# UNIVERSIDAD DE LA LAGUNA

## «Integración de problemas Stiff a través de métodos Runge-Kutta»

**Autor: Maria Soledad Pérez Rodríguez**
**Director: Dr. D. Severiano González Pinto**
**Dr. D. Juan Ignacio Montijano Torcal**

**Departamento de Análisis Matemático**

# Agradecimiento

Mi más sincero y profundo agradecimiento a los profesores Severiano González y Juan Ignacio Montijano por su labor de dirección. Gracias por todo el trabajo y tiempo que han empleado en la elaboración de esta memoria. Gracias por su apoyo y por hacerme sentir parte de su equipo desde el primer momento.

Agradezco al profesor Manuel Calvo todas sus valiosas sugerencias a lo largo de todo este tiempo, así como su gran amabilidad y disponibilidad, sobre todo durante mis estancias en su departamento.

Muchas gracias a Inmaculada Higueras y a Teo Roldán por su ayuda y colaboración desinteresada. Sus comentarios sobre las demostraciones del capítulo IV de esta memoria han sido fundamentales para el desarrollo y la claridad de las mismas.

Quiero agradecer también la inestimable ayuda que me han prestado los miembros del Departamento de Análisis Matemático, que desde un principio me han animado y han confiado en mí.

Gracias a mis padres, por enseñarme el valor del trabajo y estimularme para intentar ser cada día mejor persona. Gracias también a toda mi familia por su ayuda incondicional y su apoyo, en particular a Merci.

Y sobre todo, gracias a ti, Miguel, por darle sentido a cada día.

*A Miguel,*
*por todo*

# Prefacio

Esta memoria consta de seis capítulos más un apéndice. El primer capítulo es de tipo introductorio; los capítulos del segundo al quinto (más el apéndice A) se corresponden con la labor de investigación que hemos llevado a cabo, la cual está plasmada también en los cinco artículos que se detallan más abajo. Los capítulos II y III son prácticamente autocontenidos y pueden leerse independientemente, pues se corresponden con dos artículos ya publicados (o aceptados para publicación), mientras que los capítulos IV y V deben leerse de forma conjunta y consecutiva.

En la introducción presentamos al lector que no está especializado en el tema el concepto de sistema diferencial de tipo *stiff*, así como los métodos que se han utilizado para su integración efectiva y el estado actual de la investigación realizada en este campo, a partir de la cual hemos desarrollado nuestro trabajo. Además, hacemos una breve descripción de los resultados presentados en los siguientes capítulos de esta tesis.

Estos capítulos se corresponden con los cinco trabajos siguientes:

**C-II.** *On the numerical solution of stiff IVPs by Lobatto IIIA Runge–Kutta methods*, S. González Pinto, J.I. Montijano Torcal y S. Pérez Rodríguez. Journal of Computational and Applied Mathematics, **82** (1997) 129–148.

**C-III.** *On the implementation of high order implicit Runge–Kutta methods* S. González Pinto, J.I. Montijano Torcal y S. Pérez Rodríguez. Aceptado en Computers and Mathematics with Applications (2000).

**C-IV.** *On the starting algorithms for fully implicit Runge–Kutta methods* S. González Pinto, J.I. Montijano Torcal y S. Pérez Rodríguez. Enviado a BIT (diciembre 1998).

**C-V.** *Stabilized starting algorithms for collocation Runge–Kutta methods*, S. González Pinto, J.I. Montijano Torcal y S. Pérez Rodríguez. Preprint para enviar a Computers and Mathematics with Applications (marzo 2000).

**C-IV,V.** *Algoritmos de arranque para métodos Runge–Kutta implícitos*, S. González Pinto, J.I. Montijano y S. Pérez Rodríguez, Actas del XVI CEDYA/VI CMA, Edit. R. Montenegro, G. Montero y G. Winter, (1999) 1149-1156.

# Índice General

Capítulo I

# Introducción

.

# Capítulo I
# Introducción

En esta memoria abordamos aspectos relativos a la resolución efectiva de Problemas de Valor Inicial (para Ecuaciones Diferenciales Ordinarias) de tipo *stiff* (rígidos), mediante métodos numéricos de tipo Runge–Kutta (RK) implícitos. Los métodos de tipo RK usados principalmente serán las familias de los denominados RK-Gauss [8], Radau IIA y Lobatto IIIA [21]. Es bien conocido que estos métodos pueden alcanzar un alto orden de convergencia y poseen excelentes propiedades de estabilidad lineal y no lineal (ver por ejemplo [19], [11], [37]).

Al aplicar métodos RK implícitos a problemas stiff no lineales, los sistemas de ecuaciones algebraicas que aparecen en cada paso de integración han de resolverse por esquemas iterativos de tipo Newton. La convergencia del proceso iterativo en conjunto depende, de una parte de la "bondad" de las aproximaciones iniciales usadas y de otra, del propio esquema iterativo empleado. En esta memoria estudiamos en los capítulos II y III la construcción de esquemas iterativos optimizados para los métodos RK Lobatto-IIIA de 3, 4 y 5 etapas y los RK Gauss y Radau-IIA de 4 etapas. En los capítulos IV y V analizamos y proponemos diferentes aproximaciones iniciales (que denominaremos inicializadores) para los esquemas iterativos considerados. Las aproximaciones iniciales propuestas sólo usan información del paso de integración anterior y están vinculadas a procesos de interpolación. Además, apenas supondrán coste computacional adicional al esquema de iteración empleado.

Gran parte del trabajo desarrollado en esta memoria está inspirado en los trabajos previos de [2], [5], [9], [10], [17], [18], [28], [29] y [57]. En la mayoría de ellos se proponen esquemas iterativos de tipo Newton para implementar métodos Runge–Kutta implícitos aplicados a problemas de tipo stiff. Además, en [37] y [57] se proponen también aproximaciones iniciales para iniciar los procesos iterativos.

Los sistemas diferenciales de tipo stiff provienen de los más variados campos de las ciencias aplicadas tales como: química cinética, circuitos eléctricos, problemas en derivadas parciales dependientes del tiempo donde las variables espaciales son discretizadas mediante diferencias finitas o elementos finitos, problemas altamente oscilatorios, problemas de perturbaciones singulares, etc. Se puede ver un amplio resumen sobre problemas de este tipo y su conexión con las Ciencias Aplicadas en [1], [19] y [37]. También puede verse en [22], [23] y [37] una amplia gama de problemas de tipo stiff provenientes de las aplicaciones y que se suelen usar como problemas tests para probar los métodos numéricos.

La definición matemática rigurosa de sistema diferencial stiff no está hoy todavía suficientemente clara, aunque históricamente se han hecho bastantes esfuerzos para separar la clase stiff de la clase no–stiff. Una amplia discusión sobre ambas clases de problemas, que suele ser una referencia ineludible, es la presentada en el texto de Lambert [47]. Podemos decir que la clase stiff es de alguna manera "difusa", ya que se pasa de la clase no–stiff a la clase stiff de forma gradual y no hay una frontera clara que separe ambas clases. Así, es corriente que los especialistas digan que un problema es "midly" stiff. Existe una opinión generalizada entre los analistas

numéricos con amplia experiencia computacional (ver [37, Cap. I]) de que los problemas stiff son aquellos donde los integradores numéricos basados en métodos explícitos no funcionan satisfactoriamente. Además, hay problemas de valor inicial que son de tipo stiff en una parte del intervalo de integración siendo no–stiff en otra parte del mismo, intercalándose así zonas no–stiff con zonas stiff. Un ejemplo de este tipo de problemas es el oscilador no lineal de van der Pol, ver e.g. [37, pág.4, 21-24].

Para dar una idea simple de lo que es la clase no–stiff, podemos decir que la componen aquellos problemas de valor inicial (PVI)

$$y'(t) = f(t, y(t)), \quad y(0) = y_0, \quad t \in [0, T], \quad y,\ f \in \mathbb{R}^m, \tag{1.0.1}$$

con solución única $y(t)$ suficientemente suave en $[0, T]$, donde la constante de Lipschitz de $f(t, y)$ respecto de $y$ en un entorno tubular $\{(t, y);\ t \in [0, T],\ |y - y(t)| \leq \delta\}$ de la solución $y(t)$, digamos $L_f$, por la longitud del intervalo de integración $T$ es de tamaño moderado, i.e.,

$$L_f T = \mathcal{O}(1).$$

En cambio, la clase stiff se caracteriza porque justamente

$$L_f T >> 1,$$

y además las curvas integrales $y(t; 0, z_0)$ [1] no se separan excesivamente de $y(t; 0, y_0)$ cuando $z_0$ está próximo a $y_0$, es decir, para dos curvas integrales con valores iniciales respectivos $y_0$ y $z_0$ debe verificarse

$$|y(t_2; t_1, y_0) - y(t_2; t_1, z_0)| \leq K|y_0 - z_0|, \quad \forall\ 0 \leq t_1 < t_2 \leq T, \quad K = \mathcal{O}(1). \tag{1.0.2}$$

Las expresiones anteriores justifican por qué se dice que la clase stiff resulta algo "difusa", pues los términos $\mathcal{O}(1)$ y $>> 1$ son algo ambiguos.

**Nota I.0.1** *A lo largo de esta memoria siempre denotaremos las normas por $|\cdot|$ indistintamente del espacio normado que estemos considerando.*

Haciendo una breve reseña histórica podemos decir que los problemas stiff comenzaron a estudiarse sobre modelos lineales (ver e.g. [19, Cap.I])

$$y'(t) = Jy(t)$$

donde $J$ es una matriz cuadrada constante de orden $m$ con coeficientes constantes, imponiendo condiciones sobre sus autovalores $\lambda_i$. Así era habitual exigir

(1) Existe algún $\lambda_i$ con $Re(\lambda_i) << 0$.

(2) Existe algún autovalor verificando $|\lambda_i| = \mathcal{O}(1)$.

(3) No existen autovalores con parte real positiva grande.

(4) $|Im(\lambda_i)| >> 1$ implica $Re(\lambda_i) << 0$.

[1] $y(t; t_0, z_0)$ significa la curva integral del sistema con condición inicial $y(t_0) = z_0$.

Obsérvese que (1), (2) y (3) implican la superposición de componentes estables y superestables en las curvas integrales, y que (4) se exige a fin de evitar problemas fuertemente oscilatorios.

Es interesante observar que si en el sistema diferencial lineal anterior la dimensión $m$ se deja crecer de forma arbitraria, entonces las cuatro condiciones anteriores no garantizan para las curvas integrales un comportamiento de tipo (1.0.2), ni aun en el caso en que todos los autovalores tengan parte real negativa.

Tras varios intentos de definición de la clase stiff para problemas lineales y no lineales, podemos decir que hoy se acepta la siguiente definición, la cual se apoya en los conceptos de constantes de Lipschitz por un lado antes que en condiciones de Lipschitz, esto es, se suele exigir a $f$ la siguiente condición,

$$\langle f(t,y) - f(t,z), y - z\rangle \leq \nu |y-z|^2, \quad t \in [0,T] \quad y, z \in I\!R^m, \quad \nu = \mathcal{O}(1). \tag{1.0.3}$$

Aquí $\langle \cdot, \cdot \rangle$ denota un producto escalar cualquiera en $I\!R^m$, y la norma considerada es la euclídea asociada a dicho producto escalar. Muchas veces no se exige la condición anterior sobre todo $I\!R^m$ sino sobre un subconjunto apropiado de éste. Además, si $\nu = 0$ el sistema diferencial se suele denominar *contractivo*.

Esto abarca una clase amplia de problemas stiff, pero hay problemas que no cumplen esta condición y que podrían considerarse stiff (por ejemplo el oscilador de van der Pol [37]).

Es fácil probar, suponiendo que $f$ es continua y verifica (1.0.3), que (ver [19, Cap.I])

$$|y(t;t_0,y_0) - y(t;t_0,z_0)| \leq \exp\left(\nu(t-t_0)\right)|y_0 - z_0|, \quad \forall t \geq t_0, \quad \forall y_0, z_0. \tag{1.0.4}$$

La clase stiff podría haber sido definida usando otro tipo de normas, no sólo las que provienen de productos escalares. En este caso se hace imprescindible el concepto de norma logarítmica de una matriz $J$ (que no es propiamente una norma) que puede expresarse mediante

$$\mu[J] := \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \log\left(|\exp\left(\varepsilon J\right)|\right).$$

La norma logarítmica de matrices fue introducida independientemente por G. Dahlquist (1959) y por Lozinskij (1958). Un buen resumen sobre sus propiedades más relevantes están recogidas en [19, Cap. I]. Por otra parte, su conexión con los sistemas diferenciales cuyas curvas integrales verifican (1.0.4), puede verse en [19] y [20].

## I.1 Métodos para la integración de problemas stiff

Los métodos lineales multipaso han sido quizás los métodos más populares y eficientes para la integración de problemas de valor inicial de tipo stiff y también para Ecuaciones Diferenciales Algebraicas, especialmente a raíz del trabajo de Gear [26] que propuso las fórmulas BDF (Backward Differentiation Formulae) en el paquete DIFSUB. Se ha corroborado posteriormente que son unas de las mejores candidatas para la integración de este tipo de problemas en precisión baja y media (ver por ejemplo [37]), así como los paquetes VODE de Brown et al. [3] y LSODE de Hindsmarsh [40]. Sin embargo, estas fórmulas pierden estabilidad lineal absoluta a medida que aumenta el número de pasos $k$ y que el tamaño de paso varía a lo largo del intervalo de integración (ver por ejemplo [37]), no siendo siquiera 0–estables para $k > 6$ en el caso óptimo en que el tamaño de paso $h$ es fijo. Además sólo son L–estables para los casos $k = 1, 2$. También tienen propiedades más pobres de estabilidad lineal y no lineal que ciertas clases de métodos RK implícitos de alto orden tales como RK Gauss, Radau IA, Radau IIA, Lobatto IIIA, Lobatto

IIIC y otros muchos de tipo DIRK (Diagonally Implicit Runge–Kutta methods) y SIRK (Singly Implicit Runge–Kutta methods, ver [4], [5]).

Por otra parte, el inconveniente que presentan los métodos RK anteriormente reseñados a la hora de una implementación eficiente, es la solución de las ecuaciones algebraicas que aparecen en cada paso. Usualmente para un PVI de dimensión $m$ hay que resolver un sistema algebraico (no lineal en general) de $ms$ ecuaciones ($s$ es el número de etapas del método RK implícito considerado) que es además "fully-implicit" para métodos tales como Gauss, Radau, etc. En cambio, en el caso de los métodos BDF el sistema algebraico que aparece (en cada paso) es de la misma dimensión que la del PVI considerado, resultando por tanto los métodos RK más caros computacionalmente en cada paso. Así, en palabras de K. Burrage et al. [37, pág. 118] "Although Runge–Kutta methods present an attractive alternative, specially for stiff problems, ... it is generally believed that they will never be competitive with multistep methods". Es evidente que si se consiguieran implementaciones "baratas" de los métodos RK, podrían competir o incluso superar a los métodos BDF (o variantes) ya que, como hemos mencionado anteriormente, ciertas familias de RK implícitos poseen mejores propiedades de estabilidad y convergencia para cualquier orden por grande que sea éste. Además, aunque los BDF son más baratos computacionalmente por paso, requieren generalmente muchos más pasos que los RK para completar la integración sobre un intervalo $[0, T]$ y presentan el inconveniente de un alto "overhead", es decir, un alto coste adicional debido al manejo de la información de varios pasos y a la mayor complejidad en su uso con paso variable, en comparación con los métodos de un paso.

## I.2 Implementación de los métodos RK para problemas stiff

Un método Runge–Kutta de $s$ etapas avanza la solución numérica un tamaño de paso $h$, desde el punto $(t_n, y_n)$ hasta el punto $(t_{n+1} = t_n + h, y_{n+1})$, mediante la fórmula

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i f(t_n + c_i h, Y_i) \tag{1.2.1}$$

donde las etapas intermedias $Y_i$ se calculan de

$$Y_i = y_n + h \sum_{j=1}^{s} a_{ij} f(t_n + c_j h, Y_j), \quad 1 \le i \le s. \tag{1.2.2}$$

A la matriz $A = (a_{ij})$ y al vector $b = (b_i)$ se le denominan coeficientes del método. El vector $c$ verifica $Ae = c$, donde $e = (1, \ldots, 1) \in \mathbb{R}^s$. Además el método se suele denotar por RK $(A, b)$.

Usando el producto de Kronecker para matrices $A \otimes B := (a_{ij}B)$, y definiendo los vectores

$$Y^T := (Y_1^T, \ldots, Y_s^T), \quad F^T(Y) := (f^T(t_n + c_1 h, Y_1), \ldots, f^T(t_n + c_s h, Y_s)),$$

las ecuaciones (1.2.1) y (1.2.2) pueden escribirse de forma más compacta mediante (denotando $I$ a la matriz identidad de orden $m$),

$$y_{n+1} = y_n + h(b^T \otimes I)F(Y), \quad Y = e \otimes y_n + h(A \otimes I)F(Y).$$

Los métodos iterativos usados principalmente para la solución de (1.2.2) son de tipo Newton, ya que es bien conocido que la iteración funcional no converge cuando el problema considerado

es stiff. Entre los métodos de tipo Newton el más popular ha sido la *iteración de Newton Simplificada* (o *Modificada*) introducida en el año 1973 por Chipman [16] y que puede escribirse mediante las ecuaciones

$$\begin{cases} (I - h(A \otimes J))\Delta^k = D^k \equiv e \otimes y_n - Y^k + h(A \otimes I)F(Y^k), \\ \\ Y^{k+1} = Y^k + \Delta^k, \ \ k = 0, 1, ..., \end{cases} \tag{1.2.3}$$

donde $J \doteq \partial f / \partial y(t_n, y_n)$ e $Y^k$ es la aproximación $k-$ésima a la solución $Y$ de (1.2.2).

Este tipo de iteración fue retomada posteriormente por Butcher [9] y Bickart [2] de forma independiente. Estos autores propusieron hacer ciertas transformaciones usando factorizaciones de Jordan (o similares) de la matriz $A$, a fin de disminuir el costo algebraico necesario para la solución de los sistemas lineales que aparecen en cada iteración, aunque esto conlleva el uso de aritmética compleja en algunos casos. Esta alternativa fue considerada también por Hairer y Wanner en los años 90 para la implementación del Radau IIA de 3 etapas en su código RADAU5 [37].

En el caso en que la matriz $A$ sea triangular inferior (métodos DIRK) el sistema lineal anterior de dimensión $ms$ se puede transformar en $s$ sistemas lineales de dimensión $m$ que se van resolviendo sucesivamente en cada iteración. Por otra parte, si la matriz $A$ no es triangular inferior sino que posee autovalor único múltiple (métodos SIRK) se pueden hacer ciertas transformaciones algebraicas de modo que se reduce bastante el costo computacional por iteración (es ligeramente superior al caso de los métodos DIRK). Para el caso SIRK el esquema Newton Simplificado se puede simplificar teniendo en cuenta que

$$A = \gamma S P S^{-1}, \ \ \gamma \neq 0, \tag{1.2.4}$$

donde $S$ y $P$ son matrices regulares, siendo $P$ una matriz triangular inferior con "1" en la diagonal principal. Por tanto, después de ciertas manipulaciones, ver por ejemplo [17], se transforma el esquema anterior en:

$$\begin{cases} (I - h\gamma(I \otimes J)E^k = ((I - L)S^{-1} \otimes I)D^k + (L \otimes I)E^k \\ \\ Y^{k+1} = Y^k + (S \otimes I)E^k, \ \ k = 0, 1, ..., \end{cases} \tag{1.2.5}$$

siendo $J \doteq \partial f / \partial y(t_n, y_n)$, donde la matriz $L = I - P^{-1}$ es triangular inferior estricta. Obsérvese que si $A$ es triangular inferior con autovalor único $\gamma$, entonces se puede tomar $S = I$ y estaríamos en el caso de los métodos DIRK, los cuales evitan ciertas transformaciones matriciales que aparecen en los métodos SIRK.

El código STRIDE de Burrage et al. [4, 5] se basa precisamente en una familia de métodos SIRK encajados de orden variable, con buenas propiedades de estabilidad, siendo uno de los mejores intentos de implementar fórmulas de tipo Runge–Kutta implícitas para problemas stiff. No obstante, parece que este código es menos eficiente en general que otros que implementan fórmulas de tipo BDF (ver por ejemplo los experimentos numéricos realizados en [37]).

A la hora de implementar métodos RK de colocación de alto orden (tales como Gauss, Radau IIA, etc.) aparece la dificultad de que sus matrices de coeficientes $A$ poseen un espectro multipuntual. Es más, generalmente tienen $s$ autovalores distintos incluyendo pares de autovalores complejos conjugados. Como esto lleva a que el coste computacional de la iteración de Newton Simplificada sea muy elevado, han surgido otras perspectivas en la literatura que intentan hacer más efectivo la resolución de (1.2.2). Entre las más relevantes podemos destacar los trabajos de Frank y Ueberhuber [25] que fueron pioneros en este campo a través de la técnica de "corrección

por defecto", usando métodos simples de un paso (como Euler implícito) en cada iteración para resolver las ecuaciones algebraicas. Más adelante, Butcher [10] en 1979 generalizó esta técnica. Como fruto de experiencias e investigaciones anteriores Cooper y Butcher [17] propusieron en 1983 el siguiente tipo de esquema iterativo:

$$\begin{cases} (I - h\gamma(I \otimes J)E^k = (BS^{-1} \otimes I)D^k + (L \otimes I)E^k \\ \\ Y^{k+1} = Y^k + (S \otimes I)E^k, \;\; k = 0, 1, ..., \end{cases} \tag{1.2.6}$$

con $J \doteq \partial f/\partial y(t_n, y_n)$, donde la constante $\gamma > 0$, las matrices regulares constantes $B$ y $S$ y la matriz triangular inferior estricta $L$ son los parámetros del esquema, que deben determinarse bajo algún criterio de optimización.

El proceso de optimización seguido por estos autores fue el de exigir que la iteración tuviera una velocidad de convergencia alta sobre problemas lineales de la forma

$$y' = \lambda y, \quad Re(\lambda) \leq 0,$$

independientemente del valor de $\lambda$. Además, en [17] se dan esquemas iterativos optimizados de este tipo para los RK-Gauss de 2, 3 y 4 etapas. También hemos de reseñar que Gladwell y Thomas [27] recomiendan como esquema más eficiente para el RK-Gauss de 2 etapas, el propuesto en [17].

Por otra parte, en [28], [29] se estudia para los métodos RK Gauss de 2 y 3 etapas y el Radau IIA de 3 etapas, un esquema iterativo del tipo (1.2.6) donde ahora $B = I - L$. Se prueba también allí que esta elección presenta ciertas ventajas sobre la iteración más general (1.2.6) cuando $B \neq I - L$. Este tipo de esquema iterativo será la base de la investigación desarrollada en esta memoria en los capítulos II y III, donde se considerará la optimización de estos esquemas iterativos para los métodos Lobatto IIIA de 2, 3 y 4 etapas implícitas así como para los RK-Gauss y Radau IIA de 4 etapas.

En relación con las aproximaciones iniciales para iniciar los procesos iterativos, hemos de decir que para el caso no–stiff la investigación ha sido muy abundante, ver [52], [44], [45], [46]. Sin embargo, para el caso stiff ha sido bastante más reducida, quizás porque la teoría clásica del orden basada en B-series, la cual es válida para el caso no–stiff, deja de serlo para el caso stiff y se hace ahora necesario buscar algún marco nuevo para el estudio del orden de las aproximaciones iniciales. Entre los trabajos donde se proponen aproximaciones iniciales sin costo computacional adicional y que usan información de pasos anteriores, merece destacar [37, Cap. IV.8], [57] y más recientemente para ecuaciones diferenciales algebraicas [53]. La investigación realizada en esta memoria en los capítulos IV y V hace un extenso estudio del tipo de aproximaciones iniciales que se habían propuesto anteriormente y se añaden algunas nuevas que no habían sido consideradas. Además, se suministra un marco general para el estudio de la estabilidad y de los distintos órdenes de convergencia de las aproximaciones iniciales.

Capítulo **II**

# Implementación de los métodos RK–Lobatto IIIA

.

# Capítulo II
# Implementación de los métodos RK–Lobatto IIIA

## II.1 Introducción

Consideremos la solución numérica de problemas de valor inicial (PVI) de tipo stiff

$$y'(t) = f(t, y(t)), \quad t \in [t_0, t_f], \qquad y(t_0) = y_0 \in \mathbb{R}^m, \tag{2.1.1}$$

mediante métodos Runge–Kutta de tipo Lobatto IIIA (ver e.g. [37, Cap. IV.5]), esto es, dada una aproximación $y_n \doteq y(t_n)$, avanzamos la solución numérica un paso de tamaño $h$, mediante las fórmulas

$$Y_i = y_n + h \sum_{j=1}^{s} a_{ij} f(t_n + c_j h, Y_j), \quad i = 1, ..., s, \tag{2.1.2}$$

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i f(t_n + c_i h, Y_i). \tag{2.1.3}$$

Aquí $\{c_i, i = 1, \ldots, s\}$ son los nodos de Lobatto sobre el intervalo $[0, 1]$, que coinciden con las raíces del siguiente polinomio de grado $s$ [37, pág. 72]

$$p_s(x) = \frac{d^{s-2}}{dx^{s-2}}(x(x-1))^{s-1}. \tag{2.1.4}$$

Además, la matriz $A = (a_{ij})_{i,j=1}^{s}$ y el vector $b = (b_i)_{i=1}^{s}$ se calculan de las condiciones de colocación

$$Ac^{j-1} = \frac{1}{j} c^j, \qquad b^T c^{j-1} = \frac{1}{j}, \qquad 1 \le j \le s. \tag{2.1.5}$$

En las expresiones anteriores y a partir de ahora, entenderemos las potencias de un vector como el vector resultante de tomar las potencias correspondientes sobre cada una de sus componentes.

Es inmediato comprobar que las condiciones (2.1.4)-(2.1.5) implican

$$c_1 = 0, \quad c_s = 1, \quad a_{1j} = 0, \quad a_{sj} = b_j, \quad 1 \le j \le s.$$

Es bien conocido que para integrar problemas stiff un requisito mínimo que se suele pedir a los métodos numéricos es la A-estabilidad. Esto, junto con la consistencia, justifica la convergencia de los métodos sobre problemas stiff lineales. También muchos autores recomiendan para la integración de problemas lineales la L-estabilidad, esto es, que además de la A-estabilidad el método debe cumplir

$$R(\infty) = 0, \quad \text{donde} \quad R(z) = 1 + z b^T (I - zA)^{-1} e \quad \text{(función de estabilidad lineal)}.$$

Con esta propiedad el método amortigua mejor las componentes superestables de la solución del PVI [37, Cap. IV.1]. Si los problemas son no lineales se suele exigir además la estabilidad algebraica y la estabilidad diagonal, ver [19, Cap. 4 y 7].

Los métodos Lobatto IIIA son A–estables, pero no son ni L–estables ni algebraicamente estables (ni diagonalmente estables), por lo que sus propiedades de estabilidad y convergencia son en principio inferiores a otras familias como los Gauss, Radau IA-IIA, Lobatto IIIC [19], [37]. Teniendo esto en cuenta nos podríamos preguntar qué interés tienen este tipo de métodos en la integración de problemas stiff. Como veremos, hay otras características teóricas y prácticas que los hacen apropiados para ello.

En primer lugar, ya que la primera etapa de los Lobatto IIIA es explícita, esto es, $Y_1 = y_n$, el sistema implícito de $sm$ ecuaciones algebraicas (2.1.2) se reduce a un sistema de $(s-1)m$ ecuaciones implícitas, por lo que estos métodos tendrán un coste computacional equivalente a un RK "completamente implícito" de $s-1$ etapas, como los Gauss, Radau IIA o Lobatto IIIC. Esto significa que desde un punto de vista computacional, el Lobatto IIIA de $s$ etapas debe compararse con el Gauss, Radau IIA o Lobatto IIIC de $s-1$ etapas.

En segundo lugar, obsérvese que los métodos Lobatto IIIA son stiffly accurate. Esto significa que cuando se aplican al problema test escalar de Prothero y Robinson [55]

$$y' = \lambda(y - \phi(t)) + \phi'(t), \quad Re(\lambda) \leq 0,$$

donde $\phi(t)$ es suficientemente suave con derivadas independientes de la stiffness del problema, es decir, $\phi^{(j)}(t) = \mathcal{O}(1)$, $j \geq 0$, el error local para la fórmula de $s$ etapas se comporta como (ver [37])

$$z^{-1}h^{s+1}, \quad \text{si} \;\; z = h\lambda \to \infty.$$

Por tanto, da un resultado asintóticamente exacto para $z \to \infty$. También los errores globales para tamaños de paso variables se comportan como

$$z^{-1}h^{s}, \quad h = \max_j h_j.$$

Por supuesto, el mismo comportamiento se da en el Lobatto IIIC y el Radau IIA de $s$ etapas, aunque el error global de este último tiene un orden mayor que los otros (ver [37, pág. 242]). Luego en la integración de problemas stiff lineales los métodos Lobatto IIIA se encuentran entre los mejores candidatos. Además, como puede verse en [37], los Lobatto IIIA están entre las fórmulas con más alto orden de convergencia (con respecto al número de etapas) para problemas de perturbaciones singulares y ecuaciones diferenciales algebraicas de índices 1 y 2.

En tercer lugar, aunque los Lobatto IIIA no son B–estables y por consiguiente no son B–convergentes a paso variable en general [37, pág. 232], Calvo et al. [13] han probado que son estables y convergentes sobre ciertas clases de problemas stiff semilineales con coeficientes variables. Los mismos autores también han probado en [14] que son estables y convergentes de orden $s$ para cierta clase de problemas no lineales stiff, que verifican cierta acotación sobre la variación relativa de la matriz jacobiana de $f$ respecto de $y$. También han demostrado que esta propiedad se verifica para muchos problemas stiff no lineales que aparecen en las aplicaciones.

Por otra parte, los métodos L–estables pueden presentar algunos inconvenientes en comparación con los métodos A–estables simétricos (como los Lobatto IIIA), es decir, con aquellos que verifican

$$|R(iy)| = 1, \quad \forall y \in I\!R.$$

Así, cuando el problema tiene una onda no disipativa en la solución estacionaria, las oscilaciones de la solución exacta pueden ser amortiguadas por un método L–estable y la frecuencia de

oscilación puede ser modificada, con lo que la solución numérica puede llevarnos a conclusiones erróneas sobre el comportamiento cualitativo de la solución. Una amplia discusión sobre métodos L–estables y métodos A–estables (no L–estables) puede verse en [47, pág. 224-231].

De lo expuesto anteriormente se deduce que los métodos Lobatto IIIA pueden ser buenos candidatos para la integración de problemas stiff, problemas de perturbaciones singulares y ecuaciones diferenciales algebraicas, siempre que se consiga una buena implementación del método. Por tanto queda justificada la investigación de esquemas iterativos apropiados para la resolución de sus ecuaciones de etapa. Este será el objetivo fundamental de este capítulo, donde se obtendrán esquemas eficientes para las fórmulas de 3 y 4 etapas (los resultados aquí expuestos están publicados en [30]).

El resto del capítulo queda organizado de la siguiente manera. En la sección II.2 introduciremos el modelo de esquema iterativo a usar, el cual se basa en uno propuesto en investigaciones previas realizadas en [29]. También se lleva a cabo un estudio de las propiedades de estabilidad lineal de las iteraciones sucesivas, suministrando éste condiciones adecuadas para la optimización del esquema iterativo. En la sección II.3 obtenemos esquemas iterativos optimizados para los métodos Lobatto IIIA de 3 y 4 etapas, los cuales están basados en los requisitos de estabilidad de la sección anterior. Finalmente, en la sección II.4 presentamos algunos experimentos numéricos con el objetivo de probar que el Lobatto IIIA es un integrador eficiente para problemas stiff, así como para contrastar la eficiencia del nuevo esquema iterativo propuesto frente a la *iteración de Newton Simplificada*, que es la tradicionalmente usada al implementar los métodos implícitos.

## II.2 Una clase de esquemas iterativos para las fórmulas Lobatto IIIA

Consideremos el método Lobatto IIIA de $s$ etapas con matriz de coeficientes $A = (a_{ij})_{i=1}^{s}$. Como se vio en la sección anterior puede escribirse en la forma:

$$\begin{gathered} A = \begin{pmatrix} 0 & \mathbf{0}^T \\ w & \bar{A} \end{pmatrix}, \qquad w = (a_{21}, \ldots, a_{s1})^T \in I\!R^{s-1}, \\ \mathbf{0}^T = (0, 0, ..., 0) \in I\!R^{s-1}, \qquad \bar{A} = (a_{ij})_{i,j=2}^{s}. \end{gathered} \tag{2.2.1}$$

donde la submatriz $\bar{A}$ de dimensión $(s-1) \times (s-1)$ es no singular.

Introduciendo los vectores

$$Y = \begin{pmatrix} Y_2 \\ \vdots \\ Y_s \end{pmatrix}, \qquad F(Y) = \begin{pmatrix} f(t_n + c_2 h, Y_2) \\ \vdots \\ f(t_n + c_s h, Y_s) \end{pmatrix},$$

las ecuaciones (2.1.2)–(2.1.3) pueden reescribirse en forma matricial,

$$Y_1 = y_n, \tag{2.2.2}$$

$$Y = \bar{e} \otimes y_n + h(w \otimes f(t_n, y_n)) + h(\bar{A} \otimes I)F(Y), \tag{2.2.3}$$

$$y_{n+1} = Y_s, \tag{2.2.4}$$

donde $\bar{e} = (1, ..., 1)^T \in I\!R^{s-1}$ y $\otimes$ denota el producto de Kronecker de matrices. Obsérvese que, en lo que sigue, las dimensiones de las matrices identidad $I$ que aparecen vendrán dadas por el contexto.

En un método de tipo Newton (1.2.3) cada paso de la iteración conlleva la solución de un sistema lineal de dimensión $(s-1)m$, con matriz de coeficientes $(I - h\bar{A} \otimes J)$, donde $J = \partial f/\partial y$ está evaluado en algún punto previamente computado. Para reducir el coste computacional de factorizaciones LU, se puede introducir una transformación de semejanza para la submatriz $\bar{A}$ siguiendo las ideas de [37, Cap. IV.8]. Este esquema iterativo aún resulta costoso computacionalmente para PVIs de dimensión elevada y además involucra el uso de aritmética compleja. Por este motivo se han propuesto otros tipos de iteración [2], [9], [10], [17], [18], [28], [29], que sólo conllevan en cada paso de la iteración y para Lobatto IIIA, la solución de $s-1$ sistemas lineales de dimensión $m$ y con la misma matriz de coeficientes.

Aquí, siguiendo las ideas dadas en [29], consideraremos esquemas iterativos llamados Single–Newton en los que las iteraciones se calculan mediante

$$\begin{aligned} [I - h\gamma(I \otimes J)]\, E^k &= ((I-L)S^{-1} \otimes I)D^{k-1} + (L \otimes I)E^k, \\ Y^k &= Y^{k-1} + (S \otimes I)E^k, \quad k = 1, 2, ..., \end{aligned} \tag{2.2.5}$$

donde

$$D^{k-1} := \bar{e} \otimes y_n + h(w \otimes f(t_n, y_n)) - Y^{k-1} + h(\bar{A} \otimes I)F(Y^{k-1}).$$

La constante real $\gamma > 0$, la matriz regular $S$ y la matriz triangular inferior estricta $L$ constituyen los parámetros del esquema iterativo y deberán elegirse de manera apropiada. La matrices $L$ y $S$ tienen dimensión $(s-1) \times (s-1)$.

Obsérvese que si el esquema iterativo converge, entonces lo hace a la solución de (2.2.3), la cual se supone que es única al menos localmente, esto es, para valores positivos de $h$ de tamaño moderado.

La optimización del esquema se hace sobre problemas lineales de tipo stiff. Es por ello que se hace necesario un estudio de convergencia del mismo sobre el test $y' = \alpha y$, $Re(\alpha) \leq 0$. Denotando por $z = h\alpha$, es fácil probar que los errores en dos pasos consecutivos de la iteración están relacionados por

$$Y - Y^k = M(z)(Y - Y^{k-1}) = M(z)^k(Y - Y^0) \tag{2.2.6}$$

con

$$M(z) = z(I - zT)^{-1}(\bar{A} - T) \tag{2.2.7}$$

y

$$T = \gamma S(I-L)^{-1}S^{-1}, \quad \gamma > 0, \quad L \text{ triangular inferior estricta.} \tag{2.2.8}$$

Se sigue de (2.2.3) que

$$Y = (I - z\bar{A})^{-1}(\bar{e} + zw)y_n, \tag{2.2.9}$$

Llevando esto a (2.2.6) obtenemos

$$Y^k = M(z)^k Y^0 + (I - M(z)^k)(I - z\bar{A})^{-1}(\bar{e} + zw)y_n.$$

Teniendo en cuenta que $y_{n+1}^k = Y_s^k = e_{s-1}^T Y^k$, con $e_{s-1}^T = (0, \ldots, 0, 1) \in I\!R^{s-1}$, se deduce que

$$\begin{aligned} y_{n+1}^k = \;& e_{s-1}^T M(z)^k Y^0 + e_{s-1}^T (I - z\bar{A})^{-1}(\bar{e} + zw)y_n \\ & - e_{s-1}^T M(z)^k (I - z\bar{A})^{-1}(\bar{e} + zw)y_n. \end{aligned} \tag{2.2.10}$$

Denotando por

$$R_k(z) := y_{n+1}^k / y_n,$$

y por $R(z)$ la función de estabilidad lineal del método, esto es,

$$R(z) = e_{s-1}^T (I - z\bar{A})^{-1}(\bar{e} + zw),$$

se sigue de (2.2.10) que

$$\begin{aligned} R_k(z) = \quad & R(z) + e_{s-1}^T M(z)^k Y^0/y_n \\ & + e_{s-1}^T M(z)^k \left(I - z^{-1}\bar{A}^{-1}\right)^{-1} \bar{A}(z^{-1}\bar{e} + w). \end{aligned} \tag{2.2.11}$$

Claramente si el radio espectral de $M(z)$, que denotaremos por $\rho(M(z))$, es menor que uno entonces $M^k(z) \to 0$ cuando $k \to \infty$, y por tanto el esquema iterativo converge independientemente de la aproximación inicial $Y^0$ tomada. Además se tendría que $R_k(z)$ tiende a $R(z)$. Estamos ahora en condiciones de probar los siguientes resultados:

**Teorema II.2.1** *$R_k(\infty) = R(\infty)$ para todo $k \geq 1$ y para cualquier valor inicial $Y^0$ fijado, si y sólo si*

$$e_{s-1}^T M(\infty) = e_{s-1}^T (I - T^{-1}\bar{A}) = 0^T.$$

*Demostración.* Es una consecuencia inmediata de (2.2.11). ∎

**Teorema II.2.2** *Si $\rho(M(\infty)) = 0$, entonces para todo valor inicial $Y^0$ fijo se tiene que:*

*(i)* $R_k(\infty) = R(\infty)$ *si* $k \geq s-1$ .

*(ii)* $R_k(z) = R(z) + \mathcal{O}(1/z^q)$, $z \to \infty$ *para todo* $k > (s-2)q, \quad q = 0, 1, 2, ...$

*Demostración. (i)* Ya que $M(\infty)$ es una matriz de dimensión $(s-1) \times (s-1)$ y $\rho(M(\infty)) = 0$, entonces $M(\infty)^k = 0$ para todo $k \geq s-1$. De (2.2.11) se concluye inmediatamente el resultado.

*(ii)* Por otro lado, denotando $M := M(\infty)$, obtenemos al desarrollar $M(z)$ en potencias de $x = 1/z$ que

$$M(z) = (I + xT^{-1} + x^2T^{-2} + \cdots)M.$$

Por tanto,

$$M(z)^k = M_0^{(k)} + M_1^{(k)}x + M_2^{(k)}x^2 + \ldots,$$

con $M_0^{(k)} = M^k$ y en general

$$M_r^{(k)} = \sum_{i_1+\ldots+i_k=r; i_j \geq 0} (T^{-i_1}M)(T^{-i_2}M)\ldots(T^{-i_k}M), \quad r \geq 0, \quad k \geq 1. \tag{2.2.12}$$

Obsérvese que en el momento que aparezca una potencia $M^{s-1}$, el sumando correspondiente en (2.2.12) se anularía. Por tanto, si se verifica una de las dos opciones siguientes:

$$i_1 = i_2 = \ldots = i_{s-1} = 0, \qquad \text{ó} \tag{2.2.13}$$

$$i_p = i_{p+1} = \ldots = i_{p+s-3} = 0, \quad 2 \leq p \leq k-s+3, \tag{2.2.14}$$

se tendrá que el sumando correspondiente en (2.2.12) se anulará.

De este modo, si

$$k = k^* := (r+1)(s-2), \quad s \geq 2,$$

entonces el único sumando no nulo en (2.2.12) es el correspondiente a la siguiente distribución de índices

$$i_1 = \ldots = i_{s-2} = 0, \quad i_{s-1} = 1,\ i_s = \ldots = i_{2s-4} = 0, \quad i_{2s-3} = 1,\ i_{2s-2} = \ldots = i_{3s-6} = 0, \quad \ldots$$

$$i_{r(s-2)+1} = 1, \quad i_{r(s-2)+2} = 0 = \ldots = i_{(r+1)(s-2)} = 0.$$

Si ahora consideramos $k > k^*$, entonces aparecerá en $M_r^{(k)}$, al menos un nuevo índice $i_{k+1}$ que debe ser nulo (al tener que verificarse que la suma de todos los índices es $r$). Esto implicaría que se verifica alguna de las dos condiciones indicadas arriba, (2.2.13) o (2.2.14), y por tanto $M_r^{(k)} = 0$. En definitiva, podemos concluir que

$$k > (r+1)(s-2) \Longrightarrow M_r^{(k)} = 0.$$

De aquí se infiere inmediatamente *(ii)* al tomar, $r = 0, 1, \ldots, q-1$. ∎

**Teorema II.2.3** *Si $\rho(M(\infty)) = 0$ y $e_{s-1}^T M(\infty) = 0^T$, entonces $R_k(z) = R(z) + \mathcal{O}(1/z^q)$ para todo $k > (s-2)(q-1)$ y para cualquier valor inicial $Y^0$ fijado.*

*Demostración.* Si desarrollamos $e_{s-1}^T M(z)^k$ en potencias de $x = 1/z$ obtenemos

$$e_{s-1}^T M(z)^k = e_{s-1}^T \sum_{r \geq 0} M_r^{(k)} x^k,$$

donde los $M_r^{(k)}$ están dados por (2.2.12). Teniendo ahora en cuenta que $e_{s-1}^T M = 0^T$, y que si $i_1 = 0$ en (2.2.12), entonces los sumandos correspondientes se anulan, resulta inmediato ver que

$$e_{s-1}^T M_r^{(k)} = e_{s-1}^T \sum_{i_1 + \ldots + i_k = r; i_1 \geq 1} (T^{-i_1} M)(T^{-i_2} M) \ldots (T^{-i_k} M), \quad r \geq 0,\ k \geq 1. \tag{2.2.15}$$

Usando ahora las mismas consideraciones que en la demostración del teorema anterior no es difícil establecer que

$$k > r(s-2) \Longrightarrow e_{s-1}^T M_r^{(k)} = 0^T.$$

De aquí, se concluye la demostración al tomar $r = 0, 1, \ldots, q-1$. ∎

Podemos deducir del teorema anterior (asumiendo sus hipótesis) que para la fórmula Lobatto IIIA de tres etapas se tiene que

$$R_k(z) = R(z) + \mathcal{O}(1/z^k), \qquad \forall k \geq 1.$$

Es decir, el orden de aproximación de la función de amplificación $R_k(z)$ en el infinito se incrementa una unidad en cada iteración. Por otra parte, para el Lobatto IIIA de cuatro etapas se tendría

$$R_k(z) = R(z) + \mathcal{O}(1/z^q), \qquad \forall k \geq 2q-1.$$

Por tanto el orden de aproximación se incrementa en una unidad cada dos iteraciones del esquema iterativo.

Finalmente observemos que asumiendo que $T$ es regular se tiene

$$e_{s-1}^T M(\infty) = 0^T \Longleftrightarrow e_{s-1}^T (I - \bar{A}^{-1} T) = 0^T. \tag{2.2.16}$$

Nos proponemos ahora desarrollar esquemas iterativos del tipo Single–Newton (2.2.5) donde la matriz incógnita $T$ debe verificar (2.2.8) con $\gamma > 0$. Exigiremos para ello que se cumplan las hipótesis del teorema anterior, esto es,

$$\text{(P1)} \quad \rho(I - T^{-1}\bar{A}) = 0.$$

$$\text{(P2)} \quad e_{s-1}^T (I - \bar{A}^{-1}T) = 0^T.$$

Se pedirá además (siempre que sea compatible con (P1)-(P2)) que

$$\text{(P3)} \qquad \sup_{z \in \mathbb{C}^-} \rho(M(z)) \quad \text{sea mínimo.}$$

Se podría plantear la minimización anterior sobre todo el semiplano complejo negativo, pero como se indica en [29] esto no suele suponer una mejora en la convergencia del esquema iterativo a efectos prácticos, ya que en la mayoría de los problemas stiff que surgen de las aplicaciones, los autovalores de la matriz jacobiana causantes de la stiffness suelen estar bastante más próximos al semieje real negativo que al eje imaginario.

## II.3 Diseño de esquemas iterativos para los métodos Lobatto IIIA de 3 y 4 etapas

Construiremos aquí esquemas iterativos eficientes de la forma (2.2.5) para los métodos Lobatto IIIA de tres y cuatro etapas. Como hemos visto en la sección anterior, nuestro objetivo será encontrar para cada una de esas fórmulas una matriz $T$ de la forma (2.2.8) que cumpla (P1)-(P2) y si es posible (P3).

### II.3.1 Lobatto IIIA de tres etapas

En este caso $s = 3$ y la submatriz $\bar{A}$ y el vector $w$ son [37, pág. 75],

$$\bar{A} = \begin{pmatrix} 1/3 & -1/24 \\ 2/3 & 1/6 \end{pmatrix}, \qquad w = \begin{pmatrix} 5/24 \\ 1/6 \end{pmatrix}.$$

Si denotamos $P = I - \bar{A}^{-1}T$, la condición (P2) exige que

$$P = \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}.$$

Por otra parte, la condición (P1) es equivalente a que $\rho(P) = 0$. Ésta a su vez es equivalente a que $a = 0$.

Finalmente, $T = \bar{A}(I - P)$ tiene un único autovalor $\gamma > 0$ si y sólo si

$$\det(T) = \det(\bar{A}) = \gamma^2, \quad \operatorname{tr}(T) = 2\gamma.$$

Por tanto

$$\gamma = \sqrt{\det(\bar{A})} = 1/\sqrt{12}, \quad \operatorname{tr}(T) = \operatorname{tr}(\bar{A}(I - P)) = \frac{3 - 4b}{6} = \frac{1}{\sqrt{3}}.$$

De aquí se deduce que $b = (3-2\sqrt{3})/4$, y consecuentemente en este caso existe una única matriz $T$ (con un autovalor único) verificando las condiciones (P1)-(P2). Esta matriz viene dada por

$$T = \begin{pmatrix} \frac{1}{3} & -\frac{7}{24} + \frac{1}{2\sqrt{3}} \\ \frac{2}{3} & -\frac{1}{3} + \frac{1}{\sqrt{3}} \end{pmatrix}. \tag{2.3.1}$$

La siguiente elección de las matrices $L$ y $S$ nos dan una factorización de $T$ en la forma (2.2.8):

$$S = \begin{pmatrix} 1 & 0.0669872981077806766 \\ 0 & 1 \end{pmatrix},$$

de tal modo que $S^{-1}TS$ es triangular inferior y

$$L = I - \gamma S^{-1}T^{-1}S = \begin{pmatrix} 0 & 0 \\ 2.30940107675850306 & 0 \end{pmatrix}.$$

Hemos elegido esta factorización (la cual no es única para $S$ y $L$) a efectos de disminuir el costo computacional involucrado en el esquema iterativo (2.2.5). Observe que si $S$ fuera una matriz "llena" el costo computacional aumentaría en virtud de las transformaciones matriz por vector que aparecen en (2.2.5).

También hemos calculado para este esquema el radio espectral de la matriz $M(z)$ sobre el eje real así como sobre el eje imaginario, obteniendo los valores máximos siguientes:

$$\max\{\rho(M(z)) \,|z \in (-\infty, 0)\} = \rho(M(-2\sqrt{3})) = \frac{2-\sqrt{3}}{4} = 0.066987298...$$

$$\max\{\rho(M(iy)) \,|y \in I\!R\} = \rho(M(2\sqrt{3}i)) = \frac{2-\sqrt{3}}{2} = 0.133974596....$$

Obsérvese que estas cantidades nos dan la razón de convergencia del esquema iterativo sobre problemas lineales con espectro del jacobiano contenido en $I\!R^-$ o $\mathbb{C}^-$ respectivamente.

Las propiedades de estabilidad lineal de la iteración $k-$ésima, $R_k(z)$ (ver (2.2.11) tomando $Y^0 = \bar{e}y_n$), vienen dadas en la tabla 2.3.1. El valor $\theta_k$ nos da el mayor ángulo tal que la $k-$ésima aproximación ($R_k(z)$) es $A(\theta_k) - -estable$, esto es, el máximo $\theta_k$ tal que

$$\text{si } \; z \in \Omega_k := \{|\arg(-z)| \leq \theta_k\} \; \Rightarrow \; |R_k(z)| \leq 1.$$

En la tabla 2.3.1 presentamos los valores de $\theta_k$ para $k = 1, 2, \ldots, 6$. También incluimos el mínimo valor $\alpha_k$ tal que

$$|R_k(z)| \leq 1, \qquad \forall\, Re\; z \leq -\alpha_k.$$

Como puede observarse, la primera y la segunda iteraciones son A–estables y para $k \geq 3$ las regiones de estabilidad se aproximan rápidamente al semiplano complejo negativo, que es la región de estabilidad lineal del método Lobatto IIIA.

Nótese que la iteración de Newton Simplificada da la solución exacta de las ecuaciones de las etapas en una sola iteración cuando el método se aplica a ecuaciones diferenciales lineales de coeficientes constantes, mientras que los esquemas Single–Newton necesitan varias iteraciones para alcanzar una solución suficientemente aproximada.

| $k$ | $\theta_k$ | $\alpha_k$ |
|---|---|---|
| 1 | 90.00° | 0.00 |
| 2 | 90.00° | 0.00 |
| 3 | 89.9924594192623° | 0.00171599547911613 |
| 4 | 89.9971724227644° | 0.00042079502790675 |
| 5 | 89.9993993611608° | 0.00007243526445789 |
| 6 | 89.9998952656572° | 0.00001114772053784 |

Tabla 2.3.1: Regiones de estabilidad del Lobatto IIIA de 3 etapas.

Por otro lado, un esquema iterativo no lineal *asociado* a (2.2.5)-(2.2.8), podría definirse de la forma siguiente:

$$Y^k = \bar{e}\otimes y_n + h(w\otimes f(t_n, y_n)) + h((\bar{A}-T)\otimes I)F(Y^{k-1}) + h(T\otimes I)F(Y^k), \quad k = 1, 2, \dots \quad (2.3.2)$$

Obsérvese que ambos esquemas iterativos coinciden sobre problemas lineales.

En [28] se dan condiciones suficientes para garantizar la convergencia de (2.3.2) sobre problemas contractivos en general (no sólo sobre problemas lineales), es decir, sobre funciones $f$ verificando para algún producto escalar $\langle\cdot,\cdot\rangle$,

$$\langle f(t,y) - f(t,\tilde{y}), y - \tilde{y}\rangle \leq 0, \quad \forall\, t \in \mathbb{R},\ y, \tilde{y} \in \mathbb{R}^m.$$

Se prueba allí que si existe una matriz $D$ diagonal definida positiva tal que

(a) $M = DT + T^T D = N^T N$ es definida positiva y

(b) $2|N^{-T}D(T-\bar{A})N^{-1}|_2 < 1$,

entonces la iteración (2.3.2) converge y una medida de la razón de convergencia viene dada por la cantidad,

$$\nu = 2 \inf_{D>0} |N^{-T}D(T-\bar{A})N^{-1}|_2. \quad (2.3.3)$$

donde $D > 0$ denota el conjunto de todas las matrices diagonales definidas positivas que verifican (a).

Para el esquema propuesto para el Lobatto IIIA de 3 etapas, hemos encontrado que la matriz

$$D = \mathrm{diag}(1, 0.2470356378869429)$$

nos da el valor mínimo

$$\nu = 0.2046574854265076$$

como razón de convergencia del esquema iterativo asociado (2.3.2).

### II.3.2 Lobatto IIIA de cuatro etapas

Para este método se tiene que [37, pág. 75],

$$\bar{A} = \begin{pmatrix} \frac{25-\sqrt{5}}{120} & \frac{25-13\sqrt{5}}{120} & \frac{-1+\sqrt{5}}{120} \\ \frac{25+13\sqrt{5}}{120} & \frac{25+\sqrt{5}}{120} & \frac{-1-\sqrt{5}}{120} \\ \frac{5}{12} & \frac{5}{12} & \frac{1}{12} \end{pmatrix}, \qquad w = \begin{pmatrix} \frac{11+\sqrt{5}}{120} \\ \frac{11-\sqrt{5}}{120} \\ \frac{1}{12} \end{pmatrix}.$$

Ahora buscamos una matriz $T$ que tenga un autovalor único $\gamma > 0$ y que satisfaga las condiciones (P1)-(P2)-(P3).

Como $T$ es una matriz $3 \times 3$, tendrá autovalor triple $\gamma$ si y sólo si

$$\det(T) = \gamma^3, \quad \mathrm{tr}(T) = 3\gamma, \quad \mathrm{tr}(T^2) = 3\gamma^2. \tag{2.3.4}$$

Usando ahora [29, Teorema 3.1], $T$ satisface la condición (P1) si y sólo si

$$\gamma = \sqrt[3]{\det(\bar{A})} = \sqrt[3]{1/120}, \quad \det(\bar{A} - T) = 0, \quad \mathrm{tr}(\bar{A}^{-1}T) = 3. \tag{2.3.5}$$

Puesto que la condición (P2) implica que $\det(\bar{A} - T) = 0$, entonces (2.2.8), (P1)-(P2) resultan equivalentes a

$$\begin{aligned} \det(T) = \gamma^3, \quad \mathrm{tr}(T)/3 = \gamma := (120)^{-1/3}, \\ \mathrm{tr}(T^2) = 3\gamma^2, \quad \mathrm{tr}(\bar{A}^{-1}T) = 3, \quad e_3^T(I - \bar{A}^{-1}T) = 0^T. \end{aligned} \tag{2.3.6}$$

Lo que nos da un total de 7 ecuaciones, siendo 4 de ellas lineales. Así pues disponemos de 2 parámetros libres (en la matriz $T$) que serán usados para minimizar el radio espectral de $M(z)$ sobre el semieje real negativo, es decir, para minimizar la cantidad

$$\max\{\rho(M(z)) \,|z \in (-\infty, 0)\},$$

y también para minimizar la norma de Fröbenius (esto es la raíz cuadrada de la suma de los cuadrados de los elementos de una matriz),

$$|I - \bar{A}^{-1}T|_F.$$

**Minimización del radio espectral**

Asumiendo (2.3.6), queda claro que uno de los autovalores de la matriz $M(z)$ dada en (2.2.7) es nulo independientemente del valor de $z$. De este modo, no es difícil probar que los otros dos autovalores $\{\mu_j(z),\ j = 1, 2\}$ verifican la ecuación de segundo grado siguiente (ver [29] para detalles),

$$\mu^2(z) - b(z)\mu(z) + c(z) = 0, \tag{2.3.7}$$

donde

$$b(z) = \mathrm{tr}(M(z)), \qquad c(z) = \frac{\det(I - z\bar{A})}{(1 - \gamma z)^3} + b(z) - 1.$$

Definiendo ahora

$$\Lambda := (\mathrm{tr}^2(\bar{A}) - \mathrm{tr}(\bar{A}^2))/2,$$

se tiene que

$$\det(I - z\bar{A}) = 1 - \mathrm{tr}(\bar{A})z + \Lambda z^2 - \det(\bar{A})z^3.$$

Usando para $M(z)$ los desarrollos asintóticos siguientes:

$$\begin{aligned} &\mathrm{tr}(M(z)) = z\ \mathrm{tr}(\bar{A} - T) + \mathcal{O}(z^2), \quad (z \to 0), \\ &\mathrm{tr}(M(z)) = \mathrm{tr}(I - T^{-1}\bar{A}) + z^{-1}\mathrm{tr}(T^{-1} - T^{-2}\bar{A}) + \mathcal{O}(z^{-2}), \quad (z \to \infty), \end{aligned}$$

y recordando que $\mathrm{tr}(T^{-1}) = 3/\gamma$, resulta que los coeficientes de la ecuación de autovalores (2.3.7) pueden expresarse como funciones racionales de $z$ mediante

$$b(z) = \frac{z}{(1 - \gamma z)^3}(\xi_1 z + \mathrm{tr}(\bar{A}) - 3\gamma), \qquad c(z) = \frac{z^2}{(1 - \gamma z)^3}\xi_2,$$

donde

$$\xi_1 = \gamma^3\delta - 3\gamma^2, \quad \xi_2 = \gamma^3\delta + \Lambda - 6\gamma^2, \quad \delta = \mathrm{tr}(T^{-2}\bar{A}).$$

En definitiva, los coeficientes de la ecuación de autovalores (2.3.7) sólo dependen de la matriz $\bar{A}$ y del parámetro $\delta$. Luego, para minimizar $\rho(M(z))$ sobre el semieje real negativo tendremos que encontrar un valor $\delta = \delta_0$ de tal forma que la cantidad

$$\Gamma(\delta) = \max\{\rho(M(z)) : z \leq 0\}$$

sea mínima. Luego bastará elegir la matriz $T$ de tal forma que $\mathrm{tr}(T^{-2}\bar{A}) = \delta_0$.

**Lema II.3.1** *Para el RK Lobatto IIIA de cuatro etapas, existe un único valor*

$$\delta_0 = 17.82460709187153934...$$

*tal que* $\Gamma(\delta_0) < \Gamma(\delta)$ *para todo* $\delta \neq \delta_0$, *i.e., para* $\mathrm{tr}(T^{-2}\bar{A}) = \delta_0$ *la cantidad* $\max\{\rho(M(z)) : z \leq 0\}$ *es mínima.*

*Demostración.* Considerando el caso particular

$$c(z) \equiv 0,$$

se deduce que $\xi_2 = 0$ ó bien

$$\delta = (-\Lambda + 6\gamma^2)/\gamma^3.$$

En este caso "0" sería un autovalor doble de $M(z)$, y además

$$\xi_1 = -\Lambda + 3\gamma^2, \quad \rho(M(z)) = |b(z)|.$$

De aquí se obtiene que

$$\max\{\rho(M(z)) : z \leq 0\} = r_0 = 0.1376449122...$$

Sea ahora un $r$ cualquiera verificando

$$0 < r \leq r_0.$$

Haciendo el siguiente cambio de variables

$$t = \frac{1}{(1-\gamma z)} \in [0,1], \qquad x(t) = \frac{\mu(t)}{r}, \ \forall t \in (0,1), \tag{2.3.8}$$

resulta que la ecuación de autovalores (2.3.7) es equivalente a

$$r^2x^2(t) - rB(t)x(t) + C(t) = 0, \quad t \in (0,1) \tag{2.3.9}$$

con

$$B(t) = -(B_1t + B_2)(1-t)t, \qquad C(t) = B_3(1-t)^2t,$$

siendo

$$B_1 = \gamma\delta + \frac{\mathrm{tr}(\bar{A})}{\gamma} - 6, \quad B_2 = -\gamma\delta + 3, \quad B_3 = \gamma\delta + \frac{\Lambda}{\gamma^2} - 6. \tag{2.3.10}$$

Por tanto, para $z < 0$ las raíces $\mu_i(z)$ de la ecuación (2.3.7) tendrán módulo menor que $r$ sii las raíces $x_1(t)$ y $x_2(t)$ de (2.3.9) verifican

$$|x_i(t)| < 1, \quad \forall t \in (0,1).$$

Aplicando el Criterio de Schur-Cohn [51], esta condición se verificará si y sólo si

$$|B_3| < \frac{r^2}{(1-t)^2 t}, \qquad r|B(t)||r^2 - C(t)| < r^4 - |C(t)|^2, \quad \forall t \in (0,1).$$

Esto es equivalente a

$$|B_3| < 27r^2/4, \tag{2.3.11}$$

$$r|B(t)| < r^2 + C(t), \quad \forall t \in (0,1). \tag{2.3.12}$$

Teniendo en cuenta que

$$B_1 = B_3 + K_1, \quad B_2 = -B_3 + K_2,$$

con

$$K_1 = \frac{\mathrm{tr}(\bar{A})}{\gamma} - \frac{\Lambda}{\gamma^2}, \quad K_2 = \frac{\Lambda}{\gamma^2} - 3,$$

entonces la desigualdad (2.3.12) es equivalente a

$$B_3 > \max_{t\in(0,1)} \{P_1(r,t), P_2(r,t)\},$$

donde

$$P_1(r,t) = \frac{r}{(1+r)(1-t)} \left[ (K_1 t + K_2) - \frac{r}{(1-t)t} \right],$$

$$P_2(r,t) = \frac{r}{(1-r)(1-t)} \left[ -(K_1 t + K_2) - \frac{r}{(1-t)t} \right].$$

Por otra parte, ya que para cada $r \in (0, r_0]$, se puede probar que

$$\lim_{t\to 0^+} P_1(r,t) = \lim_{t\to 0^+} P_2(r,t) = \lim_{t\to 1^-} P_1(r,t) = \lim_{t\to 1^-} P_2(r,t) = -\infty, \quad j = 1,2, \tag{2.3.13}$$

y

$$\max_{t\in(0,1)} P_2(r,t) > \max_{t\in(0,1)} P_1(r,t).$$

Entonces, resulta que (2.3.12) es equivalente a

$$B_3 > \max_{t\in(0,1)} P_2(r,t). \tag{2.3.14}$$

Sea ahora

$$\Omega = \{r \in (0, r_0] \text{ tal que existe } B_3 \text{ verificando (2.3.11) y (2.3.14)}\}.$$

Estamos interesados en obtener el ínfimo de este conjunto, el cual es no vacío al tenerse que $r_0 \in \Omega$ (lo cual se puede comprobar). No es difícil ver teniendo en cuenta (2.3.13) que el ínfimo se obtiene al exigir

$$\max_{t\in(0,1)} P_2(r,t) = 27r^2/4.$$

De aquí se deduce inmediatamente que $r$ y $t$ deben verificar

$$27r^2/4 = P_2(r,t)\ , \qquad \frac{dP_2(r,t)}{dt} = 0.$$

Fácilmente se ve que el sistema anterior admite solución única. El valor de $r$ obtenido es

$$r = 0.08312670...$$

De aquí se deduce un valor para $B_3 = 27r^2/4$ y esto llevado a (2.3.10) arroja el siguiente valor para $\delta_0$ (con 18 cifras significativas),

$$\delta_0 = 17.8246070918715393...$$

∎

En este caso, el valor máximo del radio espectral de $M(z)$, $z \leq 0$, se alcanza en el punto $z \simeq -2.6576$ y para el eje imaginario, el máximo se alcanza en $z = 6.0907322...i$, dando el valor

$$\max\{\rho(M(iy)), y \in I\!R\} = 0.253668...$$

**Optimizando la matriz $T$**

Después de las consideraciones anteriores, vamos a buscar una matriz $T$ que verifique las siguientes condiciones,

(C1) $\det(T) = \gamma^3 := \det(\bar{A}) = 1/120$,
(C2) $\mathrm{tr}(T) = 3\gamma$,
(C3) $\mathrm{tr}(T^2) = 3\gamma^2$,
(C4) $\mathrm{tr}(\bar{A}^{-1}T) = 3$,
(C5) $\mathrm{tr}(T^{-2}\bar{A}) = \delta_0$,
(C6) $(0,0,1)\,(I - \bar{A}^{-1}T) = (0,0,0)$,
(C7) $|I - \bar{A}^{-1}T|_F$ sea tan pequeña como sea posible.

