PMM/A-261

COMISION NACIONAL DE ENERGIA ATOMICA
PRESIDENCIA DE LA NACION

PROGRAMA MULTINACIONAL DE METALURGIA
(Programa Regional de Desarrollo Científico y Tecnológico-OEA)

DEFECTOS Y DIFUSION EN METALES, COMPUESTOS IONICOS Y OXIDOS

Dra. Fanny Dyment
Dra. Ruth H. de Tendler
Lic. Adolfo Marajofsky

Buenos Aires-Argentina
1978

CNEA-AC 21/85

REPUBLICA ARGENTINA
COMISION NACIONAL DE ENERGIA ATOMICA
Dependiente de la Presidencia de la Nación
DIRECCION PLANIFICACION, COORDINACION Y CONTROL
Departamento Centro de Cálculo Científico

INTRODUCCION A LA SOLUCION NUMERICA DE ECUACIONES DIFERENCIALES ORDINARIAS

Guillermo MARSHALL

Buenos Aires
1985

PREFACIO

Estas notas, de carácter introductorio al tema, están basadas en la primera parte de un curso realizado durante el año 1984 en el Centro de Cálculo Científico de la Comisión Nacional de Energía Atómica. Fundamentalmente están dirigidas a profesionales, investigadores y estudiantes de física, matemática e ingeniería sin conocimientos previos de análisis numérico.

Quisiera agradecer a P. Binaghi, M.C. Rey y V. Visconti por la realización de ejemplos numéricos y graficación, y a A. Menéndez, E. Moyano y M.C. Rey por la atenta lectura del manuscrito. Agradezco también a los asistentes al curso que con su ayuda desinteresada en el mecanografiado de algunas secciones permitieron la rápida concreción del manuscrito.

Estas notas fueron realizadas en un procesador de la palabra en el Departamento Información Técnica de la Comisión Nacional de Energía Atómica. A sus integrantes: R. Stortini y en particular a S. Alvarez mi reconocimiento por la excelencia del trabajo realizado.

Finalmente mi reconocimiento al Dr. Tito Suter Director del Centro de Cálculo Científico por su constante apoyo a la tarea científica.

Buenos Aires, noviembre 1984 G.Marshall

INDICE

1.0 DIFERENCIAS FINITAS

1.1 Aproximaciones de una derivada de primer orden

La idea básica del método de las diferencias finitas consiste en aproximar las derivadas en un punto por cocientes en diferencias sobre un intervalo pequeño, transformando así el problema diferencial en un problema algebraico. Por ejemplo para aproximar por diferencias finitas la derivada u' de una función u(x) aproximamos la tangente por una secante como se muestra en la Fig. 1.1.

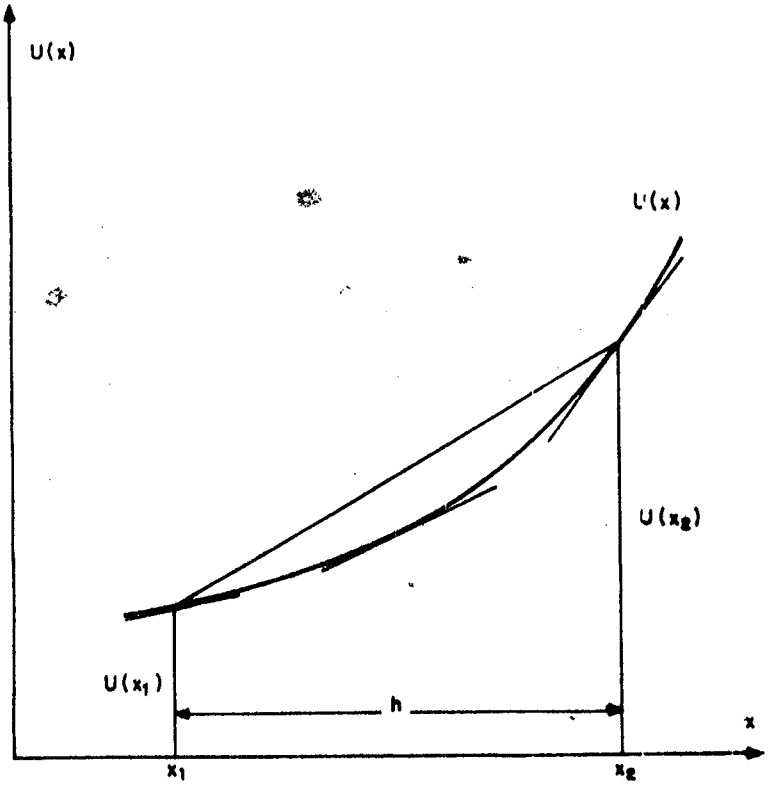

Fig. 1.1. Aproximación de la derivada primera

Definiendo el paso o intervalo $h = x_2 - x_1$, en cualquier punto del intervalo $x_1 \leq x \leq x_2$, $u'(x)$ se aproxima por

$$u'(x) = \frac{u(x_2) - u(x_1)}{h} \quad , \quad x_1 \leq x \leq x_2 \qquad (1.1)$$

Para casi todos los puntos la expresión (1.1) es una aproximación, pero por lo menos para un punto la igualdad es estricta por el teorema de Rolle. Si establecemos que

$$u'(x_1) = \frac{u(x_2) - u(x_1)}{h} \quad , \qquad (1.2)$$

la expresión (1.2) es una diferencia en adelanto. En cambio si establecemos que

$$u'(x_2) = \frac{u(x_2) - u(x_1)}{h} \quad , \qquad (1.3)$$

la expresión (1.3) es una diferencia en atraso. En la Fig. 1.2 se ilustra un esquema en diferencias centrado para la aproximación de $u'(x)$ en el punto $x_0 = (x_1+x_2)/2$,

$$u'(x_0) = \frac{u(x_2) - u(x_1)}{h} \quad , \qquad (1.4)$$

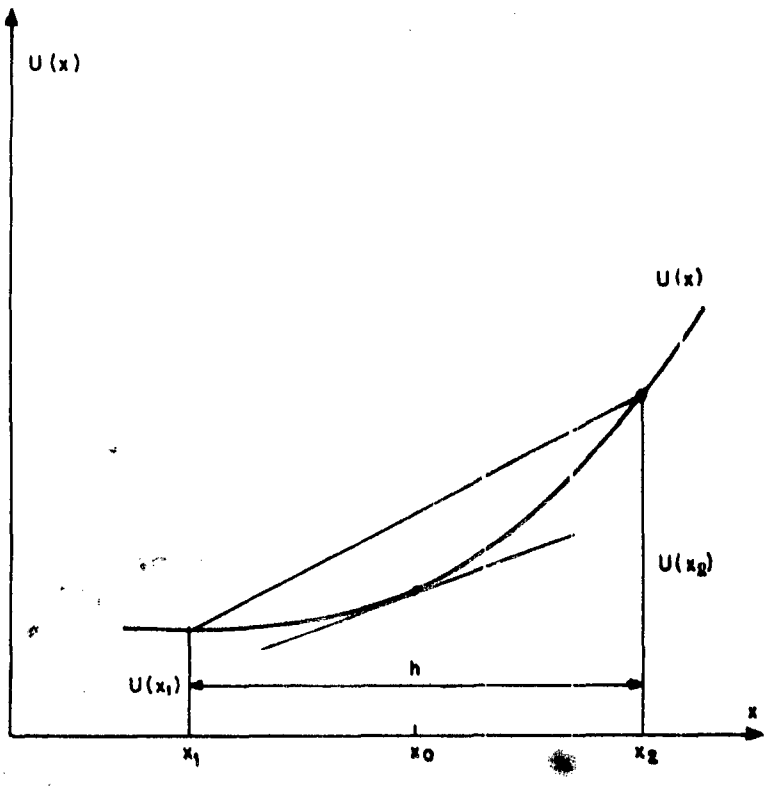

Fig. 1.2 Aproximación por diferencias centradas de la derivada primera

Por consideraciones geométricas se puede ver que la fórmula (1.4) es una mejor aproximación que las fórmulas (1.2) ó (1.3) ya que está más alineada la cuerda con la tangente.

Para poder demostrar la aseveración precedente es preciso realizar un estudio de la precisión de una aproximación en diferencias. Este consiste en el análisis del error que se comete al reemplazar una expresión diferencial por su aproximación en diferencias y de qué depende dicho error. Comenzamos analizando la aproximación de $u'(x)$ con un esquema centrado. Por conveniencia, en la Fig. 1.3 se ilustra nuevamente la aproximación de $u'(x_0)$ por un esquema centrado, donde x_0 es el punto medio entre x_1 y x_2. Desarrollando $u(x_1)$ y $u(x_2)$ alrededor de $u(x_0)$

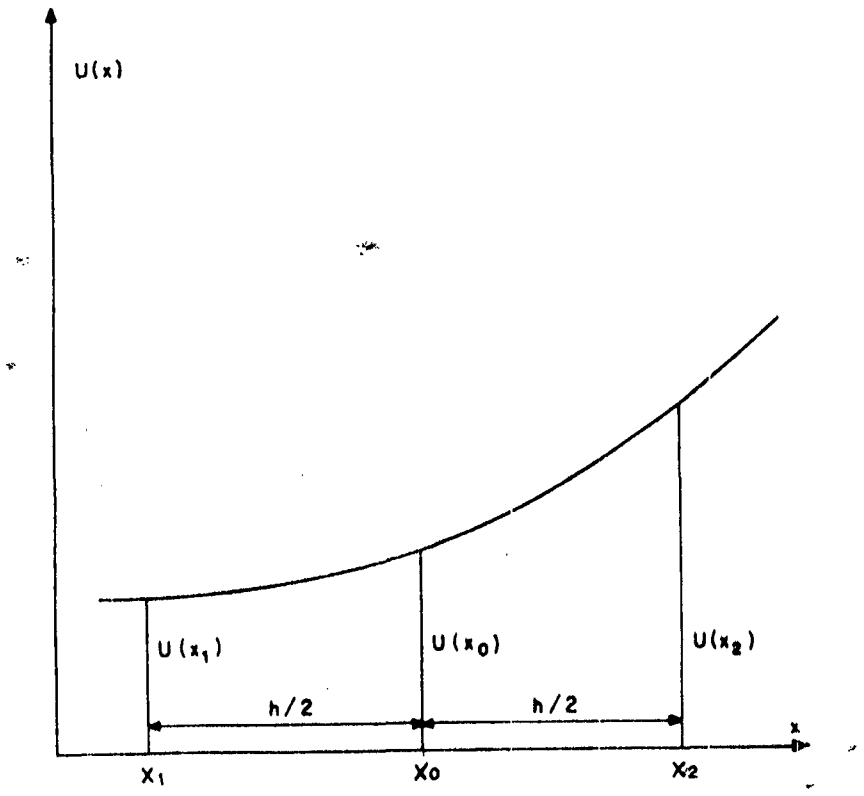

Fig. 1.3 Aproximación de $u'(x)$ por un esquema centrado

por la fórmula de Taylor

$$u(x_1) = u(x_0) - \frac{h}{2}\,u'(x_0) + \frac{1}{2!}\left(\frac{h}{2}\right)^2 u''(x_0) - \frac{1}{3!}\left(\frac{h}{2}\right)^3 u'''(\xi_1),$$

$$x_1 < \xi_1 < x_0,$$

$$u(x_2) = u(x_0) + \frac{h}{2}\,u'(x_0) + \frac{1}{2!}\left(\frac{h}{2}\right)^2 u''(x_0) + \frac{1}{3!}\left(\frac{h}{2}\right)^3 u'''(\xi_2),$$

$$x_0 < \xi_2 < x_2,$$

Restando $u(x_2)$ de $u(x_1)$ y dividiendo por h resulta

$$\frac{1}{h}\left(u(x_2) - u(x_1)\right) = u'(x_0) + \frac{1}{3!}\left(\frac{1}{2}\right)^3 h^2 \left[u'''(\xi_1) + u'''(\xi_2)\right]$$

Si tomamos módulos y reemplazamos $u'''(\xi_1) + u'''\xi_2)$ por $2 \max \left| u'''(x) \right|$ se obtiene finalmente

$$\left| \frac{1}{h}\left(u(x_2) - u(x_1)\right) - u'(x_0) \right| \leq \frac{1}{24} h^2 \max \left| u'''(x)\right| \qquad (1.5)$$

Si $u'''(x)$ está acotada entre los puntos x_1 y x_2 el error del reemplazo de $u'(x_0)$ por un esquema en diferencias centradas está acotado por una constante multiplicada por h^2 o se dice que es de orden h^2 y se escribe $O(h^2)$. Reduciendo h a la mitad la cota del error del reemplazo de $u'(x)$ por diferencias centradas se reduce en un factor 4.

Como ya lo vimos en forma geométrica los esquemas en adelanto o en atraso que aproximan $u'(x)$ son menos precisos que el esquema centrado. Refiriéndonos a la Fig. 1.1 y en el caso en que $(u(x_2) - u(x_1))/h$ aproxima a $u'(x_1)$, desarrollando en serie de Taylor alrededor de $u(x_1)$

$$u(x_2) = u(x_1) + h\, u'(x_1) + \frac{1}{2!} h^2 u''(\xi) \quad , \quad x_1 \leq \xi \leq x_2 \; ;$$

en consecuencia

$$\left| \frac{1}{h}\left(u(x_2) - u(x_1)\right) - u'(x_1) \right| \leq \frac{1}{2} h \max \left| u''(\xi) \right| \qquad (1.6)$$

Entonces el esquema en adelanto (1.2) es de $O(h)$ siempre que $u''(\xi)$ esté acotada entre x_1 y x_2. Reduciendo h a la mitad se reduce la cota del error en un factor 2. Idéntico resultado se obtiene con el análisis de la precisión del esquema en atraso (1.3).

Hasta ahora hemos visto como se pueden construir aproximaciones de $u'(x)$ de $O(h)$ y $O(h^2)$, mediante consideraciones de tipo geométrico. Vamos a presentar un método que permite obtener en forma sistemática aproximaciones de $u'(x_1)$, olvidando por un momento que la conocemos, escribiéndolas de la siguiente manera

$$u'(x_1) = \alpha\, u(x_1) + \beta\, u(x_2) \qquad (1.7)$$

donde α y β son coeficientes a ser determinados. Desarrollando $u(x_2)$ en serie de Taylor alrededor de $u(x_1)$, reemplazando en (1.7) y reordenando

$$u'(x_1) = (\alpha + \beta)\, u(x_1) + \beta\, h\, u'(x_1) + \frac{\beta\, h^2}{2} u''(\xi) \qquad (1.8)$$

Igualando los coeficientes correspondientes de (1.8) resulta $\alpha + \beta = 0$ y $\beta h = 1$ de donde $\beta = 1/h$ y $\alpha = -1/h$. Reemplazando en (1.7) se obtiene nuevamente (1.2). Con estos valores de α y β la precisión del esquema en adelanto es de $O(h)$, esto quiere decir, además, que con dichos valores se requiere que $u'(x_1)$ en (1.8) sea exacta para el término constante y lineal en u.

Entonces cualquier esquema en diferencias puede ser construído de esta manera. En el método de los coeficientes indeterminados la aproximación de la derivada $u'(x)$ es escrita como una suma ponderada de los $u(x_i)$ con coeficientes de ponderación a determinar con la condición de que el esquema sea exacto para el polinomio de mayor grado posible en $u(x_i)$.

Es instructivo presentar una aproximación en adelanto por tres puntos para $u'(x_1)$ utilizando el método de los coeficientes indeterminados. Suponiendo conocidos $u(x_1)$, $u(x_2)$ y $u(x_3)$ en los puntos x_1 , $x_2 = x_1 + h$ y $x_3 = x_2 + h$ la aproximación de $u'(x_1)$, se escribe

$$u'(x_1) = \alpha\, u(x_1) + \beta\, u(x_2) + \gamma\, u(x_3) \qquad (1.9)$$

Desarrollando $u(x_2)$ y $u(x_3)$ en serie de Taylor alrededor de $u(x_1)$, reemplazando en (1.9) y reordenando

$$u'(x_1) = (\alpha+\beta+\gamma)\, u(x_1) + (\beta h+2\gamma h)\, u'(x_1) + \left(\frac{\beta h^2}{2} + 2\gamma h^2\right) u''(x_1) + \frac{\beta h^3}{3!}\, u'''(\xi_1) + \frac{4}{3}\gamma h^3\, u'''(\xi_2), \qquad (1.10)$$

donde $x_1 < \xi_1 < x_2$ y $x_1 < \xi_2 < x_3$

Igualando los coeficientes correspondientes de (1.10) resulta $\alpha + \beta + \gamma = 0$, $\beta + 2\gamma = 1/h$ y $\beta + 4\gamma = 0$ lo que implica exigir que $u'(x_1)$ en (1.10) sea exacta para los términos constante, lineal y cuadrático. La solución del sistema anterior resulta en $\alpha = -3/2h$, $\beta = 2/h$ y $\gamma = -1/2h$. Reemplazando en (1.9) se obtiene

$$u'(x_1) = \frac{1}{2h}\,(-3\, u(x_1) + 4\, u(x_2) - u(x_3)) \qquad (1.11)$$

que es una aproximación en adelanto de tres puntos de $O(h^2)$.

Si quisieramos construir una aproximación en diferencias centradas por cuatro puntos del mayor orden posible para $u'(x_0)$ utilizando las posiciones $x_0 = 0$, $x_1 = -2h$, $x_2 = -h$, $x_3 = h$ y $x_4 = 2h$ y los valores $u(x_1)$, $u(x_2)$, $u(x_3)$ y $u(x_4)$, tendríamos que la aproximación de $u'(x_0)$ se escribe

$$u'(x_0) = \alpha\, u(x_1) + \beta\, u(x_2) + \gamma\, u(x_3) + \delta\, u(x_4) \qquad (1.12)$$

Desarrollando $u(x_1)$, $u(x_2)$, $u(x_3)$ y $u(x_4)$ en serie de Taylor alrededor de $u(x_0)$, reemplazando en (1.12) e igualando los coeficientes correspondientes obtenemos un sistema de 4 ecuaciones con las cuatro incógnitas α, β, γ y δ cuya solución es: $\alpha = (12h)^{-1}$, $\beta = -(2/3h)^{-1}$, $\gamma = (2/3h)^{-1}$ y $\delta = -(12h)^{-1}$. Reemplazando en (1.12) se obtiene

$$u'(x_0) = (12h)^{-1} (u(x_1) - 8u(x_2) + 8u(x_3) - u(x_4)) \qquad (1.13)$$

que es una aproximación por cuatro puntos centrada de $O(h^4)$.

1.2 Aproximaciones de derivadas de orden superior

Para obtener una aproximación en diferencias para una derivada segunda u"(x) hacemos una diferencia sucesiva de una derivada primera, es decir (u'(x))'. Naturalmente necesitamos, por lo menos, tres valores de u, por ejemplo, en los puntos x_1, x_2 y x_3, según se muestra en la Fig. 1.4. Llamando

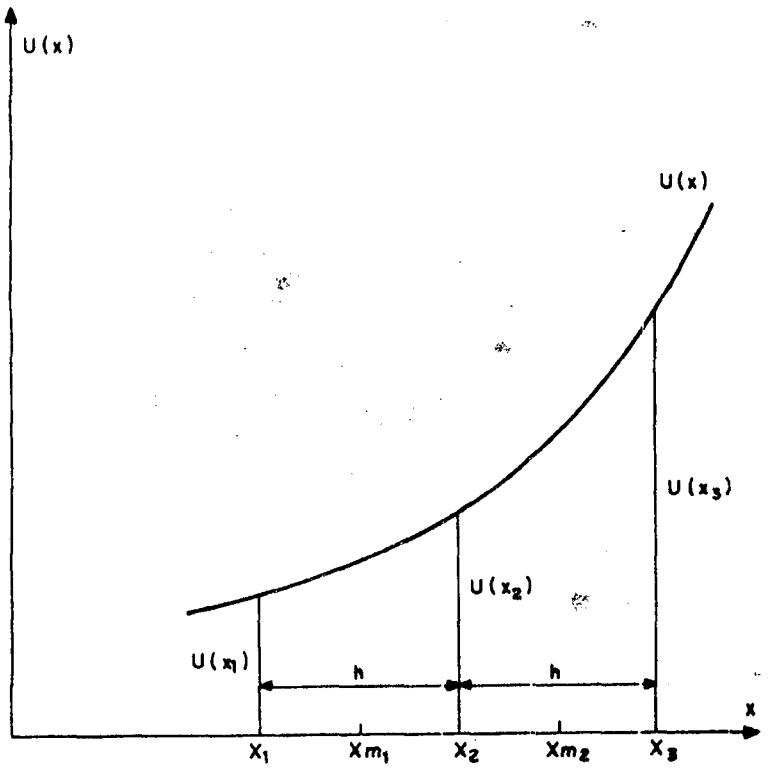

Fig. 1.4 Aproximación de la derivada segunda

$u'(xm_1)$ y $u'(xm_2)$ a los valores aproximados de la derivada primera u'(x) en los puntos xm_1 y xm_2, donde xm_1 y xm_2 son dos valores en el punto medio entre x_1 y x_2 y entre x_2 y x_3, respectivamente, tenemos que

$$u'(xm_1) = \frac{u(x_2) - u(x_1)}{h} \quad , \qquad u'(xm_2) = \frac{u(x_3) - u(x_2)}{h}$$

entonces

$$u''(x) = (u'(x))' = \frac{u'(xm_2) - u'(xm_1)}{h} = \frac{\dfrac{u(x_3)-u(x_2)}{h} - \dfrac{u(x_2)-u(x_1)}{h}}{h}$$

$$u''(x) = \frac{u(x_1) - 2\,u(x_2) + u(x_3)}{h^2} \tag{1.14}$$

Nuevamente aquí, dependiendo del punto elegido para que la ecuación (1.14) aproxime a u"(x) tendremos aproximación en adelanto, en atraso o centrada, respectivamente.

El análisis de la precisión del esquema centrado que aproxima a u"(x) es análogo al de u'(x). Refiriéndonos a la Fig. 1.4, para el análisis de la aproximación de $u''(x_2)$, expandimos $u(x_1)$ y $u(x_3)$ alrededor del $u(x_2)$ mediante la fórmula de Taylor

$$u(x_1) = u(x_2) - h\,u'(x_2) + \frac{1}{2!}h^2\,u''(x_2) - \frac{1}{3!}h^3\,u''(x_2) + \frac{1}{4!}h^4\,u^{IV}(\xi_1) \quad , \qquad x_1 < \xi_1 < x_2,$$

$$u(x_3) = u(x_2) + h\,u'(x_2) + \frac{1}{2!}h^2\,u''(x_2) + \frac{1}{3!}h^3\,u'''(x_2) + \frac{1}{4!}h^4\,u^{IV}(\xi_2) \quad , \qquad x_2 < \xi_2 < x_3,$$

Sumando ambas ecuaciones y reordenando

$$\frac{1}{h^2}(u(x_1) - 2u(x_2) + u(x_3)) = u''(x_2) + \frac{1}{4!}h^2\left[u^{IV}(\xi_1) + u^{IV}(\xi_2)\right]$$

ó

$$\left|\frac{1}{h^2}(u(x_1) - 2u(x_2) + u(x_3)) - u''(x_2)\right| \leq \frac{1}{12}h^2 \max\left|u^{IV}(x)\right| \tag{1.15}$$

La ecuación (1.15) indica que el esquema centrado que aproxima u"(x) es de $O(h^2)$ si $u^{IV}(x)$ está acotada. Si utilizáramos el es-

quema centrado que aproxima u"(x) en la modalidad en adelanto o en atraso el orden se reduciría a O(h).

2.0 SOLUCION NUMERICA DE ECUACIONES DIFERENCIALES ORDINARIAS

2.1 Introducción

Vamos a presentar una introducción a la resolución numérica de problemas de valores iniciales y de contorno en ecuaciones diferenciales ordinarias. Consideremos en primer término una ecuación diferencial de primer orden

$$\frac{du}{dt} + f(u,t) = 0 \quad , \quad 0 < t < \infty \tag{2.1}$$

donde u = u(t) es un escalar. Esta ecuación requiere una condición suplementaria , por ejemplo el valor de u en $t = t_0$, para que la solución sea única, es decir

$$u(t_0) = u_0 \tag{2.2}$$

El sistema (2.1) - (2.2) constituye un problema de valores iniciales.

Una ecuación diferencial de segundo orden (en la variable independiente x, por ejemplo) se escribe

$$\frac{d^2u}{dx^2} + f(u,u',x) = 0 \quad , \quad x_0 < x < x_n \tag{2.3}$$

donde u = u(x). Esta ecuación requiere dos condiciones suplementarias para que la solución sea única. Ahora bien, dependiendo de la manera en que las condiciones suplementarias estén definidas, se tendrán dos posibilidades. Si las dos condiciones están especificadas en un mismo punto, por ejemplo,

$$\begin{aligned} u(x_0) &= u_0 \\ u'(x_0) &= u'_0 \end{aligned} \quad , \tag{2.4}$$

el sistema (2.3) (2.4) constituye una problema de valores iniciales; si se dan dos condiciones de borde, por ejemplo, en los extremos del dominio,

$$\begin{aligned} u(x_0) &= u_0 \\ u(x_n) &= u_n \end{aligned} \quad , \tag{2.5}$$

el sistema (2.3) - (2.5) constituye un problema de contorno.

Los problemas que se presentan en la inmensa mayoría de los casos no consisten en una ecuación diferencial de primer orden sino en un sistema de p ecuaciones diferenciales de primer orden con p incógnitas. Estos se pueden escribir como

$$\begin{aligned} du_1/dt + f_1(u_1, u_2, \ldots, u_p, t) &= 0 \\ du_2/dt + f_2(u_1, u_2, \ldots, u_p, t) &= 0 \\ &\vdots \\ du_p/dt + f_p(u_1, u_2, \ldots, u_p, t) &= 0 \end{aligned} \qquad (2.6)$$

donde $0 < t < \infty$

con condiciones iniciales

$$u_i(0) = u_{i0} \quad , \quad i = 1, \ldots, p \qquad (2.7)$$

Es conveniente escribir el sistema (2.6) - (2.7) en forma vectorial

$$du/dt + f(u,t) = 0 \quad , \quad 0 < t < \infty \qquad (2.8)$$

$$u(0) = u_0$$

donde $du/dt = d/dt \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ u_p \end{bmatrix}$ $\qquad f(u,t) = \begin{bmatrix} f_1(u,t) \\ f_2(u,t) \\ \cdot \\ \cdot \\ f_p(u,t) \end{bmatrix}$;

de esta manera y como veremos más adelante, la analogía entre la aproximación numérica de una ecuación diferencial y un sistema de ecuaciones diferenciales es inmediata.

Cuando se presenta una ecuación diferencial de orden superior siempre puede ser trasformada en un sistema de ecuaciones diferenciales de primer orden. El siguiente ejemplo ilustra lo anterior: la ecuación diferencial de tercer orden

$$W''' = f(W'', W', W, t) \quad , \quad 0 < t < \infty \qquad (2.9)$$

con condiciones iniciales

$$W(0) = \alpha_1 \quad , \quad W'(0) = \alpha_2 \quad , \quad W''(0) = \alpha_3 \qquad (2.10)$$

se puede transformar en un sistema de 3 ecuaciones diferenciales de primer orden mediante la sustitución

$$u_1 = W \qquad u_2 = W' \qquad u_3 = W'' \qquad (2.11)$$

Teniendo en cuenta que $u_3 = (W')' = u'_2$ y $u_2 = (W)' = u'_1$ resulta el sistema

$$\begin{aligned} du_1/dt &= u_2 \\ du_2/dt &= u_3 \\ du_3/dt &= f(u_1,u_2,u_3,t) \end{aligned} \qquad (2.12)$$

donde $0 \leq t \leq \infty$

con condiciones iniciales

$$u_1(0) = \alpha_1 \quad , \quad u_2(0) = \alpha_2 \quad y \quad u_3(0) = \alpha_3 \qquad (2.13)$$

Naturalmente el sistema (2.12) - (2.13) también puede escribirse en forma vectorial de acuerdo a lo visto anteriormente.

