Curso 2005/06
CIENCIAS Y TECNOLOGÍAS/15
I.S.B.N.: 84-7756-707-7

DOMINGO HERNÁNDEZ ABREU

Integración numérica de sistemas diferenciales con características especiales

Director
SEVERIANO GONZÁLEZ PINTO

SERVICIO DE PUBLICACIONES
UNIVERSIDAD DE LA LAGUNA

SOPORTES AUDIOVISUALES E INFORMÁTICOS
Serie Tesis Doctorales

Agradecimientos

En primer lugar, deseo agradecer al profesor Dr. D. Severiano González Pinto las orientaciones dadas y su dedicación e interés en el desarrollo de este proyecto de Tesis Doctoral. Deseo mostrar además mi más sincera gratitud por la amistad ofrecida y por el excelente ritmo de trabajo del que he podido disfrutar en estos años.

Deseo mostrar mi agradecimiento a los profesores Dres. D. Manuel Calvo Pinilla, D. Juan Ignacio Montijano Torcal y D. Luis Rández García con quienes tuve la oportunidad de trabajar durante mis estancias en el Departamento de Matemática Aplicada de la Universidad de Zaragoza. Agradezco además el excelente trato recibido por parte de los profesores de dicho departamento.

Agradezco profundamente la excelente acogida recibida por parte de todos y cada uno de los profesores y profesoras del Departamento de Análisis Matemático de la Universidad de La Laguna durante el periodo de realización de esta memoria. Asimismo, quiero dar gracias a D. Pablo Felipe Lorenzo Cáceres, auxiliar administrativo de dicho departamento, por su inestimable ayuda para solventar detalles técnicos.

Deseo agradecer a la Universidad de La Laguna y al Ministerio de Educación y Ciencia el haberme concedido becas de investigación que, sin lugar a dudas, han hecho posible que este proyecto de investigación llegara a buen puerto.

Gracias a todos mis amigos y compañeros de doctorado Fernando, Hugo, Julio, Manolo, Pedro R., Ramón, Roberto y Rogel por haber hecho más grato el desarrollo de esta investigación.

Deseo hacer mención especial para D. José Luis Montesinos Sirera por ser el máximo responsable de mi atracción por las matemáticas.

Finalmente, deseo agradecer a mi familia, en especial a mi padre, por su apoyo e interés en que este proyecto saliera adelante.

Índice general

Prólogo

Esta memoria trata el análisis de los métodos numéricos de uno o más pasos cuando se aplican a diversos tipos de ecuaciones diferenciales ordinarias. Por un lado, en los capítulos II, III y IV analizamos la estabilidad de los principales métodos de un paso para la integración numérica de sistemas diferenciales que poseen puntos de equilibrios semiestables correspondientes al caso de variedades centro unidimensionales. Más concretamente, estamos interesados en dar condiciones para los métodos de tipo Runge-Kutta, así como para los métodos semi-implícitos (o de Rosenbrock), de modo que éstos reproduzcan el comportamiento de las órbitas estables en entornos de los equilibrios. Esto nos conduce en el capítulo III a establecer la definición de E-estabilidad para los métodos de un paso. Recomendamos al lector la lectura conjunta y consecutiva de estos tres capítulos. Tras establecer en el capítulo I diversas ideas generales de carácter introductorio, efectuamos en el capítulo II una revisión de los resultados teóricos y numéricos incluidos en el trabajo [40] en el que se analiza la estabilidad incondicional de los métodos de Runge-Kutta sobre los sistemas diferenciales con equilibrios semiestables correspondientes al caso de variedades centro unidimensionales

• S. González-Pinto, *Differential systems with semi-stable equilibria and numerical methods*, Numerische Mathematik **96** (2003), 253-268.

Seguidamente, en los capítulos III y IV damos condiciones de cara a obtener estabilidad para métodos de Runge-Kutta implícitos y métodos de Rosenbrock. Los resultados que aparecen en los capítulos III y IV se recogen fundamentalmente en los trabajos [44] y [43], respectivamente

• S. González-Pinto, D. Hernández-Abreu, *Semi-implicit methods for differential systems with semi-stable equilibria*, Applied Numerical Mathematics **56** (2006), 210-221.

• S. González-Pinto, D. Hernández-Abreu, *Stable Runge-Kutta integrations for differential systems with semi-stable equilibria*, Numerische Mathematik **97** (2004), 473-491.

El estudio realizado en estos dos capítulos se apoya en la caracterización de A-aceptabilidad fuerte de las funciones racionales establecida en el trabajo [45]

• S. González-Pinto, D. Hernández-Abreu, *Strong A-Acceptability for rational functions*, BIT Numerical Mathematics **43** (2003), 555-561.

El capítulo V puede ser leído de forma independiente y trata el estudio de estabilidad y convergencia de los Métodos Lineales Generales sobre sistemas diferenciales stiff con constantes de Lipschitz laterales negativas en intervalos temporales semi-infinitos. Los resultados teóricos que se establecen en este capítulo pueden dar pautas a seguir de cara al estudio de estabilidad de los principales métodos multipaso y multietapa sobre la clase de sistemas diferenciales con equilibrios semiestables tratada en los capítulos previos. El contenido de este capítulo se corresponde en parte con el contenido del trabajo [41], así como de los resultados presentados en el 21*st Biennial Conference on Numerical Analysis*, Dundee [42]

• S. González-Pinto, D. Hernández-Abreu, *On the contractivity and convergence of general linear methods*, preprint enviado a SIAM Journal on Numerical Analysis, abril 2006.

• S. González-Pinto, D. Hernández-Abreu, *On the contractivity of the matrices related to general linear methods*, Numerical Analysis Report NA/222, 21*st* Biennial Conference on Numerical Analysis, University of Dundee, p. 23.

El capítulo VI es autocontenido y en él diseñamos una estrategia basada en técnicas de proyección para la conservación de invariantes conocidos para un sistema diferencial autónomo. Los resultados teóricos y numéricos presentados en este capítulo corresponden al contenido de los trabajos [17, 18, 19]

• M. Calvo, D. Hernández-Abreu, J.I. Montijano, L. Rández, *Explicit Runge-Kutta methods for the preservation of invariants*, 2004. Informe técnico. Departamento Matemática Aplicada. Universidad Zaragoza.

• M. Calvo, D. Hernández-Abreu, J.I. Montijano, L. Rández, *On the preservation of invariants by explicit Runge-Kutta methods*, aceptado para publicación en SIAM Journal on Scientific Computing, diciembre de 2005.

• M. Calvo, D. Hernández-Abreu, J.I. Montijano, L. Rández, *Preservation of quadratic invariants by means of explicit Runge-Kutta methods*, Actas del congreso internacional Computational and Mathematical Methods in Science and Engineering, CMMSE-2004, Universidad de Uppsala, p. 34-38.

Esta memoria concluye con un breve apéndice en el que se reflejan las conclusiones más importantes de la investigación realizada, así como con un esbozo de las pautas a seguir en una investigación futura.

Capítulo I

Introducción.

La modelización a través de sistemas de ecuaciones diferenciales de los fenómenos que aparecen en los diversos campos aplicados es un tema de indiscutible relevancia [53, 59, 69] y, en particular, el análisis y simulación de estos fenómenos requiere frecuentemente el uso de ecuaciones diferenciales ordinarias [52, 53, 55, 88]. Muchos sistemas físicos son modelados por sistemas de ecuaciones en derivadas parciales dependientes de la variable temporal que tras ser discretizados bien por el método de líneas asociado a una adecuada red espacial o bien por métodos de elementos finitos nos conducen a la resolución numérica de un sistema de ecuaciones diferenciales ordinarias. En el campo de la resolución numérica de las ecuaciones diferenciales ordinarias disponemos de eficientes métodos de resolución, así como de interesantes resultados relativos a la estabilidad, consistencia y convergencia de los mismos. La resolución numérica de estas ecuaciones provee una información de indudable valor acerca de la evolución de los sistemas descritos. No obstante, el motivo esencial de la relevancia de la aproximación numérica de las soluciones a los sistemas diferenciales que describen los diversos fenómenos aplicados radica en el hecho de que, por norma general, es prácticamente imposible o totalmente ineficiente obtener una expresión analítica exacta de las soluciones a tales sistemas. Muchos sistemas físicos pueden ser modelados por medio de un problema de valor inicial, expresado por medio de una ecuación diferencial ordinaria, bien en forma no autónoma

$$y'(t) = f(t, y(t)), \qquad y : I \subset \mathbb{R} \to \mathbb{R}^m, \quad f : I \times \mathcal{D} \subset \mathbb{R}^{m+1} \to \mathbb{R}^m, \tag{I.1}$$

o en forma autónoma

$$y'(t) = f(y(t)), \qquad y : I \subset \mathbb{R} \to \mathbb{R}^m, \quad f : \mathcal{D} \subset \mathbb{R}^m \to \mathbb{R}^m, \tag{I.2}$$

donde, en ambos casos, I denota un intervalo de la recta real, $\mathcal{D}$ un dominio de $\mathbb{R}^m$, la variable independiente t representa el tiempo en el sistema físico, mientras que la variable dependiente y denota la solución al sistema del cual se conoce un estado inicial $y(t_0) = y_0$. Notemos que todo sistema diferencial en forma no autónoma puede ser expresado de modo equivalente en forma autónoma sin más que añadir la ecuación trivial $t' = 1$. Es de interés conocer el estado futuro del sistema en un tiempo t_N, y, de hecho, determinar si tal solución existe y es única. Para responder a estas cuestiones se asume que la función f verifica la condición de Lipschitz respecto de la variable y, esto es, existe una constante $L > 0$ y una norma $\|\cdot\|$ en $\mathbb{R}^m$ tales que

$$\|f(t, y) - f(t, z)\| \leq L \|y - z\|, \quad \forall t \in I, \quad \forall y, z \in \mathcal{D}. \tag{I.3}$$

Bajo estas condiciones se demuestra entonces el clásico resultado de existencia y unicidad de soluciones para ecuaciones diferenciales ordinarias [26, 32, 52].