Teniendo en cuenta que por el teorema de Cayley-Hamilton, $T$ debe verificar (al tener autovalor triple $\gamma$),

$$p(T) = T^3 - 3\gamma T^2 + 3\gamma^2 T - \gamma^3 I = 0,$$

y que además de (2.3.6) se deduce que

$$\mathrm{tr}(T^{-1}A) = \mathrm{tr}(A^{-1}T) = 3,$$

podemos entonces reemplazar la condición (C5) por la siguiente ecuación lineal en los coeficientes de $T$,

(C5') $\mathrm{tr}(\bar{A}T) = \gamma^3\delta_0 + 3\gamma\mathrm{tr}(\bar{A}) - 9\gamma^2$.

Para simplificar el problema de minimización (ver (C7)), en lugar de obtener directamente la matriz $T$, podemos obtener de forma equivalente la matriz $P = I - \bar{A}^{-1}T$, la cual debe verificar las siguientes condiciones que son equivalentes a (C1)-(C7):

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ 0 & 0 & 0 \end{pmatrix}$$

y además

(D1) $\mathrm{tr}(P) = 0$,
(D2) $\det(I - P) = 1$,
(D3) $\mathrm{tr}(\bar{A}P) = \mathrm{tr}(\bar{A}) - 3\gamma$,
(D4) $\mathrm{tr}(\bar{A}^2P) = \mathrm{tr}(\bar{A}^2) - K$,
(D5) $\mathrm{tr}((\bar{A}P)^2) = \mathrm{tr}(\bar{A}^2) + 3\gamma^2 - 2K$,
(D6) $|P|_F$ es mínima,

donde

$$K = \gamma^3\delta_0 + 3\gamma\mathrm{tr}(\bar{A}) - 9\gamma^2.$$

Claramente las condiciones (D1), (D3) y (D4) son lineales en los coeficientes de $P$. Por tanto, todas las condiciones anteriores se reducen a dos ecuaciones no lineales y una condición de minimización (D6).

| $k$ | $\theta_k$ | $\alpha_k$ |
|---|---|---|
| 1 | 90.00° | 0.00 |
| 2 | 89.3492085759° | 0.3737593155565 |
| 3 | 89.9272843260° | 0.0161715124278 |
| 4 | 89.9784185933° | 0.0036772308997 |
| 5 | 89.9909818221° | 0.0013306151786 |
| 6 | 89.9967005782° | 0.0004405981063 |

Tabla 2.3.2: Regiones de estabilidad del Lobatto IIIA de 4 etapas.

La construcción de la matriz $P$ ha sido llevada a cabo numéricamente usando un paquete de minimización (librerías IMSL). De este modo hemos obtenido los siguientes valores (con 16 cifras significativas):

$$P = \begin{pmatrix} -0.01832522921162447 & -0.1240694135537153 & 0.02056578515084587 \\ 0.002706662472561666 & 0.01832522921162447 & -0.1475471666586898 \\ 0 & 0 & 0 \end{pmatrix},$$

$$|P|_F = 0.1956151613768685.$$

La correspondiente matriz $T = \bar{A}(I - P)$ del esquema resulta

$$T = \begin{pmatrix} 0.1932674949117222 & -0.009750106539280771 & 0.001396313165263860 \\ 0.4582165795963249 & 0.2787104623506828 & -0.002745269684755689 \\ 0.4231744028079428 & 0.4607267434758711 & 0.1362422422949350 \end{pmatrix}. \quad (2.3.15)$$

En este caso, y de igual modo que en el Lobatto IIIA de 3 etapas, es posible encontrar una matriz $S$ triangular superior:

$$S = \begin{pmatrix} 1 & -0.0013313944847890405 & -0.021160953394204083 \\ 0 & 1 & 0.16376865269504141 \\ 0 & 0 & 1 \end{pmatrix} \quad (2.3.16)$$

que transforma $T$ en una matriz triangular (ver (2.2.8)) y

$$L = I - \gamma S^{-1}T^{-1}S = \begin{pmatrix} 0 & 0 & 0 \\ 1.91828820257772989 & 0 & 0 \\ -2.26670285249783297 & 2.26972072817430417 & 0 \end{pmatrix}. \quad (2.3.17)$$

En la tabla 2.3.2 presentamos igual que en el caso anterior los valores máximos $\theta_k$ para los que la $k-$ésima iteración, $k = 1, ..., 6$, es $A(\theta_k)-$estable, y también el mínimo $\alpha_k$ para el que $|R_k(z)| \leq 1$ para todo $Re\ z \leq -\alpha_k$. Para este método la primera iteración es A–estable y de nuevo para $k \geq 2$ las regiones de estabilidad de las iteraciones sucesivas se aproximan al semiplano complejo negativo a medida que $k$ aumenta.

En este caso, de nuevo hemos encontrado una matriz diagonal definida positiva

$$D = \mathrm{diag}(1, 0.34516405, 0.081325562)$$

tal que

$$\nu = 2 \inf_{D>0} |N^{-T}D(T - \bar{A})N^{-1}|_2 = 0.71036694...$$

es mínimo (ver (2.3.3)).

## II.4 Experimentos numéricos

En esta sección presentamos algunos experimentos numéricos con dos objetivos principales:

1. Mostrar que los nuevos esquemas iterativos propuestos en las secciones anteriores para integrar problemas stiff, son una alternativa más eficiente en muchos casos que la iteración de Newton Simplificada [37, Cap.IV.8] a la hora de resolver las ecuaciones algebraicas de las etapas.

2. Mostrar que las fórmulas Lobatto IIIA integran bien los problemas de tipo stiff.

Con este fin hemos implementado dos códigos a paso variable basados en la fórmula Lobatto IIIA de 4 etapas, uno denominado SN-LOB cuyas ecuaciones de etapa han sido resueltas usando la iteración (2.2.5), con $\gamma$ dado en (2.3.6) y las matrices $S$ y $L$ dadas respectivamente por (2.3.16) y (2.3.17). El otro código, que denominaremos MN-LOB, está basado en la iteración de Newton Simplificada tal cual se explica en [37, Cap.IV.8]. El error local en ambos casos se estima mediante la técnica de extrapolación, y la elección del cambio de tamaño de paso se ha basado en la fórmula usual

$$h_{n+2} = 0.9(\mathrm{Tol}/|\mathrm{Est}|)^{1/7} h_n,$$

después de dos pasos de tamaño $h_n$ aceptados, pero dividiendo a la mitad el paso cuando la estimación del error local "Est" es mayor que una tolerancia "Tol" prescrita por el usuario.

Para resolver las ecuaciones implícitas (2.2.3) en cada paso (en realidad damos dos pasos consecutivos del mismo tamaño) desde $t_n$ hasta $t_{n+2} = t_n + 2h_n$ hemos procedido de la siguiente manera para ambos códigos:

(1) Evaluamos la matriz jacobiana $J_n = \partial f/\partial y(t_n, y_n)$ sobre los puntos pares de red, i.e., $t_n,\ n = 0, 2, 4, \ldots$.

(2) Aplicamos la iteración considerada (Single–Newton o Newton Simplificado) para calcular las etapas internas $Y_{i,n}$ $(i = 2, 3, 4)$ del primer paso $t_n \to t_{n+1} = t_n + h_n$. Usamos como aproximaciones iniciales, $Y_{i,n}^{(0)}$ $(i = 2, 3, 4)$ las obtenidas mediante la interpolación de Lagrange de las etapas internas del paso anterior $Y_{i,n-1}$ $(i = 1, \ldots, 4)$ (para el paso inicial $n = 0$, las aproximaciones iniciales tomadas son $Y_{i,0} = y_0$, $2 \le i \le 4$). Observe que también necesitamos una factorización LU de una matriz de la forma $(I - \gamma h_n J_n)$ para el caso Single–Newton y dos factorizaciones LU (una compleja y una real) para la iteración de Newton Simplificada. Se alcanza la convergencia del esquema iterativo cuando

$$\max_{2\le i\le 4} \left( |Y_{i,n}^{(l)} - Y_{i,n}^{(l-1)}|_\infty \right) \le 0.01\ \mathrm{Tol}, \qquad l = 1, \ldots, 10.$$

Entonces tomamos $Y_{i,n} := Y_{i,n}^{(l)},\ 2 \le i \le 4$. Si no se ha obtenido convergencia después de 10 iteraciones o si algún cociente

$$\tau_r = \max_{2\le i\le 4} (|Y_{i,n}^{(r)} - Y_{i,n}^{(r-1)}|_\infty) / \max_{2\le i\le 4} (|Y_{i,n}^{(r-1)} - Y_{i,n}^{(r-2)}|_\infty), \qquad r = 2, 3, \ldots, 10,$$

es mayor que uno, entonces consideramos que no hay convergencia del esquema iterativo para el tamaño de paso tomado.

(3) Repetimos el proceso descrito en (2) para el segundo paso, i.e., el paso de $t_{n+1} \to t_{n+2} = t_n + 2h_n$. Aquí se aprovechan las factorizaciones LU ya realizadas en (2), y se toman como aproximaciones iniciales, $Y^{(0)}_{i,n+1}$ $(i = 2, 3, 4)$, las que se obtienen de la interpolación de Lagrange de las etapas del paso anterior, $Y_{i,n}$ $(i = 1, \ldots, 4)$.

(4) Calculamos un estimador del error local como es habitual al usar extrapolación. Para ello necesitamos dar un nuevo paso (doble) desde $t_n$ hasta $t_{n+2} = t_n + 2h_n$, usando esta vez como tamaño de paso $h^*_n = 2h_n$. Ahora se necesitan nuevas factorizaciones LU de matrices de la forma $(I - \gamma h^*_n J_n)$. Las aproximaciones iniciales para el paso doble se calculan de las obtenidas previamente, $Y_{i,n}, Y_{i,n+1}$, $(i = 1, \ldots, 4)$, usando interpolación de Lagrange.

Si no hubiera convergencia de la iteración considerada en alguno de los pasos (2), (3) ó (4) de arriba, entonces reiniciamos el proceso en el paso (2), pero esta vez dividiendo por dos el tamaño de paso $h_n$, con el objeto de aprovechar algunos cálculos ya realizados.

Por brevedad, aquí sólo presentaremos los resultados numéricos obtenidos con tres problemas stiff escogidos de la literatura, los dos primeros de baja dimensión y el tercero de dimensión algo mayor, y que suelen ser tomados como prototipos de esta clase de problemas [37, Cap. IV.10].

**Problema 1.-** Oregonator (ver por ejemplo [37, pág. 144]):

$$\begin{aligned} &y_1' = 77.27(y_2 + y_1(1 - 8.375 \times 10^{-6} y_1 - y_2)),\\ &y_2' = (y_3 - (1 + y_1)y_2)/77.27\\ &y_3' = 0.161(y_1 - y_3)\\ &y_1(0) = 1,\ y_2(0) = 2,\ y_3(0) = 3, \quad t \in [0, 3600] \end{aligned}$$

**Problema 2.-** El oscilador de Van der Pol (ver por ejemplo [37, pág. 144]):

$$\begin{aligned} &y_1' = y_2\\ &y_2' = ((1 - y_1^2)y_2 - y_1)/\varepsilon, \qquad \varepsilon = 10^{-6}\\ &y_1(0) = 2, y_2(0) = 0, \quad t \in [0, 20], \end{aligned}$$

**Problema 3.-** CUSP (ver por ejemplo [37, pág. 147]):

$$\begin{aligned} &y_i' = -(1/\varepsilon)\left(y_i^3 + a_i y_i + b_i\right) + D(y_{i-1} - 2y_i + y_{i+1}),\\ &a_i' = b_i + 0.07 v_i + D(a_{i-1} - 2a_i + a_{i+1}) \qquad i = 1, \ldots, N,\\ &b_i' = (1 - a_i^2)b_i - a_i - 0.4 y_i + 0.035 v_i + D(b_{i-1} - 2b_i + b_{i+1}) \end{aligned}$$

donde

$$v_i = \frac{u_i}{u_i + 1}, \quad u_i = (y_i - 0.7)(y_i - 1.3), \quad D = \frac{N^2}{100}, \quad \varepsilon = 10^{-8}, \quad N = 32$$

y

$$y_0 = y_N, \quad a_0 = a_N, \quad b_0 = b_N, \quad y_{N+1} = y_1, \quad a_{N+1} = a_1, \quad b_{N+1} = b_1.$$

Hemos tomado los valores iniciales

$$y_i(0) = 0, \quad a_i(0) = -2\cos(2i\pi/N), \quad b_i(0) = 2\sin(2i\pi/N), \quad i = 1, \ldots, N$$

y $t \in [0, 1.1]$.

Para medir la eficiencia del método Lobatto IIIA como integrador de problemas stiff, hemos comparado nuestros códigos con otros tres bien conocidos que son adecuados para este tipo de problemas:

1. El código EPISODE, un código de paso variable y orden variable basado en fórmulas BDF de órdenes de 1 a 5.

2. El código STRIDE [4, 5], basado en métodos SIRK, que también usa una estrategia de paso y orden variables, con métodos cuyos órdenes varían desde 1 hasta 12.

3. El código RADAU5 basado en la fórmula RK Radau IIA de orden 5 desarrollado por Hairer y Wanner [37]. Este código usa la iteración de Newton Simplificada para resolver las ecuaciones de etapa y hace uso de aritmética compleja.

Además, también hemos comparado nuestro código con uno basado en el método RK Radau IIA de orden 5 desarrollado por González Pinto et al. [29] (que llamaremos SN-RADAU), que usa para la resolución de las ecuaciones de las etapas un esquema iterativo del tipo (2.2.5).

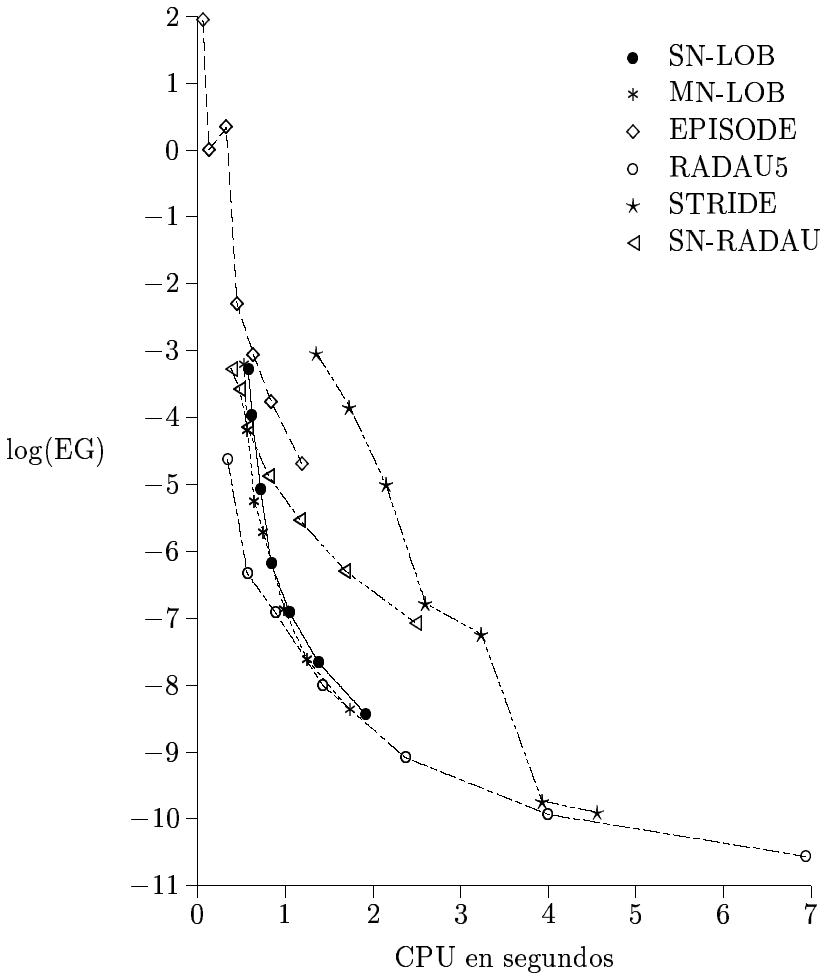
Para cada código hemos integrado cada problema con tolerancias desde $10^{-4}$ a $10^{-10}$, computando el error global máximo (EG) en el punto final del intervalo de integración así como el tiempo de CPU (en segundos) invertido en la integración. En las gráficas 2.4.1, 2.4.2 y 2.4.3 hemos dibujado los puntos (log(EG),CPU) obtenidos por cada código y tolerancia unidos por líneas poligonales.

En lo que respecta a la eficiencia del esquema iterativo propuesto, podemos observar en las gráficas que el código Lobatto IIIA basado en la iteración de Newton Simplificada es tan eficiente como el basado en el esquema Single–Newton en los problemas 1 y 2, pero sobre el problema CUSP (de dimensión elevada) el segundo es mucho más eficiente. Como ambos códigos sólo difieren en el esquema iterativo, alcanzan una precisión similar dando prácticamente el mismo número de pasos. Por tanto, por el elevado costo de las factorizaciones LU, la iteración de Newton Simplificada necesitará aproximadamente cinco veces más trabajo que el Single–Newton. También probablemente se diferenciarán en el número de iteraciones que necesitan para obtener convergencia y, en consecuencia, en el número de sistemas lineales que tienen que resolver y en el número de evaluaciones de la función derivada $f$.

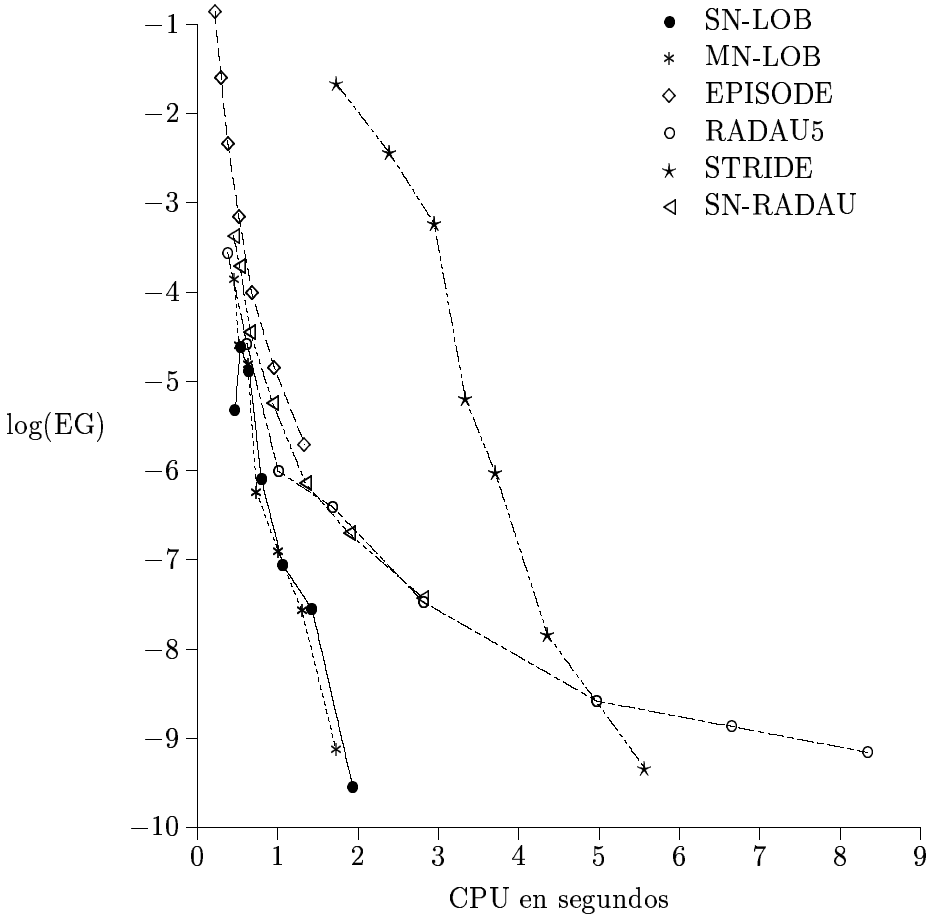
Nótese que el Single–Newton en cada iteración tiene que resolver 3 sistemas lineales de dimensión $m$ con la misma matriz de coeficientes previamente factorizada (aproximadamente $3m^2$ operaciones) mientras que la iteración de Newton Simplificada conlleva en cada iteración un sistema lineal de dimensión $m$ más un sistema lineal complejo de dimensión $m$ (aproximadamente $5m^2$ operaciones en total). En general, la iteración de Newton Simplificada necesita menos iteraciones para converger que el Single–Newton, y para problemas de dimensión pequeña, esto compensa el coste computacional realizado en las factorizaciones LU y la correspondiente resolución de sistemas lineales, por lo que ambos esquemas tienen un comportamiento similar. Sin embargo, en problemas de dimensión grande, como el CUSP, el coste computacional se debe principalmente al trabajo invertido en las factorizaciones LU y por tanto, la iteración de Newton Simplificada es claramente menos eficiente que el Single–Newton.

En las tablas 2.4.1 y 2.4.2 presentamos para los problemas Van der Pol y CUSP, con tolerancias de $10^{-4}$ a $10^{-10}$, un promedio del número de iteraciones por paso, empleado por ambos esquemas en el primer paso de la extrapolación (de $t_n$ a $t_n + h$), en el segundo paso de extrapolación (de $t_n + h$ a $t_n + 2h$) y en el paso doble (de $t_n$ a $t_n + 2h$).

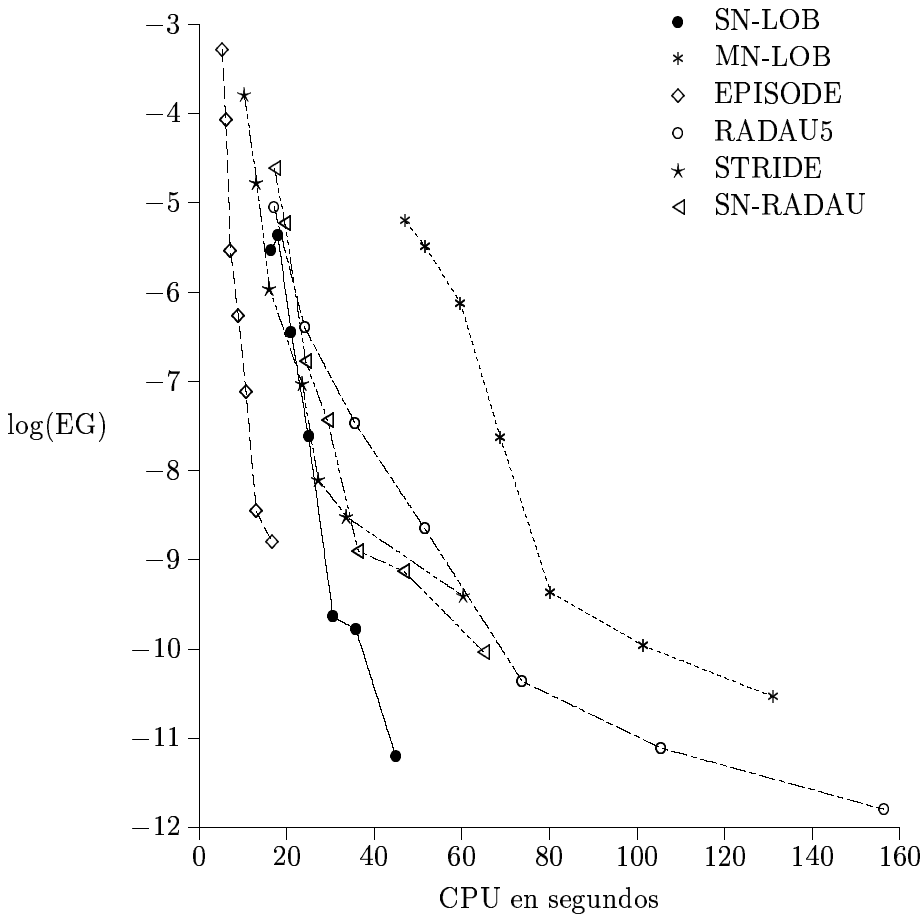
En cuanto a la eficiencia de la fórmula Lobatto IIIA como integrador de problemas stiff, hemos apreciado en nuestros experimentos que en general el código basado en dicha fórmula tiene un buen comportamiento, haciéndolo competitivo con los otros códigos usados, e incluso se prueba que es superior en algunos casos particulares. Este código usualmente da menos pasos




Gráfica 2.4.1. Problema Oregonator.




Gráfica 2.4.2. Problema Van der Pol.




Gráfica 2.4.3: Problema CUSP.

| | MN-LOB | | | SN-LOB | | |
|---|---|---|---|---|---|---|
| Tol | Primer paso | Segundo paso | Paso doble | Primer paso | Segundo paso | Paso doble |
| $10^{-4}$ | 6.6 | 7.5 | 2.5 | 8.0 | 8.4 | 2.9 |
| $10^{-5}$ | 6.4 | 7.3 | 2.6 | 7.9 | 8.2 | 3.2 |
| $10^{-6}$ | 5.8 | 6.6 | 2.8 | 7.4 | 8.1 | 3.3 |
| $10^{-7}$ | 5.4 | 6.3 | 2.8 | 7.2 | 7.7 | 3.7 |
| $10^{-8}$ | 4.9 | 5.7 | 3.0 | 6.9 | 7.3 | 3.9 |
| $10^{-9}$ | 4.6 | 5.2 | 3.1 | 6.7 | 7.0 | 4.1 |
| $10^{-10}$ | 4.4 | 4.9 | 3.1 | 6.6 | 6.8 | 4.3 |

Tabla 2.4.1: Lobatto IIIA. Número de iteraciones por paso para el problema de Van der Pol.

| | MN-LOB | | | SN-LOB | | |
|---|---|---|---|---|---|---|
| Tol | Primer paso | Segundo paso | Paso doble | Primer paso | Segundo paso | Paso doble |
| $10^{-4}$ | 6.3 | 5.4 | 1.5 | 7.1 | 5.6 | 1.7 |
| $10^{-5}$ | 6.7 | 6.1 | 1.7 | 7.6 | 6.7 | 1.5 |
| $10^{-6}$ | 7.1 | 6.8 | 1.7 | 8.1 | 7.1 | 1.8 |
| $10^{-7}$ | 6.7 | 7.1 | 2.2 | 8.0 | 7.7 | 2.4 |
| $10^{-8}$ | 6.4 | 7.1 | 2.7 | 7.7 | 8.0 | 2.9 |
| $10^{-9}$ | 5.9 | 6.6 | 2.8 | 7.3 | 7.5 | 3.3 |
| $10^{-10}$ | 5.5 | 6.3 | 2.9 | 7.1 | 7.4 | 3.5 |

Tabla 2.4.2: Lobatto IIIA. Número de iteraciones por paso para el problema CUSP.

| Tol | SN-LOB | MN-LOB | STRIDE | RADAU5 | EPISODE | SN-RADAU |
|---|---|---|---|---|---|---|
| $10^{-4}$ | 208 | 202 | 399 | 115 | 310 | 228 |
| $10^{-5}$ | 230 | 220 | 546 | 196 | 469 | 250 |
| $10^{-6}$ | 262 | 256 | 691 | 316 | 698 | 304 |
| $10^{-7}$ | 318 | 304 | 974 | 527 | 930 | 366 |
| $10^{-8}$ | 382 | 356 | 1126 | 901 | 1239 | 464 |
| $10^{-9}$ | 456 | 448 | 1305 | 1562 | 1726 | 608 |
| $10^{-10}$ | 582 | 560 | 1788 | 2764 | 2433 | 856 |

Tabla 2.4.3: Problema CUSP. Número de pasos.

que los otros manteniendo a la vez la precisión exigida. Así, el código ha sido especialmente eficiente con los problemas Van der Pol y CUSP, lo que da esperanzas de que sea apropiado para integrar problemas de perturbaciones singulares.

Observando también el comportamiento de los otros códigos tenemos que:

EPISODE en general es muy rápido, debido a que aún necesitando más pasos que los códigos basados en métodos RK, el coste computacional por paso es bastante menor que en estos últimos métodos. Sin embargo, a menudo pierde precisión en las soluciones computadas.

STRIDE normalmente realiza menos factorizaciones LU y evaluaciones de matriz jacobiana que los otros, pero en cambio, es menos eficiente para problemas de dimensión baja. Esto se explica teniendo en cuenta que suele dar más pasos que los otros códigos RK. Además, como es un método SIRK de orden $s$ tiene $s$ etapas, y por tanto han de resolverse en cada iteración $s$ sistemas lineales con la misma matriz previamente factorizada. Así, por ejemplo, un método de orden 6 necesita aproximadamente $6m^2$ operaciones, mientras que el método Lobatto IIIA de orden 6 sólo necesita $3m^2$.

También STRIDE presenta la desventaja de que requiere varios productos matriz por vector que, por ejemplo, para 6 etapas conllevan un gasto de $6^2m$ operaciones adicionales, mientras que para el Lobatto IIIA esta cantidad es sólo $3^2m$. Nótese que para los problemas 1 y 2, con $m = 3$ y $m = 2$ respectivamente, la influencia de estas transformaciones es relevante en el tiempo total de CPU.

En lo que respecta al RADAU5 vemos que presenta un buen comportamiento sobre todos los problemas integrados, pero a medida que la tolerancia decrece es menos eficiente debido a que incrementa considerablemente el número de pasos de integración.

El otro código SN-RADAU de González et al. [29], que también usa un esquema Single–Newton presenta en general buenos resultados. Sin embargo, es importante observar el hecho de que para tolerancias pequeñas el Lobatto IIIA da muchos menos pasos que el SN-RADAU. Esto puede explicarse por el hecho de que este tiene orden 5 mientras que el Lobatto IIIA tiene orden 6. También es posible que lo explique el hecho de que la fórmula Lobatto IIIA tiene orden de etapa 4, mientras que para el Radau IIA el orden de etapa es 3.

Particularmente interesante es el comportamiento de estos códigos cuando integran problemas de dimensión alta como el CUSP. En este caso el coste computacional se debe principalmente al número de factorizaciones LU que se necesitan en relación con el número de pasos dados. En la tabla 2.4.3 damos para cada tolerancia y código el número de pasos dados mientras que en la tabla 2.4.4 incluimos el número de factorizaciones LU efectuadas con el problema CUSP. Para

| Tol | SN-LOB | MN-LOB | STRIDE | RADAU5 | EPISODE | SN-RADAU |
|---|---|---|---|---|---|---|
| $10^{-4}$ | 250 | 236 | 178 | 180 | 220 | 270 |
| $10^{-5}$ | 262 | 250 | 200 | 252 | 465 | 289 |
| $10^{-6}$ | 297 | 285 | 217 | 372 | 738 | 346 |
| $10^{-7}$ | 347 | 331 | 306 | 535 | 292 | 402 |
| $10^{-8}$ | 419 | 385 | 325 | 763 | 627 | 498 |
| $10^{-9}$ | 487 | 475 | 387 | 1079 | 402 | 639 |
| $10^{-10}$ | 610 | 580 | 810 | 1582 | 9030 | 893 |

Tabla 2.4.4: Problema CUSP. Número de factorizaciones LU.

los casos del RADAU5 y el MN-LOB cada factorización LU de la tabla engloba la factorización de una matriz real $m \times m$ más la factorización de una matriz compleja $m \times m$. Por tanto, debería multiplicarse por cinco la cantidad que aparece en la tabla, a efectos de comparar con los otros códigos.

Los datos en estas tablas concuerdan con los resultados observados en la gráfica 2.4.3. La buena actuación de EPISODE y del Lobatto IIIA para tolerancias bajas refleja que necesitan pocas factorizaciones LU. Nótese que el código Lobatto IIIA evalúa la matriz jacobiana en cada punto y, por tanto, realiza una factorización LU por paso. Un refinamiento del código que reutilice la matriz jacobiana cuando sea posible, al igual que hacen EPISODE o STRIDE por ejemplo, reduciría el número de factorizaciones LU, con la consiguiente mejora en la eficiencia.

Como conclusión, de nuestros experimentos podemos deducir que los métodos Lobatto IIIA, aunque no son B–estables, pueden ser muy eficientes para la resolución numérica de sistemas stiff no lineales. También se deduce que el esquema iterativo desarrollado es fiable y eficiente, siendo competitivo con otros códigos conocidos.

Capítulo **III**

# Implementación de métodos RK de cuatro etapas implícitas

.

# Capítulo III

# Implementación de los métodos RK de cuatro etapas implícitas

## III.1 Introducción

Como hemos mencionado en el capítulo anterior, algunos métodos implícitos tales como Radau IIA, Gauss o Lobatto IIIA poseen excelentes propiedades de estabilidad y convergencia para problemas stiff [11], [19], [37]. Sin embargo, el costo computacional que conlleva la resolución de sus ecuaciones de etapa ha limitado seriamente su uso.

La forma estándar de resolver las ecuaciones de etapa (1.2.2) es a través de la *iteración de Newton Simplificada* [16] que responde al siguiente esquema iterativo:

$$(I \otimes I - h(A \otimes J))(Y^{(\nu+1)} - Y^{(\nu)}) = D(Y^{\nu}), \quad \nu = 0, 1, \ldots, \tag{3.1.1}$$

donde el residual $D(Y)$ está definido por

$$D(Y) := -Y + e \otimes y_n + h(A \otimes I)F(Y).$$

Aquí $J$ es una aproximación a la matriz jacobiana $\partial f/\partial y(t, y)$ en algún punto intermedio, normalmente el punto $(t_n, y_n)$. El principal inconveniente del esquema (3.1.1) es que involucra, en cada paso de integración (o cada 2 ó 3 pasos de integración) la factorización LU de la matriz de orden $ms$ $(I \otimes I - h(A \otimes J))$, cuyo costo en función del número de etapas $s$ varía como $s^3$. A efectos de reducir el costo en la resolución de (3.1.1), Hairer y Wanner en su código RADAU5 [37], emplean una tranformación de semejanza para la matriz $A$, que fue propuesta independientemente por Butcher [9] y Bickart [2], de tal modo que la factorización LU para la resolución de los sistemas lineales de orden $ms$ en (3.1.1) ($s = 3$ en RADAU5) es reemplazada por la factorización de dos matrices de dimensión $m$, una de ellas con valores complejos. De este modo el costo por factorizaciones LU se reduce por un factor de 5 aproximadamente como se ve en [37, pág. 122].

Otros tipos de esquemas iterativos que han recibido notable atención en los últimos años, y que pueden considerarse una alternativa viable a la iteración de Newton Simplificada son los de tipo Single–Newton, los cuales fueron ya introducidos en el capítulo II para los métodos Lobatto IIIA y responden a la siguiente formulación general:

$$\begin{aligned} (I \otimes I - \gamma I \otimes hJ)E^{(\nu)} &= (B \otimes I)D(Y^{(\nu)}) + (L \otimes I)E^{(\nu)}, \\ Y^{(\nu+1)} &= Y^{(\nu)} + (S \otimes I)E^{(\nu)}, \quad \nu = 0, 1, \ldots, \end{aligned} \tag{3.1.2}$$

donde $\gamma > 0$ y $L$, $B$ y $S$ son matrices constantes adecuadas de dimensión $s$, con $L$ triangular inferior estricta. De lo anterior se desprende que sólo se necesita una factorización LU de la matriz $(I - h\gamma J)$ (de dimensión $m$) independientemente del número de etapas $s$ del método RK considerado.

Se han desarrollado esquemas eficientes de tipo Single–Newton para los métodos Gauss, Radau IIA y Lobatto IIIA con 2 y 3 etapas implícitas en [17], [18], [29], [30], y también para el RK Gauss de 4 etapas en [17]. Este tipo de esquemas comparado con la iteración de Newton Simplificada presenta la ventaja de que sólo necesita una factorización LU de orden $m$ por paso de integración, pero presenta el inconveniente, como ha sido probado en [15], de que tiene una velocidad de convergencia más baja que aquella cuando el tamaño de paso $h$ tiende a cero.

En este capítulo pretendemos obtener información precisa del comportamiento de los esquemas de tipo Single–Newton como alternativa práctica a la iteración de Newton Simplificada, cuando se implementan en métodos Runge–Kutta tales como Gauss, Radau IIA, Lobatto IIIA, etc, considerando un número más elevado de etapas implícitas. Obsérvese que la iteración (3.1.2) es algebraicamente equivalente a

$$(I \otimes I - h(T \otimes J))(Y^{(\nu+1)} - Y^{(\nu)}) = (P \otimes I)D(Y^{\nu}), \quad \nu = 0, 1, \ldots, \tag{3.1.3}$$

donde $T = \gamma S(I - L)^{-1}S^{-1}$ y $P = \gamma^{-1}TSB$. Por lo tanto, la elección particular

$$B = (I - L)S^{-1},$$

que posee varias ventajas como se indica en [28], [29], nos da el esquema

$$(I \otimes I - h(T \otimes J))(Y^{(\nu+1)} - Y^{(\nu)}) = D(Y^{\nu}), \;\; \nu = 0, 1, \ldots. \tag{3.1.4}$$

Estas ecuaciones pueden verse como una aproximación a (3.1.1) donde la matriz $A$ se ha reemplazado por la matriz $T$ en el lado izquierdo de la ecuación (3.1.1). De este modo, cuando el número de etapas aumenta, la "distancia" entre $A$ y $T$ también aumenta, ya que la matriz $A$ para los métodos implícitos considerados posee un espectro multipuntual incluyendo pares de complejos conjugados (todos ellos distintos entre sí) y a lo más un autovalor real, mientras que la matriz $T$ posee espectro real unipuntual. Consecuentemente se espera una reducción de la velocidad de convergencia de los esquemas Single–Newton cuando el número de etapas $s$ aumenta.

El propósito de este capítulo es desarrollar esquemas de tipo Single–Newton eficientes para la implementación de métodos Runge–Kutta con cuatro etapas implícitas y compararlos, desde un punto de vista práctico, con la iteración de Newton Simplificada. En particular, estamos interesados en esta clase de esquemas para el RK Gauss y el Lobatto IIIA de orden 8, y para el Radau IIA de orden 7. Debe recordarse que los métodos de alto orden son más eficientes que los de bajo orden para integrar problemas donde se requiera una precisión alta. Por tanto, se hace necesario minimizar el costo computacional del esquema iterativo usado para resolver las ecuaciones de etapa (generalmente sistemas algebraicos no lineales).

No se ha realizado mucha experimentación con métodos Runge–Kutta implícitos de alto orden, de manera que podamos conocer en qué condiciones el alto orden de convergencia compensará el alto coste computacional que estos métodos conllevan. Sólo podemos mencionar el artículo reciente de Hairer y Wanner [38], donde se presenta un código de orden variable que implementa las fórmulas de Radau IIA de 3, 5 y 7 etapas (de órdenes 5, 9 y 13 respectivamente). En este capítulo también pretendemos comparar el comportamiento de los métodos Radau IIA de órdenes 5 y 7, usando esquemas Single–Newton para resolver sus ecuaciones de etapa.

El resto del capítulo queda organizado como sigue. En la sección III.2 comparamos el coste computacional involucrado en la iteración de Newton Simplificada y en los esquemas Single–Newton, mostrando que el Single–Newton es en general menos costoso. En la sección III.3, siguiendo las ideas desarrolladas en [29], obtenemos esquemas eficientes de tipo Single–Newton para los RK de 4 etapas implícitas, Gauss, Radau IIA y Lobatto IIIA. En la sección III.4 realizamos algunos experimentos numéricos comparando la eficiencia de los nuevos esquemas propuestos y la iteración de Newton Simplificada.

La investigación realizada en el presente capítulo de esta memoria, se encuentra también recogida en [31].

## III.2 Costo computacional de los esquemas iterativos

Consideremos un método Runge–Kutta con cuatro etapas implícitas y con una matriz de coeficientes $A$ regular. Supongamos que sus ecuaciones de etapa se resuelven mediante la iteración de Newton Simplificada (3.1.1). Con el fin de reducir el costo computacional consideremos una técnica similar a la usada por Hairer y Wanner en RADAU5 [37]. Entonces, si $Q$ es una matriz regular que transforma $A^{-1}$ en una matriz diagonal por bloques (los bloques son de dimensión 2 como máximo), $\Lambda = Q^{-1}A^{-1}Q$, entonces pre-multiplicando (3.1.1) por $A^{-1}\otimes I$ y denotando $W^{(\nu)} = (Q^{-1}\otimes I)(Y^{(\nu)} - e\otimes y_n)$ tenemos que

$$\begin{aligned} [\Lambda\otimes I - h(I\otimes J)]\Delta W^{(\nu)} &= R^{(\nu)}, \quad \nu = 0,1,\dots \\ R^{(\nu)} &= -(\Lambda\otimes I)W^{(\nu)} + (Q^{-1}\otimes I)F\Big((Q\otimes I)W^{(\nu)} + e\otimes y_n\Big) \\ W^{(\nu+1)} &= W^{(\nu)} + \Delta W^{(\nu)}. \end{aligned} \tag{3.2.1}$$

Para los métodos Gauss y Radau IIA, la matriz $A$ tiene dos pares de autovalores complejos conjugados y $\Lambda$ tiene la forma $\Lambda = \mathrm{diag}(\Lambda_1, \Lambda_2)$ con

$$\Lambda_j = \begin{pmatrix} \alpha_j & -\beta_j \\ \beta_j & \alpha_j \end{pmatrix}, \qquad j = 1,2.$$

Ahora bien, poniendo

$$R^{(\nu)} = (R_1^{(\nu)}, R_2^{(\nu)}, R_3^{(\nu)}, R_4^{(\nu)})^T,$$
$$\Delta W^{(\nu)} = (\Delta W_1^{(\nu)}, \Delta W_2^{(\nu)}, \Delta W_3^{(\nu)}, \Delta W_4^{(\nu)})^T,$$

la ecuación (3.2.1) se puede desacoplar en dos sistemas lineales complejos de dimension $m$

$$\begin{aligned} [(\alpha_1 + i\beta_1)I - hJ](\Delta W_1^{(\nu)} + i\Delta W_2^{(\nu)}) &= R_1^{(\nu)} + iR_2^{(\nu)} \\ [(\alpha_2 + i\beta_2)I - hJ]\,(\Delta W_3^{(\nu)} + i\Delta W_4^{(\nu)}) &= R_3^{(\nu)} + iR_4^{(\nu)}. \end{aligned} \tag{3.2.2}$$

Consideremos en primer lugar el costo computacional para factorizar las dos matrices complejas $(\alpha_j + i\beta_j)I - hJ$, $j = 1,2$. Sólo consideraremos las operaciones de punto flotante involucradas en sumas, productos y divisiones, asumiendo que todas ellas conllevan un costo computacional equivalente. Tampoco tendremos en cuenta el efecto que produce el uso de aritmética compleja en la propagación de errores de redondeo (para un estudio reciente del uso de aritmética de punto flotante ver por ejemplo [39]), ni tampoco consideraremos el costo de evaluar la matriz jacobiana $J$ ni la evaluación de la función derivada $f$.

| Esquema iterativo | $s=3$ | $s=4$ | $s=5$ | $s=6$ |
|---|---|---|---|---|
| Newton Simplificado | $10m^2+48m$ | $16m^2+86m$ | $18m^2+123m$ | $24m^2+177m$ |
| Single–Newton | $6m^2+35m$ | $8m^2+63m$ | $10m^2+99m$ | $12m^2+143m$ |

Tabla 3.2.1: Flops/iter. en Newton Simplificado y Single–Newton.

Teniendo en cuenta que una multiplicación compleja conlleva 4 multiplicaciones reales y 2 sumas reales, que una suma compleja involucra 2 sumas reales y que una división compleja es equivalente a 8 productos (o divisiones) reales y 3 sumas, entonces el cómputo total para las dos factorizaciones complejas LU arroja aproximadamente $16m^3/3$ flops ("floating point operations"), para ser exactos $(16m+25)m(m-1)/3$ flops.

Estudiemos ahora el número de operaciones involucradas en cada iteración. Primero, para resolver los dos sistemas complejos (3.2.2) de dimensión $m$ con la matriz previamente factorizada, ya que cada sistema necesita $m$ divisiones, $m(m-1)$ productos y $m(m-1)$ sumas complejas, entonces el número total será de $16m^2+6m$ flops.

Por otra parte, en cada iteración debe computarse el residual $R^{(\nu)}$ en (3.2.1), lo cual, si la matriz $Q$ es llena $(4\times 4)$, requiere $76m$ flops. Entonces, en total, cada iteración necesita $16m^2+86m$ flops (obsérvese que se necesitan $4m$ sumas para actualizar $W^{(\nu)}$).

El costo computacional final estará afectado en cada parte de los cálculos en un modo diferente dependiendo de la dimensión $m$ del sistema diferencial y del número de iteraciones que necesite el esquema iterativo para alcanzar la convergencia. Así, si $m$ es grande, el costo de las factorizaciones LU será el término dominante, debido al factor $m^3$. Suponiendo por ejemplo un número medio de 4 iteraciones por paso de integración y un valor de $m=100$, el número de flops involucrado en las iteraciones sería de $4(16\times 100^2+86\times 100)=674400$. Esto sería aproximadamente 1/8 del costo de las dos factorizationes LU. Por otra parte, si comparamos la influencia de la evaluación del residual y la solución de los sistemas triangulares en cada iteración para $m=100$, resulta que el costo de la evaluación del residual es 7600 flops, lo que es despreciable frente al costo de las factorizationes LU, y aproximadamente 1/21 del costo de la solución de los sistemas lineales (160600 flops). Para $m=10$, el costo de las factorizationes LU será de 5550 flops, el costo de la resolución de los sistemas 1660 flops y el costo de evaluación del residual 760 flops. Por tanto, si suponemos 4 iteraciones por paso de integración, ambos factores (las factorizationes LU y el costo de las iteraciones) tienen una influencia similar en el costo total del esquema iterativo. Finalmente, para valores de $m$ pequeños, resulta claro que el costo computacional se verá afectado principalmente por los cálculos en las iteraciones.

Siguiendo un razonamiento análogo para métodos de $s$ etapas implícitas de alto orden (Gauss, Radau IIA, Lobatto IIIA, etc.) donde la matriz $A$ si $s$ es par tiene $s/2$ pares de autovalores complejos conjugados, y si es impar, tiene un autovalor real y $(s-1)/2$ pares de autovalores complejos conjugados, obtenemos el número de flops por iteración para el Newton Simplificado, que es:

$$\text{si } s \text{ es par}: \quad 4sm^2+\frac{8s^2+11s}{2}m \quad \text{flops/iter.},$$

$$\text{si } s \text{ es impar}: \quad (4s-2)m^2+\frac{8s^2+11s-9}{2}m \quad \text{flops/iter.}$$

Así obtenemos en la tabla 3.2.1 el número total de flops que conlleva la iteración de Newton Simplificada cuando se aplica a métodos Runge–Kutta de 3, 4, 5 y 6 etapas implícitas.

Consideremos ahora el esquema Single–Newton (3.1.2) con $B=(I-L)S^{-1}$. Resulta claro

que cada factorización LU requerirá un número de operaciones de orden $2m^3/3$ (exactamente $2m^2(m-1)/3$ flops), independientemente del número de etapas $s$. Para $s=4$, esto supone 1/8 del costo empleado en la iteración de Newton Simplificada. Por tanto, el esquema Single–Newton es más eficiente cuando la dimensión $m$ es grande.

Llamando $Z^{(\nu)} = Y^{(\nu)} - e \otimes y_n$ y $W^{(\nu)} = (S^{-1} \otimes I)Z^{(\nu)}$, el esquema Single–Newton (3.1.2) puede reescribirse en la forma siguiente

$$\begin{aligned} G^{(\nu)} &= -W^{(\nu)} + h(S^{-1}A \otimes I)\, F\left(e \otimes y_n + (S \otimes I)W^{(\nu)}\right) \\ (I \otimes I - \gamma I \otimes hJ)E^{(\nu)} &= G^{(\nu)} + (L \otimes I)(E^{(\nu)} - G^{(\nu)}) \\ W^{(\nu+1)} &= W^{(\nu)} + E^{(\nu)}, \quad \nu = 0, 1, \ldots \end{aligned} \tag{3.2.3}$$

que es más interesante en la práctica porque se evita un producto por la matriz $S^{-1} \otimes I$, esto es, podemos reducir $s(s-1)m$ flops por iteración ya que, como veremos más tarde para los métodos Runge–Kutta de interés, la matriz $S$ puede elegirse triangular superior con "1" en la diagonal principal.

De (3.2.3) se sigue que el número de flops requeridos en cada iteración para los esquemas Single–Newton (aplicados a métodos Runge–Kutta implícitos de $s$ etapas) es de $2sm^2 + (4s^2 - 1)m$. Los valores resultantes para $s = 3, 4, 5$ y 6 también están representados en la tabla 3.2.1. Se ve claramente que, para el mismo número de etapas, el costo por iteración en los Single–Newton es menor que en Newton Simplificado. Además, para valores grandes de $m$ la razón entre ambos costos se aproxima a 2 para $s = 4, 6$, y a $5/3, 9/5$ y $13/7$ para $s = 3, 5$ y 7 respectivamente. Por otra parte, es necesario resaltar que para una comparación realista de ambos esquemas iterativos debería considerarse también el número de iteraciones que invierte cada esquema en alcanzar la convergencia.

Las consideraciones anteriores indican que para problemas de dimensión elevada se espera que los esquemas Single–Newton sean bastante más eficientes que Newton Simplificado. Para problemas de dimensión media también pueden ser más eficientes si no necesitan un número elevado de iteraciones para converger (menos del doble que Newton Simplificado). Finalmente, para problemas de dimensión baja la eficiencia de ambos esquemas dependerá fuertemente del número de iteraciones que necesiten para converger cada uno de los esquemas.

**Nota III.2.1** *Hay que observar que el estudio anterior cuenta el número de operaciones de punto flotante independientemente del ordenador que se use. Sin embargo, en la práctica, una parte importante del tiempo de CPU que consumen los códigos proviene de la forma en la que el compilador los optimiza, beneficiándose así de las características particulares del ordenador utilizado. Por lo tanto, no siempre un mayor número de operaciones de punto flotante implica un tiempo de CPU más alto. Un ejemplo interesante de tal optimización es el de los ordenadores con procesadores DEC Alpha de 64 bits, donde las operaciones de aritmética compleja pueden realizarse de forma particularmente optimizada, mejorando considerablemente la eficiencia de los códigos que las usan.*

## III.3 Construcción de esquemas Single–Newton para RK con 4 etapas implícitas

En esta sección desarrollaremos esquemas de tipo Single–Newton (3.1.2) con matriz $B = (I - L)S^{-1}$, para los RK Gauss, Radau IIA y Lobatto IIIA de 4 etapas implícitas. Consecuentemente,

estamos interesados en encontrar matrices $S$, $L$ (triangular inferior estricta) y un parámetro positivo $\gamma$ tal que el esquema resultante sea lo más eficiente posible.

Cuando aplicamos un esquema Single–Newton al test escalar $y' = \lambda y$, los errores en la iteraciones satisfacen

$$Y^{(\nu+1)} - Y = M(z)(Y^{(\nu)} - Y) = M(z)^{\nu}(Y^{(0)} - Y), \tag{3.3.1}$$

donde $z = h\lambda$ y la matriz de amplificación $M(z)$ está dada por

$$M(z) = z(I - zT)^{-1}(A - T), \qquad T = \gamma S(I - L)^{-1}S^{-1}. \tag{3.3.2}$$

De acuerdo con las ideas presentadas en [29], deseamos obtener una matriz $T$ tal que:

$$\left.\begin{array}{ll} \text{(P1)} & T \text{ posee espectro unipuntual positivo, } \gamma > 0, \\ \text{(P2)} & \rho(M(\infty)) = 0, \\ \text{(P3)} & b^T A^{-1} M(\infty) = 0, \\ \text{(P4)} & \max\{\rho(M(z)) | z \in \mathbb{R}^-\} \text{ sea mínimo}, \\ \text{(P5)} & \text{la "distancia" entre } A \text{ y } T \text{ sea lo más pequeña posible.} \end{array}\right\} \tag{3.3.3}$$

La condición (P1) se exige a fin de que la matriz $T$ satisfaga la relación expresada en (3.3.2). La condición (P3) garantiza que la función de amplificación asociada a la iteración $\nu$–ésima para la solución de avance del método Runge–Kutta

$$y_{n+1}^{(\nu)} = (1 - b^T A^{-1} e) y_n + (b^T A^{-1} \otimes I) Y^{(\nu)}, \quad \nu = 0, 1, ..., \tag{3.3.4}$$

aproxime a la función de estabilidad del método Runge–Kutta considerado y además coincidan cuando $z \to \infty$ [29, Teorema 2.1]. Finalmente, las condiciones (P2), (P4) y (P5) están motivadas por la optimización de la velocidad de convergencia del esquema iterativo sobre problemas lineales (ver [29] y capítulo II de esta memoria).

Para transformar las condiciones (P1) y (P2) en expresiones algebraicas más manejables y dar de paso una forma más práctica para obtener la matriz $T$, demostraremos los siguientes lemas (supondremos en adelante que las matrices $A$ y $T$ son regulares y de orden 4):

**Lema III.3.1** *$T$ satisface* (P1) *si y sólo si*

$$\begin{array}{ll} \det(T) = \gamma^4, & \operatorname{tr}(T) = 4\gamma, \\ \operatorname{tr}(T^2) = 4\gamma^2, & \operatorname{tr}(T^{-1}) = 4/\gamma \quad \acute{o} \quad \operatorname{tr}(T^3) = 4\gamma^3. \end{array}$$

La prueba de este lema se deduce inmediatamente del teorema de Cayley-Hamilton (para matrices).

**Lema III.3.2** *Si $A$ y $T$ satisfacen* (P1) *y* (P2)*, entonces*

$$\gamma = (\det A)^{(1/4)} \quad y \quad \det(A - T) = 0.$$

*Demostración.* Como $M(\infty) = T^{-1}(T - A) = (I - T^{-1}A)$, entonces (P2) implica que $\det(T - A) = 0$. Además, usando (P2) y la descomposición de Schur de $M(\infty)$, se sigue que existe una matriz ortogonal $U$ tal que

$$I - T^{-1}A = U^{-1}\Delta U,$$

siendo $\Delta$ una matriz triangular superior estricta. De aquí, despejando $A$, tomando determinantes y usando (P1) se concluye que $\det(A) = \gamma^4$. ∎

**Lema III.3.3** *Asumiendo* (P1) *y que* $\det(A-T)=0$*, se tiene que los autovalores de* $M(z)$ *son las raíces de la ecuación*

$$\mu^4 - b_3\mu^3 + b_2\mu^2 - b_1\mu = 0, \tag{3.3.5}$$

*donde*

$$\begin{array}{ll} b_1 = q(p-2) - 2 + b_3 & q = \det(I - zA)/(1-\gamma z)^4 \\ b_2 = q(p-1) - 3 + 2b_3 & p = \mathrm{tr}(Q(z)) \\ b_3 = 4 - \mathrm{tr}(Q(z)^{-1}) & Q(z) = (I - zA)^{-1}(I - zT). \end{array} \tag{3.3.6}$$

<u>*Demostración.*</u> Como $\det(A-T)=0$, se tiene $\det(M(z))=0$, y por tanto al menos uno de los autovalores de $M(z)$ se anula. Desarrollando la función

$$\phi(\mu) := \det(\mu I - M(z))/\mu = (\mu-\lambda_1)(\mu-\lambda_2)(\mu-\lambda_3),$$

(aquí $\lambda_j$, $j=1,2,3$, denotan los autovalores no nulos de $M(z)$) obtenemos que $\phi(\mu) = \mu^3 - b_3\mu^2 + b_2\mu - b_1$, donde

$$\begin{aligned} b_1 &= \lambda_1\lambda_2\lambda_3 \\ b_2 &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3 \\ b_3 &= \lambda_1 + \lambda_2 + \lambda_3. \end{aligned}$$

No es difícil ver que usando la función $\phi$, podemos escribir la fórmula alternativa

$$\begin{aligned} b_1 &= \phi'(1) - \phi(1) - 2 + b_3, \\ b_2 &= \phi'(1) - 3 + 2b_3, \\ b_3 &= \mathrm{tr}(M(z)). \end{aligned}$$

Por otra parte, usando (P1) es fácil comprobar que

$$\mu\phi(\mu) = \det(\mu I - M(z)) = \det\Big(\mu I - z(A + (\mu-1)T)\Big)/(1-z\gamma)^4.$$

De la expresión anterior tras unos cálculos sencillos se deduce

$$\begin{aligned} \phi(1) &= q \\ \phi'(1) &= \lim_{\epsilon\to 0} \frac{\phi(1+\epsilon) - \phi(1)}{\epsilon} = q(p-1). \end{aligned}$$

En virtud de que $M(z) = I - Q(z)^{-1}$, resulta claro que $b_3 = 4 - \mathrm{tr}(Q(z)^{-1})$. ∎

**Lema III.3.4** *Si* $T$ *satisface* (P1)*, entonces la condición* (P2) *es equivalente a*

$$\gamma = (\det(A))^{1/4}, \quad \det(A-T)=0, \quad \mathrm{tr}(T^{-1}A) = \mathrm{tr}(A^{-1}T) = 4.$$

<u>*Demostración.*</u> Haciendo $z\to\infty$, la ecuación de autovalores para $M(\infty)$ resulta

$$\mu^4 - b_3^*\mu^3 + b_2^*\mu^2 - b_1^*\mu + b_0^* = 0,$$

donde $b_j^* = b_j(\infty)$, $(j=1,2,3)$ (ver (3.3.6)), y $b_0^* = \det(I - T^{-1}A) = \gamma^{-4}\det(T-A)$. Ahora bien, del lema anterior se sigue que

$$b_0^* = 0 \Longleftrightarrow \det(T-A) = 0; \qquad b_3^* = 0 \Longleftrightarrow \mathrm{tr}(T^{-1}A) = 4;$$

$$b_2^* = b_1^* = 0 \iff q(\infty) = 1,\ p(\infty) = 4 \iff \gamma = (\det(A))^{1/4},\ \mathrm{tr}(A^{-1}T) = 4.$$

∎

El siguiente resultado permite expresar la condición (P4) de una forma más adecuada.