En el estudio del problema de valores iniciales (2.1) - (2.2) es importante distinguir los dos tipos de ecuaciones que se presentan según sea el signo de la derivada de la función f con respecto a la variable dependiente u: si la derivada es positiva se trata de ecuaciones de decaimiento y si la derivada es negativa se trata de ecuaciones de crecimiento. Hay un tercer caso en que la derivada es imaginaria porque la variable dependiente u es compleja y se trata de un sistema de ecuaciones de tipo oscilatorio. Ilustramos lo anterior con un ejemplo muy simple que posee solución exacta: sea la ecuación (2.1) con $f(u) = a\,u$, es decir

$$\frac{du}{dt} + a\,u = 0 \qquad 0 \leq t \leq \infty \qquad (2.14)$$

$$u(t=0) = u_0 \qquad (2.15)$$

donde a es una constante distinta de cero. Para obtener la solución exacta integramos (2.14)

$$\ln u = -at + C \qquad (2.16)$$

teniendo en cuenta la condición inicial (2.15) se obtiene el valor de la constante de integración

$$C = \ln u_0$$

reemplazando el valor de C en la ecuación (2.16) y escribiendo éste en forma exponencial resulta, finalmente

$$u(t) = u_0\, e^{-at} \qquad (2.17)$$

que es la solución exacta del problema de valores iniciales (2.14) - (2.15). Si suponemos que a es positiva la expresión (2.17) describe la solución de una ecuación de decaimiento ya que la solución decrece para valores crecientes del tiempo. Ejemplos de esta ecuación se presentan en cinética química, decaimiento radiactivo, el enfriamiento de un termómetro a la temperatura ambiente, etc. En todos los casos el sistma tiende a un estado de equilibrio representado por la línea u = 0 en un gráfico u = u(t) según se ilustra en la figura 2.1. Los sistemas

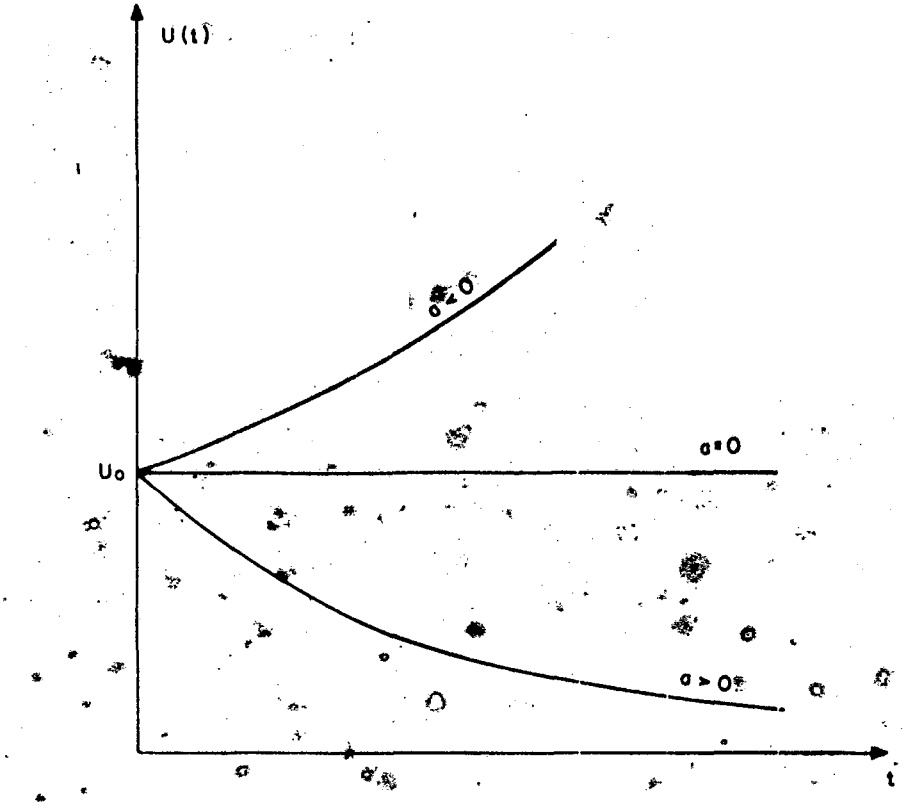

Fig. 2.1 Ejemplos de una ecuación inestable (a<0), neutra (a=0) y estable (a>0).

que presentan estas características son sistemas estables. Si suponemos que a es una constante negativa la ecuación (2.14) deviene $du/dt - a\,u = 0$ y la solución exacta (2.17) resulta entonces

$$u(t) = u_0\, e^{at} \qquad (2.18)$$

La expresión (2.18) describe en este caso una ecuación de crecimiento. Ejemplos de esta ecuación se presentan en la descripción de una ley de crecimiento demográfico (Ley de Malthus), crecimiento de bacterias, aumento del dinero en interés compuesto, etc. En todos los casos la solución crece en valor absoluto para valores crecientes del tiempo y se dice que el sistema es inestable.

De aquí en adelante vamos a suponer que el problema diferencial ya sea de valores iniciales o de contorno está correctamente propuesto en el sentido de Hadamard. Esto quiere decir que la solución existe, es única y depende en forma continua de las condiciones iniciales o de contorno.

Vamos a comenzar con el estudio de las soluciones numéricas del problema escalar de valores iniciales. Como estamos interesados en problemas físicos de valores iniciales que dependen del tiempo, la variable independiente t en el sistema (2.1) - (2.2) indicará el tiempo. Entonces vamos a integrar el problema de valores iniciales en el tiempo utilizando pequeños incrementos del mismo. El intervalo temporal $t_o \leq t \leq t_n$ es subdividido uniformemente en N segmentos o pasos de tiempo de longitud $k = (t_n - t_0)/N$ como se muestra en la figura 2.2. La unión de

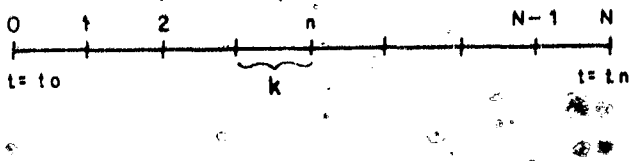

Fig. 2.2 Discretización del intervalo espacial

éstos segmentos determina los puntos nodales que se numeran secuencialmente 0,1,2,....,n,.....,N. Los valores discretos o computados de u(t) y f(u,t) en los nodos de la malla o red así determinada se denominan u^n y f^n, respectivamente, donde n es un nodo genérico. Entonces $t^n = t_0 + nk$, n = 0,..,N. A los efectos de distinguir entre las variables independientes temporal y espacial es común indicar un determinado nivel de tiempo con un supraíndice, dejando el subíndice para los niveles espaciales. Como veremos, esta convención es muy útil en la solución numérica de ecuaciones en derivadas parciales. Entonces con u^n se in-

dica el valor de la solución numérica er el nivel temporal $t^n = t_0 + nk$. Resolver numéricamente el problema de valores iniciales (2.1) significa encontrar los valores de u^n, en los nodos de la malla del dominio to $\leq t^n \leq t_n$, que satisfacen en forma aproximada la ecuación diferencial (2.1) y la condición inicial (2.2). Una manera muy sencilla de introducir los métodos de aproximaciones de problemas de valores iniciales es mediante la escritura del problema de valores iniciales (2.1) - (2.2) en su forma equivalente integral y su posterior aproximación por fórmulas de cuadratura. Entonces, integrando (2.1) en el subintervalo genérico $t^n \leq t \leq t^{n+1}$, $k = t^{n+1} - t^n$, y teniendo en cuenta la condición inicial u^n para $t = t^n$

$$u^{n+1} = u^n - \int_{t^n}^{t^{n+1}} f(u,t)\, dt \qquad (2.19)$$

Naturalmente la integral de (2.19) no puede ser evaluada en forma exacta dado que no se conoce el valor de u(t) incógnita del problema para todo el intervalo, $t^n \leq t \leq t^{n+1}$. Precisamente el método de las diferencias finitas asume diferentes aproximaciones o fórmulas de cuadratura para evaluar la integral de la ecuación (2.19) lo queda lugar a distintos métodos numéricos. Con el fin de unificar la nomenclatura con la existente para la solución numérica de ecuaciones en derivadas parciales vamos a clasificar a los métodos numéricos para ecuaciones diferenciales ordinarias en métodos de paso entero y métodos de paso fraccionario. Los métodos de paso entero a su vez pueden ser divididos en métodos de un sólo paso y métodos de paso múltiple; en los primeros la solución en el nivel temporal t^{n+1} se obtiene a partir de la solución en el nivel t^n en un sólo paso de tiempo, en los últimos la solución en t^{n+1} se obtiene utilizando la solución en los niveles t^n, t^{n-1}, t^{n-2}, etc, en un solo paso de tiempo. En los métodos de paso fraccionario la solución en t^{n+1} se obtiene a partir de la solución en t^n en dos o más pasos intermedios de tiempo. En la nomenclatura clásica de ecuaciones diferenciales ordinarias los métodos de paso fraccionario se los suele llamar métodos de tipo predictor-corrector (en su versión para un solo paso). Tanto los métodos de paso entero como los de paso fraccionario pueden ser implícitos o explícitos dependiendo que se utilice o no, en la fórmula de cuadratura, el borde superior del intervalo.

2.2 El método de Euler

La manera más sencilla de aproximar la integral de la ecuación (2.19) es suponer el valor de la función f constante para todo el intervalo $t^n \leq t \leq t^{n+1}$ e igual al valor de f en el nivel temporal t^n, según se observa en la figura 2.3, resultando

$$u^{n+1} = u^n - k\, f^n \qquad (2.20)$$

donde $f^n = f(u^n, t^n)$. La expresión (2.20) basada en una aproximación en adelanto para u, es un esquema de paso entero, explícito, y se lo denomina método de Euler.

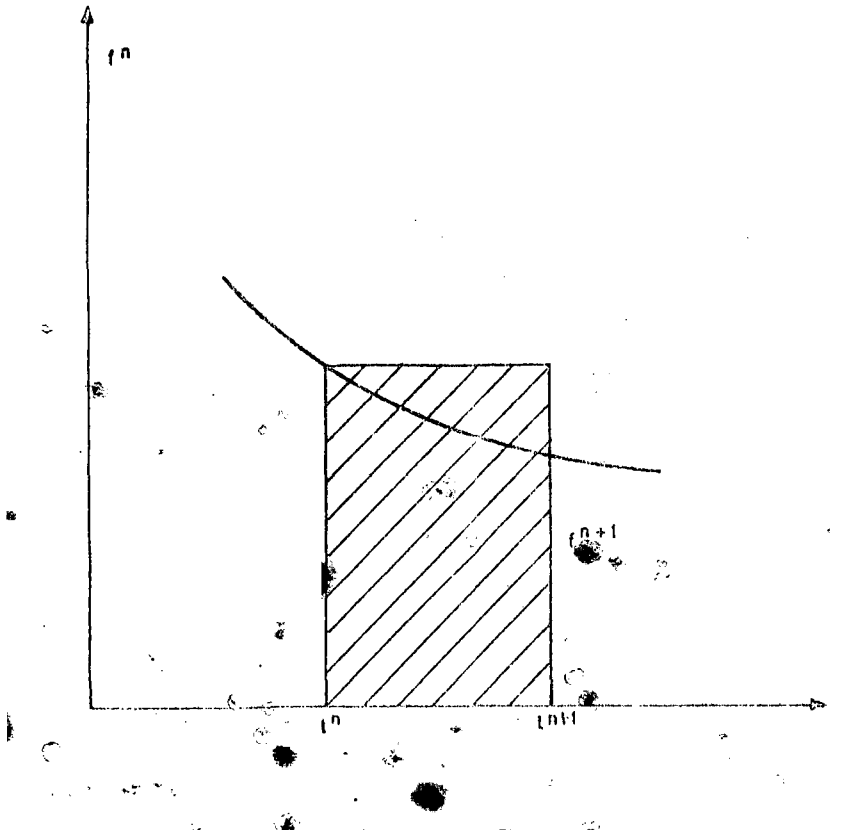

Fig. 2.3 Aproximación de la integral de la ecuación (2.19) en el método de Euler.

Existe otra manera muy sencilla de obtener el método de Euler que ilustraremos con un ejemplo. La aproximación de la ecuación de decaimiento $du/dt + a\,u = 0$ con el esquema de Euler resulta en

$$u^{n+1} = (1 - k\,a)\,u^n \qquad (2.21)$$

Esto puede ser interpretado como la solución exacta truncada de la ecuación de decaimiento con u^n como condición inicial. En efecto, la solución exacta de la ecuación de decaimiento es

$$u(t) = e^{-at}\,u_0 \qquad (2.22)$$

o en el intervalo t^n, $t^n + k$

$$u^{n+1} = e^{-ak}\,u^n \qquad (2.23)$$

Recordando que

$$e^{-ak} = 1 - k\,a + \frac{1}{2}\,k^2\,a^2 - \frac{1}{6}\,k^3\,a^3 + \ldots\ldots \qquad (2.24)$$

para k tendiendo a cero podemos escribir, en forma aproximada, que

$$e^{-ak} = 1 - k\,a \qquad (2.25)$$

que reemplazada en la ecuación (2.23) reproduce la expresión (2.21).

La retención del término de segundo orden en el desarrollo en serie de potencias de e^{-ak} sugiere un método de aproximación más preciso que el método de Euler, es decir

$$u^{n+1} = (1 - k\,a + \frac{1}{2} k^2 a^2)\, u^n \qquad (2.26)$$

más adelante veremos que la expresión (2.26) corresponde a un método predictor-corrector explícito.

2.3 Nociones de consistencia, estabilidad y convergencia

La validez de un método numérico está basada en que se cumplan ciertos requerimientos naturales de las aproximaciones en diferencias finitas: consistencia, estabilidad y convergencia. La primera condición que se exige de un método de aproximación en diferencias es que la ecuación en diferencias aproxime en un cierto sentido a la ecuación diferencial. Se dice que un esquema en diferencias que aproxima una ecuación diferencial es consistente con la misma si al refinar la malla (el paso temporal k tiende a cero), la ecuación en diferencias converge a la ecuación diferencial correcta. Está claro que la propiedad de consistencia se refiere exclusivamente a la analogía existente entre el operador diferencial y el operador en diferencias, nada se dice acerca de si los resultados de la aproximación en diferencias están próximos, en un cierto sentido, a la solución del problema diferencial; este tema pertenece al estudio de la convergencia.

La convergencia puede estudiarse analizando el comportamiento del error, medido como la diferencia entre la solución numérica y la solución exacta, cuando la malla es refinada. Diremos que el método de diferencias finitas es convergente si el error tiende a cero cuando la malla es refinada. Ahora bien, existen dos fuentes de errores que inciden en la precisión de la solución numérica tomada como aproximación de la solución exacta del problema diferencial; estos son el error de truncamiento y el error de redondeo. El error de truncamiento proviene de reemplazar el problema diferencial por una aproximación en diferencias. La magnitud del mismo puede ser determinada perfectamente y es función del tamaño de malla utilizada. El error de redondeo proviene de la manera en que una computadora "redondea" una variable cuyo número de cifras significativas es mayor que la longitud de palabra de dicha máquina. La magnitud del error de redondeo es principalmente una característica del tipo de computadora digital utilizada y su predicción puede ser estudiada por

métodos estadísticos. En general, la acumulación del error de redondeo no presenta problemas muy serios. No obstante ello, es posible que la presencia de errores de redondeo o de cualquier otro tipo de error, pueda producir la inestabilidad numérica, es decir el crecimiento sin control (no acotado) de dichos errores.

El estudio de la estabilidad numérica puede realizarse mediante el análisis del comportamiento del error medido como la diferencia entre la solución numérica y la solución exacta, cuando para un paso de tiempo k fijo, el número de pasos temporales tiende a infinito. Evidentemente se espera, bajo estas condiciones, que el error no se amplifique de tal manera que los resultados numéricos dejen de tener sentido. Diremos entonces que un método en diferencias finitas es numéricamente estable si la amplificación del error permanece acotada. Es obvio que para que un método numérico sea útil debe ser estable.

Los conceptos de estabilidad y convergencia están estrechamente vinculados: el teorema de equivalencia o teorema de Lax establece que, bajo determinadas condiciones, la estabilidad de un esquema de diferencias finitas implica la convergencia del mismo (más adelante volveremos sobre el tema).

2.4 Análisis de la consistencia en el método de Euler

Para el estudio de la consistencia es necesario analizar primero el error de truncamiento local ó error de discretización r^n que se produce cuando en el esquema en diferencias se reemplazan los valores aproximados de u^n por los valores exactos U^n . En el caso del método de Euler este resulta.

$$r^{n+1} = \frac{U^{n+1} - U^n}{k} + f(U^n, t^n) \qquad (2.27)$$

Es muy frecuente encontrar la relación (2.27) escrita como

$$U^{n+1} = U^n - k\, f(U^n, t^n) + k\, r^{n+1} \qquad (2.28)$$

Desarrollando en serie de Taylor alrededor de U^n

$$U^{n+1} = U^n + k\, dU(n)/dt + \frac{k^2}{2}\, d^2U(\xi)/dt^2 \qquad (2.29)$$

donde ξ es un punto intermedio entre t^n y t^{n+1}. Reemplazando en la expresión (2.27) resulta

$$r^{n+1} = \left[dU(n)/dt + f(U^n, t^n)\right] + \frac{1}{2}\, k\, d^2U(\xi)/dt^2 \qquad (2.30)$$

Dado que U_n es el valor exacto de U(t) en el nivel temporal $t_n = n\,k$, U_n satisface la ecuación diferencial $du/dt + f(u,t) = 0$ y por lo tanto el término entre corchetes se anula. Entonces

el residuo o error de truncamiento local $r^n = O(k)$ siempre que $d^2U(\xi)/dt^2$ sea menor que una constante entre los niveles t^n y t^{n+1}. Si $r^{n+1} \to 0$ cuando $k \to 0$ se dice que el esquema en diferencias es consistente con el problema diferencial primitivo. En este caso la consistencia del método de Euler es de O(k).

2.5 Propagación de la error

El problema de valores iniciales $du/dt + f(u,t) = 0$ con condiciones inciales $u(0) = c$ puede interpretarse como que determina la velocidad de una partícula como función de la posición u. La solución del mismo puede ser considerada como una función $u(t,c)$ donde c es un conjunto de condiciones iniciales que determina una familia de soluciones en el plano (u,t) llamadas trayectorias. Por ejemplo en el caso de la ecuación $du/dt - a\,u = 0$, $u(0) = c$, la solución es la familia $u(t,c) = c\,e^{at}$ como se ilustra en la Fig. 2.4. La elección de un

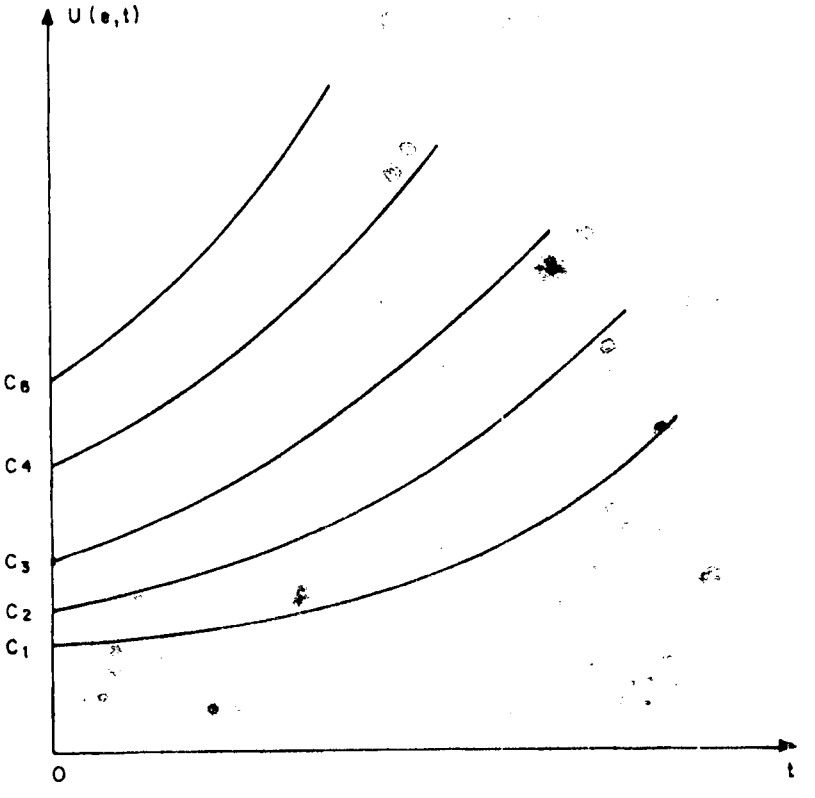

Fig. 2.4 Familia de soluciones de $du/dt - a\,u = 0$ con $u(0) = c$

valor particular de c servirá para seleccionar una de las trayectorias de la familia.

En la solución numérica del problema de valores iniciales una perturbación de las condiciones iniciales, por ejemplo,

un error de redondeo en el valor de la constante c, implica que u(t) es forzada a seguir otra trayectoria de la familia de soluciones. En los métodos numéricos de un sólo paso, por ejemplo, la solución en el nivel temporal t^{n+1} es calculada utilizando como valor inicial la solución en el nivel t^n y este proceso se repite para valores de n crecientes. En cada paso de tiempo se produce una pequeña perturbación, error de redondeo o error local de truncamiento, que produce un fenómeno similar de transición a otra trayectoria de la familia de soluciones. En la Fig. 2.5 se presenta una situación muy exagerada de lo que acontece normalmente en el cálculo. Si a medida que t aumenta las

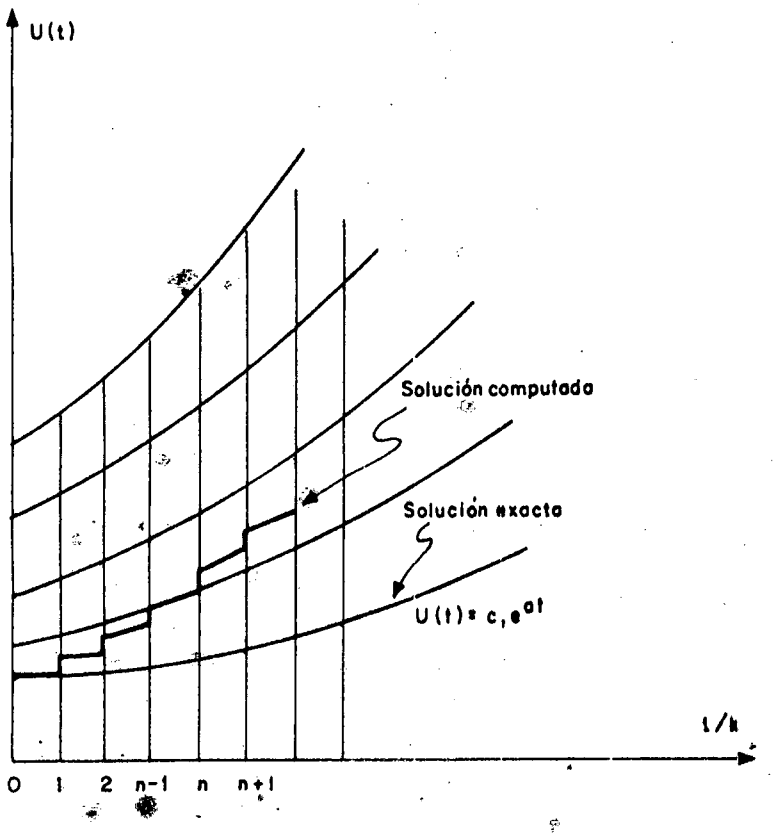

Fig. 2.5 Propagación del error durante la integración numérica

curvas de la familia de soluciones se alejan entre sí rápidamente, entonces el problema de valores iniciales es inestable; en el caso contrario el problema es estable.

Es importante distinguir entre el error global y el error local de truncamiento. El error global en el tiempo t^{n+1} es $u^{n+1} - U^{n+1}$, donde U^{n+1} es la solución exacta en t^{n+1} del problema de valores iniciales genuino ($U(t) = c_1 e^{at}$ en la Fig. 2.5). El error local en t^{n+1} es la diferencia entre el valor computado u^{n+1} y el valor en t^{n+1} de aquella solución de la ecuación diferencial que pasa por el punto (u^n, t^n). Para ser más claro, los errores locales son los saltos en la curva escalera de la figura 2.5 (Dahlquist y Bjorck, 1974).

2.6 Análisis de la convergencia en el método de Euler

Vamos a restringir nuestro análisis de la convergencia del método de Euler aplicado a la ecuación de decaimiento $du/dt + au = 0$, $u(0) = u_0 > 0$, $t > 0$. Distinguiremos entre el valor exacto de la solución que denominaremos U^n y el valor aproximado de la solución numérica que llamaremos u^n. Entre los niveles de tiempo n y n+1, por medio de un desarrollo en serie de Taylor se obtiene

$$U^{n+1} = U^n + k\, dU^n/dt + \frac{1}{2} k^2\, d^2U(\xi)/dt^2 \qquad (2.31)$$

donde ξ es un nivel intermedio de tiempo entre n y n+1. Dado que U satisface la ecuación diferencial $dU(n)/dt = -a\, U^n$ y $d^2U(\xi) = a^2\, U(\xi)$. Reemplazando en la ecuación anterior

$$U^{n+1} = (1 - a\, k)\, U^n + \frac{1}{2} k^2\, a^2\, U(\xi) \qquad (2.32)$$

Teniendo en cuenta que la solución exacta es decreciente monotónica,

$$U^{n+1} < U(\xi) < U^n \qquad (2.33)$$

Por conveniencia introducimos la notación

$$\sigma^n = \frac{1}{2} \frac{U(\xi)}{U^n} a^2\, k^2 < \frac{1}{2} a^2\, k^2 \qquad (2.34)$$

$$w = 1 - a\, k$$

que reemplazada en la ecuación (2.32) resulta

$$U^{n+1} = w\, U^n + \sigma^n\, U^n \qquad (2.36)$$

Recordando que el esquema de Euler para la ecuación de decaimiento se escribe

$$u^{n+1} = w\, u^n \quad , \quad w = 1 - a\, k \qquad (2.37)$$

comparando esta expresión con la fórmula (2.36) se deduce que el término $\sigma^n\, U^n$ de esta última es el residuo del esquema de Euler que aquí es de $O(a^2k^2)$. Restando la ecuación (2.37) de la (2.36) se obtiene una fórmula de recurrencia para el error $e^n = U^n - u^n$, es decir

$$e^{n+1} = w\, e^n + \sigma^n\, U^n \qquad (2.38)$$

Si analizamos el error desde $e^0 = U^0 - u^0$ hasta e^n comprobamos que

$$e^n = \sigma^0\, w^{n-1}\, U^0 + \sigma^1\, w^{n-2}\, U^1 + \ldots + \sigma^{n-1}\, U^{n-1} \qquad (2.39)$$

Para valores de k suficientemente pequeños o $< w < 1$ por lo que todos los términos de la ecuación (2.39) son positivos y dado que $\sigma^n < 1/2\ a^2\ k^2$ resulta que

$$e^n < \frac{1}{2} a^2 k^2 (w^{n-1} U^0 + w^{n-2} U^1 + ... + U^{n-1}) \qquad (2.40)$$

Además dado que $\sigma^n U^n > 0$, la ecuación (2.36) indica que $U^{n+1} > w U^n$ de modo que e^n en la ecuación (2.40) deviene

$$e^n < \frac{1}{2} a^2 k^2 n U^{n-1} \qquad (2.41)$$

Dado que $n k = t$ y $U^{n-1} = U^0 e^{-a(n-1)k} \cong U^0 e^{-at}$ obtendremos finalmente una cota del error

$$e^n < \frac{1}{2} U^0 a^2 k t e^{-at} \qquad (2.42)$$

Para cualquier tiempo fijo t, $e^n \to 0$ para $k \to 0$. El error de truncamiento en el esquema de Euler es de $O(k^2)$, pero luego de ser sumados los errores sobre n pasos de tiempo de longitud k, el error $e^n = U^n - u^n$ en u^n se transforma en $0(k)$. Sin embargo y como veremos a continuación es la estabilidad numérica la que gobierna el tamaño de grilla a utilizar.

