multiplicative right factor $\vec{c}=\vec{y}\,'(0)$, which is fixed by the matching method.

When $v(x)$ diverges at $x=0$, the previous procedure cannot be applied. However, when $v(x)$ has the form

$$v(x) = \frac{L(L+1)}{x^2} + \frac{\beta}{x} + \sum_{i=0}^{\infty} b_i\, x^i \ , \ L \geq 0 \ , \qquad \text{(A1.5)}$$

a very common situation in physical applications, one may write

$$y(x) = x^{\alpha+1} \sum_{i=0}^{\infty} a_i\, x^i \qquad \text{(A1.6)}$$

and use (3.1) to find $a_0$, $a_1,\ldots$ . As a result, one finds $\alpha=L$, and $a_1$, $a_2,\ldots$ proportional to $a_0$. If one truncates expression (A1.6) at appropriate points, then starting values for $y(x)$ and $y'(x)$ of any desired order of accuracy can be obtained. Again, all of them are obtained up to a multiplicative constant, $c=a_0$.

For a system, in the case that $V(x)$ diverges at $x=0$, the first procedure cannot be applied either. However, when $V(x)$ has the form

$$V(x) = \frac{L(L+1)}{x^2}\, I + \frac{1}{x}\, K + \sum_{j=0}^{\infty} x^j\, B_j \ ,$$

where $K$ and $B_j$ are constant matrices, which is also a common situation in physical applications, one writes

$$\vec{y}(x) = x^{\alpha+1} \sum_{i=0}^{\infty} x^{i} \vec{a}_{i} ,$$

and uses (3.9) to find a, $\vec{a}_0$, $\vec{a}_1$,... . As a result, one finds a = L , and $\vec{a}_1$, $\vec{a}$ ... collinear to $\vec{a}_0$. One then obtains starting values for $\vec{y}(x)$, as well as for $y'(x)$, all of them determined up to a right posi-tioned factor $\vec{c}=\vec{a}_0$, wich is fixed by the rnatching method.

## APPENDIX 2 – APPROXIMATION TO THE DERIVATIVE

Consider the linear difference operator

$$L'[g(x);h] = \sum_{j=0}^{k'} [\alpha'_j g(x+jh) - h\ \beta'_j g'(x+jh)] , \tag{A2.1}$$

associated to some LMM for a first order differential equation, with $g(x)$ being any arbitrary function possessing a firs't derivative. If $g(x)$ possesses higher order continuous derivatives, one uses Taylor's expan-sion to obtain

$$L'[g(x);h] = C'_0 + C'_1 h g^{(1)}(x) + \dots + C'_{p'+1} h^{p'+1} g^{(p'+1)}(x) + \dots, \tag{A2.2}$$

where the $C''$s are well defined functions of the $\alpha'$'s and $\beta'$'s (Ref. 3). The operator $L'$ is said to be of order $p'$ if

$$C'_0 = C'_1 = \dots = C'_{p'} = 0,\ C'_{p'+1} \neq 0 . \tag{A2.3}$$

The remainder in (A2.2) is bounded by

$$|L'[g(x);h]| \leq h^{p'+1} G' Y' , \tag{A2.4}$$

where

$$Y' = \max_{\xi \in [x, x+hk']} |g^{(p'+1)}(\xi)| , \tag{A2.5}$$

$$G' = \frac{1}{p'!} \int^{k'} \left| \sum_{j=0}^{k'} \left[ \alpha'_j (j-S)_+^{p'} - p'\beta'_j (j-S)_+^{p'-1} \right] \right| ds. \qquad \text{(A2.6)}$$

The operator L', when applied to $g'(x)$, gives

$$L'\left[g'(x);h\right] = \sum_{j=0}^{k'} \left[\alpha'_j g'(x+jh) - h\beta'_j g''(x+jh)\right] , \qquad \text{(A2.7)}$$

what suggest associating, to L', the LMM

$$\sum_{j=0}^{k'} \alpha'_j \, y'_{m+j} = h \sum_{j=0}^{k'} \beta'_j \, f_{m+j} \quad , \qquad \text{(A2.8)}$$

in order to generate the sequence $\{y'_{k'}, y'_{k'+1}, \dots\}$ as an approxirnation to the derivative of $y(x)$ (solution of 3.1), at the points $x_{k'}, x_{k'+1}, \dots$