**Teorema III.3.1** *Sea $A$ una matriz regular dada, y supongamos que $T$ denota una matriz cualquiera de modo que se verifica* (P1) *y* (P2). *Entonces, la condición de minimización* (P4) *se consigue para ciertas constantes $\Gamma_1^*, \Gamma_2^*$ y $\Gamma_3^*$ dependientes sólo de la matriz $A$ y de tal modo que*

$$\mathrm{tr}(AT) = \Gamma_1^*, \quad \mathrm{tr}(A^{-2}T) = \Gamma_2^*, \quad \mathrm{tr}(T^{-2}A) = \Gamma_3^*.$$

*Demostración.* Desarrollando $Q(z) = (I - zA)^{-1}(I - zT)$ en potencias de $z$ y $z^{-1}$ obtenemos respectivamente

$$Q(z) = I + z(A - T) + z^2A(A - T) + \mathcal{O}(z^3), \quad (z \to 0),$$
$$Q(z) = A^{-1}T + z^{-1}A^{-2}(T - A) + \mathcal{O}(z^{-2}), \quad (z \to \infty).$$

Ahora, usando los Lemas III.3.3 y III.3.4, y después de realizar algunos cálculos sencillos, se sigue que los coeficientes de la ecuación de autovalores (3.3.5) vienen dados por

$$b_1 = \frac{\gamma^4\xi_0 z^3}{(1-\gamma z)^4}, \qquad b_2 = \frac{\xi_2 z^2 + \xi_3 z^3}{(1-\gamma z)^4}, \qquad b_3 = \frac{\eta_1 z + \eta_2 z^2 + \eta_3 z^3}{(1-\gamma z)^4},$$

donde

$$\xi_0 = 12\gamma^{-1} - 3\mathrm{tr}(A^{-1}) + \Gamma_2 - \Gamma_3,$$
$$\xi_2 = \frac{(\mathrm{tr}^2(A) - \mathrm{tr}(A^2))}{2} + 6\gamma^2 - 4\gamma\mathrm{tr}(A) + \Gamma_1, \quad \xi_3 = -4\gamma^4\mathrm{tr}(A^{-1}) + 20\gamma^3 + \gamma^4\Gamma_2 - 2\gamma^4\Gamma_3,$$
$$\eta_1 = \mathrm{tr}(A) - 4\gamma, \quad \eta_2 = 12\gamma^2 - 4\gamma\mathrm{tr}(A) + \Gamma_1, \quad \eta_3 = 4\gamma^3 - \gamma^4\Gamma_3,$$

siendo

$$\Gamma_1 = \mathrm{tr}(AT), \quad \Gamma_2 = \mathrm{tr}(A^{-2}T), \quad \Gamma_3 = \mathrm{tr}(T^{-2}A).$$

Consideremos ahora la función de tres variables

$$g(\Gamma_1, \Gamma_2, \Gamma_3) := \max\{\rho(M(z)),\ z \in \mathbb{R}^-\}.$$

En virtud de que $g$ es una función continua no negativa sobre $\mathbb{R}^3$, que además satisface $g(\Gamma_1, \Gamma_2, \Gamma_3) \to \infty$ cuando algún $\Gamma_j$ $(j = 1, 2, 3)$ tiende a infinito (recuérdese que si en un polinomio con coeficiente director acotado, alguno de sus otros coeficientes tiende a infinito, alguna de sus raíces también tiende a infinito), resulta que $g$ debe poseer un mínimo absoluto que será alcanzado para unos valores, digamos $\Gamma_1^*$, $\Gamma_2^*$ y $\Gamma_3^*$. Claramente, estos valores sólo dependen de la matriz $A$. ∎

Para los métodos Runge–Kutta considerados, el cálculo de las constantes $\Gamma_j^*$ ha sido realizado usando el criterio de Schur-Cohn [51]. Esto requiere ciertos cambios de variable y un proceso de cálculos que es bastante largo y tedioso (nosotros lo hemos realizado auxiliándonos con el manipulador algebraico MAPLE V). En aras de no interrumpir la exposición y de no alargar innecesariamente este capítulo hemos omitido aquí su presentación, pero puede verse el desarrollo de los mismos con detalle en el Apéndice A de esta memoria.

En la tabla 3.3.1 presentamos los valores obtenidos para las constantes $\Gamma_j^*$ al considerar los RK Gauss y Lobatto IIIA (ambos de orden 8) y el RK Radau IIA de orden 7. En dicha tabla también incluimos los valores

$$\rho^0_{\max} = \max\{\rho(M(z)) \mid z \in \mathbb{R}^-\},$$
$$\rho^1_{\max} = \max\{\rho(M(iy)) \mid y \in \mathbb{R}\},$$
$$\rho^2_{\max} = \max\{\rho(M((1-i)y) \mid y \in \mathbb{R}^-\},$$

| | Gauss | Radau IIA | Lobatto IIIA |
|---|---|---|---|
| $\Gamma_1^*$ | 0.06316638299274640 | 0.08296548473447771 | 0.06316638299274640 |
| $\Gamma_2^*$ | 14.86450048645422 | 11.01111697854543 | 14.86450048645422 |
| $\Gamma_3^*$ | 31.73222088578815 | 27.66997861432155 | 31.73222088578815 |
| $\rho_{\max}^0$ | 0.0893204199714 | 0.104708968155 | 0.0893204199714 |
| $\rho_{\max}^1$ | 0.320182072684 | 0.378417643002 | 0.320182072684 |
| $\rho_{\max}^2$ | 0.147383853954 | 0.172953394381 | 0.147383853954 |

Tabla 3.3.1: Constantes asociadas al radio espectral mínimo de $M(z)$ sobre el semieje real negativo.

para cada uno de los tres métodos.

Obsérvese en dicha tabla 3.3.1 que los resultados obtenidos para el Gauss de 4 etapas y el Lobatto IIIA de 5 etapas son idénticos. Esto se debe a que la matriz $A$ de coeficientes del Gauss y la submatriz $\bar{A}$ del Lobatto IIIA (ver subsección III.3.3) son semejantes y las condiciones exigidas sólo dependen de los determinantes y las trazas de dichas matrices.

En virtud de los resultados anteriores, y teniendo en cuenta que (P3) implica que $\det(A-T) = 0$, las condiciones (P1) a (P5) son equivalentes a

$$
\begin{aligned}
&\text{(Q1)} \quad \det(T) = \gamma^4 = \det(A),\\
&\text{(Q2)} \quad \operatorname{tr}(T) = 4\gamma,\\
&\text{(Q3)} \quad \operatorname{tr}(T^2) = 4\gamma^2,\\
&\text{(Q4)} \quad \operatorname{tr}(T^{-1}) = 4/\gamma \ \text{ ó } \ \operatorname{tr}(T^3) = 4\gamma^3,\\
&\text{(Q5)} \quad \operatorname{tr}(A^{-1}T) = 4,\\
&\text{(Q6)} \quad \operatorname{tr}(T^{-1}A) = 4,\\
&\text{(Q7)} \quad \operatorname{tr}(AT) = \Gamma_1^*,\\
&\text{(Q8)} \quad \operatorname{tr}(A^{-2}T) = \Gamma_2^*,\\
&\text{(Q9)} \quad \operatorname{tr}(T^{-2}A) = \Gamma_3^*,\\
&\text{(Q10)} \quad b^T A^{-1}(I - T^{-1}A) = 0,\\
&\text{(Q11)} \quad \text{la “distancia” entre } A \text{ y } T \text{ es lo más corta posible.}
\end{aligned}
$$

Tenemos que determinar los 16 coeficientes de la matriz $T$ de tal modo que se satisfagan las 13 ecuaciones expresadas en las condiciones (Q1) a (Q10). De este modo quedan 3 parámetros libres que deben ajustarse para satisfacer la condición (Q11). Para ello hemos considerado como “distancia” la cantidad

$$D_{AT} = \nu_1 |I - A^{-1}T|_F^2 + \nu_2 |I - TA^{-1}|_F^2 + \nu_3 |e_4^T (A - T)|_2^2, \quad e_4^T = (0,0,0,1),$$

donde $|\cdot|_F$ es la norma de Fröbenius y $\nu_i$ son pesos que deberán ser elegidos de modo que el esquema iterativo resultante dé los mejores resultados sobre los problemas stiff del paquete DETEST [23], teniendo al mismo tiempo buenas propiedades de estabilidad para la iteraciones correspondientes a la solución de avance resultante.

Debido al elevado número de ecuaciones y parámetros involucrados en el proceso, hemos utilizado el manipulador algebraico MAPLE V y paquetes de FORTRAN (librerías IMSL) procediendo de la siguiente forma.

- Primero, obtenemos una matriz inicial $T_0$ que verifica las condiciones (Q1) a (Q10).
- Luego fijamos un conjunto de valores $\nu_1$, $\nu_2$ y $\nu_3$.

- Para cada uno de estos valores $\nu_i$, $1 \leq i \leq 3$, y comenzando con $T_0$ como aproximación inicial, conseguimos una matriz $T$ que minimiza $D_{AT}$ y verifica las condiciones (Q1) a (Q10).
- Entre todas las matrices $T$ obtenidas, seleccionamos aquella de modo que la iteración correspondiente (3.1.4) tenga buenas propiedades de estabilidad (es decir, que la solución de avance dada por la iteración $\nu$–ésima (3.3.4) tenga regiones de estabilidad lo más parecidas posible a la región de estabilidad del método RK considerado) y dé los mejores resultados en general sobre los problemas stiff del paquete DETEST [23].

### III.3.1 RK Gauss de orden 8

Para este método hemos obtenido el mejor esquema Single–Newton para los valores $\nu_1 = 1$, $\nu_2 = \nu_3 = 0$. Los valores computados en este caso, con 16 cifras significativas, son:

$$D_{AT} = 0.2186117957792797, \quad |A - T|_F = 0.1866324635411617.$$

con matriz $T = (t_{ij})$:

$$\begin{array}{ll}
t_{11} = 0.07056898453975971 & t_{31} = 0.1097496189565937 \\
t_{12} = -0.01381201242940272 & t_{32} = 0.3953973119562834 \\
t_{13} = 0.01374509656255927 & t_{33} = 0.2550102453783648 \\
t_{14} = 0.001273397980705694 & t_{34} = -0.03800926472551498 \\
t_{21} = 0.1359096681314922 & t_{41} = 0.1026795079784531 \\
t_{22} = 0.2039916522067102 & t_{42} = 0.3643735550837732 \\
t_{23} = 0.01953742322502287 & t_{43} = 0.4333395062278329 \\
t_{24} = -0.007041562052392658 & t_{44} = 0.09521710525921647\ .
\end{array}$$

Para esta matriz es posible obtener una triangular superior

$$S = \begin{pmatrix}
1 & -0.6677448107835342 & 0.1296306965460327 & 0.01526277075698497 \\
0 & 1 & -0.2153491783691625 & 0.07296098377515141 \\
0 & 0 & 1 & 0.07575507029183779 \\
0 & 0 & 0 & 1
\end{pmatrix}$$

que transforma $T$ en una matriz triangular semejante. Con $T$ y $S$, la matriz $L = I - \gamma S^{-1} T^{-1} S$ es

$$L = \begin{pmatrix}
0 & 0 & 0 & 0 \\
0.9627423789846739 & 0 & 0 & 0 \\
-1.194428300588649 & 1.918753137082504 & 0 & 0 \\
1.649572580382698 & -2.628995768624925 & 2.357166809194904 & 0
\end{pmatrix}.$$

Luego, las matrices $L$ y $S$ junto con

$$\gamma = (\det(A))^{1/4} = 0.1561969968460128$$

nos dan los parámetros del esquema Single–Newton.

Al igual que para los esquemas estudiados en el capítulo II (ver (2.3.3)), también aquí hemos encontrado una matriz $D$ diagonal definida positiva

$$D = \mathrm{diag}(1, 0.4207246, 0.260849252, 0.046279706)$$

que nos da una medida de la razón de convergencia del esquema no lineal asociado (2.3.2) sobre problemas no lineales contractivos, a saber,

$$\nu = 2|N^{-T} D (T - A) N^{-1}|_2 = 0.829777178...$$

### III.3.2 RK Radau IIA de orden 7

En este caso, el mejor esquema Single–Newton fue obtenido para $\nu_1 = \nu_2 = 1$ y $\nu_3 = 100$, dando los valores

$$D_{AT} = 1.334357031843827, \quad |A - T|_F = 0.1198832704310612.$$

Las matrices correspondientes $T = (t_{ij})$, $S$ y $L$ vienen dadas por:

$$\begin{array}{ll} t_{11} = 0.1187824099582517 & t_{31} = 0.2267733612906856 \\ t_{12} = 0.01022763543870539 & t_{32} = 0.4394654798955388 \\ t_{13} = 0.02251934010521350 & t_{33} = 0.2423196391476349 \\ t_{14} = -0.002140831122870532 & t_{34} = -0.01672793262894805 \\ t_{21} = 0.2463531839329877 & t_{41} = 0.2303363939912873 \\ t_{22} = 0.2880948365341910 & t_{42} = 0.4140965520644702 \\ t_{23} = -0.02947965404901304 & t_{43} = 0.3882107808506906 \\ t_{24} = -0.002392091968997757 & t_{44} = 0.09380543432526635\ , \end{array}$$

$$S = \begin{pmatrix} 1 & -0.3746257695117888 & 0.07689675270074446 & 0.04190406032755296 \\ 0 & 1 & 0.05051271922734543 & -0.01257194014862304 \\ 0 & 0 & 1 & 0.2253907333361419 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$L = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1.294297023384814 & 0 & 0 & 0 \\ -1.014023314466600 & 1.510766557167087 & 0 & 0 \\ 1.286041959197947 & -1.706853680903114 & 2.297920385846297 & 0 \end{pmatrix}.$$

Estas matrices $L$ y $S$ junto con

$$\gamma = (\det(A))^{1/4} = 0.1857505799913360$$

determinan el esquema Single–Newton.

En este caso también hemos encontrado la matriz $D$ diagonal definida positiva

$$D = \mathrm{diag}(1, 0.75395702, 0.46923933, 0.10190215)$$

que nos da el mínimo valor

$$\nu = 0.943202598...$$

### III.3.3 RK Lobatto IIIA de orden 8

En el caso del método Lobatto IIIA, la primera fila de la matriz de coeficientes $A$ es nula. Entonces, denotando

$$A = \begin{pmatrix} 0 & \mathbf{0}^T \\ w & \bar{A} \end{pmatrix}$$

la ecuación (2.2.3) puede reescribirse

$$\bar{Y} = e \otimes y_n + h(w \otimes f(t_n, y_n)) + h(\bar{A} \otimes I)F(\bar{Y}),$$

donde $\bar{Y} = (Y_2, Y_3, Y_4, Y_5)^T$. De este modo, el sistema (2.2.3) se reduce a un sistema de $4m$ ecuaciones que puede estudiarse como en los casos anteriores, pero tomando ahora la submatriz $\bar{A}$ en lugar de la matriz $A$.

Esta vez, el mejor esquema Single–Newton se obtuvo para $\nu_1 = 1$, y $\nu_2 = \nu_3 = 0$, resultando

$$D_{\bar{A}T} = 0.07066885046321667, \quad |\bar{A} - T|_F = 0.09799263272854765.$$

La matriz $T = (t_{ij})$ para este método es

$$\begin{array}{ll} t_{11} = 0.1205065476893790 & t_{31} = 0.2675367041374556 \\ t_{12} = -0.001249676535040056 & t_{32} = 0.4217726039803753 \\ t_{13} = 0.003900830554640007 & t_{33} = 0.1739257710023307 \\ t_{14} = -0.0006329622087931463 & t_{34} = 0.009132813977995455 \\ t_{21} = 0.3079578502684815 & t_{41} = 0.2775596403310148 \\ t_{22} = 0.2327971369316140 & t_{42} = 0.3986701534245386 \\ t_{23} = -0.02614746695545937 & t_{43} = 0.2894006793595838 \\ t_{24} = 0.006158162143340951 & t_{44} = 0.09755853176072735 \end{array}$$

y la matriz triangular superior $S$ que transforma a $T$ en una matriz triangular semejante, y la matriz $L$ vienen dadas por,

$$S = \begin{pmatrix} 1 & -0.1345492788488319 & -0.7907579166890781E-3 & 0.01048164212642994 \\ 0 & 1 & 0.1654189391431284 & -0.03863351412430941 \\ 0 & 0 & 1 & 0.2457879968605093 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

y

$$L = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1.829166626367437 & 0 & 0 & 0 \\ -2.201612484488081 & 1.901230267943492 & 0 & 0 \\ 2.551217615151542 & -2.009365789995880 & 2.273595510125324 & 0 \end{pmatrix}.$$

Así, el esquema Single–Newton queda determinado por las matrices $L$ y $S$ y por el parámetro

$$\gamma = (\det(\bar{A}))^{1/4} = 0.1561969968460128$$

Finalmente, obtenemos la matriz $D$ diagonal definida positiva

$$D = \mathrm{diag}(1, 0.40505882, 0.1458211752, 0.0324047056)$$

que nos da la medida de la razón de convergencia del esquema sobre problemas no lineales contractivos (ver (2.3.3))

$$\nu = 2|N^{-T} D (T - \bar{A}) N^{-1}|_2 = 0.915550194...$$

## III.4 Experimentos Numéricos

En esta sección presentamos los resultados obtenidos al realizar algunos experimentos numéricos con objeto de mostrar el comportamiento de los esquemas Single–Newton y Newton Simplificado cuando se usan para resolver las ecuaciones de etapa de las fórmulas Radau IIA de órdenes 5 y 7 respectivamente.

Para comparar el rendimiento de los algoritmos hemos implementado los siguientes cuatro códigos a paso variable:

- R5SN basado en Radau IIA de orden 5 + Single–Newton,
- R5SNC basado en Radau IIA de orden 5 + Newton Simplificado,
- R7SN basado en Radau IIA de orden 7 + Single–Newton,
- R7SNC basado en Radau IIA de orden 5 + Newton Simplificado.

Estas implementaciones, no pensadas para competir con códigos estándar, estiman el error local por la técnica clásica de extrapolación de Richardson. Además, en su programación se han seguido fielmente los pasos indicados en el capítulo II de esta memoria (sección II.4).

Presentamos aquí los resultados obtenidos con los tres siguientes problemas stiff tomados de la literatura (dos de ellos considerados previamente en el capítulo II), elegidos por tener tres dimensiones bien distintas:

- El oscilador de Van der Pol [37, pág.144], de dimensión 2, integrado desde 0 hasta 20.
- El Ring Modulator [49], que tiene dimensión 15.
- El problema CUSP [37, pág.147], tomando $N = 32$ el problema tiene dimensión 96.

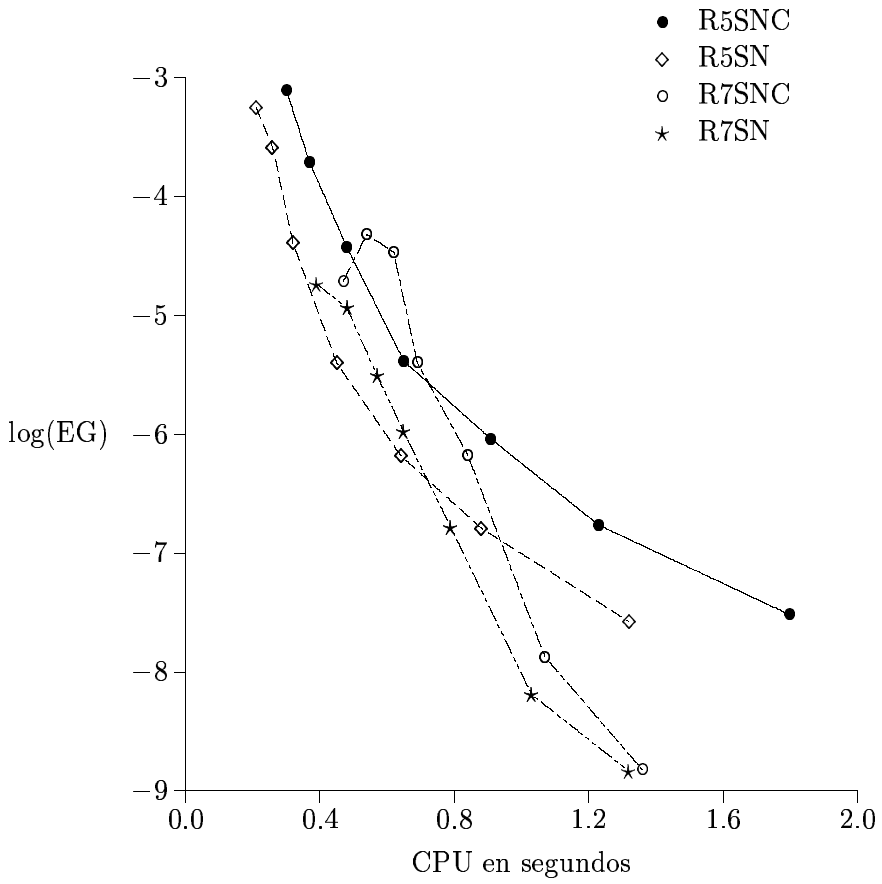
En las gráficas 3.4.1 a 3.4.3 representamos el tiempo de CPU (medido en segundos) frente al logaritmo del error global sobre el punto final de integración, para los cuatro códigos arriba mencionados.

Para entender más claramente el comportamiento de los códigos, hemos representado en las tablas 3.4.1 a 3.4.3, y para algunas tolerancias, el número total de iteraciones (NIT) requeridas por cada código así como el número total de factorizaciones LU (NLU). También aparecen el número de paso aceptados (NACC) y el número de pasos donde se produjo divergencia del esquema iterativo (NREJ). Obsérvese que para R5SNC la mitad de las factorizaciones LU corresponden a matrices complejas $m \times m$, mientras que para R7SNC todas las factorizaciones LU corresponden a matrices complejas $m \times m$.

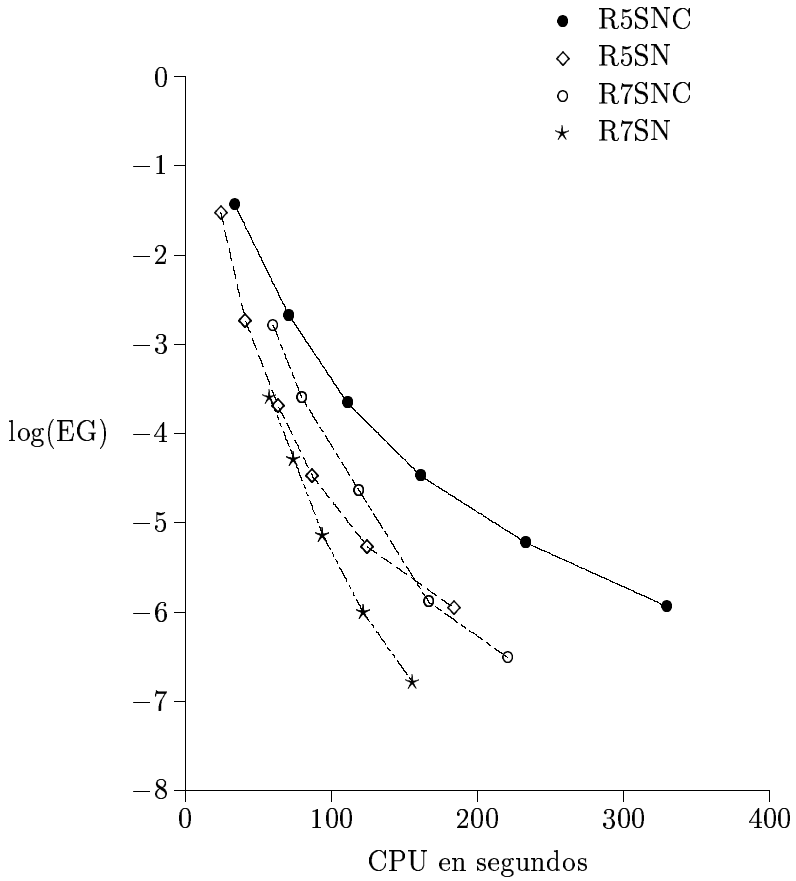
En general, los esquemas Single–Newton necesitan más iteraciones para converger que la iteración de Newton Simplificada. Además, los códigos que usan los esquemas Single–Newton integran los problemas tomando algunos pasos más que aquellos que usan el Newton Simplificado, debido a que con estos últimos esquemas se rechazan menos pasos por convergencia del esquema iterativo. Por otra parte, comparando los códigos de orden 7 con los de orden 5, podemos observar que los primeros integran los problemas dando algunos pasos menos que los últimos (suponiendo que se usa el mismo tipo de esquema iterativo en ambos casos).

Comentamos ahora brevemente el comportamiento de los códigos con respecto a cada uno de los 3 problemas considerados.

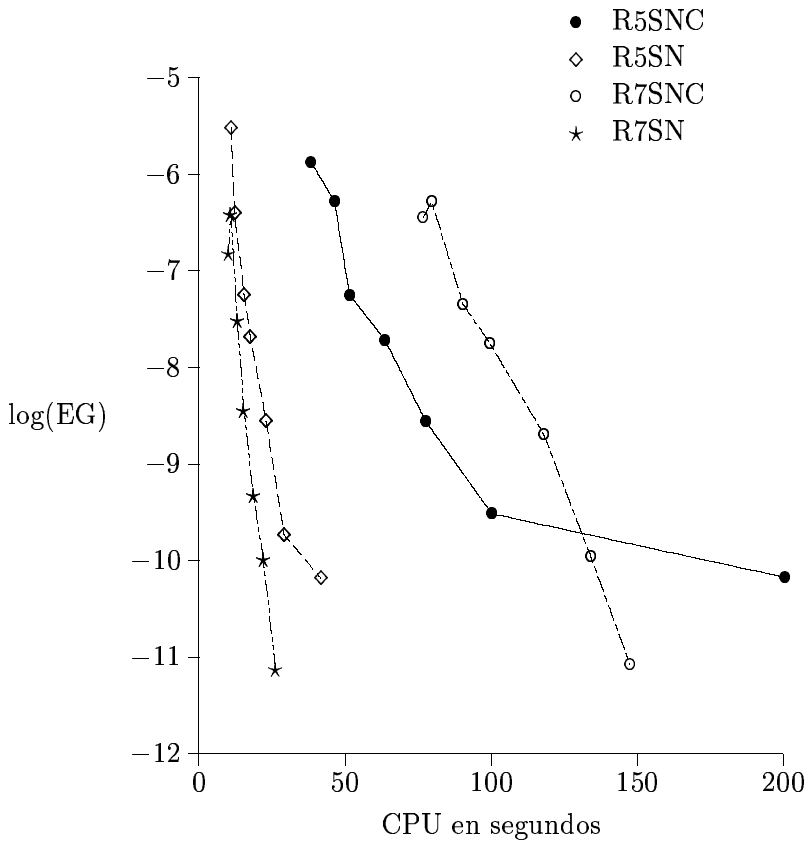
En primer lugar, ya que el Van der Pol es un problema de dimensión 2, el costo computacional está afectado principalmente por el costo de las iteraciones (mucho más que por el costo de las factorizaciones LU). Como puede verse en la gráfica 3.4.1 y en la tabla 3.4.1, aunque los esquemas Single–Newton convergen usando más iteraciones que Newton Simplificado, como el costo de cada iteración es mayor en el último caso, los esquemas Single–Newton son más eficientes que Newton Simplificado para este problema. En general, podemos decir que para problemas de dimensión baja no hemos apreciado diferencias notables con respecto al tipo de esquema usado (Single–Newton o Newton Simplificado). Con respecto al comportamiento de los códigos de




Gráfica 3.4.1: Problema Van der Pol.




Gráfica 3.4.2: Problema Ring Modulator.




Gráfica 3.4.3: Problema CUSP.

| | Tol | NIT | NLU | NACC | NREJ | Log(EG) |
|---|---|---|---|---|---|---|
| R5SNC | $10^{-5}$ | 24050 | 7112 | 3418 | 312 | -3.7171 |
| | $10^{-7}$ | 39196 | 13406 | 6538 | 183 | -5.3862 |
| | $10^{-9}$ | 71524 | 26686 | 13190 | 25 | -6.7644 |
| R5SN | $10^{-5}$ | 24728 | 3594 | 3438 | 362 | -3.5898 |
| | $10^{-7}$ | 49543 | 6653 | 6492 | 195 | -5.4003 |
| | $10^{-9}$ | 96309 | 13327 | 13176 | 27 | -6.7905 |
| R7SNC | $10^{-5}$ | 20539 | 5120 | 2410 | 464 | -4.3188 |
| | $10^{-7}$ | 25996 | 6852 | 3324 | 287 | -5.3982 |
| | $10^{-9}$ | 39064 | 11434 | 5588 | 178 | -7.8742 |
| R7SN | $10^{-5}$ | 27775 | 2965 | 2724 | 592 | -4.9363 |
| | $10^{-7}$ | 37929 | 3857 | 3678 | 464 | -5.9823 |
| | $10^{-9}$ | 60473 | 6306 | 6158 | 295 | -8.1881 |

Tabla 3.4.1: Problema de Van der Pol.

| | Tol | NIT | NLU | NACC | NREJ | Log(EG) |
|---|---|---|---|---|---|---|
| R5SNC | $10^{-5}$ | 154299 | 92988 | 44958 | 51 | -2.6725 |
| | $10^{-7}$ | 349226 | 213530 | 102834 | 20 | -4.4781 |
| | $10^{-9}$ | 714432 | 436514 | 211490 | 7 | -5.2241 |
| R5SN | $10^{-5}$ | 331168 | 48259 | 46690 | 777 | -2.7365 |
| | $10^{-7}$ | 695589 | 108887 | 104654 | 890 | -4.4843 |
| | $10^{-9}$ | 1384071 | 219704 | 212666 | 1129 | -5.2742 |
| R7SNC | $10^{-5}$ | 82350 | 41404 | 20322 | 114 | -2.7866 |
| | $10^{-7}$ | 162526 | 82512 | 40116 | 57 | -4.6380 |
| | $10^{-9}$ | 301844 | 153424 | 74380 | 23 | -5.8876 |
| R7SN | $10^{-5}$ | 251423 | 30715 | 28760 | 2697 | -3.5881 |
| | $10^{-7}$ | 410514 | 49696 | 47586 | 2334 | -5.1354 |
| | $10^{-9}$ | 677947 | 81573 | 78834 | 1569 | -6.0069 |

Tabla 3.4.2: Problema Ring Modulator.

| | Tol | NIT | NLU | NACC | NREJ | Log(EG) |
|---|---|---|---|---|---|---|
| R5SNC | $10^{-5}$ | 1770 | 544 | 250 | 55 | -6.2808 |
| | $10^{-7}$ | 2524 | 740 | 350 | 37 | -7.7230 |
| | $10^{-9}$ | 3715 | 1184 | 584 | 23 | -9.5152 |
| R5SN | $10^{-5}$ | 1833 | 269 | 248 | 55 | -6.4022 |
| | $10^{-7}$ | 2905 | 373 | 350 | 46 | -7.6808 |
| | $10^{-9}$ | 4610 | 608 | 596 | 23 | -9.7380 |
| R7SNC | $10^{-5}$ | 1376 | 464 | 212 | 55 | -6.2861 |
| | $10^{-7}$ | 2080 | 552 | 262 | 51 | -7.7474 |
| | $10^{-9}$ | 2969 | 764 | 368 | 36 | -9.9585 |
| R7SN | $10^{-5}$ | 1712 | 246 | 222 | 58 | -6.4162 |
| | $10^{-7}$ | 2642 | 306 | 290 | 61 | -8.4424 |
| | $10^{-9}$ | 3906 | 411 | 392 | 51 | -9.9907 |

Tabla 3.4.3: Problema CUSP.

órdenes 5 y 7 podemos decir que, en general, el de orden 5 se comporta mejor para tolerancias grandes o medianas ($10^{-2}$ hasta $10^{-6}$) mientras que la fórmula de orden 7 es preferible cuando se requiere una precisión más alta.

Con el problema Ring Modulator, que tiene dimensión $m = 15$, tanto el costo de las factorizaciones LU como el de las iteraciones contribuyen de forma similar en el costo computacional final. Por tanto, aunque el número total de iteraciones de los códigos que usan esquemas de tipo Single–Newton es considerablemente mayor que para los códigos que implementan la iteración de Newton Simplificada, el costo involucrado en las factorizaciones LU hace que estos últimos sean menos eficientes que los primeros.

Con respecto a la eficiencia de las fórmulas de órdenes 5 y 7, podemos decir que los métodos de orden más alto (7 en este caso) resultaron más eficientes para todos los valores de las tolerancias, independientemente del tipo de esquema iterativo usado. Esto puede explicarse teniendo en cuenta que la fórmula de más alto orden (orden 7) necesita muchos menos pasos (aproximadamente la mitad) para completar la integración que la fórmula de orden más bajo (orden 5). Consecuentemente, se invierte mucho menos trabajo en el cómputo de factorizaciones LU. Además, aunque los métodos de orden 5 necesitaron menos iteraciones por paso de integración (ver tabla 3.4.2) que los métodos de orden 7, en el cómputo global de iteraciones los métodos de orden 5 usaron casi el doble de iteraciones que los métodos de orden 7.

Comentemos finalmente los resultados obtenidos al integrar el problema CUSP. En este caso la dimensión es $m = 96$, y el costo computacional viene determinado prácticamente por las factorizaciones LU. Como puede verse en la gráfica 3.4.3, los esquemas Single–Newton son mucho más eficientes que la iteración de Newton Simplificada para ambos métodos (fórmulas de órdenes 5 y 7), como era de esperar. Además, el esquema Single–Newton resulta más eficiente (comparado con Newton Simplificado) en la fórmula de orden 7 que en la de orden 5, pues obsérvese que independientemente del número de etapas, el esquema Single–Newton sólo involucra una factorización LU por paso de integración (siempre que no haya problemas de convergencia), mientras que la iteración de Newton Simplificada conlleva dos factorizaciones, siendo una real y otra compleja para la fórmula de orden 5 y las dos complejas para la fórmula de orden 7.

Comparando las fórmulas de órdenes 5 y 7 observamos que cuando se implementan esquemas de tipo Single–Newton, la fórmula de orden 7 resulta siempre más eficiente que la de orden 5. Esto se justifica por el menor número de pasos invertido en la integración cuando se usa la fórmula de orden más alto. Por otra parte, cuando se implementa la iteración de Newton Simplificada, la fórmula de orden 5 resulta casi siempre más eficiente, lo que es debido al coste de las factorizaciones LU, que resultan mucho más costosas en la fórmula de orden 7 que en la de orden 5. En este caso, aunque el número de pasos de integración tomados por la fórmula de orden 7 es sensiblemente menor que el tomado por la fórmula de orden 5, esto no compensa el gasto extra en factorizaciones LU que necesita la fórmula de mayor orden.

Como conclusión y a la vista de todo lo expuesto anteriormente, podemos decir que los esquemas de tipo Single–Newton (optimizados) son bastante eficaces a la hora de resolver las ecuaciones de etapa de los métodos Runge–Kutta "fully–implicit" de alto orden, cuando se aplican sobre problemas de tipo stiff. Además parecen ser una mejor alternativa que la iteración de Newton Simplificada para integrar este tipo de problemas, particularmente cuando la dimensión de los mismos es elevada. Además los esquemas de tipo Single–Newton suelen hacer que los métodos de orden más alto sean más eficientes que los métodos de orden más bajo.

De los experimentos aquí realizados y de los del capítulo II se vislumbra la posibilidad de que los métodos Runge–Kutta de alto orden tales como Radau IIA o Lobatto IIIA (con iteración Single–Newton) pueden ser competitivos e incluso superiores a los otros tipos de métodos

(Runge–Kutta o multipaso) usados para la resolución de sistemas stiff, principalmente cuando se requiera precisión media o alta.

Capítulo **IV**

# Inicializadores sin coste adicional para RK implícitos

.

# Capítulo IV

# Inicializadores sin coste adicional para RK implícitos

## IV.1 Introducción

Hemos dedicado los dos capítulos anteriores de la memoria a optimizar esquemas iterativos de tipo Single–Newton para la resolución eficiente de las ecuaciones de etapa de algunos métodos Runge–Kutta implícitos para problemas stiff. De los fundamentos teóricos anteriormente expuestos y los experimentos numéricos realizados se infiere que el tipo de iteración Single–Newton parece ser robusta y bastante eficiente para los métodos de colocación basados en los nodos de Gauss, Radau y Lobatto. También se observa en las tablas del capítulo II (ver tabla 2.4.1 y tabla 2.4.2) que el número de iteraciones por paso de integración para obtener convergencia del proceso iterativo es relativamente alto sobre problemas no lineales en general, tanto si se usa la iteración de Newton Simplificada como la iteración Single–Newton. Con el fin de mejorar la eficiencia del proceso iterativo, es natural plantearse la obtención de buenas aproximaciones iniciales con el objeto de reducir si es posible el número de iteraciones.

Consideremos el Problema de Valor Inicial (PVI):

$$y'(t) = f(t, y(t)), \quad y(0) = z_0 \in I\!R^m, \quad t \in [0, T], \tag{4.1.1}$$

donde $f : [0, T] \times I\!R^m \to I\!R^m$ admitirá tantas derivadas parciales como se necesite en el análisis, en un entorno tubular de la solución única $y(t),\ t \in [0, T]$, de (4.1.1).

Como hemos visto previamente, un método Runge–Kutta de $s$ etapas avanza un paso de tamaño $h$, digamos de $(t_0, y_0)$ a $(t_1 = t_0 + h, y_1)$[1], mediante la fórmula

$$y_1 = y_0 + h \sum_{i=1}^{s} b_i f(t_0 + c_i h, X_i), \tag{4.1.2}$$

donde las etapas internas $X_i$ se calculan del sistema

$$X_i = y_0 + h \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, X_j), \quad (i = 1, \ldots, s). \tag{4.1.3}$$

Una vez que se ha resuelto el sistema algebraico (4.1.3), normalmente mediante una iteración de tipo Newton, queremos computar buenas aproximaciones $Y_i^0$ a las etapas internas del RK para

[1] Preferimos usar en adelante las notaciones $t_0,\ t_1$ en lugar de $t_n,\ t_{n+1}$, para simplificar la exposición.

el siguiente paso $(t_1 = t_0 + h, y_1) \to (t_2 = t_1 + rh, y_2)$ a efectos de comenzar de nuevo el proceso iterativo para el sistema

$$Y_i = y_1 + \bar{h} \sum_{j=1}^{s} a_{ij} f(t_1 + c_j \bar{h}, Y_j), \quad (i = 1, \ldots, s), \tag{4.1.4}$$

donde $r = \bar{h}/h$ se puede suponer de tamaño moderado.

Con este fin vamos a considerar los dos siguientes tipos de inicializadores (o aproximaciones iniciales) que pueden implementarse sin coste adicional en el sentido de que no requieren evaluaciones de función ni resoluciones de sistemas adicionales:

**Inicializadores de Tipo I**

$$Y_i^0 = \gamma_i y_0 + \sum_{j=1}^{s} \alpha_{ij} X_j, \qquad i = 1, ..., s. \tag{4.1.5}$$

**Inicializadores de Tipo II**

$$Y_i^0 = y_0 + h\delta_i f(t_0, y_0) + h \sum_{j=1}^{s} \beta_{ij} f(t_0 + c_j h, X_j), \qquad i = 1, ..., s. \tag{4.1.6}$$

Los coeficientes $\{\gamma_i, \alpha_{ij}\}$ y $\{\delta_i, \beta_{ij}\}$ sólo van a depender de la razón de cambio de paso $r$ y de los coeficientes del método Runge–Kutta.

Naturalmente interesa obtener inicializadores $Y_i^0$ $(1 \le i \le s)$ que aproximen los más posible a las etapas internas $Y_i$ $(1 \le i \le s)$ del método RK considerado, pues de este modo se conseguirá en general una convergencia más rápida para el proceso iterativo empleado. Mediremos el orden de esta aproximación con respecto al tamaño de paso $h$. Así, denotando por

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_s \end{pmatrix}, \qquad Y^0 = \begin{pmatrix} Y_1^0 \\ \vdots \\ Y_s^0 \end{pmatrix},$$

diremos que la aproximación inicial $Y^0$ es de *orden clásico* $q$ sii

$$\max_{1 \le i \le s} |Y_i - Y_i^0| = \mathcal{O}(h^{q+1}).$$

El término $\mathcal{O}(h^{q+1})$ puede depender de potencias positivas de la razón de paso $r$ y de los coeficientes del método RK. Como veremos más adelante, este tipo de orden puede no ser relevante para problemas de tipo stiff, pues para este tipo de problemas las diferenciales elementales sucesivas de $f$ pueden crecer arbitrariamente en un entorno de $(t_0, y_0)$. De todos modos el orden clásico de un inicializador es interesante porque nos da una primera idea de la aproximación obtenida cuando el problema es no stiff, y nos suministrará además una cota superior del orden stiff del inicializador. Por otra parte, hay muchos problemas considerados stiff que combinan intervalos donde el problema es stiff con intervalos donde no lo es, y por tanto también se hace necesario que el inicializador posea un orden clásico lo más alto posible. El orden stiff será definido de forma precisa en la sección IV.3.

Obsérvese de (4.1.3) que en realidad los inicializadores de Tipo I son un caso particular de los inicializadores de Tipo II con $\delta_i = 0,\ i = 1, \ldots, s$, si

$$\gamma_i + \sum_{j=1}^{s} \alpha_{ij} = 1. \tag{4.1.7}$$

En este caso, denotando $\mathcal{A} = (\alpha_{ij})$ y $\mathcal{B} = (\beta_{ij})$, los coeficientes de los inicializadores correspondientes están relacionados por $\mathcal{B} = \mathcal{A}A$. Además, si $A$ es no singular, todos los inicializadores de Tipo II con $\delta_i = 0,\ i = 1, \ldots, s$, pueden también expresarse como un inicializador de Tipo I con $\mathcal{A} = \mathcal{B}A^{-1}$. Por otra parte, como $Y_i = y_0 + \mathcal{O}(h)$ y $X_i = y_0 + \mathcal{O}(h)$, la condición (4.1.7) será necesaria para que $Y^0$ tenga orden al menos 0. Por tanto, vamos a considerar solamente inicializadores de Tipo I que satisfagan (4.1.7).

A pesar de esto, en lo que sigue preferiremos mantener por separado las notaciones de Tipo I y II. Procedemos así ya que la notación de Tipo I hace más patente la relación entre los inicializadores y la interpolación polinómica, que generalmente suele ser la más aplicada.

Si la matriz $A$ es no singular, por motivos de ahorro computacional y sobre todo para reducir la propagación de errores (tales como errores de redondeo, errores globales del método RK, errores de convergencia del proceso iterativo empleado en la solución del sistema algebraico de las etapas, etc), es mejor usar una versión algebraicamente equivalente de (4.1.6) que consiste en eliminar, usando (4.1.3), las derivadas del paso anterior $f(t_0 + c_j h, X_j)$, expresando entonces el inicializador $Y_i^0$ como una combinación lineal de $X_i,\ y_0$ y $f(t_0, y_0)$. Además, si el método es stiffly accurate (i.e. $b_j = a_{sj},\ j = 1, \ldots, s$), es conveniente expresar también $f(t_0, y_0)$ como una combinación lineal de las etapas internas del paso anterior, todo ello con el propósito de reducir las amplificaciones de errores que se producen al evaluar la función derivada, sobre todo en problemas stiff que son en realidad aquéllos para los que pretendemos los inicializadores. Sin embargo, a efectos del análisis teórico vamos a mantener el inicializador como se ha expresado en (4.1.6).

Nørsett y Thomsen en [52] ya han considerado inicializadores de estos tipos para el caso no stiff. También se han estudiado inicializadores similares en [44], [45], [46], principalmente para arrancar las iteraciones de los métodos RK–Gauss aplicados a problemas hamiltonianos no stiff. En estos artículos la autora también propone algunas variantes que consisten en añadir una o dos evaluaciones adicionales de la función derivada para ganar orden y reducir así el número de iteraciones del proceso iterativo empleado. Por otro lado, Sand [57] estudió inicializadores de Tipo I para problemas stiff, considerando además de las condiciones de orden usuales (lo que aquí hemos denominado orden clásico) lo que él denomina condiciones de orden inverso, i.e., desarrollos de las etapas intermedias en potencias de $h^{-1}$. Esto puede tener sentido, como reconoce el mismo autor, cuando el radio espectral de la matriz $h^{-1}J^{-1}(t_0, y_0)$ es pequeño (aquí $J = \partial f/\partial y$ denota la matriz jacobiana). Sin embargo, aunque da algunas ideas sobre qué condiciones de orden elegir cuando se considera un método SDIRK tomado de [52] y se integran problemas stiff lineales, no está claro si esas condiciones de orden serán apropiadas para métodos RK completamente implícitos (como los Gauss, Radau, Lobatto o métodos RK de colocación) sobre problemas no lineales en general. Además, la teoría del orden inverso no está plenamente justificada y la de orden clásico (basada en la comparación de desarrollos en potencias de $h$) no es satisfactoria para los problemas stiff, ya que para estos problemas el tamaño de algunas diferenciales elementales de $f$ puede crecer arbitrariamente.

En [41] se consideran varios tipos de algoritmos de arranque, como predictores de ciertos métodos *Runge–Kutta iterados* (orientados a la computación en paralelo) los cuales están basados a su vez en los métodos de Gauss, RadauIA-IIA o LobattoIIIA-C, y se analiza el orden de los métodos predictor-corrector resultantes, así como algunas propiedades importantes de estabilidad lineal. Nuestro objetivo aquí es bastante diferente, ya que nosotros estamos interesados en inicializadores para la computación secuencial y además, resolvemos el sistema (4.1.4) hasta que dos iteraciones sucesivas estén lo suficientemente cerca, i.e., $\max_i |Y_i^{k+1} - Y_i^k| < \text{Tol}$, y no con un número fijo de iteraciones como se hace en [41]. También se observa que las aproximaciones iniciales más interesantes desarrolladas en [41] conllevan un costo computacional adicional que

no es despreciable como lo es en nuestro caso.

Por otro lado, en [37, Cap. IV.8], los autores recomiendan como inicializadores para la iteración de Newton Simplificada la interpolación polinómica de Lagrange de las etapas internas $X_i$ e $y_0$ del paso anterior evaluada en los puntos $t = 1 + rc_i$, antes que tomar $Y_i^0 = y_1$ (ambos casos son de Tipo I). Los autores basan sus consideraciones en el comportamiento de tales inicializadores sobre una gran variedad de problemas stiff integrados por su código RADAU5 (que usa la fórmula Radau IIA de tres etapas). Es bien conocido que los inicializadores basados en la interpolación de Lagrange de las etapas internas del paso anterior $X_i$ e $y_0$ dan buenos valores iniciales en general cuando el método RK usado es de colocación. Aquí intentamos dar un soporte teórico a este hecho y analizar además otros inicializadores más generales de posible interés para la integración de problemas stiff. Por otra parte, también se han hecho estudios sobre los inicializadores para ecuaciones diferenciales algebraicas en [53] y [56]. Sin embargo, aunque existe una conexión entre cierta clase de problemas stiff y cierto tipo de DAE's, el estudio es bastante diferente en ambos casos.

### IV.1.1 Algoritmos de arranque y procesos iterativos de tipo Newton

Relacionando los inicializadores con los procesos iterativos más usados, es decir, la iteración de Newton Simplificada y los métodos de tipos Single–Newton podemos decir lo siguiente:

**Problemas no stiff:**

Supondremos que el inicializador considerado tiene orden clásico (no stiff) $q$, es decir,

$$Y - Y^0 = \mathcal{O}(h^{q+1}),$$

y que el jacobiano se ha evaluado en el punto $t_0$ o en alguno próximo, es decir,

$$J = \frac{\partial f}{\partial y}(t_0, y_0) + \mathcal{O}(h).$$

En este caso, considerando la iteración de Newton Simplificada (3.1.1) se tiene que tras $k$ iteraciones, ver por ejemplo [48] o [20, fórmula (4.2c)],

$$Y - Y^k = \mathcal{O}(h^2)(Y - Y^{k-1}) = \mathcal{O}(h^{2k})(Y - Y^0) = \mathcal{O}(h^{2k+q+1}), \quad k = 1, 2, \ldots.$$

Por otra parte, si consideramos la iteración Single–Newton, entonces se sigue de [28, sección 3] que

$$Y - Y^k = \mathcal{O}(h)(Y - Y^{k-1}) = \mathcal{O}(h^k)(Y - Y^0) = \mathcal{O}(h^{k+q+1}), \quad k = 1, 2, \ldots.$$

De esto se infiere claramente que un mayor orden clásico para el inicializador garantiza una convergencia más rápida del proceso iterativo, al menos cuando medimos la velocidad de convergencia con respecto al tamaño de paso $h$.

**Problemas stiff:**

Consideremos ahora problemas stiff contractivos, es decir aquellos cuya $f$ satisface (1.0.3) con $\nu = 0$. Supondremos además que el jacobiano de $f$ varía relativamente de forma acotada, esto es,

$$J(t + \Delta t, y + \Delta y) - J(t, y) = J(t, y)E(t, y, \Delta t, \Delta y), \quad \forall t, y, \Delta t, \Delta y,$$

donde

$$|E(t, y, \Delta t, \Delta y)| \leq K_1|\Delta t| + K_2|\Delta y|, \quad K_j = \mathcal{O}(1), \quad j = 1, 2.$$

Este tipo de hipótesis ha sido ampliamente usada como puede verse en [20], [58], [14], [15] y es verificada por muchos problemas stiff de interés.

Consideramos la clase anterior de problemas stiff y

$$J = \frac{\partial f}{\partial y}(t_0, y_0) + \mathcal{O}(h),$$

donde se asume que $\mathcal{O}(h)$ es independiente de la stiffness del problema considerado (esto también se cumplirá para las cantidades $\mathcal{O}(h)$ que aparecen a continuación en el resto de esta subsección). Entonces tenemos lo siguiente:

(a) Iteración de Newton Simplificada:

Se deduce inmediatamente de los resultados obtenidos en [15, Teorema 3.1] o [20, fórmula (4.5c)] que tras $k$ iteraciones

$$Y - Y^k = \mathcal{O}(h)(Y - Y^{k-1}) = \mathcal{O}(h^k)(Y - Y^0), \quad k = 1, 2, \ldots.$$

(b) Iteración Single–Newton:

En este caso tras $k$ iteraciones, se obtiene de los resultados presentados en [15, sección 2] que

$$\max_{1\le i\le s} |Y_i - Y_i^k| \le (\eta^k + \mathcal{O}(h)) \max_{1\le i\le s} |Y_i - Y_i^0|, \quad k = 0, 1, \ldots,$$

donde

$$\eta = \sup_{Re(z)\le 0} |M(z)|_2,$$

con $M(z)$ dada en el capítulo anterior mediante (3.3.2).

También se puede dar una acotación en función del radio espectral de $M(z)$

$$\beta = \sup_{Re(z)\le 0} \rho(M(z)),$$

asumiendo que $\beta < 1$, mediante la siguiente fórmula [15, sección 2],

$$\max_{1\le i\le s} |Y_i - Y_i^k| \le C_s(\alpha)(\alpha^k + \mathcal{O}(h)) \max_{1\le i\le s} |Y_i - Y_i^0|, \quad k = 0, 1, \ldots,$$

donde $\alpha$ es cualquiera en el intervalo $(\beta, 1)$, y

$$C_s(\alpha) = (1 + \eta/(\alpha - \beta))^{s-1}.$$

Aquí $s$ denota el número de etapas implícitas del método RK considerado.

De todo lo anterior se desprende que, como era de esperar, también en el caso stiff un inicializador de más alto orden (para problemas stiff) conducirá a una convergencia más rápida del proceso iterativo, independientemente del tipo de iteración considerada.

El resto de este capítulo se organiza de la siguiente manera. En la sección IV.2 estudiamos el orden clásico de los inicializadores propuestos mediante la teoría de las series de Butcher. En la sección IV.3 consideramos el orden stiff, en primer lugar analizando el orden sobre el modelo de Prothero y Robinson, para concluir con algunos resultados de orden sobre problemas contractivos en general. Finalizamos el capítulo con la sección IV.4, donde se realizan experimentos numéricos para confirmar algunos de los resultados teóricos obtenidos en las secciones anteriores y profundizar un poco más en el comportamiento de los inicializadores más interesantes cuando se usan con fines prácticos.

Hemos de reseñar también que lo esencial de la investigación realizada en el presente capítulo, se encuentra recogida en los trabajos [32] y [33].

## IV.2 Orden no stiff

Vamos a usar las condiciones simplificadoras

$$B(p):\ \ b^T c^{j-1} = 1/j,\ \ (1 \le j \le p), \qquad C(q):\ \ Ac^{j-1} = c^j/j,\ \ (1 \le j \le q),$$

donde $c^j = (c_1^j, \ldots, c_s^j)^T$, y llamamos *orden de etapa* del RK $(A, b)$ al mayor entero $\tau$ tal que se verifican $B(\tau)$ y $C(\tau)$. Además, diremos que un RK es *no confluente* si todos los $c_i$ son distintos entre sí.

Para simplificar el estudio veamos en primer lugar los siguientes lemas:

**Lema IV.2.1** *Sea $1 \le q \le s$ y supongamos que tenemos un método RK $(A, b)$ no confluente que verifica las condiciones $B(q)$ y $C(q)$. Entonces, un inicializador de Tipo II tiene orden $\ge q+1$ si y sólo si para todo $i = 1, \ldots, s$,*

$$\delta_i + \beta_i^T e = (1 + rc_i), \tag{4.2.1}$$

$$\beta_i^T c^{j-1} = \Delta_{ij}(r), \quad j = 2, ..., q+1, \tag{4.2.2}$$

*donde $\beta_i^T = (\beta_{i1}, ..., \beta_{is})$ y*

$$\Delta_{ij}(r) = b^T c^{j-1} + r A_i^T (e + rc)^{j-1}, \tag{4.2.3}$$

*siendo $A_i^T = (a_{i1}, \ldots, a_{is})$ la i–ésima fila de la matriz $A$.*

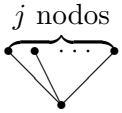
<u>*Demostración.*</u> Si consideramos $y_0$ en (4.1.3) como una etapa adicional $X_0 \equiv y_0$, el inicializador $Y_i^0$ puede interpretarse como la aproximación dada por un método RK de $s+1$ etapas cuya tabla de Butcher [36, Cap. II.1] es

$$\begin{array}{c|c} \tilde{c} & \tilde{A} \\ \hline Y_i^0 & \tilde{p}_i^T \end{array} \quad = \quad \begin{array}{c|cc} 0 & 0 & \mathbf{0}^T \\ c & 0 & A \\ \hline Y_i^0 & \delta_i & \beta_i^T \end{array} \tag{4.2.4}$$

De la misma forma, la solución exacta $Y_i$ del sistema implícito (4.1.4) puede verse como la solución dada por el RK de $2s$ etapas con tabla de Butcher

$$\begin{array}{c|c} \bar{c} & \bar{A} \\ \hline Y_i & p_i^T \end{array} \quad = \quad \begin{array}{c|cc} c & A & \mathbf{0} \\ e + rc & eb^T & rA \\ \hline Y_i & b^T & rA_i^T \end{array} \tag{4.2.5}$$

Luego, comparando las series de Butcher [36, Cap. II.2] correspondientes a $Y_i$ e $Y_i^0$ respectivamente y teniendo en cuenta las condiciones $B(q)$ y $C(q)$, las condiciones para obtener orden $q+1$ (i.e. $Y_i - Y_i^0 = \mathcal{O}(h^{q+2})$) se reducen a las correspondientes a los árboles de máxima ramificación para $j = 0, 1, ..., q$:

$\tau_j \ \equiv$ 


lo que nos lleva directamente a las ecuaciones (4.2.1), (4.2.2). ∎

**Nota IV.2.1** *Como los inicializadores de Tipo I que verifican la condición (4.1.7) son un caso particular de los de Tipo II, las condiciones para que un algoritmo de Tipo I tenga orden $\geq q+1$ (asumiendo orden de etapa $q$ al menos) se reducen a:*

$$\gamma_i + \alpha_i^T e = 1, \quad \alpha_i^T A c^{j-1} = \Delta_{ij}(r), \quad 1 \leq j \leq q+1, \tag{4.2.6}$$

*donde $\alpha_i^T = (\alpha_{i1}, ..., \alpha_{is})$ y $\Delta_{ij}$ dado por (4.2.3).*

**Lema IV.2.2** *Supongamos que un RK $(A, b)$ de $s$ etapas verifica $B(p)$ y $C(q)$. Entonces,*

$$\Delta_{ij}(r) = \frac{(1+rc_i)^j}{j}, \quad i = 1, ..., s, \quad j = 1, ..., u.$$

*con $u = \min\{p, q\}$.*

*Demostración.* Se sigue inmediatamente de la definición de $\Delta_{ij}(r)$ usando las condiciones $B(p)$ y $C(q)$. ∎

**Lema IV.2.3** *Sea un RK $(A, b)$ de $s$ etapas no confluente que satisface la condición $C(q)$ con $q \geq s-1$. Denotemos $V = [e, c, ..., c^{s-1}]$ (matriz de orden $s$) y $e_1^T = (1, 0, ..., 0)$. Entonces*

$$A \text{ es singular} \quad \Longleftrightarrow \quad e_1^T V^{-1} A c^{s-1} = 0.$$

*Demostración.* Si $s = 1$ es trivial. Para $s > 1$ denotemos $K = e_1^T V^{-1} A c^{s-1}$. Como todos los nodos $c_i$ son distintos, la matriz $V$ es no singular, por lo que existen números reales $\lambda_j$, $j = 0, \ldots, s-1$, tales que

$$A c^{s-1} = \sum_{j=0}^{s-1} \lambda_j c^j.$$

Por otra parte, $V^{-1} c^j = e_{j+1}$ $(j = 0, 1, ..., s-1)$ donde $e_j^T = (0, ..., \overset{(j)}{1}, ..., 0) \in \mathbb{R}^s$. De esto se sigue que

$$K = e_1^T \sum_{j=0}^{s-1} \lambda_j V^{-1} c^j = e_1^T \sum_{j=0}^{s-1} \lambda_j e_{j+1} = \lambda_0.$$

Ahora bien, usando $C(s-1)$ podemos poner

$$A[e, c, \ldots, c^{s-1}] = [c, c^2, \ldots, c^{s-1}, e] \begin{bmatrix} 1 & 0 & \cdots & 0 & \lambda_1 \\ 0 & 1/2 & 0 & \cdots & \lambda_2 \\ & & & & \\ 0 & & 0 & 1/(s-1) & \lambda_{s-1} \\ 0 & & & 0 & \lambda_0 \end{bmatrix},$$

con lo que tomando determinantes se concluye la demostración. ∎

**Teorema IV.2.1** *Supongamos que un RK $(A, b)$ de $s$ etapas no confluente satisface $B(s-1)$ y $C(s-1)$. Entonces, para los inicializadores de Tipo I tenemos que*

*(a) Existe una familia s-paramétrica de orden $s-1$ (con $\gamma_j$, $j=1,...,s$ como parámetros) que viene determinada por el sistema lineal*

$$\alpha_i^T e = 1-\gamma_i, \quad \alpha_i^T c^j = (1+rc_i)^j, \quad 1\le i\le s,\ \ 1\le j\le s-1. \tag{4.2.7}$$

*En particular, eligiendo $\gamma_j = 0; j=1,\dots,s$, los coeficientes del inicializador de Tipo I resultante son*

$$\alpha_{ij} = \bar{l}_j(1+rc_i), \quad 1\le i,j\le s, \tag{4.2.8}$$

*donde $\{\bar{l}_j(x); j=1,\dots,s\}$ es la base de polinomios fundamentales de Lagrange asociada a los nodos $\{c_1,\dots,c_s\}$.*

*(b) Si $A$ es singular, ningún inicializador de este tipo puede alcanzar orden $s$.*

*(c) Si $A$ es no singular, existe un único inicializador de orden máximo $s$. Si además se verifican $B(s)$ y $C(s)$ los coeficientes de dicho inicializador vienen dados por*

$$\begin{array}{l}\gamma_i = l_0(1+rc_i), \quad 1\le i\le s,\\ \alpha_{ij} = l_j(1+rc_i), \quad 1\le i,j\le s,\end{array} \tag{4.2.9}$$

*donde $\{l_0(x), l_1(x), ..., l_s(x)\}$ es la base de polinomios fundamentales de Lagrange asociada a los nodos $\{c_0=0, c_1, ..., c_s\}$.*

<u>*Demostración.*</u> *(a)* Es una consecuencia inmediata del Lema IV.2.1, la Nota IV.2.1 y el Lema IV.2.2, junto con la condición de que $c_i\neq c_j$ para $i\neq j$.

Además, si elegimos $\gamma_i = 0,\ i=1,\dots,s$, el sistema lineal (4.2.7) se reduce a

$$\alpha_i^T c^j = (1+rc_i)^j, \quad j=0,\dots,s-1, \quad i=1,\dots,s. \tag{4.2.10}$$

Por otro lado, la base de Lagrange asociada a los nodos $\{c_1,\dots,c_s\}$

$$\bar{l}_j(t) = \pi(t)/((t-c_j)\pi'(c_j)), \quad j=1,\dots,s; \quad \pi(t) = (t-c_1)\cdots(t-c_s), \tag{4.2.11}$$

satisface

$$t^j = \sum_{k=1}^{s} \bar{l}_k(t) c_k^j, \quad j=0,1,...,s-1. \tag{4.2.12}$$

Por tanto, tomando $t = 1+rc_i,\ i=1,...,s$, se sigue de la unicidad de solución del sistema (4.2.10) que $\alpha_{ik} = \bar{l}_k(1+rc_i),\ i,k=1,\dots,s$.