Teorema I.0.1 *Sea $f : I \times \mathcal{D} \subset \mathbb{R}^{m+1} \to \mathbb{R}^m$ una función continua en $I \times \mathcal{D}$ y acotada por $M > 0$, satisfaciendo la condición de Lipschitz (I.3) en $\{(t, y)/t_0 \leq t \leq T,\ \|y - y_0\| \leq b\}$. Si $T - t_0 \leq b/M$ entonces*

i) las poligonales de Euler

$$\begin{array}{rcl} y_h(t) & := & y_i + (t - t_i) f(t_i, y_i), \quad t_i \leq t \leq t_{i+1} := t_i + h, \quad i = 0, 1, 2, \ldots, \\ y_{i+1} & := & y_i + h f(t_i, y_i), \quad i = 0, 1, 2, \ldots, \end{array}$$

convergen uniformemente a una función continua $y(t)$ cuando $|h| \to 0$;

ii) la función $y(t)$ es continuamente diferenciable y es una solución del problema (I.1), con dato inicial $y(t_0) = y_0$, en el intervalo $t_0 \leq t \leq T$;

iii) el problema de valor inicial

$$\left\{ \begin{array}{rcl} y' & = & f(t, y) \\ y(t_0) & = & y_0 \end{array} \right.$$

no admite otra solución distinta en el intervalo $t_0 \leq t \leq T$.

La continuidad y la condición de Lipschitz para el dato f también permiten estudiar la sensibilidad de las soluciones frente a perturbaciones del valor inicial y_0 y del dato f. En este sentido, debemos comentar que una solución $y(t)$ de (I.1) se dice estable si las soluciones que

parten próximas a $y(t)$ permanecen próximas en todo tiempo, y que $y(t)$ es asintóticamente estable si es estable y las soluciones que parten de un entorno del valor inicial convergen en tiempo infinito a $y(t)$. Las soluciones de (I.1) que no sean estables se dirán inestables. Más formalmente tenemos las siguientes definiciones.

Definición I.0.2 *Una solución $y(t)$ de (I.1) se dice estable (en sentido de Lyapunov) si para cada $\varepsilon > 0$ existe $\delta = \delta(\varepsilon) > 0$ tal que para cualquier otra solución $z(t)$ de (I.1) verificando $\|z(t_0) - y(t_0)\| < \delta$ entonces $\|z(t) - y(t)\| < \epsilon$, para todo $t \in I,\ t \geq t_0$.*

Definición I.0.3 *Una solución $y(t)$ de (I.1) se dice asintóticamente estable si es estable (en sentido de Lyapunov) y si existe una constante $M > 0$ tal que si $\|z(t_0) - y(t_0)\| < M$ entonces $lim_{t\to\infty} \|z(t) - y(t)\| = 0$, siendo $z(t)$ solución de (I.1).*

Es bien conocido el siguiente resultado que asegura la estabilidad de las soluciones de (I.1) respecto del valor inicial y_0 y del dato f.

Teorema I.0.4 *Todo sistema diferencial (I.1), con condición inicial $y(t_0) = y_0$, y dato f continuo en $I \times \mathcal{D}$, I acotado, verificando la condición de Lipschitz (I.3) respecto de y, es totalmente estable, esto es, dado el sistema perturbado*

$$\begin{cases} z'(t) &= f(t,z) + \delta(t), \\ z_0 &= y_0 + \eta, \end{cases}$$

siendo δ continua en I, entonces existe $K > 0$ de modo que para todo $\varepsilon > 0$ se tiene que $\|y(t) - z(t)\| \leq K\varepsilon$, siempre que $\max\{\|\delta(t)\|, \|\eta\|\} \leq \varepsilon$,

En particular, en virtud del Lema de Gronwall [26, 32] se puede deducir una cota de estabilidad del tipo

$$\|y(t) - z(t)\| \leq e^{Lm(I)} \max_{t\in[t_0,t]} (\|\eta\| + \int_{t_0}^{t} \|\delta(s)\|\, ds), \tag{I.4}$$

donde $m(I)$ denota la longitud del intervalo I y L la constante de Lipschitz del problema. Sin embargo, esta cota de estabilidad puede carecer de valor práctico para aquellos problemas diferenciales en los que la constante $Lm(I)$ es extremadamente grande. En este punto surge de modo natural la clase de sistemas diferenciales de tipo *stiff*, o rígidos, y que viene caracterizada por el hecho de que $Lm(I) >> 1$ para los problemas de la clase a pesar de que las órbitas de valores iniciales próximos permanecen próximas a lo largo del intervalo de integración, esto es,

$$\|y(t; t_0, y_0) - y(t; t_0, z_0)\| \leq \kappa \|y_0 - z_0\|, \quad \forall t \in I, \quad t \geq t_0,$$

siendo κ una constante de tamaño moderado. Aquí $y(t; t_0, y_0)$ denota la órbita del sistema diferencial en consideración que parte en tiempo t_0 del valor y_0.

Los sistemas diferenciales de tipo stiff han sido objeto de riguroso estudio en las últimas décadas y los métodos numéricos más comunes para la integración efectiva de este tipo de problemas han sido los métodos BDF (Backward Differentiation Formulae) así como los métodos de tipo Runge-Kutta implícitos. Sin embargo, aunque el significado práctico e intuitivo de sistema diferencial stiff es bien conocido, carecemos aún hoy día de una definición matemática precisa para este concepto. Curtiss & Hirschfelder [29] fueron pioneros en destacar que la resolución numérica eficiente de ciertos sistemas diferenciales ordinarios requería el uso de tipos especiales de métodos numéricos. Esto constituye el origen del concepto de *stiffness*, y desde entonces es común entre los analistas numéricos la distinción entre problemas de tipo stiff y problemas de tipo no stiff. Sin embargo, la frontera que delimita ambas clases de problemas stiff y no stiff no puede ser establecida de modo riguroso [33, 68]; más aún, se pueden dar ejemplos de problemas de valor inicial que presentan características no stiff en ciertas regiones del intervalo de integración y características stiff en la región complementaria. En otras palabras, aunque hablemos aquí de sistemas diferenciales stiff, una ecuación por sí misma no es stiff, sino que adquiere esta condición para determinados valores iniciales, y este es el principal inconveniente que impide obtener una definición satisfactoria de stiffness.

En palabras de Curtiss & Hirschfelder: *las ecuaciones de tipo stiff son ecuaciones donde ciertos métodos implícitos, en particular los métodos BDF, realizan una mejor integración, normalmente bastante mejor, que los métodos explícitos.* No obstante, esta apreciación no es del todo satisfactoria por cuanto los métodos implícitos también son preferibles sobre los explícitos en problemas de tipo oscilatorio, y este tipo de problemas no se engloba dentro de la clase stiff. De modo general, la stiffness de un sistema diferencial ordinario viene determinada en la práctica por los autovalores de la matriz Jacobiana $\partial f/\partial y$ del dato f que define el sistema. A grosso modo, un sistema diferencial ordinario de tipo stiff suele verificar que la matriz Jacobiana cumple las siguientes condiciones [33]

i) existen autovalores λ_i tales que Re $\lambda_i << 0$, y autovalores μ_j de tamaño moderado, tales que $|\mu_j/\lambda_i|$ es pequeño;

ii) si λ es autovalor tal que Im $\lambda >> 1$ entonces Re $\lambda << 0$;

iii) no hay autovalores λ tales que Re $\lambda >> 1$.

Bajo estas condiciones, aunque no se excluye la existencia de soluciones inestables, las soluciones del sistema diferencial (I.1) poseen componentes estacionarias y transitorias en intervalos temporales de tamaño moderado, de modo que las componentes transitorias (que corresponden a autovalores de módulo grande y parte real negativa) están dominadas por las componentes estacionarias y tienden rápidamente a ellas. Además la condición *ii*) evita la existencia de soluciones altamente oscilatorias. Así, estos sistemas superestables muestran un comportamiento muy favorable para evitar la propagación de errores, pero esto no ocurre con las órbitas numéricas si se efectúa una integración por medio de métodos explícitos.

A pesar de la regularidad de las soluciones, la principal razón por la que los métodos explícitos no son adecuados para la integración de problemas stiff radica en que estos métodos poseen un dominio de estabilidad acotado, lo cual fuerza al método a tomar pasos temporales excesivamente pequeños para poder concluir la integración con un mínimo de precisión. De este modo, es la estabilidad de los métodos numéricos, y no su precisión, lo que restringe el tamaño de paso en la integración de sistemas stiff. Todo esto condujo a Dahlquist [30] a introducir el concepto de A-estabilidad para los métodos numéricos, de tal modo que un método numérico se dirá A-estable si provee integraciones estables sobre toda la clase de problemas lineales $y' = Jy$, donde la matriz J posee todos sus autovalores con parte real no positiva. En otras palabras los métodos A-estables son aquellos cuyo dominio de estabilidad lineal incluye a todo el semiplano complejo negativo.