2.7 Análisis de la estabilidad en el método de Euler

Utilizando el ejemplo del método de Euler estudiaremos bajo qué condiciones un método numérico para problemas de valores iniciales es estable. Para ello supondremos la existencia de un pequeño error ε^n o perturbación de la variable dependiente u^n en el tiempo t^n y trataremos de determinar el error que se produce en la variable dependiente u^{n+1} como consecuencia del mismo, a través de una linealización del problema. Escribamos nuevamente el problema de valores iniciales (2.1) - (2.2) para el caso de una ecuación escalar.

$$\frac{du}{dt} + f(u,t) = 0 \qquad (2.43)$$

El método de Euler se escribe entonces

$$u^{n+1} = u^n - k f(u^n, t^n) \qquad (2.44)$$

En consecuencia

$$u^{n+1} + \varepsilon^{n+1} = u^n + \varepsilon^n - k f(u^n - \varepsilon^n, t^n) \qquad (2.45)$$

Como hemos supuesto que el error es pequeño es válido el siguiente desarrollo en serie de Taylor

$$f(u^n + \varepsilon^n, t^n) = f(u^n, t^n) + \varepsilon^n \, (\partial f/\partial u)_n + O(\varepsilon^n) \qquad (2.46)$$

Reemplazando (2.46) en (2.45) y restando al resultado la ecuación (2.44), obtenemos

$$\varepsilon^{n+1} = \varepsilon^n - k \, (\partial f/\partial u)_n \, \varepsilon^n + O(\varepsilon^n) \qquad (2.47)$$

La expresión (2.47) representa la ecuación para la amplificación de un error pequeño, que escribimos como

$$\varepsilon^{n+1} = \left[1 - k \, (\partial f/\partial u)_n\right] \varepsilon^n + O(\varepsilon^n) \qquad (2.48)$$

El término entre corchetes

$$g = 1 - k \, (\partial f/\partial u)_n \qquad (2.49)$$

se denomina factor de amplificación. La expresión (2.48) puede ser escrita, despreciando términos de orden superior, como

$$\varepsilon^{n+1} = g \, \varepsilon^n \qquad (2.50)$$

Para que el método numérico sea estable es necesario que el error no se amplifique del paso t^n al t^{n+1}; para ello se requiere que

$$| \varepsilon^{n+1} | < | \varepsilon^n | \qquad (2.51)$$

Entonces teniendo en cuenta la ecuación (2.50) la estabilidad numérica exige que se cumpla que

$$| g | < 1 \qquad (2.52)$$

Conviene hacer notar que la condición de estabilidad (2.52) no tiene en cuenta aquellos casos en que la solución exacta del problema de valores iniciales admite un crecimiento exponencial. Para estos casos es necesario establecer una condición de estabilidad que permita el crecimiento no acotado del error siempre que este sea menor al crecimiento de la solución exacta (más adelante volveremos sobre el tema).

Para el método de Euler la condición de estabilidad (2.52) exige que

$$-1 < 1 - k \, (\partial f/\partial u)_n < 1 \qquad (2.53)$$

Naturalmente, la ecuación (2.53) está ligada a la ecuación diferencial primitiva a través de la derivada f con respecto a u.

En el caso de que estemos tratando con una ecuación de decaimiento la condición de estabilidad en el método de Euler exige que

$$k < \frac{2}{(\partial f/\partial u)_n} \qquad (2.54)$$

expresión que se obtiene de la fórmula (2.53). La ecuación (2.54) nos dice que existe una cota superior para el valor del paso temporal y que si dicha cota es superada, de acuerdo al análisis lineal de la estabilidad, cualquier error se incrementará en forma no acotada. Diremos en este caso que el método es inestable.

La aplicación del concepto de estabilidad puede extenderse a un sistema de ecuaciones diferenciales ordinarias en cuyo caso se obtiene en lugar del factor de amplificación una matriz de amplificación cuyo radio espectral (el mayor de sus autovalores) cumple una función similar a la derivada de f con respecto a u en el caso escalar.

El siguiente ejemplo ilustra lo dicho más arriba. Si aplicamos el esquema de Euler a la ecuación de decaimiento $du/dt + a\,u = 0$ se obtiene

$$u^{n+1} = (1 - a\,k)\,u^n \qquad (2.55)$$

El factor de amplificaciones resulta

$$g = 1 - a\,k \qquad (2.56)$$

Para visualizar el factor de amplificación definimos una nueva variable $y = a\,k$ y en la Fig. 2.6 presentamos el gráfico de g(y) como función del argumento y. Aquí se ve que los valores estables de y están comprendidos en el intervalo $0 < y < 2$. La condición de estabilidad exige entonces que

$$k < \frac{2}{a} \qquad (2.57)$$

Suponiendo que $a = 1$ y $u_0 = 1$ en la tabla 2.1 y 2.2 se muestran los resultados obtenidos utilizando el esquema (2.55). La tabla 2.1 contiene tres conjuntos de valores correspondientes a diferentes tamaños de grilla ($k = 0.1$, 0.2, y 0.9), dentro del límite de estabilidad determinada por la ecuación (2.57) para $a = 1$, juntamente con la solución exacta. Estos resultados son para los primeros 10 pasos de tiempo. Si comparamos los resultados, en el mismo tiempo, para diferentes pasos de tiempo vemos que el error entre la solución exacta y la numérica decrece en

forma lineal para valores de k decrecientes, como era de esperar. En la tabla 2.2 tenemos los resultados para valores de k superiores al límite de estabilidad. Naturalmente estos resultados no tienen ninguna relación con la solución exacta. No obstante ello, formalmente existe una diferencia entre los resultados de los dos primeras columnas y los correspondientes a la tercera: en los primeros la solución es estable (oscilante pero acotada para k = 2), el error se mantiene acotado; en la última columna el error crece en forma no acotada. El primer caso corresponde a una estabilidad neutra mientras que el segundo caso corresponde a una franca inestabilidad. La conclusión principal de los resultados numéricos presentados es que la estabilidad numérica es una propiedad del método de diferencias finitas utilizado. En el caso particular del método de Euler diremos que es un método numérico condicionalmente estable. Esta es una característica de todos los métodos explícitos de paso entero.

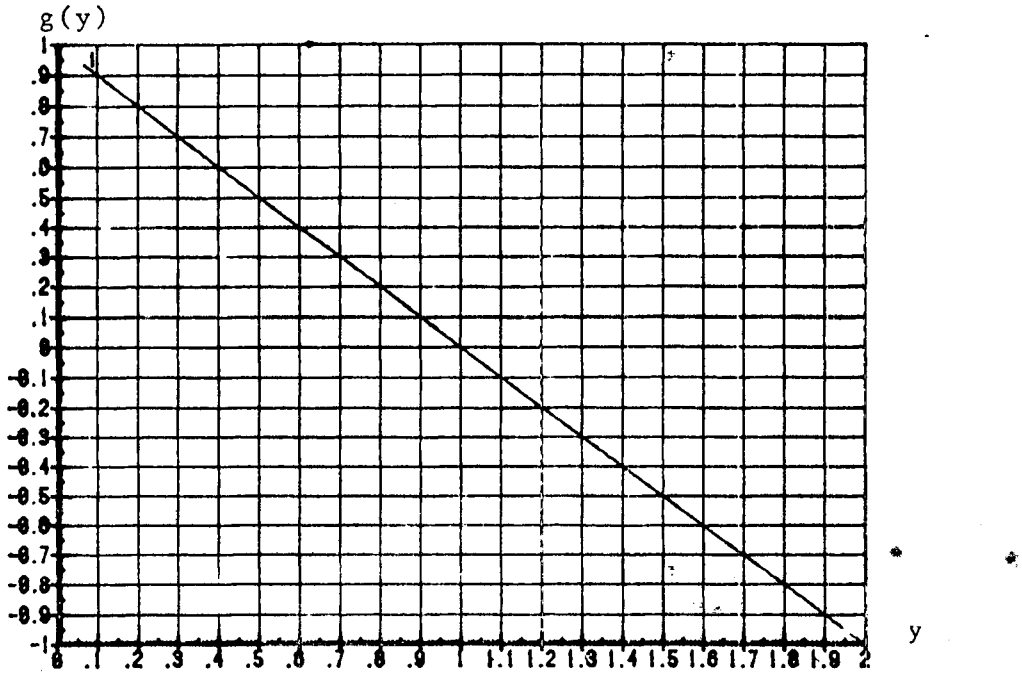

Fig. 2.6 Factor de amplificación en el método de Euler

Conviene hacer notar que el factor de amplificación g es función del tamaño de grilla. Para valores de k tendiendo a cero, el error de truncamiento local tiende a cero y el factor de amplificación tiende al valor unidad. Entonces está claro que cuanto más se acerque g a la unidad como cota superior, más precisos serán los resultados numéricos. Esto se puede observar en la tabla 2.3 donde se comparan para un mismo tiempo, t = 1, las

soluciones numérica y exacta, para valores decrecientes del factor de amplificación (g = 1 - k). Estos datos fueron tomados (de derecha a izquierda) de la tabla 2.1.

Tabla 2.1 Soluciones estables de la ecuación de decaimiento utilizando el método de Euler

n	k = 0.1		k = 0.2		k = 0.9		t	Solución Exacta $U = e^{-t}$
	t=n k	u^n	t=n k	u^n	t=n k	u^n		
1	0.1	0.9000	0.2	0.8000	0.9	0.1000	0.1	0.9048
2	0.2	0.8100	0.4	0.6400	1.8	0.0100	0.2	0.8187
3	0.3	0.7290	0.6	0.5120	2.7	0.0010	0.3	0.7408
4	0.4	0.6561	0.8	0.4096	3.6	0.0001	0.4	0.6703
5	0.5	0.5905	1.0	0.3277	4.5	0.0000	0.5	0.6065
6	0.6	0.5314	1.2	0.2621	5.4	0.0000	0.6	0.5488
7	0.7	0.4783	1.4	0.2097	6.3	0.0000	0.7	0.4966
8	0.8	0.4305	1.6	0.1678	7.2	0.0000	0.8	0.4493
9	0.9	0.3874	1.8	0.1342	8.1	0.0000	0.9	0.4066
10	1.0	0.3487	2.0	0.1074	9.0	0.0000	1	0.3679

Tabla 2.2 Soluciones neutralmente estables e inestables de la ecuación de decaimiento utilizando el método de Euler

n	k = 1		k = 2		k = 11	
	t=n k	U^n	t=n k	U^n	t=n k	U^n
1	2	0.	4	-1.	11	-10^1
2	4	0.	8	1.	22	10^2
3	6	0.	12	-1.	33	-10^3
4	8	0.	16	1.	44	10^4
5	10	0.	20	-1.	55	-10^5
6	12	0.	24	1.	66	10^6
7	14	0.	28	-1.	77	-10^7
8	16	0.	32	1.	88	10^8
9	18	0.	36	-1.	99	-10^9
10	20	0.	40	1.	110	10^{10}

Tabla 2.3 Comparación de las soluciones numéricas y exactas para t = 1

	k = 0.9 g = 0.1	k = 0.2 g = 0.8	k = 0.1 g = 0.9	Solución Exacta
U(t=1)	0.0900	0.3277	0.3487	0.3679

Hasta aquí nos hemos referido a la ecuación de decaimiento. Cuando la constante a es negativa hemos visto que la solución del problema de valores iniciales es inestable (algunos autores suelen denominarlas ecuaciones intrínsicamente inestables). Si aplicamos el método de Euler a la ecuación de crecimiento comprobaremos, por un análisis lineal de la estabilidad, que el factor de amplificaciones es siempre mayor que la unidad, para cualquier tamaño de malla. Esto puede deducirse fácilmente de la condición (2.53) que en este caso se transforma en

$$|g| = |1 + a\,k| > 1 \qquad (2.58)$$

Entonces el método de Euler aplicado a una ecuación de crecimiento produce errores que se amplifican en forma no acotada para tiempos crecientes. Es más, podemos decir que cualquier método de diferencias finitas aplicado a una ecuación de crecimiento produce errores que se amplifican en forma no acotada. Sin embargo, la misma solución exacta de la ecuación de crecimiento crece en forma no acotada y una condición de estabilidad como la representada por (2.52) no permite dicho crecimiento; de hecho no se puede satisfacer la condición de estabilidad (2.52) sin violar la consistencia. Este dilema queda resuelto introduciendo la condición de estabilidad de Von Neumann (ver Richtmyer y Morton, 1967) que establece que

$$|g| < 1 + 0(k) \qquad (2.59)$$

Esta condición que es más generosa que la (2.52) permite el crecimiento no acotado legítimo. Volveremos sobre este tema más adelante.

En la práctica la solución de una ecuación inestable implica forzosamente la utilización de una malla muy fina es decir con valores del factor de amplificación cercanos a la unidad si se quieren obtener resultados razonables.

2.8 Esquema fuertemente implícito

El método de Euler que acabamos de presentar es un método explícito condicionalmente estable. En determinados problemas es conveniente utilizar métodos de paso entero implícitos que no posean restricciones de estabilidad.

Para obtener un método implícito aproximamos la integral de la ecuación (2.19) suponiendo el valor de f constante para todo el intervalo $t^n < t < t^{n+1}$ e igual al valor de f en el nivel temporal t^{n+1}, según se ilustra en la figura 2.7, resultando

$$u^{n+1} = u^n - k\,f^{n+1} \qquad (2.60)$$

donde $f^{n+1} = f(u^{n+1}, t^{n+1})$. La expresión (2.60) es un esquema de paso entero, fuertemente implícito (f se calcula solamente en t^{n+1}) y con una aproximación en adelanto para u; posee primer orden de precisión y se denomina método fuertemente implícito.

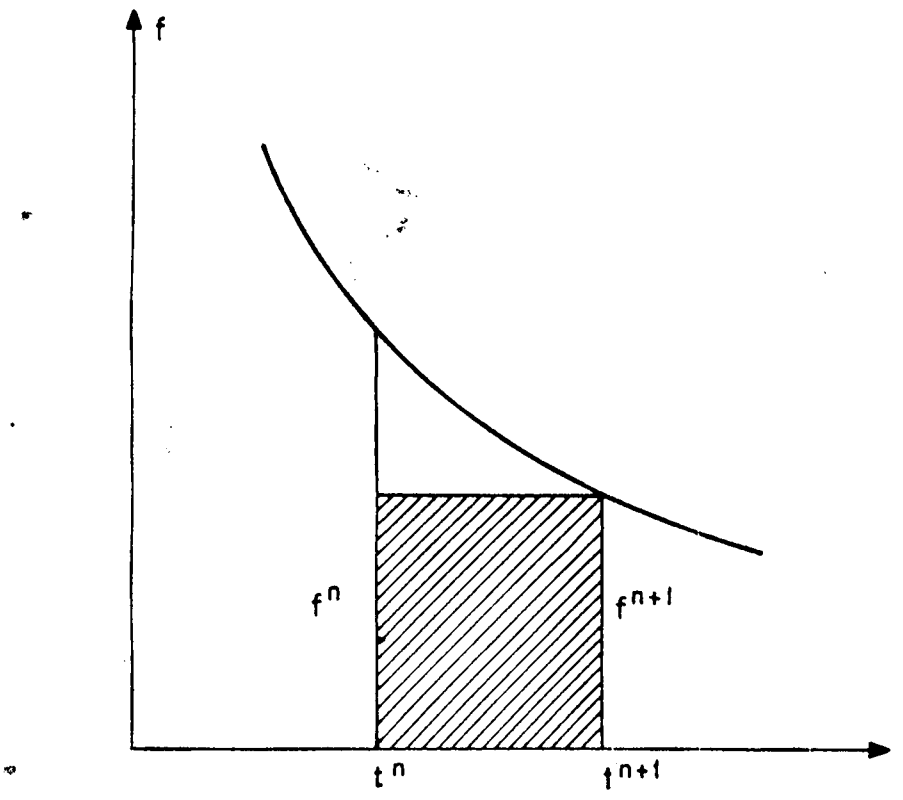

Fig. 2.7 Aproximación de la integral t^{n+1} de la ecuación (2.19) en el método fuertemente implícito

Vamos a realizar un análisis lineal de la estabilidad de este método utilizando la técnica de perturbación ya vista. Entonces introduciendo un pequeño error ε^n en la ecuación (2.60)

$$u^{n+1} + \varepsilon^{n+1} = u^n + \varepsilon^n - k\, f(u^{n+1} + \varepsilon^{n+1}, t^{n+1}) \qquad (2.61)$$

Por medio de un desarrollo en serie de Taylor se obtiene

$$f(u^{n+1}+\varepsilon^{n+1}, t^{n+1}) = f(u^{n+1}, t^{n+1}) + \varepsilon^{n+1}\,(M)_{n+1} + T.O.S. \qquad (2.62)$$

donde M es la derivada parcial de f con respecto a u. Reemplazando (2.62) en (2.61) y luego restándole la ecuación (2.60) resulta

$$\varepsilon^{n+1} = \frac{1}{1 + k\,(M)_{n+1}}\ \varepsilon^n \qquad (2.63)$$

El factor de amplificación es entonces

$$g = \frac{1}{1 + k\ (M)_{n+1}} \qquad (2.64)$$

Para ecuaciones de decaimiento $M > 0$ y el módulo del factor de amplificación es siempre menor que la unidad, para cualquier tamaño de malla; por lo tanto, el método fuertemente implícito es incondicionalmente estable. La mayor estabilidad del método se logra a costa de una mayor complicación del cálculo: en general, si f(u,t) es una función complicada, la obtención de u^{n+1} en forma implícita implica recurrir a un proceso iterativo.

En el caso de la ecuación de decaimiento $du/dt + a\ u = 0$ el factor de amplificación resulta

$$g = \frac{1}{1 + k\ a} \qquad (2.65)$$

Definiendo una nueva variable $y = k\ a$, la figura 2.8 nos muestra el gráfico de g(y) en función de y.

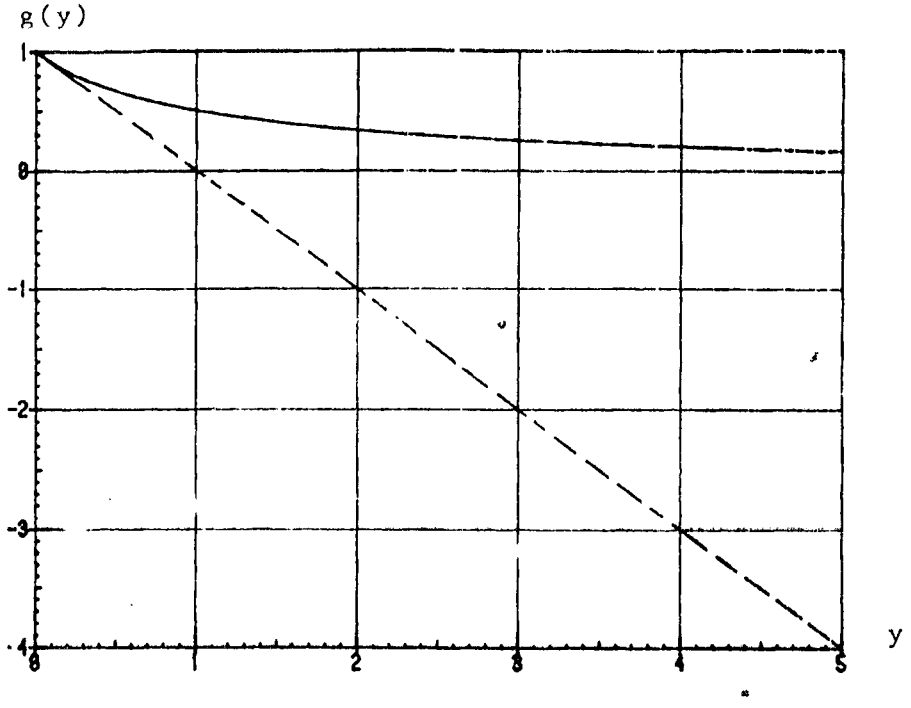

Fig. 2.8 Variación del factor de amplificación en función de y para el esquema fuertemente implícito (en línea punteada se indica el factor de amplificación del esquema de Euler).

Los esquemas de este tipo para los cuales el factor de amplificación tiende a cero para valores crecientes de y, suelen denominarse esquemas super estables. Su utilización es muy difundida como veremos más adelante en el caso de sistemas de ecuaciones diferenciales rígidas (stiff).

Para ecuaciones de crecimiento (M < 0) el factor de amplificación se transforma en

$$g = \frac{1}{1 - k\,(M)_{n+1}} \qquad (2.66)$$

y el módulo del factor de amplificación es mayor que la unidad para cualquier tamaño de malla, por lo que el método fuertemente implícito produce errores que se amplifican en forma no acotada. Para la utilización de este método en ecuaciones de crecimiento valen las mismas consideraciones que para el método de Euler. En realidad conviene más, en estas circunstancias, emplear el método de Euler ya que en el mismo los errores crecen de una manera menos catastrófica. Esto se puede verificar comparando los factores de amplificación de ambos métodos. El siguiente ejemplo ilustra lo anterior. Supongamos a = 1 y u_0 = 1 en la ecuación de crecimiento du/dt - a u = 0. Aplicando a la misma el método de Euler y el método fuertemente implícito respectivamente, resulta

$$u^{n+1} = (1 + k)\,u^n = g_E\,u^n \quad \text{(Euler)} \qquad (2.67)$$

$$u^{n+1} = \frac{1}{1 - k}\,u^n = g_I\,u^n \quad \text{(Fuertemente implícito)} \qquad (2.68)$$

Definiendo una nueva variable y = ka los factores de amplificación de ambos métodos se comparan en la fig. 2.9. En las tablas 2.4 y 2.5 se presentan los resultados obtenidos por ambos métodos para k = 0.5 y 0.1, respectivamente.

Tabla 2.4 Soluciones "inestables" de la ecuación de crecimiento

n	t	k = 0.5		Solución exacta e^t
		$g_I = 2$	$g_E = 1.5$	
		U_I	U_E	
0	0	1	1	1
1	0.5	2	1.5	1.649
2	1	4	2.25	2.718
3	1.5	8	3.37	4.487
4	2	16	5	7.389
5	2.5	32	7.59	12.18

Tabla 2.5 Soluciones "estables" de la ecuación de crecimiento

n	t	k = 0.1		Solución exacta e^t
		$g_I = 1.111$	$g_E = 1.1$	
		U_I	U_E	
0	0	1	1	1
1	0.1	1.234	1.1	1.105
2	0.2	1.372	1.21	1.221
3	0.3	1.524	1.331	1.350
4	0.4	1.693	1.464	1.492
5	0.5	1.881	1.610	1.649

En la tabla 2.4 para k = 0.5 ambos métodos producen resultados inestables pero mientras el método de Euler diverge lentamente el método implícito lo hace en forma catastrófica. En la tabla 2.5 para k = 0.1 y para valores del factor de amplificación mucho más cercanos a la unidad, ambos métodos producen resultados razonables, siendo más precisos los correspondientes al método de Euler.

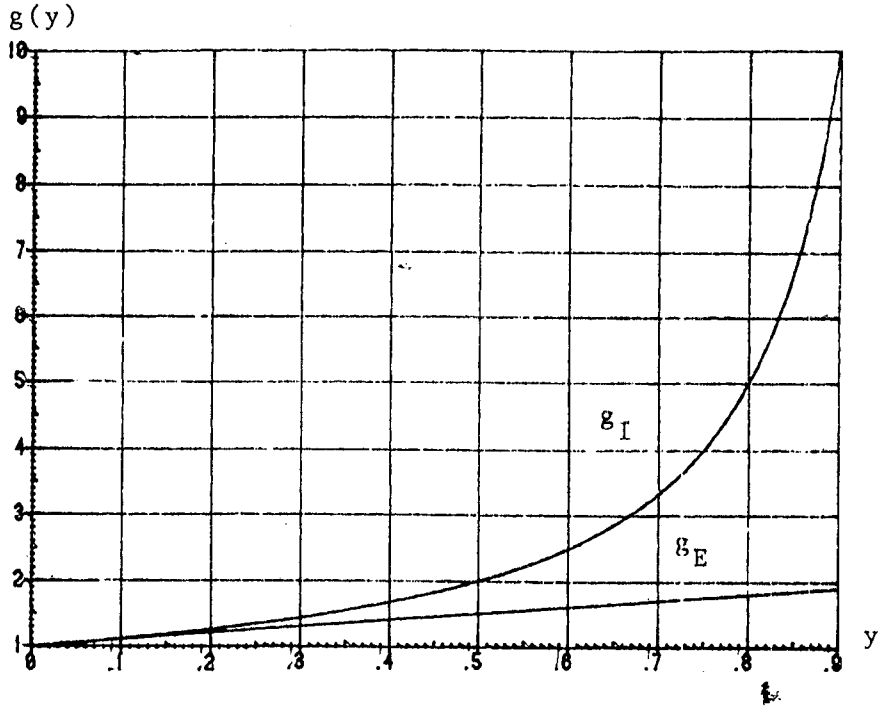

Fig. 2.9 Comparación de los factores de amplificación del método de Euler y fuertemente implícito para la ecuación de crecimiento

2.9 Método implícito ponderado

Para obtener una familia de métodos de paso entero implícito ponderado se aproxima la integral de la ecuación (2.19) suponiendo el valor de f constante para todo el intervalo $t^n \leq t \leq t^{n+1}$ e igual al valor promedio ponderado de f en los niveles temporales t^n y t^{n+1}, es decir

$$u^{n+1} = u^n - k \left[\beta f^{n+1} + (1 - \beta) f^n\right] \qquad (2.69)$$

donde $f^n = f(u^n, t^n)$, $f^{n+1} = f(u^{n+1}, t^{n+1})$ y β es una constante de ponderación con valores entre 0 y 1. Para $\beta = 0$ se tiene el método explícito de Euler, para $\beta = 0.5$ se obtiene el método implícito de segundo orden de precisión llamado de Crank-Nicolson y para $\beta = 1$ se tiene el método fuertemente implícito. La ventaja del método implícito ponderado es que permite una mayor precisión que los métodos de Euler o fuertemente implícito.