The coefficients of our LMM(A2.8) coincide then with those of the LMM associated to the first order differential equation. Ref. 3 quotes all convergent LMM's for $k'$ up to 4. Some of them are

i) $y'_{n+1} - y'_n = hf_n \quad , p = 1,$ (A2.9.1)

ii) $y'_{n+2} - y'_n = 2hf_{n+1} \quad , p = 2 ,$ (A2.9.2)

iii) $y'_{n+1} - y'_n = \frac{h}{2}(f_{n+1} + f_n) \quad , p = 2,$ (A2.9.3)

iv) $y'_{n+2} - y'_n = \frac{h}{3}(f_{n+2} + 4f_{n+1} + f_n) \quad , \quad p = 4,$ (A2.9.4)

v) $y'_{n+4} - \frac{8}{19}(y'_{n+3} - y'_{n+1}) - y'_n = \frac{6h}{19}(f_{n+4} + 4f_{n+3} + 4f_{n+1} + f_n), \; p = 6 \; .$
(A2.9.5)

One looks now for a bound for the discretization error for the derivative, $e'_n = y'(x_n) - y'_n$ , where $y'_n$ is obtained solving (A2.8) exactly.

Applying (A2.1) to $y'(x_i)$, and taking $(3.1)$ and (A2.4) into account , one finds

$$\sum_{j=0}^{k'} \alpha'_j y'(x_i+jh) = h \sum_{j=0}^{k'} \beta'_j f(x_i+jh,\ y(x_i+jh)) + R_i \ , \qquad \text{(A2.10)}$$

where

$$|R_i| \le h^{p'+1} G'Y' \ , \qquad \text{(A2.11)}$$

$$Y' = \max_{x\varepsilon[a,b]} |y^{(p'+2)}(x)| \ . \qquad \text{(A2.12)}$$

Subtracting (A2.8) from (A2.10), one obtains

$$\sum_{j=0}^{k'} \alpha'_j e'_{i+j} = h \sum_{j=0}^{k'} \beta'_j t_{i+j} + R_i, \qquad \text{(A2.13)}$$

where

$$t_i = f(x_i,\ y(x_i)) - f(x_i, y_i) \ . \qquad \text{(A2.14)}$$

In order to obtain $e'_n$, from (A2.13), for a fixed value of $n$, and $\ell'=0,1,\ldots, n-k'$, one multiples (A2.13) corresponding to $i= n-k'-\ell'$ by constants $\gamma'_{\ell'}$, and add the resulting equations. In doing so, in the left side there results

$$(\alpha'_{k'}\gamma'_0)e'_n + (\alpha'_{k'}\gamma'_1+\alpha'_{k'-1}\gamma'_0)e'_{n-1}+\ldots+(\alpha'_{k'}\gamma'_{n-k'}+\alpha'_{k'-1}\gamma'_{n-k'-1}+$$

$$+\ldots+\alpha'_0\gamma'_{n-2k'})e'_{k'}+\ldots+(\alpha'_0\gamma'_{n-k'})e'_0 \ , \qquad \text{(A2.15)}$$

where the convention $\gamma'_{-4}=0$, for $q=1,2,\ldots$, was used. Putting the coefficient of $e'_n$ equal to 1, and those of $e'_{n-1}, e'_{n-2},\ldots,e'_{k'}$, equal to nought, one obtains $n-k'+1$ equations,

$$\gamma_0 = \alpha_0$$

$$\sum_{j=0}^{k'} \alpha'_{k'-j}\gamma'_{i-j} = 0 \; ; \; i = 1,2,\ldots,n-k' \; ,$$

which determine the constants $\gamma'_0, \gamma'_1, \ldots, \gamma'_{n-k'}$. (A2.16)

These are precisely the relations which determine $\gamma'_0, \gamma'_1, \ldots,$ in the expansion

$$\frac{1}{\alpha'_{k'}+\alpha'_{k'-1}\xi+\alpha'_{k'-2}\xi^2+\ldots+\alpha'_0\xi^{k'}} = \gamma'_0 + \gamma'_1\xi + \gamma'_2\xi^2+\ldots \quad . \tag{A2.17}$$

A problem raised by (~2.16)and (A2.17) is to see whether $|\gamma'_n|$ is limited as $n$ goes to infinity. From (A2.16), one obtains firstly $\gamma'_0, \gamma'_1, \ldots, \gamma'_{k'-1}$. Then, for $i \geq k'$, (A2.16) becomes a k'th-order linear, homogeneous, difference equation with constant coefficients:

$$\alpha'_{k'}\gamma_{n+k'} + \alpha'_{k'-1}\gamma'_{n+k'-1} +\ldots+ \alpha'_0\gamma'_n = 0 \; , \quad n = 0,1,2,\ldots \; . \tag{A2.18}$$

As can be seen in Refs. 1 and 2 , the necessary and sufficient condition for $\gamma'_n$ , given by (A2.18), beingfiniteasngoes to infinity is that the roots $r_1, r_2, \ldots r_{k'}$ of the polynomial

$$\rho(r) = \sum_{j=0}^{k'} \alpha'_j \, r^j$$

have moduli not greater than 1, and those with modulus 1 being simple roots. This is precisely the necessary and sufficient condition for stability of a LMM, for a first order differential equation. Then, taking a stable LMM, the $\gamma'_n$'s in (A2.16) and (A2.17) will be finite.