Con vistas a generalizar los estudios de estabilidad lineal de los métodos numéricos a problemas no lineales, Dahlquist [31] propone en 1975 la clase de sistemas diferenciales stiff con constantes de Lipschitz laterales, esto es, sistemas diferenciales satisfaciendo una condición del tipo

$$\langle f(t,y) - f(t,z), y - z\rangle \leq \nu \left\|y - z\right\|^2, \quad \forall t \in I,\ \forall y, z \in \mathcal{D},\ \nu \text{ constante}, \tag{I.5}$$

con respecto a un determinado producto interior en $\mathbb{R}^m$. Notemos que si f posee una constante de Lipschitz clásica L (I.3), entonces L es también una constante de Lipschitz lateral en el sentido de (I.5). Sin embargo, el recíproco no es cierto en general, y además el uso de constantes de Lipschitz laterales permite obtener una cota para la diferencia entre dos soluciones de (I.1) más precisa que la obtenida a través de la constante de Lipschitz clásica (I.4), esto es

$$\left\|y(t) - z(t)\right\| \leq e^{\nu(t-t_0)} \left\|y(t_0) - z(t_0)\right\|, \quad \forall t \geq t_0$$

siendo $y(t),\ z(t)$ soluciones de (I.1).

De este modo, Dahlquist establece la definición de G-estabilidad, o estabilidad no lineal, para los métodos lineales multipaso aplicados a sistemas diferenciales que satisfacen una condición de Lipschitz lateral, y da una serie de resultados preliminares acerca de la estabilidad de esta familia de métodos. En el mismo año, Butcher [10], influenciado por las ideas de Dahlquist, extiende la idea de estabilidad no lineal a la clase de métodos de Runge-Kutta introduciendo el concepto de B-estabilidad. No será hasta finales de la década de los 80, cuando Butcher y Burrage [6, 7, 8, 11, 12] logran establecer la unificación de los estudios de estabilidad no lineal en el marco de los métodos lineales generales aplicados a sistemas diferenciales stiff satisfaciendo una condición de Lipschitz lateral.

No obstante, muchos de los métodos numéricos anteriormente citados integran satisfactoriamente sobre amplios intervalos de tiempo problemas diferenciales de tipo stiff en principio no cubiertos por la teoría lineal y no lineal de estabilidad y convergencia. En este sentido, los sistemas diferenciales autónomos (I.2) conducentes en el tiempo a *equilibrios semiestables* pueden ser integrados eficientemente por algunos de los métodos arriba mencionados, con secuencias de paso extraordinariamente grandes. Acuñamos aquí el término *equilibrio semiestable* para referirnos a un punto de equilibrio inestable en el sentido de Lyapunov pero que posee la propiedad de ser atractor para una variedad de órbitas de una cierta vecindad de la que éste es punto frontera, pero no es atractor para las órbitas de la vecindad complementaria. De esta manera, introduciremos en los capítulos II, III y IV una clase particular de sistemas diferenciales autónomos stiff con la presencia de un tipo particular de equilibrios semiestables, y dedicaremos especial atención al estudio de estabilidad de los principales métodos de un paso cuando son aplicados a esta clase de problemas diferenciales.

Debemos recordar que dado un sistema diferencial autónomo (I.2), un punto $\overline{y} \in \mathbb{R}^m$ se dice dice *punto de equilibrio* para f si $f(\overline{y}) = 0$. Es de interés entonces caracterizar el comportamiento de las órbitas del sistema diferencial (I.2) en entornos de la *solución estacionaria* $\overline{y}$. La estabilidad de las órbitas de un tal sistema puede describirse estudiando el sistema linealizado asociado $y' = Df(\overline{y})y$, siendo $Df(\overline{y})$ la matriz Jacobiana del dato f en el equilibrio. De hecho es bien conocido que si $Df(\overline{y})$ posee todos sus autovalores con parte real negativa entonces el equilibrio es asintóticamente estable, mientras que la existencia de al menos un autovalor con parte real positiva implica inestabilidad [49, 55, 88]. Así pues, la linealización de sistemas diferenciales autónomos en entornos de un punto de equilibrio de tipo hiperbólico, esto es, un punto de equilibrio para el que la matriz Jacobiana no posee autovalores con parte real nula,

es suficiente para determinar el tipo de estabilidad del punto de equilibrio.

Notemos que para un sistema diferencial autónomo (I.2) una *variedad localmente positivamente (resp. negativamente) invariante* es un conjunto $S \subset \mathbb{R}^m$ no vacío tal que si $y_0 \in S$, entonces la solución $y(t)$ de (I.2), tal que $y(0) = y_0$, está totalmente contenida en S para $0 \leq t \leq T$ (resp. $T \leq t \leq 0$). Si $T = +\infty$ (resp. $T = -\infty$) diremos que la variedad es positivamente (resp. negativamente) invariante. De este modo, para un sistema diferencial autónomo con un punto de equilibrio de tipo hiperbólico existen una variedad positivamente invariante por el flujo ϕ_t del sistema diferencial $W^s_{loc}(\overline{y})$, denominada *variedad estable local*, y una variedad negativamente invariante $W^u_{loc}(\overline{y})$, denominada *variedad inestable local*, ambas con la misma regularidad que el dato f, definidas por

$$\begin{aligned} W^s_{loc}(\overline{y}) &= \{y \in U / \phi_t(y) \to \overline{y}, t \to +\infty,\ \phi_t(y) \in U, t \geq 0\}, \\ W^u_{loc}(\overline{y}) &= \{y \in U / \phi_t(y) \to \overline{y}, t \to -\infty,\ \phi_t(y) \in U, t \leq 0\}, \end{aligned}$$

siendo $U \subset \mathbb{R}^m$ un entorno del punto de equilibrio $\overline{y}$, y cuyas dimensiones respectivas n_s y $n_u = m - n_s$ coinciden con las dimensiones de los espacios vectoriales E^s y E^u generados por los autovectores generalizados de $Df(\overline{y})$ asociados a los autovalores con parte reales negativas y positivas, respectivamente. Además las variedades $W^s_{loc}(\overline{y})$ y $W^u_{loc}(\overline{y})$ son tangentes en el punto $\overline{y}$ a los espacios E^s y E^u, respectivamente. Este resultado es conocido en la literatura como *teorema de la variedad estable* [24, 49, 74, 78, 88].

Como consecuencia del teorema de la variedad estable, se tiene el siguiente comportamiento de estabilidad o inestabilidad para las órbitas locales alrededor del equilibrio. Dado un punto de equilibrio $\overline{y}$ para (I.2) tal que la matriz Jacobiana del sistema en el equilibrio posee autovalores con Re $\lambda_j < \alpha < 0$, para $j = 1, \ldots, n_s$, y Re $\mu_k > \beta > 0$, para $k = 1, \ldots, n_u$, y donde $W^s_{loc}(\overline{y})$ y $W^u_{loc}(\overline{y})$ denotan las variedades estable e inestable asociadas al equilibrio $\overline{y}$, entonces para cada $\varepsilon > 0$ existe $\delta > 0$ tal que si $y_0 \in W^s_{loc}(\overline{y})$, $\|y_0 - \overline{y}\| < \varepsilon$, se tiene que

$$\|\phi_t(y_0) - \overline{y}\| \leq \varepsilon e^{\alpha t}, \quad \text{para todo t} \geq 0,$$

mientras que si $y_0 \in W^u_{loc}(\overline{y})$, $\|y_0 - \overline{y}\| < \varepsilon$, entonces se tiene que

$$\|\phi_t(y_0) - \overline{y}\| \leq \varepsilon e^{\beta t}, \quad \text{para todo t} \leq 0,$$

donde $\phi_t(\cdot)$ denota el flujo del sistema diferencial.

De modo más general tenemos el siguiente

Teorema I.0.5 [Teorema de la Variedad Centro] [24, 49, 74, 88] *Sea $\overline{y}$ un punto de equilibrio para el sistema diferencial (I.2) con dato f de clase C^r, y consideremos la descomposición $\sigma[Df(\overline{y})] = \sigma_s \cup \sigma_c \cup \sigma_u$, donde $\sigma[Df(\overline{y})]$ denota el espectro de la matriz $Df(\overline{y})$, y σ_s, σ_c, σ_u son los conjuntos formados por los autovalores de $Df(\overline{y})$ que poseen parte real negativa, nula y positiva, respectivamente. Entonces existen variedades invariantes por el flujo del sistema W^s, W^u y W^c de clase C^r tangentes en el punto $\overline{y}$ a los espacios vectoriales E^s, E^u y E^c generados por los autovectores generalizados de σ_s, σ_u y σ_c, respectivamente.*

Las variedades W^s y W^u, denominadas variedades estable e inestable, respectivamente, son únicas verificando la condición de tangencia a los espacios de autovectores generalizados. Sin embargo, la variedad W^c, denominada *variedad centro*, tangente al espacio de autovectores generalizados asociados a los autovalores de $Df(\overline{y})$ con parte real nula, no tiene porqué ser única.