*(b)* Procedemos por reducción al absurdo. Si suponemos que $A$ es singular y que un inicializador alcanza orden $s$, entonces los parámetros $\{\alpha_{ij},\gamma_j\}$ deben satisfacer el sistema lineal (4.2.7) y la ecuación en (4.2.6) correspondiente a $j=s$ (Nota IV.2.1), i.e.,

$$\alpha_i^T A c^{s-1} = \Delta_{is}(r). \tag{4.2.13}$$

Denotando $V = [e, c, \dots, c^{s-1}]$, $u_i^T = (1, (1+rc_i), \dots, (1+rc_i)^{s-1})$ y $e_1^T = (1,0,\dots,0)$, las condiciones (4.2.7) pueden escribirse de la forma

$$\alpha_i^T V = u_i^T - \gamma_i e_1^T, \quad i=1,\dots,s \tag{4.2.14}$$

Ahora, multiplicando estas ecuaciones por el vector $V^{-1}Ac^{s-1}$ y sustituyendo en (4.2.13) obtenemos

$$u_i^T V^{-1} A c^{s-1} - \gamma_i e_1^T V^{-1} A c^{s-1} = \Delta_{is}(r). \tag{4.2.15}$$

Por el Lema IV.2.3, como $A$ es singular, se tiene que $e_1^T V^{-1} A c^{s-1} = 0$, lo que implica

$$u_i^T V^{-1} A c^{s-1} = b^T c^{s-1} + r A_i^T (e + rc)^{s-1}, \ \ \forall r > 0. \tag{4.2.16}$$

En virtud de que el lado izquierdo de esta ecuación es un polinomio en $r$ de grado a lo más $s-1$, obtenemos que $A_i^T c^{s-1} = 0$ para todo $i = 1, ..., s$, i.e., $Ac^{s-1} = \mathbf{0}$. Llevando esto a (4.2.16) se tendría que $b^T c^{s-1} + r A_i^T (e + rc)^{s-1} \equiv 0$, lo que nos lleva a que $A \equiv \mathbf{0}$. Esto es claramente imposible si $s > 1$ ya que $Ae = c$. Para el caso $s = 1$ también es imposible pues entonces $X_1 = y_0$, lo que implica de (4.2.7) que $Y_1^0 = y_0$, resultando un inicializador de orden 0 pero no de orden $s = 1$.

*(c)* Como $A$ es no singular se tiene del Lema IV.2.3 que $e_1^T V^{-1} A c^{s-1} \neq 0$, por lo que el único inicializador de orden $s$ (al menos) satisface (4.2.14)-(4.2.15) para cada $i = 1, \ldots, s$. Por tanto, viene explícitamente determinado por

$$\gamma_i = \frac{u_i^T V^{-1} A c^{s-1} - \Delta_{is}(r)}{e_1^T V^{-1} A c^{s-1}}, \quad 1 \leq i \leq s, \tag{4.2.17}$$

y el sistema lineal (4.2.14).

Veamos que es imposible alcanzar orden $s+1$ (en general). Para obtener orden $s+1$ el inicializador tendría que verificar además de (4.2.14) y (4.2.17) la condición (ver Nota IV.2.1)

$$\alpha_i^T A c^s = b^T c^s + r A_i^T (e + rc)^s.$$

Esto implica que $Ac^s = \mathbf{0}$ ya que el lado izquierdo de la ecuación anterior es un polinomio en $r$ de grado $\leq s$, mientras que el derecho es de grado $s+1$. Por tanto, $A$ debería ser cero, lo que es imposible.

Por otro lado, si también suponemos $B(s)$ y $C(s)$, tenemos por el Lema IV.2.2 que $\Delta_{is}(r) = (1 + rc_i)^s / s$. Luego, las condiciones para alcanzar orden $s$ se reducen a los sistemas lineales siguientes (para cada $i = 1, \ldots, s$) cuya dimensión es $s+1$,

$$(\gamma_i, \alpha_i^T) W = v_i^T, \quad i = 1, \ldots, s \tag{4.2.18}$$

donde

$$W = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ e & c & \cdots & c^s \end{bmatrix}, \quad v_i^T = (1, (1 + rc_i), \cdots, (1 + rc_i)^s). \tag{4.2.19}$$

Procediendo de forma similar a la demostración de la parte *(a)*, es fácil ver que la base de Lagrange asociada a los nodos $\{c_0 = 0, c_1, \ldots, c_s\}$ verifica las ecuaciones (4.2.18)-(4.2.19), por lo que se concluye la demostración de la unicidad de solución de (4.2.18). ∎

**Teorema IV.2.2** *Supongamos que un RK de s etapas no confluente satisface $B(s-1)$ y $C(s-1)$. Entonces para los inicializadores de Tipo II tenemos que*

*(a) Existe una familia s–paramétrica de orden s (con $\delta_j$, $j = 1, \ldots, s$ como parámetros libres) cuyos coeficientes vienen determinados por el sistema lineal*

$$\left.\begin{array}{l} \beta_i^T e = (1 + rc_i) - \delta_i, \\ \beta_i^T c^{j-1} = \dfrac{(1 + rc_i)^j}{j}, \quad j = 2, ..., s-1, \\ \beta_i^T c^{s-1} = \Delta_{is}(r). \end{array}\right\} \ i = 1, ..., s. \tag{4.2.20}$$

*Además si se verifican $B(s)$ y $C(s)$ y elegimos $\delta_j = 0$, $j = 1, \ldots, s$, el único inicializador de orden $s$ es*

$$\beta_{ij} = \int_0^{1+rc_i} \bar{l}_j(t)dt, \quad 1 \le i, j \le s,$$

*donde $\{\bar{l}_j(t);\ j = 1, \ldots, s\}$ es la base de Lagrange asociada a los nodos $\{c_1, \ldots, c_s\}$. Si $A$ es no singular dicho inicializador coincide con el de Tipo I dado en el Teorema IV.2.1–(c).*

*(b) Suponiendo que se verifican $B(s)$ y $C(s)$, existe un único inicializador de orden $s+1$ si y sólo si $A$ es no singular o el método RK considerado es el Euler explícito.*

Demostración. *(a)* La primera parte es una consecuencia inmediata de los Lemas IV.2.1 y IV.2.2. Además, si se satisfacen $B(s)$ y $C(s)$ las condiciones (4.2.20) pueden escribirse de la forma

$$\beta_i^T V = (1 + rc_i, (1 + rc_i)^2/2, ..., (1 + rc_i)^s/s); \quad i = 1, \ldots, s. \tag{4.2.21}$$

La base de Lagrange (4.2.11) asociada a los nodos $\{c_j;\ j = 1, \ldots, s\}$ verifica (4.2.12), por lo que la integración de esta última ecuación en $[0, 1 + rc_i]$ nos da

$$\sum_{k=1}^{s} c_k^{j-1} \int_0^{1+rc_i} \bar{l}_k(t)dt = (1 + rc_i)^j/j, \quad i, j = 1, \ldots, s.$$

Ahora bien, de la unicidad de solución del sistema lineal (4.2.21) se sigue que

$$\beta_{ik} = \int_0^{1+rc_i} \bar{l}_k(t)dt; \quad 1 \le i, k \le s.$$

Si la matriz $A$ del RK es no singular, tenemos que este inicializador es equivalente al de Tipo I dado en el Teorema IV.2.1–*(c)*, ya que ambos pueden expresarse de la forma

$$Y_i^0 = y_0 + h \sum_{j=1}^{s} \beta_{ij} f(t_0 + hc_j, X_j), \quad 1 \le i, j \le s,$$

para parámetros $\beta_{ij}$ adecuados y además ambos son de orden $s$. Por tanto, la equivalencia surge inmediatamente de la unicidad de la solución de (4.2.21).

*(b)* Las condiciones para alcanzar orden $s + 1$ con este tipo de inicializadores (teniendo en cuenta $B(s)$, $C(s)$ y el Lema IV.2.1) vienen dadas por

$$\delta_i e_1^T + \beta_i^T V = \bar{u}_i^T \equiv (1 + rc_i, (1 + rc_i)^2/2, ..., (1 + rc_i)^s/s), \ \ 1 \le i \le s, \tag{4.2.22}$$

y por la ecuación

$$\beta_i^T c^s = \Delta_{i,s+1}(r). \tag{4.2.23}$$

Esta última pueden reemplazarse (usando (4.2.22)) por

$$\delta_i e_1^T V^{-1} c^s + \Delta_{i,s+1}(r) = \bar{u}_i^T V^{-1} c^s. \tag{4.2.24}$$

Luego si $A$ es no singular, usando el Lema IV.2.3, $\bar{K} = e_1^T V^{-1} c^s \neq 0$ y por tanto los coeficientes del único inicializador de orden $\ge s + 1$ se obtienen de

$$\begin{cases} \delta_i = \dfrac{\bar{u}_i^T V^{-1} c^s - \Delta_{i,s+1}(r)}{e_1^T V^{-1} c^s}, \\ \beta_i^T = \bar{u}_i^T V^{-1} - \delta_i e_1^T V^{-1}, \end{cases} \quad 1 \le i \le s. \tag{4.2.25}$$

Recíprocamente, si el inicializador tiene orden $s+1$ se verifican (4.2.22) y (4.2.24). Luego, si $A$ fuera singular, la ecuación (4.2.24) queda

$$\Delta_{i,s+1}(r) = \bar{u}_i^T V^{-1} c^s, \quad 1 \leq i \leq s.$$

Comparando el coeficiente de $r^{s+1}$ de los polinomios en $r$ que aparecen en ambos miembros de la igualdad anterior, se llega a que tiene que ser $Ac^s = \mathbf{0}$. Como la matriz $V$ es no singular, existen números reales $\lambda_i$ tales que

$$c^s = \sum_{j=1}^{s} \lambda_j c^{j-1}. \tag{4.2.26}$$

De esto resulta que

$$\mathbf{0} = Ac^s = \sum_{j=1}^{s} \lambda_j A c^{j-1} = \sum_{j=1}^{s} \frac{1}{j} \lambda_j c^j,$$

y reemplazando $c^s$ dado por (4.2.26) obtenemos

$$\mathbf{0} = \lambda_1 \lambda_s e + \sum_{j=1}^{s-1} (\frac{s}{j} \lambda_j + \lambda_s \lambda_{j+1}) c^j.$$

Se sigue entonces que $\lambda_j = 0$, $\forall j = 1, \ldots, s$, lo que implica $c^s = \mathbf{0}$, o de forma equivalente, $c = \mathbf{0}$. Entonces, si $s > 1$ esto es imposible; si $s = 1$ tenemos que $A = c = 0$ y como se verifica $B(1)$, el método RK es el explícito de Euler, como queríamos demostrar.

Para ver que no se puede alcanzar orden $s+2$ cuando $A$ es no singular, demostraremos que si añadimos las condiciones correspondientes a la potencia $s+2$ de $h$ llegamos a una contradicción. Obsérvese que esto implicaría que

$$\beta_i^T A c^s = b^T A c^s + r A_i^T \left( e b^T c^s + r A (e + rc)^s \right), \quad i = 1, \ldots, s. \tag{4.2.27}$$

Por (4.2.25) el primer miembro de (4.2.27) es un polinomio en $r$ de grado $\leq s+1$, lo que implica que $A^2 c^s = \mathbf{0}$, o equivalentemente $\mathbf{c} = 0$. Pero esto ya hemos visto que es imposible. ∎

**Nota IV.2.2** *Para los métodos RK de $s$ etapas no confluentes que verifican $B(s)$ y $C(s)$ y que tienen matriz $A$ no singular, el inicializador de Tipo II con orden máximo $s+1$ puede calcularse por la fórmula alternativa siguiente, la cual es más práctica y más intuitiva que (4.2.25), y que además está relacionada con la interpolación polinómica:*

$$Y_i^0 = y_1 + rh \sum_{j=1}^{s} a_{ij} g_j, \quad i = 1, ..., s, \tag{4.2.28}$$

*donde*

$$g_j = P(1 + rc_j), \quad j = 1, ..., s$$

*siendo $P(x)$ el polinomio de interpolación definido por las condiciones:*

$$P(0) = f(t_0, y_0), \quad P(c_j) = f(t_0 + c_j h, X_j), \quad j = 1, \ldots, s.$$

*Esto puede verificarse fácilmente expresando el inicializador (4.2.28) como uno de Tipo II y demostrando que verifica las condiciones de orden $s+1$, dadas por (4.2.22) y (4.2.23).*

## IV.3 Orden stiff

Los resultados relativos al orden de los inicializadores propuestos, y que fueron deducidos en la sección anterior se basan en la teoría de las series de Butcher y son relevantes, en principio, sólo para problemas no stiff. Es bien sabido que los métodos Runge–Kutta pueden experimentar cierto fenómeno de reducción de orden cuando se aplican a sistemas stiff [37, Cap. IV.15]. Por tanto no es de extrañar que esta misma reducción de orden se transmita en el caso de los inicializadores. Como una primera aproximación al caso stiff, vamos a estudiar el comportamiento de los inicializadores cuando el método RK se aplica a la ecuación lineal de Prothero y Robinson. Posteriormente estudiaremos el orden sobre una clase más general de problemas no lineales stiff.

### IV.3.1 El modelo de Prothero y Robinson

Consideremos el modelo de Prothero y Robinson [55]

$$y' = f(t,y) := \lambda(y-\phi(t)) + \phi'(t), \quad y(0) = z_0 = \phi(0), \quad Re(\lambda) \leq 0, \tag{4.3.1}$$

donde $\phi(t)$ se supone suficientemente suave en $[0,T]$ (por ejemplo, podemos considerar que $\phi(t)$ es una función analítica). Si aplicamos un método RK $(\hat{A},\hat{b})$ para avanzar del punto $(t_0,y_0)$ a $t_1 = t_0 + h$, obtenemos (ver por ejemplo [50]):

$$y_{RK}(t_0+h) \equiv y_1 = \hat{R}(z)(y_0-\phi(t_0)) + \phi(t_0) + \sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!} v_j(z) h^j \tag{4.3.2}$$

donde $z = \lambda h$, y se supone que $z^{-1} \notin \sigma(\hat{A})$. Aquí $\hat{R}(z)$ es la función de estabilidad del método RK considerado, i.e.,

$$\hat{R}(z) = 1 + z\hat{b}^T(I - z\hat{A})^{-1}e, \tag{4.3.3}$$

y los coeficientes $v_j(z)$ vienen dados por

$$\begin{cases} v_j(z) = -z\mu_j + j\mu_{j-1}, & j = 1,2,... \\ \mu_j = \mu_j(z) = \hat{b}^T(I - z\hat{A})^{-1}\hat{c}^j, & j = 0,1,... \end{cases} \tag{4.3.4}$$

**Nota IV.3.1** *Obsérvese que si el RK $(\hat{A},\hat{b})$ tiene orden de etapa $q \geq 1$, entonces, usando el Lema IV.3.1 dado a continuación, es inmediato probar que*

$$v_j(z) = 1, \quad 1 \leq j \leq q, \quad \forall\, z \text{ con } z^{-1} \notin \sigma(\hat{A}).$$

**Lema IV.3.1** *Si un RK $(\hat{A},\hat{b})$ verifica $C(q)$, $q \geq 1$, entonces*

$$(I - z\hat{A})^{-1}(-z\hat{c}^j + j\hat{c}^{j-1}) = j\hat{c}^{j-1}, \quad 1 \leq j \leq q, \quad \forall\, z \text{ con } z^{-1} \notin \sigma(\hat{A}).$$

*Demostración.* Teniendo en cuenta que la igualdad anterior es equivalente a

$$-z\hat{c}^j + j\hat{c}^{j-1} = (I - z\hat{A})j\hat{c}^{j-1}, \quad 1 \leq j \leq q,$$

se concluye la demostración sin mas que usar la condición $C(q)$. ∎

Por otra parte, como la solución exacta del problema en $t_0 + h$ puede expresarse como

$$y(t_0 + h) = \phi(t_0 + h) = \phi(t_0) + \sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!} h^j,$$

el error global sobre $t_1$ viene dado por

$$y(t_1) - y_1 = \hat{R}(z)(\phi(t_0) - y_0) + d_h(t_0)$$

donde

$$d_h(t_0) = \sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!}(1 - v_j(z))h^j$$

es el error local y el término $\hat{R}(z)(\phi(t_0)-y_0)$ representa la propagación del error del paso anterior.

Si el método RK $(\hat{A}, \hat{b})$ es ASI–estable [6], [13], es decir, $I - z\hat{A}$ es no singular para todo $z$ con $Re\ z \leq 0$ y

$$\sup_{Re(z)\leq 0} |(I - z\hat{A})^{-1}|_2 < \infty,$$

y AS–estable [6, 13], esto es,

$$\sup_{Re(z)\leq 0} |z\hat{b}^T(I - z\hat{A})^{-1}|_2 < \infty,$$

está claro que los coeficientes $v_j(z),\ j \geq 1$, están acotados uniformemente en $Re\ z \leq 0$ y el orden del error local $d_h(t_0)$ es independiente de la stiffness del problema.

Por otro lado, cada etapa interna $Y_i$ de (4.1.4) puede interpretarse como un método RK $(\bar{A}, p_i)$ con la tabla de Butcher dada en (4.2.5). Luego, para el modelo de Prothero y Robinson tenemos para $z$ con $z^{-1} \notin \sigma(\bar{A})$:

$$\left.\begin{array}{c} Y_i = \mathcal{R}_i(z)(y_0 - \phi(t_0)) + \phi(t_0) + \displaystyle\sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!} v_{i,j}(z)h^j, \\ v_{i,j}(z) = -z\mu_{i,j} + j\mu_{i,j-1}, \quad j \geq 1, \\ \mu_{i,k} = p_i^T(I - z\bar{A})^{-1}\bar{c}^k, \quad k \geq 0 \\ \mathcal{R}_i(z) = 1 + zp_i^T(I - z\bar{A})^{-1}e. \end{array}\right\} \tag{4.3.5}$$

Si el RK $(A, b)$ es AS y ASI–estable, entonces $\mathcal{R}_i(z)$ está uniformemente acotada en $Re\ z \leq 0$. Es más, un cálculo directo nos da

$$(I - z\bar{A})^{-1} = \begin{pmatrix} (I - zA)^{-1} & \mathbf{0} \\ u(r, z) & (I - zrA)^{-1} \end{pmatrix}, \quad u(r, z) = z(I - zrA)^{-1}eb^T(I - zA)^{-1}. \tag{4.3.6}$$

De aquí, denotando $A_i^T = e_i^T A$, en virtud de que

$$1 + zA_i^T(I - zA)^{-1}e = e_i^T(I + zA(I - zA)^{-1})e = e_i^T(I - zA)^{-1}e,$$

se sigue que

$$\mathcal{R}_i(z) = R(z)T_i(rz) \tag{4.3.7}$$

con

$$\left\{\begin{array}{l} R(z) = 1 + zb^T(I - zA)^{-1}e \quad \text{(función de amplificación del RK)}, \\ T_i(z) = e_i^T(I - zA)^{-1}e. \end{array}\right. \tag{4.3.8}$$

**Nota IV.3.2** *Obsérvese que si el método RK $(A,b)$ de $s$ etapas tiene orden de etapa $q \geq 1$, de forma inmediata se tiene que el RK $(\bar{A}, p_i)$, $(1 \leq i \leq s)$, de $2s$ etapas, verifica $C(q)$. Por tanto, de (4.3.5) usando los Lemas IV.2.2 y IV.3.1, se tiene, para $z$ con $z^{-1} \notin \sigma(\bar{A})$,*

$$\begin{aligned} v_{i,j}(z) &= p_i^T(I - z\bar{A})^{-1}(-z\bar{c}^j + j\bar{c}^{j-1}) = jp_i^T\bar{c}^{j-1} \\ &= j\Delta_{ij}(r) = (1 + rc_i)^j, \quad 1 \leq i \leq s, \quad 1 \leq j \leq q. \end{aligned} \tag{4.3.9}$$

Además, bajo las condiciones de AS-ASI estabilidad se puede asegurar (como veremos a continuación) que

$$\sup_{Re(z)\leq 0} \max_{1\leq i\leq s} |v_{i,j}(z)| \leq K_j < \infty, \quad j \geq 1. \tag{4.3.10}$$

Para probar esto, es suficiente ver que $|z\mu_{i,j}|$ está uniformemente acotada en $Re\ z \leq 0$ para cualquier $1 \leq i \leq s$ y para cada $j \geq 1$ fijo. Para ello, a su vez es suficiente demostrar que la norma euclídea de la siguiente matriz de dimensión $s \times 2s$

$$N(z) \equiv z[eb^T, rA](I - z\bar{A})^{-1}$$

está uniformemente acotada en $Re\ z \leq 0$. Después de algunos cálculos inmediatos y de usar (4.3.6) obtenemos

$$N(z) = [e(zb^T(I - zA)^{-1}) + (zrA(I - zrA)^{-1})e(zb^T(I - zA)^{-1})\ ,\ zrA(I - zrA)^{-1}].$$

Teniendo en cuenta la AS y ASI-estabilidad del método RK $(A,b)$, se obtiene la cota deseada.

Por otro lado, la aproximación $Y_i^0$ dada por el inicializador de Tipo II (4.1.6) puede considerarse como la aproximación dada por el método RK $(\tilde{A}, \tilde{p}_i)$, que tiene la tabla de Butcher (4.2.4) y que por tanto, para $z$ con $z^{-1} \notin \sigma(\tilde{A})$, tiene el desarrollo

$$\left.\begin{aligned} Y_i^0 = \tilde{R}_i(z)(y_0 - \phi(t_0)) + \phi(t_0) + \sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!}\tilde{v}_{i,j}(z)h^j, \\ \tilde{v}_{i,j}(z) = -z\tilde{\mu}_{i,j} + j\tilde{\mu}_{i,j-1}, \quad j \geq 1, \\ \tilde{\mu}_{i,k} = \tilde{p}_i^T(I - z\tilde{A})^{-1}\tilde{c}^k, \quad k \geq 0 \\ \tilde{R}_i(z) = 1 + z\tilde{p}_i^T(I - z\tilde{A})^{-1}e \end{aligned}\right\} \tag{4.3.11}$$

Teniendo en cuenta el valor de $\tilde{p}_i$ en (4.2.4), la función de amplificación de error se puede expresar en términos de los coeficientes del inicializador como

$$\tilde{R}_i(z) = 1 + z\delta_i + z\beta_i^T(I - zA)^{-1}e, \tag{4.3.12}$$

y los coeficientes $\tilde{\mu}_{i,j}$ como

$$\begin{aligned} \tilde{\mu}_{i,0} &= \delta_i + \beta_i^T(I - zA)^{-1}e \\ \tilde{\mu}_{i,j} &= \beta_i^T(I - zA)^{-1}c^j, \quad \text{para } j \geq 1. \end{aligned}$$

**Nota IV.3.3** *No es difícil ver que si el método RK $(A,b)$ de $s$ etapas tiene orden de etapa $q \geq 1$, entonces el método $(\tilde{A}, \tilde{p}_i)$, $(1 \leq i \leq s)$, de $s+1$ etapas, también verifica $C(q)$. Por tanto, de (4.3.11) aplicando el Lema IV.3.1, se tiene, para $z$ con $z^{-1} \notin \sigma(\tilde{A})$,*

$$\begin{aligned} \tilde{v}_{i,j}(z) &= \tilde{p}_i^T(I - z\tilde{A})^{-1}(-z\tilde{c}^j + j\tilde{c}^{j-1}) = j\tilde{p}_i^T\tilde{c}^{j-1} \\ &= j(\delta_{j1}\delta_i + \beta_i^T c^{j-1}), \quad 1 \leq i \leq s, \quad 1 \leq j \leq q, \end{aligned} \tag{4.3.13}$$

*donde $\delta_{11} = 1$, $\delta_{j1} = 0$, $2 \leq j \leq q$.*

De (4.3.11) y (4.3.5), el error $Y_i - Y_i^0$ resulta

$$Y_i - Y_i^0 = (\mathcal{R}_i(z) - \tilde{R}_i(z))(y_0 - \phi(t_0)) + \sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!}(v_{i,j}(z) - \tilde{v}_{i,j}(z))h^j.$$

Para analizar el comportamiento de los inicializadores nos vamos a centrar en tres aspectos:

1. La acotación de los coeficientes $\tilde{v}_{i,j}(z)$.

2. El orden del inicializador, esto es, el máximo $j \geq 1$ tal que $v_{i,j}(z) - \tilde{v}_{i,j}(z) = 0$ para todo $i = 1, \ldots, s$ y todo $z$ con $Re\ z \leq 0$.

3. La acotación de la diferencia $\mathcal{R}_i(z) - \tilde{R}_i(z)$.

Diremos pues que un inicializador Tipo II tiene *orden q sobre la ecuación de Prothero y Robinson* sii

$$\tilde{v}_{i,j}(z) - v_{i,j}(z) \equiv 0, \quad 1 \leq i \leq s, \quad 1 \leq j \leq q, \quad \forall\ Re\ z \leq 0$$

y

$$\sup_{Re(z)\leq 0} |\tilde{v}_{i,j}(z)| < \infty, \quad 1 \leq i \leq s, \quad j \geq 1.$$

De este modo desligamos el orden del inicializador de su función de amplificación de error, la cual será estudiada separadamente.

**Teorema IV.3.1** *Supongamos que un método RK $(A, b)$ es ASI–estable. Entonces, los coeficientes $\tilde{v}_{i,j}(z)$ de (4.3.11) satisfacen*

$$\sup_{Re(z)\leq 0} \max_{1\leq i\leq s} |\tilde{v}_{i,j}(z)| \leq \tilde{K}_j < \infty, \quad \forall j \geq 1,$$

*si se da alguna de las siguientes condiciones:*

*a) Los coeficientes $\beta_{ij}$ del inicializador de Tipo II son de la forma $\beta_i^T = \alpha_i^T A$ $(i = 1, \ldots, s)$ para ciertos vectores $\alpha_i^T$.*

*b) El método RK es no confluente y verifica $C(s-1)$ con $c_k = 0$ para algún $1 \leq k \leq s$.*

<u>*Demostración.*</u> Si $\beta_i^T = \alpha_i^T A$ entonces

$$z\tilde{\mu}_{i,j} = z\beta_i^T (I - zA)^{-1} c^j = z\alpha_i^T A(I - zA)^{-1} c^j$$

y como $zA(I - zA)^{-1} = ((I - zA)^{-1} - I)$, la ASI–estabilidad nos da la cota buscada.

Si $c_k = 0$ para algún $k$, se tiene que para todo $j \geq s$, el vector $c^j$ puede escribirse como una combinación lineal de los $s - 1$ vectores $c, c^2, \ldots, c^{s-1}$. Consecuentemente, $\tilde{\mu}_{i,j}$ puede ponerse como una combinación lineal de $\tilde{\mu}_{i,1}, \ldots, \tilde{\mu}_{i,s-1}$. Pero por $C(s-1)$, $c^j = jAc^{j-1}$ para $1 \leq j \leq s - 1$ y por tanto,

$$z\tilde{\mu}_{i,j} = j\beta_i^T ((I - zA)^{-1} - I)c^{j-1}, \quad 1 \leq j \leq s - 1.$$

Luego tenemos acotación uniforme en $Re\ z \leq 0$ por la ASI-estabilidad. De aquí se obtiene la acotación uniforme en $Re\ z \leq 0$ de $|\tilde{v}_{ij}|$ para $1 \leq i \leq s$ y para cualquier $j \geq 1$ fijo. ∎

**Nota IV.3.4** .

- *Los inicializadores de Tipo I son un caso particular de los de Tipo II con* $\beta_i^T = \alpha_i^T A$. *Luego para esta clase de inicializadores los coeficientes* $\tilde{v}_{i,j}$ *están unifomemente acotados para todo* $1 \leq i \leq s$ *y* $j \geq 1$.
- *Si la matriz* $A$ *del RK es no singular, para cada* $\beta_i^T$ *existe un único vector* $\alpha_i^T$ *tal que* $\beta_i^T = \alpha_i^T A$. *Como consecuencia, en este caso, cada inicializador de Tipo II tiene sus coeficientes* $\tilde{v}_{i,j}$ *uniformemente acotados.*

**Lema IV.3.2** *Si un RK* $(A, b)$ *de* $s$ *etapas no confluente verifica* $B(q)$ *y* $C(q)$ *(*$1 \leq q \leq s$*), entonces un inicializador de Tipo II con coeficientes* $\tilde{v}_{i,j}$ *uniformemente acotados tiene orden* $q$ *sobre la ecuación de Prothero y Robinson si y sólo si*

$$\delta_i + \beta_i^T e = (1 + rc_i), \quad \beta_i^T c^{j-1} = \Delta_{ij}(r) = (1 + rc_i)^j / j, \quad j = 2, \ldots, q, \quad i = 1, \ldots, s. \quad (4.3.14)$$

Demostración. En virtud de que los coeficientes $\tilde{v}_{i,j}$ están uniformemente acotados y que el RK $(A, b)$ tiene orden de etapa $q$, el lema se demuestra inmediatamente igualando las ecuaciones (4.3.9) y (4.3.13). ∎

**Teorema IV.3.2** *Si un RK* $(A, b)$ *de* $s$ *etapas no confluente verifica*

*(i) las condiciones simplificadoras* $B(s-1)$ *y* $C(s-1)$,

*(ii) es ASI–estable y AS–estable,*

*entonces*

*(a) La familia* $s$*–paramétrica de inicializadores de Tipo I de orden clásico* $s-1$ *dada en el apartado (a) del Teorema IV.2.1 tiene también orden* $s-1$ *sobre la ecuación de Prothero y Robinson.*

*(b) Si* $A$ *es no singular existe un único inicializador de la familia anterior que alcanza además orden* $s$ *sobre cuadraturas, i.e., para el caso* $\lambda = 0$ *(sus coeficientes vienen dados por (4.2.7) y (4.2.17)). Si además se verifican* $B(s)$ *y* $C(s)$ *dicho inicializador alcanza orden máximo* $s$ *sobre la ecuación de Prothero y Robinson, y sus coeficientes están dados por (4.2.9).*

Demostración. *(a)* Es una consecuencia inmediata del Teorema IV.3.1, la Nota IV.3.4 y el Lema IV.3.2 con $q = s - 1$.

*(b)* Si la matriz $A$ es no singular, las ecuaciones para alcanzar orden $s-1$ para todo $Re\, z \leq 0$ y orden $s$ para $z = 0$ son las dadas en (4.2.7) y (4.2.17). Si además imponemos al método las condiciones $B(s)$ y $C(s)$, usando el Lema IV.3.2 con $q = s$ y procediendo como en el Teorema IV.2.1–*(c)* se concluye fácilmente la demostración. ∎

**Teorema IV.3.3** *Supongamos que estamos bajo las mismas hipótesis que en el Teorema IV.3.2 y que además se verifica una de las condiciones a) o b) del Teorema IV.3.1. Entonces para los inicializadores de Tipo II tenemos:*

*(a) Existe una familia* $2s$*–paramétrica de inicializadores con orden* $s-1$ *sobre la ecuación de Prothero y Robinson. Si además buscamos orden* $s$ *sobre cuadraturas (i.e., para* $z = 0$*) obtenemos la familia* $s$*–paramétrica dada en el Teorema IV.2.2-(a).*

*(b) Si se verifican las condiciones $C(s)$ y $B(s)$, la familia $s$–paramétrica anterior alcanza orden $s$ sobre la ecuación de Prothero y Robinson. Además, si elegimos $\delta_i = 0,\ i = 1, \ldots, s$, los coeficientes del único inicializador de orden máximo $s$ vienen dados por*

$$\beta_{ij} = \int_0^{1+rc_i} \bar{l}_j(t)dt, \quad 1 \le i, j \le s,$$

*donde $\{\bar{l}_j(t); j = 1, \ldots, s\}$ es la base de Lagrange asociada a los nodos $\{c_1, \ldots, c_s\}$, como en el Teorema IV.2.2.*

*(c) Suponiendo $C(s)$, $B(s)$ y que la matriz $A$ es no singular, el orden $s+1$ sobre la ecuación de Prothero y Robinson, no se puede alcanzar. Sin embargo, si pedimos orden $s$ en general y orden $s+1$ en los puntos particulares $z = 0$ ó $z = \infty$, podemos obtener un inicializador único para cada caso.*

Demostración. *(a)* Del Lema IV.3.2 el orden $s-1$, independiente de la stiffness, es equivalente a (4.3.14) con $q = s-1$ (lo que es equivalente también a (4.2.1)-(4.2.2) con $q = s-2$). Por tanto, las condiciones de orden se reducen a un sistema lineal de $s-1$ ecuaciones con $s+1$ incógnitas (para cada $i = 1, \ldots, s$), que puede escribirse como

$$\beta_i^T V^* = u_i^{*T} - \delta_i e_{s-1,1}^T, \quad V^* = [e, c, ..., c^{s-2}],$$

$$u_i^{*T} = (1 + rc_i, ..., (1 + rc_i)^{s-1}/(s-1)), \quad e_{s-1,1}^T = (1, 0, ..., 0) \in I\!R^{s-1}.$$

Como el rango de la matriz $V^*$ es $s-1$ obtenemos una familia $2s$–paramétrica de inicializadores.

Por otra parte, si pedimos orden $s$ para cuadraturas, i.e., para $z = 0$, tenemos de (4.3.5) y (4.3.11) que

$$\begin{aligned} v_{i,s}(0) - \tilde{v}_{i,s}(0) = &\ s\ (p_i^T \bar{c}^{s-1} - \tilde{p}_i^T \tilde{c}^{s-1}) \\ = &\ s\left(b^T c^{s-1} + rA_i^T (e + rc)^{s-1} - \beta_i^T c^{s-1}\right) = 0. \end{aligned}$$

Luego se verifican las ecuaciones (4.2.1)–(4.2.2) para $q = s$. Esto nos lleva a la misma familia $s$–paramétrica de inicializadores del Teorema IV.2.2–*(a)*.

*(b)* Si suponemos $B(s)$ y $C(s)$, las condiciones para orden stiff $s$ son equivalentes a (4.3.14) con $q = s$. De aquí obtenemos una familia $s$–paramétrica con los $\delta_i,\ (i = 1, ..., s)$ como parámetros. Si elegimos $\delta_i = 0,\ i = 1, ..., s$, tenemos el inicializador dado en el Teorema IV.2.2-*(a)*.

*(c)* Si además la matriz $A$ es no singular, para obtener orden $s+1$ para $z = 0$ es necesario que

$$v_{i,s+1}(0) - \tilde{v}_{i,s+1}(0) = (s+1)(b^T c^s + rA_i^T (e + rc)^s - \beta_i^T c^s) = 0,$$

lo que nos lleva a las ecuaciones de orden (4.2.1)-(4.2.2) con $q = s$. La existencia y unicidad de solución, que viene dada en (4.2.25), está asegurada por la no singularidad de $A$.

Para estudiar el orden $s+1$ para $z = \infty$, es fácil ver que

$$v_{i,s+1}(z) - \tilde{v}_{i,s+1}(z) = d_i^T (I - z\bar{A})^{-1}(-z\bar{c}^{s+1} + (s+1)\bar{c}^s),$$

con $d_i^T = (b^T - \beta_i^T, rA_i^T)$. Por tanto, después de algunos cálculos directos obtenemos que $v_{i,s+1}(\infty) - \tilde{v}_{i,s+1}(\infty) = 0$ si y sólo si

$$\beta_i^T A^{-1} c^{s+1} = b^T A^{-1} c^{s+1} - A_i^T A^{-1} e b^T A^{-1} c^{s+1} + A_i^T A^{-1} (e + rc)^{s+1}.$$

Teniendo en cuenta que $A_i^T A^{-1} = (0, ..., 0, \overset{(i)}{1}, 0, ..., 0)$, lo anterior es equivalente a

$$\beta_i^T A^{-1} c^{s+1} = (1 + rc_i)^{s+1}, \quad 1 \leq i \leq s. \tag{4.3.15}$$

Por tanto el orden $s$ en general junto con orden $s+1$ en el infinito, es equivalente a (4.3.14) con $q = s$ y (4.3.15). No es difícil ver que estas ecuaciones tienen solución única, dada por

$$\begin{cases} \delta_i = \dfrac{\bar{u}_i^T V^{-1} A^{-1} c^{s+1} - (1 + rc_i)^{s+1}}{e_1^T V^{-1} A^{-1} c^{s+1}}, \qquad \beta_i^T = \bar{u}_i^T V^{-1} - \delta_i e_1^T V^{-1} \\ \qquad V = [e, c, \ldots, c^{s-1}], \qquad \bar{R}_i^T = (1 + rc_i, \ldots, (1 + rc_i)^s / s). \end{cases} \tag{4.3.16}$$

Para ver que no se puede alcanzar orden stiff $s + 1$ en general sobre el modelo de Prothero y Robinson, vamos a proceder por contradicción. Si existiera un inicializador de Tipo II con orden $s + 1$, entonces alcanzaría ese orden para $z = 0$ y para $z = \infty$ simultáneamente. Por tanto, después de igualar las expresiones de $\delta_i$ en (4.3.16) y (4.2.25) (denotando $K_1 = e_1^T V^{-1} c^s$ y $K_2 = e_1^T V^{-1} A^{-1} c^{s+1}$) obtendríamos

$$K_2(\bar{u}_i^T V^{-1} A^{-1} c^{s+1} - (1 + rc_i)^{s+1}) = K_1(\bar{u}_i^T V^{-1} c^s - \Delta_{i,s+1}(r)),$$

esto es,

$$\bar{u}_i^T V^{-1}(K_2 A^{-1} c^{s+1} - K_1 c^s) = K_2(1 + rc_i)^{s+1} - K_1(b^T c^s + r A_i^T (e + rc)^s), \quad \forall r > 0. \tag{4.3.17}$$

Comparando los coeficientes principales de los polinomios en $r$ de esta última igualdad resulta que $K_2 c_i^{s+1} - K_1 A_i^T c^s = 0$ para cada $i = 1, ..., s$, i.e., $K_2 c^{s+1} = K_1 A c^s$, o lo que es lo mismo, $K_2 A^{-1} c^{s+1} = K_1 c^s$. Luego el primer miembro de (4.3.17) es nulo, por lo que podemos poner

$$K_2(1 + rc_i)^{s+1} - K_1(b^T c^s + r A_i^T (e + rc)^s) = 0, \quad \forall r > 0,$$

o de forma equivalente,

$$(K_2 - K_1 b^T c^s) + \sum_{j=1}^{s} \binom{s+1}{j} r^j c_i^j \left( K_2 - \frac{K_1}{s+1} \right) + r^{s+1}(K_2 c_i^{s+1} - K_1 A_i^T c^s) \equiv 0.$$

Esto implicaría $B(s + 1)$ y $C(s + 1)$, ya que $K_1$ es no nulo (ver Lema IV.2.3 y recordar que la matriz $A$ es no singular). Esto nos lleva a una contradicción pues las condiciones $B(s + 1)$ y $C(s + 1)$ simultáneamente no pueden alcanzarse para ningún método RK de $s$ etapas, como se prueba en el siguiente lema. ∎

**Lema IV.3.3** *Un método RK $(A, b)$ de colocación, con $s$ etapas verifica* $\det(A) = (\Pi_{j=1}^s c_j)/s!$. *Además, si se verifica $C(s + 1)$, el método debe ser Euler explícito.*

*Demostración.* Ya que un RK de colocación verifica $C(s)$ y $B(s)$ [36, pág. 212], es inmediato comprobar que

$$AV = CVD,$$

donde las matrices $V$, $C$ y $D$ están definidas por

$$V = [e, c, \ldots, c^{s-1}], \quad C = \operatorname{diag}(c_1, c_2, \ldots, c_s), \quad D = \operatorname{diag}(1, 1/2, \ldots, 1/s).$$

En virtud de que la matriz $V$ es regular, concluimos la primera parte del teorema.

Para ver la segunda parte, asumiremos en primer lugar y junto con $C(s+1)$, que todos los nodos $c_i$ sean no nulos. Entonces podemos poner

$$ACV = C^2V \text{ diag } (1/2, 1/3, \ldots, 1/(s+1)),$$

lo cual implica por la regularidad de las matrices $V$ y $C$ que

$$\det(A) = (\Pi_{j=1}^{s} c_j)/(s+1)!.$$

Esto contradice la primera proposición del teorema.

Por otra parte, si asumimos por ejemplo que el primer nodo $c_1$ es nulo y que $s > 1$, entonces considerando la submatriz $\bar{A} = [a_{ij}]_{2\le i,j\le s}$ y el vector $\bar{c}^T = (c_2, \ldots, c_s)$, podemos poner en virtud de $C(s+1)$ que

$$\bar{A}[\bar{c}, \bar{c}^2, \ldots, \bar{c}^s] = [\bar{c}^2/2, \bar{c}^3/3, \ldots, \bar{c}^{s+1}/(s+1)].$$

De aquí se deduce que

$$\bar{A}[\bar{c}, \bar{c}^2, \ldots, \bar{c}^{s-1}] = [\bar{c}^2/2, \bar{c}^3/3, \ldots, \bar{c}^s/s], \tag{4.3.18}$$

y

$$\bar{A}[\bar{c}^2, \bar{c}^3, \ldots, \bar{c}^s] = [\bar{c}^3/3, \bar{c}^4/4, \ldots, \bar{c}^{s+1}/(s+1)]. \tag{4.3.19}$$

De (4.3.18) se sigue que $\det(\bar{A}) = (\Pi_{j=2}^{s} c_j)/s!$ y de (4.3.19) se deduce que $\det(\bar{A}) = 2(\Pi_{j=2}^{s} c_j)/(s+1)!$. Esto implica $s = 1$, lo que es imposible por hipótesis. Por tanto, el único caso posible es que el método considerado sea el Euler explícito. ∎

**Comparación de las funciones de amplificación del error**

Las funciones de amplificación del error de la solución exacta $Y_i$, $\mathcal{R}_i(z)$, y la de la aproximación dada por el inicializador $Y_i^0$, $\tilde{R}_i(z)$, vienen dadas en (4.3.7)-(4.3.8) y (4.3.12) respectivamente por

$$\left.\begin{array}{c} \mathcal{R}_i(z) = R(z)T_i(rz) = (1 + zb^T(I - zA)^{-1}e)(e_i^T(I - rzA)^{-1}e) \\ \tilde{R}_i(z) = 1 + z\delta_i + z\beta_i^T(I - zA)^{-1}e. \end{array}\right\} \tag{4.3.20}$$

Para el caso particular de los inicializadores de Tipo I, se tiene que $\delta_i = 0$ y $\beta_i^T = \alpha_i^T A$. Por tanto, sus funciones de amplificación $\hat{R}_i(z)$ resultan

$$\hat{R}_i(z) := \tilde{R}_i(z) = 1 + z\alpha_i^T A(I - zA)^{-1}e. \tag{4.3.21}$$

Si el método RK es AS y ASI–estable está claro que $\mathcal{R}_i(z)$ y $\hat{R}_i(z)$ están uniformemente acotadas en $Re\ z \le 0$. Sin embargo, la funciones de amplificación de los inicializadores de Tipo II en general no estarán acotadas si algún $\delta_i \neq 0$ y $A$ es no singular. Si la matriz $A$ fuera singular, es posible la acotación uniforme de $\tilde{R}_i(z)$, pero en ese caso, $\delta_i$ y $\beta_i^T$ deberán verificar cierta condición que veremos en el siguiente teorema para el caso del método Lobatto IIIA.

Este hecho podría llevarnos a descartar los inicializadores de Tipo II con $\delta_i \neq 0$. Por otra parte debemos observar que si el RK $(A, b)$ es stiffly accurate, A–estable y su matriz $A$ es no singular, los errores globales para la ecuación de Prothero y Robinson satisfacen $\phi(t_n) - y_n = \mathcal{O}(z^{-1})$ $(z \to \infty)$ (ver [37, pág. 225-227]) y, en consecuencia, el producto $\tilde{R}(z)(\phi(t_n) - y_n)$ está acotado cuando $z$ tiende a infinito. Por tanto, los inicializadores de Tipo II con funciones de amplificación no acotadas pudieran ser apropiados en la práctica cuando se usan con métodos stiffly accurate, como es el caso de los familias Radau IIA, Lobatto IIIC o Lobatto IIIA.

El siguiente teorema estudia el comportamiento de las funciones de amplificación del error en el infinito para diferentes métodos RK basados en fórmulas de cuadratura de alto grado de precisión.

**Teorema IV.3.4** *(a) Para los RK Gauss, Radau IA, Radau IIA y Lobatto IIIC de $s$ etapas (y en general para cualquier RK con $A$ no singular) tenemos para $1 \le i \le s$ que*

$$\mathcal{R}_i(\infty) = 0, \quad \hat{R}_i(\infty) = \gamma_i, \quad \tilde{R}_i(\infty) = \begin{cases} 1 - \beta_i^T A^{-1} e & \text{si } \delta_i = 0 \\ \infty & \text{si } \delta_i \neq 0. \end{cases}$$

*(b) Para el método Lobatto IIIA de $s$ etapas, denotando*

$$A = \begin{pmatrix} 0 & \mathbf{0}^T \\ w & \bar{A} \end{pmatrix}, \quad w^T = (a_{21}, \ldots, a_{s1}), \quad \bar{A} = [a_{ij}]_{i,j=2}^{s},$$

$$\bar{b}^T = (b_2, ..., b_s), \ \bar{c} = (c_2, ..., c_s)^T, \ e_{s-1,j}^T = (0, ..., \overset{(j)}{1}, ..., 0) \in I\!R^{s-1}, \ \beta_i^T = (\beta_{i1}, \bar{\beta}_i^T)$$

*tenemos para cada $2 \le i \le s$ que*

$$\mathcal{R}_i(\infty) = (-1)^s e_{s-1,i-1}^T \bar{A}^{-1} w, \quad \hat{R}_i(\infty) = 1 - \bar{\alpha}_i^T \bar{A}^{-1} \bar{c}, \tag{4.3.22}$$

$$\tilde{R}_i(\infty) = \begin{cases} 1 - \bar{\beta}_i^T \bar{A}^{-2} \bar{c}, & \text{si } \delta_i = -\beta_{i1} + \bar{\beta}_i^T \bar{A}^{-1} w \\ \infty, & \text{en otro caso.} \end{cases} \tag{4.3.23}$$

*Demostración. (a)* Se tiene tomando límites cuando $z \to \infty$ en las expresiones de las funciones de amplificación.

*(b)* Para el Lobatto IIIA, como $A$ es singular, necesitamos algunos cálculos adicionales. Directamente se tiene

$$(I - zA)^{-1} = \begin{pmatrix} 1 & \mathbf{0}^T \\ z(I - z\bar{A})^{-1} w & (I - z\bar{A})^{-1} \end{pmatrix}, \quad (I - zA)^{-1} e = \begin{pmatrix} 1 \\ (I - z\bar{A})^{-1}(zw + e) \end{pmatrix}.$$

Como $\bar{A}$ es no singular y $\bar{A}e + w = \bar{c}$, se sigue que

$$(I - z\bar{A})^{-1}(zw + e) = -\bar{A}^{-1} w - z^{-1} \bar{A}^{-2} \bar{c} + \mathcal{O}(z^{-2}). \tag{4.3.24}$$

Por otro lado, para este método es bien sabido que su función de estabilidad lineal $R(z)$ verifica $R(\infty) = (-1)^{s-1}$. Además,

$$T_i(z) = e_{s,i}^T (I - zA)^{-1} e = e_{s-1,i-1}^T (I - z\bar{A})^{-1}(zw + e), \quad 2 \le i \le s.$$

De esto se deduce inmediatamente que

$$\mathcal{R}_i(\infty) = R(\infty) T_i(\infty) = (-1)^s e_{s-1,i-1}^T \bar{A}^{-1} w, \quad 2 \le i \le s.$$

Para los inicializadores de Tipo I tenemos, usando la notación $\alpha_i^T = (\alpha_{i1}, \bar{\alpha}_i^T)$, que

$$\hat{R}_i(z) = 1 + z\alpha_i^T A (I - zA)^{-1} e = 1 + z\bar{\alpha}_i^T \left( w + \bar{A}(I - z\bar{A})^{-1}(zw + e) \right).$$

De (4.3.24) se sigue que $\hat{R}_i(\infty) = 1 - \bar{\alpha}_i^T \bar{A}^{-1} \bar{c}, \quad 2 \le i \le s.$
Para los inicializadores de Tipo II resulta

$$\tilde{R}_i(z) = 1 + z\delta_i + z(\beta_{i1}, \bar{\beta}_i^T)(I - zA)^{-1} e = 1 + z\delta_i + z\beta_{i1} + z\bar{\beta}_i^T (I - z\bar{A})^{-1}(zw + e). \tag{4.3.25}$$

De (4.3.24) se deduce que

$$\tilde{R}_i(z) = z(\delta_i + \beta_{i1} - \bar{\beta}_i^T \bar{A}^{-1} w) + (1 - \bar{\beta}_i^T \bar{A}^{-2} \bar{c}) + \mathcal{O}(z^{-1}), \quad 2 \le i \le s,$$

de donde se concluye la demostración. ∎

### IV.3.2 Orden stiff sobre problemas contractivos

Vamos a estudiar el orden de los inicializadores de Tipo I y II sobre una clase más general de ecuaciones diferenciales no lineales que suelen considerarse habitualmente cuando se estudia la convergencia de los métodos Runge–Kutta o los métodos multipaso (ver por ejemplo [24], [19], [37]).

Consideremos la clase de PVI *contractivos*, esto es, los problemas para los cuales la función derivada $f$ verifica una condición de Lipschitz por un lado

$$\langle f(t,y) - f(t,z), y - z\rangle \leq 0, \quad \forall t \in [0,T], \quad \forall y, z \in I\!R^m. \tag{4.3.26}$$

Convenimos pues que $f \in \mathcal{F}_0$ sii $f$ verifica la condición anterior.

También consideraremos que las derivadas de la solución exacta $y(t)$ del PVI (4.1.1) verifican

$$|y^{(l)}(t)| \leq M_l = \mathcal{O}(1), \quad \forall t \in [0,T], \quad l = 1, \ldots, q+1, \tag{4.3.27}$$

donde $q$ es un natural (la norma considerada es la norma euclídea asociada al producto escalar $|v|^2 = \langle v, v\rangle$). Generalmente $q$ será el orden de etapa del RK considerado.

Estamos ahora en condiciones de definir el orden stiff sobre problemas contractivos. Diremos pues que un inicializador $Y^0$ tiene *orden stiff* $q$ sii

$$\max_{1\leq i\leq s} |Y_i - Y_i^0| \leq K h^{q+1}, \quad \forall h > 0,$$

cuando se considera $f \in \mathcal{F}_0$ e $y_0 = y(t_0)$. La constante $K$ sólo dependerá de las cantidades $M_l$ $(1 \leq l \leq q+1)$, de los coeficientes del método RK considerado y de la razón de paso $r$, siendo siempre de tamaño moderado.

De la definición anterior se observa, para los inicializadores considerados, que el orden stiff siempre estará acotado superiormente por el orden sobre la ecuación de Prothero y Robinson y por el orden clásico. Estableceremos de modo preciso el orden stiff para los inicializadores más interesantes estudiados previamente.

Los siguientes teoremas confirman que los resultados principales de orden dados en los Teoremas IV.3.2 y IV.3.3 para la ecuación de Prothero y Robinson son también válidos para la familia de problemas no lineales contractivos que satisfacen además (4.3.27). En el caso anterior supusimos que el método RK era AS y ASI–estable, que son propiedades de estabilidad que surgen de forma natural cuando se estudia la convergencia de los métodos sobre sistemas lineales stiff. Aquí, en cambio, para probar los resultados de orden sobre ecuaciones no lineales stiff vamos a suponer que el método RK es diagonalmente estable. i.e., existe una matriz diagonal definida positiva $D$ tal que $DA + A^T D$ es definida positiva. Nótese que esta condición implica que la matriz $A$ es no singular y que además el método es AS–estable y ASI–estable.

**Teorema IV.3.5** *Sea un RK $(A, b)$ de $s$ etapas no confluente que verifica*
*(i) $B(s-1)$, $C(s-1)$ y*
*(ii) es diagonalmente estable.*

*Entonces, la familia de inicializadores $s$–paramétrica de Tipo I de orden clásico $s-1$ dada en el apartado (a) del Teorema IV.2.1 tiene orden stiff $s-1$. Además, si se verifican $B(s)$ y $C(s)$, el único inicializador de Tipo I de orden $s$ dado por el polinomio de interpolación de Lagrange, como se indica en el apartado (c) del Teorema IV.2.1, tiene también orden stiff $s$.*

*Demostración.* Como la teoría de las series de Butcher no es válida para estudiar el orden stiff, vamos a hacerlo de una forma alternativa usando las hipótesis (4.3.26) y (4.3.27). Denotemos por

$$\hat{X}_i := y(t_0 + c_i h), \quad \hat{Y}_i := y(t_0 + (1 + rc_i)h), \quad 1 \le i \le s, \tag{4.3.28}$$

donde $y(t)$ es la solución exacta del problema de valor inicial, y definamos los vectores

$$Z_i := \gamma_i y_0 + \sum_{j=1}^{s} \alpha_{ij} \hat{X}_j, \quad 1 \le i \le s. \tag{4.3.29}$$

Con estas notaciones podemos poner

$$Y_i - Y_i^0 = \underbrace{Y_i - \hat{Y}_i}_{(1)} + \underbrace{\hat{Y}_i - Z_i}_{(2)} + \underbrace{Z_i - Y_i^0}_{(3)}. \tag{4.3.30}$$

Vamos a analizar por separado cada uno de estos sumandos.

(1) De $B(s-1)$, $C(s-1)$ y la estabilidad diagonal del método RK, procediendo por ejemplo como en la demostración del Teorema 14.3 de [37, pág. 219], podemos asegurar que $Y_i - \hat{Y}_i = \mathcal{O}(h^s)$.

(3) $Z_i - Y_i^0 = \sum_{j=1}^{s} \alpha_{ij}(\hat{X}_j - X_j) = \mathcal{O}(h^s)$, ya que los $\alpha_{ij}$ están uniformemente acotados para $r \in (0, r_0]$, con $r_0 = \mathcal{O}(1)$ y $(\hat{X}_j - X_j) = \mathcal{O}(h^s)$ (ver Teorema 14.3 en [37]).

(2) Desarrollando en potencias de $h$ en torno a $t_0$ tenemos que

$$\begin{aligned} \hat{Y}_i - Z_i = &\quad y(t_0 + (1 + rc_i)h) - \gamma_i y_0 - \sum_{j=1}^{s} \alpha_{ij} y(t_0 + c_j h) \\ = &\quad \gamma_i (y(t_0) - y_0) + \left(1 - \gamma_i - \sum_{j=1}^{s} \alpha_{ij}\right) y(t_0) + \sum_{k \ge 1} \frac{y^{(k)}(t_0)}{k!} \left((1 + rc_i)^k - \sum_{j=1}^{s} \alpha_{ij} c_j^k\right) h^k. \end{aligned}$$

Por tanto, las condiciones de orden (4.2.7) junto con la hipótesis $y(t_0) = y_0$ nos llevan a $\hat{Y}_i - Z_i = \mathcal{O}(h^s)$.

Además, asumiendo $B(s)$ y $C(s)$, resulta que para los inicializadores de orden orden $s$, los tres sumandos en (4.3.30) son $\mathcal{O}(h^{s+1})$. Se infiere de aquí el orden stiff $s$.

Por otro lado, no se puede alcanzar orden stiff $s+1$ ya que en el caso particular del modelo de Prothero y Robinson es imposible alcanzarlo, tal como se vio en el Teorema IV.3.2. ∎

**Teorema IV.3.6** *Si consideramos un RK $(A, b)$ de $s$ etapas no confluente y suponemos las hipótesis (i)-(ii) del Teorema IV.3.5, entonces se tiene para los inicializadores de Tipo II que:*

*(a) La familia $2s$–paramétrica considerada en el Teorema IV.3.3-(a) tiene orden stiff $s - 1$.*

*(b) Si también se tiene $B(s)$ y $C(s)$, existe una familia $s$–paramétrica de inicializadores con orden stiff $s$, que viene dada por las ecuaciones (4.2.22) (tomando como parámetros a $\delta_i\,(i = 1, \ldots, s)$ por ejemplo). Si elegimos $\delta_i = 0\,(i = 1, \ldots, s)$, entonces los coeficientes $\beta_{ij}$ vienen dados por la interpolación polinómica de Lagrange como se indica en el Teorema IV.3.3-(b).*

*Demostración.* Usando las notaciones dadas en (4.3.28) y reemplazando $Z_i$ en (4.3.29) por

$$Z_i = y_0 + h\delta_i f(t_0, y_0) + h\sum_{j=1}^{s} \beta_{ij} f(t_0 + c_j h, \hat{X}_j), \tag{4.3.31}$$

podemos poner

$$Y_i - Y_i^0 = \underbrace{Y_i - \hat{Y}_i}_{(1)} + \underbrace{\hat{Y}_i - Z_i}_{(2)} + \underbrace{Z_i - Y_i^0}_{(3)}.$$

(1) $Y_i - \hat{Y}_i = \mathcal{O}(h^s)$ usando los mismos argumentos que en la demostración del Teorema IV.3.5.
(3) Usando la acotación uniforme de los $\beta_{ij}$ y el Teorema 14.3 de [37, pág. 219] (recordando que $A$ es no singular por la hipótesis *(ii)*) es fácil ver que

$$Z_i - Y_i^0 = h\sum_{j=1}^{s} \beta_{ij}(f(t_0 + c_j h, \hat{X}_j) - f(t_0 + c_j h, X_j)) = \mathcal{O}(h^s).$$

(2) Considerando los desarrollos de Taylor de $y'(t_0 + c_j h)$ e $y(t_0 + (1 + rc_i)h)$ en torno a $t_0$, llegamos a

$$\begin{aligned}\hat{Y}_i - Z_i \quad &= y(t_0) - y_0 + h\delta_i(y'(t_0) - f(t_0, y_0)) \\ &+ \left(1 + rc_i - \delta_i - \sum_{j=1}^{s} \beta_{ij}\right) h y'(t_0) + \sum_{k\geq 2} \frac{y^{(k)}(t_0)}{k!} \left((1 + rc_i)^k - k\sum_{j=1}^{s} \beta_{ij} c_j^{k-1}\right) h^k.\end{aligned}$$

Usando de nuevo la hipótesis $y(t_0) = y_0$, resulta para los inicializadores de orden $s - 1$ que $\hat{Y}_i - Z_i = \mathcal{O}(h^s)$. De aquí se concluye el orden stiff $s - 1$.

*(b)* La demostración es similar a la del apartado *(a)*. ∎

A continuación vamos a dar un teorema que nos da el orden de los inicializadores sobre problemas stiff contractivos cuando $y_0 \neq y(t_0)$, es decir cuando no partimos en $t_0$ de la solución exacta del PVI (4.1.1), sino de la solución numérica obtenida por el método considerado al avanzar la integración hasta ese punto. Obsérvese que ésta es la situación real que se da en la práctica. Es de esperar que la aproximación obtenida por los inicializadores tras $n$ pasos dados por el método quede afectada por el orden de convergencia (B-convergencia) del método.

Para tal fin vamos a introducir las notaciones adecuadas. Dada una partición cualquiera $\mathcal{P}$ del intervalo $[0, T]$, con puntos de red

$$\mathcal{P}: \quad 0 = \tau_0 < \tau_1 < \ldots < \tau_n < \ldots < \tau_N = T, \tag{4.3.32}$$

definimos

$$h_n = \tau_n - \tau_{n-1}, \quad \bar{h}_n = \max\{h_j,\ j = 1, \ldots, n\}, \quad r_n = h_{n+1}/h_n, \quad n = 1, 2, \ldots, N. \tag{4.3.33}$$

Denotaremos la $i-$ésima etapa del paso $n$ por $Y_{i,n}$, teniéndose por tanto que el método Runge–Kutta avanza mediante las fórmulas

$$\left.\begin{aligned} Y_{i,n} &= y_{n-1} + h_n \textstyle\sum_{j=1}^{s} a_{ij} f(\tau_{n-1} + c_j h_n, Y_{j,n}), \quad i = 1, \ldots, s \\ y_n &= y_{n-1} + h_n \textstyle\sum_{j=1}^{s} b_j f(\tau_{n-1} + c_j h_n, Y_{j,n}) \end{aligned}\right\} \ n = 1, 2, \ldots, N \tag{4.3.34}$$

También denotaremos por $Y^0_{i,n}$ la aproximación inicial considerada (Tipo I o Tipo II), correspondiente a la etapa $Y_{i,n}$. Así, si por ejemplo consideramos inicializadores de Tipo I se tendría de acuerdo con las notaciones usadas que (téngase en cuenta que siempre $Y^0_{i,1} = y_0,\ 1 \le i \le s$)

$$Y^0_{i,n} = \gamma_i y_{n-2} + \sum_{j=1}^{s} \alpha_{ij} Y_{j,n-1},\ \ i = 1, ..., s, \quad n = 2, 3, \ldots, N. \tag{4.3.35}$$

Si usamos inicializadores de Tipo II, tendríamos

$$\begin{aligned} Y^0_{i,n} = \ & y_{n-2} + h_{n-1}\delta_i f(\tau_{n-2}, y_{n-2}) \\ & + h_{n-1} \sum_{j=1}^{s} \beta_{ij} f(\tau_{n-2} + c_j h_{n-1}, Y_{j,n-1}), \quad i = 1, ..., s, \quad n = 2, 3, \ldots, N. \end{aligned} \tag{4.3.36}$$

Naturalmente en cualquiera de los dos casos los coeficientes $\gamma_i$, $\alpha_{ij}$, $\delta_i$, $\beta_{ij}$ dependerán de la razón de paso $r_{n-1} = h_n/h_{n-1}$, pero ésta se supondrá siempre de tamaño moderado, tal como hemos asumido desde el principio del capítulo. Por lo tanto, los inicializadores considerados en el siguiente teorema tendrán sus coeficientes uniformemente acotados, tal como ocurre para los inicializadores previamente considerados en este capítulo.