Un análisis lineal de la estabilidad muestra que la ecuación para la amplificación de un error pequeño se escribe (ver Marshall, 1984).

$$\varepsilon^{n+1} = \frac{1 - k(1 - \beta)\,(M)_n}{1 + k\,\beta\,(M)_{n+1}}\,\varepsilon^n \qquad (2.70)$$

Vamos a estudiar el factor de amplificación proveniente de la aplicación del esquema (2.69) a la ecuación de decaimiento $du/dt + au = 0$. En este caso la expresión (2.70) deviene

$$\varepsilon^{n+1} = \frac{1 - k\,(1 - \beta)\,a}{1 + k\,\beta\,a}\,\varepsilon^n \qquad (2.71)$$

Definiendo una nueva variable $y = ka$ el factor de amplificación se escribe

$$g(y) = \frac{1 - (1 - \beta)\,y}{1 + \beta\,y} \qquad (2.72)$$

En la figura 2.10 se presenta el gráfico de g(y) como función del argumento y para distintos valores del factor de ponderación β. El límite de g(y) para y tendiendo a infinito es $(\beta - 1)/\beta$. Si $\frac{1}{2} \leq \beta \leq 1$, el valor asintótico de g(y) no es menor que -1 y el método implícito ponderado es incondicionalmente estable. Pero si $0 \leq \beta < \frac{1}{2}$, el valor del argumento y debe acotarse por el valor al cual intersecta la línea $g(y) = -1$, para la estabilidad del método. En resumen, la condición de estabilidad del método implícito ponderado para la ecuación de decaimiento es

$$k < \frac{2}{a(1 - 2\beta)} \quad \text{si} \quad 0 < \beta < 1/2$$

$$\text{sin restricción} \quad \text{si} \quad 1/2 < \beta < 1 \qquad (2.73)$$

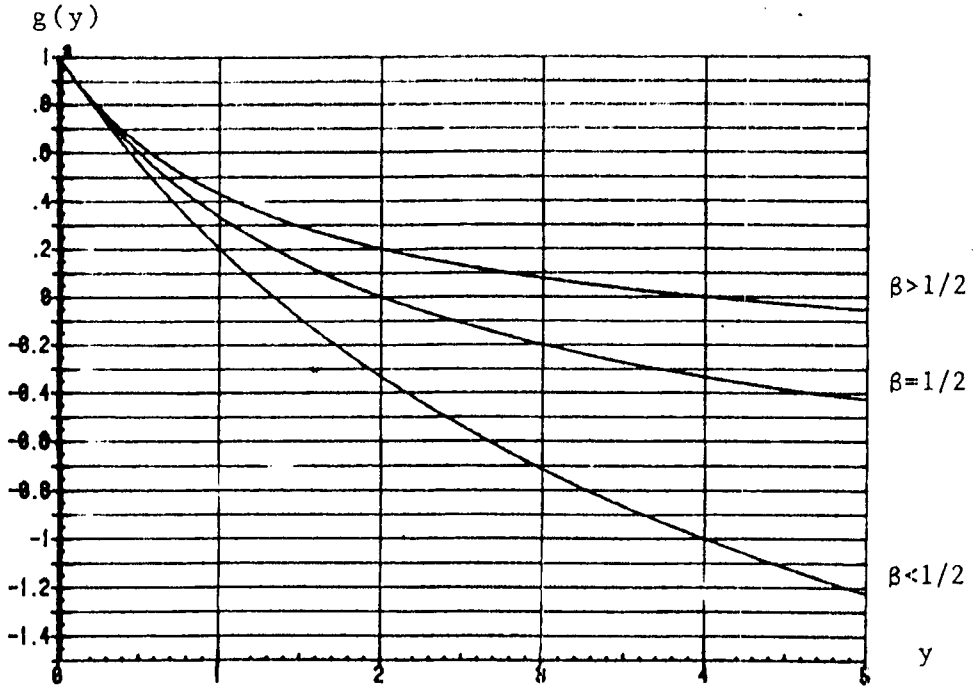

Fig. 2.10 Factor de amplificación para el método implícito ponderado

Para ecuaciones de crecimiento el módulo del factor de amplificación es mayor que la unidad para cualquier tamaño de malla por lo que el método implícito ponderado produce errores que se amplifican en forma no acotada. Al respecto, las consideraciones sobre el método de Euler y el método fuertemente implícito son válidas aquí.

2.10 El método de extrapolación de Richardson

El conocimiento de la velocidad de convergencia de un método numérico permite mejorar los resultados obtenidos con el mismo mediante una extrapolación a $k = 0$. Supongamos que el error de un determinado método de integración numérica es $O(k^q)$ y que k es pequeño, entonces podemos escribir

$$u - U = c\,k^q + \delta \qquad (2.74)$$

donde u es la solución aproximada y U es la solución exacta; el error dominante es $c\,k^q$ y δ es un error mucho más pequeño. Si ahora realizamos dos cálculos numéricos con el mismo método y problema utilizando dos mallas diferentes k_1 y k_2, tendremos

$$u_1 - U = c\,k_1^q + \delta_1 \qquad (2.75)$$

$$u_2 - U = c\,k_2^q + \delta_2 \qquad (2.76)$$

Eliminando la constante c resulta

$$U = \frac{1}{k_2^q - k_1^q}(u_1 k_2^q - u_2 k_1^q) + \frac{1}{k_2^q - k_1^q}(\delta_1 k_2^q - \delta_2 k_1^q) \qquad (2.77)$$

Si despreciamos los términos que contienen los errores δ_1 y δ_2 por ser pequeños, la ecuación (2.77) nos da una mejor aproximación de U.

La técnica de extrapolación descripta se debe a Richardson (ver Henrici, 1962), también se la suele denominar el método de extrapolación al límite o el método de las correcciones diferidas.

El siguiente ejemplo ilustra la aplicación del método de Richardson a los resultados numéricos obtenidos con el método de Euler. Hemos visto que con dicho método se obtiene un error de O(k), si quisiéramos reducir el error a la mitad tendríamos que reducir la malla a k/2, siempre que el término de orden superior sea despreciable. Consideremos dos integraciones de una ecuación diferencial utilizando en la primera una malla con paso k y en la segunda una malla con paso k/2. LLamando u_1 y u_2 a los respectivos resultados numéricos y U a la solución exacta, las ecuaciones (2.75) y (2.76) resultan en este caso (q = 1)

$$u_1 - U = c\,k + O(k^2)$$

$$u_2 - U = c\,\frac{k}{2} + O(k^2)$$

Eliminando la constante c obtenemos la expresión correspondiente a la ecuación (2.77), es decir

$$U = 2\,u_2 - u_1 + O(k^2) \qquad (2.78)$$

Entonces podemos utilizar la ecuación (2.78) para obtener una mejor aproximación numérica ya que la misma constituye un método de segundo orden de precisión debido a que el error es $O(k^2)$.

En la tabla 2.6 se muestra el resultado del método de Richardson aplicado a los resultados obtenidos utilizando el método de Euler en la ecuación du/dt = -u con condición inicial u(0) = 1 en el intervalo $0 \leq t \leq 1$.

Tabla 2.6 Resultados del método de extrapolación de Richardson

	k = 0.25	k = 0.125	k = 0.0625	k = 0.03125
$u_k(1)$ Euler	0.3164	0.3436	0.3561	0.3621
(2.78)	-	0.3708	0.3685	0.3680
Error A	-	0.0243	0.0118	0.0058
Error B	-	0.0029	0.0006	0.0001
Error/k^2	-	0.0469	0.0422	0.0401

La primera fila corresponde a los resultados del método de Euler para t = 1, utilizando diferentes mallas ($k = 0.25 \times 2^{-i}$, i = 0,1,2,3). La segunda fila corresponde a una nueva aproximación calculada con la extrapolación de Richardson representada por la ecuación (2.78). El error A es la diferencia entre la solución exacta (u = 0.3679) y la obtenida por el método de Euler, el error B es la diferencia entre la solución exacta y la aproximación de Richardson. Está claro que esta última es más exacta que el método de Euler al mismo tiempo que converge con mayor velocidad a medida que se afina la malla ($O(k^2)$ vs. $O(k)$).

2.11 Método predictor-corrector explícito

Una manera sencilla de mejorar la precisión del método de Euler es mediante su utilización dentro de un esquema de pasos fraccionarios. Este consiste en utilizar un procedimiento de paso doble: en el primer paso, denominado predictor, se predice una solución auxiliar con la fórmula de Euler; en el segundo paso, corrector, se calcula la solución final con un esquema centrado utilizando la solución auxiliar. El método predictor-corrector explícito suele denominarse también método de paso doble.

En el contexto de la ecuación (2.19) el predictor aproxima la integral del segundo miembro suponiendo un valor constante para el intervalo $t^n \leq t \leq t^{n+1/2}$ e igual al valor de f en el nivel temporal t^n (método de Euler), mientras que el corrector aproxima la integral suponiendo un valor constante para el intervalo $t^n \leq t \leq t^{n+1}$ e igual al valor de f en el nivel temporal $t^{n+1/2}$ (esquema centrado) según se ilustra en la figura 2.11, es decir

$$\text{predictor} \qquad u^{n+1/2} = u^n - \frac{k}{2} f^n \qquad (2.79)$$

$$\text{corrector} \qquad u^{n+1} = u^n - k\, f^{n+1/2} \qquad (2.80)$$

donde $f^{n+1/2} = f(u^{n+1/2}, t^{n+1/2})$. El método predictor-corrector explícito es consistente con el problema diferencial y posee segundo orden de precisión.

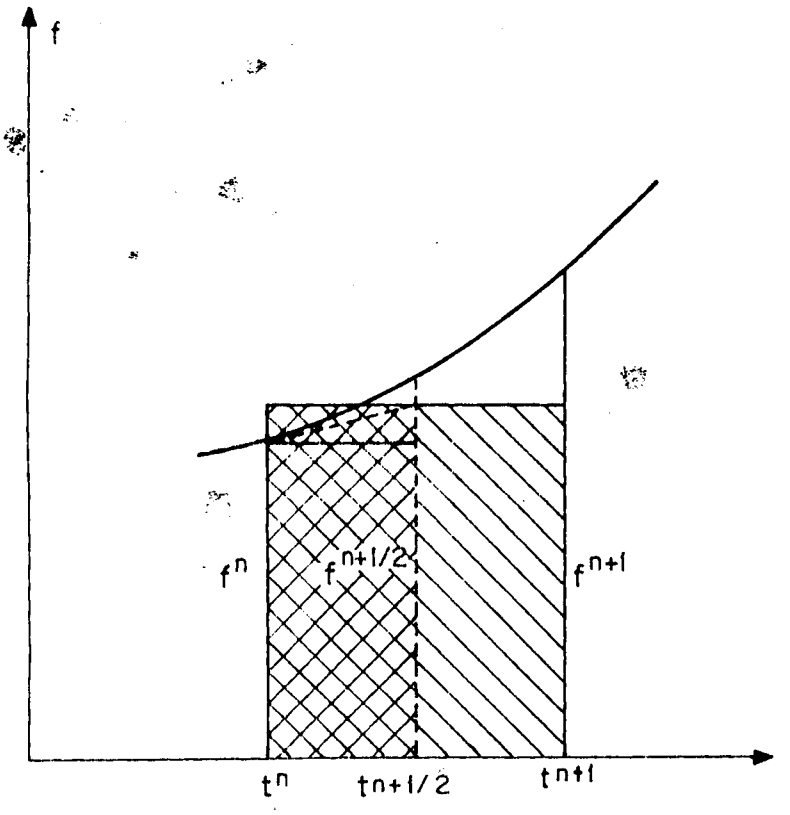

Fig. 2.11 Aproximación de la integral de la ecuación (2.19) en el método predictor-corrector explícito

Un análisis lineal de la estabilidad muestra que las ecuaciones para la amplificación del error están dadas por

$$\varepsilon^{n+1/2} = \varepsilon^n - \frac{k}{2} M \varepsilon^n \qquad (2.81)$$

$$\varepsilon^{n+1} = \varepsilon^n - k M \varepsilon^{n+1/2} \qquad (2.82)$$

donde M representa la derivada de f con respecto a u supuesta constante en el tiempo. Eliminando el error intermedio, el factor de amplificación del paso entero equivalente resulta

$$g = 1 - k M + \frac{1}{2} k^2 M^2 \qquad (2.83)$$

Para el caso de la ecuación de decaimiento du/dt + au = 0 la condición de estabilidad exige que

$$k < 2/a \qquad (2.84)$$

Definiendo una nueva variable y = ka la figura 2.12 muestra el gráfico del factor de amplificación en función de y para la ecuación de decaimiento

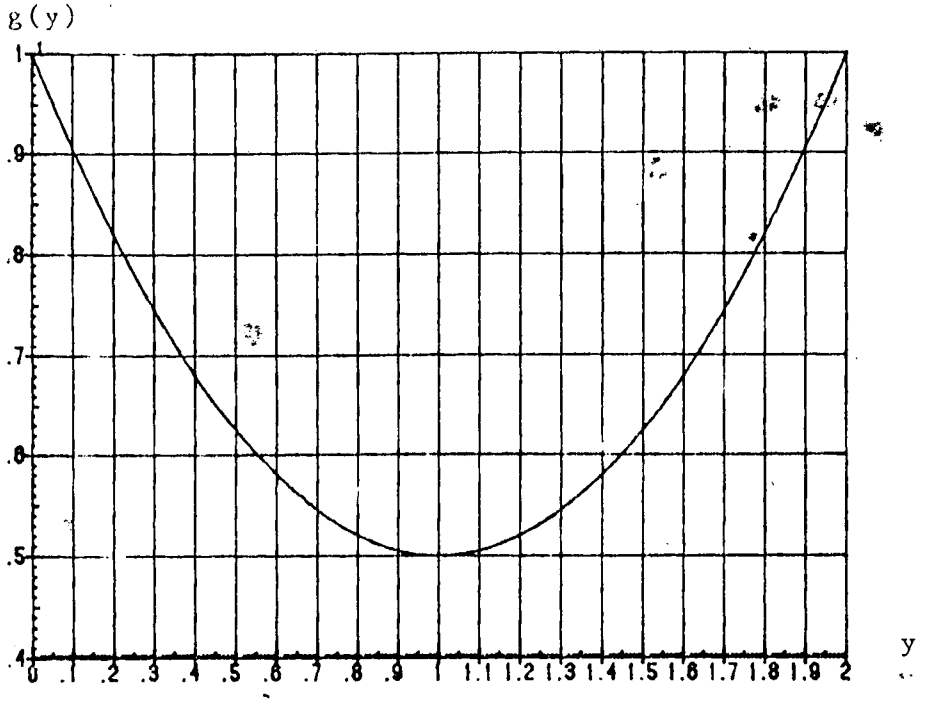

Fig. 2.12 Factor de amplificación del esquema predictor-corrector para la ecuación de decaimiento

Para ecuaciones de crecimiento, M < 0, el módulo del factor de amplificación es mayor que la unidad para cualquier tamaño de malla de donde el método predictor-corrector produce errores que se amplifican en forma no acotada; para su utilización en estos casos rigen las mismas consideraciones que para los esquemas ya vistos.

2.12 El método predictor-corrector implícito

Es interesante analizar la utilización de un esquema predictor-corrector implícito con el que cabría la posibilidad de mejorar la precisión y la estabilidad del cálculo. Refiriéndonos nuevamente a la ecuación (2.19) el predictor está constituído por el método de Euler, al igual que con la ecuación (2.79), mientras que el corrector está constituído por un esquema implícito ponderado, es decir

$$\text{predictor} \qquad u^{*n+1} = u^n - k\, f^n \qquad (2.85)$$

corrector $$u^{n+1} = u^n - k\left[\beta f^{*n+1} + (1 - \beta) f^n\right] \qquad (2.86)$$

donde $f^{*n+1} = f(u^{*n+1}, t^{n+1})$ y β es un coeficiente de ponderación $0 < \beta < 1$. Vamos a analizar el caso en que el predictor es utilizado una sola vez (para cada paso de tiempo), pero el corrector es utilizado en forma iterativa, es decir

predictor $$u^{n+1}_{m=0} = u^n - k\, f^n \qquad (2.87)$$

corrector $$u^{n+1}_{m+1} = u^n - k(1 - \beta)\, f^n - k\, \beta\, f^{n+1}_m \qquad (2.88)$$

donde el subíndice m indica el número de iteración. Vamos a analizar qué sucede con la precisión del método, cuando el corrector es utilizado m veces en forma iterativa, en el ejemplo de la ecuación de decaimiento $du/du + au = 0$. Además vamos a suponer que $\beta = 0.5$, es decir, que el corrector es un esquema de Crank-Nicolson. Entonces el esquema de cálculo resulta

$$u^{n+1}_0 = (1 - ak)\, u^n$$

$$u^{n+1}_1 = (1 - a\frac{k}{2})u^n - a\frac{k}{2}\, u^{n+1}_0$$

$$u^{n+1}_2 = (1 - a\frac{k}{2})\, u^n - a\frac{k}{2} u^{n+1}_1$$

$$\vdots$$

$$u^{n+1}_{m+1} = (1 - a\frac{k}{2})\, u^n - a\frac{k}{2}\, u^{n+1}_m \qquad (2.89)$$

Si expresamos u^{n+1}_{m+1} en forma recursiva en función de u^n se obtiene

$$u^{n+1}_{m+1} = (1 - ak + a^2\frac{k^2}{2} - a^3\frac{k^3}{4} + \ldots)\, u^n \qquad (2.90)$$

Comparando la expresión (2.90) con el siguiente desarrollo de Taylor

$$u^{n+1}_{m+1} = (1 - ak + a^2\frac{k^2}{2} - a^3\frac{k^3}{6} + \ldots)\, u^n\,, \qquad (2.91)$$

vemos que bajo el punto de vista de la precisión no tiene mayor sentido iterar más de dos veces.

A continuación analizamos la estabilidad del predictor-corrector implícito. Es natural suponer que si el proceso de iteración se efectúa hasta que $u_{m+1}^{n+1} = u_m^n$, el esquema predictor-corrector

resultará incondicionalmente estable tal como sucede con el corrector. Para obtener el factor de amplificación de la ecuación (2.89) la relacionaremos con el esquema de Crank-Nicolson sumando a ambos miembros la cantidad $ak/2\ u_{m+1}^{n+1}$ resultando

$$(1 + \frac{ak}{2})u_{m+1}^{n+1} = (1 - \frac{ak}{2})u^n - \frac{ak}{2}(u_m^{n+1} - u_{m+1}^{n+1}) \qquad (2.92)$$

Por otra parte la diferencia entre dos iteraciones sucesivas en función del valor inicial está dada por

$$u_m^{n+1} - u_{m+1}^{n+1} = -2\left[\frac{-ak}{2}\right]^{m+2} u^n \qquad (2.93)$$

Reemplazando en la expresión anterior y resolviendo para u_{m+1}^{n+1} obtenemos, finalmente, el factor de amplificación

$$g = \frac{1 - ak/2}{1 + ak/2} - \frac{2\,(-ak/2)^{m+2}}{1 + ak/2} \qquad (2.94)$$

El factor de amplificación (2.94) del predictor-corrector converge al del esquema de Crank-Nicolson sólo si se cumple que $k < 2/a$, que es precisamente la estabilidad del esquema de Euler. Introduciendo la nueva variable $y = ak$ en la Fig. 2.13 se ilustra la variación del factor de amplificación en función de y para diferentes valores del número de iteración m, en el esquema predictor-corrector implícito. El gráfico muestra, quizás con mayor claridad, la imposibilidad de extender la estabilidad del predictor-corrector en su modalidad iterativa, más allá de la estabilidad correspondiente al esquema de Euler.

Existe otra manera de interpretar el esquema predictor-corrector (2.87)-(2.88) y consiste en resolver el corrector (para $\beta = 1/2$)

$$(1 + \frac{1}{2}ak)u^{n+1} = (1 - \frac{1}{2}ak)u^n \qquad (2.95)$$

mediante el siguiente proceso iterativo

$$u_{m+1}^{n+1} = -\frac{1}{2}ak\,u_m^{n+1} + (1 - \frac{1}{2}ak)u^n \qquad (2.96)$$

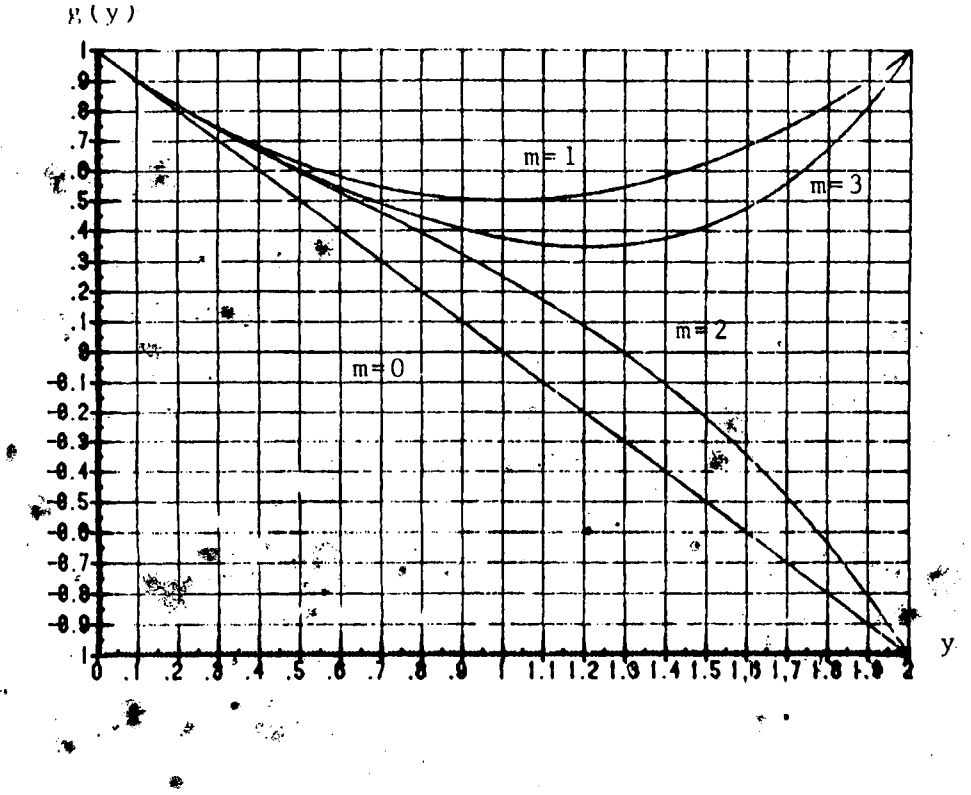

Fig. 2.13 Factor de amplificación para distintos valores del número de iteraciones m en el esquema predictor-corrector implícito. (Crank-Nicolson en línea de puntos).

En este caso el factor de amplificación del proceso iterativo está dado por $g = -ak/2$. Para que el proceso iterativo representado por la ecuación (2.96) sea convergente es necesario que se cumpla la condición de estabilidad $k < a/2$.

2.13 El problema de valores de contorno

Introducimos el tratamiento numérico de problemas de contorno con la siguiente ecuación diferencial de segundo orden

$$u'' - p(x)\,u' - q(x) = r(x) \qquad 0 < x < 1 \tag{2.97}$$

$$u(0) = \alpha \;,\; u(1) = \beta \tag{2.98}$$

donde $u = u(x)$ y $q(x) > 0$. De acuerdo a la convención adoptada, la variable independiente x representa una variable espacial. El problema de contorno (2.97) - (2.98) que, como veremos más adelante, también representa una ecuación lineal elíptica en una

dimensión, tiene una gran variedad de aplicaciones en Física e Ingeniería. La condición $q(x) > 0$ es necesaria para la existencia de soluciones del problema de contorno: supondremos entonces que la solución existe y es única. Con el objeto de aproximar el problema diferencial por un esquema en diferencias finitas subdividimos el intervalo espacial $0 \leq x \leq 1$ en N+1 segmentos iguales de longitud $h = 1/N+1$, donde h es el paso espacial, según se ilustra en la Fig. 2.14. La unión de cada uno de estos segmentos determina un nodo de la malla o red. Para discretizar la ecuación

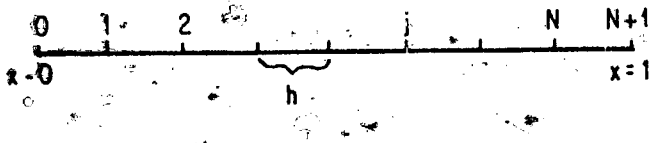

Fig. 2.14 Discretización del dominio

diferencial reemplazamos cada término de la misma por su aproximación en diferencias finitas en los nodos de la malla; en particular utilizaremos esquemas en diferencias centradas. Entonces una aproximación de la ecuación (2.97) se puede escribir (para un nodo genérico)

$$\frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} - p_j \frac{u_{j+1} - u_{j-1}}{2h} - q_j u_j = r_j \qquad (2.99)$$

$$j = 1,2,..,N$$

donde $p_j = p(x_j)$, $q_j = q(x_j)$ y $r_j = r(x_j)$. El esquema representado por la ecuación (2.99) es consistente con el problema

diferencial (2.97) y la consistencia es de $O(h^2)$. Si ahora escribimos para cada nodo interior la ecuación (2.99), resulta

$$j = 1 : h^{-2}(u_0 - 2u_1 + u_2) - p_1(2h)^{-1}(u_2 - u_0) - q_1 u_1 = r_1$$

$$j = 2 : h^{-2}(u_1 - 2u_2 + u_3) - p_2(2h)^{-1}(u_3 - u_1) - q_2 u_2 = r_2$$

$$j = 3 : h^{-2}(u_2 - 2u_3 + u_4) - p_3(2h)^{-1}(u_4 - u_2) - q_3 u_3 = r_3$$

$$\vdots$$

$$j = N : (u_{N-1} - 2u_N + u_{N+1}) - p_N (2h)^{-1} (u_{N+1} - u_{N-1}) -$$

$$q_N u_N = r_N \qquad (2.100)$$

El sistema (2.100) es un sistema de N ecuaciones algebraicas con N+2 incógnitas u_j. Las dos ecuaciones faltantes están dadas por las condiciones de borde: $u_0 = \alpha$ y $u_{N+1} = \beta$. Entonces si reemplazamos los valores de u_0 y u_{N+1} en el sistema (2.100) y reagrupamos los coeficientes de las incógnitas podemos escribir el sistema resultante en la forma vectorial siguiente

$$A\,U = R \qquad (2.101)$$

donde: U y R son vectores de dimensión N, y A es una matriz tridiagonal de orden N, definidas por

$$U = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_j \\ \vdots \\ u_N \end{bmatrix} \qquad R = -\frac{h^2}{2} \begin{bmatrix} r_1 + b_1\alpha \\ r_2 \\ r_3 \\ \vdots \\ r_j \\ \vdots \\ r_N + c_{N+1}\beta \end{bmatrix}$$

$$A = \begin{bmatrix} a_1 & -c_1 & & & & \\ -b_2 & a_2 & -c_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -b_j & a_j & -c_j & \\ & & & \ddots & \ddots & \\ & & & & -b_N & a_N \end{bmatrix}$$

donde

$$a_j = 1 + \frac{h^2}{2} q_j \; ; \; b_j = \frac{1}{2}\left(1 + \frac{h}{2} p_j\right) \; y \; c_j = \frac{1}{2}\left(1 - \frac{h}{2} p_j\right)$$

(2.102)

En síntesis, la solución numérica por el método de las diferencias finitas del problema diferencial (2.97) - (2.98) se reduce a la solución del sistema algebraico lineal representado por la ecuación (2.101), es decir

$$U = A^{-1} R$$

Suponiendo por el momento que la inversión numérica de A es posible, existe una gran variedad de programas de cálculo que permiten la resolución del sistema (2.101) por métodos directos o iterativos. A continuación presentaremos un método directo de factorización para la solución numérica de sistemas lineales con matrices de estructura similar a la matriz A y luego veremos bajo qué condiciones dicho método es aplicable al sistema (2.101).