La relevancia del estudio de las variedades centro asociada a un punto de equilibrio de un sistema diferencial autónomo (I.2) dado radica en que hace posible el estudio de estabilidad del equilibrio por medio de un sistema diferencial asociado de dimensión reducida. Más concretamente, consideremos, sin pérdida de generalidad, un sistema diferencial autónomo (I.2) con equilibrio $\overline{y} = 0 \in \mathbb{R}^m$, de modo que la variedad inestable es vacía, esto es, la matriz Jacobiana $Df(0)$ posee únicamente autovalores con parte real no positiva, y sean n_s y n_c, $n_s + n_c = m$, las dimensiones de los espacios vectoriales E^s y E^c generados por los autovectores generalizados de $Df(0)$ asociados a los autovalores con parte real negativa y nula, respectivamente. Entonces, ya que una variedad centro es tangente en el equilibrio al espacio E^c podemos representarla localmente como el grafo de una aplicación diferenciable C^r de modo que

$$W^c = \{(y_1, y_2) \in \mathbb{R}^{n_c} \times \mathbb{R}^{n_s} / \ \ y_2 = h(y_1)\} \quad h(0) = 0, \ Dh(0) = 0,$$

donde $h : U_0 \subset \mathbb{R}^{n_c} \to \mathbb{R}^{n_s}$ está definida en algún entorno U_0 del origen en $\mathbb{R}^{n_c}$. Asumamos además que la parte lineal del sistema (I.2) ha sido descompuesta en una matriz diagonal por bloques, esto es,

$$\begin{array}{rcll} y_1' & = & Ay_1 + f_1(y_1, y_2), & \\ y_2' & = & By_2 + f_2(y_1, y_2), & (y_1, y_2) \in \mathbb{R}^{n_c} \times \mathbb{R}^{n_s} \end{array} \tag{I.6}$$

donde $A \in \mathbb{R}^{n_c, n_c}$ tiene todos sus autovalores con parte real nula, $B \in \mathbb{R}^{n_s, n_s}$ tiene todos sus autovalores con parte real negativa, mientras que f_1, f_2 y sus derivadas parciales de primer orden se anulan en el origen. Notemos que si f_1 y f_2 son idénticamente nulas entonces las soluciones de (I.6) tienden exponencialmente rápido a las soluciones del sistema lineal $y_1' = Ay_1$. Más

generalmente, bajo las condiciones anteriores y considerando la proyección del sistema (I.6) sobre la variedad centro $y_2 = h(y_1)$ obtenemos la *ecuación reducida a la variedad centro*

$$y_1' = Ay_1 + f_1(y_1, h(y_1)), \quad y_1 \in \mathbb{R}^{n_c}. \tag{I.7}$$

El siguiente teorema describe la estabilidad del equilibrio $(y_1, y_2) = (0, 0)$ del sistema (I.6) en términos de la estabilidad del equilibrio $y_1 = 0$ para la ecuación reducida a la variedad centro (I.7).

Teorema I.0.6 [24, 49, 88] *Si la solución nula del sistema (I.7) es estable (asintóticamente estable) (inestable), entonces la solución nula de (I.6) es estable (asintóticamente estable) (inestable).*

Si la solución nula de (I.7) es estable, y $(y_1(t), y_2(t))$ denota la solución de (I.6) que parte de $(y_1(0), y_2(0))$ suficientemente pequeño, entonces existe una solución $u(t)$ del sistema (I.7) tal que

$$\begin{aligned} y_1(t) &= u(t) + \mathcal{O}(e^{-\gamma t}) \\ y_2(t) &= h(u(t)) + \mathcal{O}(e^{-\gamma t}) \end{aligned}$$

para $t \to +\infty$, siendo $\gamma > 0$ una constante.

Con esto, el sistema diferencial sobre la variedad centro determina el comportamiento asintótico de las soluciones del sistema inicial modulo términos exponenciales, esto es, las soluciones locales de (I.6) en entornos del origen decaen exponencialmente rápido a la variedad centro del sistema. En otro orden de cosas, notemos que derivando en el tiempo la expresión $y_2 = h(y_1)$ obtenemos a partir de (I.6) la ecuación en derivadas parciales

$$\mathcal{N}(h(y_1)) \equiv Dh(y_1)[Ay_1 + f_1(y_1, h(y_1))] - Bh(y_1) - f_2(y_1, h(y_1)) = 0,$$

junto con las condiciones $h(0) = 0$ y $Dh(0) = 0$. Esta ecuación, que se conoce como *ecuación de la variedad centro*, es en general más complicada de resolver que el sistema original (I.6); no obstante, su solución puede ser aproximada por polinomios de Taylor en el origen hasta el grado de precisión deseado. Más precisamente, tenemos el siguiente

Teorema I.0.7 *Sea $\phi : \mathbb{R}^{n_c} \to \mathbb{R}^{n_s}$ una función C^1 en un entorno del origen, con $\phi(0) = 0$ y $D\phi(0) = 0$, tal que $\mathcal{N}(\phi(y_1)) = \mathcal{O}(|y_1|^q)$, $y_1 \to 0$, para cierto $q > 1$. Entonces*

$$h(y_1) = \phi(y_1) + \mathcal{O}(|y_1|^q), \ y_1 \to 0.$$

Este resultado nos da un método de cómputo del desarrollo de Taylor de la función que define localmente la variedad centro. Notemos entonces que aunque la variedad centro pueda

no ser única, en virtud de este resultado todas las posibles variedades centro diferenciables en el equilibrio admiten el mismo desarrollo en serie de Taylor. Sin embargo, tal desarrollo en serie no siempre existe por cuanto W^c puede no ser analítica en el equilibrio.

En relación a la discretización numérica a través de métodos de un paso de este tipo de sistemas diferenciales (véase, por ejemplo, [85]), Beyn y Lorenz [2] estudian la cuestión de la aproximación de la variedad centro del sistema continuo por medio de la variedad centro que posee el sistema discreto. Por otro lado, Kloeden y Lorenz en [61] estudian las propiedades de estabilidad de un conjunto atractor asociado a un sistema dinámico discreto (obtenido por discretización con un método de un paso) en entornos de un conjunto atractor del sistema dinámico continuo. Estos tratamientos conectan con los estudios de estabilidad de los métodos numéricos que llevaremos a cabo en los capítulos II, III y IV de esta memoria. Dentro de la clase de sistemas diferenciales stiff consideraremos la integración numérica de sistemas diferenciales autónomos con un punto de equilibrio inestable para el cual existe una variedad centro unidimensional que rige el comportamiento de las órbitas estables e inestables en entornos del equilibrio. Interesantes y clásicos problemas de la literatura stiff poseen puntos de equilibrio que presentan este tipo de dinámica local [37, 38, 53, 76]. Es nuestro interés analizar condiciones prácticas sobre los métodos numéricos que permitan asegurar la conservación de las órbitas locales estables del problema.

La teoría de las variedades centro constituye una técnica fundamental para el análisis de las propiedades cualitativas del flujo de un sistema diferencial en entornos de un punto de equilibrio (no hiperbólico), de tal modo que el comportamiento del flujo del sistema en entornos de un equilibrio depende de su comportamiento sobre una variedad centro. Otra técnica de reducción está basada en la teoría de *formas normales*, cuyos orígenes se remontan a la tesis doctoral de H. Poincaré [75]. El método de las formas normales consiste en determinar sucesivas transformaciones de coordenadas que permitan simplificar *tanto como sea posible* la expresión analítica del campo vectorial del sistema diferencial reducido a la variedad centro [49, 64, 88]. Así, a través de sucesivos cambios de coordenadas analíticos se obtiene un sistema diferencial reducido *en forma normal* topológicamente equivalente al sistema sobre la variedad centro, con la ventaja de que para el estudio cualitativo de las órbitas sólo son considerados los términos esenciales (o *términos de resonancia*) del desarrollo de Taylor del campo vectorial reducido a la variedad centro. Aunque no es nuestra intención entrar en detalles acerca del proceso de determinación de formas normales de campos vectoriales, debemos señalar que recientemente

han sido llevados a cabo diversos estudios para la obtención de algoritmos de determinación y cómputo de formas normales, véase, por ejemplo, [35, 46, 47, 65, 66, 67].

Estas técnicas de reducción de sistemas diferenciales poseen un carácter no lineal, y puesto que en general los métodos numéricos no preservan cambios de variable no lineales, nos vemos obligados a tratar el análisis de estabilidad de los métodos a través del sistema original introduciendo un marco hipotético adecuado para llevar a cabo dicho análisis. No obstante, las técnicas de reducción de sistemas diferenciales constituyen la idea básica que subyace en el marco hipotético que introduciremos en el segundo capítulo de esta memoria para los sistemas diferenciales que serán considerados a posteriori en los capítulos II, III y IV.

Con la idea de una futura generalización del estudio de estabilidad de los métodos de un paso a métodos multipaso aplicados a los sistemas diferenciales con equilibrios semiestables, presentamos en el capítulo V de esta memoria una serie de resultados que describen el comportamiento contractivo así como la convergencia en intervalos semi-infinitos de la amplia clase métodos lineales generales sobre los sistemas diferenciales con constantes de Lipschitz laterales negativas. Aunque, los resultados de contractividad que damos en el capítulo V se refieren a una integración a paso fijo de este tipo de problemas, debemos señalar que, en el caso de integraciones a paso variable, tan sólo conocemos resultados más débiles relativos a la A_0-estabilidad (o estabilidad sobre el eje real negativo) de los métodos (véase, por ejemplo, [4, 13, 16, 20, 21, 25, 48, 79]).

En otro orden de cosas, el avance en el estudio y tratamiento de ciertos problemas prácticos ha generado nuevos requerimientos en los métodos de resolución numérica, lo que conlleva la necesidad de desarrollar nuevos algoritmos de integración con propiedades específicas. Así, en las dos últimas décadas los analistas numéricos han manifestado un creciente interés en el campo de la integración numérica geométrica [23, 50]. De esta manera, los integradores numéricos, además de poseer buenas propiedades de aproximación y estabilidad, deben reproducir las características geométricas presentadas por el flujo del problema diferencial, [22, 23, 34, 50, 60, 70, 80]. En este sentido, muchos problemas de interés práctico requieren una modelización no sólo basada en un sistema de ecuaciones diferenciales ordinarias sino también en restricciones algebraicas. Este tipo de restricciones incluyen leyes de conservación cuadráticas o con carácter no lineal general, como por ejemplo conservación de masa, energía o momentos, u otras restricciones de carácter geométrico. Estas propiedades del espacio de fases asociado al sistema diferencial permiten obtener una visión más acertada del comportamiento global y la estructura geométrica

de las soluciones. Así, por medio de la conservación numérica de las diferentes propiedades geométricas de las soluciones se obtiene no sólo un mejor comportamiento cualitativo para las soluciones numéricas sino que en añadidura se consigue una integración más precisa a largo tiempo.