**Teorema IV.3.7** *Sea un RK(A,b) no confluente de s etapas, algebraicamente estable, diagonalmente estable y que verifique $B(q)$ y $C(q)$. Sea un PVI (4.1.1) contractivo cualquiera cuya solución exacta $y(t)$ cumple las acotaciones dadas en (4.3.27). Entonces se tiene lo siguiente:*

*(a) Los inicializadores de Tipo I con orden stiff $l-1$ satisfacen*

$$\max_{1 \le i \le s} |Y_{i,n} - Y^0_{i,n}| \le K(\bar{h}_{n-1})^{\mu}, \quad \mu = \min\{q, l\}, \quad n = 2, 3, \ldots, N \tag{4.3.37}$$

*donde $K$ es una constante de tamaño moderado que sólo depende de los coeficientes del RK considerado, de las constantes $\{M_j,\ j = 1, \ldots, q+1\}$ y de un valor $\bar{r} = \mathcal{O}(1)$, pero es independiente de la red tomada con tal que las razones de paso (ver(4.3.33)) verifiquen*

$$r_n \in (0, \bar{r}], \quad n = 1, 2, \ldots, N.$$

*(b) Si el método RK es además stiffly accurate, entonces los inicializadores de Tipo II con orden stiff $l-1$ también satisfacen (4.3.37), con constante $K$ en las mismas condiciones que en el caso anterior.*

<u>*Demostración.*</u> Sea una partición $\mathcal{P}$ cualquiera de $[0, T]$ como se indica en (4.3.32), verificando

$$0 < r_n \le \bar{r} = \mathcal{O}(1), \quad n = 1, 2, \ldots, N.$$

Denotemos la solución exacta del PVI por

$$\hat{y}_n = y(\tau_n), \quad n = 0, 1, \ldots, N,$$

e

$$\hat{Y}_{i,n} = y(\tau_{n-1} + c_i h_n), \quad 1 \le i \le s, \quad n = 1, 2, \ldots, N \tag{4.3.38}$$

Por otro lado, sea $\bar{Y}_{i,n}$ la $i-$ésima etapa del RK considerado tras dos pasos consecutivos y que parte del punto $(\tau_{n-2}, \hat{y}_{n-2})$, i.e.,

$$\bar{Y}_{i,n} = \bar{y}_{n-1} + h_n \sum_{j=1}^{s} a_{ij} f(\tau_{n-1} + c_j h_n, \bar{Y}_{j,n}), \quad i = 1, \ldots, s, \quad n = 2, 3, \ldots, N,$$

donde

$$\left.\begin{array}{c} \bar{y}_{n-1} = \hat{y}_{n-2} + h_{n-1} \sum_{j=1}^{s} b_j f(\tau_{n-2} + c_j h_{n-1}, X_{j,n-1}), \\ X_{i,n-1} = \hat{y}_{n-2} + h_{n-1} \sum_{j=1}^{s} a_{ij} f(\tau_{n-2} + c_j h_{n-1}, X_{j,n-1}),\ 1 \le i \le s, \end{array}\right\} \ n = 2, 3, \ldots, N. \tag{4.3.39}$$

Denotemos también por $\bar{Y}^0_{i,n}$ el algoritmo de arranque de Tipo I o de Tipo II que parte del punto $(\tau_{n-2}, \hat{y}_{n-2})$, es decir, aquellos expresados en (4.3.35) o (4.3.36) con $y_{n-2}$ reemplazada por $\hat{y}_{n-2}$ y las etapas $Y_{j,n-1}$ reemplazadas por $X_{j,n-1}$.

Con estas notaciones, demostraremos el teorema basándonos en la siguiente igualdad:

$$Y_{i,n} - Y^0_{i,n} = (Y_{i,n} - \bar{Y}_{i,n}) + (\bar{Y}_{i,n} - \bar{Y}^0_{i,n}) + (\bar{Y}^0_{i,n} - Y^0_{i,n}). \tag{4.3.40}$$

Teniendo en cuenta la estabilidad algebraica, la estabilidad diagonal y las condiciones $B(q)$ y $C(q)$ para el Runge–Kutta considerado, se sigue la B–convergencia de orden $q$, ver p.e. [37, Teorema 15.3, pág. 230]. Por tanto tenemos que

$$|\hat{y}_n - y_n| \le C_1 (\bar{h}_n)^q, \quad n = 1, 2, \ldots, \tag{4.3.41}$$

donde $C_1$ es una constante que sólo depende de $M_{q+1}$ y de los coeficientes del método RK considerado. Nótese también que las constantes $C_j$ que aparezcan en adelante en la demostración del teorema sólo dependerán de los coeficientes del RK, de $M_j$, $(j = 1, \ldots, q+1)$, y eventualmente de $\bar{r}$, pero serán siempre de tamaño moderado e independientes de la red elegida.

Por otra parte, convendremos que $h_0 = 0$ con el objeto de que todas las fórmulas presentadas en esta demostración sean consistentes con la notación usada.

Usando la estabilidad diagonal del método, la estabilidad algebraica y (4.3.41), se sigue inmediatamente (ver, p.e. [37, Teorema 14.3,pág. 219])

$$\max_{1 \le i \le s} |Y_{i,n} - \bar{Y}_{i,n}| \le C_2 |\bar{y}_{n-1} - y_{n-1}| \le C_2 |\hat{y}_{n-2} - y_{n-2}| \le C_1 C_2 (\bar{h}_{n-2})^q, \quad n = 2, 3, \ldots \tag{4.3.42}$$

Además, usando los mismos argumentos se tiene también que

$$\max_{1 \le i \le s} |X_{i,n-1} - Y_{i,n-1}| \le C_2 |\hat{y}_{n-2} - y_{n-2}| \le C_1 C_2 (\bar{h}_{n-2})^q, \quad n = 2, 3, \ldots \tag{4.3.43}$$

Por otra parte, por tener los algoritmos de arranque orden stiff $l - 1$ se sigue que

$$\max_{1 \le i \le s} |\bar{Y}_{i,n} - \bar{Y}^0_{i,n}| \le C_3 (h_{n-1})^l, \quad n = 2, 3, \ldots \tag{4.3.44}$$

*(a)* Para los algoritmos de Tipo I se tiene que

$$\bar{Y}^0_{i,n} - Y^0_{i,n} = \gamma_i (\hat{y}_{n-2} - y_{n-2}) + \sum_{j=1}^{s} \alpha_{ij} (X_{j,n-1} - Y_{j,n-1}), \quad i = 1, ..., s, \quad n = 2, 3, \ldots \tag{4.3.45}$$

Se sigue ahora de la acotación uniforme de los coeficientes $\{\gamma_i, \alpha_{ij}\}$, y de (4.3.41), (4.3.43) y (4.3.45) que

$$\max_{1\le i\le s} |\bar{Y}^0_{i,n} - Y^0_{i,n}| \le C_4(\bar{h}_{n-2})^q, \quad n = 2, 3, \dots \tag{4.3.46}$$

Por tanto, tomando normas en (4.3.40) y usando (4.3.42), (4.3.44) y (4.3.46) se concluye el apartado *(a)*.

*(b)* De (4.3.36) se tiene

$$\bar{Y}^0_{i,n} - Y^0_{i,n} = (\hat{y}_{n-2} - y_{n-2}) + h_{n-1}\left(\delta_i(f(\tau_{n-2}, \hat{y}_{n-2}) - f(\tau_{n-2}, y_{n-2})) + \sum_{j=1}^{s} \beta_{ij}(\bar{f}_{j,n-2} - f_{j,n-2})\right), \tag{4.3.47}$$

donde hemos usado la notación

$$\bar{f}_{j,n-2} \equiv f(\tau_{n-2} + c_j h_{n-1}, X_{j,n-1}), \quad f_{j,n-2} \equiv f(\tau_{n-2} + c_j h_{n-1}, Y_{j,n-1}).$$

En virtud de que $A$ es inversible podemos poner (denotando el elemento genérico de $A^{-1}$ mediante $a^{(-1)}_{ij}$)

$$h_{n-1}(\bar{f}_{i,n-2} - f_{i,n-2}) = \sum_{j=1}^{s} a^{(-1)}_{ij}[(X_{j,n-1} - Y_{j,n-1}) + (y_{n-2} - \hat{y}_{n-2})], \quad 1 \le i \le s, \quad n \ge 2.$$

Se sigue ahora de (4.3.41) y (4.3.43) que

$$h_{n-1} \max_{1\le j\le s} |\bar{f}_{j,n-2} - f_{j,n-2}| \le C_5(\bar{h}_{n-2})^q, \quad n \ge 2. \tag{4.3.48}$$

Por otra parte, veamos que

$$h_n|f(\tau_n, \hat{y}_n) - f(\tau_n, y_n)| \le C_6(\bar{h}_n)^q, \quad n \ge 0. \tag{4.3.49}$$

Para este fin, teniendo en cuenta que el método es stiffly accurate, podemos escribir (despejando la derivada de la última etapa)

$$h_n f(\tau_n, y_n) = \sum_{j=1}^{s} w_j(Y_{j,n} - y_{n-1}), \tag{4.3.50}$$

y (ver (4.3.38))

$$h_n f(\tau_n, \hat{y}_n) = \sum_{j=1}^{s} w_j(\hat{Y}_{j,n} - \hat{y}_{n-1} - R_{j,n}), \tag{4.3.51}$$

donde

$$w^T = (w_1, \dots, w_s) = e_s^T A^{-1},$$

y

$$R_{i,n} = y(\tau_{n-1} + c_i h_n) - y(\tau_{n-1}) - h_n \sum_{j=1}^{s} a_{ij} y'(\tau_{n-1} + c_j h_n). \tag{4.3.52}$$

Además, se sigue de la condición $C(q)$ que

$$|R_{i,n}| \le K M_{q+1}(h_n)^{q+1}, \quad 1 \le i \le s, \quad n \ge 1, \tag{4.3.53}$$

donde la constante $K$ sólo depende de los coeficientes del método RK. Por otra parte usando la estabilidad algebraica, la estabilidad diagonal y la condición $C(q)$ no es difícil probar que

$$|\hat{Y}_{i,n} - Y_{i,n}| \leq C_7(\bar{h}_n)^q, \quad 1 \leq i \leq s, \quad n \geq 1. \tag{4.3.54}$$

Restando ahora (4.3.50) de (4.3.51), tomando normas y usando las acotaciones dadas en (4.3.41), (4.3.42), (4.3.53) y (4.3.54), se obtiene la expresión indicada en (4.3.49).

En consecuencia, con esta acotación (4.3.49) y teniendo en cuenta que $h_{n-1} = r_{n-2}h_{n-2} \leq \bar{r}h_{n-2}$ se tiene que

$$h_{n-1}|f(\tau_{n-2}, \hat{y}_{n-2}) - f(\tau_{n-2}, y_{n-2})| \leq C_8(\bar{h}_{n-2})^q, \quad n \geq 2. \tag{4.3.55}$$

Tomando normas en (4.3.47) y usando (4.3.41), (4.3.48) y (4.3.55) se concluye inmediatamente

$$\max_{1\leq i\leq s} |\bar{Y}^0_{i,n} - Y^0_{i,n}| \leq C_9(\bar{h}_{n-2})^q, \quad n = 2, 3, \ldots. \tag{4.3.56}$$

La demostración del teorema se deduce ahora de forma clara de (4.3.40), (4.3.42), (4.3.44) y (4.3.56). ∎

**Nota IV.3.5** *Del teorema anterior se infiere que si, por ejemplo, usamos tamaño de paso fijo $h$, entonces considerando métodos de colocación de $s$ etapas que sean además algebraicamente estables y diagonalmente estables, se tiene que los inicializadores de Tipo I con orden stiff $s$ suministran aproximaciones iniciales que verifican*

$$Y_{i,n} - Y^0_{i,n} = \mathcal{O}(h^s), \quad 1 \leq i \leq s, \quad n \geq 2,$$

*donde el término $\mathcal{O}(h^s)$ no está afectado por la stiffness del problema. Sin embargo, para los inicializadores de Tipo II con orden stiff $s$ no se puede garantizar lo mismo al menos que el método sea stiffly accurate. Así para el RK Gauss de $s$ etapas no es aconsejable usar inicializadores de Tipo II, pues considerando el modelo de Prothero y Robinson y haciendo $\lambda \to -\infty$, es fácil ver que el error del inicializador tiende a infinito.*

*Otra cuestión interesante es determinar si se podrían encontrar inicializadores de Tipo I o II que cumplan*

$$Y_{i,n} - Y^0_{i,n} = \mathcal{O}(h^{s+1}), \quad 1 \leq i \leq s, \quad n \geq 2, \tag{4.3.57}$$

*con $\mathcal{O}(h^{s+1})$ independiente de la stiffness, para métodos RK implícitos de colocación de $s$ etapas (algebraicamente estables y diagonalmente estables) que además verifiquen $B(s+1)$. A este respecto podemos decir que si el método tiene orden de B–convergencia $s+1$, entonces para los inicializadores de Tipo I dados por (4.2.9) y para los de Tipo II (si el método es además stiffly accurate) dados en la Nota IV.2.2, se tiene una expresión del tipo (4.3.57). La demostración de este hecho es análoga a la realizada en el teorema anterior. Por otra parte, los métodos que alcanzan orden de B–convergencia $s+1$ han sido caracterizados por Burrage et al. en [7] y poseen a lo más orden clásico 3, debiendo verificar que*

$$c^{s+1} - (s+1)Ac^s = \gamma e, \quad \gamma \in \mathbb{R}.$$

*De este modo la mayoría de los métodos de interés quedan excluidos, y por tanto sospechamos que la respuesta a la cuestión planteada es negativa fuera del conjunto particular de métodos Runge–Kutta indicados en el artículo anterior.*

### IV.3.3 Algunos inicializadores importantes

En esta sección vamos a resumir los resultados de orden obtenidos para algunos inicializadores importantes de Tipos I y II cuando se consideran los siguientes métodos Runge–Kutta: Gauss, Radau IA, Radau IIA, Lobatto IIIA y Lobatto IIIC. También damos una breve guía para su uso en la práctica, indicando las ecuaciones que deben verificar dichos inicializadores, así como una notación especial para ellos.

**Tipo I**

$\mathcal{L}_s^0$: denota el obtenido a través de la interpolación de Lagrange sobre las etapas internas $X_1, \ldots, X_s$ e $y_0$, como se indica en el Teorema IV.2.1 por (4.2.9). Como se probó en la sección anterior, este algoritmo tiene orden $s$ (stiff y no stiff) para el Gauss y el Radau IIA, que satisfacen $C(s)$, y orden $s-1$ para Radau IA y Lobatto IIIC, ya que sólo verifican $C(s-1)$. Para el Lobatto IIIA, aunque verifica $C(s)$, su matriz $A$ es singular y el inicializador sólo tiene orden $s-1$ (clásico y stiff). En lo que respecta a las funciones de amplificación del error, están uniformemente acotadas en todos los casos, siendo $\hat{R}_i(\infty) = \gamma_i \neq 0$.

$\mathcal{L}_s^1$: denota el obtenido a través de la interpolación de Lagrange de las etapas internas $X_1, \ldots, X_s$ ($\gamma_i = 0$), como se indica en el Teorema IV.2.1 por (4.2.8). Tiene orden $s-1$ en todos los casos y las funciones de amplificación (todas uniformemente acotadas) verifican $\hat{R}_i(\infty) = 0$, excepto para el Lobatto IIIA. Nótese que para este método en particular se tiene $y_0 = X_1$, por lo que este inicializador es esquivalente a $\mathcal{L}_s^0$.

$\mathcal{M}_s^I$: denota el inicializador de Tipo I de orden $s$ (clásico) cuando $A$ es no singular y el orden de etapa del RK es $s-1$ (ver Teorema IV.2.1), y viene dado por (4.2.7) y (4.2.17). Este sólo se considera en los casos del Radau IA y Lobatto IIIC (para los otros métodos, este inicializador es el mismo que el $\mathcal{L}_s^0$). Tiene orden stiff $s-1$ y sus funciones de amplificación están acotadas.

**Tipo II**

$\mathcal{M}_{s+1}^{II,1}$: para el Gauss y el Radau IIA, denota el inicializador de Tipo II de orden clásico $s+1$ dado en (4.2.25). Su orden stiff es $s$ y su función de amplificación no está acotada cuando $z$ tiende a infinito (ver Teorema IV.3.4).

$\mathcal{M}_{s+1}^{II,2}$: para el Gauss y el Radau IIA, denota el inicializador de orden clásico $s$ y orden $s+1$ en el caso particular $z = \infty$ (sobre el modelo de Prothero y Robinson). Sus coeficientes vienen dados en (4.3.16) (ver Teorema IV.3.3-(c)). Como en el caso anterior, su orden stiff es $s$ y las funciones de amplificación no están acotadas cuando $z \to \infty$.

$\mathcal{M}_{s+1}^{II,3}$: para el Lobatto IIIA, denota el inicializador de orden clásico y stiff $s$ dado en el Teorema IV.3.3-(b). En este caso la función de amplificación no está acotada cuando $z$ tiende a infinito (ver Teorema IV.3.4).

**Nota IV.3.6** *Para los métodos Radau IA y Lobatto IIIC el orden de etapa es $s-1$, por lo que el orden máximo (clásico) para los inicializadores es $s$. Además, tenemos una familia $s$–paramétrica de este orden. Esta familia sólo alcanza orden $s-1$ sobre el modelo de Prothero y Robinson y su orden stiff es $s-1$ (en el caso del Lobatto IIIC debe realizarse un análisis especial pues dichos métodos no son diagonalmente estables para $s > 2$, y su orden stiff sería sólo $s-2$ para problemas que verifican (1.0.3) con $\nu < 0$).*

En las tablas 4.3.1, 4.3.2 y 4.3.3 resumimos las propiedades de los inicializadores considerados.

| | orden no stiff | orden stiff | orden P-R en $z=0$ | orden P-R en $z=\infty$ | $\tilde{R}_i(\infty)$ |
|---|---|---|---|---|---|
| $\mathcal{L}_s^1$ | $s-1$ | $s-1$ | $s-1$ | $s-1$ | 0 |
| $\mathcal{L}_s^0$ | $s$ | $s$ | $s$ | $s$ | acotado |
| $\mathcal{M}_{s+1}^{II,1}$ | $s+1$ | $s$ | $s+1$ | $s$ | $\infty$ |
| $\mathcal{M}_{s+1}^{II,2}$ | $s$ | $s$ | $s$ | $s+1$ | $\infty$ |

Tabla 4.3.1: Gauss y Radau IIA ($A$ no singular, $C(s)$).

| | orden no stiff | orden stiff | orden P-R en $z=0$ | orden P-R en $z=\infty$ | $\tilde{R}_i(\infty)$ |
|---|---|---|---|---|---|
| $\mathcal{L}_s^1$ | $s-1$ | $s-1$ | $s-1$ | $s-1$ | 0 |
| $\mathcal{L}_s^0$ | $s-1$ | $s-1$ | $s-1$ | $s-1$ | acotado |
| $\mathcal{M}_s^I$ | $s$ | $s-1$ | $s$ | $s-1$ | acotado |

Tabla 4.3.2: Radau IA y Lobatto IIIC ($A$ no singular, $C(s-1)$).

## IV.4 Experimentos numéricos

En esta sección vamos a presentar varios experimentos numéricos con dos objetivos principales. En primer lugar, mediante experimentos de tipo local, es decir, sólo se avanzan dos pasos consecutivos de tamaños $h$ y $rh$, pretendemos confirmar el orden teórico obtenido para los inicializadores, así como su factor de amplificación de error. En segundo lugar, y con propósitos más prácticos nos proponemos estudiar la eficacia de los distintos inicializadores aquí considerados. Para ello, realizaremos integraciones de problemas stiff en intervalos "largos", mediante métodos Runge–Kutta de colocación de tal forma que los códigos sólo difieran en el inicializador elegido.

En ambos casos, como método Runge–Kutta hemos elegido el Radau IIA de tres etapas, y como inicializadores los denotados por $\mathcal{L}_3^1$, $\mathcal{L}_3^0$, $\mathcal{M}_4^{II,1}$ y $\mathcal{M}_4^{II,2}$, cuyas propiedades más relevantes están dadas en la tabla 4.3.1 al considerar $s=3$. Para contrastar los resultados hemos incluido también el inicializador $Y_i^0 = y_1$, $i = 1,2,3$. Obsérvese que este último inicializador es de Tipo I con $\gamma_i = 0$ y $\alpha_i^T = (0,0,1)$, por lo que su orden es 0 (clásico y stiff) y su función de amplificación es $\hat{R}_i(z) = R(z)$, o sea, igual a la función de estabilidad lineal del RK Radau IIA. Esto lo convierte en un inicializador muy estable, es decir, su función de amplificación hereda las buenas propiedades de estabilidad lineal del método Radau IIA.

Con respecto al primer objetivo propuesto, y a efectos de avanzar sólo dos pasos de tamaños $h$ y $rh$ respectivamente, hemos realizado un programa que procede de la siguiente forma sobre

| | orden no stiff | orden stiff | orden P-R en $z=0$ | orden P-R en $z=\infty$ | $\tilde{R}_i(\infty)$ |
|---|---|---|---|---|---|
| $\mathcal{L}_s^1 \equiv \mathcal{L}_s^0$ | $s-1$ | $s-1$ | $s-1$ | $s-1$ | acotado |
| $\mathcal{M}_{s+1}^{II,3}$ | $s$ | $s$ | $s$ | $s$ | $\infty$ |

Tabla 4.3.3: Lobatto IIIA ($A$ singular, $C(s)$).

el PVI considerado:

$$y' = f(t,y), \quad y(0) = z_0, \quad y, f \in I\!R^m.$$

(1) Calcula con 16 cifras de exactitud las etapas internas $X_i$ $(1 \le i \le 3)$ correspondientes al primer paso de tamaño $h$, partiendo de valor inicial exacto, esto es, $y_0 = z_0$.

(2) Calcula también con 16 cifras de exactitud las etapas $Y_i$ $(1 \le i \le 3)$ correspondientes al segundo paso de tamaño $\bar{h} = rh$.

(3) Obtiene las aproximaciones iniciales $Y_i^0$ $(1 \le i \le 3)$ correspondientes a cada uno de los inicializadores considerados.

(4) Como medida del error en la aproximación dada por los inicializadores toma

$$\varepsilon(h) := \max_{1 \le i \le 3} |Y_i - Y_i^0|_\infty. \tag{4.4.1}$$

(5) Para medir cómo se propaga el error del dato inicial $y_0 - z_0$, es decir, cómo amplifica el error inicial cada inicializador, se repite el proceso anterior desde el paso (1) hasta el paso (4), pero considerando ahora como valor inicial

$$y_0 = z_0 + \Delta, \qquad \Delta = (\Delta_1, ..., \Delta_m)^T, \qquad \Delta_i = 10^{-3} \max\{1, |z_{0,i}|\},\ i = 1, ..., m. \tag{4.4.2}$$

Por brevedad sólo presentamos aquí los resultados obtenidos con los dos problemas stiff siguientes:

**Problema 1.-** Problema de Prothero y Robinson (4.3.1) con

$$\lambda = -10^6, \quad \phi(t) = \exp(2t), \quad z_0 = 1.$$

Este problema se ha integrado varias veces con el código anterior tomando como valores de tamaño de paso y razón de paso

$$h = h_0/2^k, \quad k = 0, 1, ..., 7, \quad h_0 = 0.4, \quad r = 1.$$

**Problema 2.-** Problema escalar no lineal del tipo propuesto en [58, pág. 202]:

$$y' = \lambda(y^3 - \phi^3(t)) + \phi'(t), \quad y(0) = z_0, \quad Re\ \lambda \le 0. \tag{4.4.3}$$

En este caso hemos tomado

$$\lambda = -10^6, \quad \phi(t) = 1 + \exp(t), \quad z_0 = 2,$$

realizando varias integraciones con los tamaños de paso

$$h = h_0/2^k, \quad k = 0, 1, ..., 6, \quad h_0 = 0.2, \quad r = 1.$$

Hemos elegido estos dos problemas por su extrema sencillez y porque además creemos que en cierto modo, pueden ser "representativos" de los sistemas stiff de dimensión más elevada. Es decir, en el caso de sistemas stiff lineales es patente que el modelo de Prothero y Robinson es el modelo resultante cuando el sistema diferencial se desacopla en forma conveniente, mientras que para los sistemas stiff no lineales en general, no podemos decir en absoluto que el problema 2 (o similares) represente la complejidad del problema no lineal. Sin embargo, creemos que por la

relativa sencillez del problema 2, éste podría considerarse como una primera aproximación para el caso no lineal en general. Veremos también más adelante en esta sección, que algunas conclusiones sobre el comportamiento de los inicializadores sobre el problema 2 se pueden extrapolar a ciertos sistemas stiff no lineales.

En las gráficas 4.4.1, 4.4.2, 4.4.3 y 4.4.4, dadas en la páginas siguientes, hemos representado el logaritmo decimal del error dado en (4.4.1) con respecto al logaritmo decimal del tamaño de paso $h$, para los distintos tamaños de paso propuestos en los problemas 1 y 2, y para los inicializadores previamente considerados. Estos pares de puntos, $(\log(h), \log(\varepsilon(h)))$, se han unido en las mencionadas gráficas mediante líneas poligonales.

Las gráficas 4.4.1 y 4.4.3 se corresponden respectivamente con los problemas 1 y 2 en el caso en que no se considera perturbación en el valor inicial, es decir, cuando se toma $y_0 = z_0 = \phi(0)$. Obsérvese que en este caso el valor $\varepsilon(h)$ dado por (4.4.1) nos daría prácticamente el error de cada inicializador en el caso $\lambda \to -\infty$ o $z = h\lambda \to -\infty$. Por tanto, si un inicializador es de orden $p$ en $\lambda = -\infty$, entonces su error se comporta de la forma siguiente:

$$\varepsilon(h) = Kh^{p+1} + \mathcal{O}(h^{p+2}), \quad K \text{ cte}, \quad h \to 0^+.$$

Esto implica que

$$\log(\varepsilon(h)) \doteq \log(K) + (p+1)\log(h), \quad h \to 0^+.$$

Se tiene así que al representar los pares de puntos $(\log(h), \log(\varepsilon(h)))$, obtengamos "rectas" cuya pendiente será $p+1$, es decir, el orden del inicializador más uno. Este hecho se corrobora perfectamente en las gráficas 4.4.1 y 4.4.3, en las que se aprecia que el inicializador denotado por $y_1$ (de órdenes clásico y stiff cero) nos proporciona "rectas" de pendiente uno. Asimismo, el inicializador $\mathcal{L}_3^1$ (de orden 2) suministra rectas de pendiente 3, y los inicializadores denotados por $\mathcal{L}_3^0$ y $\mathcal{M}_4^{II,1}$ (ambos de orden 3, para $\lambda \to -\infty$) dan lugar a rectas de pendiente 4. El inicializador denotado por $\mathcal{M}_4^{II,2}$ (de orden 4 para $\lambda = -\infty$) proporciona rectas de pendiente 5, como era de esperar. Con estos experimentos podemos decir que se corrobora experimentalmente la teoría del orden stiff basada en el modelo de Prothero y Robinson, en el caso de ausencia de perturbación en el valor inicial $z_0 = \phi(0)$.

En las gráficas 4.4.2 y 4.4.4 representamos respectivamente los valores obtenidos sobre los problemas 1 y 2, en el caso de introducir una perturbación en el valor inicial, tal cual se especifica en (4.4.2). Pretendemos ahora medir el efecto que producen pequeñas perturbaciones en el dato inicial sobre la aproximación suministrada por los inicializadores. Dicho de otro modo, nos interesa medir la magnitud de la función de amplificación de error para los distintos inicializadores.

En primer lugar podemos observar que los inicializadores de Tipo II ($\mathcal{M}_4^{II,1}$ y $\mathcal{M}_4^{II,2}$) nos dan errores bastante grandes. Esto se debe sin duda al hecho de que los errores de amplificación dominan a los errores locales (aquéllos que aparecen en ausencia de perturbación). Naturalmente, es de esperar que los errores de amplificación producidos por los inicializadores de Tipo II sean de magnitud elevada, en virtud de que sus funciones de amplificación se comportan del modo siguiente para $(\lambda \to -\infty)$ (ver fórmula (4.3.20)),

$$\tilde{R}_i(\lambda h) \doteq \delta_i \lambda h, \quad 1 \le i \le 3, \quad \text{con} \ \ \delta_i \neq 0,$$

lo que implica que el error del inicializador considerado se comporte como

$$Y_i(h) - Y_i^0(h) \doteq \tilde{R}_i(\lambda h)(y_0 - \phi(0)), \quad 1 \le i \le 3.$$

No es de extrañar por tanto que al representar los puntos $(\log(h),\ \log(\varepsilon(h)))$ en las gráficas 4.4.2 y 4.4.4, , aparezcan rectas de pendiente 1.

En cuanto a los inicializadores de Tipo I, observamos que el denotado por $y_1$ se comporta prácticamente igual en ausencia o presencia de perturbación del dato inicial. Esto lo convierte en un inicializador muy estable a pesar de su bajo orden de aproximación. El inicializador denotado por $\mathcal{L}_3^0$ sí que queda afectado por las pertubaciones introducidas y en este caso al dominar los errores de amplificacion a los errores locales, se observa en las gráficas 4.4.2 y 4.4.4, que el error $\varepsilon(h)$ dado en (4.4.1) se mantiene prácticamente constante al variar $h$. Esto se justifica teniendo en cuenta que en este caso las funciones de amplificación vienen dadas por

$$\hat{R}_i(\lambda h) \doteq K_i, \quad (K_i \text{ ctes}), \quad 1 \leq i \leq 3, \quad \lambda \to -\infty.$$

Por otra parte, el error $\varepsilon(h)$ dado por el inicializador $\mathcal{L}_3^1$ presenta un comportamiento cualitativamente distinto antes y después de cierto tamaño de paso "crítico" $h^* \doteq 0.1 \times 2^{-3}$ (en ambos problemas). Es decir, para $h > h^*$ los errores del inicializador están dominados por los errores locales (aquellos que aparecen en ausencia de perturbación del dato inicial $z_0$), mientras que para $h < h^*$ (pero $h$ no suficientemente próximo a cero) los errores están dominados por los errores de amplificación del inicializador. Se observa (en las gráficas 4.4.2 y 4.4.4) que para este último rango de valores de $h$, el error crece al disminuir $h$. Esto se puede justificar teniendo en cuenta que las funciones de amplificación vienen dadas por

$$\hat{R}_i(\lambda h) \doteq \bar{K}_i(\lambda h)^{-1}, \quad 1 \leq i \leq 3, \quad (\lambda \to -\infty,\ h \to 0^+).$$

Consecuentemente,

$$\log(\varepsilon(h)) \doteq \log(\bar{K}) - \log(h), \quad \bar{K} = \max_{1 \leq i \leq 3} |\bar{K}_i \lambda^{-1}(y_0 - \phi(0))|, \quad h \to 0^+.$$

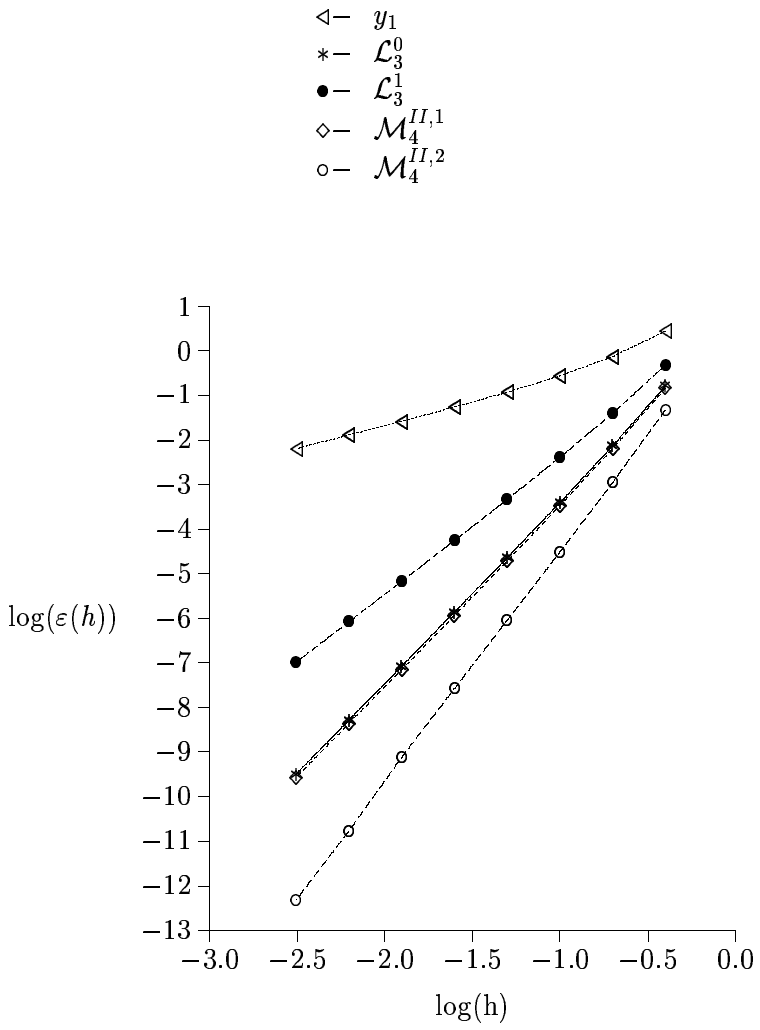
Por tanto, la representación de los puntos $(\log(h), \log(\varepsilon(h)))$, nos debe llevar a "rectas" de pendiente -1.

Respecto al segundo objetivo planteado, y a efectos de comparar los inicializadores arriba considerados con fines prácticos, hemos construido un código a paso variable basado en el Radau IIA de 3 etapas que estima el error local por extrapolación. Este código es similar al denotado por SN–RADAU en el capítulo II (sección II.4) de esta memoria, y en su programación se han seguido los pasos detallados en las primeras páginas de la mencionada sección II.4. Las principales novedades del nuevo código es que permite resolver las ecuaciones de etapa del Radau IIA (de 3 etapas) mediante la iteración de Newton Simplificada o bien mediante esquemas Single–Newton. Además, el código se ha preparado para que nos permita elegir el inicializador deseado de entre los cinco considerados anteriormente y algunos más que veremos en el próximo capítulo.

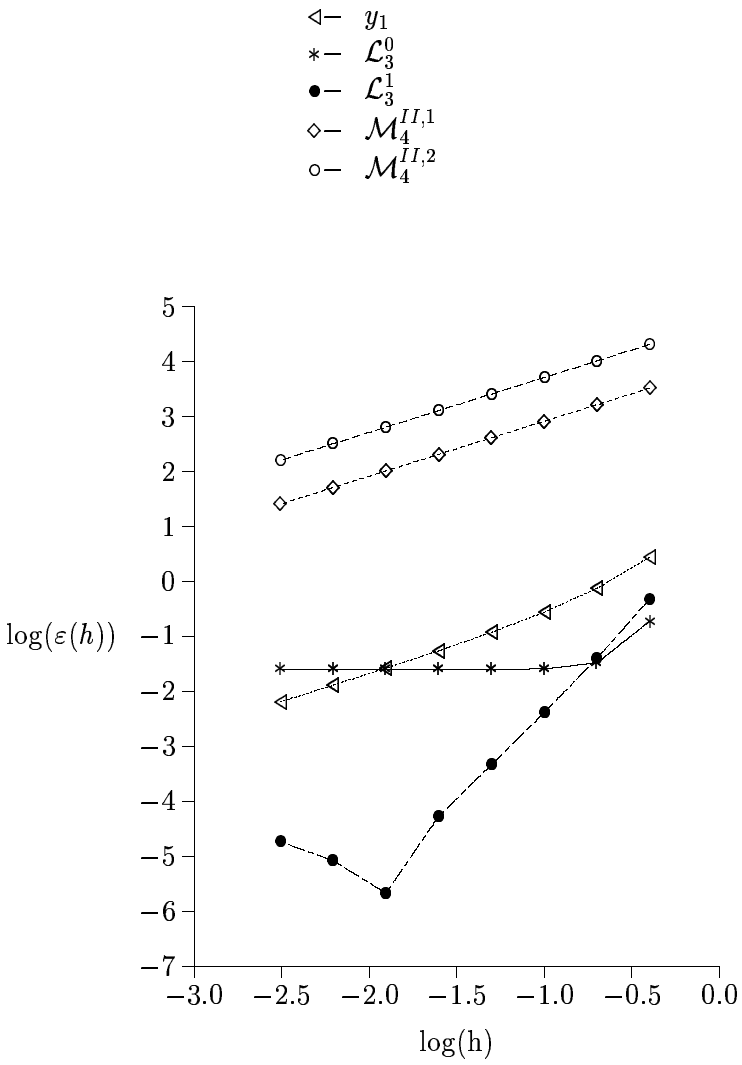
Hemos realizado numerosas integraciones para todos los problemas stiff propuestos en el paquete DETEST [23] así como para los también propuestos como tests en [37, Cap.IV.10] y [49]. Presentaremos aquí los resultados obtenidos para los tres problemas detallados debajo, dos de cuales han sido considerados previamente en los capítulos anteriores de esta memoria. Además, estos problemas pueden ser considerados en cierto modo representativos de la clase stiff.

**Problema 3.-** El oscilador de Van der Pol (ver p.e. [37, pág. 144]):

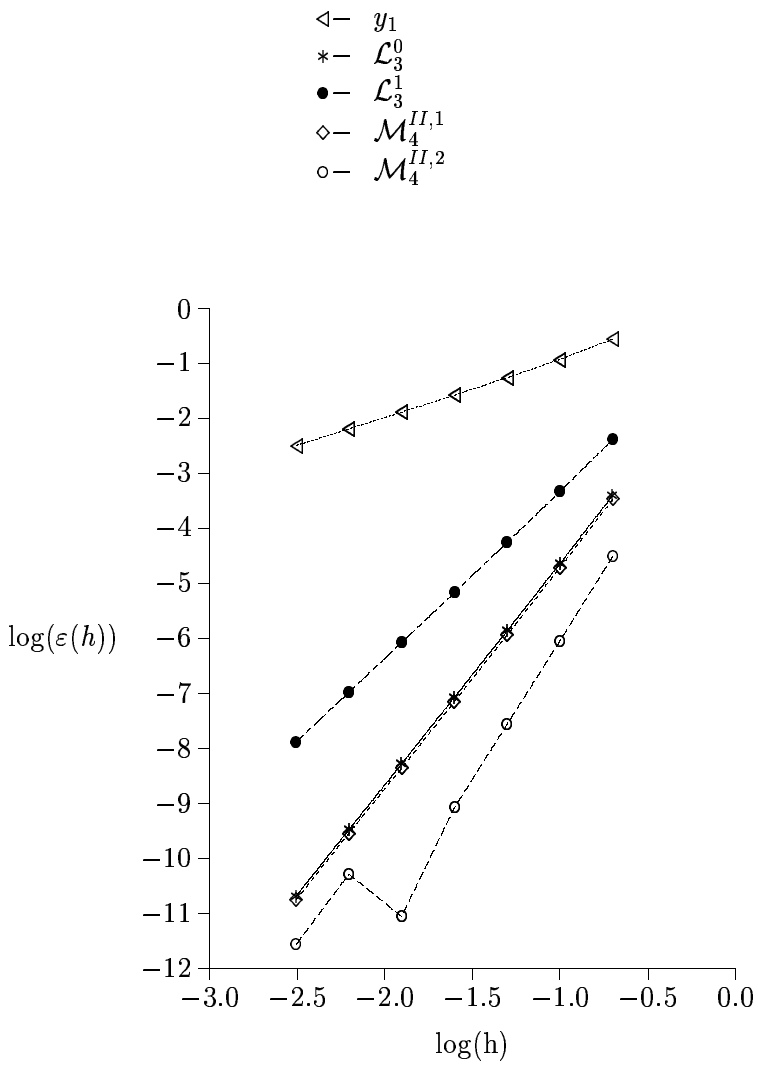
$$\begin{array}{ll} y_1' = y_2 & y_1(0) = 2 \\ y_2' = ((1 - y_1^2)y_2 - y_1)/\varepsilon & y_2(0) = 0, \\ t \in [0, 2], \qquad \varepsilon = 10^{-6}, \qquad h_0 = 10^{-8}. & \end{array}$$



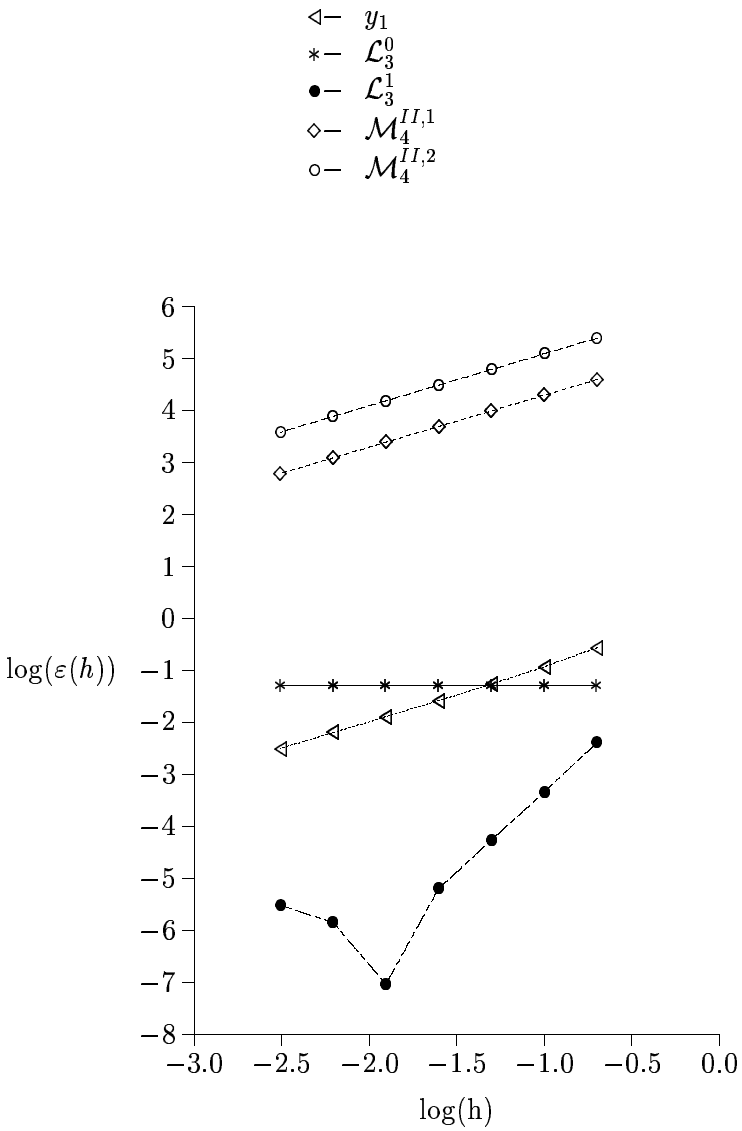

Gráfica 4.4.1: Problema de Prothero y Robinson ($y_0 = z_0$).




Gráfica 4.4.2: Problema de Prothero y Robinson ($y_0 = z_0 + \Delta$).




Gráfica 4.4.3: Problema de Spijker ($y_0 = z_0$).




Gráfica 4.4.4: Problema de Spijker ($y_0 = z_0 + \Delta$).

**Problema 4.-** E5 del paquete DETEST [37, pág.145]:

$$\begin{aligned}
y_1' &= -7.89\cdot 10^{-10}y_1 - 1.1\cdot 10^{7}y_1y_3 && y_1(0) = 1.76\cdot 10^{-3}\\
y_2' &= 7.89\cdot 10^{-10}y_1 - 1.13\cdot 10^{9}y_2y_3 && y_2(0) = 0\\
y_3' &= 7.89\cdot 10^{-10}y_1 - 1.1\cdot 10^{7}y_1y_3 - 1.13\cdot 10^{9}y_2y_3 + 1.13\cdot 10^{3}y_4 && y_3(0) = 0\\
y_4' &= 1.1\cdot 10^{7}y_1y_3 - 1.13\cdot 10^{3}y_4 && y_4(0) = 0
\end{aligned}$$

$$t \in [0, 1000], \qquad h_0 = 10^{-2}.$$

**Problema 5.-** Problema Ring Modulator considerado en [49]. Este problema es de dimensión 15, y es un problema stiff altamente oscilatorio, que requiere ser integrado con tamaños de paso muy pequeños ($h_n \doteq 10^{-8}$). En la integración de este problema fallan muchos códigos, tales como el RADAU5 [37] cuando se usan tolerancias altas (ver [49]). Este problema ha sido integrado por nuestro código en el intervalo $[0, 10^{-3}]$, con tamaño de paso inicial $h_0 = 10^{-8}$ (tanto el intervalo de integración considerado como el tamaño de paso inicial propuesto son los habituales, ver [49]).

Las tablas 4.4.1, 4.4.2 y 4.4.3 (dadas más adelante) se corresponden respectivamente con los problemas 3, 4 y 5, y hemos representado en ellas los siguientes valores:

- RTOL: tolerancia relativa. La tolerancia para cada paso de integración se ha calculado por la fórmula siguiente:

  $$\mathrm{TOL} = 0.01\cdot \mathrm{RTOL}\cdot \max\{\mathrm{ATOL}, |y_n|_\infty\},$$

  con valores de ATOL= 1 para todos los problemas salvo el 5, en el que se toma ATOL= 0.001 (esto se debe a que las soluciones de este problema tienen norma infinito del orden de 0.001).

- INI: el inicializador considerado.

- EG: el error global cometido en el punto final del intervalo de integración.

- NPA: el número de pasos aceptados.

- NR-ES: el número de pasos rechazados por el estimador.

- NR-SN: el número de pasos rechazados por no converger el esquema iterativo Single–Newton.

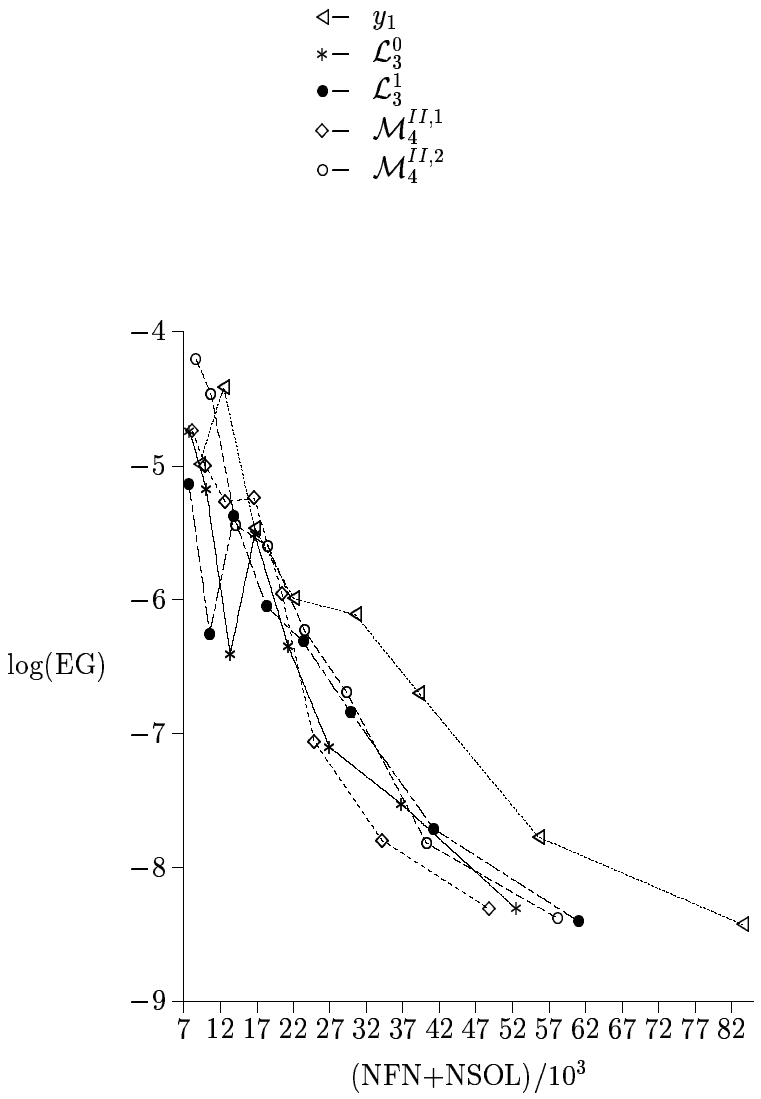
- NLU: el número de factorizaciones LU.

- NSOL: el número de sistemas triangulares resueltos.

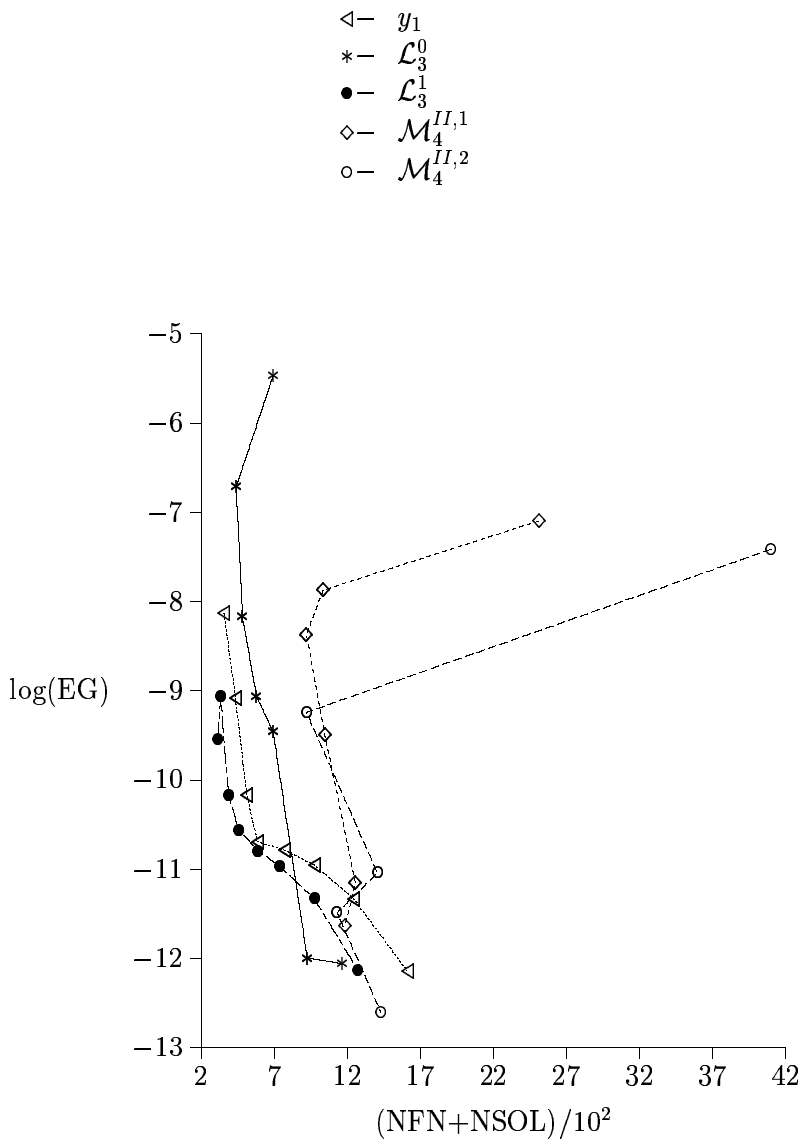
- NFN: el número de evaluaciones de la función derivada.

- NITER: el número medio de iteraciones que se han necesitado para alcanzar convergencia en cada paso.

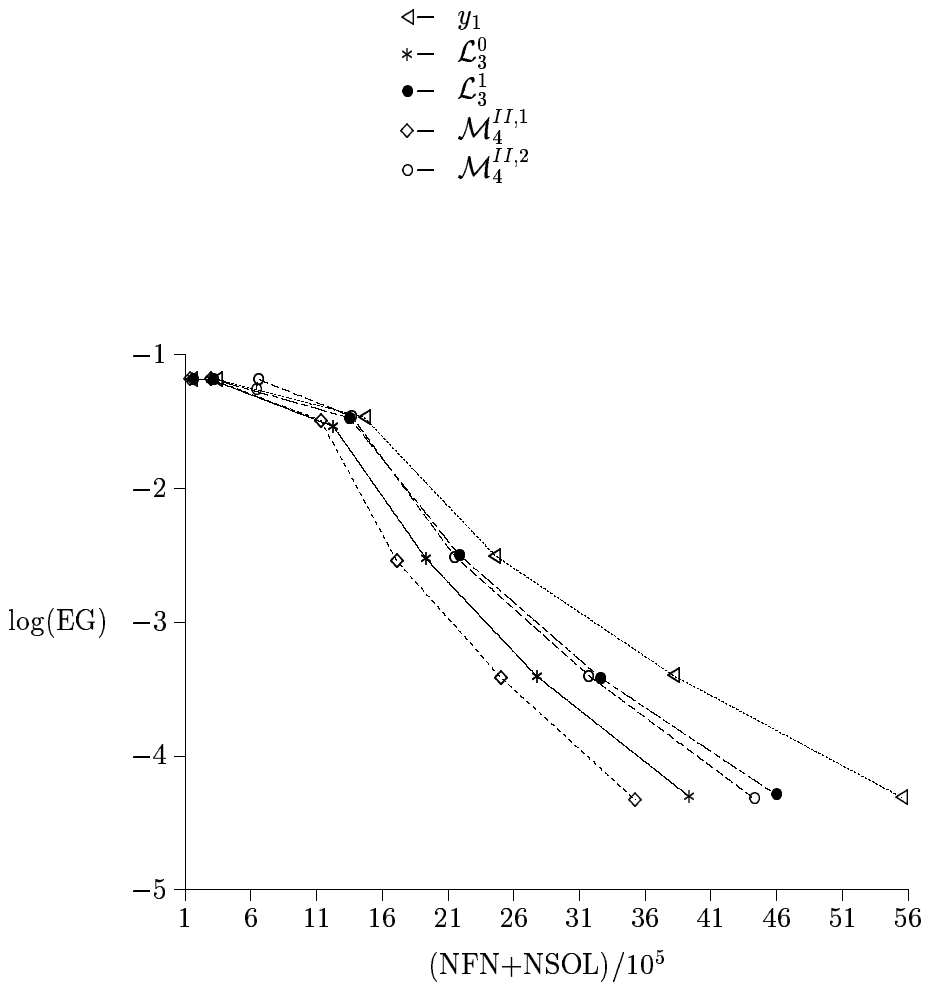
Con respecto al problema 3 (Van der Pol), podemos decir que para cada tolerancia (RTOL) todos los inicializadores hacen que el código dé errores globales (EG) similares, así como números parecidos de pasos (NPA, NR-ES, NR-SN), y números similares de factorizaciones LU (NLU), evaluaciones de función derivada (NFN) y de iteraciones por paso (NITER). Sin embargo, debemos resaltar que los inicializadores con un orden clásico mayor $\mathcal{L}_3^0$, $\mathcal{M}_4^{II,1}$ son ligeramente más




Gráfica 4.4.5: Problema de Van der Pol.




Gráfica 4.4.6: Problema E5 stiff.




Gráfica 4.4.7: Problema Ring Modulator.

| RTOL | INI | EG | NPA | NR-ES | NR-SN | NLU | NSOL | NFN | NITER |
|---|---|---|---|---|---|---|---|---|---|
| $10^{-2}$ | $y_1$ | .1024E-04 | 230 | 0 | 39 | 242 | 4545 | 4603 | 3.76 |
| | $\mathcal{L}_3^0$ | .1793E-04 | 224 | 0 | 39 | 237 | 3894 | 3940 | 3.13 |
| | $\mathcal{L}_3^1$ | .7192E-05 | 226 | 0 | 39 | 240 | 3858 | 3895 | 3.04 |
| | $\mathcal{M}_4^{II,1}$ | .1811E-04 | 224 | 0 | 39 | 240 | 4071 | 4114 | 3.37 |
| | $\mathcal{M}_4^{II,2}$ | .6206E-04 | 226 | 0 | 40 | 240 | 4332 | 4363 | 3.97 |
| $10^{-4}$ | $y_1$ | .3415E-05 | 310 | 4 | 40 | 323 | 8301 | 8404 | 5.30 |
| | $\mathcal{L}_3^0$ | .3876E-06 | 294 | 10 | 34 | 311 | 6645 | 6733 | 3.97 |
| | $\mathcal{L}_3^1$ | .4156E-05 | 296 | 4 | 36 | 308 | 6888 | 6985 | 4.23 |
| | $\mathcal{M}_4^{II,1}$ | .5370E-05 | 286 | 8 | 34 | 299 | 6306 | 6391 | 3.91 |
| | $\mathcal{M}_4^{II,2}$ | .3564E-05 | 300 | 12 | 33 | 314 | 7065 | 7114 | 4.67 |
| $10^{-6}$ | $y_1$ | .1015E-05 | 392 | 30 | 32 | 412 | 10980 | 11068 | 5.77 |
| | $\mathcal{L}_3^0$ | .4416E-06 | 468 | 46 | 20 | 495 | 10644 | 10702 | 4.35 |
| | $\mathcal{L}_3^1$ | .4817E-06 | 470 | 42 | 22 | 495 | 11646 | 11710 | 4.93 |
| | $\mathcal{M}_4^{II,1}$ | .1112E-05 | 460 | 42 | 21 | 486 | 10191 | 10252 | 4.16 |
| | $\mathcal{M}_4^{II,2}$ | .5814E-06 | 466 | 44 | 22 | 494 | 11751 | 11809 | 5.09 |
| $10^{-8}$ | $y_1$ | .1673E-07 | 924 | 22 | 35 | 939 | 27741 | 27844 | 7.18 |
| | $\mathcal{L}_3^0$ | .2963E-07 | 838 | 28 | 15 | 852 | 18369 | 18412 | 4.68 |
| | $\mathcal{L}_3^1$ | .1924E-07 | 846 | 24 | 17 | 860 | 20607 | 20656 | 5.49 |
| | $\mathcal{M}_4^{II,1}$ | .1592E-07 | 840 | 30 | 12 | 855 | 17049 | 17083 | 4.19 |
| | $\mathcal{M}_4^{II,2}$ | .1509E-07 | 842 | 28 | 14 | 858 | 20106 | 20146 | 5.35 |

Tabla 4.4.1: Problema de Van der Pol.

| RTOL | INI | EG | NPA | NR-ES | NR-SN | NLU | NSOL | NFN | NITER |
|---|---|---|---|---|---|---|---|---|---|
| $10^{-1}$ | $y_1$ | .5529E-07 | 32 | 0 | 0 | 32 | 171 | 169 | 1.28 |
| | $\mathcal{L}_3^0$ | *** | ** | ** | ** | ** | ** | ** | **** |
| | $\mathcal{L}_3^1$ | .2811E-07 | 32 | 0 | 0 | 32 | 144 | 142 | 1.00 |
| | $\mathcal{M}_4^{II,1}$ | *** | ** | ** | ** | ** | ** | ** | **** |
| | $\mathcal{M}_4^{II,2}$ | *** | ** | ** | ** | ** | ** | ** | **** |
| $10^{-2}$ | $y_1$ | .7387E-08 | 32 | 0 | 0 | 32 | 180 | 178 | 1.38 |
| | $\mathcal{L}_3^0$ | *** | ** | ** | ** | ** | ** | ** | **** |
| | $\mathcal{L}_3^1$ | .2871E-09 | 32 | 0 | 0 | 32 | 159 | 157 | 1.16 |
| | $\mathcal{M}_4^{II,1}$ | *** | ** | ** | ** | ** | ** | ** | **** |
| | $\mathcal{M}_4^{II,2}$ | *** | ** | ** | ** | ** | ** | ** | **** |
| $10^{-3}$ | $y_1$ | .8391E-09 | 32 | 0 | 0 | 32 | 219 | 217 | 1.78 |
| | $\mathcal{L}_3^0$ | .3388E-05 | 56 | 0 | 4 | 58 | 348 | 346 | 1.38 |
| | $\mathcal{L}_3^1$ | .8560E-09 | 32 | 0 | 0 | 32 | 168 | 166 | 1.25 |
| | $\mathcal{M}_4^{II,1}$ | *** | ** | ** | ** | ** | ** | ** | **** |
| | $\mathcal{M}_4^{II,2}$ | *** | ** | ** | ** | ** | ** | ** | **** |
| $10^{-4}$ | $y_1$ | .6743E-10 | 32 | 0 | 0 | 32 | 258 | 256 | 2.09 |
| | $\mathcal{L}_3^0$ | .1944E-06 | 32 | 0 | 0 | 32 | 222 | 220 | 1.69 |
| | $\mathcal{L}_3^1$ | .6748E-10 | 32 | 0 | 0 | 32 | 195 | 193 | 1.44 |
| | $\mathcal{M}_4^{II,1}$ | .7970E-07 | 140 | 0 | 21 | 155 | 1257 | 1255 | 2.09 |
| | $\mathcal{M}_4^{II,2}$ | .5278E-06 | 1144 | 0 | 219 | 1311 | 12387 | 12385 | 2.56 |
| $10^{-6}$ | $y_1$ | .1634E-10 | 34 | 0 | 0 | 34 | 387 | 385 | 3.00 |
| | $\mathcal{L}_3^0$ | .8633E-09 | 34 | 0 | 0 | 34 | 291 | 289 | 2.06 |
| | $\mathcal{L}_3^1$ | .1600E-10 | 34 | 0 | 0 | 34 | 297 | 295 | 2.12 |
| | $\mathcal{M}_4^{II,1}$ | .4270E-08 | 48 | 0 | 3 | 49 | 462 | 460 | 2.29 |
| | $\mathcal{M}_4^{II,2}$ | .5710E-09 | 40 | 0 | 1 | 40 | 462 | 460 | 2.98 |
| $10^{-8}$ | $y_1$ | .4638E-11 | 40 | 0 | 0 | 40 | 624 | 622 | 4.03 |
| | $\mathcal{L}_3^0$ | .1004E-11 | 40 | 0 | 0 | 40 | 465 | 463 | 2.70 |
| | $\mathcal{L}_3^1$ | .4654E-11 | 40 | 0 | 0 | 40 | 489 | 487 | 2.90 |
| | $\mathcal{M}_4^{II,1}$ | .6977E-11 | 62 | 0 | 4 | 65 | 630 | 628 | 2.35 |
| | $\mathcal{M}_4^{II,2}$ | .3228E-11 | 40 | 0 | 0 | 40 | 567 | 565 | 3.55 |

Tabla 4.4.2: Problema E5 stiff.

| RTOL | INI | EG | NPA | NR-ES | NR-SN | NLU | NSOL | NFN | NITER |
|---|---|---|---|---|---|---|---|---|---|
| $10^{-2}$ | $y_1$ | .6515E-01 | 2590 | 0 | 427 | 2690 | 73632 | 74569 | 5.93 |
| | $\mathcal{L}_3^0$ | .6551E-01 | 3700 | 0 | 600 | 3991 | 78321 | 79426 | 4.11 |
| | $\mathcal{L}_3^1$ | .6526E-01 | 3872 | 0 | 655 | 4187 | 82635 | 83728 | 4.34 |
| | $\mathcal{M}_4^{II,1}$ | .6533E-01 | 3050 | 0 | 498 | 3310 | 69933 | 70852 | 4.45 |
| | $\mathcal{M}_4^{II,2}$ | .5445E-01 | 17166 | 0 | 2967 | 18922 | 318645 | 323530 | 4.07 |
| $10^{-3}$ | $y_1$ | .6513E-01 | 5392 | 32 | 345 | 5504 | 171138 | 172087 | 6.92 |
| | $\mathcal{L}_3^0$ | .6514E-01 | 5538 | 22 | 302 | 5663 | 148608 | 149257 | 5.40 |
| | $\mathcal{L}_3^1$ | .6513E-01 | 5510 | 24 | 305 | 5624 | 158235 | 158935 | 6.02 |
| | $\mathcal{M}_4^{II,1}$ | .6514E-01 | 5942 | 30 | 359 | 6096 | 151056 | 151723 | 4.95 |
| | $\mathcal{M}_4^{II,2}$ | .6514E-01 | 12624 | 0 | 1367 | 13390 | 327843 | 330130 | 5.65 |
| $10^{-4}$ | $y_1$ | .3381E-01 | 22314 | 1236 | 193 | 22991 | 731598 | 732160 | 7.00 |
| | $\mathcal{L}_3^0$ | .2890E-01 | 22966 | 1200 | 413 | 23774 | 615045 | 615427 | 5.20 |
| | $\mathcal{L}_3^1$ | .3360E-01 | 22476 | 1176 | 286 | 23159 | 675822 | 676219 | 6.20 |
| | $\mathcal{M}_4^{II,1}$ | .3226E-01 | 23232 | 1144 | 412 | 23968 | 565887 | 566215 | 4.45 |
| | $\mathcal{M}_4^{II,2}$ | .3459E-01 | 22724 | 854 | 580 | 23482 | 682812 | 683599 | 6.32 |
| $10^{-5}$ | $y_1$ | .3143E-02 | 39782 | 2626 | 150 | 41132 | 1229328 | 1229764 | 6.82 |
| | $\mathcal{L}_3^0$ | .2950E-02 | 40942 | 2258 | 654 | 42420 | 968532 | 968767 | 4.66 |
| | $\mathcal{L}_3^1$ | .3176E-02 | 40294 | 2380 | 384 | 41672 | 1093275 | 1093558 | 5.71 |
| | $\mathcal{M}_4^{II,1}$ | .2866E-02 | 41118 | 2298 | 613 | 42641 | 856041 | 856267 | 3.78 |
| | $\mathcal{M}_4^{II,2}$ | .3088E-02 | 40670 | 2634 | 487 | 42262 | 1077186 | 1077430 | 5.46 |
| $10^{-6}$ | $y_1$ | .3999E-03 | 63924 | 7352 | 167 | 67612 | 1906410 | 1906888 | 6.49 |
| | $\mathcal{L}_3^0$ | .3930E-03 | 63902 | 5268 | 1122 | 67373 | 1388517 | 1388662 | 4.22 |
| | $\mathcal{L}_3^1$ | .3806E-03 | 64454 | 6980 | 469 | 68271 | 1632039 | 1632253 | 5.19 |
| | $\mathcal{M}_4^{II,1}$ | .3825E-03 | 64506 | 7118 | 365 | 68323 | 1253067 | 1253197 | 3.45 |
| | $\mathcal{M}_4^{II,2}$ | .3963E-03 | 65600 | 7066 | 939 | 69679 | 1587594 | 1587748 | 4.84 |
| $10^{-7}$ | $y_1$ | .4903E-04 | 93856 | 10204 | 164 | 98971 | 2775627 | 2776092 | 6.59 |
| | $\mathcal{L}_3^0$ | .4948E-04 | 93578 | 8092 | 1273 | 98845 | 1968534 | 1968594 | 4.09 |
| | $\mathcal{L}_3^1$ | .5121E-04 | 94148 | 9564 | 683 | 99311 | 2298426 | 2298504 | 5.04 |
| | $\mathcal{M}_4^{II,1}$ | .4670E-04 | 94740 | 9840 | 680 | 99999 | 1763409 | 1763466 | 3.32 |
| | $\mathcal{M}_4^{II,2}$ | .4775E-04 | 95422 | 9724 | 1110 | 101064 | 2216715 | 2216775 | 4.67 |

Tabla 4.4.3: Problema Ring Modulator.

eficientes que los otros cuando la tolerancia disminuye. Este hecho se refleja de forma más patente en la gráfica 4.4.5, donde se aprecia que el inicializador de orden clásico máximo $\mathcal{M}_4^{II,1}$ es ligeramente más eficiente que los otros. Debemos también hacer notar, que este problema presenta dos subintervalos de variación rápida ("transient zones") dentro del intervalo de integración $[0, 2]$. Una se corresponde con el subintervalo $[0.80, 0.81]$ y la otra con el subintervalo $[1.60, 1.61]$ (ver por ejemplo [37, pág. 113-114, 125-127]). Además, el código toma el 80% de los pasos aproximadamente en estas zonas de variación rápida, por lo que aquí los tamaños de paso deben ser fuertemente reducidos, y se puede decir que el problema deja de ser stiff en estas zonas (de hecho se puede integrar satisfactoriamente en esas zonas con métodos explícitos). Por tanto, no es de extrañar que los inicializadores con un mayor orden clásico permitan obtener integraciones más eficientes sobre este problema. De todos modos debemos resaltar que salvo el inicializador denotado por $y_1$ (el cual funciona bastante peor) todos los demás considerados no presentan diferencias sustancialmente significativas.