2.14 Matriz tridiagonal o de Jacobi

En la solución del sistema algebraico A X = f proveniente de la aproximación de ecuaciones diferenciales ordinarias o en derivadas parciales es muy frecuente la aparición de matrices con forma tridiagonal también llamadas de Jacobi (ver Isaacson y Keller, 1966 o Jannenko, 1968) en los que $a_{i,j} = 0$ si $|i - j| > 1$, es decir

$$A = \begin{bmatrix} a_1 & c_1 & & & \\ b_2 & a_2 & c_2 & & \\ & \cdot & \cdot & \cdot & \\ & & b_{n-1} & a_{n-1} & c_{n-1} \\ & & & b_n & a_n \end{bmatrix} \qquad (2.103)$$

Supongamos que la matriz A puede ser factorizada en la forma

$$A = L\,U \qquad (2.104)$$

donde L y U son matrices de forma bidiagonal definidas por

$$L = \begin{bmatrix} \alpha_1 & & & \\ b_2 & \alpha_2 & & \\ & \cdot & \cdot & \\ & & b_{n-1} & \alpha_{n-1} \\ & & & b_n \quad \alpha_n \end{bmatrix} \qquad (2.105a)$$

$$U = \begin{bmatrix} 1 & \gamma_1 & & & \\ & 1 & \gamma_2 & & \\ & & \cdot & \cdot & \\ & & & 1 & \gamma_{n-1} \\ & & & & 1 \end{bmatrix} \qquad (2.105b)$$

$$LU = \begin{bmatrix} \alpha_1 & \alpha_1\gamma_1 & & & \\ b_2 & b_2\gamma_1+\alpha_2 & c_2\gamma_2 & & \\ & \cdot & \cdot & \cdot & \\ & & b_{n-1} & b_{n-1}\gamma_{n-2}+\alpha_{n-1} & \\ & & & b_n & b_n\gamma_{n-1}+\alpha_n \end{bmatrix} \qquad (2.106)$$

Teniendo en cuenta la ecuación (2.104) y la (2.106) obtenemos

$$\begin{aligned} \alpha_1 &= a_1 & \gamma_1 &= c_1/\alpha_1 \\ \alpha_i &= a_i - b_i \gamma_{i-1} & i &= 2,3,..,n \\ \gamma_i &= c_i/\alpha_i & i &= 2,3,..n-1 \end{aligned} \qquad (2.107)$$

Entonces, si se cumple que α_i es distinto de cero, las fórmulas de recurrencia representadas en (2.107) permiten obtener la factorización (2.104). Una vez obtenida dicha factorización aún nos resta resolver el sistema algebraico primitivo A X = f a través del sistema equivalente L U X = f. Para ello primero se resuelve el sistema auxiliar o intermedio Lg = f y luego se obtiene la solución final resolviendo el sistema U X = g. La solución intermedia se obtiene por

$$\begin{aligned} g_1 &= f_1/\alpha_1 \\ g_i &= (f_i - b_i \gamma_{i-1})/\alpha_i & i &= 2,\ldots,n \end{aligned} \qquad (2.108)$$

La solución final está dada por

$$\begin{aligned} x_n &= g_n \\ x_j &= g_j - \gamma_j x_{j+1}, & j &= n-1,n-2,\ldots,1. \end{aligned} \qquad (2.109)$$

Se puede demostrar que si se cumplen las siguientes condiciones en la matriz A

$$|a_1| > |c_1| > 0$$

$$|a_i| > |b_i| + |c_i| \quad , \; b_i\, c_i \neq 0 \; , \; i = 2,3,\ldots,n-1; \qquad (2.110)$$

$$|a_n| > |b_n| > 0$$

A es una matriz no singular y por lo tanto el procedimiento representado en la expresión (2.107) es válido (el lector interesado en la demostración de lo anterior puede consultar Isaacson-Keller, 1966).

2.15 Solución directa del problema de contorno

Habíamos visto que la solución numérica del problema de contorno (2.97) - (2.98) se reducía a la solución del sistema algebraico lineal de orden N (2.101) cuya matriz de coeficientes es tridiagonal. Veamos ahora las condiciones que se deben cumplir para que el método de factorización sea aplicable a dicho sistema. A tal efecto vamos a requerir que la malla sea lo sufientemente pequeña para que se cumpla

$$\frac{h}{2}|p_j| < 1 \qquad j = 1,2,\ldots,N \qquad (2.111)$$

De la ecuación (2.102) se deduce entonces

$$|b_j| + |c_j| = b_j + c_j = 1 \qquad (2.212)$$

Recordando que

$$q(x) > 0 \quad , \quad 0 \leq x \leq 1 \quad , \qquad (2.113)$$

resulta $a_j > 1$. De lo anterior se deduce que

$$|a_1| > |c_1|$$

$$|a_j| > |b_j| + |c_j| \quad , \quad 2 \leq j \leq N-1 \qquad (2.114)$$

$$|a_N| > |b_N|$$

Vemos entonces que se cumplen las condiciones para que la factorización descripta en la sección anterior sea válida. La solu-

ción de la ecuación (2.101) de la sección correspondiente al problema de contorno puede entonces ser obtenida por el proceso de recurrencia anteriormente descripto.

2.16 El método de Runge-Kutta

Habíamos visto que para mejorar la precisión del método de Euler podíamos utilizar un método de paso fraccionario explícito que denominamos predictor-corrector explícito. El método de Runge-Kutta utiliza la misma idea de varios pasos fraccionarios para obtener una mayor precisión. Utilizando la terminología de los textos sobre ecuaciones diferenciales ordinarias (ver Henrici, 1962, o Dahlquist y Bjorck, 1974) el método de Runge-Kutta de segundo orden para la ecuación $du/dt = f(u,t)$ se escribe

$$\begin{aligned} q_1 &= k\, f(u^n, t^n) \\ q_2 &= k\, f(u^n + q_1, t^{n+1}) \\ u^{n+1} &= u^n + \frac{1}{2}(q_1 + q_2) \end{aligned} \qquad (2.115)$$

donde q_1 y q_2 son valores aproximados de la derivada de u con respecto a t calculados en el intervalo $t^n \leq t \leq t^{n+1}$ (multiplicada por k). Este esquema suele denominarse también método de Heun. Pero el más conocido de los métodos de Runge-Kutta es el de cuarto orden que se escribe

$$\begin{aligned} q_1 &= k\, f(u^n, t^n) \\ q_2 &= k\, f(u^n + \frac{1}{2} q_1, t^{n+1/2}) \\ q_3 &= k\, f(u^n + \frac{1}{2} q_2, t^{n+1/2}) \\ q_4 &= k\, f(u^n + q_3, t^{n+1}) \\ u^{n+1} &= u^n + \frac{1}{6}(q_1 + 2\, q_2 + 2\, q_3 + q_4) \end{aligned} \qquad (2.116)$$

El lector interesado en la fundamentación del método de Runge-Kutta puede consultar Henrici, 1962, Isaacson y Keller, 1966 ó Gear, 1971. Aquí nos contentamos presentando un análisis heurístico del método de Runge-Kutta que sirve para su mejor comprensión. Si suponemos que la función f es independiente de la incógnita u el método de Runge-Kutta de segundo orden resulta

$$u^{n+1} = u^n + \frac{k}{2}(f^n + f^{n+1}) \qquad (2.117)$$

donde $f^n = f(t^n)$ y $f^{n+1} = f(t^{n+1})$. Entonces, para el caso particular analizado, el método de Runge-Kutta de segundo orden es equivalente a la aproximación de la integral de la ecuación (2.19) con el método implícito de Crank-Nicolson. En la nomenclatura clásica de ecuaciones diferenciales se lo llama método de los trapecios. Nuevamente hacemos notar que las fórmulas de Runge-Kutta presentadas corresponden a la aproximación de la ecuación diferencial du/dt = f(u,t) de ahí el signo positivo en el segundo miembro (2.117).

Con igual argumento el método de Runge-Kutta de cuarto orden resulta

$$u^{n+1} = u^n + \frac{1}{6} k \, (f^n + 4 f^{n+1/2} + f^{n+1}) \qquad (2.118)$$

donde $f^n = f(t^n)$, $f^{n+1/2} = f(t^{n+1/2})$ y $f^{n+1} = f(t^{n+1})$

En este caso el método de Runge-Kutta de cuarto orden resulta equivalente a la aproximación de la integral de la ecuación (2.19) con la fórmula de cuadratura de Simpson.

2.17 Métodos de paso múltiple

Habíamos visto que en los métodos de un sólo paso como por ejemplo en el método de Euler la solución global del problema de valores iniciales para la ecuación de primer orden du/dt = f(u,t) se obtiene por medio de una integración sucesiva sobre una serie de intervalos pequeños de tiempo, es decir

$$\begin{aligned} u^1 &= u^0 + \int_{t^0}^{t^1} f^0 \, dt \\ u^2 &= u^1 + \int_{t^1}^{t^2} f^1 \, dt \\ &\vdots \\ u^{n+1} &= u^n + \int_{t^n}^{t^{n+1}} f^n \, dt \end{aligned} \qquad (2.119)$$

Para una integración en que estén involucrados p+1 intervalos de tiempo finalizando en t^{n+1}, este proceso puede escribirse como

$$u^{n+1} = u^{n-p} + \int_{t^{n-p}}^{t^{n+1}} \phi^n(t) \, dt \qquad (2.120)$$

donde $\phi^n(t)$ es una función constante a trozos con valores $\phi^i = f(u^i,t^i)$ en los intervalos semiacotados $[t^i,t^{i+1}]$, $i = n-p,n-p+1,..,n-1,n$. La integral en la ecuación (2.120) puede ser interpretada como el área entre los límites t^{n-p} y t^{n+1} como se muestra en la Fig. 2.15.

Está claro que un reemplazo de la función constante a trozos $\phi^n(t)$ en la ecuación (2.120), por una función de interpolación polinomial p(t) que pase por los puntos (u^i,t^i), para valores de i cercanos a n, puede dar lugar a un método de integración más preciso.

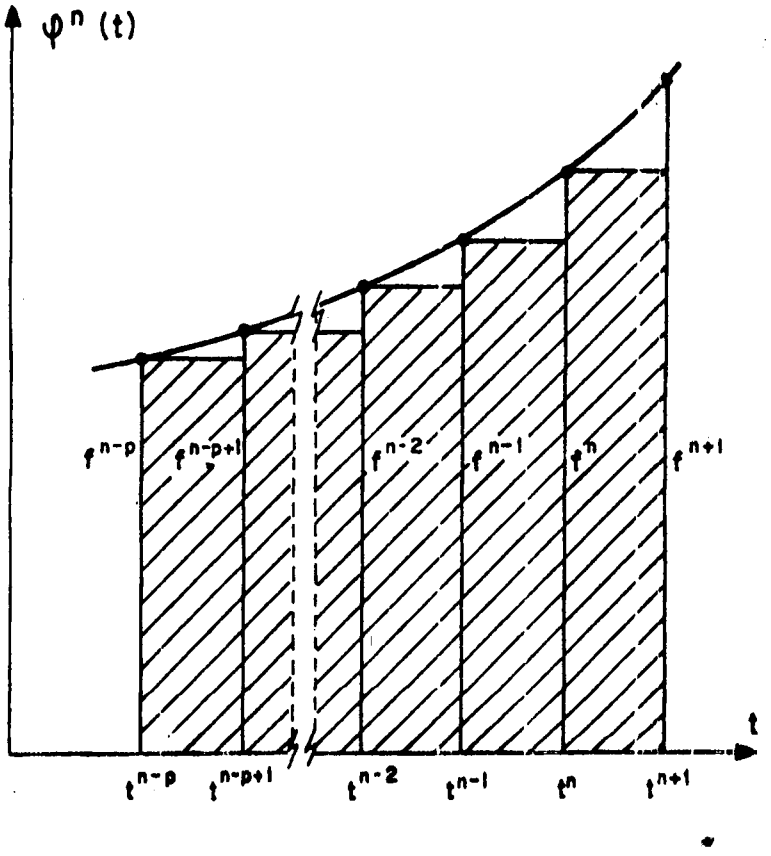

Fig. 2.15 Aproximación de la integral en la ecuación (2.120) por un esquema de paso múltiple

De la enorme variedad de métodos de paso múltiple originados por las distintas variantes de la aproximación polinómica mencionada sólo presentaremos aquí una clase especial de métodos de paso múltiple que constituyen la base de muchas bibliotecas de rutinas para la integración de ecuaciones diferenciales ordinarias.

Los métodos de Adams están basados en la fórmula

$$u^{n+1} = u^n + \int_{t^n}^{t^{n+1}} p(t)\, dt \qquad (2.121)$$

donde p(t) interpola f(u,t) en varios puntos. Como es usual la fórmula final obtenida no contiene el polinomio p(x) en forma explícita sino que expresa u^{n+1} como una combinación lineal de u^n y de $f^n, f^{n-1}, \ldots, f^{n-p}$. Las fórmulas particulares obtenidas

dependerán de los puntos de interpolación elegidos. Existen dos tipos de elección que son los más comunes: el método de Adams-Bashforth de orden p+1 utiliza la interpolación de f(u,t) en $t^n, t^{n-1}, \ldots, t^{n-p}$, y el método de Adams-Moulton de orden p+1 utiliza la interpolación de f(u,t) en $t^{n+1}, t^n, t^{n-1}, \ldots, t^{n-p+1}$. Naturalmente el primero da lugar a métodos explícitos mientras que del segundo resultan métodos implícitos. Las tablas 2.7 y 2.8 presentan las fórmulas más comunes para ambos métodos junto con los errores locales correspondientes (las derivadas sucesivas de u(t) se evalúan en un punto ξ intermedio desconocido entre t^n y t^{n-p}). En general el orden del método de Adams está determinado por el grado del polinomio utilizado. El error local de una fórmula de orden p es de orden $O(k^{p+1})$ dado que t^{p+1} es el orden del polinomio próximo no utilizado. En los métodos de Adams-Bashforth y Adams-Moulton el orden está también determinado por el número de términos de f(u,t) utilizados en la obtención de u^{n+1}.

Tabla 2.7 Fórmulas de Adams-Bashforth (Explícitas)

ORDEN	FORMULA	ERROR LOCAL
1	$u^{n+1} = u^n + kf^n$	$\frac{1}{2} k\, u'(\xi)$
2	$u^{n+1} = u^n + k/2(3f^n - f^{n-1})$	$\frac{5}{12} k^3\, u''(\xi)$
3	$u^{n+1} = u^n + k/12(23f^n - 16f^{n-1} + 5f^{n-2})$	$\frac{3}{8} k^4\, u^{IV}(\xi)$
4	$u^{n+1} = u^n + k/24(55f^n - 59f^{n-1} + 37f^{n-2} - 9f^{n-3})$	$\frac{251}{720} k^5\, u^{V}(\xi)$

En la tabla 2.7 se observa que el método de Adams-Bashforth de orden 1 no es mas que el esquema de Euler, mientras que en la tabla 2.8 los métodos de Adams-Moulton de orden 1 y 2 coinciden con el método fuertemente implícito y el método de Crank-Nicolson, respectivamente. El lector interesado en la fundamentación de estos métodos puede consultar Henrici, 1962 e Isaacson y Keller, 1966, entre otros.

La principal ventaja de los métodos de paso múltiple frente a los métodos de un solo paso es que, en general, a igual precisión las fórmulas son más sencillas y por ende el esfuerzo computacional por paso de tiempo es menor. La principal desventaja radica en que no se puede iniciar el cálculo con dichos

métodos ya que se requieren n-p valores previos. La solución consiste en iniciar el cálculo con un método de un solo paso o con un método de paso fraccionario. Es importante tener en cuenta que para iniciar el cálculo se debe utilizar un método del mismo orden o mayor que el método de paso múltiple que se utilizará luego. Por ejemplo para utilizar el método explícito de Adams-Bashforth de orden 4 bastaría iniciar el cálculo con un método de Runge-Kutta de orden 4. También se puede iniciar el cálculo con un método de menor orden pero con una malla más fina de modo que el error sea del mismo orden.

Tabla 2.8 Fórmulas de Adams-Moulton (Implícitas)

ORDEN	FORMULA	ERROR LOCAL
1	$u^{n+1} = u^n + k\ f^{n+1}$	$-\frac{1}{2} k^2\ u''(\xi)$
2	$u^{n+1} = u^n + k/2(f^{n+1} + f^n)$	$-\frac{1}{12} k^3\ u''(\xi)$
3	$u^{n+1} = u^n + k/12(5f^{n+1} + 8f^n - f^{n-1})$	$-\frac{1}{24} k^4\ u^{IV}(\xi)$
4	$u^{n+1} = u^n + k/24(9f^{n+1} + 19f^n - 5f^{n-1} + f^{n-2})$	$-\frac{19}{720} k^5\ u^{V}(\xi)$

En general se prefiere utilizar los métodos de paso múltiple implícitos para el cálculo de la solución numérica ya que para el mismo orden, estos tienen un error local mucho menor que los métodos de paso múltiple explícitos. Esto se puede apreciar comparando en las tablas 2.7 y 2.8 los métodos de orden 4 de Adams-Bashforth y Adams-Moulton. Suponiendo que las derivadas de orden cinco obtenidas mediante el cálculo y las derivadas exactas son aproximadamente iguales (para k suficientemente pequeño) el método implícito es considerablemente más preciso que su homólogo explícito.

En los métodos de paso múltiple implícitos, al igual que en los explícitos, es necesario utilizar un algoritmo de un solo paso entero o fraccionario para iniciar los cálculos. A esto se agrega la necesidad de tener que iterar por la característica intrínseca del método implícito. Por ejemplo tratándose de la fórmula de Adams-Moulton de orden 4 un posible procedimiento iterativo sería

$$u^{n+1}_{m+1} = u^n + k/24(9\ f^{n+1}_m + 19\ f^n - 5\ f^{n-1} + f^{n-2}) \qquad (2.122)$$

donde el subíndice m indica el número de la iteración. Este esquema iterativo es parecido al estudiado en el método predictor-corrector implícito. En la práctica, los métodos de paso múltiple implícito sólo pueden competir con el método de Runge-Kutta, por ejemplo, si la convergencia es muy rápida es decir si no se necesitan más de dos iteraciones. Precisamente para lograr una convergencia rápida es necesario comenzar el proceso iterativo con una buena estimación inicial y esto se logra utilizando un método de paso múltiple explícito. Entonces el procedimiento usual consiste en utilizar, en cada paso de tiempo, en la modalidad predictor-corrector, un método de paso múltiple explícito como predictor y un método de paso múltiple implícito como corrector, ambos con error local de truncamiento comparables. Por ejemplo un esquema de cálculo podría ser el siguiente: se inicializa el cálculo con un método de Runge-Kutta de orden 4, luego para cada paso de tiempo se utiliza como predictor, el método de Adams-Bashforth de orden 4 y como corrector, en modo iterativo, el método de Adams-Moulton de orden 4.

Existen muchos otros métodos de paso múltiple en la modalidad predictor-corrector y el lector interesado puede consultar Henrici, 1962, Hamming, 1962, Carnahan, Luther y Wilkes, 1969, e Isaacson y Keller, 1966, entre otros.

Otro método comúnmente utilizado es el esquema predictor-corrector de orden 4 de Milne que se escribe

predictor

$$u^{n+1} = u^{n-3} + \frac{4}{3} k(2 f^n - f^{n-1} + 2 f^{n-2}) \quad , \quad r = - \frac{14}{45} k^5 y^V(\xi_1) \tag{2.123}$$

corrector

$$u^{n+1} = u^{n-1} + \frac{k}{3} (f^{n+1} + 4 f^n + f^{n-1}) \quad , \quad r = \frac{1}{90} k^5 y^V(\xi_2) \tag{2.124}$$

donde ξ_1 es un punto intermedio entre t^{n-3} y t^{n+1} mientras ξ_2 está comprendido entre t^{n-1} y t^{n+1}.

Una de las ventajas adicionales del método predictor-corrector es la estimación del error local de truncamiento y su posterior corrección. Ilustramos el método con un ejemplo. Supongamos el problema de valores iniciales $du/dt = f(u,t)$ resuelto con el método de Milne de orden 4. Conocidos los valores iniciales u^0, u^1, u^2 u^3 y los correspondientes f^n (calculados, por ejemplo, con el método de Runge-Kutta de cuarto orden) y fijada la malla en un valor k, el cálculo prosigue de la siguiente manera:

a) Se calcula $u^{n+1}_{m=0}$ utilizando el predictor

$$u_{m=0}^{n+1} = u^{n-3} + \frac{4}{3} k(2 f^n - f^{n-1} + 2 f^{n-2}) \qquad (2.125)$$

b) Utilizando el valor del paso previo se calcula el corrector en forma iterativa

$$u_{m+1}^{n+1} = u^{n-1} + \frac{k}{3} (f_m^{n+1} + 4 f^n + f^{n-1}) \qquad (2.126)$$

$$m = 0, 1, 2, \ldots$$

La iteración se continúa hasta que el criterio de convergencia elegido se cumple, por ejemplo

$$\left| \frac{u_{m+1}^{n+1} - u_m^{n+1}}{u_{m+1}^{n+1}} \right| < \varepsilon \qquad (2.127)$$

donde ε es un número positivo pequeño. En el caso de que la convergencia falle es necesario detener el proceso y utilizar una malla más fina, volveremos sobre esto más adelante.

c) Para estimar el error local de truncamiento en el uso del método de Milne, en la integración sobre el intervalo t^n, t^{n+1}, se supone que los errores locales de truncamiento dados en las expresiones (2.123) y (2.124) son válidos. Los valores de las derivadas de orden 5 son desconocidos pero es razonable suponer que son iguales (siempre que k sea pequeño); denominaremos M a su valor. Luego la diferencia entre el valor obtenido por el predictor $u_{m=0}^{n+1}$ y el valor del corrector u_{mf}^{n+1} (donde mf es el número de iteraciones necesarios para satisfacer el criterio de convergencia), resulta

$$\left| u_{m=0}^{n+1} - u_{mf}^{n+1} \right| = \frac{1}{90} k^5 M + \frac{14}{45} k^5 M \qquad (2.128)$$

Esto implica que

$$k^5 M = \frac{90}{29} \left| u_{m=0}^{n+1} - u_{mf}^{n+1} \right| \qquad (2.129)$$

y el error local en u^{n+1} es estimado como

$$\frac{1}{90} k^5 M = \frac{\left| u_{m=0}^{n+1} - u_{mf}^{n} \right|}{29} \qquad (2.130)$$

Esta estimación debe ser utilizada con mucha cautela para que las suposiciones efectuadas acerca de las derivadas de orden 5 no causen problemas. Una posición más conservadora sería adoptar la siguiente expresión como estimación del error

$$1/3 \left| u_{m=0}^{n+1} - u_{m+1}^{n+1} \right| \tag{2.131}$$

añadiendo un factor 10 para compensar la incertidumbre

d) Vamos a modificar el método de Milne de cuarto orden teniendo en cuenta la estimación del error obtenida en la sección c). Para ello escribimos nuevamente las expresiones para el error de truncamiento

$$U^{n+1} - u_{m=0}^{n+1} = \frac{14}{45} k^5 M \tag{2.132}$$

$$U^{n+1} - u_{mf}^{n+1} = -\frac{1}{90} k^5 M \tag{2.133}$$

donde U^{n+1} es la solución exacta en los nodos de la malla y M es la derivada quinta en un punto intermedio supuesta de igual valor tanto para el predictor como para el corrector. Dividiendo la ecuación (2.132) por la (2.133) resulta

$$U^{n+1} = \frac{1}{29} (28\, u_{mf}^{n+1} + u_{m=0}^{n+1}) \tag{2.134}$$

Entonces la ecuación (2.134) es utilizada para modificar el cálculo del corrector de Milne de la siguiente manera:

$$u^{n+1} = \frac{1}{29} (28\, u_{mf}^{n+1} + u_{m=0}^{n+1}) \tag{2.135}$$

La ecuación (2.135) puede escribirse, sumando y restando $u_{m=0}^{n+1}$,

$$u^{n+1} = u_{m=0}^{n+1} + \frac{28}{29} (u_{mf}^{n+1} - u_{m=0}^{n+1}) \tag{2.136}$$

Llamando r_t a la estimación del error local de truncamiento dado por la ecuación (2.130) la ecuación (2.136) resulta

$$u^{n+1} = u_{m=0}^{n+1} + 28\, r_t \tag{2.137}$$

También podemos modificar el cálculo del predictor de modo de encontrar un valor mejorado del mismo. Si suponemos que el error local de truncamiento varía lentamente de modo que r_t es aproxi-

madamente igual en los intervalos $[t^n, t^{n+1}]$ y $[t^{n-1}, t^n]$, llamando $u^{*n+1}_{m=0}$ al valor mejorado de $u^{n+1}_{m=0}$ y teniendo en cuenta que r_t en el intervalo $[t^{n-1}, t^n]$ es, de acuerdo a la expresión (2.130), aproximadamente $(u^n_{mf} - u^n_{m=0})/29$, podemos estimar $u^{*n+1}_{m=0}$ de la ecuación (2.136) como

$$u^{*n+1}_{m=0} = u^{n+1}_{m=0} + \frac{28}{29}(u^n_{mf} - u^n_{m=0}) \qquad (2.138)$$

Entonces utilizando la expresión (2.138) en vez de $u^{n+1}_{m=0}$ como valor inicial para comenzar la iteración con el corrector se puede acelerar la velocidad de convergencia de la solución iterativa (2.126) ya que en la mayoría de los casos la ecuación (2.138) da un valor mejorado de $u^{n+1}_{m=0}$. El método presentado aquí puede ser extendido a otros algoritmos de tipo predictor-corrector.

El control del tamaño de malla o paso temporal es un problema de distinta naturaleza. El paso de tiempo debe ser suficientemente pequeño para satisfacer restricciones debidas a criterios de convergencia y o a la magnitud del error local de truncamiento. Por otro lado el paso de tiempo debe ser, en lo posible, suficientemente grande para minimizar el error de redondeo y el número de evaluaciones de la derivada. En general las expresiones (2.127) y (2.130) nos permiten determinar cuando hay necesidad de actuar sobre la malla haciéndola más fina o más gruesa. Desafortunadamente el mecanismo para dichos cambios no es tan inmediato como se desea. La dificultad radica en que cuando se cambia el paso de malla, en general, no se poseen los valores iniciales necesarios para la utilización del predictor y el corrector. Ilustraremos lo anterior con un ejemplo. En la Fig. 2.16 se muestran 3 mallas diferentes durante la solución numérica de una ecuación diferencial ordinaria. Supongamos que se usa un método de paso entero múltiple que requiere tres valores previos para obtener el nuevo valor y que el paso k en t^n debe ser cambiado. Si k es cambiado en forma arbitraria la situación es idéntica a la del punto inicial. Si el valor de k es duplicado entonces toda la información está disponible (siempre que se la haya guardado) para ser utilizada en el nuevo paso de longitud 2 k. Si el valor de k es reducido a la mitad entonces algunos de los valores necesarios están disponibles. Los valores intermedios faltantes pueden ser obtenidos por interpolación polinomial, el grado del polinomio debe ser del mismo orden que el del método de integración a los efectos de mantener la precisión. En la Fig. 2.16 se interpola en t^n, t^n-k, t^n-2k y t^n-3k para obtener un polinomio cúbico para poder evaluar en t^n-k/2. Luego los valores en t^n, t^n-k/2 y t^n-k pueden ser utilizados para continuar con el cálculo.

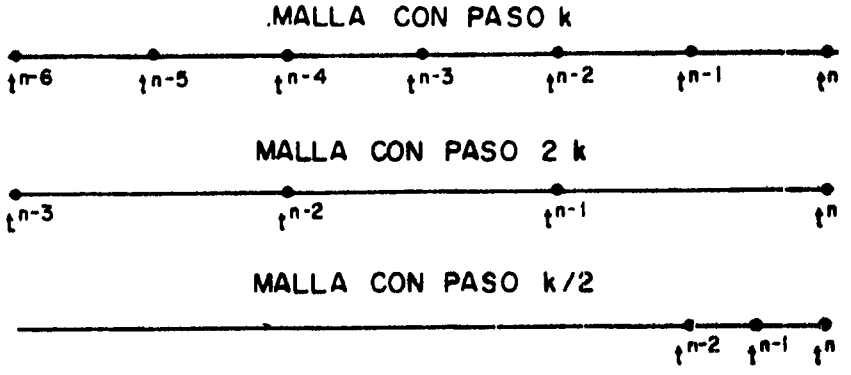

Fig. 2.16 Cambio de paso en los métodos de paso entero múltiple

En la mayoría de los programas se utilizan reglas adicionales que evitan el cambio de paso en demasía o muy seguido. Algunas reglas típicas son: a) no incrementar k en más de un factor 2; b) no incrementar k más de una vez cada cinco pasos.