Aunque este tipo de ecuaciones pueden ser tratadas por medio de la teoría correspondiente a los sistemas diferenciales algebraicos [5, 53], nosotros nos interesaremos en el capítulo final de esta memoria por la adaptación de métodos de tipo Runge-Kutta explícitos para la conservación de tales restricciones algebraicas. Es bien conocido que ningún método de Runge-Kutta explícito puede conservar por sí mismo invariantes cuadráticos [27, 34, 50]. De hecho, un método Runge-Kutta implícito irreducible conserva cualquier invariante cuadrático si y sólo si el método es *simpléctico* [27, 50]. Más aún, no existen métodos Runge-Kutta implícitos capaces de conservar todos los invariantes polinomiales de grado mayor o igual a tres. Sin embargo, cuando para un problema diferencial se conoce un invariante, es posible llevar a cabo una integración efectiva combinando métodos explícitos con técnicas de proyección ortogonal sobre la variedad que definen los invariantes del sistema diferencial. Sin embargo, mientras que un método de tipo Runge-Kutta preserva por sí mismo cualquier invariante afín, el proceso de proyección ortogonal destruye la conservación de estos invariantes. Además, la proyección ortogonal de un método de tipo Runge-Kutta requiere para su implementación el conocimiento explícito del gradiente de la función que define los invariantes. En este sentido, diseñamos en el capítulo VI de esta memoria una técnica de proyección, compatible con los códigos Runge-Kutta a paso variable, basada en pares encajados de métodos explícitos que hace posible la conservación de cualquier invariante multidimensional no lineal general sin destruir los invariantes afines del problema y que no requiere conocimiento alguno acerca del gradiente para su implementación.

Capítulo II

Métodos numéricos para sistemas diferenciales con equilibrios semiestables.

II.1. Consideraciones preliminares

Muchos problemas de tipo stiff, tales como el *problema de Robertson*, [53, p. 3, p. 144], [76], o la reacción química conocida como *E5*, [38], [53, p. 145], se consideran frecuentemente como problemas modelos para estudiar la eficiencia y el comportamiento de los métodos numéricos. Muchos integradores numéricos implícitos presentan un comportamiento totalmente inestable y poco realista cuando este tipo de problemas son integrados en intervalos temporales extremadamente amplios con las estrategias usuales de selección de paso variable. De este modo, para evitar este comportamiento anómalo y evitar el cómputo de soluciones numéricas alejadas de la solución exacta, los métodos se ven forzados a tomar un elevado número de pasos temporales para poder concluir la integración, lo cual redunda en una integración poco eficiente desde un punto de vista computacional. Más aún, muchos de estos métodos pueden resultar incapaces de concluir la integración para un problema dado. Véase [53, p. 144-145] y las referencias allí indicadas.

Una dificultad a la que los métodos deben hacer frente en la integración es la posible presencia de puntos de equilibrio *semiestables*, esto es, puntos de equilibrio para los que en todo entorno existen tanto órbitas estables como inestables. Esto conlleva a que las soluciones

numéricas que parten de la zona estable puedan atravesar la zona inestable ante la presencia de pequeños errores en el cómputo y llevar a cabo una integración cualitativamente deficiente. De este modo, resulta que algunos métodos con buenas propiedades de convergencia y estabilidad, como los métodos de la familia Gauss (que poseen orden de convergencia maximal, y tienen excelentes propiedades de estabilidad lineal y no lineal), no garantizan una integración eficiente frente a este tipo de problemas con equilibrios semiestables sobre grandes intervalos temporales.

Por otro lado, desde un punto de vista práctico y computacional, es un hecho bien conocido que los métodos de la familia Radau IIA o los métodos BDF (Backward Differentiation Formulae) de uno y dos pasos integran satisfactoriamente esta clase de problemas, a pesar de no ser *incondicionalmente estables*, en el sentido de que ciertas condiciones sobre la selección del tamaño de paso deben ser impuestas para garantizar una integración estable. En el siguiente capítulo probaremos que, esencialmente, cualquier método de un paso *fuertemente A-estable* puede integrar satisfactoriamente este tipo de problemas sobre ciertas redes temporales verificando que las razones de paso están acotadas por alguna constante mayor que la unidad.

En este capítulo dedicaremos nuestra atención al estudio de métodos de interés práctico que resulten ser incondicionalmente estables sobre una clase de sistemas diferenciales con equilibrios semiestables. De hecho probaremos que el método de Euler implícito provee integraciones estables sin restricción alguna sobre los tamaños de paso que se consideren en la integración; asimismo ilustraremos con diversos ejemplos cómo otros métodos de gran interés práctico no verifican tal propiedad de estabilidad incondicional. Debemos señalar que el contenido de este capítulo constituye una revisión de los resultados teóricos y numéricos incluidos en el trabajo [40].

A continuación especificamos la clase de problemas a tratar en los capítulos II, III y IV de esta memoria. Consideremos el caso de un punto de equilibrio semiestable, pongamos $y = 0$ (esto no supone pérdida de generalidad por cuanto un punto de equilibrio $y = y_0$ se lleva al origen por medio de una traslación sin que ello afecte a la stiffness del problema en consideración), para sistemas diferenciales autónomos de la forma

$$y'(t) = f(y), \quad t \in [0, \infty),\ y, f(y) \in \mathbb{R}^m,\ m \geq 2. \tag{II.1}$$

En el resto del capítulo se entenderá que f es una función al menos tres veces continuamente diferenciable en un pequeño entorno del origen $y = 0$ de radio $\eta > 0$ cumpliendo que $f(0) = 0$. Si denotamos por J a la matriz Jacobiana de f en el origen, $J = f'(0)$, tenemos que $f(y)$ puede

ser expresada como

$$f(y) := Jy + N(y) + R(y), \quad N(y) = \frac{1}{2}f''(0)[y, y], \tag{II.2}$$

$$R(y) = \frac{1}{2}\left(\int_0^1 (1-s)^2 f'''(sy)[y, y, y]ds\right), \tag{II.3}$$

para $\|y\| \leq \eta$, donde $f''(z)[\cdot,\cdot]$ y $f'''(z)[\cdot,\cdot,\cdot]$ representan la segunda y tercera derivadas de Frèchet de f en el punto z, respectivamente.

Asumiremos que el sistema (II.1)-(II.3) satisface

(A1) $\lambda = 0$ es un autovalor simple de $J = f'(0)$ y los restantes autovalores de J pertenecen al semiplano complejo negativo $\mathbb{C}^- := \{z \in \mathbb{C} \ : \ \mathrm{Re}(z) < 0\}$,

y

(A2) la constante

$$\vartheta := \frac{p^T N(v)}{p^T v} \neq 0,$$

donde v y p denotan, respectivamente, un autovector derecho e izquierdo de J asociados al autovalor nulo.

Según la teoría de las variedades centro (ver, por ejemplo, [24, 46, 49, 88]), estas suposiciones garantizan la existencia de un conjunto conexo m-dimensional Γ, con $y = 0$ como punto frontera, tal que para cada $y_0 \in \Gamma$ la órbita positiva $y(t; 0, y_0)$ de (II.1) permanece en Γ; más aún

$$\lim_{t\to\infty} y(t; 0, y_0) = 0,$$

y la órbita es tangente en el origen al subespacio centro $E^c := span\{v\} = \{\mu v, \quad \mu \in \mathbb{R}\}$.

Usando la técnica de reducción de formas normales descrita en [46, p. 61-62],[49, p. 138-145], el sistema diferencial (II.1) puede ser reducido a

$$\begin{aligned} x_1' &= \vartheta x_1^2 + \mathcal{O}(\|x\|^3), \\ x_j' &= \sum_{k=2}^{m} l_{jk}x_k + \mathcal{O}(\|x\|^2), \quad j = 2, \ldots, m. \end{aligned} \tag{II.4}$$

Arriba J se considera semejante a la matriz diagonal por bloques $\bar{L} := \text{Block-Diag}(0, L)$, con $L := (l_{jk})_{j,k=2}^m$.

Ambos problemas diferenciales (II.1) y (II.4) presentan la misma dinámica en entornos del equilibrio $y = 0$ incluso después de suprimir el término $\mathcal{O}(\|x\|^3)$ en la primera componente de

(II.4) y los términos $\mathcal{O}(\|x\|^2)$ en las restantes componentes. Sin embargo esta última propiedad no se aplica a la dinámica de las soluciones propuestas por los métodos numéricos, por cuanto éstos no preservan cambios de variable distintos de los lineales. En consecuencia, aunque desde un punto de vista teórico, existan técnicas de carácter no lineal, como la teoría de las variedades centro o la teoría de las formas normales, para reducir el estudio de la dinámica de problemas diferenciales con puntos de equilibrio a un menor número de casos, estamos forzados a trabajar con el modelo original (II.1) con vistas a estudiar la estabilidad de los métodos numéricos sobre tales problemas.

En otro orden de cosas, aunque pudiera parecer que el estudio de las órbitas del problema test simple

$$\begin{aligned} y_1' &= \vartheta y_1^2, \quad \vartheta \neq 0, \\ y_2' &= \lambda y_2, \quad \lambda \in \mathbb{C}^-, \end{aligned} \tag{II.5}$$

es suficiente de cara a estudiar estabilidad de los métodos sobre la totalidad de problemas que conforman la clase delimitada por las hipótesis $(A1) - (A2)$, esto no es en absoluto cierto tal como se verá en la sección final de este capítulo.