El problema 4 (problema E5) podemos catalogarlo como un problema stiff "delicado", que proviene de un modelo de reacción química, ver [37, pág. 145]. Este problema esta *mal escalado* y muchos códigos convencionales usados para la integración de problemas stiff, tales como LSODE [40], fallan en la integración del mismo cuando se consideran tolerancias medio–altas (RTOL$\leq 10^{-5}$), ver por ejemplo [37, pág. 153]. En la tabla 4.4.2 podemos apreciar que los inicializadores de Tipo II hacen que nuestro código falle en la integración del problema para RTOL$= 10^{-1}, 10^{-2}, 10^{-3}$. Incluso el inicializador $\mathcal{L}_3^0$ falla también para RTOL$= 10^{-1}, 10^{-2}$. Los únicos inicializadores que integran el problema satisfactoriamente en todas las tolerancias son los denotados por $\mathcal{L}_3^1$ y $y_1$, cuyos órdenes respectivos (clásico y stiff) son 2 y 0. Obsérvese que estos últimos inicializadores poseen funciones de amplificación nulas en $z = \infty$, es decir, $\hat{R}_i(\infty) = 0,\ 1 \leq i \leq 3$. Este hecho confirma en cierto modo que a la hora de elegir inicializadores para integrar problemas stiff "delicados" hay que prestar especial atención a las funciones de amplificación de error de los mismos. Incluso sería preferible elegir un inicializador con orden (clásico y/o stiff) más bajo con tal de conseguir amplificaciones de error pequeñas. En tolerancias más bajas (RTOL$\leq 10^{-6}$) todos los inicializadores presentan un comportamiento más uniforme, pero de todos modos se aprecia que los denotados por $\mathcal{L}_3^1$ y $\mathcal{L}_3^0$ son algo más eficientes. Este hecho se hace más patente en la gráfica 4.4.6.

Con respecto al problema Ring Modulator (problema 5), podemos observar en la tabla 4.4.3, que todos los inicializadores presentan un comportamiento similar, en el sentido de que el código toma números parecidos de pasos y de factorizaciones LU, salvo en el caso del inicializador $\mathcal{M}_4^{II,2}$, que conlleva un costo computacional bastante más elevado para las tolerancias más altas RTOL$= 10^{-2}, 10^{-3}$. Sin embargo, a medida que la tolerancia disminuye, este último inicializador se comporta de forma similar a los otros. Con respecto al número medio de iteraciones por paso, observamos que el inicializador denotado por $y_1$ (de orden 0) es el que presenta una cantidad más elevada. Esto se debe sin duda a su bajo orden de aproximación. En este problema se aprecia de nuevo, y de forma parecida que en el caso del problema 3, que los inicializadores de orden más alto son ligeramente más eficientes. Esto se puede ver mejor en la gráfica 4.4.7. Sospechamos que esto se debe a que el código necesita tomar pasos muy pequeños ($h_n \doteq 10^{-8}$) para la integración del problema y, por tanto, los inicializadores de más alto orden clásico pueden tener ventaja a la hora de lograr una convergencia más rápida en el proceso iterativo empleado. Esto se hace más patente, observando el número medio de iteraciones por paso (NITER), a medida que la tolerancia disminuye.

Como conclusión general de los experimentos numéricos aquí realizados, podemos inferir que

aumentando el orden de los inicializadores podemos obtener integraciones más eficientes, pero si queremos obtener un integrador robusto debemos prestar especial atención a las funciones de amplificación de error del inicializador.

En base a nuestra experimentación, y tomando como integrador la fórmula Radau IIA de 3 etapas (sospechamos que las conclusiones pueden ser generalizadas para otros métodos de la misma familia al menos, e incluso para métodos de colocación que sean stiffly accurate y que posean buenas propiedades de estabilidad), podemos decir:

1. El inicializador de orden cero, denotado por $y_1$, es bastante robusto pues hizo que el código no fallara nunca en las integraciones. Por otra parte, suele hacer las integraciones muy costosas cuando la tolerancia es baja. De todas formas, este inicializador podría recomendarse para tolerancias altas y medio–altas.

2. El inicializador denotado por $\mathcal{L}_3^1$ es bastante robusto (tampoco falló en las integraciones) y da mejores resultados en general que el $y_1$. Este inicializador se puede recomendar para la gama de tolerancias medias y altas.

3. El inicializador denotado por $\mathcal{L}_3^0$ (recomendado en [37, Cap. IV.8]) suele dar mejores resultados que el $\mathcal{L}_3^1$, cuando el primer inicializador no presenta problemas en la convergencia del proceso iterativo. Sin embargo, para cierta clase de problemas stiff el inicializador $\mathcal{L}_3^1$ es preferible por su mayor estabilidad. En cualquier caso parece que para las tolerancias más bajas puede usarse el $\mathcal{L}_3^0$, pero para las más altas sería preferible el $\mathcal{L}_3^1$.

4. El inicializador denotado por $\mathcal{M}_4^{II,1}$ suele hacer las integraciones ligeramente menos costosas, pero nos puede llevar a veces a integraciones inestables. Por tanto, este inicializador sólo se recomienda en general para las tolerancias muy bajas, o bien si tenemos la certeza de que las curvas integrales del problemas stiff considerado no se alejan mucho frente a ligeras perturbaciones en los datos iniciales. En ese caso podría usarse para una gama mayor de tolerancias. Por otra parte, el inicializador denotado por $\mathcal{M}_4^{II,2}$ no debería usarse, pues como hemos visto experimentalmente, no presenta ventaja alguna respecto de los otros.

En vista de lo anteriormente expuesto podemos decir que si tuvieramos que inclinarnos definitivamente por un inicializador específico para propósitos generales, debemos elegir el denotado por $\mathcal{L}_3^1$ y excepcionalmente, para las tolerancias más bajas, podríamos considerar el denotado por $\mathcal{L}_3^0$. Por otro lado, observamos también que si los inicializadores denotados por $\mathcal{L}_3^0$ y $\mathcal{M}_4^{II,1}$ pudieran ser *estabilizados* de alguna forma, manteniendo al mismo tiempo sus órdenes de aproximación, entonces podrían ser ambos candidatos idóneos como aproximaciones iniciales. En el próximo capítulo dirigiremos la investigación a la consecución de estos objetivos.

Capítulo V

# Inicializadores estabilizados para métodos RK de colocación

.

# Capítulo V

# Inicializadores estabilizados para métodos RK de colocación

## V.1 Introducción

En el capítulo IV se analizaron, desde varias perspectivas, dos tipos de inicializadores que apenas suponían coste computacional adicional, incluyendo además los algoritmos de arranque más utilizados en la literatura (ver [37, Cap. IV.8], [57]), para comenzar las iteraciones de alguna variante del método de Newton. Asimismo, vimos que los experimentos numéricos realizados no nos permitían inclinarnos definitivamente por un tipo particular de inicializador, pues los de mayor orden resultaron ser menos robustos. Es por ello que en este capítulo pretendemos profundizar un poco más en la investigación y proponer algunos inicializadores alternativos, de tal modo que sean robustos y posean al mismo tiempo los órdenes mayores posibles, tanto orden clásico como orden stiff. Estos nuevos inicializadores están inspirados en los previamente presentados en el capítulo anterior aunque involucran ciertas modificaciones de los mismos. Además, conllevarán un ligero coste computacional adicional por paso de integración, pero se espera que con ellos el esquema iterativo converja en menos iteraciones por paso de integración dado (en término medio), y de este modo se pueda disminuir el coste computacional global tomado por el código en la integración del problema considerado.

Este capítulo es una continuación natural de la investigación realizada en el capítulo anterior y por tanto seguiremos refiriéndonos a las notaciones allí usadas, salvo algunas excepciones que serán explicitadas aquí de forma adecuada.

Entrando en materia, vamos a considerar en principio solamente métodos RK de colocación de $s$ etapas, aunque la mayor parte del trabajo puede adaptarse a otros métodos tales como el Radau IA, Lobatto IIIC, etc.

Como ya se estableció en el capítulo anterior, si cero no es un nodo de colocación, el algoritmo de arranque obtenido de la interpolación de las etapas del paso anterior y de $y_0$ está dado por

$$Y_i^0 = l_0(1+rc_i)y_0 + \sum_{j=1}^{s} l_j(1+rc_i)X_j, \quad 1 \le i \le s, \tag{5.1.1}$$

donde $\{l_j(t)\}_{j=0}^{s}$ son los polinomios fundamentales de Lagrange con respecto a los nodos $\{c_0 = 0,\ c_1, \ldots,\ c_s\}$ dados por:

$$l_j(t) = \frac{\Pi(t)}{(t-c_j)\Pi'(c_j)}, \quad j = 0, \ldots, s, \quad \Pi(t) = (t-c_0)\cdots(t-c_s). \tag{5.1.2}$$

Este es el único inicializador de Tipo I con órdenes máximos (orden clásico $s$ y orden stiff $s$). Además, es el más usado en los códigos actuales, ver por ejemplo [37, Cap. IV.8].

Por otro lado, si aplicamos el inicializador $Y_i^0$ de (5.1.1) al problema test de Prothero y Robinson [55] dado en (4.3.1), entonces (ver sección IV.3)

$$Y_i^0 := Y_i^0(z) = R_i^0(z)(y_0 - \phi(t_0)) + \phi(t_0) + \sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!} v_{i,j}^0(z) h^j, \quad 1 \leq i \leq s, \tag{5.1.3}$$

siendo $z = \lambda h$ y $R_i^0(z)$ $(1 \leq i \leq s)$ las funciones de amplificación, que en este caso pueden expresarse como (ver (4.3.21))

$$R_i^0(z) = l_0(1 + rc_i) + \alpha_i^T (I - zA)^{-1} e, \quad \alpha_i^T = (l_1(1 + rc_i), \ldots, l_s(1 + rc_i)), \quad 1 \leq i \leq s. \tag{5.1.4}$$

Los coeficientes $v_{i,j}^0(z)$ en (5.1.3) se calculan de acuerdo con (4.3.11) obteniéndose en este caso

$$v_{i,j}^0(z) = \alpha_i^T A (I - zA)^{-1} (-zc^j + jc^{j-1}), \quad 1 \leq i \leq s, \quad j = 1, 2, \ldots \tag{5.1.5}$$

Por otra parte, para las etapas internas $Y_i$ del Runge–Kutta bajo consideración, se obtiene el desarrollo en potencias de $h$ dado en (4.3.5)-(4.3.7)-(4.3.8). También se demostró que en el caso de métodos RK de colocación, sus coeficientes se pueden expresar (ver(4.3.9)) como

$$v_{i,j}(z) = (1 + rc_i)^j, \quad (1 \leq i, j \leq s). \tag{5.1.6}$$

Además, suponiendo que el método es ASI–estable (recuérdese que si la matriz $A$ del método RK es regular, la ASI–estabilidad implica la AS–estabilidad), se tiene la acotación (4.3.10). También es fácil probar que

$$v_{i,j}(\infty) = (1 + rc_i)^j, \quad 1 \leq i \leq s, \quad j > s. \tag{5.1.7}$$

Se sigue entonces de (4.3.5) que

$$Y_i(\infty) = \phi(t_0) + \sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!} (1 + rc_i)^j h^j = \phi(t_0 + (1 + rc_i)h), \quad 1 \leq i \leq s.$$

Por otra parte, usando (5.1.3)–(5.1.4)–(5.1.5) es inmediato probar

$$Y_i^0(\infty) = l_0(1 + rc_i)(y_0 - \phi(t_0)) + \phi(t_0) + \sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!} (\alpha_i^T c^j) h^j, \quad 1 \leq i \leq s. \tag{5.1.8}$$

Ahora teniendo en cuenta que

$$\alpha_i^T c^j = \sum_{k=1}^{s} l_k (1 + rc_i) c_k^j = \sum_{k=0}^{s} l_k (1 + rc_i) c_k^j = (1 + rc_i)^j, \quad 1 \leq i, j \leq s,$$

se obtiene de (5.1.8) que

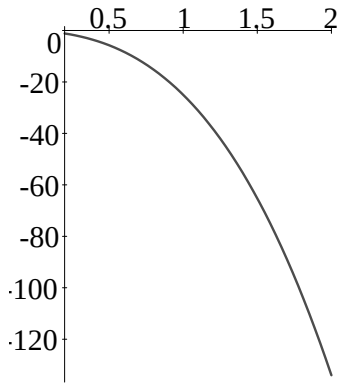
$$\left.\begin{aligned} Y_i^0(\infty) = l_0(1 + rc_i)(y_0 - \phi(t_0)) + \phi(t_0 + (1 + rc_i)h) + C_i(t_0, h) h^{s+1}, \quad 1 \leq i \leq s, \\ C_i(t_0, h) = \phi^{(s+1)}(t_0) \Pi (1 + rc_i)/(s + 1)! + \mathcal{O}(h). \end{aligned}\right\} \tag{5.1.9}$$

De las expresiones anteriores para $Y_i(\infty)$ e $Y_i^0(\infty)$, vemos claramente que el error global en el punto actual $t_0$, $y_0 - \phi(t_0)$, puede afectar de manera negativa al inicializador si $|l_0(1 + rc_i)|$

es una cantidad grande. Este factor también contribuye a amplificar los errores de redondeo y cualquier otro tipo de errores acumulados sobre el punto $t_0$. Para hacernos una idea de este factor de amplificación del error $l_0(1+rc_i)$, consideremos el caso del Radau IIA de tres etapas, para el que $c_1 = (4-\sqrt{6})/10,\ c_2 = (4+\sqrt{6})/10,\ c_3 = 1$. Así tenemos

$$l_0(1+rc_3) = -\frac{(1-c_1+r)(1-c_2+r)r}{c_1c_2c_3}.$$

En la gráfica 5.1.1 hemos representado los valores de $l_0(1+rc_3)$ en función de $r \in [0,2]$.




GRÁFICA 5.1.1: Valores de $l_0(1+rc_3)$ en función de $r \in [0,2]$.

En particular, para $r = 1/2,\ 1,\ 3/2,\ 2$ resulta

$$l_0(1+(1/2)c_3) = -5.75, \quad l_0(1+c_3) = -25, \quad l_0(1+(3/2)c_3) = -65.25, \quad l_0(1+2c_3) = -134.$$

Obsérvese también que a medida que el número de etapas $s$ del método Runge–Kutta considerado aumenta, el valor de $|l_0(1+rc_s)|$ suele hacerse mayor. Este hecho puede provocar ciertas dificultades en la convergencia de las iteraciones de tipo Newton, e incluso puede llevar a convergencia numérica (para la iteración considerada) hacia una solución que no es lo suficientemente aproximada (a la solución exacta del sistema algebraico), sobre todo cuando se integra cierta clase de problemas stiff "delicados" (por ejemplo, problemas stiff "mal escalados" o problemas stiff cuya solución exacta está cerca de una región de inestabilidad). A veces este hecho puede hacer que el código no complete la integración del PVI o dé una solución totalmente errónea en el punto final, tal como se reflejó en la tabla 4.4.2 para el problema E5 [37, pág. 145].

A la vista de lo expuesto anteriormente y a efectos de reducir la amplificación de error, un posible remedio es considerar la interpolación de Lagrange de las etapas internas $X_i$. En este caso, el inicializador es de la forma

$$\hat{Y}_i^0 = \sum_{j=1}^{s} \hat{l}_j(1+rc_i)X_j, \quad 1 \le i \le s, \tag{5.1.10}$$

donde $\{\hat{l}_j(t)\}_{j=1}^s$ son los polinomios básicos de Lagrange con respecto a los nodos $\{c_1,\ldots,c_s\}$, esto es,

$$\hat{l}_j(t) = \frac{\hat{\Pi}(t)}{(t-c_j)\hat{\Pi}'(c_j)}, \quad j = 1,\ldots,s, \quad \hat{\Pi}(t) = (t-c_1)\cdots(t-c_s). \tag{5.1.11}$$

Para el problema test de Prothero y Robinson (4.3.1) obtenemos ahora (ver sección IV.3)

$$\hat{Y}_i^0 = \hat{R}_i^0(z)(y_0 - \phi(t_0)) + \phi(t_0) + \sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!}\hat{v}_{i,j}^0(z)h^j, \quad 1 \leq i \leq s, \tag{5.1.12}$$

donde en este caso las funciones de amplificación de error vienen dadas por la fórmula

$$\hat{R}_i^0(z) = \hat{\alpha}_i^T(I - zA)^{-1}e, \quad \hat{\alpha}_i^T = (\hat{l}_1(1 + rc_i), \ldots, \hat{l}_s(1 + rc_i)), \quad 1 \leq i \leq s, \tag{5.1.13}$$

y los coeficientes $\hat{v}_{i,j}^0(z)$ se obtienen de

$$\hat{v}_{ij}^0(z) = \hat{\alpha}_i^T A(I - zA)^{-1}(-zc^j + jc^{j-1}), \quad 1 \leq i \leq s, \quad j = 1, 2, \ldots \tag{5.1.14}$$

Por tanto, en el punto $z = \infty$ se tiene que

$$\hat{Y}_i^0(\infty) = \phi(t_0) + \sum_{j\geq 1} \frac{\phi^{(j)}(t_0)}{j!}(\hat{\alpha}_i^T c^j)h^j, \quad 1 \leq i \leq s.$$

Se sigue ahora, procediendo de forma similar al caso anterior, que

$$Y_i(\infty) - \hat{Y}_i^0(\infty) = h^s\phi^{(s)}(t_0)\hat{\Pi}(1 + rc_i)/s! + \mathcal{O}(h^{s+1}), \quad 1 \leq i \leq s.$$

Por tanto, en este caso el factor de amplificación del error global acumulado, $(y_0 - \phi(t_0))$, se anula y por consiguiente, desde el punto de vista de la estabilidad (propagación de errores) este inicializador parece ser más apropiado que el dado en (5.1.1). Este hecho ha sido corroborado experimentalmente en el capítulo anterior (tabla 4.4.2).

El principal inconveniente del inicializador (5.1.10) es que sólo posee órdenes clásico y stiff $s - 1$. Otro inconveniente menor es que para métodos RK interesantes como el Radau IIA de tres etapas, y para algunos valores particulares de $z$ en el semieje real negativo, la función de amplificación del error $|\hat{R}_3^0(z)|$ es algo mayor que uno, para valores de $r$ cercanos a 1. Esto puede verse en la gráfica 5.2.2 dada más adelante.

Por otra parte, el inicializador sin coste adicional más estable que hemos encontrado para los métodos Radau IIA de $s$ etapas es

$$\dot{Y}_i^0 = X_s, \quad 1 \leq i \leq s. \tag{5.1.15}$$

(este inicializador fue denotado para $s = 3$ por $y_1$, en los experimentos numéricos realizados en el capítulo anterior).

Obsérvese que las funciones de amplificación de error vienen dadas por

$$\dot{R}_i^0(z) = R(z) = 1 + zb^T(I - zA)^{-1}e = e_s^T(I - zA)^{-1}e, \quad 1 \leq i \leq s.$$

Por tanto estas funciones de amplificación, independientemente de la razón de paso $r$ tomada, poseen las buenas propiedades de la función de estabilidad lineal del Radau IIA. La gran desventaja de este inicializador es que sólo posee órdenes clásico y stiff igual a cero.

El resto de este capítulo ha sido organizado de la manera siguiente. En la sección V.2 proponemos varias alternativas para estabilizar los algoritmos más prometedores desarrollados en el capítulo IV y analizamos sus propiedades de convergencia y estabilidad. Finalmente, en la sección V.3 realizamos varios experimentos numéricos a efectos de contrastar la teoría aquí desarrollada.

La investigación llevada a cabo en este capítulo está sintetizada en el trabajo [34].

## V.2 Construcción de inicializadores estabilizados

Cuando el método Runge–Kutta considerado es completamente implícito (*fully–implicit*), como es el caso de los métodos de colocación, y el PVI (4.1.1) es stiff y no lineal, el sistema algebraico (4.1.3) se resuelve invariablemente usando alguna modificación de la iteración de Newton, tal como la considerada en [5] para métodos SIRK, cuya matriz de coeficientes $A$ posee un espectro unipuntual. Cuando la matriz $A$ del RK tiene un espectro multipuntual se suele utilizar la *iteración de Newton Simplificada*, que ha sido usada entre otros por Hairer y Wanner [37] en su código RADAU5, o bien otras clases de iteración de tipo *Single–Newton*, como puede verse en [17], [18], [29], [30] y en los capítulos II y III de esta memoria. En todos los casos anteriores tenemos a nuestra disposición una factorización LU reciente de una matriz de la forma $(I - h\beta J)$, donde $J \doteq \partial f/\partial y(t_0, y_0)$, y $\beta$ es un parámetro positivo fijado por la iteración propuesta.

Utilizando esta factorización LU $= (I - h\beta J)$, vamos a estabilizar los inicializadores de mayor orden desarrollados en el capítulo anterior, de tal modo que al mismo tiempo pretendemos mantener si es posible los órdenes clásico, Prothero–Robinson y stiff del inicializador considerado.

**Estabilización de Tipo I**

Si denotamos

$$\begin{aligned} P(\tau) &:= l_0(\tau)y_0 + \textstyle\sum_{j=1}^{s} l_j(\tau)X_j, \\ \hat{P}(\tau) &:= \textstyle\sum_{j=1}^{s} \hat{l}_j(\tau)X_j, \end{aligned} \tag{5.2.1}$$

donde $\{l_j(\tau),\ j = 0, \ldots, s\}$ y $\{\hat{l}_j(\tau),\ j = 1, \ldots, s\}$ se definen en (5.1.2) y (5.1.11) respectivamente, tenemos que los inicializadores (5.1.1) y (5.1.10) vienen dados por

$$Y_i^0 = P(1 + rc_i), \quad \hat{Y}_i^0 = \hat{P}(1 + rc_i), \quad 1 \le i \le s. \tag{5.2.2}$$

Definimos ahora un nuevo inicializador como sigue:

$$E_i^0 = \hat{Y}_i^0 + (I - h\beta J)^{-1}(Y_i^0 - \hat{Y}_i^0), \quad 1 \le i \le s. \tag{5.2.3}$$

Este inicializador puede reescribirse de manera tal que para calcularlo sólo se necesite la solución de un sistema lineal adicional independientemente del número de etapas del RK considerado, donde la matriz $(I - h\beta J) =$ LU, es la matriz de coeficientes de dicho sistema lineal. Para ver esto, usamos la siguiente igualdad que se demuestra directamente (siendo $l_j$ y $\hat{l}_j$ los polinomios de Lagrange definidos en (5.1.2) y (5.1.11) respectivamente),

$$l_j(\tau) - \hat{l}_j(\tau) = \frac{\hat{\Pi}(\tau)}{c_j\hat{\Pi}'(c_j)}, \quad 1 \le j \le s, \qquad l_0(\tau) = \frac{(-1)^s}{c_1 \cdots c_s}\hat{\Pi}(\tau). \tag{5.2.4}$$

De aquí obtenemos

$$P(\tau) - \hat{P}(\tau) = \hat{\Pi}(\tau)\left(\frac{(-1)^s}{c_1 \cdots c_s}y_0 + \sum_{j=1}^{s}\frac{1}{c_j\hat{\Pi}'(c_j)}X_j\right).$$

Por tanto, el nuevo inicializador puede expresarse como

$$\begin{aligned} E_i^0 &= \hat{Y}_i^0 + \hat{\Pi}(1 + rc_i)(I - h\beta J)^{-1}V, \quad 1 \le i \le s, \quad \text{con} \\ V &:= (-1)^s(c_1 \cdots c_s)^{-1}y_0 + \textstyle\sum_{j=1}^{s}(c_j\hat{\Pi}'(c_j))^{-1}X_j. \end{aligned} \tag{5.2.5}$$

Obsérvese que también pueden escribirse los inicializadores anteriores mediante una fórmula alternativa que use diferencias divididas.

Nuestro principal resultado respecto al orden y la estabilidad de este nuevo algoritmo de arranque es el siguiente:

**Teorema V.2.1** *Consideremos un método RK de colocación (sobre $s$ nodos distintos y no nulos) y supongamos que dicho método es ASI–estable. Entonces, tomando $J = \partial f/\partial y(t_0, y_0)$ y cualquier constante $\beta > 0$, tenemos que*

*(a) El inicializador (5.2.3) tiene orden clásico $s$.*

*(b) Alcanza orden $s-1$ sobre la ecuación de Prothero y Robinson. Además, sus funciones de amplificación del error $S_i^0(z)$ verifican*

$$S_i^0(\infty) = 0, \quad i = 1, \ldots, s.$$

*(c) Posee orden stiff $s-1$ si el método RK subyacente es diagonalmente estable, i.e., si existe una matriz diagonal definida positiva $D$ tal que $Q = DA + A^T D$ es definida positiva.*

Demostración. *(a)* El inicializador se puede escribir de la forma

$$E_i^0 = -\beta hJ(I - \beta hJ)^{-1}\hat{Y}_i^0 + (I - h\beta J)^{-1}Y_i^0, \quad 1 \le i \le s,$$

y las etapas internas del RK $(A, b)$ considerado como

$$Y_i = -\beta hJ(I - \beta hJ)^{-1}Y_i + (I - h\beta J)^{-1}Y_i, \quad 1 \le i \le s.$$

De esto tenemos

$$Y_i - E_i^0 = -\beta hJ(I - \beta hJ)^{-1}(Y_i - \hat{Y}_i^0) + (I - h\beta J)^{-1}(Y_i - Y_i^0), \quad 1 \le i \le s. \tag{5.2.6}$$

Ahora bien, considerando que $Y_i - \hat{Y}_i^0 = \mathcal{O}(h^s)$, $Y_i - Y_i^0 = \mathcal{O}(h^{s+1})$, y que $\beta hJ(I - \beta hJ)^{-1} = \mathcal{O}(h)$ para el caso no stiff (ver sección IV.2), obtenemos el resultado buscado.

*(b)* Para el modelo de Prothero y Robinson (4.3.1) tenemos

$$E_i^0 = S_i^0(z)(y_0 - \phi(t_0)) + \phi(t_0) + \sum_{j \ge 1} \frac{\phi^{(j)}(t_0)}{j!} w_{i,j}^0(z) h^j, \quad 1 \le i \le s, \tag{5.2.7}$$

donde

$$S_i^0(z) = \hat{R}_i^0(z) + (1 - \beta z)^{-1}(R_i^0(z) - \hat{R}_i^0(z)), \quad 1 \le i \le s, \tag{5.2.8}$$

y

$$w_{i,j}^0(z) = \hat{v}_{i,j}^0(z) + (1 - \beta z)^{-1}(v_{i,j}^0(z) - \hat{v}_{i,j}^0(z)), \quad 1 \le i \le s,\ j \ge 1, \tag{5.2.9}$$

siendo $R_i^0(z)$, $v_{i,j}^0(z)$, $\hat{R}_i^0(z)$, $\hat{v}_{i,j}^0(z)$ las funciones definidas en (5.1.4), (5.1.5), (5.1.13) y (5.1.14) respectivamente.

De estas igualdades y considerando de nuevo (5.2.6), la afirmación dada en el teorema se prueba inmediatamente después de usar los resultados dados en la sección IV.3 (Teorema IV.3.2).

*(c)* Si el RK $(A, b)$ es diagonalmente estable, entonces considerando $y_0 = y(t_0)$, donde $y(t)$ es la solución exacta de (4.1.1), y suponiendo que el sistema diferencial es contractivo (ver [19, Cap. I]), puede verse que (ver detalles en el Teorema IV.3.5 del capítulo anterior)

$$Y_i - \hat{Y}_i^0 = \mathcal{O}(h^s), \quad Y_i - Y_i^0 = \mathcal{O}(h^{s+1}), \quad 1 \le i \le s,$$

donde las cantidades $\mathcal{O}(h^s)$, $\mathcal{O}(h^{s+1})$ son independientes de la stiffness del problema. Además, como consecuencia del Teorema de von Neumann (ver por ejemplo, [37, Cap. IV.11]) tenemos que,

$$|-\beta hJ(I-\beta hJ)^{-1}|_2 \leq \sup_{\mathrm{Re}\, z\leq 0} |-z(1-z)^{-1}| = 1, \qquad |(I-\beta hJ)^{-1}|_2 \leq \sup_{\mathrm{Re}\, z\leq 0} |(1-z)^{-1}| = 1.$$

De aquí se deduce directamente que el inicializador $E_i^0$ alcanza orden stiff $s-1$. ∎

**Estabilización de Tipo II**

Ahora vamos a considerar la estabilización de otro algoritmo de arranque interesante propuesto en el capítulo anterior en el Teorema IV.2.2 y en la Nota IV.2.2. Dicho inicializador es el único de Tipo II que tiene orden clásico $s+1$ y orden stiff $s$ para métodos RK de colocación (siempre que cero no sea uno de sus nodos). El principal inconveniente de este algoritmo es que sus funciones de amplificación no están acotadas cuando $z \to \infty$. Como ya vimos en el capítulo IV, dicho inicializador se puede expresar como sigue:

$$\begin{gathered} \tilde{Z}_i^0 = y_1 + rh\textstyle\sum_{j=1}^s a_{ij}\tilde{g}_j \quad (i=1,\ldots,s) \qquad \text{donde} \\ \tilde{g}_i = l_0(1+rc_i)f(t_0,y_0) + \textstyle\sum_{j=1}^s l_j(1+rc_i)f(t_0+hc_j,X_j), \end{gathered} \tag{5.2.10}$$

con los $l_j(t)$ definidos por (5.1.2).

Junto con $\tilde{Z}_i^0$ también vamos a considerar el inicializador

$$\begin{gathered} Z_i^0 = y_1 + rh\textstyle\sum_{j=1}^s a_{ij}g_j \quad (i=1,\ldots,s), \qquad \text{donde} \\ g_i = \textstyle\sum_{j=1}^s \hat{l}_j(1+rc_i)f(t_0+hc_j,X_j), \end{gathered} \tag{5.2.11}$$

con los $\hat{l}_j(t)$ definidos en (5.1.11). Como se vio en el Teorema IV.2.2, este inicializador coincide con el de la interpolación de Lagrange de las etapas $X_i$ $(1 \leq i \leq s)$ y de $y_0$, el cual hemos denotado en el presente capítulo por $Y_i^0$, $(1 \leq i \leq s)$, ver (5.1.1).

A partir de estos dos algoritmos de arranque definiremos ahora una nueva familia de inicializadores dependientes de $s$ parámetros $\theta_i$ $(i=1,\ldots,s)$, donde dichos parámetros podrán depender a su vez de $r$ para algunos casos particulares (debido principalmente a razones de estabilidad), pero siempre serán de tamaño moderado. La nueva familia de inicializadores viene dada por

$$\bar{E}_i^0 = Z_i^0 + \theta_i(I-\beta hJ)^{-1}(\tilde{Z}_i^0 - Z_i^0), \quad 1 \leq i \leq s. \tag{5.2.12}$$

Nótese que usando (5.2.4) los nuevos algoritmos de esta familia pueden reescribirse de la siguiente manera:

$$\begin{gathered} \bar{E}_i^0 = Z_i^0 + r\theta_i\left(\textstyle\sum_{j=1}^s a_{ij}\hat{\Pi}(1+rc_j)\right)(I-h\beta J)^{-1}W, \quad 1 \leq i \leq s, \quad \text{con} \\ W := h\left((-1)^s(c_1\cdots c_s)^{-1}f(t_0,y_0) + \textstyle\sum_{j=1}^s (c_j\hat{\Pi}'(c_j))^{-1}f(t_0+hc_j,X_j)\right). \end{gathered} \tag{5.2.13}$$

Así se ve claramente que para implementar los nuevos inicializadores sólo se necesita la solución de un sistema lineal adicional con respecto a (5.2.10).

Por otro lado, desde el punto de vista computacional, es mejor expresar los nuevos inicializadores (5.2.13) como una combinación lineal de las etapas $X_j$ $(j=1,\ldots,s)$, de $y_0$ y de

$hf(t_0, y_0)$. Para este fin, usando (4.1.3) y teniendo en cuenta que $Z_i^0 = Y_i^0$, $(1 \le i \le s)$, es inmediato comprobar tras unos cálculos sencillos que,

$$\begin{aligned} \bar{E}_i^0 &= Y_i^0 + r\theta_i \left(\textstyle\sum_{j=1}^s a_{ij}\hat{\Pi}(1+rc_j)\right)(I - h\beta J)^{-1}W, \quad 1 \le i \le s, \quad \text{con} \\ W &:= (-1)^s(c_1 \cdots c_s)^{-1}hf(t_0, y_0) - \kappa y_0 + (\tau^T A^{-1} \otimes I)X, \qquad \text{y} \\ \tau &= ((c_1\hat{\Pi}'(c_1))^{-1}, \ldots, (c_s\hat{\Pi}'(c_s))^{-1})^T, \quad \kappa = \tau^T A^{-1} e. \end{aligned} \tag{5.2.14}$$

Además en el caso de que el método RK considerado sea stiffly accurate, también se puede computar fácilmente la cantidad $\bar{h}f(t_1, y_1)$ (sin evaluar $f$) la cual es necesaria para el siguiente paso de integración, expresándola como una combinación lineal de las etapas internas del paso anterior, mediante la fórmula

$$\bar{h}f(t_1, y_1) = r\left(\sum_{j=1}^s u_j X_j - (e_s^T A^{-1} e)y_0\right), \quad u^T = (u_1, \ldots, u_s) = e_s^T A^{-1}.$$

Sin embargo, si el RK subyacente no fuera stiffly accurate, los inicializadores (5.2.10) y (5.2.13) necesitarían, en comparación con (5.2.11), la computación de una función derivada adicional por paso de integración.

El siguiente resultado nos da los órdenes de la nueva familia de inicializadores así como sus funciones de amplificación del error.

**Teorema V.2.2** *Consideremos un método RK de colocación (sobre $s$ nodos distintos y no nulos) y supongamos que dicho método es ASI–estable. Entonces, tomando $J = \partial f/\partial y(t_0, y_0)$ y cualquier constante $\beta > 0$, se tiene que*

*(a) La familia de inicializadores (5.2.12) tiene orden clásico $s$. Además, para el caso particular $\theta_i = 1$, $(i = 1, \ldots, s)$, el algoritmo resultante alcanza orden clásico $s + 1$.*

*(b) Toda la familia alcanza orden $s$ sobre la ecuación de Prothero y Robinson, y sus funciones de amplificación de error $\bar{S}_i^0(z)$ verifican*

$$\bar{S}_i^0(\infty) = l_0(1 + rc_i) - \theta_i\beta^{-1}r\sum_{j=1}^s a_{ij}l_0(1 + rc_j), \quad i = 1, \ldots, s. \tag{5.2.15}$$

*Además, no existe ninguna elección posible de los parámetros tal que el algoritmo resultante consiga orden $s+1$ sobre la ecuación de Prothero y Robinson. Por otra parte, para el caso*

$$\theta_i = \beta l_0(1 + rc_i)(r\sum_{j=1}^s a_{ij}l_0(1 + rc_j))^{-1}, \quad i = 1, \ldots, s, \tag{5.2.16}$$

*suponiendo que el denominador no se anula (ver el Lema V.2.1 dado a continuación), el inicializador resultante verifica*

$$\bar{S}_i^0(\infty) = 0, \quad i = 1, \ldots, s. \tag{5.2.17}$$

*(c) Todos los algoritmos de la familia poseen orden stiff $s$ si el método RK es diagonalmente estable.*

*Demostración.* *(a)-(c)* Como las etapas internas del método RK satisfacen

$$Y_i = Y_i + \theta_i(I - h\beta J)^{-1}(Y_i - Y_i), \quad 1 \le i \le s,$$

restando (5.2.12) de esta última igualdad se sigue

$$Y_i - \bar{E}_i^0 = (I - \theta_i(I - \beta hJ)^{-1})(Y_i - Z_i^0) + \theta_i(I - h\beta J)^{-1}(Y_i - \tilde{Z}_i^0), \quad 1 \le i \le s. \tag{5.2.18}$$

Con esto concluimos la demostración de *(a)-(c)* (respecto al orden $s$) simplemente usando que para los casos no stiff y stiff tenemos (ver Teoremas IV.2.2 y IV.3.3),

$$Y_i - \tilde{Z}_i^0 = \mathcal{O}(h^{s+1}), \quad Y_i - Z_i^0 = \mathcal{O}(h^{s+1}), \quad (I - \beta hJ)^{-1} = \mathcal{O}(1).$$

Para probar el orden clásico $s+1$ (para el caso no stiff) para el caso particular $\theta_i = 1$, $(i = 1, \ldots, s)$ tenemos del Teorema IV.2.2 que

$$Y_i - \tilde{Z}_i^0 = \mathcal{O}(h^{s+2}), \quad Y_i - Z_i^0 = \mathcal{O}(h^{s+1}), \quad I - (I - \beta hJ)^{-1} = \mathcal{O}(h),$$

con lo que se obtiene directamente el resultado.

*(b)* Si consideramos el modelo de Prothero y Robinson (4.3.1) obtenemos las siguientes expresiones para las funciones de amplificación del error $\bar{S}_i^0(z)$ y los coeficientes $\bar{w}_{i,j}^0(z)$ de los algoritmos de la familia (5.2.12),

$$\bar{E}_i^0 = \bar{S}_i^0(z)(y_0 - \phi(t_0)) + \phi(t_0) + \sum_{j \ge 1} \frac{\phi^{(j)}(t_0)}{j!} \bar{w}_{i,j}^0(z) h^j, \quad 1 \le i \le s, \tag{5.2.19}$$

donde

$$\bar{S}_i^0(z) = R_i^0(z) + \theta_i(1 - \beta z)^{-1}(\tilde{R}_i^0(z) - R_i^0(z)), \quad 1 \le i \le s, \tag{5.2.20}$$

y

$$\bar{w}_{i,j}^0(z) = v_{i,j}^0(z) + \theta_i(1 - \beta z)^{-1}(\tilde{v}_{i,j}^0(z) - v_{i,j}^0(z)), \quad 1 \le i \le s,\ j \ge 1, \tag{5.2.21}$$

denotando por $\tilde{R}_i^0(z)$, $R_i^0(z)$, $\tilde{v}_{i,j}^0(z)$ y $v_{i,j}^0(z)$ las funciones de amplificación y los coeficientes de los algoritmos (5.2.10) y (5.2.11) (ó (5.1.1)) respectivamente, cuando se aplican al modelo de Prothero y Robinson.

Como $Z_i^0 = Y_i^0$, $(i = 1, \ldots, s)$, tenemos de (5.1.4) que

$$R_i^0(\infty) = l_0(1 + rc_i), \quad 1 \le i \le s. \tag{5.2.22}$$

Por otra parte, no es difícil deducir que (ver Sección IV.3 o aplicar directamente (5.2.10) sobre el test $y' = \lambda y$),

$$\tilde{R}_i^0(z) = R(z) + zr \sum_{j=1}^{s} a_{ij} \left( l_0(1 + rc_j) + \sum_{k=1}^{s} l_k(1 + rc_j) T_k(z) \right), \quad 1 \le i \le s, \tag{5.2.23}$$

donde $R(z)$, $T_i(z)$ están definidos en (4.3.8) y los $l_j(\tau)$ en (5.1.2). Por consiguiente,

$$\lim_{z \to \infty} \frac{\tilde{R}_i^0(z)}{1 - \beta z} = -r\beta^{-1} \sum_{j=1}^{s} a_{ij} l_0(1 + rc_j), \quad i = 1, \ldots, s.$$

Usando ahora (5.2.22) se sigue inmediatamente de (5.2.20) las expresiones dadas en (5.2.15) y en (5.2.16)-(5.2.17).

Finalmente, para probar que no se puede alcanzar el orden $s+1$ sobre todo el semiplano complejo negativo, vamos a proceder por reducción al absurdo. Así, supongamos que se alcanza orden $s+1$ en los puntos particulares $z=0$ y $z=\infty$ para una elección fija de los $\theta_i$–parámetros. Esto implicaría de (5.2.21) que

$$\bar{w}^0_{i,s+1}(0) = v^0_{i,s+1}(0) + \theta_i(\tilde{v}^0_{i,s+1}(0) - v^0_{i,s+1}(0)) = v_{i,s+1}(0), \quad 1 \le i \le s, \tag{5.2.24}$$

donde los $v_{i,j}(z)$ vienen dados por (4.3.5). Por otro lado, como el algoritmo (5.2.10) posee orden clásico $s+1$, tenemos que $\tilde{v}^0_{i,s+1}(0) = v_{i,s+1}(0)$. Además, ya que el inicializador (5.2.11) (ó (5.1.1)) tiene exactamente orden $s$, no es difícil probar que $v^0_{i,s+1}(0) \neq \tilde{v}^0_{i,s+1}(0)$. Por tanto, de (5.2.24) obtenemos que $\theta_i = 1,\ (i = 1, \ldots, s)$.

Ahora bien, considerando el caso $z=\infty$ tenemos de (5.2.21) que

$$\bar{w}^0_{i,s+1}(\infty) = v^0_{i,s+1}(\infty).$$

Por otra parte, teniendo en cuenta (5.1.5), resulta

$$v^0_{i,s+1}(\infty) = \alpha_i^T c^{s+1}. \tag{5.2.25}$$

De (5.1.7) se sigue inmediatamente que

$$v_{i,s+1}(\infty) - v^0_{i,s+1}(\infty) = (1+rc_i)^{s+1} - \alpha_i^T c^{s+1} = \Pi(1+rc_i) \neq 0, \quad 1 \le i \le s,$$

donde $\Pi(t)$ está dado por (5.1.2). Esto nos lleva a una contradicción con el hecho de haber asumido orden $s+1$ en $z=\infty$. ∎

**Lema V.2.1** *Sea un método RK $(A,b)$ de colocación sobre $s$ nodos, $c_1 < c_2 < \ldots < c_s$. Entonces, una condición suficiente para que*

$$r\sum_{j=1}^{s} a_{ij} l_0(1+rc_j) \neq 0, \quad \forall r > 0, \quad 1 \le i \le s,$$

*es:*

$$0 < c_i \le 1, \qquad e_i^T A c^s \ge 0, \quad 1 \le i \le s. \tag{5.2.26}$$

*Demostración.* De (5.1.11) y (5.2.4) tenemos que

$$r\sum_{j=1}^{s} a_{ij} l_0(1+rc_j) = r\frac{(-1)^s}{c_1 \cdots c_s} \sum_{j=1}^{s} a_{ij} \hat{\Pi}(1+rc_j),$$

por tanto basta probar que,

$$\rho_i(r) := \sum_{j=1}^{s} a_{ij} \hat{\Pi}(1+rc_j) \neq 0, \quad \forall r > 0, \quad 1 \le i \le s. \tag{5.2.27}$$

Para demostrar esto último, denotemos por

$$d_j(r) := \hat{\Pi}(1+rc_j),\ 1 \le j \le s.$$

Al ser los $d_j(r)$ polinomios en $r$ de grado menor o igual que $s$, usando el desarrollo de Taylor (en torno a $r=0$) obtenemos para $j = 1, \ldots s$,

$$d_j(r) = \sum_{k=0}^{s} d_j^{(k)}(0)\frac{r^k}{k!},$$
$$d_j^{(k)}(0) = c_j^k \hat{\Pi}^{(k)}(1), \quad 0 \le k \le s.$$

Llevando esto a (5.2.27) se sigue que,

$$\rho_i(r) = \sum_{j=1}^{s} a_{ij} \sum_{k=0}^{s} c_j^k \hat{\Pi}^{(k)}(1) \frac{r^k}{k!} = \sum_{k=0}^{s} \left[ \sum_{j=1}^{s} a_{ij} c_j^k \right] \hat{\Pi}^{(k)}(1) \frac{r^k}{k!}.$$

Usando ahora la condición $C(s)$ se deduce inmediatamente que,

$$\left.\begin{aligned} \rho_i(r) &= \sum_{k=0}^{s} \xi_{ik} \frac{r^k}{k!}, \\ \xi_{ik} &= \frac{c_i^{k+1}}{k+1} \hat{\Pi}^{(k)}(1), \quad 0 \le k \le s-1, \\ \xi_{is} &= A_i^T c^s \hat{\Pi}^{(s)}(1). \end{aligned}\right\} \tag{5.2.28}$$

De las hipótesis impuestas se sigue que

$$\xi_{ik} \ge 0, \quad 1 \le i \le s, \quad 0 \le k \le s.$$

Por lo tanto, si $s = 1$ de (5.2.28) se sigue que

$$\xi_{i1} = (e_1^T A c) \hat{\Pi}'(1) = c_1^2 > 0.$$

Por otra parte, si $s > 1$, entonces también se deduce de (5.2.28) que

$$\xi_{i1} = \frac{c_i^2}{2} \hat{\Pi}'(1) > 0.$$

Luego, en cualquiera de las situaciones tenemos que,

$$\rho_i(r) > 0, \qquad \forall r > 0, \qquad (1 \le i \le s).$$

Esto concluye la demostración del lema. ∎

Un resultado curioso que hemos observado, tras representar las gráficas de amplificación de error de los dos tipos de inicializadores estabilizados (5.2.3) y (5.2.12) con $\theta_i$ dadas por (5.2.16), para el Radau IIA de 3 etapas, es que ambas coincidían (ver gráficas 5.2.4 y 5.2.6, más adelante). Como veremos en el siguiente teorema esto hecho no es casual.

**Teorema V.2.3** *Para métodos Runge–Kutta de colocación sobre nodos no nulos, se tiene que las funciones de amplificación de error para los inicializadores (5.2.3), coinciden respectivamente con las de los inicializadores (5.2.12)–(5.2.16).*

*Demostración.* Sean $S_i^0(z)$ y $\bar{S}_i^0(z)$ las funciones de amplificación de los inicializadores (5.2.3) y (5.2.12)–(5.2.16) respectivamente, cuyas expresiones vienen dadas por (5.2.8) y (5.2.20)-(5.2.16). Tenemos que comprobar que

$$S_i^0(z) \equiv \bar{S}_i^0(z), \quad \forall z \text{ con } Re\ z \le 0. \tag{5.2.29}$$

Para este fin, denotemos $\theta_i = \beta \eta_i$ con (ver (5.2.16))

$$\eta_i = l_0(1 + r c_i)/(r \sum_{j=1}^{s} a_{ij} l_0(1 + r c_j)). \tag{5.2.30}$$

Obtenemos ahora de (5.2.8) y (5.2.20) que (5.2.29) resulta equivalente a

$$\hat{R}_i^0(z) - R_i^0(z) = (1-\beta z)^{-1}\left(\beta\eta_i(\tilde{R}_i^0(z) - R_i^0(z)) + (\hat{R}_i^0(z) - R_i^0(z))\right).$$

Si multiplicamos esta ecuación por $\beta^{-1} - z$, entonces hemos de probar que

$$z(R_i^0(z) - \hat{R}_i^0(z)) = \eta_i(\tilde{R}_i^0(z) - R_i^0(z)). \tag{5.2.31}$$

A efectos de probar esta última igualdad, obtengamos primero una expresión adecuada para $\tilde{R}_i^0(z)$. Para ello, teniendo en cuenta (4.3.8), se sigue de (5.2.23) que

$$\tilde{R}_i^0(z) = 1 + zr\sum_{j=1}^{s} a_{ij} l_0(1+rc_j) + z(b^T + r\sum_{j=1}^{s} a_{ij}\alpha_j^T)(I-zA)^{-1}e. \tag{5.2.32}$$

Si usamos ahora (5.2.30) resulta

$$\tilde{R}_i^0(z) = 1 + z\eta_i^{-1} l_0(1+rc_i) + z(b^T + r\sum_{j=1}^{s} a_{ij}\alpha_j^T)(I-zA)^{-1}e. \tag{5.2.33}$$

Demostraremos a continuación las dos igualdades siguientes:

$$\begin{gathered} b^T + r\textstyle\sum_{j=1}^{s} a_{ij}\alpha_j^T = \alpha_i^T A - \dfrac{l_0(1+rc_i)}{\eta_i} e_1^T V^{-1}, \\ \alpha_i^T - \hat{\alpha}_i^T = -l_0(1+rc_i)e_1^T V^{-1}, \end{gathered} \tag{5.2.34}$$

donde $V = [e, c, \ldots, c^{s-1}]$, $e_1^T = (1, 0, \ldots, 0) \in I\!R^s$.

Para ver esto téngase en cuenta que de (5.1.4) y (5.1.13) se tiene respectivamente que

$$\alpha_i^T V = (1, (1+rc_i), \ldots, (1+rc_i)^{s-1}) - l_0(1+rc_i)e_1^T, \quad 1 \le i \le s,$$

$$\hat{\alpha}_i^T V = (1, (1+rc_i), \ldots, (1+rc_i)^{s-1}), \quad 1 \le i \le s.$$

Por tanto, la segunda igualdad en (5.2.34) es inmediata. Por otra parte, la primera igualdad en (5.2.34) multiplicada por la derecha por $V$, es justamente la identidad expresada en (4.2.6) (ver también (4.2.3)), teniendo en cuenta el valor de $\eta_i$ dado por (5.2.30).

Podemos escribir ahora la siguiente expresión para $\tilde{R}_i^0(z)$:

$$\tilde{R}_i^0(z) = 1 + z\eta_i^{-1} l_0(1+rc_i) + z(\alpha_i^T A - \eta_i^{-1} l_0(1+rc_i)e_1^T V^{-1})(I-zA)^{-1}e.$$

De la expresión anterior y de (5.1.4) es inmediato probar

$$\tilde{R}_i^0(z) - R_i^0(z) = \frac{z}{\eta_i} l_0(1+rc_i)\left(1 - e_1^T V^{-1}(I-zA)^{-1}e\right). \tag{5.2.35}$$

Por otra parte de (5.1.4), (5.1.13) y (5.2.34) se sigue que

$$R_i^0(z) - \hat{R}_i^0(z) = l_0(1+rc_i)\left(1 - e_1^T V^{-1}(I-zA)^{-1}e\right).$$

De esta última ecuación y de (5.2.35) se obtiene la expresión buscada en (5.2.31) ∎

### V.2.1 Gráficas de las funciones de amplificación para el Radau IIA de 3 etapas

Para ver más claramente la magnitud de las funciones de amplificación de error correspondientes a los distintos inicializadores, hemos representado en las siguientes gráficas la amplificación de error sobre la tercera etapa (o solución de avance) para el Radau IIA de 3 etapas. Hemos tomado también cuatro razones de paso distintas y que podrían ser consideradas representativas cuando el método se implementa en un código de paso variable. Estas razones de paso son $r = 1/2,\ 1,\ 3/2,\ 2$.

En la gráfica 5.2.1 representamos, a efectos de referencia, la amplificación exacta correspondiente a la tercera etapa del Radau IIA (de 3 etapas),

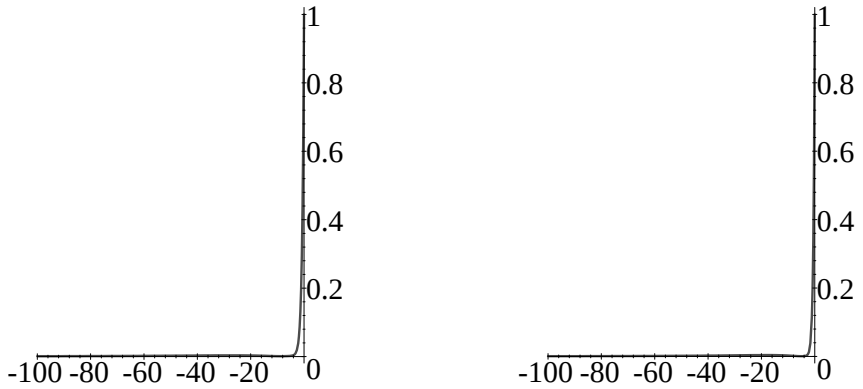
$$\mathcal{R}_3(z) = ((e_3^T(I - zA)^{-1})e)((e_3^T(I - rzA)^{-1})e).$$

para $z$ variando en el semieje real negativo.

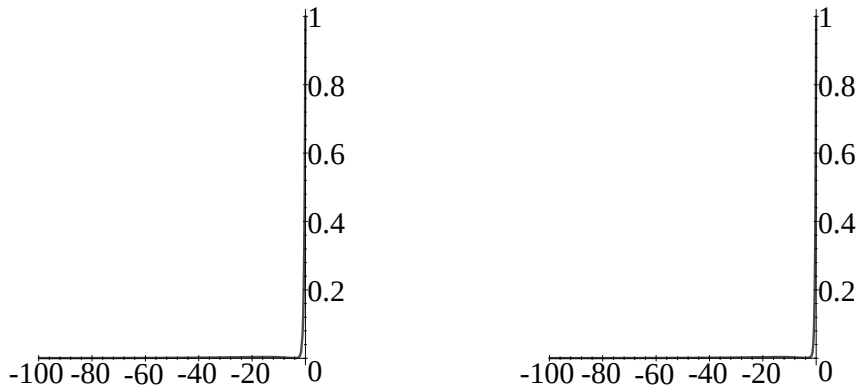
En la gráfica 5.2.2 representamos la amplificación de la tercera etapa del inicializador dado por (5.1.10). Asimismo en la gráfica 5.2.3 representamos la del inicializador dado por (5.1.1). Se observa que las amplificaciones correpondientes al inicializador (5.1.10) son en general algo menos de la décima parte de las correspondientes al (5.1.1). Por otra parte, para el valor $r = 1$ se observa que la amplificación de error dada por (5.1.10) excede en valor absoluto a la unidad para valores de $z \in [-50, -5]$, tomando un máximo aproximado de 2.2.

En las gráficas 5.2.4 y 5.2.6 representamos respectivamente las funciones de amplificación de los inicializadores estabilizados (5.2.3) y (5.2.12)-(5.2.16) (hemos tomado $\beta = 60^{-1/3}r$, pues este es el parámetro correspondiente al esquema Single–Newton considerado en [29] para el Radau IIA de 3 etapas). Dichas gráficas coinciden tal cual se expresa en el Teorema V.2.3. Se observa que desde el punto de vista de la amplificación son las más parecidas a las dadas en la gráfica 5.2.1, y por tanto son las más estables. Además para $r = 1$ se mantienen acotadas en valor absoluto por uno en todo el semieje real negativo.

Por último, consideramos en la gráfica 5.2.5 la amplificación de error de la tercera etapa, del inicializador estabilizado dado por (5.2.12) con $\theta_i = 1,\ (i = 1, 2, 3)$, el cual posee el orden máximo (clásico y stiff). Este inicializador posee amplificaciones acotadas, pero no son nulas para $z = \infty$. Se puede decir que presenta amplificaciones de error comparables al dado por (5.1.1), aunque para el primero las amplificaciones de error son en general del orden de la quinta parte de las amplificaciones del segundo, lo cual lo convierte en ligeramente más estable.