A continuación presentamos un análisis simplificado de la estabilidad numérica de los métodos de paso múltiple utilizando como modelo el problema de valores iniciales $du/dt = f(u,t)$ con $f(u,t) = au$, $u(0) = U_0$ y solución exacta $U(t) = e^{at}U_0$, y el predictor-corrector de cuarto orden de Milne. Al respecto efectuaremos el análisis del corrector solamente suponiendo que este ha sido iterado hasta obtener la convergencia. Si el corrector es utilizado una sola vez se debe entonces hacer el análisis del predictor y del corrector en conjunto. El corrector de Milne es, entonces,

$$u^{n+1} = u^{n-1} + \frac{k}{3}(f^{n+1} + 4\,f^{n} + f^{n-1}) \qquad (2.139)$$

y teniendo en cuenta que $f(u,t) = au$ resulta

$$\left(1 - \frac{ak}{3}\right) u^{n+1} - 4 \left(\frac{ak}{3}\right) u^{n} - \left(1 + \frac{ak}{3}\right) u^{n-1} = 0 \qquad (2.140)$$

que constituye una ecuación lineal y homogénea en diferencias de orden 2. Para k fijo los coeficientes de (2.140) son constantes y se puede demostrar que el criterio de convergencia exige que (ver Dahlquist y Bjorck, 1974)

$$\frac{h}{3} \left| a \right| < 1 \qquad (2.141)$$

La teoría de la solución de las ecuaciones lineales en diferencias está estrechamente ligada a la teoría de la solución de ecuaciones diferenciales ordinarias lineales. Daremos una breve síntesis de esta última (el lector interesado puede consultar las obras de Collatz, 1966 y Dahlquist y Bjorck, 1974 entre otros). Si suponemos que u^n es una función discreta definida para el conjunto de enteros n, una ecuación lineal en diferencias de orden q (el orden de una ecuación en diferencias es igual a la diferencia entre el mayor y menor de los supraíndices) es una combinación lineal involucrando $u^n, u^{n+1}, \ldots, u^{n+q}$ de la forma

$$a_0 u^n + a_1 u^{n+1} + a_2 u^{n+2} + \ldots + a_q u^{n+q} = b \qquad (2.142)$$

donde los coeficientes $a_0, \ldots, a_q$ y b pueden ser funciones de n pero no de u. Si b es nulo se trata de una ecuación homogénea. La teoría de la solución de ecuaciones lineales homogéneas en diferencias con coeficientes constantes es similar a la teoría de la solución de ecuaciones lineales homogéneas diferenciales con coeficientes constantes. En ese caso se buscan soluciones de la forma $u^n = \gamma^n$. Substituyendo dicha solución en la ecuación (2.142) suponiendo b = 0 y coeficientes constantes; resulta

$$a_0 \gamma^n + a_1 \gamma^{n+1} + \ldots + a_q \gamma^{n+q} = 0 \qquad (2.143)$$

dividiendo por γ^n resulta la ecuación característica

$$a_0 + a_1 \gamma + \ldots + a_q \gamma^q = 0 \qquad (2.144)$$

que es un polinomio de grado q en la variable γ. Si las q raíces $\gamma_1, \gamma_2, \ldots, \gamma_q$ son distintas, entonces $\gamma_1^n, \gamma_2^n, \ldots, \gamma_q^n$ son todas soluciones de la ecuación (2.142). La solución general de la ecuación (2.142) es entonces la combinación lineal

$$u^n = c_1 \gamma_1^n + c_2 \gamma_2^n + \ldots + c_q \gamma_q^n \qquad (2.145)$$

donde $c_1, c_2, \ldots, c_n$ son constantes. Cada término de (2.145) es una solución particular de la ecuación homogénea.

El caso de raíces múltiples o complejas puede ser tratado de una manera similar (ver por ej. Isaacson-Keller, 1966, entre otros) a la usada para ecuaciones diferenciales ordinarias.

Volviendo a la ecuación lineal homogénea en diferencias (2.140) la ecuación característica resulta

$$(1 - ak/3)\,\gamma^2 - (4\,ak/3)\,\gamma - (1 + ak/3) = 0 \qquad (2.146)$$

Esta es una ecuación cuadrática cuyas raíces están dadas por

$$\gamma_{1,2} = \frac{2\,ak/3 \pm (1 + a^2k^2/3)^{1/2}}{1 - ak/3} \qquad (2.147)$$

Definiendo una nueva variable $y = ak$ en la Fig. 2.17 se muestra el gráfico de $\gamma_{1,2}$ en función de y en el intervalo $-1 < y < 1$.

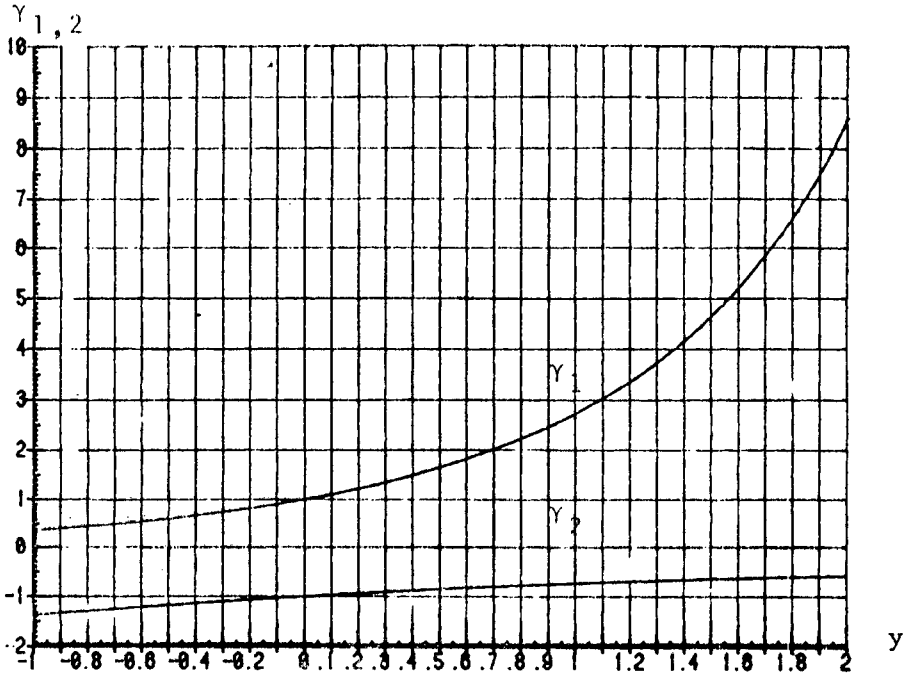

Fig. 2.17 Gráfico de las raíces de la ecuación característica del corrector de Milne

Podemos observar que la raíz γ_1 es monotónica creciente con un valor mayor que 1 para valores positivos de y mientras que la raíz γ_2 es menor que -1 para valores de y negativos.

Expandiendo γ_1 y γ_2 en serie de potencias resulta

$$\gamma_1 = 1 + ak + \frac{(ak)^2}{2} + \frac{(ak)^3}{6} + \frac{(ak)^4}{24} + \frac{(ak)^5}{72} + \ldots.$$

$$= e^{ak} + \frac{(ak)^5}{180} + \ldots. \qquad (2.148)$$

$$\gamma_2 = -1 + \frac{ak}{3} - \frac{(ak)^2}{18} - \frac{(ak)^3}{54} + \frac{5(ak)^4}{648} + \frac{5(ak)^5}{1944} + \ldots$$

(2.149)

La solución de la ecuación en diferencias (2.140) toma la forma de la ecuación (2.145) es decir

$$u^n = c_1 \gamma_1^n + c_2 \gamma_2^n \qquad (2.150)$$

o teniendo en cuenta la ecuación (2.148)

$$u^n \cong c_1 e^{nak} + c_2 \gamma_2^n = c_1 e^{at} + c_2 \gamma_2^n \qquad (2.151)$$

De la comparación entre la solución exacta de la ecuación en diferencias y la solución de la ecuación diferencial surge que la solución particular de la ecuación homogénea en diferencias correspondiente a la raíz γ_1 es la solución de la ecuación diferencial. La ecuación en diferencias posee una segunda solución particular $c_2 \gamma_2^n$ totalmente extraña al problema diferencial primitivo y que por ello se denomina solución parásita. Este tipo de soluciones parásitas surgen siempre que se aproxima un problema diferencial de primer orden con una ecuación en diferencias de un orden mayor. Es preciso analizar el comportamiento de las soluciones parásitas para ver si pueden llegar a tener una influencia negativa sobre la solución verdadera. Analicemos primero el caso de una ecuación de crecimiento (en este caso $a > 0$). La solución asociada a la raíz γ_1 crece en forma exponencial pudiendo aproximar la ecuación diferencial (el método es numéricamente inestable). La solución parásita asociada a la raíz γ_2 decrece para valores positivos de y de modo que cualquier error introducido no crecerá en forma no acotada debido a la presencia de dicha solución. Cuando se trata de una ecuación de decaimiento (en este caso $a < 0$), la solución asociada a la raíz γ_1, decrece para valores negativos de y aproximando el problema diferencial (el método es numéricamente estable para valores de y mayores que -1). La solución parásita asociada a la raíz γ_2 crece para valores negativos de y dado que la raíz γ_2 es menor que -1. En realidad debido a que γ_2 es negativa, la solución resulta de signo oscilante y cualquier error introducido crecerá en forma no acotada y oscilatoria. Eventualmente la presencia de la solución parásita hará que los resultados obtenidos

dejen de ser valederos. Los métodos numéricos que poseen estas características suelen denominarse métodos débilmente inestables o débilmente estables.

Si establecemos que γ_1 es la raíz de la ecuación característica asociada a la solución verdadera del problema diferencial de valores iniciales dado, entonces se denomina a γ_1 la raíz principal. Si se cumple que

$$|\gamma_i| < |\gamma_1| \qquad i = 2,3,\ldots,q \tag{2.152}$$

el método es relativamente estable. En este caso, la aproximación a la solución verdadera domina las soluciones parásitas asociadas a las raíces γ_i.

Se denomina fuertemente estable al método para el que se cumple que

$$|\gamma_i| < 1 \quad , \quad i = 2,3,\ldots,q \tag{2.153}$$

En este caso todas las soluciones parásitas decrecen para valores crecientes de n. Dahlquist y Bjorck, 1974, presentan el siguiente criterio llamado "condición de la raíz" como condición necesaria y suficiente para la estabilidad de los métodos lineales de paso múltiple: todas las raíces γ_i deben estar ubicadas dentro o sobre el círculo de radio unidad y las raíces de módulo unidad deben ser simples.

2.18 Polialgoritmos para ecuaciones diferenciales ordinarias

La construcción de un programa eficiente y confiable para resolver ecuaciones diferenciales ordinarias es una tarea mucho más compleja que la simple elección y programación del mejor de los métodos numéricos presentados. Un buen programa necesita muchas componentes integradas cuidadosamente entre sí y que conforman un polialgoritmo. Los componentes principales de un programa para resolver ecuaciones diferenciales ordinarias son (Rice, 1979):

1. Inicialización: determinación del paso inicial y de los métodos de paso entero múltiple de inicialización.
2. Control de la longitud de paso: determinación de la longitud de paso a los efectos de controlar el error local, cambiando el tamaño de malla para métodos de paso múltiple.
3. Control del orden del método: selección y cambio del orden del método particular a ser utilizado entre una misma familia de métodos.
4. Salida: Cálculo de valores de la solución en un conjunto predeterminado de puntos en vez de en aquellos puntos determinados por el programa.
5. Control del error global: Estimación del efecto acumulativo de los errores locales y control del cálculo de manera de mantener el error global dentro de una tolerancia predeterminada.
6. Chequeo: Cada programa está basado en un modelo que representa una clase de problemas a los que va a resolver. Un buen pro-

grama efectúa varios chequeos para ver si la entrada y el problema son consistentes con el modelo. Uno de los logros principales de la década del 70 en computación fue el desarrollo de programas eficientes y confiables para la solución numérica de ecuaciones diferenciales ordinarias. Aquí sólo discutiremos brevemente las características principales de las mismas.

Inicialización

Los métodos de un solo paso entero no presentan dificultad computacional alguna en el primer paso, salvo la elección del tamaño del mismo. El próximo paso es el control de la longitud de paso mucho de lo cual se aplica aquí. Los métodos de control de la longitud de paso, en general, consisten en comparar resultados actuales (por ejemplo, estimación del error local) con resultados previos. Dado que no existen resultados previos para el primer paso se deben extremar las precauciones. El peligro es que un primer paso muy grande puede producir un error inicial que arruina la precisión de todo el cálculo. En general las estrategias suelen ser muy conservativas y se elige un tamaño de paso menor que el necesario. Por supuesto que si el paso es demasiado pequeño hay un derroche de esfuerzo computacional y el método de control de la longitud de paso debe incrementarlo rápidamente.

Muchos programas piden al usuario que determine un paso de tiempo inicial. En general el usuario no tiene idea del tamaño de malla apropiado y por lo tanto el programa que depende del usuario para una información tan importante no es confiable. Algunos programas piden al usuario que sugiera un paso de tiempo pero sólo lo usan como punto de partida para su propia determinación. Esta metodología es muy razonable cuando se resuelven un conjunto de ecuaciones diferenciales de características similares: se determina el paso inicial óptimo para alguna de ellas y para el resto se comienza con el mismo valor.

Los métodos de paso múltiple, como ya hemos visto, tienen la dificultad adicional de necesitar varios valores previos en cada paso de tiempo. En general se comienza con un Runge-Kutta para los primeros pasos y luego se cambia al método de paso múltiple. Otra metodología consiste en tener una familia de métodos de paso múltiple de distinto orden y que incluyen un método de un sólo paso. El cálculo comienza con un método de un sólo paso para el primer paso, luego se cambia a un método de dos pasos para el segundo paso. El número de pasos en el método es incrementado en un paso hasta obtener el número deseado. Una vez que el cálculo ha comenzado se comienza a utilizar el método general para el control del paso de malla y para la selección del orden del método.

Control de la longitud de paso

Aquí el problema consiste en estimar el error local de truncamiento o de discretización. La mayoría de los programas sólo intentan mantener el error dentro de los límites de la tolerancia preestablecida. En general no se puede estimar con pre-

cisión el error sin resolver el problema dos veces o de una manera más precisa que la deseada. Por ejemplo, podríamos querer calcular u^{n+1} con una fórmula de $O(k^4)$ y u^{*n+1} con una fórmula de $O(k^5)$; salvo que k sea un valor totalmente no realista, u^{*n+1} es una mejor estimación de la solución exacta en t^{n+1} que u^{n+1} por lo que $| u^{* n+1} - u^{n+1} |$ es una estimación razonable del error en u^{n+1}. La idea es que aunque el error en u^{*n+1} no es conocido es sabido que es menor que el error en u^{n+1} .

La estimación del error local es comparado con la tolerancia requerida. La mayoría de los programas utilizan dos tipos de tolerancia:

TOLABS = valor absoluto de la tolerancia
TOLREL = tolerancia del error relativo

Si U^{n+1} es la solución exacta o genuina en los nodos de la malla el usuario espera que el error global de truncamiento o discretización satisfaga:

$$| u^{n+1} - U^{n+1} | < \text{TOLABS} + \text{TOLREL} | U^{n+1} | \qquad (2.154)$$

Sin embargo la mayoría de los programas intentan controlar el error local por cada paso de tiempo

$$| u^{n+1} - U^{* n+1} | < \text{TOLABS} + \text{TOLREL} | U^{* n+1} | \qquad (2.155)$$

donde $U^{* n+1}$ es la solución de la ecuación diferencial que pasa por el punto (u^n, t^n) .

Un método para obtener un control parcial del error global es distribuir el requerimiento del usuario sobre el intervalo $[a,b]$ de integración, reemplazando el miembro de la derecha de la ecuación (2.155) por TOLABS*k/(a-b)+TOLREL*k/(a-b). Esto es denominado criterio del error por unidad de paso. Otra alternativa para estimar el error local de truncamiento es utilizar fórmulas con el mismo orden cuyos coeficientes de errores son conocidos (ver punto c de la sección anterior). Una vez obtenida una buena estimación del error para cada paso de tiempo el problema consiste en determinar en base al mismo si hay necesidad de afinar o agrandar la malla. Si el error es muy grande k debe ser achicado. Se podría calcular una estimación de cuánto se debe cambiar k; si el error local de truncamiento es de $O(k^5)$ entonces duplicando o reduciendo k a la mitad el error cambia por un factor de aproximadamente 32. Este procedimiento es razonable para métodos de un solo paso. Para métodos de paso múltiple la situación se complica ligeramente (ver punto d de la sección anterior).

Control del orden del método

La idea básica de un método de orden variable consiste en hacer una proyección del error local de truncamiento en cada paso de tiempo utilizando el método corriente de $O(kP)$ y un

método de $O(k^{p+1})$ y otro de $O(k^{p-1})$, todos de la misma familia. Luego se utiliza aquel método que resulte en un menor error. La esencia del método está basada en el hecho de que existen modos de calcular las tres estimaciones del error con un pequeño esfuerzo adicional al necesario para calcular solamente uno. Estos métodos que son muy complejos para ser descriptos aquí han resultado ser en la práctica confiables y eficientes. Hacemos notar que existe una interacción entre el cambio de orden y de paso de malla. Uno quisiera saber si es mejor, por ejemplo, cambiar el orden y la malla simultáneamente. Los buenos programas poseen tests para esta interacción. La interacción entre cambio de orden y de malla no es simple, los cambios deben hacerse lentamente para no introducir inestabilidades en el cálculo.

Control de la salida

Cuando se utilizan polialgoritmos complejos no es posible predecir la posición y el número de puntos donde la solución es computada. Muchos programas permiten que el usuario especifique el conjunto de puntos donde desea obtener valores de la solución. Esto es muy útil, por ejemplo, cuando se desea dibujar la solución y se está usando una rutina de dibujo que requiere datos equiespaciados.

Control del error global

No existen programas que realmente controlen el error global. Para tener una idea de la dificultad imaginemos una ecuación diferencial $du/dt - f(u,t) = 0$ que tiene dos puntos especiales t_1 y t_2. En t_1 la ecuación diferencial se torna inestable y las trayectorias rápidamente se alejan entre sí. En t_2 ocurre lo contrario y las trayectorias se acercan unas a otras y se mantienen así por un cierto tiempo. Ahora consideremos un esquema numérico que tiene el error global bajo control y que llega al punto t_1. La solución computada no está exactamente en la trayectoria de la solución genuina de modo que pasado el punto t_1 comienzan a producirse grandes errores debido justamente a que las trayectorias se alejan. La única manera de mantenerse dentro del límite requerido por el error global es volver atrás al punto inicial y rehacer los cálculos con una precisión mayor.

Un programa práctico no recomenzará por si mismo el cálculo cada vez que encuentra un error global demasiado grande, dado que esto significaría decenas de reintegraciones sobre una parte del intervalo. Por el contrario irá hasta el final del intervalo de integración y luego estimará la precisión necesaria sobre la marcha antes de recomenzar el cálculo. Con suerte, una o dos iteraciones con este procedimiento producirán la precisión global deseada.

El comportamiento en el punto t_2 tiene el efecto contrario; se descubre que la precisión es repentinamente mucho mayor de la requerida. Raramente se volverán a realizar los cálcu-

los para obtener resultados de menor precisión pero por una cuestión de principios obtener demasiada precisión es tan malo como obtener demasiado poca. Todo lo anterior explica el porqué de la inexistencia de programas standard que controlen el error global. Sin embargo no existe una dificultad fundamental en la construcción de un programa que controle el error global siempre que se pueda estimar el mismo en forma confiable.

Esto nos lleva a la cuestión central: ¿cómo se puede estimar el error de truncamiento global? Primero hacemos notar que la idea del control del error por unidad de paso de tiempo no controla realmente el error global a pesar de que es un intento válido para gobernar el error local en un contexto global. El método básico para estimar el error global en ecuaciones diferenciales ordinarias es el mismo que para otros métodos numéricos; se resuelve el problema más de una vez y luego se comparan los resultados. El conocimiento de los términos del error permite hacer estimaciones cuantitativas del mismo modo que las que se hicieron para el error local. Desgraciadamente la incertidumbre en la estimación global del error es mucho mayor debido a que los puntos medios en los términos de los errores pueden estar ubicados en cualquier lugar en intervalos grandes por lo que resulta mucho menos probable que términos como $u^{(4)}(\xi_1)$ y $u^{(4)}(\xi_2)$ sean casi iguales.

El método más simple consiste en resolver la ecuación dos veces con el mismo programa utilizando distintas tolerancias para el error. La trampa aquí radica en que el cambio de la tolerancia puede tener efectos impredecibles -o ningún efecto- en el cálculo.

Lo más apropiado es tener un programa que esté específicamente diseñado para obtener estimaciones del error global. La mejor manera de realizar esto es haciendo que la ecuación sea resuelta en paralelo por dos métodos. Este método es relativamente nuevo pero las estimaciones del error global son mas confiables que reutilizando un programa standard. Este método está implementado en el programa GERK (ACM algoritmo número 504).

Chequeo

Hay tres tipos de chequeo. El primer tipo es para controlar errores triviales en los datos de entrada tales como: ¿es la precisión requerida positiva?, ¿es la longitud del intervalo de tiempo nula?, ¿es el número de ecuaciones positivo?, ¿es el dimensionamiento de los arreglos compatible con el número de ecuaciones especificada?, etc.

El segundo tipo de chequeo es para constatar la razonabilidad del problema presentado: ¿es el requerimiento de precisión relativa compatible con la precisión de la computadora?, ¿es el paso inicial significativo, es el volumen de salida razonable?.

El tercer tipo de chequeo se refiere al modelo de ecuación diferencial que el programa contiene. Ejemplos de este tipo de chequeo son: ¿ha decrecido el paso en forma repentina a un valor casi insignificante (presencia de una posible singularidad)? ¿Hay un progreso razonable hacia la solución del problema? Si se han realizado mil pasos y sólo se ha cubierto el uno por ciento del intervalo de integración es muy posible que algo ande mal. ¿Se ha vuelto inestable el cálculo?.

Estos ejemplos ilustran el tipo y extensión del chequeo que debe esperarse de un paquete de software de buena calidad.

2.19 Sistemas de ecuaciones diferenciales

Vamos a presentar la extensión de algunos de los métodos para ecuaciones diferenciales escalares a sistemas de ecuaciones diferenciales de primer orden. Recordando que un problema de valores iniciales para un sistema de p ecuaciones diferenciales de primer orden se puede escribir como

$$\begin{aligned} du_1/dt + f_1(u_1,u_2,\ldots,u_p,t) &= 0 \\ du_2/dt + f_2(u_1,u_2,\ldots,u_p,t) &= 0 \\ &\vdots \\ du_i/dt + f_i(u_1,u_2,\ldots,u_i,\ldots,u_p,t) &= 0 \\ &\vdots \\ du_p/dt + f_p(u_1,u_2,\ldots,u_p,t) &= 0 \end{aligned} \qquad (2.156)$$

donde $0 < t < \infty$, con condiciones iniciales

$$u_i(0) = u_{i0} \quad i = 1,\ldots,p \; ; \qquad (2.157)$$

o en forma vectorial

$$du/dt + f(u,t) = 0 \quad , \quad 0 < t < \infty \qquad (2.158)$$

$$u(0) = u_0$$

donde ahora u y f son vectores con p componentes u_i y f_i, respectivamente. El método de Euler en forma vectorial se escribe

$$u^{n+1} = u^n - kf^n \qquad (2.159)$$

Si escribimos la expresión anterior para la componente i tendremos

$$u_i^{n+1} = u_i^n - kf_i(u_1^n,u_2^n,\ldots,u_i^n,\ldots,u_p^n)$$

$$i = 1,\ldots,p \qquad (2.160)$$

El método fuertemente implícito en forma vectorial resulta

$$u^{n+1} = u^n - kf^{n+1} \qquad (2.161)$$

o para la componente i

$$u_i^{n+1} = u_i^n - kf_i(u_1^{n+1}, u_2^{n+1}, \ldots, u_i^{n+1}, \ldots, u_p^{n+1})$$

$$i = 1, \ldots, p \qquad (2.162)$$

El método de Runge-Kutta de orden 2 en forma vectorial se escribe (para la ecuación du/dt = f(u,t))

$$q_1 = k\ f(u^n, t^n) \qquad (2.163)$$

$$q_2 = k\ f(u^n + q_1, t^n + k) \qquad (2.164)$$

$$u^{n+1} = u^n + 1/2\ (q_1 + q_2) \qquad (2.165)$$

o para la componente i

$$q_{1i} = k\ f_i(u_1^n, u_2^n, \ldots, u_i^n, \ldots, u_p^n, t^n) \qquad (2.166)$$

$$q_{2i} = k\ f_i(u_1^n + q_{11}, u_2^n + q_{12}, \ldots, u_i^n + q_{1i}, \ldots, u_p^n + q_{1p}, t^n + k) \qquad (2.167)$$

$$u_i^{n+1} = u_i + 1/2\ (q_{1i} + q_{2i}) \qquad (2.168)$$

Finalmente el método predictor-corrector implícito centrado (Crank-Nicolson) en forma vectorial se escribe

$$u_{m=0}^{n+1} = u^n - k\ f(u^n, t^n) \qquad (2.169)$$

$$u_{m+1}^{n+1} = u^n - \frac{k}{2}\left[f(u_m^{n+1}, t^{n+1}) + f(u^n, t^n)\right] \qquad (2.170)$$

o para la componente i

$$u_{i,m=0}^{n+1} = u_i^n - k\ f_i(u_1^n, u_2^n, \ldots, u_i^n, \ldots, u_p^n, t^n) \qquad (2.171)$$

$$u_{i,m+1}^{n+1} = u_i^n - \frac{k}{2}\left[f_i(u_{1,m}^{n+1}, u_{2,m}^{n+1}, \ldots, u_{i,m}^{n+1}, \ldots, u_{p,m}^{n+1}, t^{n+1}) + f_i(u_1^n, u_2^n, \ldots, u_i^n, \ldots, u_p^n, t^n)\right] \qquad (2.172)$$

Está claro que la solución de un sistema de p ecuaciones diferenciales con p incógnitas u_i consiste en una familia de p curvas o trayectorias en el plano u,t. También observamos que la

solución numérica de sistemas diferenciales se obtiene extendiendo las fórmulas básicas para escalares por medio del reemplazo de ciertas cantidades escalares por vectores. Las nuevas dificultades que se presentan con sistemas son: la más obvia es que el volumen de cálculo por paso de tiempo se incrementa notablemente por la evaluación, por ejemplo, de p funciones de p variables de f(u,t); el trabajo puede llegar a ser p^2 veces más que el correspondiente a una sola función de una sola variable. Asimismo hay un incremento de $O(p^2)$ de memoria. Una dificultad mucho más importante es la posibilidad que las ecuaciones y soluciones del sistema tengan un comportamiento muy diferente entre sí. Por ejemplo es muy frecuente la aparición de problemas denominados rígidos (stiff problems, en la literatura sajona) producidos por la existencia de ecuaciones diferenciales que describen procesos con escalas de tiempo muy diferentes. En física e ingeniería se utiliza el término "constante de tiempo" para referirse a la velocidad de decaimiento de un determinado proceso. Por ejemplo, en la ecuación de decaimiento $du/dt + au = 0$ cuya solución es $u(t) = u_0 e^{-at}$, la función $u(t)$ decae en un factor e^{-1} en el tiempo $1/a$. El término $1/a$ es la constante de tiempo; cuanto mayor es el valor de a más breve es la constante de tiempo. En el caso de sistemas de ecuaciones diferenciales, distintas componentes pueden decaer con velocidades muy diferentes. Por ejemplo en determinados procesos químicos la constante de tiempo puede ser del orden de una centésima de segundo para una determinada reacción y del orden del minuto para otra. En procesos químicos complejos donde se producen decenas de reacciones es fácil obtener sistemas de ecuaciones diferenciales con más de cien componentes ($p=100$) cuyas soluciones tienen un comportamiento muy diferente entre si. Para sistemas gobernados por la ecuación $du/dt + f(u,t) = 0$ la velocidad de decaimiento se puede relacionar localmente con los autovalores de $\partial f/\partial u$. Si como hemos dicho algunas de las reacciones son rápidas y otras son lentas, las primeras gobiernan el proceso de estabilidad numérica aún en el caso que hayan decaido a niveles insignificantes y que el error de truncamiento esté determinado por los componentes lentos. Este es el problema de ecuaciones "rígidas". A continuación haremos un análisis introductorio al tema.