El resto del capítulo se organiza del modo siguiente. En la sección segunda introduciremos el marco de las H-hipótesis como alternativa equivalente a las A-hipótesis (esto es, *(A1)-(A2)*) y más adecuada para el análisis de estabilidad de los métodos. En la sección 3 estudiaremos la dinámica local del sistema diferencial (II.1) alrededor del equilibrio. A continuación, en la sección cuarta de este capítulo, se probará la estabilidad incondicional del método de Euler implícito, para concluir en la sección quinta con el análisis de estabilidad para otros métodos de interés práctico.

II.2. H-hipótesis

Con vistas a simplificar el análisis de estabilidad de los métodos sobre la clase de problemas delimitados por las A-hipótesis introducimos en esta sección las siguientes *H-hipótesis*.

H-*Hipótesis*

Existe un producto interior $\langle \cdot, \cdot \rangle$ en $\mathbb{R}^m$, con su correspondiente norma inducida $\|u\| = \langle u, u \rangle^{1/2}$, y un vector unitario v satisfaciendo

(H1) $Jv = 0$.

(H2) $Jw \in v^{\perp}, \quad \forall w \in v^{\perp} := \{u \in \mathbb{R}^m : \langle u, v\rangle = 0\}$.

(H3) $\sup\limits_{\{w\in v^{\perp},\, w\neq 0\}} \dfrac{\langle w, Jw\rangle}{\|w\|^2} = -\delta_1 < 0.$

(H4)

$$\sup_{x\neq 0} \frac{\|N(x)\|}{\|x\|^2} = \delta_2 > 0 \quad \text{y} \quad \sup_{0<\|x\|\leq\eta} \frac{\|R(x)\|}{\|x\|^3} = \delta_3 \geq 0.$$

(H5) $\langle v, N(v)\rangle = -\nu < 0$.

Aunque *(H4)* es algo redundante teniendo en cuenta la regularidad de la función f que define el sistema (II.1), es conveniente mantener esta hipótesis por cuanto gran parte de las constantes involucradas en los desarrollos sucesivos dependerán única y exclusivamente de las constantes que aparecen en el marco hipotético recién introducido. Por consiguiente, los resultados que se derivarán en lo que sigue no se verán afectados por la *stiffness* de la matriz Jacobiana J ni por la dimensión m del sistema diferencial.

Nota II.2.1

1. Téngase en cuenta que la hipótesis *(H5)* puede ser reemplazada por

   $$\langle v, N(v)\rangle \neq 0,$$

   ya que de esta última suposición si $\langle v, N(v)\rangle > 0$, entonces tomando $-v$ en lugar de v, se obtiene *(H5)*, sin afectar a las restantes H-hipótesis.

2. Bajo las H-hipótesis, tenemos que

   $$0 < \nu \leq \delta_2.$$

   En efecto, esta propiedad se deduce teniendo en cuenta que

   $$\nu = \langle -v, N(-v)\rangle \leq \|N(v)\| \leq \delta_2.$$

Para probar la equivalencia entre las A-hipótesis y las H-hipótesis haremos uso de los dos siguientes lemas.

Lema II.2.2 *Sea J una matriz verificando $(A1)$, siendo p y v autovectores izquierdo y derecho, respectivamente, asociados al autovalor nulo de J. Entonces existe un producto interior $\langle\cdot,\cdot\rangle$ en $\mathbb{R}^m$ de modo que $v^{\perp} = \{y \in \mathbb{R}^m \,/\, p^T y = 0\}$. En particular, $Jv^{\perp} \subset v^{\perp}$.*

Demostración. Puesto que $\lambda = 0$ es autovalor simple de la matriz J, existen $m-1$ vectores $v_2, \ldots, v_m \in \mathbb{R}^m$ tales que $\{v, v_2, \ldots, v_m\}$ es un sistema linealmente independiente y

$$JV = V \begin{pmatrix} 0 & 0^T \\ 0 & \Delta \end{pmatrix}, \tag{II.6}$$

siendo $V := [v, v_2, \ldots, v_m]$ y Δ una matriz regular de dimensión $m-1$.

Definiendo el producto interior $\langle x, y\rangle := (V^{-1}x)^T(V^{-1}y)$ sigue que $\langle v, v_j\rangle = e_1^T e_j = 0$, $j = 2, \ldots, m$, siendo $e_1, \ldots, e_m$ los vectores de la base canónica de $\mathbb{R}^m$. En consecuencia, sigue que $v^\perp = span\{v_2, \ldots, v_m\}$ y, a partir de (II.6), $Jv^\perp \subset v^\perp$.

Por otro lado, ya que $p^T J = 0$, sigue de (II.6) que

$$0^T = (p^T v, p^T v_2, \ldots, p^T v_m) \begin{pmatrix} 0 & 0^T \\ 0 & \Delta \end{pmatrix}.$$

Luego debe ser que $p^T v_j = 0,\ j = 2, \ldots, m$, ya que Δ es una matriz regular. Además, $p^T v \neq 0$, por cuanto $p \neq 0$.

En definitiva, para $y = y_1 v + \sum_{j=2}^{m} y_j v_j \in \mathbb{R}^m$, con $y_1, \ldots, y_m$ escalares reales, tenemos que $p^T y = y_1(p^T v) = \langle y, v\rangle(p^T v)$. Esto concluye la prueba. □

Lema II.2.3 *Sea J una matriz verificando $(A1)$, siendo v autovector derecho asociado al autovalor nulo de J. Si existen dos productos interiores $\langle\cdot,\cdot\rangle_a$ y $\langle\cdot,\cdot\rangle_b$, con subespacios ortogonales a v dados respectivamente por $v_a^\perp$ y $v_b^\perp$, tales que $Jv_a^\perp \subset v_a^\perp$ y $Jv_b^\perp \subset v_b^\perp$ entonces*

$$v_a^\perp = Jv_a^\perp = Jv_b^\perp = v_b^\perp.$$

En consecuencia, $\langle v, x\rangle_a \neq 0$ si y sólo si $\langle v, x\rangle_b \neq 0$.

Demostración. Puesto que el subespacio generalizado asociado al autovalor $\lambda = 0$ está dado por $span\{v\}$ (por cuanto $\lambda = 0$ es simple para la matriz J), se tiene que J como aplicación lineal de $v_a^\perp$ en sí mismo es inyectiva (debe quedar claro que $span\{v\} \cap v_a^\perp = \{0\}$). Luego, en virtud del primer teorema de isomorfía se deduce que $dim\ Jv_a^\perp = dim\ v_a^\perp = m-1$ y, por tanto, $v_a^\perp = Jv_a^\perp$. Del mismo modo deducimos que $v_b^\perp = Jv_b^\perp$.

Tomemos ahora $\{v_2^a, \ldots, v_m^a\}$ y $\{v_2^b, \ldots, v_m^b\}$ bases de $v_a^\perp$ y $v_b^\perp$, respectivamente. Puesto que $\{v, v_2^b, \ldots, v_m^b\}$ constituye una base de $\mathbb{R}^m$ podemos expresar

$$v_i^a = \alpha_{i1} v + \sum_{j=2}^{m} \alpha_{ij} v_j^b, \quad \alpha_{ij} \in \mathbb{R},\ 2 \leq i \leq m,\ 1 \leq j \leq m$$

y de aquí se deduce que $Jv_a^{\perp} \subset Jv_b^{\perp} = v_b^{\perp}$. En definitiva, $v_a^{\perp} \subset v_b^{\perp}$.

Un razonamiento análogo permite deducir que $v_b^{\perp} \subset v_a^{\perp}$, lo cual concluye la prueba. □

Lema II.2.4 *Sea J una matriz verificando $(A1)$. Entonces existe una transformación de semejanza real P tal que*

$$P^{-1}JP = \begin{pmatrix} 0 & 0^T \\ 0 & \Delta \end{pmatrix},$$

siendo Δ una matriz real regular de dimensión $m-1$ con norma logarítmica euclídea negativa, $\mu_2[\Delta] := \lambda_{max}(\frac{1}{2}(\Delta + \Delta^T)) < 0$.

Demostración. En efecto, la matriz J es semejante a su descomposición de Jordan real $\begin{pmatrix} 0 & 0^T \\ 0 & \Gamma \end{pmatrix}$, con Γ matriz diagonal por bloques de dimensión $m-1$ de la forma

$$\Gamma = BlockDiag(K_1, \ldots, K_s, L_1, \ldots, L_t),$$

siendo K_i, $1 \le i \le s$, y L_j, $1 \le j \le t$, bloques de Jordan expresados en forma real asociados a los autovalores reales y complejos de la matriz J, respectivamente,

$$K_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{pmatrix}_{k_i \times k_i}, \quad L_j = \begin{pmatrix} A_j & I_2 & & & \\ & A_j & I_2 & & \\ & & \ddots & \ddots & \\ & & & A_j & I_2 \\ & & & & A_j \end{pmatrix}_{2l_j \times 2l_j},$$

para $1 \le i \le s$, $1 \le j \le t$, donde I_2 denota la matriz identidad de orden 2, $A_j := \begin{pmatrix} a_j & b_j \\ -b_j & a_j \end{pmatrix}$; λ_i, $1 \le i \le s$, son los autovalores reales de J, $a_j + ib_j$, $1 \le j \le t$ (aquí i representa la unidad imaginaria $i = \sqrt{-1}$), denotan los autovalores no reales de J; y $k_1 + \ldots + k_s + 2l_1 + \ldots + 2l_t = m-1$.