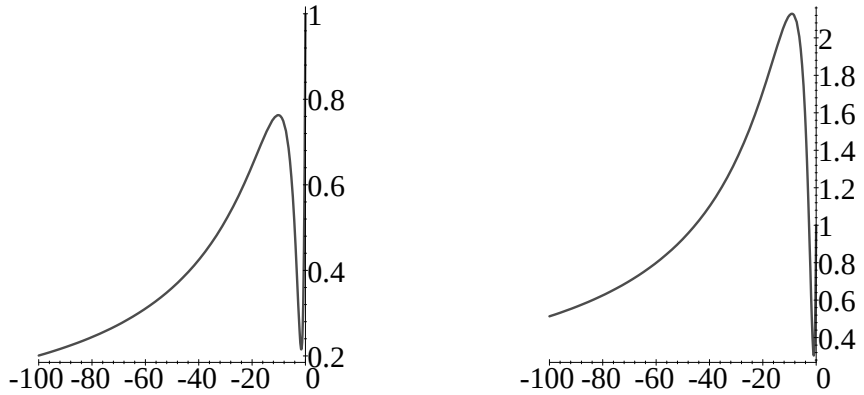


(a) (b)

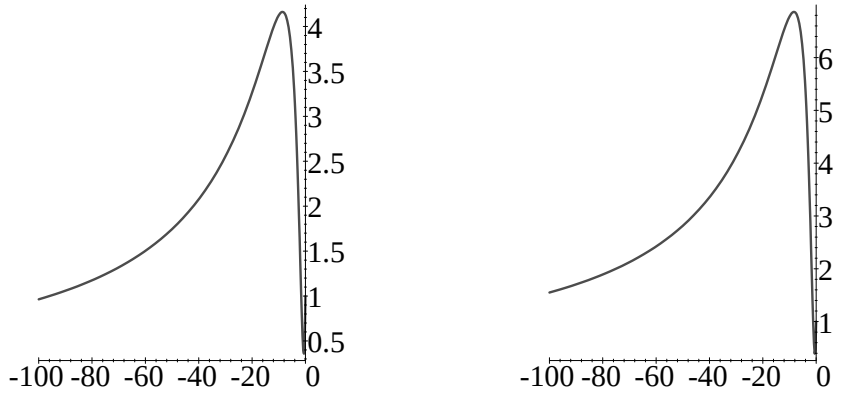



(c) (d)

Gráfica 5.2.1: Función de amplificación $\mathcal{R}_3(z)$: (a) $r = 0.5$, (b) $r = 1$, (c) $r = 1.5$, (d) $r = 2$.

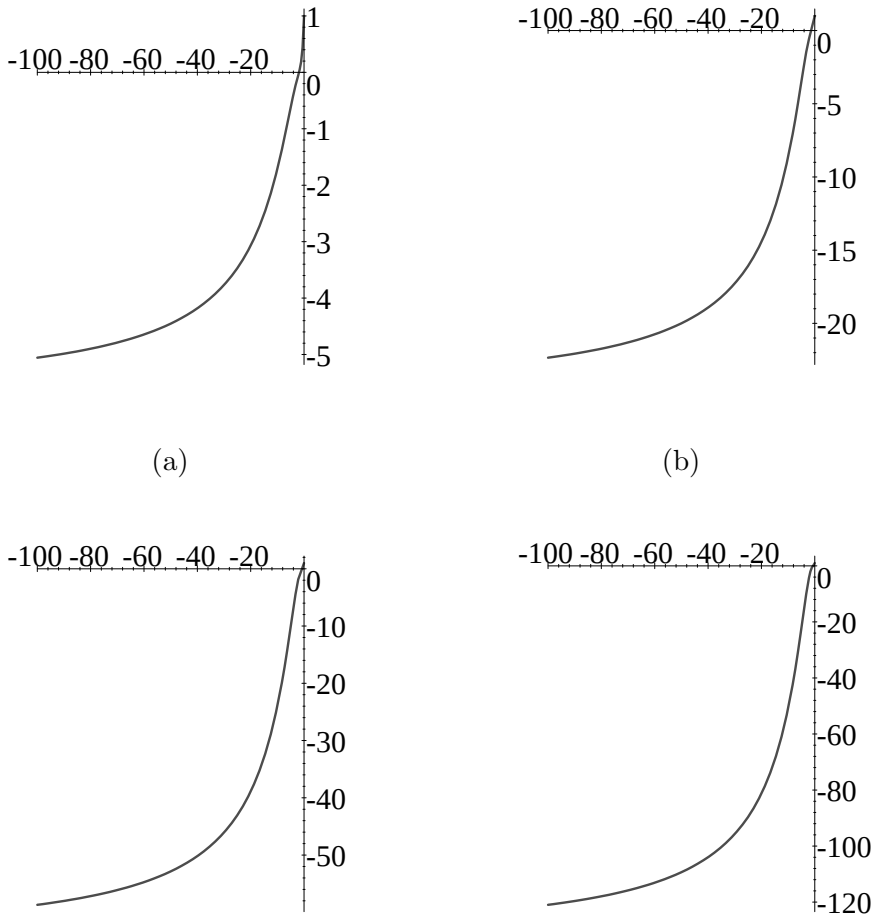



(a) (b)




(c) (d)

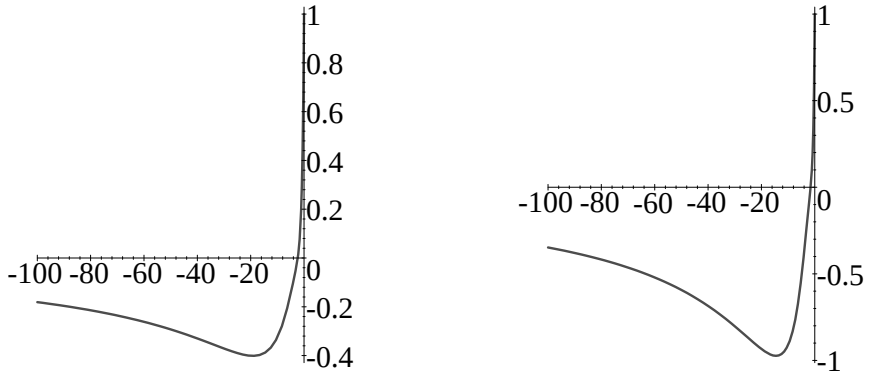
Gráfica 5.2.2: Interpolación de Lagrange $X_i$: (a) $r = 0.5$, (b) $r = 1$, (c) $r = 1.5$, (d) $r = 2$.



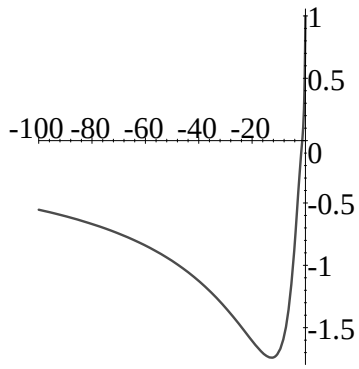

(a) (b)

(c) (d)

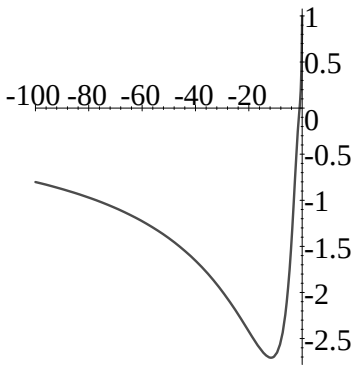
Gráfica 5.2.3: Interpolación de Lagrange $y_0, X_i$: (a) $r = 0.5$, (b) $r = 1$, (c) $r = 1.5$, (d) $r = 2$.




(a) (b)

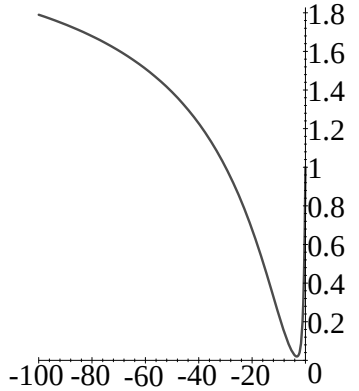



(c)

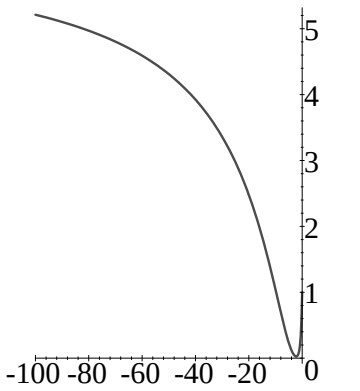



(d)

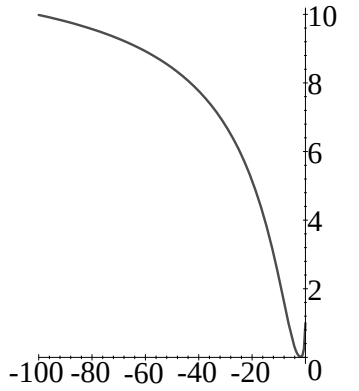
Gráfica 5.2.4: Tipo I estabilizado: (a) $r = 0.5$, (b) $r = 1$, (c) $r = 1.5$, (d) $r = 2$.



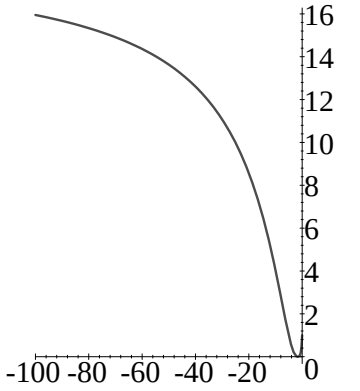

(a)




(b)

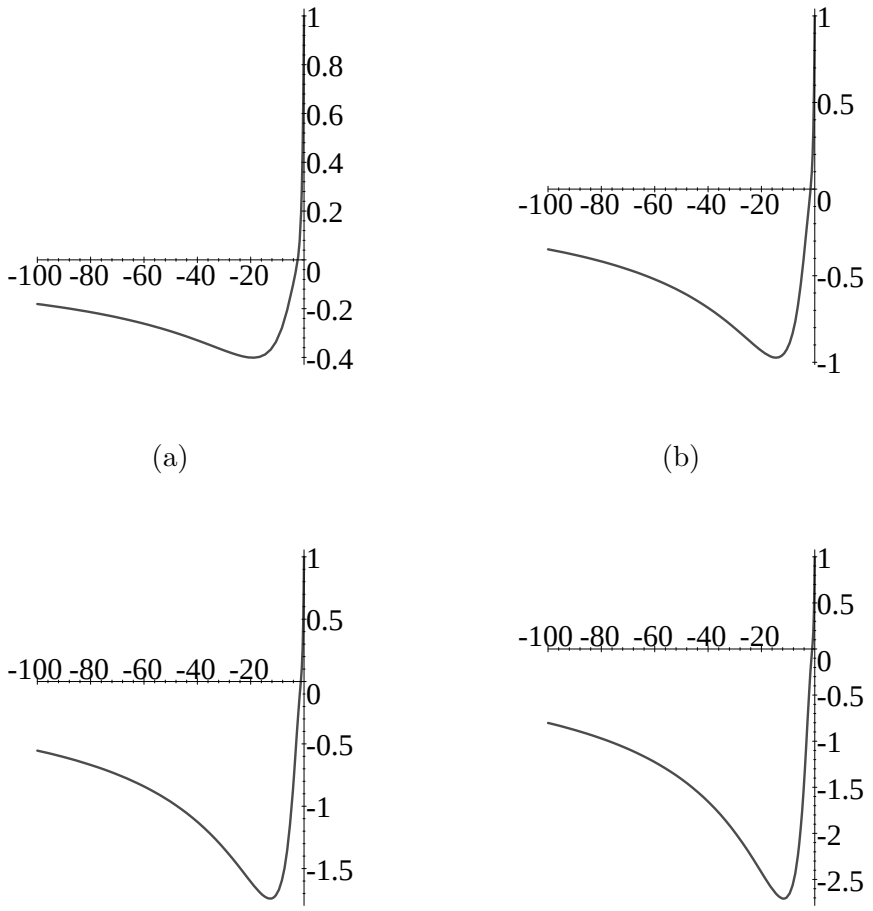



(c)




(d)

Gráfica 5.2.5: Tipo II estabilizado de orden máximo: (a) $r = 0.5$, (b) $r = 1$, (c) $r = 1.5$, (d) $r = 2$.




Gráfica 5.2.6: Tipo II estabilizado con máxima estabilidad: (a) $r = 0.5$, (b) $r = 1$, (c) $r = 1.5$, (d) $r = 2$.

### V.2.2 Inicializadores estabilizados para los métodos Lobatto IIIA

Como ya vimos en la introducción del capítulo II, los métodos Lobatto IIIA presentan varias ventajas respecto a otros métodos Runge–Kutta completamente implícitos cuando se usan para integrar problemas stiff. También se desarrollaron en los capítulos II y III iteraciones de tipo Single–Newton para resolver sus ecuaciones de etapa. De los experimentos numéricos realizados se infería que el Lobatto IIIA de 4 etapas era un integrador prometedor para problemas stiff. Por todo esto consideramos interesante realizar una investigación particularizada sobre inicializadores para arrancar los procesos iterativos cuando se aplica este método. Obsérvese que aunque los métodos Lobatto IIIA son métodos de colocación, su matriz $A$ es singular ya que el cero es un nodo de colocación. Por tanto, en este caso no se pueden aplicar directamente los teoremas vistos en las secciones anteriores de este capítulo.

Introducimos ahora la siguiente notación para los coeficientes del método Lobatto IIIA de $s$

etapas:

$$A = \begin{pmatrix} 0 & \mathbf{0}^T \\ w & \bar{A} \end{pmatrix}, \quad w := \begin{pmatrix} w_2 \\ \vdots \\ w_s \end{pmatrix} = \begin{pmatrix} a_{21} \\ \vdots \\ a_{s1} \end{pmatrix}, \quad \bar{A} = (a_{ij})_{i,j=2}^{s}, \quad \det \bar{A} \neq 0,$$
$$b^T = (b_1, \bar{b}^T), \quad c = \begin{pmatrix} 0 \\ \bar{c} \end{pmatrix}, \quad \bar{c} = (c_i)_{i=2}^{s}. \tag{5.2.36}$$

En este caso, el único inicializador de Tipo I (ver sección IV.3, teorema IV.2.1) con orden clásico máximo $s-1$ viene dado por

$$\begin{aligned} L_i^0 &= \sum_{j=1}^{s} \alpha_{ij} X_j, \quad 2 \leq i \leq s, \\ \alpha_{ij} &= \hat{l}_j(1 + rc_i), \quad 2 \leq i \leq s,\ 1 \leq j \leq s, \end{aligned} \tag{5.2.37}$$

donde los $\hat{l}_j(t)$ están dados en (5.1.11).

Las funciones de amplificación del error verifican para este inicializador (ver en la sección IV.3 el Teorema IV.3.4*(b)*),

$$R_i^0(\infty) = 1 - \bar{\alpha}_i^T \bar{A}^{-1} \bar{c}, \qquad \bar{\alpha}_i^T = (\alpha_{i2}, \ldots, \alpha_{is}), \qquad 2 \leq i \leq s. \tag{5.2.38}$$

No vamos a intentar estabilizar este inicializador ya que en este caso tanto $R_i^0(\infty)$ como $\mathcal{R}_i(\infty)$ están acotados y son no nulos (recuérdese que $\mathcal{R}_i(z)$ denota la función de amplificación del error de la etapa interna $Y_i$).

Por otro lado, en el caso de algoritmos de Tipo II, el único que alcanza orden clásico máximo $s$ es (ver Teorema IV.2.2*(a)*)

$$\begin{aligned} \hat{L}_i^0 &= y_0 + h \sum_{j=1}^{s} \beta_{ij} f(t_0 + c_j h, X_j), \quad 2 \leq i \leq s \\ \beta_{ij} &= \int_0^{1+rc_i} \hat{l}_j(t) dt, \quad 2 \leq i \leq s,\ 1 \leq j \leq s. \end{aligned} \tag{5.2.39}$$

Este algoritmo tiene la desventaja de que sus funciones de amplificación del error $\hat{R}_i^0(z)$ no están acotadas cuando $z$ tiende a infinito, como demostraremos en el próximo teorema. Presentaremos además una fórmula alternativa para computar dicho inicializador, la cual no necesita evaluar las derivadas $f(t_0 + c_j h, X_j),\ j = 2, \ldots, s$. Para simplificar la presentación de los resultados, daremos a continuación el siguiente lema que nos será muy útil en las demostraciones posteriores. En el desarrollo de esta subsección, entenderemos que $c_i$ $(1 \leq i \leq s)$ denotan los nodos correspondientes a la fórmula de cuadratura de Lobatto de $s$ nodos sobre el intervalo $[0, 1]$, y por tanto $c_1 = 0,\ c_s = 1$.

**Lema V.2.2** *El polinomio de interpolación $p(t)$ que satisface*

$$p(c_j) = y_j, \quad 1 \leq j \leq s, \quad p'(c_1) = y_1',$$

*viene dado por*

$$\left.\begin{aligned} p(t) &= \sum_{j=1}^{s} p_j(t) y_j + q_1(t) y_1', \\ p_1(t) &= \hat{l}_1(t)(1 - \hat{l}_1'(c_1)(t - c_1)), \\ p_j(t) &= \frac{t - c_1}{c_j - c_1} \hat{l}_j(t), \quad (j = 2, \ldots, s), \\ q_1(t) &= (t - c_1)\hat{l}_1(t). \end{aligned}\right\} \tag{5.2.40}$$

*Además se tiene que*

$$\sum_{j=1}^{s} p_j(t) = 1, \qquad q_1(t) + \sum_{j=1}^{s} p_j(t) c_j = t, \qquad \sum_{j=1}^{s} p_j(t) c_j^l = t^l, \quad 2 \le l \le s, \tag{5.2.41}$$

*y*

$$\begin{aligned} p_1(t) - \hat{l}_1(t) &= -\frac{\hat{l}_1'(c_1)}{\hat{\Pi}'(c_1)} \hat{\Pi}(t), \qquad q_1(t) = \frac{\hat{\Pi}(t)}{\hat{\Pi}'(c_1)}, \\ p_j(t) - \hat{l}_j(t) &= \frac{1}{(c_j - c_1)\hat{\Pi}'(c_j)} \hat{\Pi}(t), \quad 2 \le j \le s. \end{aligned} \tag{5.2.42}$$

*donde $\hat{l}_j(t)$ y $\hat{\Pi}(t)$ vienen dados en (5.1.11).*

<u>*Demostración.*</u> Es inmediato probar que el polinomio $p(t)$ dado en (5.2.40) satisface las condiciones de interpolación.

Para demostrar (5.2.41) usamos el hecho de que para todo polinomio $q(t)$ de grado $s$ como máximo se tiene que

$$p(t) = q(t) \qquad \text{si} \qquad y_j = q(c_j) \ \ (j = 1, \ldots, s), \quad y_1' = q'(c_1).$$

Tomando entonces $q(t) = t^l$ $(l = 0, 1, \ldots, s)$, se sigue claramente el resultado buscado.

Las identidades de (5.2.42) se deducen inmediatamente de (5.1.11). ∎

**Teorema V.2.4** *Sea $P(t)$ el polinomio de interpolación que verifica*

$$P(c_i) = X_i \ (1 \le i \le s), \quad P'(0) = h f(t_0, y_0).$$

*Entonces,*

*(a)* $\hat{L}_i^0 = P(1 + r c_i), \quad 2 \le i \le s.$

*(b)* $\hat{R}_i^0(\infty) = \infty, \quad 2 \le i \le s,$

*donde $\hat{R}_i^0(z)$ denota la función de amplificación del error del inicializador $\hat{L}_i^0$ dado en (5.2.39).*

<u>*Demostración.*</u> *(a)* Del lema anterior tenemos

$$P(t) = \sum_{j=1}^{s} p_j(t) X_j + h q_1(t) f(t_0, y_0). \tag{5.2.43}$$

Usando la expresión para las etapas $X_j$ dada en (4.1.3), teniendo en cuenta (5.2.36) y la primera ecuación de (5.2.41) se sigue que

$$P(t) = y_0 + h\left(\sum_{j=2}^{s} w_j p_j(t) + q_1(t)\right) f(t_0, y_0) + h\sum_{k=2}^{s}\left(\sum_{j=2}^{s} p_j(t) a_{jk}\right) f(t_0 + c_k h, X_k).$$

De aquí,

$$P(1+rc_i) = y_0 + h\sum_{k=1}^{s} \bar{\beta}_{ik} f(t_0 + c_k h, X_k), \quad 2 \le i \le s, \tag{5.2.44}$$

donde

$$\begin{aligned} \bar{\beta}_{i1} &= q_1(1+rc_i) + \textstyle\sum_{j=2}^{s} p_j(1+rc_i) w_j, \quad 2 \le i \le s \\ \bar{\beta}_{ik} &= \textstyle\sum_{j=2}^{s} p_j(1+rc_i) a_{jk}, \quad 2 \le i \le s, \quad 2 \le k \le s. \end{aligned} \tag{5.2.45}$$

Por otra parte, usando (5.2.41) y $C(s)$ (i.e. $Ac^{l-1} = c^l/l,\ 1 \le l \le s$), es inmediato probar que

$$\sum_{k=1}^{s} \bar{\beta}_{ik} c_k^{l-1} = (1+rc_i)^l/l, \quad 1 \le l \le s, \quad 2 \le i \le s. \tag{5.2.46}$$

Como las $\beta_{ik}$ (ver (5.2.39)) satisfacen también el sistema lineal (5.2.46), por la unicidad de solución del anterior sistema lineal se tiene que

$$\bar{\beta}_{ik} = \beta_{ik}, \quad 2 \le i \le s, \quad 1 \le k \le s,$$

con lo que se demuestra el apartado *(a)*.

*(b)* Del Teorema IV.3.4*(b)* se sigue que

$$\hat{R}_i^0(\infty) = \infty \iff \bar{\beta}_{i1} - \bar{\beta}_i^T \bar{A}^{-1} w \neq 0, \quad \bar{\beta}_i^T = (\bar{\beta}_{i2}, \ldots, \bar{\beta}_{is}), \quad 2 \le i \le s.$$

Si denotamos $\bar{A}^{-1} = (\gamma_{ij})_{i,j=2}^{s}$ obtenemos usando (5.2.45) que,

$$\begin{aligned} \bar{\beta}_{i1} - \bar{\beta}_i^T \bar{A}^{-1} w &= \bar{\beta}_{i1} - \textstyle\sum_{k=2}^{s}\left(\sum_{j=2}^{s} p_j(1+rc_i) a_{jk} \sum_{l=2}^{s} \gamma_{kl} w_l\right) \\ &= \bar{\beta}_{i1} - \sum_{j=2}^{s} p_j(1+rc_i) w_j = q_1(1+rc_i) \neq 0, \end{aligned}$$

lo que completa la demostración. ∎

**Nota V.2.1** *No es difícil probar que el anterior inicializador $\hat{L}_i^0$ también puede expresarse como*

$$\hat{L}_i^0 = P_m(1+rc_i), \qquad 2 \le i \le s,$$

*donde $P_m(t)$ es el polinomio de interpolación que satisface las condiciones*

$$P_m(c_j) = X_j, \quad j = 1, ..., s \qquad y \qquad P_m'(c_m) = hf(t_0 + c_m h, X_m),$$

*para cualquier $m = 2, 3, ..., s$ fijo.*

**Estabilización del inicializador de orden $s$**

Vamos a abordar ahora el problema de la estabilización del inicializador anterior manteniendo al mismo tiempo los mayores órdenes posibles (clásico y stiff). Con este objetivo definimos una familia de inicializadores que depende de $s-1$ parámetros $\theta_i,\ i=2,\ldots,s$, que se supondrán de tamaño moderado y que pueden depender de la razón de paso $r$ en algún caso,

$$\tilde{E}_i^0 = L_i^0 + \theta_i (I-\beta h J)^{-1}(\hat{L}_i^0 - L_i^0), \qquad 2 \le i \le s. \tag{5.2.47}$$

Naturalmente, $L_i^0$ y $\hat{L}_i^0$ hacen referencia a los inicializadores previamente definidos mediante (5.2.37) y (5.2.39) respectivamente.

Obsérvese que para computar esta nueva familia de algoritmos sólo se necesita resolver un sistema lineal adicional, pues de (5.2.37), (5.2.43), (5.2.42) y del Teorema V.2.4 se sigue sin mucha dificultad que

$$\hat{L}_i^0 - L_i^0 = \hat{\Pi}(1+rc_i)\tilde{W}, \quad 2 \le i \le s, \tag{5.2.48}$$

donde

$$\tilde{W} := (-1)^s (c_2 \ldots c_s)^{-1} (\hat{l}_1'(0) y_0 - h f(t_0, y_0)) + \sum_{j=2}^{s} (c_j \hat{\Pi}'(c_j))^{-1} X_j. \tag{5.2.49}$$

Por tanto tenemos la siguiente fórmula alternativa para la familia de inicializadores (5.2.47):

$$\tilde{E}_i^0 = L_i^0 + \theta_i \hat{\Pi}(1+rc_i)(I-\beta h J)^{-1}\tilde{W}, \quad 2 \le i \le s. \tag{5.2.50}$$

Además, como los métodos Lobatto IIIA son stiffly accurate, podemos computar casi sin ningún costo adicional la cantidad $\bar{h} f(t_1, y_1)$ para el siguiente paso por medio de

$$\bar{h} f(t_1, y_1) = r \left( \sum_{j=2}^{s} u_j X_j - (u^T e) y_0 - h (u^T w) f(t_0, y_0) \right), \quad u^T = e_{s-1}^T \bar{A}^{-1},$$

siendo $e_{s-1}^T = (0, ..., 1) \in I\!R^{s-1}$.

El siguiente teorema nos da las principales propiedades de convergencia y estabilidad de esta nueva familia de algoritmos de arranque.

**Teorema V.2.5** *Consideremos el Lobatto IIIA de $s$ etapas. Entonces, tomando cualquier constante $\beta > 0$ y $J = \partial f/\partial y\ (t_0, y_0)$, se tiene que:*

*(a) Toda la familia de inicializadores (5.2.47) tiene orden clásico $s-1$. Además, si elegimos $\theta_i = 1,\ (i = 2, ..., s)$, el algoritmo resultante alcanza orden clásico $s$.*

*(b) Dicha familia alcanza orden $s-1$ sobre el modelo de Prothero y Robinson, y sus funciones de amplificación $\tilde{S}_i^0(z)$ del error verifican*

$$\tilde{S}_i^0(\infty) = R_i^0(\infty) - \theta_i \beta^{-1} q_1 (1+rc_i), \quad i = 2, ..., s, \tag{5.2.51}$$

*donde $R_i^0(\infty)$ viene dado en (5.2.37)–(5.2.38).*
*Además, no existe ningún algoritmo de la familia con orden $s$ en todo el semiplano complejo negativo. No obstante, podemos hacer varias elecciones interesantes de los parámetros $\theta_i$, a saber,*

*(b.1)* $\theta_i = 1,\ 2 \le i \le s$. *En este caso el algoritmo resultante posee orden* $s$ *sobre cuadraturas, i.e., en el caso* $z = 0$, *y sus correspondientes funciones de amplificación del error están acotadas y verifican*

$$\tilde{S}_i^0(\infty) = R_i^0(\infty) - \beta^{-1} q_1 (1 + r c_i), \quad 2 \le i \le s.$$

*(b.2)* $\theta_i = \beta (q_1(1 + r c_i))^{-1} R_i^0(\infty),\ 2 \le i \le s$. *Sus funciones de amplificación satisfacen*

$$\tilde{S}_i^0(\infty) = 0, \quad 2 \le i \le s.$$

*(b.3)* $\theta_i = \beta (q_1(1 + r c_i))^{-1} (R_i^0(\infty) - \mathcal{R}_i(\infty)),\ 2 \le i \le s$, *donde*

$$\mathcal{R}_i(\infty) = (-1)^s e_{i-1}^T \bar{A}^{-1} w, \quad 2 \le i \le s,$$

*son las funciones de amplificación del error de las etapas internas* $Y_i$ *en el infinito (ver Teorema IV.3.4(b)). Entonces tenemos*

$$\tilde{S}_i^0(\infty) = \mathcal{R}_i(\infty), \quad 2 \le i \le s.$$

*(c) Todos los algoritmos de la familia poseen orden stiff* $s - 1$.

<u>Demostración.</u> *(a)* Se sigue inmediatamente de los Teoremas IV.2.1 y IV.2.2 que

$$Y_i - L_i^0 = \mathcal{O}(h^s), \qquad Y_i - \hat{L}_i^0 = \mathcal{O}(h^{s+1}). \tag{5.2.52}$$

De aquí se concluye inmediatamente el apartado *(a)* el tenerse que $I - (I - h\beta J)^{-1} = \mathcal{O}(h)$.

*(c)* Obsérvese que en el caso stiff suponemos que el PVI (4.1.1) es contractivo y que $y_0 = y(t_0)$. Luego, usando que para los métodos Lobatto IIIA existe una matriz diagonal definida positiva $D$ tal que $D\bar{A} + \bar{A}^T D$ es definida positiva (ver por ejemplo [37, pág. 222]), podemos probar (usando los mismos argumentos que en el Teorema IV.3.5 del capítulo IV) que también se tiene (5.2.52). Por tanto, ahora sólo se alcanza orden stiff $s-1$ al tenerse que $I-(I-h\beta J)^{-1} = \mathcal{O}(1)$.

*(b)* Cuando aplicamos los inicializadores al problema de Prothero y Robinson (4.3.1) se obtiene el siguiente desarrollo en potencias de $h$ (tomando $z = \lambda h$),

$$\tilde{E}_i^0(z) = \tilde{S}_i^0(z)(y_0 - \phi(t_0)) + \phi(t_0) + \sum_{j \ge 1} \frac{\phi^{(j)}(t_0)}{j!} \tilde{w}_{i,j}^0(z) h^j, \quad 2 \le i \le s,$$

con funciones de amplificación de error dadas por

$$\tilde{S}_i^0(z) = R_i^0(z) + \theta_i (1 - \beta z)^{-1} (\hat{R}_i^0(z) - R_i^0(z)), \quad 2 \le i \le s \tag{5.2.53}$$

y coeficientes

$$\tilde{w}_{i,j}^0(z) = v_{i,j}^0(z) + \theta_i (1 - \beta z)^{-1} (\hat{v}_{i,j}^0(z) - v_{i,j}^0(z)), \quad 2 \le i \le s,\ j \ge 1,$$

donde $R_i^0(z),\ \hat{R}_i^0(z),\ v_{i,j}^0(z)$ y $\hat{v}_{i,j}^0(z)$ denotan respectivamente las funciones de amplificación y los coeficientes de los desarrollos de los inicializadores dados en (5.2.37) y (5.2.39) cuando se aplican al modelo de Prothero y Robinson. El hecho de que $\tilde{E}_i^0$ tiene orden $s-1$ se sigue inmediatamente de que el inicializador (5.2.37) alcanza orden $s - 1$ sobre la ecuación de Prothero y Robinson y el (5.2.39) alcanza orden $s$ sobre la misma ecuación (ver el apartado *(b)* del Teorema IV.3.3).

Además, en el caso $z = 0$ no es difícil ver que $\hat{v}^0_{i,s}(0) \neq v^0_{i,s}(0)$ $(i = 2, \ldots, s)$, por lo que entonces el orden $s$ se alcanza si y sólo si $\theta_i = 1$, $\forall i = 2, ..., s$.

Para probar que no se puede alcanzar orden $s$ sobre la ecuación de Prothero y Robinson para ningún inicializador de la familia, vamos a proceder por contradicción. Supondremos que se alcanza orden $s$ para los puntos particulares $z = 0$ y $z = \infty$. Como acabamos de ver, el orden $s$ para $z = 0$ implica que $\theta_i = 1$, $2 \leq i \leq s$, y el orden $s$ para $z = \infty$ nos conduce a

$$\tilde{w}^0_{i,s}(\infty) = v^0_{i,s}(\infty) = \bar{\alpha}_i^T \bar{c}^s, \quad 2 \leq i \leq s$$

donde $\bar{\alpha}_i^T$ son los vectores dados en (5.2.37)-(5.2.38) (para más detalles, ver la sección IV.3 del capítulo anterior). Ahora bien, de (5.1.7) se sigue que

$$v_{i,s}(\infty) - v^0_{i,s}(\infty) = (1 + rc_i)^s - \bar{\alpha}_i^T \bar{c}^s, \qquad 2 \leq i \leq s.$$

Este hecho nos lleva a una contradicción ya que el segundo término de la igualdad anterior es un polinomio en $r$ de grado exactamente $s$ que no puede ser idénticamente nulo.

Finalmente, para demostrar la expresión (5.2.51), no es difícil probar (ver por ejemplo, (4.3.25) en la demostración del Teorema IV.3.4) que

$$\hat{R}^0_i(z) = 1 + z\bar{\beta}_{i1} + z\bar{\beta}_i^T (I - z\bar{A})^{-1}(zw + e), \quad 2 \leq i \leq s.$$

con los $\bar{\beta}_i^T = (\bar{\beta_{i2}}, \ldots, \bar{\beta_{is}})$ dados en (5.2.45). Ahora un cálculo directo nos da

$$\hat{R}^0_i(z) = 1 + zq_1(1 + rc_i) + z \sum_{j=2}^{s} p_j(1 + rc_i) \left( w_j + \sum_{k=2}^{s} a_{jk} \hat{\Delta}_k(z) \right),$$

donde

$$\hat{\Delta}_k(z) = e_{k-1}^T (I - z\bar{A})^{-1}(zw + e), \quad e_{k-1}^T = (0, ..., \overset{(k-1)}{1}, ..., 0) \in I\!R^{s-1}, \quad 2 \leq k \leq s.$$

Teniendo en cuenta que

$$w_j + \sum_{k=2} a_{jk} \hat{\Delta}_k(\infty) = w_j + \sum_{k=2} a_{jk}(-e_{k-1}^T \bar{A}^{-1} w) = 0,$$

llegamos a que

$$\lim_{z \to \infty} \frac{\hat{R}^0_i(z)}{1 - \beta z} = -\beta^{-1} q_1 (1 + rc_i). \tag{5.2.54}$$

Ahora llevando ésto a (5.2.53) obtenemos la expresión buscada en (5.2.51). De aquí se deduce inmediatamente los apartados *(b.1), (b.2)* y *(b.3)*, para las diferentes elecciones de los parámetros $\theta_i$. ∎

## V.3 Experimentos numéricos

De igual modo que en el capítulo IV, vamos a presentar varios experimentos numéricos realizados desde dos puntos de vista diferentes. En primer lugar, estudiamos el comportamiento local de los nuevos inicializadores propuestos, mediante experimentos en que sólo se avanzan dos pasos consecutivos de tamaños $h$ y $rh$. Con ellos pretendemos confirmar el orden teórico obtenido para los inicializadores, así como su factor de amplificación de error. En segundo lugar,

estudiamos la eficacia de los distintos inicializadores aquí considerados realizando integraciones de problemas en intervalos "largos".

Hemos usado aquí los mismos códigos empleados en la sección IV.4 del capítulo anterior, los cuales están basados en el método de colocación Radau IIA de tres etapas y orden cinco. A estos códigos sólo les hemos incluido los inicializadores más relevantes propuestos en el presente capítulo.

Respecto a los experimentos de tipo local, consideramos aquí de nuevo los problemas 1 y 2 de la sección IV.4 (así como los mismos tamaños de paso), esto es:

**Problema 1.-** Problema de Prothero y Robinson (4.3.1) con

$$\lambda = -10^6, \quad \phi(t) = e^{2t}, \quad z_0 = 1.$$

**Problema 2.-** Problema escalar no lineal del tipo propuesto en [58, pág. 202] que viene dado en (4.4.3), donde del mismo modo hemos tomado

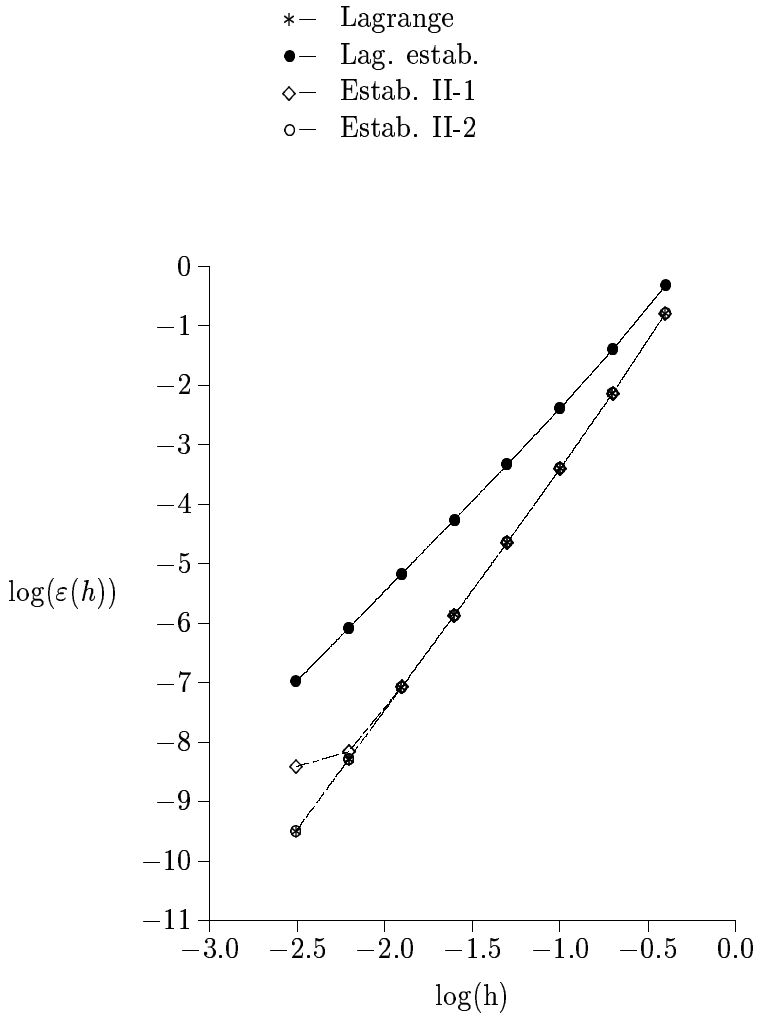
$$\lambda = -10^6, \quad \phi(t) = 1 + e^t, \quad z_0 = 2.$$

En este caso, representamos en las gráficas 5.3.1 a 5.3.4 los resultados obtenidos para los distintos inicializadores estabilizados, donde hemos enfrentado el logaritmo (decimal) del error dado por (4.4.1) con respecto al logaritmo de los tamaños de paso $h = h_k$ considerados. Los inicializadores que hemos tomado son los siguientes:

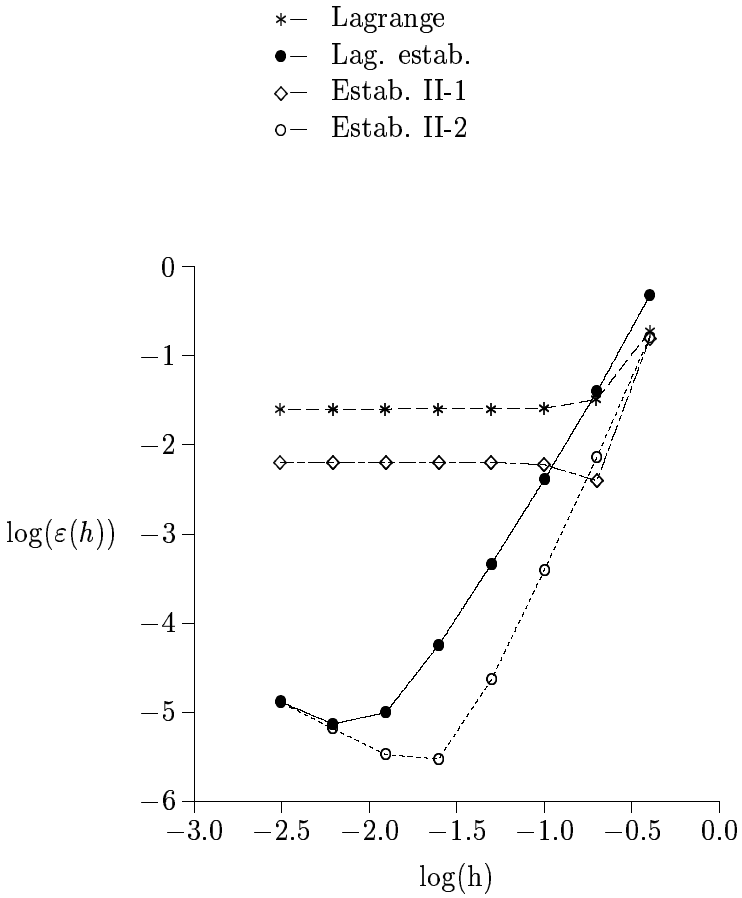
- **Lagrange (Lag):** interpolación de Lagrange de $y_0$ y $X_i$, $i = 1, 2, 3$, dado en (5.1.1).
- **Lagrange estabilizado (Lag. estab.):** algoritmo estabilizado propuesto en (5.2.3).
- **Estabilizado Tipo II-1 (Estab. II-1):** algoritmo estabilizado de la familia (5.2.12) con $\theta_i = 1$, $1 \leq i \leq s$.
- **Estabilizado Tipo II-2 (Estab. II-2):** algoritmo estabilizado de la familia (5.2.12) con $\theta_i$, $1 \leq i \leq s$ dados por (5.2.16).

Claramente, cuando integramos los problemas 1 y 2, partiendo del valor inicial exacto $y_0 = \phi(0)$, las pendientes de las rectas que aparecen en las gráficas 5.3.1 y 5.3.3 nos dan respectivamente (para cada problema) el orden de aproximación para cada uno de los inicializadores considerados. Así por ejemplo, podemos ver en ambas gráficas que el inicializador de Lagrange y los de Tipo II estabilizados alcanzan orden 3, es decir su error (4.4.1) es de orden $h^4$, lo que en las gráficas se traduce en rectas de pendiente 4 para ambos casos; de hecho, dan lugar prácticamente a las mismas rectas. Sin embargo, para el inicializador de Lagrange estabilizado la pendiente de las rectas correspondientes es 3, lo que se traduce en que se está comportando como un inicializador de orden 2, tal cual se predijo con anterioridad en la teoría.

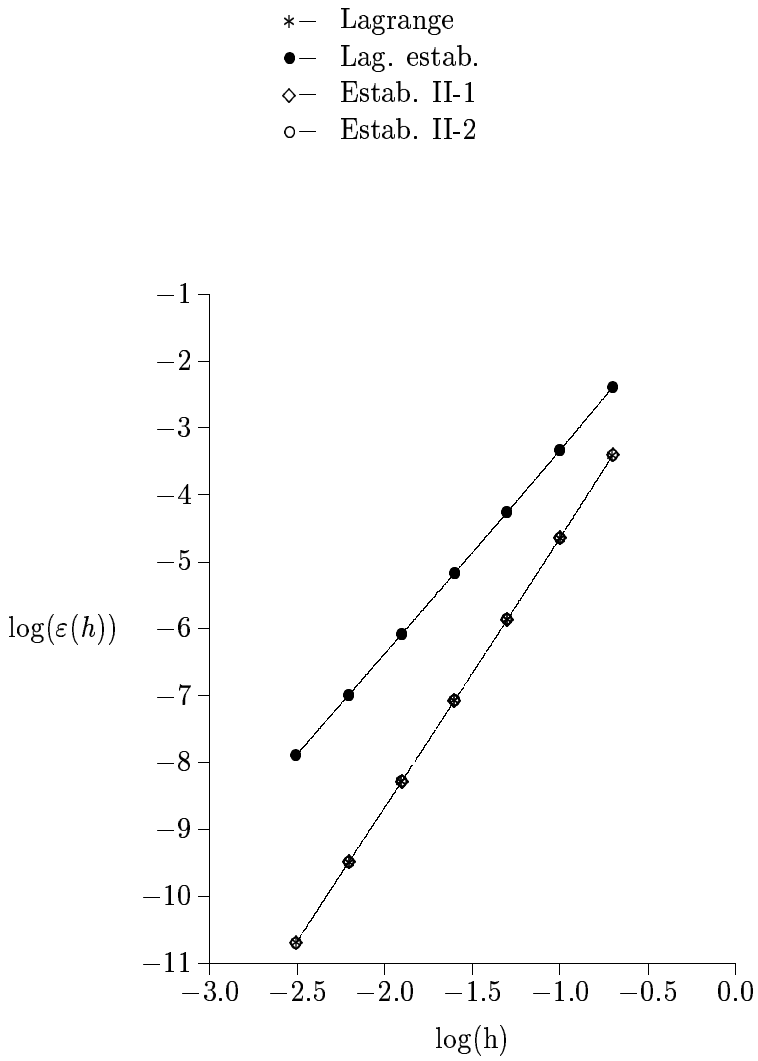
Por otra parte, al considerar el dato inicial $y_0$ perturbado, según se indica en (4.4.2), se observa en las gráficas 5.3.2 y 5.3.4 que el único inicializador, que se comporta prácticamente igual que sin la perturbación es el Lagrange estabilizado (**Lag. estab.**). Esto concuerda con el hecho de que este inicializador tiene amplificación cero para $z = \infty$, y por tanto, para $z = \lambda h$ grandes en módulo, se espera un comportamiento de propagación casi nula de los errores del dato inicial, manteniéndose así el orden de forma prácticamente independiente del error del dato inicial. El otro inicializador que posee amplificación cero para $z = \infty$ es el Tipo II estabilizado y que hemos denotado por **Estab. II-2**. En este caso se observa en la gráfica 5.3.2 que tiene



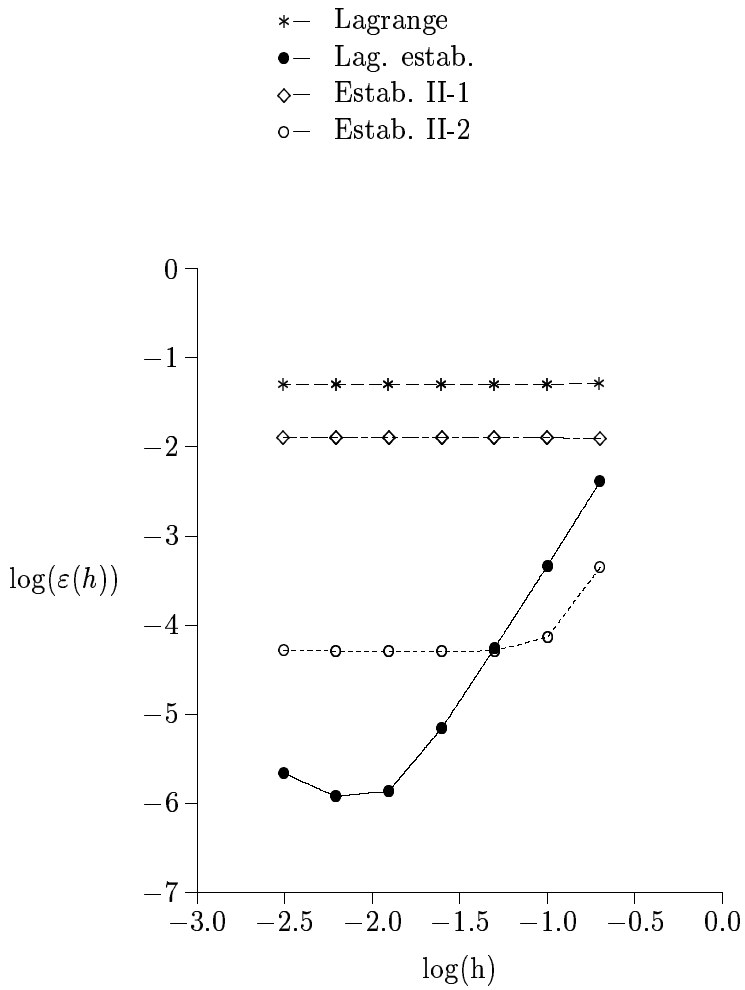

Gráfica 5.3.1: Problema de Prothero y Robinson ($y_0 = z_0$).




Gráfica 5.3.2: Problema de Prothero y Robinson ($y_0 = z_0 + \Delta$).




Gráfica 5.3.3: Problema no lineal de Spijker ($y_0 = z_0$).




Gráfica 5.3.4: Problema no lineal de Spijker ($y_0 = z_0 + \Delta$).

un comportamiento similar al Lagrange estabilizado en presencia de una perturbación en el dato inicial (de hecho coinciden sus funciones de amplificación de error), pero si comparamos las gráficas 5.3.3 y 5.3.4, las curvas correspondientes al caso perturbado y no perturbado se diferencian antes. Esto es debido a que para $h \doteq h^* = h_0/2^2$ (ver gráfica 5.3.4) los errores de amplificación y de truncamiento del inicializador se hacen del mismo tamaño. Por tanto, aunque $h$ disminuya ($h < h^*$), el error total no puede disminuir, pues el error de amplificación se mantiene prácticamente constante en el problema 2 (caso no lineal) y aumenta ligeramente en el problema 1.

Es interesante notar que la perturbación introducida afecta de un modo más sensible al inicializador **Estab. II-2** que al **Lag. estab.** Esto es razonable, ya que ambos inicializadores propagan las perturbaciones de igual modo en problemas lineales, pero no exactamente del mismo modo para problemas no lineales. Se observa en este caso que el **Lag. estab.** parece ser ligeramente más estable que el denotado por **Estab. II-2**.

Con respecto a los inicializadores que poseen función de amplificación acotada, podemos ver en las gráficas que, cuando introducimos una perturbación en el dato inicial, los errores se mantienen prácticamente invariables para los distintos tamaños de paso considerados. Esto se debe a que en estos casos los errores globales (errores de truncamiento más errores de amplificación) están dominados por los errores de amplificación que se mantienen prácticamente constantes para los tamaños de paso considerados.

Por otra parte, y a efectos de comparar los inicializadores arriba considerados con fines más prácticos, hemos usado en la integración de los problemas stiff, el código de paso variable previamente considerado en la sección IV.4, el cual está basado en la fórmula Runge–Kutta Radau IIA de tres etapas implementado con el esquema Single–Newton propuesto en [29] que estima el error local mediante extrapolación. Este código se ha preparado de modo que permita elegir al usuario entre cualquiera de los inicializadores (estabilizados o no) considerados en este capítulo y en el anterior.

Con este código hemos realizado numerosas integraciones para todos los problemas stiff propuestos en el paquete DETEST [23], así como para los también propuestos como tests en [37, Cap.IV.10] y [49]. Aquí presentaremos los resultados obtenidos para los cuatro problemas siguientes, los cuales ya han sido considerados en los capítulos anteriores de esta memoria:

**Problema 3.-** Problema de Van der Pol [37, pág.144] de dimensión 2 (con $h_0 = 10^{-8}$).
**Problema 4.-** E5 del paquete DETEST [37, pág.145] de dimensión 4 (con $h_0 = 10^{-2}$) .
**Problema 5.-** El Problema denotado por CUSP en [37, pág. 147]. Este problema proviene de un sistema no lineal en derivadas parciales dependiente del tiempo, siendo una combinación del modelo Zeeman's "cusp catastrophe" y del oscilador de Van der Pol (ver detalles en la referencia anterior). Las variables espaciales se discretizan mediante diferencias finitas y de este modo aparece un sistema diferencial de dimensión relativamente elevada (dimensión 96). Este problema lo hemos integrado en el intervalo $[0, 1.1]$, y hemos considerado paso inicial $h_0 = 10^{-8}$.
**Problema 6.-** Problema Ring Modulator [49] de dimensión 15 (con $h_0 = 10^{-8}$).

En las tablas 5.3.1, 5.3.2, 5.3.3 y 5.3.4 dadas a continuación, hemos representado para cada caso los mismos parámetros que en las tablas de la sección IV.4.

Con respecto al problema 3 (tabla 5.3.1) podemos decir que los cuatro inicializadores considerados hacen que el código integre el problema con errores globales similares para cada tolerancia. Asimismo se observa que obtenemos números muy parecidos de pasos, de factorizaciones LU, de evaluación de función derivada y de iteraciones por paso. Podemos observar también que el inicializador denotado por **Estab. II-1**, es ligeramente más eficiente que los otros, ya que suele

| RTOL | INI | EG | NPA | NR-ES | NR-SN | NLU | NSOL | NFN | NITER |
|---|---|---|---|---|---|---|---|---|---|
| $10^{-2}$ | Lag. | .1793E-04 | 224 | 0 | 39 | 237 | 3894 | 3940 | 3.13 |
| | Lag. estab. | .2863E-04 | 232 | 0 | 39 | 245 | 4123 | 3880 | 2.69 |
| | Estab. II-1 | .2393E-04 | 222 | 0 | 39 | 238 | 4040 | 3787 | 3.00 |
| | Estab. II-2 | .6974E-04 | 220 | 0 | 39 | 233 | 4055 | 3811 | 3.05 |
| $10^{-3}$ | Lag. | .6677E-05 | 246 | 0 | 38 | 259 | 5007 | 5086 | 3.66 |
| | Lag. estab. | .1011E-04 | 248 | 0 | 39 | 264 | 5112 | 4867 | 3.28 |
| | Estab. II-1 | .8822E-05 | 244 | 0 | 39 | 258 | 5065 | 4831 | 3.45 |
| | Estab. II-2 | .1904E-04 | 246 | 0 | 39 | 260 | 5227 | 4990 | 3.48 |
| $10^{-4}$ | Lag. | .3878E-06 | 294 | 10 | 34 | 311 | 6645 | 6733 | 3.97 |
| | Lag. estab. | .1331E-04 | 294 | 6 | 35 | 307 | 6839 | 6574 | 3.79 |
| | Estab. II-1 | .8799E-06 | 296 | 10 | 33 | 308 | 6617 | 6334 | 3.70 |
| | Estab. II-2 | .5852E-05 | 288 | 8 | 35 | 305 | 6799 | 6526 | 3.93 |
| $10^{-5}$ | Lag. | .3057E-05 | 360 | 30 | 26 | 379 | 8331 | 8404 | 4.27 |
| | Lag. estab. | .2592E-05 | 368 | 30 | 28 | 391 | 8918 | 8554 | 4.17 |
| | Estab. II-1 | .3157E-05 | 360 | 28 | 28 | 380 | 8369 | 8011 | 3.91 |
| | Estab. II-2 | .8874E-06 | 368 | 40 | 24 | 395 | 8999 | 8620 | 4.13 |
| $10^{-6}$ | Lag. | .4416E-06 | 468 | 46 | 20 | 495 | 10644 | 10702 | 4.35 |
| | Lag. estab. | .3527E-06 | 470 | 42 | 22 | 496 | 11310 | 10825 | 4.36 |
| | Estab. II-1 | .4016E-06 | 470 | 48 | 20 | 499 | 10757 | 10264 | 4.03 |
| | Estab. II-2 | .9756E-06 | 464 | 42 | 23 | 489 | 11046 | 10564 | 4.29 |
| $10^{-7}$ | Lag. | .7882E-07 | 602 | 30 | 18 | 621 | 13452 | 13504 | 4.60 |
| | Lag. estab. | .1429E-06 | 608 | 28 | 21 | 627 | 14387 | 13777 | 4.66 |
| | Estab. II-1 | .8425E-07 | 600 | 30 | 18 | 618 | 13210 | 12601 | 4.14 |
| | Estab. II-2 | .7298E-07 | 602 | 30 | 18 | 620 | 13947 | 13336 | 4.51 |
| $10^{-8}$ | Lag. | .2962E-07 | 838 | 28 | 15 | 852 | 18369 | 18412 | 4.68 |
| | Lag. estab. | .2721E-07 | 844 | 24 | 17 | 858 | 19684 | 18835 | 4.82 |
| | Estab. II-1 | .2486E-07 | 840 | 26 | 15 | 853 | 17931 | 17080 | 4.17 |
| | Estab. II-2 | .2767E-07 | 838 | 28 | 14 | 852 | 19030 | 18178 | 4.61 |
| $10^{-9}$ | Lag. | .4875E-08 | 1238 | 20 | 9 | 1248 | 26250 | 26275 | 4.64 |
| | Lag. estab. | .4095E-08 | 1246 | 16 | 12 | 1254 | 28905 | 27655 | 4.96 |
| | Estab. II-1 | .4877E-08 | 1238 | 20 | 9 | 1248 | 25661 | 24412 | 4.15 |
| | Estab. II-2 | .4862E-08 | 1238 | 20 | 9 | 1248 | 27194 | 25945 | 4.56 |

Tabla 5.3.1: Problema Van der Pol (Single–Newton con extrapolación)

| RTOL | INI | EG | NPA | NR-ES | NR-SN | NLU | NSOL | NFN | NITER |
|---|---|---|---|---|---|---|---|---|---|
| $10^{-1}$ | Lag. | *** | ** | ** | ** | ** | ** | ** | ** |
| | Lag. estab. | .3192E-08 | 32 | 0 | 0 | 32 | 174 | 142 | 1.00 |
| | Estab. II-1 | *** | ** | ** | ** | ** | ** | ** | ** |
| | Estab. II-2 | *** | ** | ** | ** | ** | ** | ** | ** |
| $10^{-2}$ | Lag. | *** | ** | ** | ** | ** | ** | ** | ** |
| | Lag. estab. | .3192E-08 | 32 | 0 | 0 | 32 | 174 | 142 | 1.00 |
| | Estab. II-1 | *** | ** | ** | ** | ** | ** | ** | ** |
| | Estab. II-2 | *** | ** | ** | ** | ** | ** | ** | ** |
| $10^{-3}$ | Lag. | .3388E-05 | 56 | 0 | 4 | 58 | 348 | 346 | 1.38 |
| | Lag. estab. | .1312E-08 | 32 | 0 | 0 | 32 | 183 | 151 | 1.09 |
| | Estab. II-1 | .3699E-05 | 38 | 0 | 1 | 39 | 265 | 226 | 1.42 |
| | Estab. II-2 | .2566E-05 | 38 | 0 | 1 | 39 | 268 | 229 | 1.45 |
| $10^{-4}$ | Lag. | .1944E-06 | 32 | 0 | 0 | 32 | 222 | 220 | 1.69 |
| | Lag. estab. | .2585E-09 | 32 | 0 | 0 | 32 | 204 | 172 | 1.22 |
| | Estab. II-1 | .1121E-07 | 32 | 0 | 0 | 32 | 243 | 211 | 1.63 |
| | Estab. II-2 | .3014E-07 | 32 | 0 | 0 | 32 | 240 | 208 | 1.59 |
| $10^{-5}$ | Lag. | .6809E-08 | 32 | 0 | 0 | 32 | 243 | 241 | 1.91 |
| | Lag. estab. | .2601E-10 | 32 | 0 | 0 | 32 | 237 | 205 | 1.53 |
| | Estab. II-1 | .2023E-08 | 32 | 0 | 0 | 32 | 264 | 232 | 1.81 |
| | Estab. II-2 | .1542E-08 | 32 | 0 | 0 | 32 | 267 | 235 | 1.84 |
| $10^{-6}$ | Lag. | .8633E-09 | 34 | 0 | 0 | 34 | 291 | 289 | 2.06 |
| | Lag. estab. | .1904E-10 | 34 | 0 | 0 | 34 | 290 | 256 | 1.74 |
| | Estab. II-1 | .3409E-10 | 34 | 0 | 0 | 34 | 305 | 271 | 1.88 |
| | Estab. II-2 | .1355E-08 | 34 | 0 | 0 | 34 | 320 | 286 | 2.03 |
| $10^{-7}$ | Lag. | .3444E-09 | 36 | 0 | 0 | 36 | 348 | 346 | 2.25 |
| | Lag. estab. | .1102E-10 | 36 | 0 | 0 | 36 | 358 | 322 | 2.06 |
| | Estab. II-1 | .4328E-10 | 36 | 0 | 0 | 36 | 367 | 331 | 2.14 |
| | Estab. II-2 | .5850E-10 | 36 | 0 | 0 | 36 | 385 | 349 | 2.31 |
| $10^{-8}$ | Lag. | .1004E-11 | 40 | 0 | 0 | 40 | 465 | 463 | 2.70 |
| | Lag. estab. | .4649E-11 | 40 | 0 | 0 | 40 | 470 | 430 | 2.42 |
| | Estab. II-1 | .1902E-11 | 40 | 0 | 0 | 40 | 479 | 439 | 2.50 |
| | Estab. II-2 | .3361E-12 | 40 | 0 | 0 | 40 | 488 | 448 | 2.58 |
| $10^{-9}$ | Lag. | .8734E-12 | 46 | 0 | 0 | 46 | 585 | 583 | 2.93 |
| | Lag. estab. | .7169E-12 | 46 | 0 | 0 | 46 | 602 | 556 | 2.74 |
| | Estab. II-1 | .4297E-12 | 46 | 0 | 0 | 46 | 602 | 556 | 2.74 |
| | Estab. II-2 | .1334E-11 | 46 | 0 | 0 | 46 | 623 | 577 | 2.87 |

Tabla 5.3.2: Problema E5–stiff (Single–Newton con extrapolación)

| RTOL | INI | EG | NPA | NR-ES | NR-SN | NLU | NSOL | NFN | NITER |
|---|---|---|---|---|---|---|---|---|---|
| $10^{-2}$ | Lag. | .4681E-05 | 230 | 0 | 53 | 260 | 2826 | 2824 | 2.20 |
| | Lag. estab. | .7464E-05 | 232 | 0 | 53 | 258 | 2798 | 2491 | 1.64 |
| | Estab. II-1 | .6821E-05 | 232 | 0 | 53 | 255 | 3073 | 2761 | 2.13 |
| | Estab. II-2 | .8115E-05 | 232 | 0 | 53 | 261 | 3105 | 2794 | 2.13 |
| $10^{-3}$ | Lag. | .2040E-05 | 236 | 0 | 53 | 259 | 3393 | 3412 | 2.52 |
| | Lag. estab. | .2104E-05 | 238 | 0 | 53 | 267 | 3698 | 3427 | 2.19 |
| | Estab. II-1 | .1861E-05 | 236 | 0 | 53 | 263 | 3826 | 3544 | 2.53 |
| | Estab. II-2 | .2033E-05 | 232 | 0 | 53 | 263 | 3790 | 3511 | 2.54 |
| $10^{-4}$ | Lag. | .1332E-04 | 256 | 0 | 53 | 282 | 4578 | 4630 | 3.02 |
| | Lag. estab. | .2707E-05 | 252 | 0 | 53 | 279 | 4649 | 4381 | 2.79 |
| | Estab. II-1 | .1331E-04 | 260 | 2 | 51 | 279 | 4837 | 4558 | 2.86 |
| | Estab. II-2 | .1334E-04 | 252 | 0 | 53 | 277 | 4757 | 4477 | 2.95 |
| $10^{-5}$ | Lag. | .4535E-06 | 276 | 4 | 51 | 297 | 5520 | 5587 | 3.54 |
| | Lag. estab. | .4508E-06 | 282 | 4 | 51 | 305 | 6020 | 5737 | 3.29 |
| | Estab. II-1 | .4493E-06 | 278 | 6 | 50 | 301 | 5846 | 5551 | 3.39 |
| | Estab. II-2 | .4532E-06 | 272 | 2 | 52 | 292 | 5650 | 5356 | 3.46 |
| $10^{-6}$ | Lag. | .5265E-07 | 320 | 6 | 50 | 336 | 7206 | 7324 | 3.93 |
| | Lag. estab. | .1617E-07 | 330 | 18 | 45 | 344 | 7669 | 7327 | 3.73 |
| | Estab. II-1 | .1621E-07 | 320 | 10 | 49 | 340 | 7319 | 7003 | 3.75 |
| | Estab. II-2 | .2087E-07 | 330 | 12 | 48 | 348 | 7648 | 7327 | 3.74 |
| $10^{-7}$ | Lag. | .2222E-07 | 386 | 20 | 45 | 408 | 8820 | 8915 | 4.21 |
| | Lag. estab. | .2193E-07 | 396 | 26 | 42 | 417 | 9526 | 9131 | 4.16 |
| | Estab. II-1 | .2236E-07 | 390 | 26 | 41 | 413 | 9172 | 8789 | 3.97 |
| | Estab. II-2 | .2195E-07 | 392 | 26 | 42 | 413 | 9382 | 8993 | 4.05 |
| $10^{-8}$ | Lag. | .1397E-08 | 494 | 48 | 37 | 525 | 11784 | 11879 | 4.39 |
| | Lag. estab. | .1200E-08 | 506 | 54 | 35 | 541 | 12792 | 12266 | 4.45 |
| | Estab. II-1 | .1033E-08 | 498 | 50 | 36 | 531 | 11892 | 11378 | 4.07 |
| | Estab. II-2 | .1040E-08 | 502 | 56 | 31 | 535 | 12260 | 11729 | 4.24 |
| $10^{-9}$ | Lag. | .1907E-09 | 638 | 58 | 31 | 677 | 14982 | 15057 | 4.61 |
| | Lag. estab. | .1342E-09 | 654 | 60 | 36 | 697 | 16350 | 15675 | 4.66 |
| | Estab. II-1 | .1346E-09 | 636 | 58 | 30 | 674 | 14827 | 14163 | 4.19 |
| | Estab. II-2 | .2233E-09 | 640 | 62 | 30 | 681 | 15593 | 14922 | 4.49 |

Tabla 5.3.3: Problema CUSP (Single–Newton con extrapolación)

| RTOL | INI | EG | NPA | NR-ES | NR-SN | NLU | NSOL | NFN | NITER |
|---|---|---|---|---|---|---|---|---|---|
| $10^{-2}$ | Lag. | .6551E-01 | 3700 | 0 | 600 | 3991 | 78321 | 79426 | 4.11 |
| | Lag. estab. | .6522E-01 | 2170 | 0 | 365 | 2278 | 51822 | 49732 | 4.08 |
| | Estab. II-1 | .6536E-01 | 2086 | 0 | 350 | 2189 | 53734 | 51814 | 4.54 |
| | Estab. II-2 | .6514E-01 | 2196 | 0 | 366 | 2316 | 55839 | 53740 | 4.55 |
| $10^{-3}$ | Lag. | .6514E-01 | 5538 | 22 | 302 | 5663 | 148608 | 149257 | 5.40 |
| | Lag. estab. | .6514E-01 | 5304 | 34 | 271 | 5393 | 148238 | 143080 | 5.33 |
| | Estab. II-1 | .6516E-01 | 5034 | 30 | 238 | 5134 | 138249 | 133429 | 5.11 |
| | Estab. II-2 | .6515E-01 | 5222 | 40 | 256 | 5332 | 145851 | 140773 | 5.29 |
| $10^{-4}$ | Lag. | .2890E-01 | 22966 | 1200 | 413 | 23774 | 615045 | 615427 | 5.20 |
| | Lag. estab. | .3558E-01 | 22080 | 1242 | 142 | 22728 | 632834 | 609622 | 5.34 |
| | Estab. II-1 | .3391E-01 | 21902 | 1246 | 121 | 22560 | 582861 | 559834 | 4.67 |
| | Estab. II-2 | .3449E-01 | 22032 | 1260 | 138 | 22699 | 630610 | 607417 | 5.32 |
| $10^{-5}$ | Lag. | .2950E-02 | 40942 | 2258 | 654 | 42420 | 968532 | 968767 | 4.66 |
| | Lag. estab. | .3143E-02 | 39428 | 2510 | 92 | 40700 | 1009318 | 967477 | 4.82 |
| | Estab. II-1 | .3151E-02 | 39424 | 2492 | 87 | 40694 | 898916 | 857080 | 3.94 |
| | Estab. II-2 | .3135E-02 | 39646 | 2718 | 83 | 41022 | 1012237 | 969955 | 4.79 |
| $10^{-6}$ | Lag. | .3930E-03 | 63902 | 5268 | 1122 | 67373 | 1388517 | 1388662 | 4.22 |
| | Lag. estab. | .3991E-03 | 63654 | 7402 | 71 | 67366 | 1504080 | 1433098 | 4.30 |
| | Estab. II-1 | .3961E-03 | 63374 | 7236 | 53 | 66995 | 1351274 | 1280701 | 3.63 |
| | Estab. II-2 | .3963E-03 | 63458 | 7462 | 51 | 67194 | 1500455 | 1429576 | 4.28 |
| $10^{-7}$ | Lag. | .4948E-04 | 93578 | 8092 | 1273 | 98845 | 1968534 | 1968594 | 4.09 |
| | Lag. estab. | .4859E-04 | 93364 | 10246 | 31 | 98493 | 2111732 | 2008155 | 4.13 |
| | Estab. II-1 | .4885E-04 | 93514 | 10324 | 33 | 98682 | 1871762 | 1767933 | 3.36 |
| | Estab. II-2 | .4905E-04 | 93472 | 10330 | 23 | 98641 | 2106752 | 2002977 | 4.10 |

Tabla 5.3.4: Problema Ring Modulator (Single–Newton con extrapolación)

conllevar un número menor de evaluaciones de función derivada (NFN), lo que es consecuencia de un promedio de iteraciones por paso ligeramente más bajo. Esto se debe sin duda, al hecho de que además de ser un inicializador "estabilizado", posee un orden clásico más elevado que los otros. No debemos olvidar que en el problema 3 (Van der Pol) el 80% aproximadamente de los pasos se dan en zonas de variación rápida de la curva integral ("transient zones") y por tanto ahí el problema es no-stiff. En la gráfica 5.3.5, se puede apreciar también el comportamiento de los distintos inicializadores sobre este problema, y se puede constatar nuevamente que el inicializador **Estab. II-1** es ligeramente más eficiente que los otros.