2.20 Ecuaciones diferenciales rígidas

Vamos a presentar un análisis suscinto del problema de ecuaciones diferenciales ordinarias; el lector interesado en el tema puede consultar Gear, 1971, Dahlquist y Ljorck, 1971, ó Henrici, 1962, entre otros. Una excelente introducción al tema se encuentra en el artículo de Shampine y Gear, 1979. Consideremos el siguiente ejemplo tomado de Gear

$$du_1/dt = 998u_1 + 1998u_2 \qquad (2.173)$$

$$du_2/dt = -999u_1 - 1999u_2 \qquad (2.174)$$

con condiciones iniciales

$$u_1(0) = 1 \tag{2.175}$$

$$u_2(0) = 0 \tag{2.176}$$

La solución exacta del sistema (2.173)-(2.176) se escribe

$$u_1 = 2\,e^{-t} - e^{-1000t} \tag{2.177}$$

$$u_2 = -e^{-t} + e^{-1000t} \tag{2.178}$$

Ambas soluciones contienen componentes rápidas y lentas. La figura 2.18 ilustra las soluciones (2.177) y (2.178).

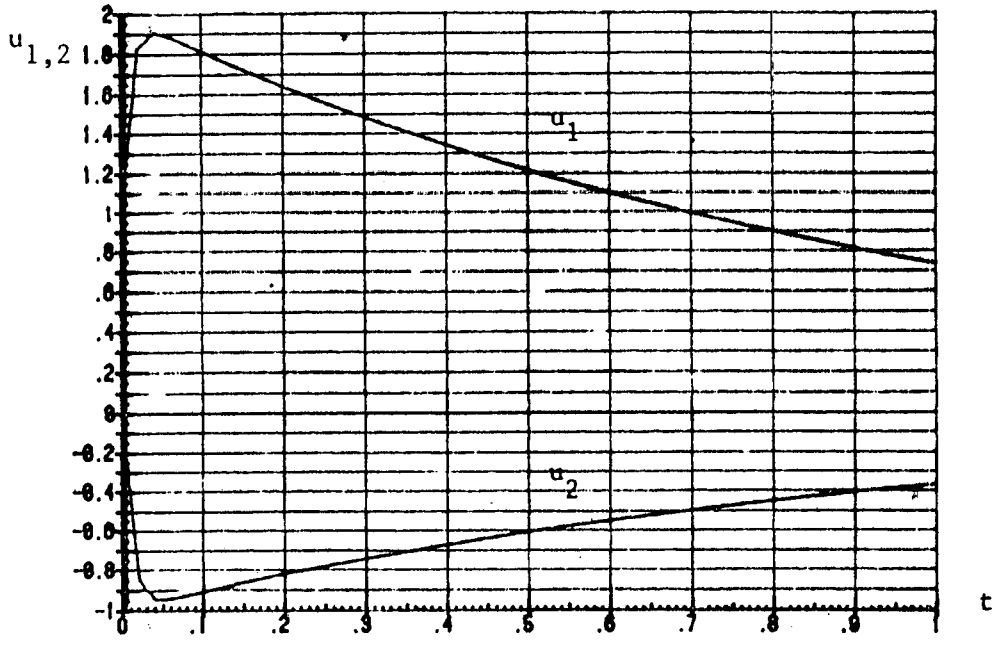

Fig. 2.18 solución exacta del problema (2.173)-(2.176)

Por ejemplo, la componente rápida en la solución de $u_1(t)$ es insignificante para $t=0.01$ ($e^{-1000t}=10^{-5}$), lo mismo sucede para $u_2(t)$. La solución de este problema mediante la mayoría de los métodos numéricos presentados genera requerimientos en la condición de estabilidad que hacen prácticamente imposible su aplicación. Recordemos que la aplicación del método de Euler a una ecuación escalar $du/dt + f(u,t) = 0$ produce una restricción en la estabilidad del tipo $k < 2/M$ donde $M = \partial f/\partial u$, $M > 0$. Por ejemplo si $f(u) = 1000\,u$ el paso temporal debe cumplir la condición $k<2/1000$. La aplicación del método de Euler al sistema (2.173)-(2.174) produce las mismas restricciones en el valor de k aún cuando la componente rápida se haya vuelto insignificante.

El problema de la rigidez también puede manifestarse en el caso de una ecuación escalar. Por ejemplo, la ecuación

$$du/dt = a[u - f(t)] + df(t)/dt \ , \ u(0) = f(0) + c_0 \qquad (2.179)$$

con $a<0$ tiene un comportamiento similar si $f(t)$ es una función lisa.

La solución de la ecuación (2.179) es

$$u(t) = [u(0) - f(0)]e^{at} + f(t) \qquad (2.180)$$

Aún en el caso en que el término entre corchetes sea distinto de cero, at adquirirá rápidamente un valor suficientemente negativo de modo tal que el valor del primer término del segundo miembro o componente rápida se tornará despreciable frente al segundo término o componente lenta. Si utilizamos el método de Euler para obtener una solución numérica de (2.179) veríamos que cuando at tiende a $-\infty$ el error de truncamiento está determinado por el valor de k y de la derivada de la función f mientras que la estabilidad del método depende de a y k .

En el caso de sistemas de ecuaciones diferenciales de la forma

$$du/dt = A[u - f(t)] + df(t)/dt \qquad (2.181)$$

los autovalores de A juegan el papel de a. En este caso se deben considerar valores de a complejos. Si todos los autovalores de A tienen parte real negativa, la solución numérica de (2.181) converge a f(t) para t tendiendo a infinito.

Si bien es cierto que se pueden utilizar algunos de los métodos numéricos presentados para resolver ecuaciones rígidas también es cierto que los pasos de tiempo deben ser tan pequeños para asegurar la estabilidad que el volumen de cálculo y los errores de redondeo se vuelven críticos. La solución del problema pasa por desarrollar métodos en los que no haya restricción del paso temporal por problemas de estabilidad. De acuerdo a Dahlquist y Bjorck, 1974, los métodos que dan mejores resultados son naturalmente los métodos implícitos: por ejemplo el método fuertemente implícito ya visto, combinado con un procedimiento de alisado y extrapolación de Richardson. Se hace notar que en el caso de tener que realizar iteraciones internas el método de iteración de punto fijo presentado con anterioridad no es muy apropiado ya que la condición de convergencia depende de $\partial f/\partial u$ lo que implica que la componente rápida limita la velocidad de convergencia y por lo tanto el paso de malla. En general se utilizan en estos casos variantes del método de Newton-Raphson.

2.21 Método del tiro

En la literatura sajona se da el nombre de métodos de tiro (Shooting) a los métodos de valores iniciales utilizados para resolver problemas de contorno. Este nombre surge de la metáfora de la práctica del tirador al blanco de corregir el tiro en base al resultado del tiro precedente. Un ejemplo sencillo ilustra el método. Supongamos querer resolver el siguiente problema de contorno:

$$-u'' = f(u) \quad , \quad 0 < x < 1 \quad , \quad u(0) = u_0 \quad , \quad u(1) = u_1 \qquad (2.182)$$

En cambio, resolvemos el siguiente problema de valores iniciales:

$$-u'' = f(u) \quad , \quad u(0) = u_0 \quad y \quad u'(0) = \alpha_1 \qquad (2.183)$$

donde α_1 es el primer ensayo para $u'(0)$. Supongamos que resolvemos el problema de valores iniciales (2.183) por cualquiera de las técnicas vistas y que obtenemos un valor en $x=1$ de $u(1)=\beta_1$, según se ilustra en al figura 2.19. En general $\beta_1 \neq u_1$ y volvemos a ensayar un nuevo disparo $u'(0)=\alpha_2$ y esta vez resolviendo el problema (2.183) obtenemos $u(1)=\beta_2$.

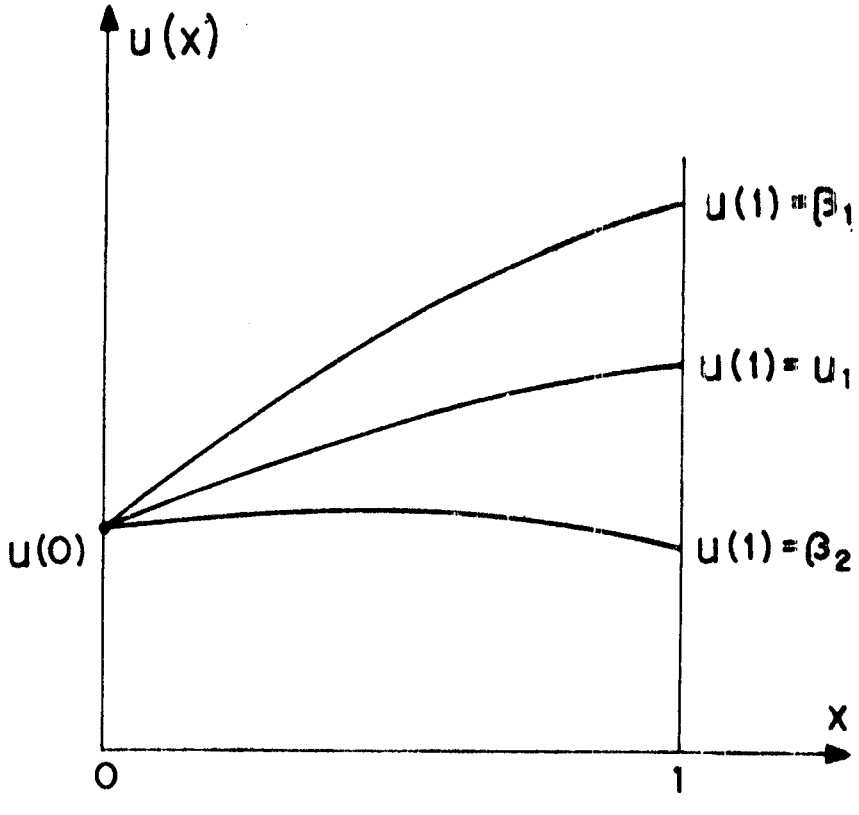

FIG. 2.19 Esquematización del método de tiro.

Si suponemos que existe una relación lineal entre α y β resulta que:

$$\beta = \frac{\alpha_2\beta_1 - \alpha_1\beta_2}{\alpha_2 - \alpha_1} + \frac{\beta_2 - \beta_1}{\alpha_2 - \alpha_1}\,\alpha \qquad (2.184)$$

y el próximo ensayo de tiro se ejecuta con un ángulo α correspondiente a $\beta=u_1$. En general esto no basta y se deben hacer varios ensayos más con un grado variado de sofisticación del método esperando que el proceso converja. El lector interesado puede consultar el libro de Na, 1979.

2.22 El método de Zadunaisky

En numerosas ramas de ciencia y tecnología (tales como en Astronáutica, Astronomía, Balística, etc.) es preciso resolver ecuaciones diferenciales con errores muy pequeños, de allí la necesidad de contar con métodos de integración de alta precisión y de una buena estimación de los errores globales de los mismos. En este contexto presentaremos un método relativamente reciente introducido por Zadunaisky, 1976 (existe una versión resumida del mismo en la compilación de Marshall, 1978), que consiste en un conjunto unificado de procedimientos destinados a estimar y corregir los errores globales propagados en la integración numérica de ecuaciones diferenciales ordinarias.

Dado un problema de valores iniciales

$$du/dt - f(u,t) = 0\ , \quad u(0) = u_0\ , \quad t=0 \qquad (2.185)$$

y una solución aproximada u^n obtenida mediante un método de diferencias finitas, el error global se define como

$$e^n = u^n - U^n \qquad (2.186)$$

donde U^n es la solución exacta en $t=t_n$. El objeto es encontrar una buena estimación del error global e^n, para ello se supone, en lo que sigue, que aquel es debido únicamente a errores de truncamiento. La idea básica del método de Zadunaisky consiste en, una vez resuelto numéricamente el problema original (2.185), construir un pseudo problema de la forma

$$dw/dt - F(w,t) = 0\ , \quad w(0) = w_0\ , t>0 \qquad (2.187)$$

de modo tal que su solución exacta $W(t_n)$ sea conocida de antemano y que además difiera poco de la solución numérica u^n del problema original (2.185). El pseudo problema (2.187) se integra con el mismo método utilizado para el problema original y se obtiene la solución numérica w^n. El error global esta definido de manera exacta por

$$E^n = w^n - W(t_n) \qquad (2.188)$$

Si para cada paso de tiempo la solución numérica w^n del pseudo problema (2.187) difiere poco de la solución numérica u^n

del problema original, es razonable esperar que los errores tengan el mismo comportamiento y por lo tanto tomar E^n como una buena aproximación de e^n.
El pseudo problema puede ser construido de la siguiente manera. Una vez obtenida la solución numérica u^n es posible encontrar una función polinomial P(x) (por ejemplo) que ajusta los valores de u^n de la mejor manera posible.
Entonces conociendo P'(t) y fijando

$$w(t) = P(t) \tag{2.189}$$

$$F(w,t) - f(w,t) - D(t) = 0 \tag{2.190}$$

donde

$$D(t) = P'(t) - f(P,t) \tag{2.191}$$

el pseudo problema adopta la forma

$$du/dt - f(w,y) - D(t) = 0 \tag{2.192}$$

$$w(0) = P(0) = P_0 \tag{2.193}$$

cuya solución, evidentemente, es la expresión (2.189). Además si la función P(t) es elegida apropiadamente puede llegar a diferir de la solución numérica u^n en una cantidad arbitrariamente pequeña. Entonces, teniendo en cuenta la igualdad (2.189), la fórmula (2.188) para la estimación del error global resulta finalmente

$$E^n = w^n - P(t_n) \tag{2.194}$$

Como lo hace notar Zadunaisky, la idea básica del método es relativamente simple pero su implementación requiere técnicas numéricas apropiadas. El lector interesado puede consultar detalles sobre la implementación del método aplicada a problemas de valores iniciales y de contorno en ecuaciones diferênciales ordinarias y a ecuaciones en derivadas parciales en las referencias citadas más arriba. Al respecto cabe mencionar un trabajo de Echebest y Lieberman, que se encuentra en la compilación de Marshall, 1978, y que presenta la solución numérica de ecuaciones en derivadas parciales mediante la transformada rápida de Fourier y una estimación de los errores globales mediante el método de Zadunaisky.

2.23 Software para ecuaciones diferenciales ordinarias

Existen numerosas revistas que publican programas individuales de computación científica. Estos cubren una variada gama de tópicos, entre ellos la resolución numérica de ecuaciones diferenciales ordinarias. Las revistas más importantes son ACM Transactions on Mathematical Software, Applied Statistics, BIT, The Computer Journal y Numerische Mathematik. La serie de algoritmos del ACM contiene más de 500 programas (comenzó en

1960) y están agrupados en Collected Algorithms of the Association for Computing Machinery (ACM). Estos algoritmos se pueden obtener de: ACM Algorithms Distribution Center, c/o IMSL, Inc, Sixth Floor, NBC Building, 7500 Bellaire Boulevard, Houston, Texas 77036, USA.

En general, el software disponible para la solución numérica de ecuaciones diferenciales ordinarias puede ser dividido en varios grupos: Collected Algorithms of the ACM, Biblioteca H.S.L. de Subrutinas de Harwell, UKAEA, Inglaterra, Biblioteca NAG del Numerical Algorithm Group de la Universidad de Oxford, Inglaterra (esta biblioteca es comparable, en términos generales, con la H.S.L. pero posee, por ejemplo, muchas más rutinas estadísticas); Biblioteca IMSL de la International Mathematical and Statistical Library, Inc., Texas; Biblioteca SL/MATH de la Empresa IBM, Biblioteca SSP/CNEA del Centro de Cálculo Científico de la Comisión Nacional de Energía Atómica; y programas de diversas fuentes. A continuación presentamos una lista de dichos programas en lenguaje Fortran y cómo pueden obtenerse. Hacemos notar que existe una superposición de programas en los grupos mencionados y que muchas versiones son similares pero con algunas modificaciones. Es difícil determinar apriori cuál es mejor para un determinado problema.

DIFSUB, Algoritmo 407 del ACM que fue desarrollado por G.W. Gear, es una subrutina que integra un sistema de ecuaciones diferenciales de primer orden para un sólo paso de tiempo (en cada llamado a la misma). El paso de tiempo puede ser especificado por el usuario, pero es automáticamente modificado si la lógica interna del programa así lo requiere. DIFSUB utiliza un método predictor-corrector de Adams o un método de paso múltiple (asimismo apropiado para sistemas rígidos). El programa está muy bien documentado en el libro de G.W. Gear, 1971. En la SSP/CNEA del Centro de Cálculo Científico de la CNEA, Visconti, 1984, ha implemantado una versión de DIFSUB en simple y doble precisión que lleva los nombres DIFSUB y DDIFSU, respectivamente.

GERK, Algoritmo 504 del ACM, resuelve un sistema de ecuaciones diferenciales ordinarias por el método de Runge-Kutta-Fehlberg. Tiene la ventaja de ofrecer una estimación realista del error global a un costo marginal de entre el 20 y el 50 por ciento.

STINT, Algoritmo 534 del ACM, diseñado para sistemas diferenciales rígidos, utiliza un método de orden variable, paso variable y método de paso múltiple. Los programas mencionados pueden ser solicitados a la IMSL, Inc.

DEPACK, Differential Equation Package, es un conjunto de programas Fortran para la integración de sistemas de ecuaciones diferenciales ordinarias que se puede obtener del National Energy Software Center, Argonne National Laboratory, Argonne, Illinois 60439.

DE/STEP, es un programa Fortran (en dos versiones) para la solución de sistemas de ecuaciones diferenciales ordinarias que utiliza un método de Adams con paso variable y orden variable, fue realizado por L.F. Shampine y M.K. Gordon y se puede obtener del National Energy Software Center, Argonne National Laboratory, Argonne, Illinois 60439.

EPISODE, es un programa Fortran (en dos versiones: para ecuaciones diferenciales no rígidas y para ecuaciones rígidas). Utiliza el método de Adams con paso variable y orden variable. Está documentado en Byrne y Hindmarsh, 1975. El programa se puede obtener del National Energy Software Center.

RKF45, es un programa Fortran que utiliza el método de Runge-Kutta-Fehlberg de orden cuatro y cinco, fue desarrollado por Shampine y Watts y esta documentado en Forsythe, Malcolm y MOLER, 1977.

DVERK, resuelve un sistema de ecuaciones diferenciales ordinarias por el método de Runge-Kutta de orden cinco y seis, (en Fortran), se puede obtener de la IMSL, Inc.

DGEAR, es un programa que utiliza un método de Adams de orden variable y paso variable, escrito en Fortran y para ecuaciones diferenciales rígidas. Se puede obtener de la IMSL, Inc.

DREBS, es un programa Fortran que utiliza un método de extrapolación desarrollado por Bulrish-Stoer y se puede obtener de la IMSL, Inc.

DTPTB/DVCPR, son dos programas Fortran para un sistema de ecuaciones diferenciales de contorno. Utilizan un método de tiro en el cual el problema de valores iniciales usa el programa DVERK y un método de diferencias finitas con paso variable y correcciones diferidas respectivamente. Se pueden obtener de la IMSL, Inc.

RKMERS, es un programa Fortran para un sistema de ecuaciones diferenciales ordinarias de primer orden que utiliza Runge-Kutta-Merson de quinto orden y fue implementado por Visconti en la Biblioteca SS/PCNEA. La versión original del mismo es el algoritmo de Kutta-Merson 218 del ACM.

DC01AD, es un paquete de subrutinas para integrar un sistema de valores iniciales de primer orden utilizando un predictor-corrector desarrollado por Gear. Posee ajuste automático del paso y orden del método. Es apropiado incluso para problemas rígidos. El programa DC02AD es una interfase que facilita el uso del programa DC01AD. Esta versión en Fortran pertenece a la Biblioteca H.S.L. de Harwell.

DD01A, resuelve una ecuación de contorno lineal mediante un método de diferencias finitas. Está basada en un trabajo de Fox, 1947. La rutina DD02A resuelve un problema de contorno no lineal por un método iterativo linealizando en cada iteración y llamando a la rutina DD01A. Ambos programas Fortran pertenecen a la Biblioteca H.S.L. de Harwell.

DD03AD/DD04AD, resuelven sistemas de contorno utilizando el primero un método de tiro múltiple y Runge-Kutta de cuarto orden (Osborne, 1969; Keller, 1968); y el segundo, que es asimismo apropiado para problemas altamente no lineales, utiliza una técnica desarrollada por Lentini y Pereyra, 1977. Estos programas Fortran pertenecen a la Biblioteca H.S.L de Harwell.

DA01A, integra un sistema de valores iniciales por el método de Runge-Kutta-Merson de cuarto orden. Cada llamado a la subrutina avanza un paso de la integración. La subrutina DA01A, resuelve el mismo problema pero controlando automáticamente el paso y llamando a la DA01A. Estos programas Fortran pertenecen a la Biblioteca H.S.L de Harwell.

RK1/RK2, resuelven un problema de valores iniciales escalar por el método de Runge-Kutta de cuarto orden. El programa Fortran pertenece a la Biblioteca SL/MATH de IBM.

RKGS/DRKGS, resuelve un sistema de valores iniciales por el método de Runge-Kutta-Gill con ajuste automático del paso de malla. El programa Fortran pertenece a la Biblioteca SL/MATH de IBM.

HPCG/DHPCG, resuelven un sistema de valores iniciales por el método predictor-corrector de Hamming de cuarto orden. Tiene la ventaja sobre el programa anterior que sólo requiere evaluar dos veces la función del miembro de la derecha del sistema diferencial. Posee control automático del paso de malla. La versión Fortran de este programa pertenece a la Biblioteca SL/MATH de IBM.

LHVP/DLBVP, resuelven un sistema lineal de primer orden de ecuaciones diferenciales de contorno mediante el método de las ecuaciones adjuntas, que generan sucesivamente problemas adjuntos de valores iniciales. Estos últimos son resueltos por el método predictor-corrector de Hamming. La versión Fortran de este programa pertenece a la Biblioteca SL/MATH de IBM.

2.24 Referencias

G. D. Byrne and A. Hindmarsh, ACM Transactions Math. Software 1, 1975.

B. Carnahan, H. A. Luther, and J. O. Wilkes, Applied Numerical Methods, Wiley, 1969.

L. Collatz, The Numerical Treatment of Differential Equations, Springer, 1966.

G. Dahlquist and H. Bjorck, Numerical Methods, Prentice Hall, 1983.

G. Forsythe, M. Malcolm, and C. Moler, Computer Methods for Mathematical Computations, Prentice Hall, 1977.

L. FOX, Proc. Roy. Soc. A 190, 1947.

C. W. Gear, Numerical Initial Value Problems in Ordinary Differential Equations, Prentice Hall, 1971.

R. W. Hamming, Numerical Methods for Scientists and Engineers, Mc-Graw Hill, 1962.

P. Henrici, Discrete Variable Methods in Ordinary Differential Equations, Wiley, 1962.

E. Isaacson and H. B. Keller, Analysis of Numerical Methods, Mc-Graw Hill, 1966.

N. N. Jannenko, Methodes a Pas Fractionnaires, A. Colin, 1968.

H. B. Keller, Numerical Methods for Two-Point Boundary Value Problems, Blaisdell, 1968.

L. Lapidus and J. H. Seinfeld, Numerical Solution of Ordinary Differential Equations, Academic Press, 1971.

M. Lentini and V. Pereyra, SIAM J. Num. Anal. 14, 1977.

G. Marshall (Comp), Métodos Numéricos de la Mecánica del Continuo, EUDEBA, 1978

G. Marshall, A Front Tracking Method for One-Dimensional Moving Boundary Problems, SIAM Journal on Scientific and Statistical Computing, 1985 (en prensa); ver también CNEA-NT 22, 1983.

T. Y. Na, Computational Methods in Engineering Boundary Value Problems, Academic Press, 1979.

J. R. Rice, Software for Numerical Computation, in Research Directions in Software Technology, P. Wegner, Ed., M.I.T. Press, 1979.

R. D. Richtmyer and K. W. Morton, Difference Methods for initial value problems, Interscience, 1967.

L. F. Shampine and C. W. Gear, SIAM Review 21, No. 1, 1979.

V. Visconti, Una Comparación de Rutinas Para Integrar Sistemas de Ecuaciones Diferenciales Ordinarias Basadas en Métodos de un Solo Paso y en Métodos de Paso Múltiple Disponibles en la Comisión Nacional de Energía Atómica, CNEA-NT (en preparación).

P. E. Zadunaisky, Numerische Math. 27, 21, 1976.

2.25 Apéndice

Presentamos un ejemplo del uso de subrutinas científicas para resolver un problema de valores iniciales consistente en la modelización de la interacción de poblaciones de coyotes y conejos (el conocido problema predador-presa). El ejemplo fue tomado de la obra de Rice antes citada y para su resolución se utilizó la subrutina DDIFSU perteneciente al ACM (ver la sección sobre software para ecuaciones diferenciales ordinarias). La implementación numérica fue realizada por la licenciada María C. Rey del Centro de Cálculo Científico de la Comisión Nacional de Energía Atómica. El problema predador-presa consiste en la resolución del siguiente sistema de ecuaciones diferenciales ordinarias o problema de valores iniciales

$$R'(t) = P_r * R(t) - LUNCH(t) * C(t) \qquad (A.1)$$

$$C'(t) = P_c * C(t) - STARVE(t) * C(t) \qquad (A.2)$$

$$LUNCH(t) = EATRAB * \min (1 , R(t)/\alpha) \qquad (A.3)$$

$$STARVE(t) = e^{-\beta S(t)} \qquad (A.4)$$

$$S(t) = \frac{R(t)}{EATRAB * C(t)} \qquad (A.5)$$

donde: R(t) = densidad de conejos (por milla cuadrada) en función del tiemp; C(t) = densidad de coyotes; P_r = velocidad de crecimiento natural de los conejos (suponiendo que no hay coyotes que comen conejos); P_c = velocidad de crecimiento natural de los coyotes (suponiendo que hay infinitos conejos para comer); EATRAB = número de conejos que un coyote desearía comer por año; LUNCH(t) = número de conejos que un coyote realmente come por año; STARVE(t) = proporción de coyotes que se mueren por inanición por año ante la escasez de conejos; α = una vez que la densidad de conejos cae debajo de α los coyotes comienzan a comer menos conejos por año; y β = índice de inanición de los coyotes.

El rol de β está ilustrado por el hecho de que cuando S(t)=1 (hay suficientes conejos para que los coyotes se alimenten en un año), $STARVE(t) = e^{-\beta}$, por lo tanto la elección STARVE(t) = 2 por ciento nos da $\beta = -\ln(.02) = 1.7$.

Los parámetros α y β nos permiten ajustar el modelo y experimentar con el mismo.

A continuación se muestran distintas soluciones consistentes en un estado inicial transitorio seguido de un estado estacionario.

En la figura A1 se ve como la superpoblación inicial de conejos casi causa la explosión de ambas poblaciones.

En la figura A2 se ve como después de un pequeño crecimiento de la población de coyotes las dos poblaciones decrecieron a un estado estacionario.

En la figura A3 una superpoblación inicial de coyotes casi causa la desaparición de ambas poblaciones.

En la figura A4.a un cambio inicial rápido se produce antes de alcanzar un estado estacionario en 35 años.