Teniendo en cuenta que

$$Diag(1, \epsilon, \ldots, \epsilon^{k_i-1})^{-1} K_i Diag(1, \epsilon, \ldots, \epsilon^{k_i-1}) = \begin{pmatrix} \lambda_i & \epsilon & & & \\ & \lambda_i & \epsilon & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & \epsilon \\ & & & & \lambda_i \end{pmatrix}$$

y

$$Diag(1,1,\epsilon,\epsilon,\ldots,\epsilon^{l_j-1},\epsilon^{l_j-1})^{-1}L_jDiag(1,1,\epsilon,\epsilon,\ldots,\epsilon^{l_j-1},\epsilon^{l_j-1}) = \begin{pmatrix} A_j & \epsilon I_2 & & & \\ & A_j & \epsilon I_2 & & \\ & & \ddots & \ddots & \\ & & & A_j & \epsilon I_2 \\ & & & & A_j \end{pmatrix},$$

tenemos que para todo $\epsilon > 0$ existe una transformación de semejanza real P_ϵ tal que

$$P_\epsilon^{-1} J P_\epsilon = \begin{pmatrix} 0 & 0^T \\ 0 & \Gamma_\epsilon \end{pmatrix},$$

siendo Γ_ϵ matriz regular real de dimensión $m-1$ de la forma

$$\Gamma_\epsilon = BlockDiag(\lambda_1,\ldots,\lambda_s,A_1,\ldots,A_t) + \epsilon N,$$

con N matriz nilpotente.

En definitiva, puesto que

$$\frac{1}{2}(\Gamma_\epsilon + \Gamma_\epsilon^T) = Diag(Re\nu,\ \nu \in \sigma[J] \setminus \{0\}) + \frac{\epsilon}{2}(N + N^T),$$

se tiene para la norma logarítmica euclídea de Γ_ϵ, con $\epsilon > 0$ suficientemente pequeño, que

$$\mu_2[\Gamma_\epsilon] = \lambda_{\max}\left(\frac{1}{2}(\Gamma_\epsilon + \Gamma_\epsilon^T)\right) < 0,$$

y se deduce entonces la propiedad anunciada con $\Delta := \Gamma_\epsilon$ y $P := P_\epsilon$. □

Estos lemas previos nos permiten probar ahora de modo sencillo la equivalencia entre ambos marcos hipotéticos introducidos.

Proposición II.2.5 *El sistema diferencial (II.1)-(II.3) satisface las H-hipótesis si y sólo si satisface las A-hipótesis.*

Demostración. "⇒" Supongamos que (II.1)-(II.3) satisface las H-hipótesis. Entonces de $(H3)$ se deduce que $\lambda = 0$ posee multiplicidad geométrica igual a 1 como autovalor de la matriz J, ya que si $u = \alpha v + z \in \mathbb{R}^m \equiv span\{v\} \oplus v^\perp$, con $\alpha \in \mathbb{R}$ y $z \neq 0,\ z \in v^\perp$, fuera autovector de J asociado al autovalor nulo entonces se tendría que $Jz = 0$ y entonces

$$\sup_{\{w \in v^\perp,\, w \neq 0\}} \frac{\langle w, Jw\rangle}{\|w\|^2} \geq 0.$$

En particular, si existiera un vector x tal que $J^2x = 0$, entonces según lo anterior debería ser que $Jx \in \ span\{v\}$. Sin embargo por $(H1)-(H3)$ se tendría que $Jx \in v^\perp$. En consecuencia debe ser que $Jx = 0$, y $x \in \ span\{v\}$. Esto nos dice que la multiplicidad algebraica de $\lambda = 0$ como autovalor de J es igual a 1.

Por otro lado, si $\lambda \in \mathbb{C} \setminus \{0\}$ es autovalor de J y $u \in \mathbb{C}^m$, $u \neq 0$, es autovector asociado, entonces haciendo uso de la factorización $\mathbb{R}^m = span\{v\} \oplus v^\perp$ para expresar

$$\begin{aligned} Re\ u &= \alpha_1 v + w_1, \quad \alpha_1 \in \mathbb{R},\ w_1 \in v^\perp, \\ Im\ u &= \alpha_2 v + w_2, \quad \alpha_2 \in \mathbb{R},\ w_2 \in v^\perp, \end{aligned}$$

concluimos en virtud de $(H2)-(H3)$ que

$$\begin{aligned} 0 &> \langle Jw_1, w_1\rangle + \langle Jw_2, w_2\rangle \\ &= \langle Re(\lambda u), Re\ u\rangle + \langle Im(\lambda u), Im\ u\rangle \\ &= Re\lambda(\|Re\ u\|^2 + \|Im\ u\|^2). \end{aligned}$$

En consecuencia, $Re\lambda < 0$ y queda demostrada $(A1)$.

Finalmente, teniendo en cuenta $(H5)$ y los lemas II.2.2, II.2.3, tenemos que $p^Tv \neq 0$ y $p^TN(v) \neq 0$, y de aquí que (II.1)-(II.3) satisface también $(A2)$.

"$\Leftarrow$" Si (II.1)-(II.3) satisface $(A1)$, entonces según el lema II.2.4 existe una transformación de semejanza P real tal que

$$P^{-1}JP = \begin{pmatrix} 0 & 0^T \\ 0 & \Delta \end{pmatrix} \equiv \overline{\Delta},$$

siendo Δ una matriz real regular de dimensión $m-1$ con norma logarítmica euclídea negativa, $\mu_2[\Delta] := \lambda_{max}(\frac{1}{2}(\Delta + \Delta^T)) < 0$.

En añadidura, podemos suponer, sin pérdida de generalidad, que $Pe_1 = v$. Pongamos además $v_j := Pe_j$, $2 \leq j \leq m$, siendo $e_1, \ldots, e_m$ los vectores de la base canónica de $\mathbb{R}^m$.

Con todo, definiendo $\langle x, y\rangle := (P^{-1}x)^T(P^{-1}y)$ obtenemos que $v^\perp = span\{v_2, \ldots, v_m\} \equiv P \cdot span\{e_2, \ldots, e_m\}$ y $Jv^\perp \subset v^\perp$. En virtud de los lemas II.2.2, II.2.3, concluimos entonces que

$$v^\perp = \{y \in \mathbb{R}^m \ / \ p^Ty = 0\},$$

siendo $p \neq 0$ autovector izquierdo asociado al autovalor nulo (simple) de J. Por lo tanto, teniendo la hipótesis $(A2)$ en cuenta, se tiene que $N(v) \notin v^\perp$ y por tanto se deduce $(H5)$.

Finalmente $(H3)$ queda demostrada teniendo en cuenta que

$$\begin{aligned}\sup_{\{w\in v^{\perp},\, w\neq 0\}} \frac{\langle w, Jw\rangle}{\|w\|^2} &= \sup_{\{w\in\, P\cdot span\{e_2,\ldots,e_m\},\, w\neq 0\}} \frac{(P^{-1}w)^T(P^{-1}Jw)}{(P^{-1}w)^T(P^{-1}w)} \\ &= \sup_{\{y\in\, span\{e_2,\ldots,e_m\},\, y\neq 0\}} \frac{y^T\overline{\Delta}y}{y^Ty} \\ &= \sup_{\{z\in\mathbb{R}^{m-1},\, z\neq 0\}} \frac{z^T\Delta z}{z^Tz} \\ &= \mu_2[\Delta].\end{aligned}$$

□

Para el desarrollo posterior es preciso considerar una aplicación bilineal simétrica $M(x,y)$, que extiende a $N(x)$ en el sentido de que $M(x,x) = N(x)$. Más precisamente, debemos considerar el siguiente

Lema II.2.6 *Dada la aplicación $N(x) = \frac{1}{2}f''(0)[x,x]$ (II.2), existe una única aplicación bilineal y simétrica $M(x,y) : \mathbb{R}^m \times \mathbb{R}^m \longmapsto \mathbb{R}^m$ tal que $M(x,x) = N(x)$, $\forall x \in \mathbb{R}^m$. Esta aplicación viene dada por*

$$M(x,y) := N\left(\frac{x+y}{2}\right) - N\left(\frac{x-y}{2}\right), \tag{II.7}$$

y satisface

1. $$\|M(x,y)\| \leq \delta_2 \|x\| \cdot \|y\|, \quad \forall x,\, y \in \mathbb{R}^m, \tag{II.8}$$

2. $$\|M(x,x) - M(y,y)\| \leq \delta_2 \|x+y\| \cdot \|x-y\|, \quad \forall x,\, y \in \mathbb{R}^m. \tag{II.9}$$

Demostración. En efecto, la aplicación M definida por $M(x,y) = \frac{1}{2}f''(0)[x,y]$ es bilineal y simétrica, y verifica $M(x,x) = N(x)$. Teniendo en cuenta la bilinealidad y simetría del operador derivada segunda, es sencillo comprobar que esta aplicación M así definida coincide con la aplicación dada en (II.7).

Además, si $\tilde{M}$ es otra tal aplicación bilineal y simétrica extendiendo a N, se tiene por un argumento de polarización que

$$N(x+y) = \tilde{M}(x+y, x+y) = N(x) + N(y) + 2\tilde{M}(x,y).$$

De aquí que M y $\tilde{M}$ cumplen la misma identidad, y, en consecuencia, $\tilde{M} = M$.

Por otro lado, ya que la norma considerada está inducida por un producto interior, si x e y son vectores unitarios, sigue de (II.7) que

$$\begin{aligned} \|M(x,y)\| &\leq \left\| N\left(\frac{x+y}{2}\right)\right\| + \left\| N\left(\frac{x-y}{2}\right)\right\| \\ &\leq \delta_2\left(\left\|\frac{x+y}{2}\right\|^2 + \left\|\frac{x-y}{2}\right\|^2 \right) \\ &= \frac{\delta_2}{2}(\|x\|^2 + \|y\|^2) \\ &= \delta_2. \end{aligned}$$

De aquí que para $x,\ y \in \mathbb{R}^m \setminus \{0\}$ tengamos por bilinealidad que

$$M(x,y) = \|x\|\,\|y\|\, M\left(\frac{x}{\|x\|}, \frac{y}{\|y\|}\right),$$

propiedad ésta de la cual deducimos (II.8).