Para el problema 4 (E5) podemos observar en la tabla 5.3.2, que para las tolerancias más altas, RTOL= $10^{-1}$, $10^{-2}$, el único inicializador que permitió concluir la integración fue el denotado por **Lag. estab.**. De nuevo se hacen patentes en este problema las mejores propiedades de estabilidad (para problemas lineales y no lineales) de la función de amplificación de error del inicializador **Lag. estab.**. Estas buenas propiedades de estabilidad fueron ya observadas en los experimentos de tipo local realizados previamente.

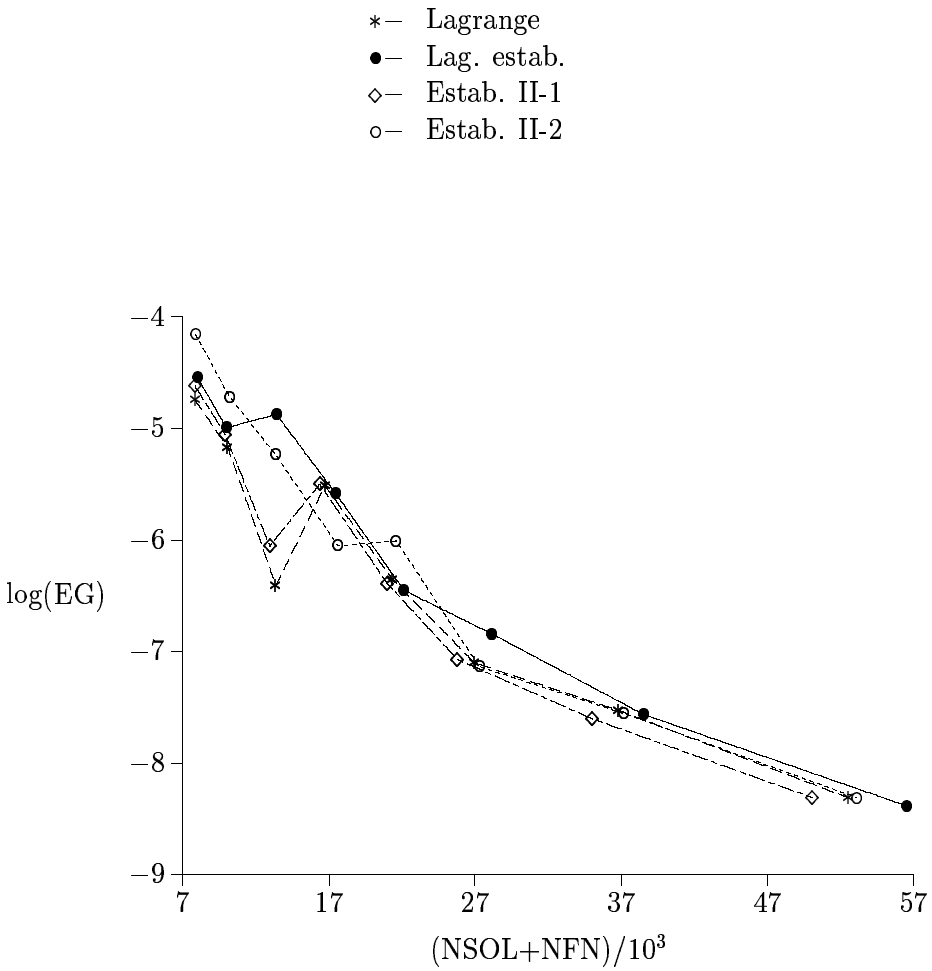
También podemos observar que para RTOL= $10^{-3}$, el inicializador **Lag. estab.** es el mejor, y que los denotados por **Estab. II-1** y **Estab. II-2** funcionan mejor que el **Lag.**. Esto último se puede justificar en virtud de que los inicializadores de Tipo II estabilizados poseen funciones de amplificación de error más próximas a cero que las del **Lag.** (veánse las gráficas 5.2.3, 5.2.4, 5.2.5 y 5.2.6). También resulta curioso el hecho de que el inicializador **Estab. II-2**, el cual posee una función de amplificación de error idéntica al **Lag. estab.** y un orden mayor que este último para el modelo de Prothero y Robinson, se comporte peor que **Lag. estab.** para todos los valores de RTOL, y especialmente para RTOL= $10^{-1}$, $10^{-2}$, $10^{-3}$. Creemos que esto se puede justificar, teniendo en cuenta que las funciones de amplificación de error son idénticas para ambos inicializadores en problemas lineales, pero no lo son para problemas no lineales. Este hecho fue constatado previamente en los experimentos de tipo local realizados sobre el problema 2 (ver gráficas 5.3.3 y 5.3.4), y se puso de manifiesto allí que el inicializador denotado por **Lag. estab.** era menos sensible a perturbaciones en el dato inicial que el denotado por **Estab. II-2**.

Se observa también en la tabla 5.3.2, que cuando las tolerancias disminuyen (RTOL$\leq 10^{-5}$), todos los inicializadores hacen que el rendimiento del código sea similar. Este hecho se constata de forma más clara en la gráfica 5.3.6.

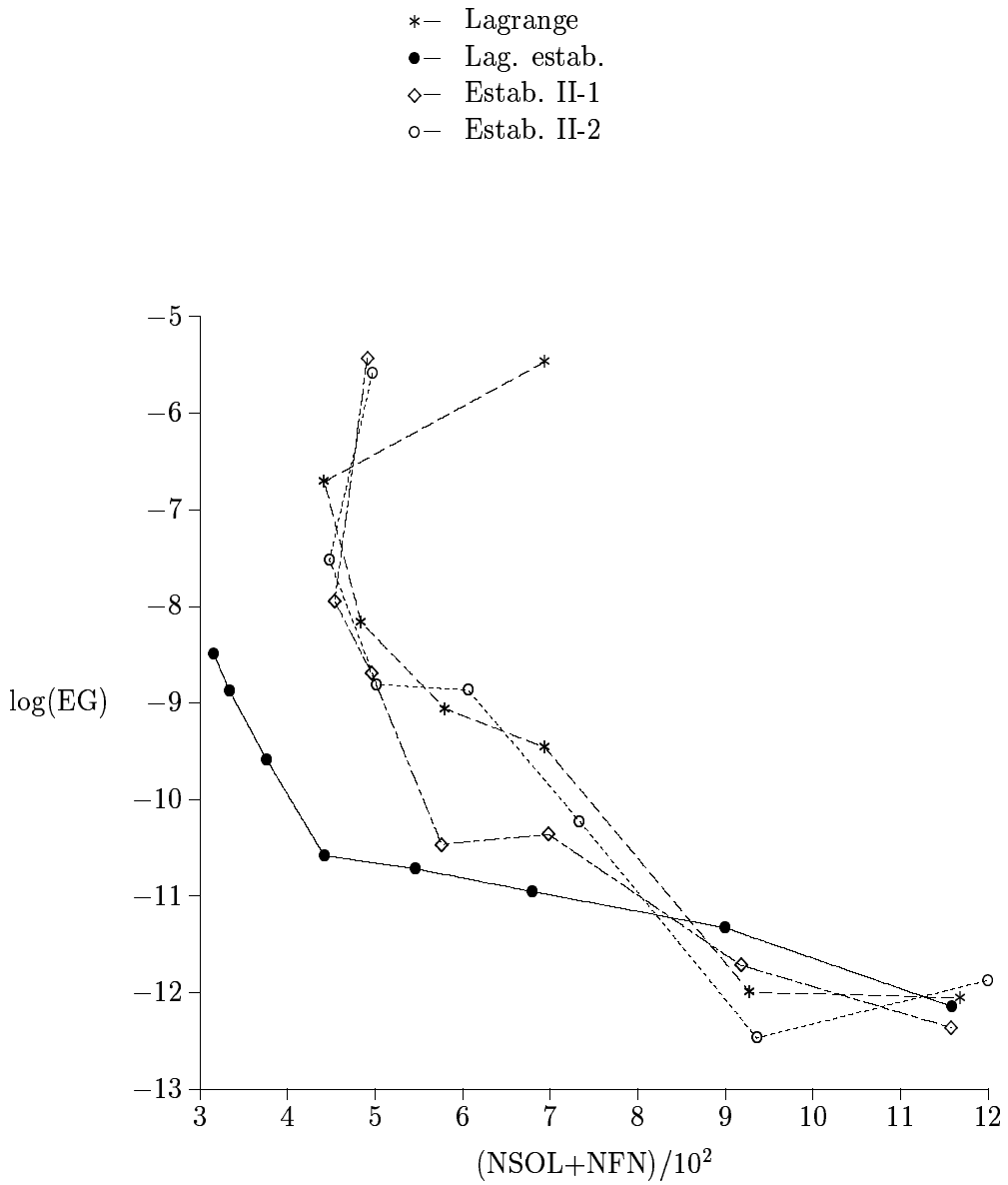
Con respecto al problema 5 (denotado por CUSP) hemos de decir que posee dos zonas de variación rápida (para las curvas integrales próximas a la solución exacta) dentro del intervalo de integración $[0, 1.1]$, las cuales están localizadas en los subintervalos de $t$, $[0.84, 0.85]$ y $[1.06, 1.07]$. En estas zonas de variación rápida, el código toma el 60% de los pasos, y los tamaños de los pasos son bastante reducidos, con lo cual no es de extrañar que los inicializadores de más alto orden den lugar a integraciones más eficientes, especialmente cuando las tolerancias son disminuidas.

En tolerancias bajas, se constata en la tabla 5.3.3 y en la gráfica 5.3.7 un mejor rendimiento para el inicializador **Estab. II-1**, sobre todo para RTOL$\leq 10^{-6}$. Por otra parte, para las tolerancias mayores, RTOL= $10^{-2}$, $10^{-3}$, $10^{-4}$, el inicializador que resultó más apropiado es el **Lag. estab.**, pues conlleva un número menor de iteraciones por paso, lo que implica un número inferior de funciones derivada (NFN) y de sistemas lineales (NSOL). El inicializador denotado por **Lag.** es quizás el más apropiado para las tolerancias medio–bajas, RTOL= $10^{-5}, 10^{-6}, 10^{-7}$.

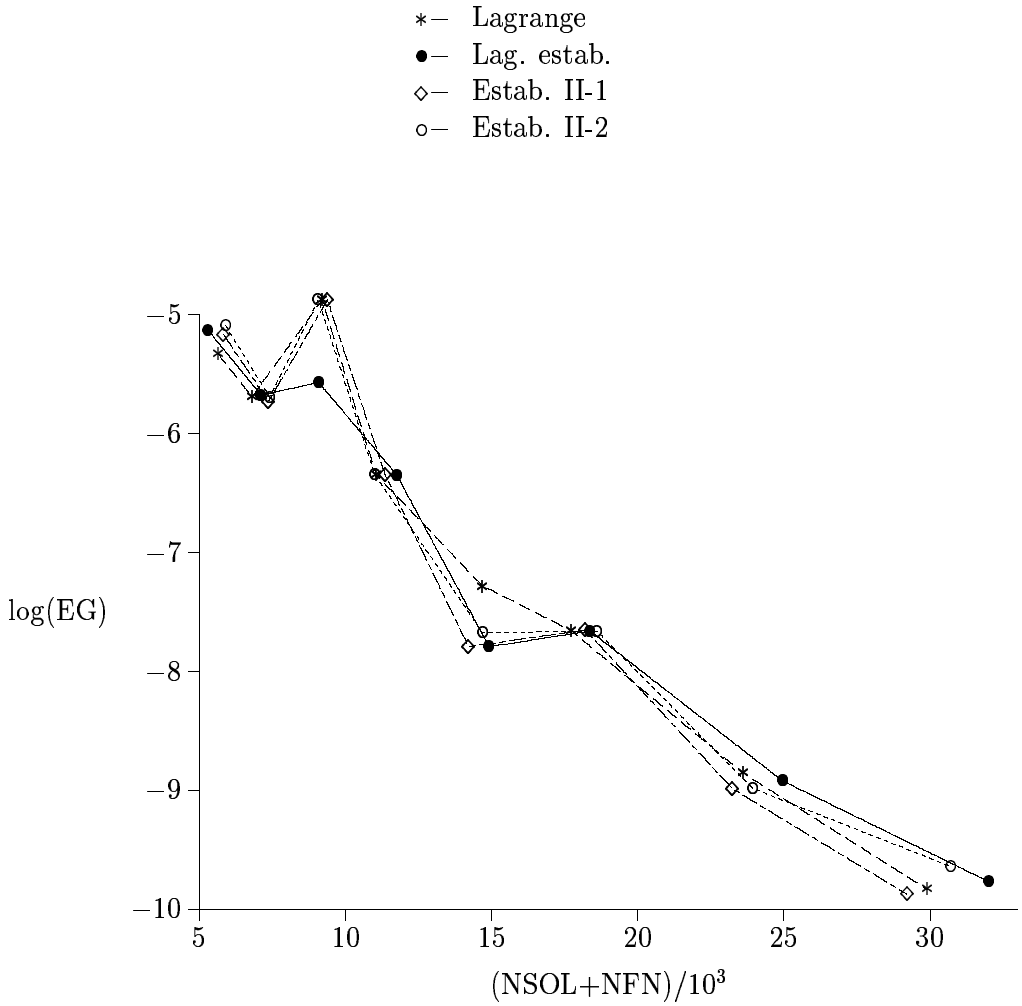
Para el problema 6 (Ring Modulator) podemos apreciar en la tabla 5.3.4, que todos los inicializadores hacen que el código complete la integración del problema satisfactoriamente, con errores globales (EG) similares para todas las tolerancias consideradas. Se constata también que el inicializador más eficiente en este caso, es el denotado por **Estab. II-1**, pues involucra un



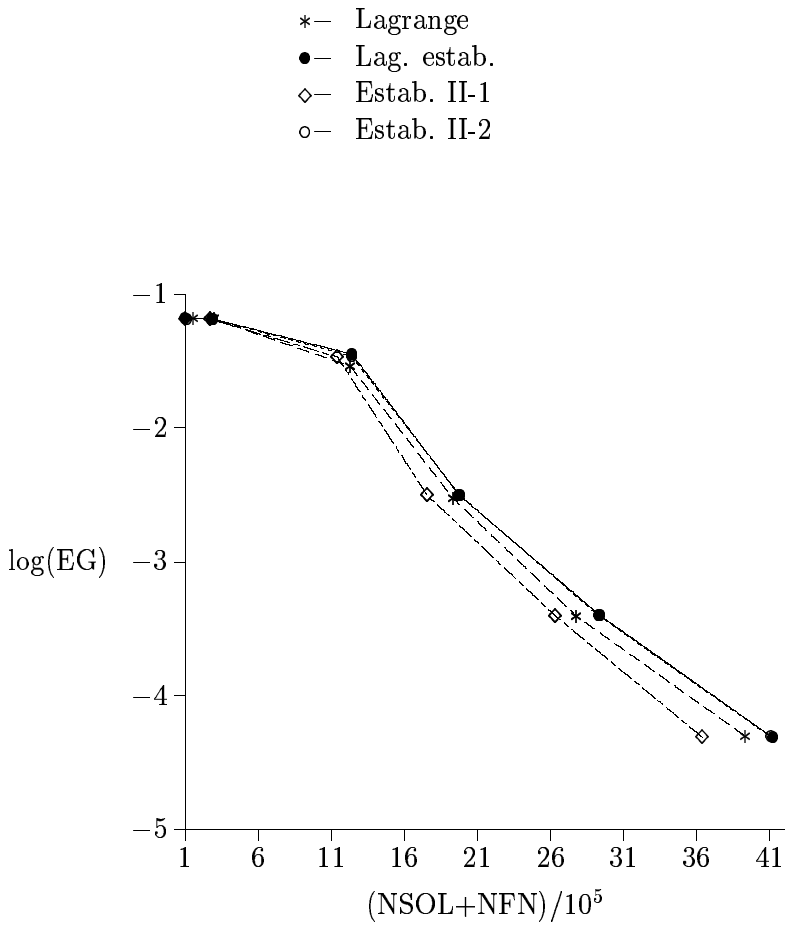

Gráfica 5.3.5: Problema de Van der Pol.




Gráfica 5.3.6: Problema E5–stiff.




Gráfica 5.3.7: Problema CUSP.




Gráfica 5.3.8: Problema Ring Modulator.

número ligeramente menor de pasos, lo que conlleva un número menor de factorizaciones LU, de soluciones de sistemas lineales y de evaluaciones de función derivada.

Se aprecia también en la tabla 5.3.4, que el promedio de iteraciones por paso para cada inicializador aparece en orden inverso al orden clásico del inicializador, lo cual es de esperar (al menos cuando las tolerancias son bajas) porque la integración de este problema requiere el uso de tamaños de paso muy reducidos. Este mismo hecho fue constatado en los experimentos numéricos realizados en el capítulo anterior para este mismo problema.

Como conclusiones generales de la investigación desarrollada en este capítulo podemos decir que:

1. Los experimentos numéricos avalan la teoría aquí desarrollada, la cual se fundamenta en estabilizar los inicializadores en base a modelos stiff lineales. Asimismo, se han estabilizado los inicializadores de más alto orden considerados previamente en el capítulo anterior.

2. La estabilización de los inicializadores cuyas funciones de amplificación poseen propiedades "pobres" de estabilidad, parece ser una alternativa viable y recomendable a efectos de conseguir inicializadores más fiables para los procesos iterativos empleados en la resolución de las ecuaciones de etapa de los métodos RK implícitos. Además, la estabilización de los inicializadores, la cual está basada en modelos lineales, parece ser relevante también para modelos no lineales.

3. Creemos que entre todos los inicializadores presentados en este capítulo, los más eficientes cuando se usa como método Runge–Kutta el Radau IIA de tres etapas son los dados por:

   (a) **Lag. estab.** (fórmula (5.2.3)), es el más robusto y el más recomendable para tolerancias altas y medio–altas.

   (b) **Lag.** (fórmula (5.1.1)), se recomienda para tolerancias medias.

   (c) **Estab. II-1** (fórmula (5.2.14) con $\theta_i = 1,\ 1 \leq i \leq 3$), se recomienda para tolerancias medias y bajas.

4. Para aprovechar mejor las potenciales ventajas de cada inicializador, creemos que un código de paso variable debería programarse de modo que tome el inicializador más conveniente en cada paso de la integración, antes que inclinarse por un inicializador determinado. Naturalmente, esto requeriría un criterio de selección que habría que estudiar con cierta profundidad. Este criterio de selección habría de actuar de modo que no supusiera un costo computacional excesivo con respecto al proceso iterativo empleado. Creemos que este último objetivo requiere una experimentación exhaustiva, la cual rebasa el marco de lo pretendido en este memoria. Creemos que ésto constituye un problema abierto para futuras investigaciones.

Capítulo **VI**

# Conclusiones e investigación futura

.

# Capítulo VI
# Conclusiones e investigación futura

En base a la investigación y experimentación desarrollada en esta memoria, podemos concluir lo siguiente:

1. La iteración de tipo Single–Newton para resolver las ecuaciones de etapa de los métodos Runge–Kutta de alto orden tales como Gauss, Radau IIA y Lobatto IIIA, es bastante robusta y eficiente para la integración de problemas stiff. Además parece ser una mejor alternativa que la iteración de Newton Simplificada para problemas de dimensión media o elevada. Todo esto ha quedado reflejado en la teoría y experimentación numérica realizada en los capítulos II y III de esta memoria.

2. Se confirma que los métodos Runge–Kutta de la familia Lobatto IIIA pueden ser buenos candidatos como integradores de problemas stiff. Estos métodos han sido descartados tradicionalmente por no poseer propiedades de estabilidad no-lineal tan buenas como otras familias tales como Radau IIA, Gauss, Lobatto IIIC, etc. Sin embargo, por la experimentación realizada en el capítulo II de esta memoria y por el soporte teórico a favor de los métodos Lobatto IIIA dado en los trabajos [13], [14] se puede avalar el uso de estos métodos. De hecho ateniéndonos a los resultados prácticos, en algunos problemas son más eficientes que los de la familia Radau IIA. Esto se debe seguramente a que cuando ambas familias operan con un mismo número de etapas implícitas, y por lo tanto implican un costo computacional por paso similar, resulta que los Lobatto IIIA poseen mayor orden clásico y mayor orden de etapa.

3. En los capítulos IV y V de esta memoria se establece un marco teórico general para el estudio y desarrollo de algoritmos de arranque. Todo esto se hace en base al orden alcanzado por dichos algoritmos, el cual se mide desde tres perspectivas distintas. En primer lugar se estudia el orden clásico a través de desarrollos en series de Butcher; en segundo lugar se estudia el orden sobre el test de Prothero y Robinson, haciéndose notar aquí el efecto que los errores globales acumulados (en los pasos anteriores de la integración) producen sobre el orden alcanzado por los algoritmos de arranque. En tercer lugar, se analiza el orden alcanzado sobre problemas no lineales contractivos.

   Se concluye, de modo general, y para los tipos de inicializador propuestos, que los tres órdenes de convergencia (clásico, Prothero y Robinson, y stiff) suelen coincidir o estar muy próximos entre sí. Esto supone un respaldo teórico, al uso con fines prácticos de inicializadores de tipo "interpolación de Lagrange" de las etapas previamente calculadas en el paso anterior y que se usan tradicionalemente como aproximaciones iniciales en los códigos actuales.

4. Se aporta también una nueva gama de inicializadores, principalmente los de Tipo II, que cuando se estabilizan suelen dar una mayor robustez a los códigos, involucrando además un menor costo computacional global en general.

**Problemas abiertos e investigación futura**

1. En lo referente a la iteración de tipo Single–Newton, creemos que sería interesante obtener esquemas específicos para métodos de colocación de alto orden tales como Gauss, Radau IIA y Lobatto IIIA de 5 etapas implícitas y de órdenes respectivos 10, 9 y 10.

   Estos métodos son interesantes a efectos de explotar las posibles ventajas de los métodos de mayor orden frente a los métodos de orden más bajo, sobre todo cuando se requiere una alta precisión en la integración.

2. También creemos muy prometedor el desarrollo de la iteración Single–Newton para métodos implícitos de tipo Runge–Kutta Nyström, a efectos de integrar sistemas diferenciales de segundo orden de tipo stiff de la forma:

$$y''(t) = f(t, y), \quad y(0) = y_0, \quad y'(0) = y'_0, \quad t \in [0, T].$$

   Obsérvese que este problema puede expresarse como uno equivalente de primer orden (doblando la dimensión) y luego se podría aplicar un Runge–Kutta (por ejemplo el Gauss de $s$ etapas) sobre el PVI correspondiente, del cual se podrían resolver sus ecuaciones de etapa mediante una iteración Single–Newton de las ya estudiadas. Todo esto proceso es mucho más costoso computacionalmente que intentar aplicar directamente un posible esquema de tipo "Single–Newton", sobre el método Runge–Kutta Nyström usado para la solución del PVI de segundo orden, sin tener que pasar al problema de primer orden. A este respecto, podemos decir que van der Howen et al. [42], [43] han desarrollado algún tipo de esquema con alguna semejanza al Single–Newton, pero para proceder en computación en paralelo.

   Por otra parte, consideramos también que una línea de investigación muy interesante sería la extensión a DAE's (Differential Algebraic Equations) de la iteración Single–Newton para resolver las ecuaciones de etapa de métodos Runge–Kutta. En [37, Cap. VII.3] y las referencias allí dadas se trata el caso de la iteración de Newton Simplificada para DAE's de índice 2.

3. En cuanto a las aproximaciones iniciales podemos concluir del estudio realizado en los capítulos IV y V, que la eficiencia de éstas, no depende sólo de su orden, pues los inicializadores con pobres propiedades de estabilidad pueden generar propagaciones de los errores poco satisfactorias. Por esta razón, los inicializadores con mayor orden no siempre dan lugar a menores errores. Pensamos, por tanto, que sería interesante investigar la posibilidad de encontrar algoritmos que nos permitan seleccionar en cada paso de la integración el inicializador más adecuado. Esto sería factible posiblemente mediante el diseño de estimadores del error cometido por cada inicializador.

Apéndice A

# Minimización del radio espectral

.

# Apéndice A
# Minimización del radio espectral

En el capítulo III hemos visto el teorema III.3.1 que nos dice que la condición de minimización (P4) del radio espectral de la matriz $M(z)$ (ver (3.3.3)) en el semieje real negativo se consigue para unas constantes $\Gamma_1^*, \Gamma_2^*, \Gamma_3^*$ (que dependen sólo de la matriz $A$ del método RK) de tal modo que

$$\mathrm{tr}(AT) = \Gamma_1^*, \quad \mathrm{tr}(A^{-2}T) = \Gamma_2^*, \quad \mathrm{tr}(T^{-2}A) = \Gamma_3^*.$$

Ahora vamos a demostrar que para los métodos de 4 etapas implícitas Gauss, Radau IIA y Lobatto IIIA, estas constantes se definen de forma unívoca (mediante unas ecuaciones que detallaremos debajo) y son (en el caso del Lobatto IIIA en lugar de la matriz $A$ del método usamos su submatriz $\bar{A}$):

| | Gauss | Radau IIA | Lobatto IIIA |
|---|---|---|---|
| $\Gamma_1^*$ | 0.06316638299274640 | 0.08296548473447771 | 0.06316638299274640 |
| $\Gamma_2^*$ | 14.86450048645422 | 11.01111697854543 | 14.86450048645422 |
| $\Gamma_3^*$ | 31.73222088578815 | 27.66997861432155 | 31.73222088578815 |

obteniendo respectivamente las cantidades mínimas

$$\rho^0_{\max} = \max\{\rho(M(z)) \mid z \in I\!R^-\}$$

siguientes:

| | Gauss | Radau IIA | Lobatto IIIA |
|---|---|---|---|
| $\rho^0_{\max}$ | 0.0893204199714 | 0.104708968155 | 0.0893204199714 |

Para demostrarlo, recordemos que en la sección III.3 vimos que si una matriz cualquiera $T$ de orden 4 verifica las condiciones (P1) y (P2), el cero es autovalor de $M(z)$ y los otros tres restantes son las raíces de la ecuación de autovalores

$$\mu^3 - b_3\mu^2 + b_2\mu - b_1 = 0 \tag{A.1}$$

con

$$b_1 = \frac{\gamma^4\xi_0 z^3}{(1-\gamma z)^4}, \qquad b_2 = \frac{\xi_2 z^2 + \xi_3 z^3}{(1-\gamma z)^4}, \qquad b_3 = \frac{\eta_1 z + \eta_2 z^2 + \eta_3 z^3}{(1-\gamma z)^4}, \tag{A.2}$$

donde

$$\xi_0 = 12\gamma^{-1} - 3\mathrm{tr}(A^{-1}) + \Gamma_2 - \Gamma_3,$$

$$\xi_2 = \frac{(\mathrm{tr}^2(A) - \mathrm{tr}(A^2))}{2} + 6\gamma^2 - 4\gamma\mathrm{tr}(A) + \Gamma_1, \quad \xi_3 = -4\gamma^4\mathrm{tr}(A^{-1}) + 20\gamma^3 + \gamma^4\Gamma_2 - 2\gamma^4\Gamma_3,$$

| | Gauss | Radau IIA | Lobatto IIIA |
|---|---|---|---|
| $r_0$ | 0.2053294868454444 | 0.2383872813698223 | 0.2053294868454444 |

TABLA A.1: Valores mínimos del radio espectral en el caso $b_1 \equiv b_2 \equiv 0$.

$$\eta_1 = \mathrm{tr}(A) - 4\gamma, \quad \eta_2 = 12\gamma^2 - 4\gamma \mathrm{tr}(A) + \Gamma_1, \quad \eta_3 = 4\gamma^3 - \gamma^4\Gamma_3,$$

denotando

$$\Gamma_1 = \mathrm{tr}(AT), \quad \Gamma_2 = \mathrm{tr}(A^{-2}T), \quad \Gamma_3 = \mathrm{tr}(T^{-2}A).$$

Debido al alto número de parámetros que intervienen en este problema, haremos una primera aproximación con el caso particular en que el cero es autovalor doble de $M(z)$, para luego pasar al caso general.

**Primera aproximación**

Consideremos el caso

$$b_1(z) \equiv 0$$

de lo que se deduce $\xi_0 = 0$ o, lo que es lo mismo:

$$\Gamma_2 = \Gamma_3 - 12\gamma^{-1} + 3\mathrm{tr}A^{-1}. \tag{A.3}$$

Consecuentemente, el cero sería un autovalor doble de $M(z)$ y la ecuación de autovalores (A.1) se reduciría a

$$\mu^2 - b_3\mu + b_2 = 0, \tag{A.4}$$

con lo que tenemos un problema del mismo tipo al que aparece en el capítulo II (subsección II.3.2) para el caso del Lobatto IIIA de 4 etapas (ver ecuación (2.3.7)).

Por tanto, para estudiarlo vamos a seguir un proceso análogo al seguido en la demostración del Lema II.3.1, con la única diferencia de que ahora disponemos de dos parámetros libres (en el Lema II.3.1 sólo se dispone de un parámetro).

Así, en primer lugar consideramos el caso

$$b_2(z) \equiv 0,$$

con lo cual se tiene $\xi_2 = \xi_3 = 0$, de lo que salen unívocamente determinados unos valores para $\Gamma_1 = \Gamma_1^0$ y $\Gamma_3 = \Gamma_3^0$ y se consigue para cada uno de los métodos un valor mínimo del radio espectral

$$0 < r_0 = \max\{|b_3(z)| \ / \ z \in (-\infty, 0)\} < 1$$

cuyas cantidades se reflejan en la tabla A.1.

En caso contrario ($b_2 \neq 0$), sea un $r$ cualquiera tal que

$$0 < r \leq r_0.$$

Haciendo ahora el cambio de variables (2.3.8) obtenemos la ecuación de autovalores transformada:

$$r^2x^2(t) - rB_3(t)x(t) + B_2(t) = 0, \qquad t \in (0,1), \tag{A.5}$$

donde

$$\begin{aligned} B_2 &= (at+b)(t-1)^2t \\ B_3 &= B_2 + \Pi(t) \end{aligned} \tag{A.6}$$

con

$$a = \gamma^{-2}\Gamma_1 - \gamma\Gamma_3 + K_1, \qquad b = \gamma\Gamma_3 + K_2, \tag{A.7}$$

donde las constantes $K_1$ y $K_2$ y la función $\Pi(t)$ sólo dependen de la matriz $A$ del método y son:

$$\left.\begin{array}{c} K_1 = (\mathrm{tr}^2 A - \mathrm{tr}A^2)/(2\gamma^2) - 4\gamma^{-1}\mathrm{tr}A - \gamma\mathrm{tr}A^{-1} + 14, \qquad K_2 = \gamma\mathrm{tr}A^{-1} - 8, \\ \Pi(t) = \kappa_1(t-1)t^3 + \kappa_2(t-1)^2t^2 + \kappa_3(t-1)^3t, \\ \kappa_1 = \gamma^{-1}\mathrm{tr}A - 4, \quad \kappa_2 = -(\mathrm{tr}^2 A - \mathrm{tr}A^2)/(2\gamma^2) + 6, \quad \kappa_3 = \gamma\mathrm{tr}A^{-1} - 4. \end{array}\right\} \tag{A.8}$$

Con esto se tiene que el radio espectral de $M(z)$ será menor o igual a $r$ en todo el semieje real negativo si todas las raíces de la ecuación de autovalores transformada (A.5) están en el disco unidad para todo $t \in (0,1)$. Aplicando entonces el criterio de Schur–Cohn [51] y simplificando, esto resulta equivalente a que se verifiquen las siguientes desigualdades para todo $t \in (0,1)$:

$$0 < |at + b| < R(r,t), \tag{A.9}$$

$$at + b > \max\{P_1(r,t), P_2(r,t)\}, \tag{A.10}$$

donde

$$R(r,t) = \frac{r^2}{(1-t)^2 t},$$

$$P_1(r,t) = -\frac{r}{(1+r)}\frac{r+\Pi(t)}{(1-t)^2 t}, \qquad P_2(r,t) = -\frac{r}{(1-r)}\frac{r-\Pi(t)}{(1-t)^2 t}.$$

Con esta información queremos hallar el ínfimo del conjunto

$$\Omega_0 = \{r \in (0, r_0] \text{ tal que existen } a \text{ y } b \text{ que verifican (A.9) y (A.10)}\},$$

que, como $r_0 \in \Omega_0$, es no vacío. Por tanto, escogemos un $r \in \Omega_0$ y estudiamos con más detalle las desigualdades.

De la primera (A.9) claramente se deduce que

$$at + b < 0, \ \forall t \in (0,1) \quad \text{ó} \quad at + b > 0, \ \forall t \in (0,1).$$

Además, hay que tener en cuenta que para los tres métodos estudiados el polinomio $\Pi(t)$ es siempre positivo en $(0,1)$, anulándose solamente en los extremos 0 y 1. A partir de esto, estudiamos las dos alternativas posibles.

(1) Si $at + b > 0, \ \forall t \in (0,1)$: como $P_1(r,t) < 0$ será necesario que

$$R(r,t) - P_2(r,t) > 0, \ \forall t \in (0,1),$$

o de forma equivalente, que $2r - r^2 > \Pi(t), \ \forall t \in (0,1)$. En consecuencia el valor de $r = r_1$ menor posible se obtiene al exigir

$$2r_1 - r_1^2 = \Pi_0 \quad \text{con} \quad \Pi_0 = \max\{\Pi(t), \ t \in (0,1)\}.$$

Con esto conseguimos de forma unívoca los valores de $r_1$ para cada uno de los tres métodos dados en la tabla A.2.

(2) Si $at + b < 0, \ \forall t \in (0,1)$: de $P_2(r,t) < at + b < 0$ se tendría que $\Pi(t) < r, \ \forall t \in (0,1)$, o lo que es lo mismo $r > \Pi_0 > r_1$, con lo cual $r_1$ es el ínfimo de $\Omega_0$.

| | Gauss | Radau IIA | Lobatto IIIA |
|---|---|---|---|
| $r_1$ | 0.1085570611898058 | 0.1272957438913353 | 0.1085570611898058 |
| $a_1$ | 0.1711167002572258 | 0.2314020933796865 | 0.1711167002572258 |
| $b_1$ | 0.008615576167832601 | 0.01374029830773092 | 0.008615576167832601 |
| $\Gamma_1^1$ | 0.06325114414874897 | 0.08315415968797673 | 0.06325114414874897 |
| $\Gamma_3^1$ | 31.2725323654151098 | 27.14247793294383 | 31.2725323654151098 |

TABLA A.2: Valores mínimos del radio espectral en el caso $b_2 \equiv 0$.

Además, si $t_0 \in (0,1)$ es tal que $\Pi(t_0) = \Pi_0$ (que en los métodos estudiados es único), resulta que

$$R(r_1,t_0) = P_2(r_1,t_0) > P_1(r_1,t_0), \qquad \frac{d}{dt}R(r_1,t_0) = \frac{d}{dt}P_2(r_1,t_0),$$

y en consecuencia existe una única recta $a_1 t + b_1$ que verifica las desigualdades (A.9) y (A.10) para $r = r_1$, que está univocamente determinada por

$$a_1 = \frac{d}{dt}R(r_1,t_0), \qquad b_1 = R(r_1,t_0) - a_1 t_0.$$

Evidentemente, de aquí se obtienen unos únicos valores $\Gamma_1 = \Gamma_1^1$, $\Gamma_3 = \Gamma_3^1$ en el caso de que el cero sea autovalor doble de la matriz $M(z)$, con lo que conjuntamente con (A.3) se demuestra lo que queremos. Los valores de estos parámetros correspondientes a cada uno de los métodos considerados se recogen en la tabla A.2.

**Minimización en general**

Ahora consideramos el caso general de la ecuación de autovalores de $M(z)$ (A.1) y elegimos un valor $r$ cualquiera tal que

$$0 < r \leq r_1.$$

Haciendo el mismo cambio de variable (2.3.8) a la ecuación completa se obtiene:

$$r^3 x^3(t) - r^2 B_3(t) x^2(t) + r B_2(t) x(t) - B_1(t) = 0, \quad t \in (0,1) \tag{A.11}$$

donde

$$\begin{aligned} B_1 &= d(t-1)^3 t \\ B_2 &= (at+b)(t-1)^2 t \\ B_3 &= B_2 - B_1 + \Pi(t) \end{aligned} \tag{A.12}$$

con

$$a = \gamma^{-2}\Gamma_1 + \gamma\Gamma_2 - 2\gamma\Gamma_3 + \bar{K}_1, \quad b = -\gamma\Gamma_2 + 2\gamma\Gamma_3 + \bar{K}_2, \quad d = \gamma\Gamma_2 - \gamma\Gamma_3 + \bar{K}_3, \tag{A.13}$$

donde las constantes $\bar{K}_i$, $i = 1,2,3$ sólo dependen de la matriz $A$ del método y vienen dadas por

$$\bar{K}_1 = (\mathrm{tr}^2 A - \mathrm{tr}A^2)/(2\gamma^2) - 4\gamma^{-1}\mathrm{tr}A - 4\gamma\mathrm{tr}A^{-1} + 26,$$

$$\bar{K}_2 = 4\gamma\mathrm{tr}A^{-1} - 20, \quad \bar{K}_3 = -3\gamma\mathrm{tr}A^{-1} + 12.$$

Del mismo modo que en el caso anterior, el radio espectral de $M(z)$ será menor que $r$ para todo $z \leq 0$ si y sólo si todas las raíces de esta ecuación transformada caen en el disco unidad para todo $t \in (0,1)$. Para demostrarlo usamos de nuevo el criterio de Schur–Cohn pero ahora

para ecuaciones de tercer grado, con lo que, después de algunas simplificaciones, tenemos que demostrar que existen valores de $a$, $b$ y $d$ tal que se verifican las desigualdades siguientes:

$$0 < |d| < \frac{r^3}{(1-t)^3 t}, \tag{A.14}$$

$$|\Lambda_0| < \Lambda_2, \tag{A.15}$$

$$|\Lambda_1| < \Lambda_0 + \Lambda_2, \tag{A.16}$$

donde

$$\Lambda_0 = B_2 r^4 - B_1 B_3 r^2, \quad \Lambda_1 = B_1 B_2 r - B_3 r^5, \quad \Lambda_2 = r^6 - B_1^2.$$

Desarrollando estas desigualdades llegamos a que son equivalentes a

$$0 < |d| < \frac{256}{27} r^3, \tag{A.17}$$

$$\max\left\{\frac{F_1^*(r,t)}{(t-1)^2 t}, \frac{F_2(r,t)}{(t-1)^2 t}, \frac{F_2^*(r,t)}{(t-1)^2 t}\right\} < at + b < \frac{F_1(r,t)}{(t-1)^2 t}. \tag{A.18}$$

siendo

$$\begin{aligned}
F_1(r,t) &:= \frac{1}{r^2(r^2 - B_1)}\left(-(1+r^2)B_1^2 + r^2\Pi(t)B_1 + r^6\right), \\
F_1^*(r,t) &:= \frac{1}{r^2(r^2 - B_1)}\left((1-r^2)B_1^2 + r^2\Pi(t)B_1 - r^6\right), \\
F_2(r,t) &:= \frac{1+r}{r}B_1 + \frac{r}{1-r}(\Pi(t) - r) \\
F_2^*(r,t) &:= -\frac{1-r}{r}B_1 - \frac{r}{1+r}(\Pi(t) + r).
\end{aligned}$$

Debido a la gran complejidad de las desigualdades que resultan, en lo que sigue nos vamos a restringir a los tres métodos de cuatro etapas implícitas que nos interesan: el Gauss, el Radau IIA y el Lobatto IIIA (recordando siempre que en este último caso la matriz $A$ del método se sustituye por la submatriz regular $\bar{A}$).

En el apartado anterior hallamos el $r = r_1$ ínfimo que minimizaba el radio espectral para el caso $b_1(z) \equiv B_1(t) \equiv 0$ (o lo que es lo mismo $d = 0$), obteniendo unos valores de $a = a_1$ y $b = b_1$ de forma única (ver tabla A.2). Ahora queremos hallar el $r$ ínfimo del conjunto

$$\Omega_1 = \{r \in (0, r_1] \quad \text{tal que existen } a,\ b \text{ y } d \text{ que verifiquen (A.17) y (A.18)}\}$$

que es no vacío pues $r_1 \in \Omega_1$. Claramente, por razones de continuidad, los valores óptimos de $(a, b, d)$ estarán en un entorno de $(a_1, b_1, 0)$. Representando gráficamente las funciones $F_1(r,t)$, $F_2(r,t)$, $F_1^*(r,t)$, $F_2^*(r,t)$ en un entorno de los valores $r_1$, $a_1$, $b_1$, $d_1 = 0$, hemos deducido empíricamente que:

(1) Si $d \in (-256/27\ r^3, 0)$, si se verifican las condiciones:

$$\begin{aligned}
f_1(t; r, a, b, d) &:= \frac{F_1(r,t)}{(t-1)^2 t} - (at + b) > 0, \\
f_2(t; r, a, b, d) &:= (at + b) - \frac{F_2(r,t)}{(t-1)^2 t} > 0,
\end{aligned}$$

se tiene (A.18).

| | Gauss | Radau IIA | Lobatto IIIA |
|---|---|---|---|
| $r_2$ | 0.08932041997140538 | 0.1047089681545739 | 0.08932041997140538 |
| $a_2$ | 0.08933347954571903 | 0.1175291657755426 | 0.08933347954571903 |
| $b_2$ | 0.08692462330964289 | 0.1221449062444267 | 0.08692462330964289 |
| $d_2$ | -0.006507080774954180 | -0.01042105042495914 | -0.006507080774954180 |
| $t_1$ | 0.2118947594963050 | 0.2102642865738790 | 0.2118947594963050 |
| $t_2$ | 0.5738267823492197 | 0.5722382251308269 | 0.5738267823492197 |
| $t_3$ | 0.3375332575647970 | 0.3474279357113105 | 0.3375332575647970 |
| $t_4$ | 0.6624667424352030 | 0.6525720642886895 | 0.6624667424352030 |
| $\Gamma_1^*$ | 0.06316638299274640 | 0.08296548473447771 | 0.06316638299274640 |
| $\Gamma_2^*$ | 14.86450048645422 | 11.01111697854543 | 14.86450048645422 |
| $\Gamma_3^*$ | 31.73222088578815 | 27.66997861432155 | 31.73222088578815 |

TABLA A.3: Valores mínimos del radio espectral.

Estudiando las gráficas de estas funciones en $t \in (0,1)$ en un entorno de los $r_1$, $a_1$, $b_1$, y $d_1 = 0$, se observa que el valor $r = r_2$ óptimo es el valor para el que existen $a_2, b_2, d_2 \in \mathbb{R}$ y $t_1, t_2, t_3, t_4 \in (0,1)$ de tal forma que se verifican las ocho ecuaciones siguientes:

$$f_1(t_1; r_2, a_2, b_2, d_2) = f_1(t_2; r_2, a_2, b_2, d_2) = 0,$$

$$\frac{d}{dt} f_1(t_1; r_2, a_2, b_2, d_2) = \frac{d}{dt} f_1(t_2; r_2, a_2, b_2, d_2) = 0,$$

$$f_2(t_3; r_2, a_2, b_2, d_2) = f_2(t_4; r_2, a_2, b_2, d_2) = 0,$$

$$\frac{d}{dt} f_2(t_3; r_2, a_2, b_2, d_2) = \frac{d}{dt} f_2(t_4; r_2, a_2, b_2, d_2) = 0.$$

Resolviendo estas ecuaciones obtenemos de forma unívoca los valores de $r_2$, $a_2$, $b_2$, $d_2$ correspondientes a cada uno de los métodos, y por tanto, este $r_2$ es el menor de los de $\Omega_1$ que se puede obtener con $d \in (-256/27\, r^3, 0)$.

(2) Si $d \in (0, 256/27\, r^3)$ se obtienen unos valores de $r \in \Omega_1$ mayores que $r_2$, por lo que se tiene que $r_2$ es el ínfimo de $\Omega_1$, y, por tanto,

$$\rho^0_{\max} = \max\{\rho(M(z)) \mid z \in \mathbb{R}^-\} = r_2.$$

Los valores de estos parámetros óptimos $r_2$, $a_2$, $b_2$, $d_2$ y los puntos $t_1$, $t_2$, $t_3$, $t_4$ para cada uno de los métodos considerados se dan en la tabla A.3, así como los correspondientes $\Gamma_1^*$, $\Gamma_2^*$ y $\Gamma_3^*$ que resultan unívocamente determinados de (A.13).

# Bibliografía


[1] R. Aiken. *Stiff computation*, Oxford University Press, Oxford (1985).

[2] Bickart, T. A., *An efficient solution process for implicit Runge–Kutta methods*, SIAM J. Numer Anal., **14** (1977) 1022–1027.

[3] P.N. Brown, G.D. Byrne & A.C. Hindsmarsh, *VODE: a variable coefficient ODE solver*, SIAM J. Sci. Stat. Comput., **10** (1989) 1039–1051.

[4] K. Burrage, J.C. Butcher & F.H. Chipmann. *STRIDE: Stable Runge-Kutta integrator for differential equations*, Report Series No. 150, Dept. Mathematics, University of Auckland (1979).

[5] K. Burrage, J.C. Butcher and F.H. Chipman, *An implementation of singly–implicit Runge–Kutta methods*, BIT, **20** (1980) 326–340.

[6] K. Burrage, W. H. Hundsdorfer and J.G. Verwer, *A study of B-convergence of Runge-Kutta methods*, Computing **36** (1986) 17–34.

[7] K. Burrage and W. H. Hundsdorfer *The order of B-convergence of algebraically stable Runge-Kutta methods*, BIT **27** (1987) 62–71.

[8] J.C. Butcher, *Implicit Runge-Kutta processes* Math. Comput. **18** (1964) 50–64.

[9] J.C. Butcher, *On the implementation of implicit Runge–Kutta methods*, BIT, **16** (1976) 237–240.

[10] J.C. Butcher, *Some implementation schemes for implicit Runge–Kutta methods*, Proceedings of the Dundee Conference on Numerical Analysis, 1979. Lecture Notes in Mathematics **773** (1979) 12–24.

[11] J.C. Butcher, *The numerical analysis of ordinary differential equations* John Wiley & Sons, Chichester, (1987).

[12] G.D. Byrne, and A.C. Hindmarsh, *A polyalgorithm for the numerical solution of ordinary differential equations*, ACM Trans. Math. Soft., **1** (1975) 71–96.

[13] M. Calvo, S. González Pinto and J.I. Montijano, *Runge–Kutta methods for the numerical solution of stiff semilinear initial value problems*, enviado a BIT (1998).

[14] M. Calvo, S. González Pinto and J.I. Montijano, *Convergence of Runge–Kutta methods for stiff nonlinear differential equations*, Numer. Math. **81** (1998) 31–51.

[15] M. Calvo, S. González Pinto, and J.I. Montijano, *On the iterative solution of the algebraic equations in fully implicit Runge–Kutta methods*, enviado a Numerical Algorithms (1998).

[16] F.H. Chipman, *The implementation of Runge–Kutta processes*, BIT, **13** (1973) 391–393.

[17] G.J. Cooper and J.C. Butcher, *An iteration scheme for implicit Runge-Kutta methods,* IMA J. of Num. Anal., **3** (1983) 127-140.

[18] G.J. Cooper and R. Vignesvaran, *A scheme for the implementation of implicit Runge-Kutta methods*, Computing, **45** (1990) 321-332.

[19] K. Dekker and J.G. Verwer, *Stability of Runge–Kutta methods for stiff nonlinear differential equations*, North Holland, Amsterdam (1984).

[20] J.L.M. Dorsselaer and M.N. Spijker, *The error committed by stopping the Newton iteration in the numerical solution of stiff initial value problems*, IMA J. Numer. Anal., **14** (1994) 183-209.

[21] B.L. Ehle, *On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial stiff problems*, Research Report CSRR 2010, Univ. of Waterloo, Ontario, Canada (1969).

[22] W.H. Enright, T.E. Hull and B. Lindberg, *Comparing numerical methods for stiff systems of ODEs*, BIT, **15** (1975) 10–48.

[23] W.H. Enright and J.D. Pryce, *Two FORTRAN packages for assessing initial value methods,* ACM Transactions on Mathematical Software, **13** (1) (1987) 1-27.

[24] R. Frank, J. Schneid and C.W. Ueberhuber, *The concept of B-convergence*, SIAM J. Numer.Anal., **18** (1981) 753–780.

[25] R. Frank and C.W. Ueberhuber. *Iterated defect correction for efficient solution of stiff systems of ordinary differential equations*, BIT, **17**, (1977) 146-159.

[26] C. W. Gear, *Numerical initial value problems in ordinary differential equations*, Prentice–Hall, Englewood Cliffs, New Jersey (1971).

[27] I. Gladwell and R.M. Thomas, *Efficiency of methods for second order problems*, IMA J. Numer. Anal. **10** (1990) 181–207.

[28] S. González Pinto, C. González Concepcion and J.I. Montijano, *Iterative schemes for Gauss methods*, Computers Math. Applic., **27** (7) (1994) 67–81.

[29] S. González Pinto, J.I. Montijano and L. Rández, *Iterative schemes for three–stage implicit Runge–Kutta methods*, Appl. Numer. Math., **17** (1995) 363-382.

[30] S. González Pinto, J.I. Montijano and S. Pérez Rodríguez, *On the numerical solution of stiff IVPs by Lobatto IIIA Runge–Kutta methods*, J. Comp. and Appl. Math., **82** (1997) 129-148.

[31] S. González Pinto, J.I. Montijano and S. Pérez Rodríguez, *On the implementation of high order implicit Runge-Kutta methods*, aceptado en Computers Math. Applic. (febrero 2000).

[32] S. González Pinto, J.I. Montijano y S. Pérez Rodríguez, *Algoritmos de arranque para métodos Runge–Kutta implícitos*, Actas del XVI CEDYA/VI CMA, Edit. R. Montenegro, G. Montero y G. Winter, Las Palmas de Gran Canaria (1999) 1149-1156.

[33] S. González Pinto, J.I. Montijano and S. Pérez Rodríguez, *On the starting algorithms for fully implicit Runge-Kutta methods*, enviado a BIT (1998).

[34] S. González Pinto, J.I. Montijano and S. Pérez Rodríguez, *Stabilized starting algorithms for collocation Runge-Kutta methods*, preprint (Computers Math. Applic., marzo 2000).

[35] E. Hairer, Ch. Lubich and M. Roche, *The numerical solution of differential algebraic systems by Runge–Kutta methods*, Lecture Notes in Math., **1049**, Springer Verlag (1989).

[36] E. Hairer, S.P. Nørsett and G. Wanner, *Solving ordinary differential equations I*, Springer Verlag, Berlin (1993).

[37] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II* Springer–Verlag, Berlin (1996).

[38] E. Hairer and G. Wanner, *Stiff differential equations solved by Radau methods*, ponencia presentada en Coimbra, Portugal (1998).

[39] N. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia (1996).

[40] A.C. Hindmarsh, *LSODE and LSODI, two new initial value ordinary differential equations solvers*, ACM-SIGNUM Newsletter **15** (1980) 10–11.

[41] P.J. van der Houwen and B.P. Sommeijer, *Preconditioning in parallel Runge-Kutta methods for stiff initial value problems*, J. Comput. Math. Appl. **28**, No. 10–12 (1994) 17–31.

[42] P.J. van der Houwen, B.P. Sommeijer and J.Kok *The iterative solution of fully implicit discretizations of 3-dimensional transport models*, Appl. Numer. Math. **25** (1997) 243-256.

[43] P.J. van der Houwen and E. Messina, *Splitting methods for second order initial value problems*, Numerical Algorithms **18** (1998) 233–257.

[44] M.P. Laburta, *Inicializadores para las ecuaciones implícitas de los métodos RK para sistemas hamiltonianos*, Tesis Doctoral, Universidad de Zaragoza (1996).

[45] M.P. Laburta, *Starting algorithms for IRK methods*, J. Comput. Appl. Math., **83** (1997) 269–288.

[46] M.P. Laburta, *Construction of starting algorithms for the RK-Gauss methods*, J. Comput. Appl. Math., **90** (1998) 239–261.

[47] J.D. Lambert, *Numerical Methods for Ordinary Differential Systems*, John Wiley & Sons Ltd., England (1991).

[48] W. Liniger, *A stopping criterion for the Newton-Raphson method in implicit multistep integration algorithms for nonlinear systems of ordinary differential equations*, Commun. ACM, **14** (1971) 600–601.

[49] W.M. Lioen, J.J.B. de Swart and W.A. van der Veen, *Test set for IVP solvers*, http://www.cwi.nl/cwi/projects/IVPtestset.shtml, Test set for IVP solvers (1996).

[50] C. Mac Donald and W. Enright, *Implications of order reduction for implicit Runge–Kutta methods*, Numerical Algorithms, **2** (1992) 351–370.

[51] M. Marden, *Geometry of polynomials*, American Mathematical Society, Providence, Rhode Island, 2nd edit. (1966).

[52] S.P. Nørsett and P.G. Thomsen, *Local error control in SDIRK-methods*, BIT, **26** (1986) 100–113.

[53] H. Olsson and G. Söderlind, *Stage value predictors and efficient Newton iterations in implicit Runge–Kutta methods*, SIAM Journal on Scientific Computing, **20**, 1 (1999) 185–202.

[54] S. Pérez Rodríguez, *Problemas stiff y métodos Runge–Kutta*, Tesina de Licenciatura, Universidad de La Laguna (1995).

[55] A. Prothero and A. Robinson, *On the stability and accuracy of one–step methods for solving stiff systems of ordinary differential equations*, Math. of Comp., **28** (1974) 145-162.

[56] T. Roldán and I. Higueras, *IRK methods for DAE: starting algorithms*, Dpto. de Matemáticas e Informática, Univ. Pública de Navarra (1998).

[57] J. Sand, *Starting methods for the iteration of IRK's*, Dept. of Computer Science, Univ. Copenhagen, Denmark (1989).

[58] M.N. Spijker, *On the error commited by stopping the Newton iteration in implicit Runge-Kutta methods*, Annals of Numer. Math., **1** (1994) 199-212.