The right hand side of the equation which determines $e'_n$ is

$$h[\gamma'_0(\beta'_0 t_{n-k'}+\beta'_1 t_{n-k'+1}+\ldots+\beta'_{k'} t_n)+ \gamma'_1(\beta'_0 t_{n-k'-1}+\beta'_1 t_{n-k'}+\ldots+\beta'_{k'} t_{n-1})$$

$$+\dots+\gamma'_{n-k'}(\beta'_0 t_0+\beta'_1 t_1+\dots+\beta'_{k'} t_{k'})]+(\gamma'_0 R_{n-k'}+\gamma'_1 R_{n-k'-1}+\dots+\gamma'_{n-k'} R_0) \quad .$$

Taking moduli and using

$$|t_i| \leq L\ |y(x_i)-y_i| = L\ e_i\ , \tag{A2.19}$$

one finds

$$|e'_n| \leq \Gamma'\left[k'A'\delta' + (n-k'+1)h^{p'+1}\ G'Y' + (n-k'+1)hB'L|e_n|\right]\ , \tag{A2.20}$$

where

$$\begin{aligned} A' &= |\alpha'_0| + |\alpha'_1|+\dots+|\alpha'_{k'}| \\ B' &= |\beta'_0| + |\beta'_1|+\dots+|\beta'_{k'}| \\ \Gamma' &= \max\ \{\gamma'_0,\gamma'_1,\dots\} \\ Y' &= \max_{x\in[a,b]} |y^{(p'+2)}(x)| \end{aligned} \tag{A2.21}$$

and $\delta'$ is such that

$$|y'(x_i) - y'_i| \leq \delta'\ ,\quad i = 0,1,\dots,k'-1. \tag{A2.22}$$

For the accumulated round-off error, $r'_n=y'_n - \tilde{y}'_n$, the calculation is exactly the same as that for first order differential equations. Assuming that the local round-off errors are bounded by $\varepsilon'$, the accurnulated round-off error, for $h<L^{-1}|\alpha'_{k'}/\beta'_{k'}|$, satisfies[1]

$$|r'_n| \leq \frac{\varepsilon'}{h}\ \Gamma'^*(x_n-a)\exp\left[(x_n-a)\Gamma'^*LB'\right],$$

where

$$\Gamma'^* = \frac{\Gamma'}{1-hL|\beta'_{k'}/\alpha'_{k'}|} \tag{A2.23}$$

For the case of a system of differential equations, one uses the vector version of (A2.8),

$$\sum_{j=0}^{k'} \alpha'_j \vec{y}'_{i+j} = h \sum_{j=0}^{k'} \beta'_j \vec{f}_{i+j} , \tag{A2.24}$$

and finds, for the discretization and round-off errors,

$$\| \vec{e}'_n \| \leq \Gamma' \left[ k'A'\delta' + (n-k'+1)h^{p'+1} G'Y' + (n-k'+1)hB' \|e_n\| \right], \tag{A2.25}$$

$$\| \vec{r}'_n \| \leq \frac{E'}{h} \Gamma'^* f(x_n-a) \exp \left[ (x_n-a)\Gamma'^* LB' \right] , \tag{A2.26}$$

where

$$Y' = \max_{x \in [a,b]} \| \vec{y}^{(p'+2)}(x) \| , \tag{A2.27}$$

$$\| \vec{y}'(x_i) - \vec{y}'_i \| \leq \delta' \quad i = 0,1,\ldots,k'-1 . \tag{A2.28}$$

**REFERENCES AND NOTES**

1. P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, New York, 1962.

2. P. HENRICI, *Error Propagation for Difference Methods*, John Wiley, New York, 1963.

3. J.D. LAMBERT, *Computational Methods in Ordinary Differential Equations*, John Wiley, New York, 1973.

4. See, for instance,
J.W. COOLEY, Math. Comp. 15, 363 (1961);
S.M. PEREZ, in *A Computer Programme for the Calculation of the Nuclear Shell Model Single-Particle Wavefunctions*, Nuclear Physics Theoretical Group Report nº 38, University of Oxford, 1967.

5, C. FROESE FISHER, Can.J.Phys. 41, 1895 (1963); Phys. Rev. 137A, 1644 (1965); Computer Phys. Commun. 1, 151 (1969). D.F. MAYERS and F.O'BRIEN, J. Phys. B1, 145 (1968).
6. J. RAYNAL, in *Computing as a Language of Physics*, pp.281-322, IAEA, Vienna, 1972.
7. Usually one takes $x_\ell$ as the last zero of $v(x)$, since for $x > x_\ell$, $y(x)$ must have an exponentially decreasing behavior.