Al modelo anterior lo hemos complicado agregando la siguiente dificultad numérica :
Migración de conejos : el valor de R salta de golpe a M en el tiempo C. Esto lo hacemos agregando un término a la primera ecuación del sistema. La función DIFFUN es modificada agregando las sentencias :

```
REAL*8  MIGRAT
DATA C , RATE , D / 28.65 , 20. , 1.0 /
MIGRAT = RATE
IF ( DABS(T-C) + DABS(T-C-D) .GT. D ) MIGRAT = 0.
```

En la figura A4.b se observa el efecto de una migración de 20 conejos por milla cuadrada en un año a partir del tiempo 28.65 teniendo hasta ese momento poblaciones estacionarias.

En la figura A5 se presenta un listado del programa FORTRAN juntamente con las subrutinas y los parámetros numéricos utilizados.

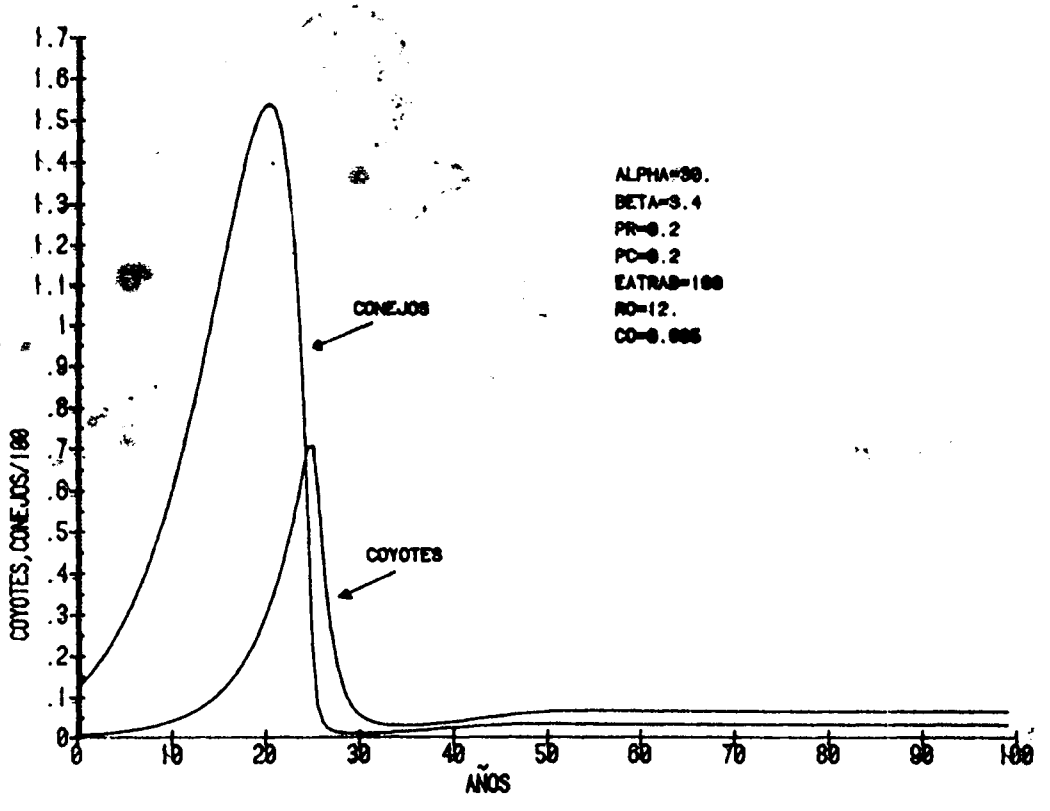

Fig. A.1 La superpoblación inicial de conejos casi causa la explosión de ambas poblaciones.

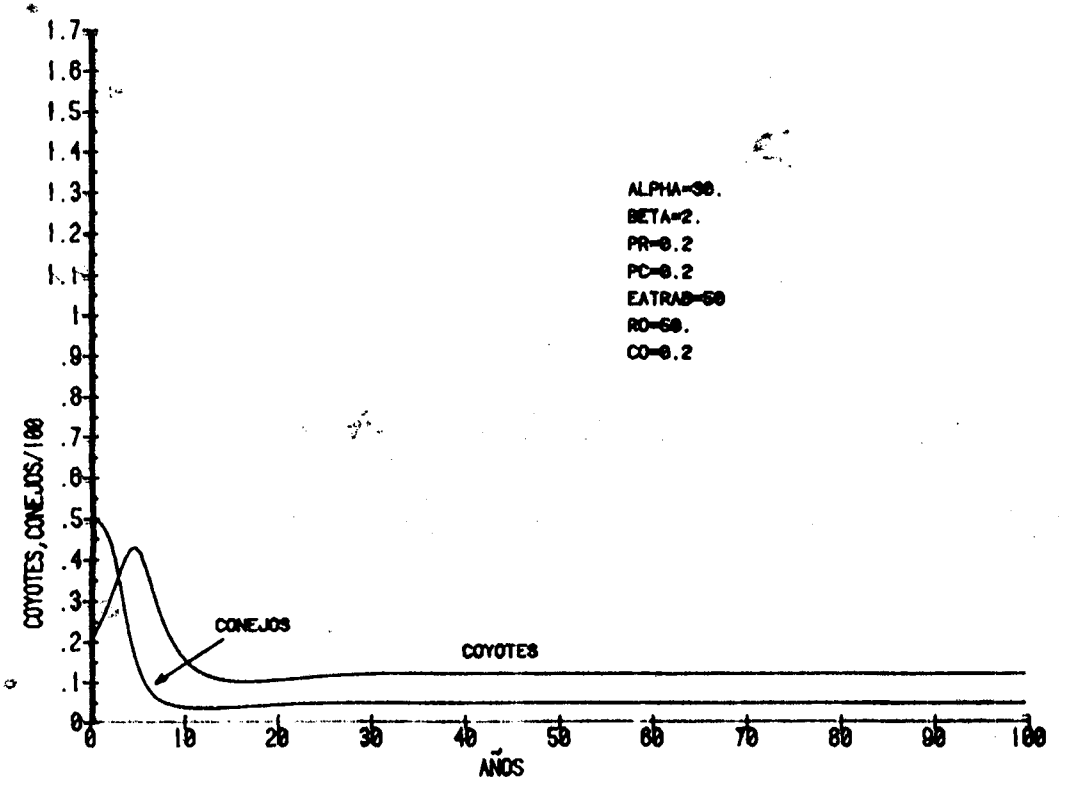

Fig. A.2 Luego de un pequeño crecimiento de la población de coyotes las dos poblaciones decrecen a un estado estacionario.

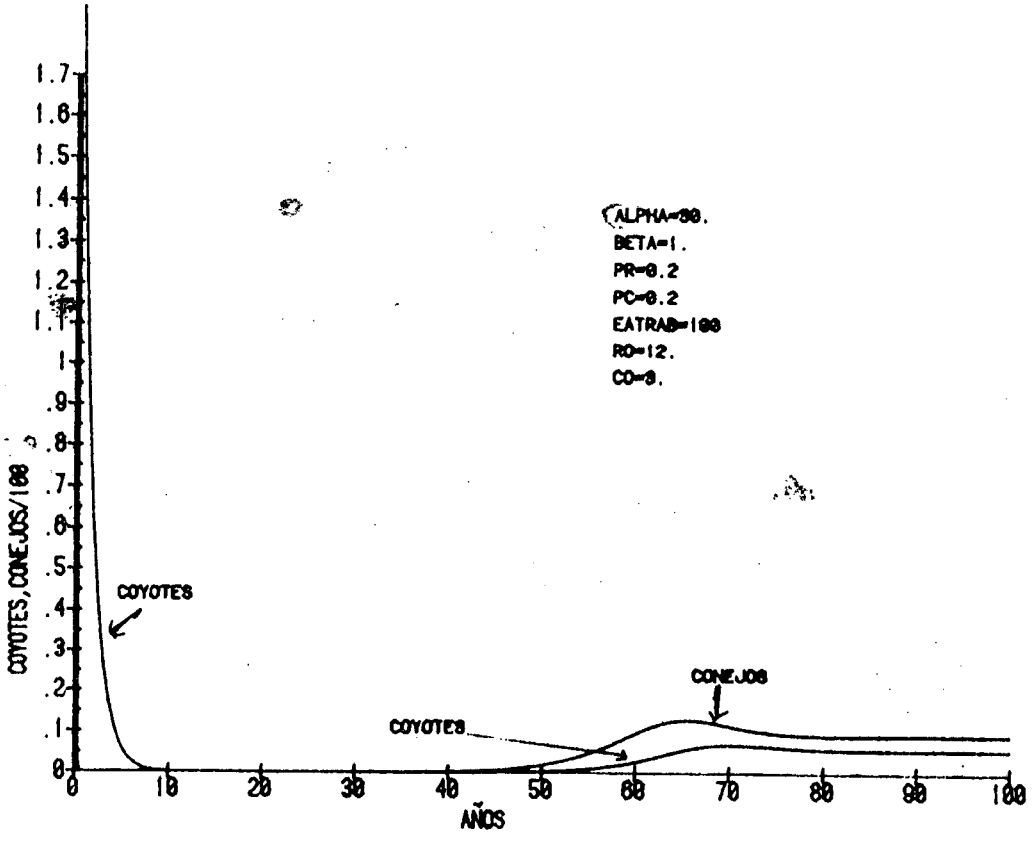

Fig. A.3 La superpoblación inicial de coyotes casi causa la desaparición de ambas poblaciones.

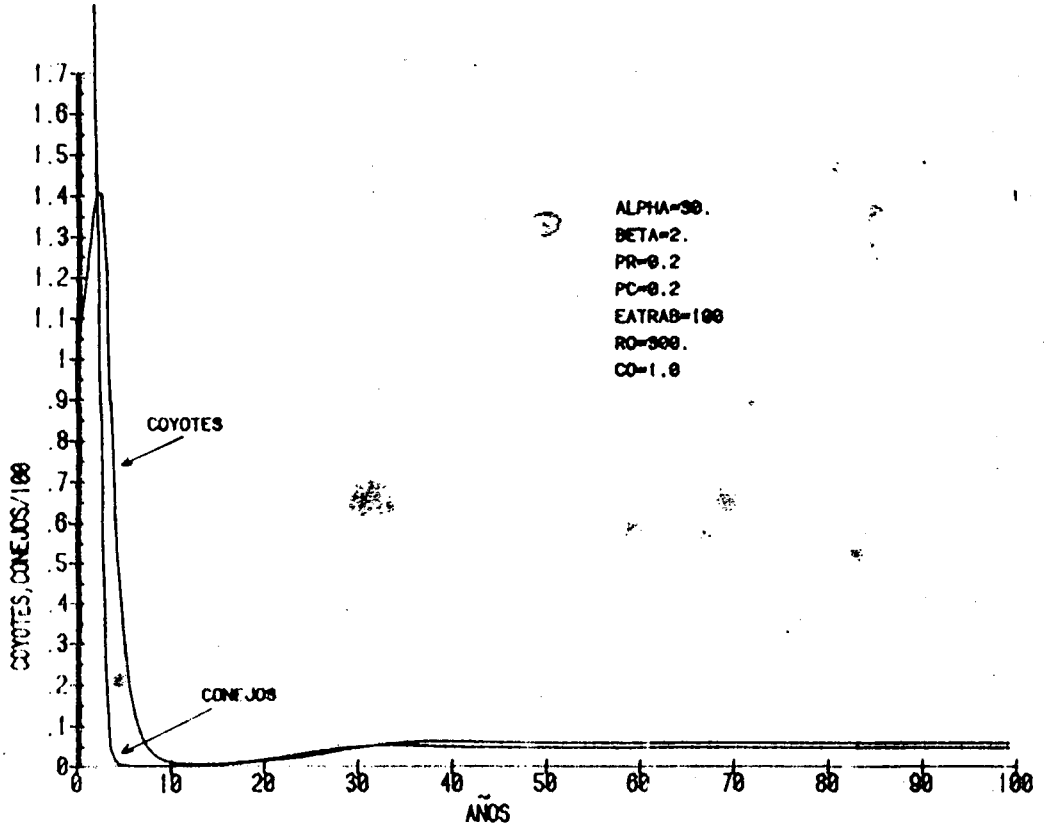

Fig. A.4.a Luego de un cambio inicial rápido se produce un estado estacionario en 36 años.

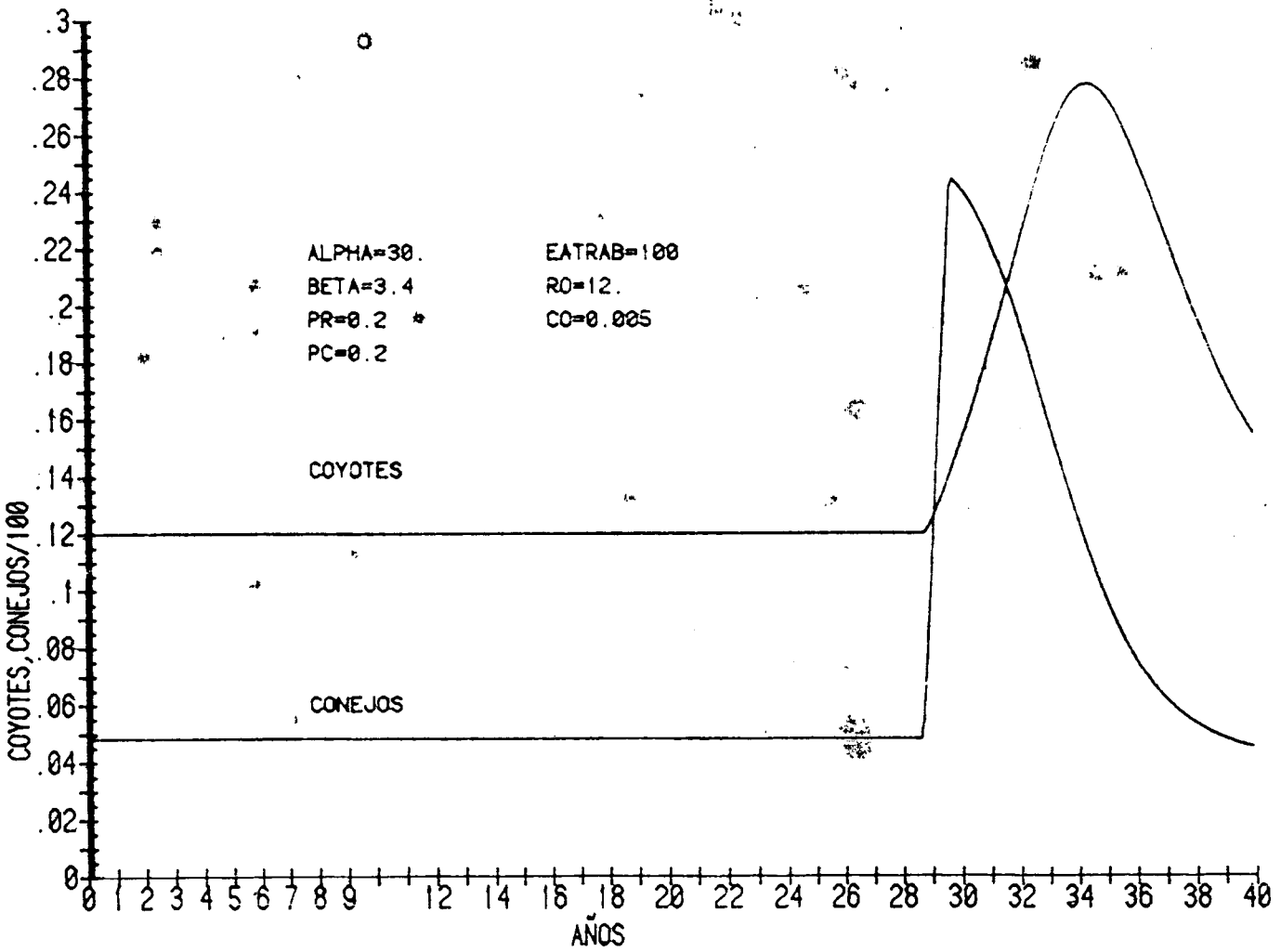

Fig. A.4.b Efecto de una migración de 20 conejos por milla cuadrada en un año a partir del tiempo 28.65 teniendo hasta ese momento poblaciones estacionarias.

Figura A.5 LISTADO DEL PROGRAMA FORTRAN

```
C
C       USO DE LA RUTINA DDIFSU PARA RESOLVER EL PROBLEMA PREDADOR-PRESA
C       PARA LOS COYOTES Y LOS CONEJOS. LAS ECUACIONES DIFERENCIALES DEL
C       MODELO DE INTERACCION DE LAS 2 POBLACIONES SON :
C
C               R'(T) = PR * R(T) - LUNCH(T) * C(T)
C               C'(T) = PC * C(T) - STARVE(T) * C(T)
C
C       R       = DENSIDAD DE CONEJOS = Y(1,1)
C       C       = DENSIDAD DE COYOTES = Y(1,2)
C       PR      = VELOCIDAD DE CRECIMIENTO NATURAL DE LOS CONEJOS
C       PC      = VELOCIDAD DE CRECIMIENTO NATURAL DE LOS COYOTES
C       RO,CO   = DENSIDAD INICIAL DE CONEJOS Y COYOTES
C       EATRAB  = CANTIDAD DE CONEJOS QUE UN COYOTE NORMALMENTE COME
C                 POR AÑO ( SUPONIENDO QUE HAY INFINITOS CONEJOS )
C       ALPHA   = DENSIDAD DE LOS CONEJOS A PARTIR DE LA CUAL LOS COYOTES
C                 COMIENZAN A COMER MENOS CONEJOS POR AÑO
C       BETA    = FACTOR DE INANICION DE LOS COYOTES
C
C
C   LUNCH(T)    = NÚMERO DE CONEJOS QUE UN COYOTE REALMENTE COME EN
C                 EN UN AÑO
C   STARVE(T)   = PROPORCION DE COYOTES QUE SE MUEREN DE HAMBRE POR
C                 AÑO ANTE LA ESCASEZ DE CONEJOS
C
C         LUNCH(T)  = EATRAB * MIN ( 1 , R(T)/ALPHA )
C         STARVE(T) = EXP ( -BETA * S(T) )
C
C                   S(T) = R(T) / (EATRAB * C(T))
C
C
C
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION Y(8,2) , SAVE(24) , YMAX(2) , ERROR(2)
      REAL*4 PW(4)
      COMMON / MODEL / PR,PC,EATRAB,BETA,ALPHA
C
C
C                 PARAMETROS DEL PROBLEMA
C
C
      ALPHA  = 30.
      BETA   = 3.4
      PR     = .2
      PC     = .2
      EATRAB = 110.
      RO     = 12.
      CO     = 0.4
```

Figura A.5 (continuación)

```
C        INICIALIZACION DE LOS PARAMETROS DE LA SUBRUTINA DDIFSU
C
C
      N        = 2
      T        = 0.
      Y(1,1)   = RO
      Y(1,2)   = CO
      H        = 0.25
      HMIN     = 0.0001
      HMAX     = 0.5
      EPS      = 1.D-4
      MF       = 1
      YMAX(1)  = 1.
      YMAX(2)  = 1.
      JSTART   = 0
      MAXDER   = 1
C
C
C              IMPRESION IDENTIFICACION DEL PROBLEMA
C
C
      WRITE ( 7 , 10 ) PR,PC,BETA,ALPHA,EATRAB,RO,CO,EPS,YEARS,TSTEP
 10   FORMAT(//,10X,'PROBLEMA DE LOS COYOTES Y LOS CONEJOS',/
     *  10X,'PARAMETROS DEL MODELO :'//
     *    15X,'PR = ',F8.4/15X,'PC = ',F8.4/15X,'BETA = ',F8.4/
     *    15X,'ALPHA = ',F8.4/15X,'EATRAB = ',F8.4/15X,'RO = ',F8.4/
     *  15X,'CO = ',F8.4/15X,'EPS = ',E7.1/15X,'YEARS =',F4.0/
     *    15X,'TSTEP = ',F6.4)
C
      R = RO/100
      WRITE(7,15) T,R,CO
 15   FORMAT(///2X,'T',16X,'R(T)',17X,'C(T)'//1X,F6.2,2X,F10.5,2X,F10.5)
C
C
C              LOOP DE INTEGRACION
C
C
      TANT = T
 100  IF (T.GE.100.) STOP
C
         CALL DDIFSU ( N , T , Y , SAVE , H , HMIN , HMAX , EPS ,
     *                 MF , YMAX , ERROR , KFLAG , JSTART , MAXDER , PW )
C
         JSTART = 1
         R = Y(1,1)/100
         IF (DABS(T-TANT).LT.0.5) GO TO 100
         WRITE(7,20) T,R,Y(1,2)
         TANT = T
 20      FORMAT(1X,F6.2,2X,F10.5,2X,F10.5)
      GO TO 100
      END
```

Figura A.5 (continuación)

```
C
C
C        SUBRUTINAS DEL USUARIO REQUERIDAS POR DDIFSU
C
      SUBROUTINE  DIFFUN (T,Y,DY)
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION Y(8,2) , DY(2)
      REAL*8 LUNCH
      COMMON / MODEL / PR,PC,EATRAB,BETA,ALPHA
      R = Y(1,1)/(EATRAB*Y(1,2))
      LUNCH = EATRAB * DMIN1(1.D0,Y(1,1)/ALPHA)
      DY(1) = PR * Y(1,1) - LUNCH * Y(1,2)
      DY(2) = PC * Y(1,2) - DEXP(-R*BETA) * Y(1,2)
      RETURN
      END
C
C
      SUBROUTINE  PEDERV(T,Y,PW,N)
      IMPLICIT REAL*8 (A-H,O-Z)
      REAL*8 LUNCH , Y(8,2)
      REAL*4 PW(4)
      COMMON / MODEL / PR,PC,EATRAB,BETA,ALPHA
      R = Y(1,1)/(EATRAB*Y(1,2))
      LUNCH = EATRAB * DMIN1(1.D0,Y(1,1)/ALPHA)
      PW(1) = PR
      PW(2) = 0.
      PW(3) = -LUNCH
      PW(4) = PC - DEXP(-R*BETA)
      RETURN
      END
C
      SUBROUTINE DDIFSU(N,T,Y,SAVE,H,HMIN,HMAX,EPS,MF,YMAX,ERROR,
     *KFLAG,JSTART,MAXDER,PW)
C ---INTEGRACION DE SISTEMAS DE ECUACIONES DIFERENCIALES ORDINARIAS---
C                  ----- DE A UN PASO. -----
C                ----- ALGORITMO ACM 407 -----
C  LA RUTINA EMPLEA UN METODO DE MULTIPASOS (PREDICTOR-CORRECTOR),QUE
C  PUEDE SER DE ADAMS O UNO ADECUADO PARA ECUACIONES STIFF,CON CONTROL
C  AUTOMATICO DE LA LONGITUD DEL PASO Y ORDEN ELEGIDO AUTOMATICAMENTE A
C  MEDIDA QUE SE DESARROLLA LA INTEGRACION.
C                    *** PARAMETROS ***
C  N : NUMERO DE ECUACIONES EN EL SISTEMA.
C  T : VARIABLE INDEPENDIENTE.
C  Y : MATRIZ DE DIMENSION 8 POR N CONTENIENDO LAS VARIABLES DEPENDIEN-
C      TES Y SUS DERIVADAS ESCALADAS.
C      Y(J+1,I) CONTIENE LA DERIVADA J-ESIMA DE Y(I) ESCALADA POR
C          H**J/FACTORIAL(J)  DONDE H ES LA LONGITUD DEL PASO CORRIENT
C    SOLO Y(1,I) DEBE SER DADO EN LA PRIMERA ENTRADA.
C  SAVE : ALMACENAMIENTO AUXILIAR DE AL MENOS 12*N UBICACIONES.
C  H : LONGITUD DEL PASO DE INTEGRACION QUE SERA AJUSTADO (O NO)
C      POR LA RUTINA. PARA AHORRAR TIEMPO DE COMPUTACION,EL USUARIO
C      DEBE PROVEER H CLARAMENTE PEQUE"O EN LA PRIMERA LLAMADA.
C      LUEGO SERA AUTOMATICAMENTE INCREMENTADO.
C  HMIN : LONGITUD MINIMA DEL PASO DE INTEGRACION A SER USADA.
```

Figura A.5 (continuación)

```
C  HMAX : LONGITUD MAXIMA CON LA QUE SE INCREMENTARA EL PASO DE INTEGRA
C  EPS : PRECISION REQUERIDA.
C  MF : INDICADOR DEL METODO A SER USADO.
C       0 : SE USA UN METODO PREDICTOR-CORRECTOR DE ADAMS.
C       1 : SE USA UN METODO ADECUADO PARA SISTEMAS STIFF QUE TAMBIEN
C           TRABAJA CON SISTEMAS NO STIFF.
C           EL USUARIO DEBE PROVEER LA  SUBROTINE PEDERV(T,Y,PW,N)  PARA
C           EVALUAR LAS DERIVADAS PARCIALES DE LAS ECUACIONES DIFERENC.
C           CON RESPECTO A LAS Y,ALMACENANDO LA DERIVADA PARCIAL DE LA
C           I-ESIMA ECUACION CON RESPECTO A LA J-ESIMA VARIABLE EN
C            PW(I+N*(J-1)), CON PW DE SIMPLE PRECISION.
C  YMAX : VECTOR DE DIMENSION N QUE CONTIENE EL MAXIMO DE CADA Y.
C         DEBE SER NORMALMENTE PUESTO EN 1 PARA CADA Y ANTES DE LA
C         PRIMERA ENTRADA.
C  ERROR : VECTOR DE DIMENSION N QUE CONTIENE AL RETORNO
C          EL ERROR ESTIMADO POR PASO PARA CADA COMPONENTE.
C  KFLAG : CODIGO DE RETORNO.
C           1 : LA EJECUCION FUE EXITOSA.
C          -1 : EL PASO FUE TOMADO CON H=HMIN PERO NO SE ALCANZO LA
C               PRECISION REQUERIDA.
C          -2 : SE ENCONTRO QUE EL ORDEN MAXIMO ESPECIFICADO
C               ERA DEMASIADO GRANDE.
C          -3 : NO PUDO ALCANZARSE LA CONVERGENCIA DEL CORRECTOR
C               PARA H MAYOR QUE HMIN .
C          -4 : EL ERROR REQUERIDO ES MENOR QUE EL QUE
C               PUEDE SER ALCANZADO POR ESTE PROBLEMA.
C  ISTART : INDICADOR DE ENTRADA.
C           -1 : REPETIR EL ULTIMO PASO CON UN NUEVO H.
C            0 : EJECUTAR EL PRIMER PASO.
C            1 : TOMAR UN NUEVO PASO CONTINUANDO DESDE EL U TIMO.
C           AL RETORNO,CONTIENE EL VALOR NO : ORDEN ALCANZADO POR EL
C                                             METODO U ORDEN DE LA
C                                             MAXIMA DERIVADA.
C  MAXDER : ORDEN DEL METODO A SER USADO.
C           DADO QUE EL ORDEN ES IGUAL A LA MAS ALTA DERIVADA USADA,
C           ESTO RESTRINGE EL ORDEN. DEBE SER MENOR QUE :
C            8   PARA METODOS DE ADAMS.
C            7   PARA METODOS PARA ECUACIONES STIFF.
C  PW : MATRIZ DE DIMENSION N POR N (A SER USADA POR PEDERV)
C     (DEBE SER DEFINIDA EN SIMPLE PRECISION).
C
C   EL USUARIO DEBE PROVEER LA  SUBROUTINE DIFFUN(T,Y,DY)  PARA EVALUAR
C   LAS DERIVADAS DE LAS VARIABLES DEPENDIENTES Y CON RESPECTO A LA
C   VARIABLE INDEPENDIENTE T DEJANDOLAS EN EL VECTOR DY. Y SERA DE
C   DIMENSION 8 POR N Y LOS VALORES DE FUNCION ESTARAN EN Y(1,I) PARA
C   I=1 A N.
```