La propiedad (II.9) es consecuencia inmediata de la propiedad (II.8) y de la identidad $M(x,x) - M(y,y) = M(x-y, x+y)$. □

Ejemplos

En este punto verificaremos las suposiciones introducidas previamente sobre algunos problemas interesantes de la literatura stiff que poseen cierta relevancia en las aplicaciones. En el año 1987, Enright y Pryce [38] proponen una colección de rutinas numéricas así como una amplia colección de ecuaciones diferenciales ordinarias modelo con vistas a determinar el dominio de problemas sobre el que un determinado método numérico resulta adecuado para una integración eficiente, así como, además, para detectar posibles debilidades que puedan existir en la implementación del mismo.

En particular, Enright y Pryce proponen, dentro del marco de los problemas stiff, modelos de problemas pertenecientes a seis diferentes clases de ecuaciones diferenciales ordinarias: una clase A de problemas lineales con autovalores reales, una clase B de problemas lineales con autovalores no reales, una clase C de problemas no lineales con acoplamiento, una clase D de problemas no lineales con autovalores reales, una clase E de problemas no lineales con autovalores no reales, así como, finalmente, una clase F de problemas diferenciales que surgen en el campo de la Cinética Química. Tras analizar el número y naturaleza de los puntos de equilibrio que cada uno de estos problemas presenta, hallamos que este paquete de problemas modelo contiene diversos e interesantes problemas que se ajustan al marco de las A/H-hipótesis.

El comportamiento de las soluciones a cada uno de estos problemas ha sido simulado considerando el método de Euler implícito como integrador numérico. Tal como veremos en la parte final de este capítulo, este método resulta ser *incondicionalmente estable* sobre la clase de sistemas diferenciales bajo las H-hipótesis, en el sentido de que no impone restricciones en el tamaño de paso para garanatizar la estabilidad numérica.

De este modo, cada problema se acompaña de ilustraciones gráficas sobre el comportamiento de las componentes de la solución, así como de la órbita y pendiente numérica resultantes de aplicar el método de Euler implícito. El estudio de la órbita y pendiente que provee el método resultan de especial interés, por cuanto la solución exacta tiende al punto de equilibrio en tiempo infinito formando ángulo cero con el autovector derecho asociado al autovalor nulo de la matriz Jacobiana de la función derivada $f(y)$ en el equilibrio.

Ejemplo 1 (Problema *D2*, [38]).

Este problema de dimensión 3 con matriz Jacobiana con autovalores reales

$$\begin{aligned} y_1' &= -0.04y_1 + 0.01y_2y_3, & y_1(0) &= 1, \\ y_2' &= 400y_1 - 100y_2y_3 - 3000y_2^2, & y_2(0) &= 0, \\ y_3' &= 30y_2^2, & y_3(0) &= 0, \end{aligned}$$

integrado hasta $T_f = 10^{30}$, posee un invariante lineal $100y_1 + 0.01y_2 + y_3 = 100$, que permite despejar la tercera componente de la solución en función de las dos primeras y así reducir el sistema a una dimensión menor

$$\begin{aligned} y_1' &= -0.04y_1 + y_2 - y_1y_2 - 10^{-4}y_2^2, & y_1(0) &= 1 \\ y_2' &= 400y_1 - 10^4y_2 + 10^4y_1y_2 - 2999y_2^2, & y_2(0) &= 0. \end{aligned} \tag{II.10}$$

El problema reducido (II.10) tiene un único punto de equilibrio $(0,0)^T$, y la matriz Jacobiana en dicho punto posee autovalores $\lambda_1 = 0$ y $\lambda_2 = -10000.04$. Los autovectores derechos e izquierdo asociados al autovalor nulo son respectivamente

$$v := (1, 0.04)$$

y

$$p := (10000, 1).$$

Así, la hipótesis $(A2)$ se cumple para el valor $\vartheta = \dfrac{-4.8}{10000.04} \simeq -4.8 \cdot 10^{-4}$.

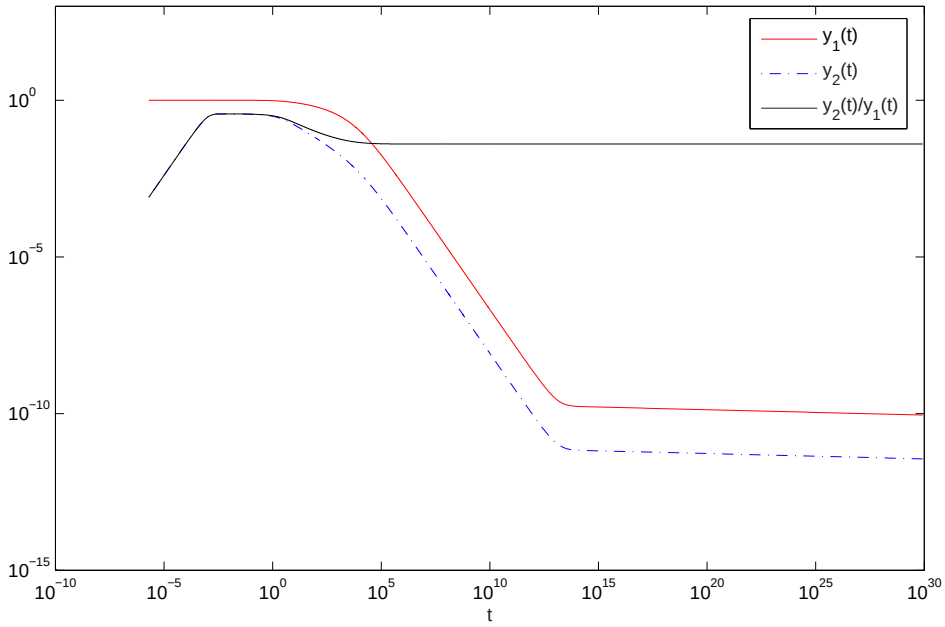

Figura II.1: Ejemplo 1. Componentes de la solución y pendiente numérica (escala logarítmica).

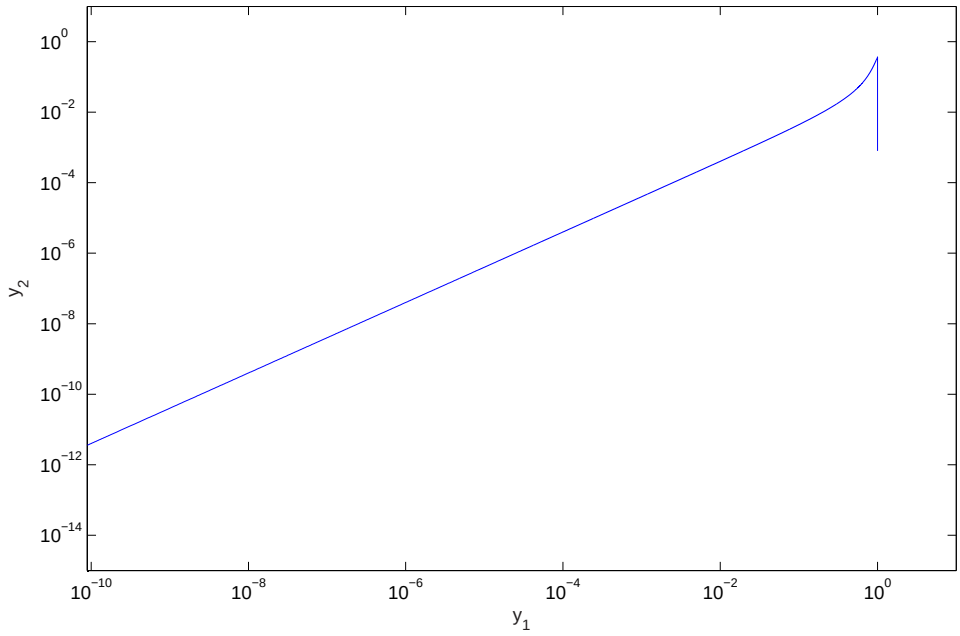

Figura II.2: Ejemplo 1. Órbita numérica (escala logarítmica).

Ejemplo 2 (Problema *E5*, [38], [53, p. 145]).

Este problema de dimensión 4, que modela una reacción química,

$$\begin{aligned}
y_1' &= -7.89\cdot 10^{-10}y_1 - 1.1\cdot 10^{7}y_1y_3, & y_1(0) &= 1.76\cdot 10^{-3},\\
y_2' &= 7.89\cdot 10^{-10}y_1 - 1.13\cdot 10^{9}y_2y_3, & y_2(0) &= 0,\\
y_3' &= 7.89\cdot 10^{-10}y_1 - 1.1\cdot 10^{7}y_1y_3 + 1.13\cdot 10^{3}y_4 - 1.13\cdot 10^{9}y_2y_3, & y_3(0) &= 0,\\
y_4' &= 1.1\cdot 10^{7}y_1y_3 - 1.13\cdot 10^{3}y_4, & y_4(0) &= 0,
\end{aligned}$$

tomando $T_f = 10^{12}$, tiene el invariante lineal $y_2 - y_3 - y_4 = 0$, que después de ser insertado en el sistema diferencial y hacer el cambio lineal

$$x_1 = y_1, \quad x_2 = 10^9 y_2, \quad x_3 = 10^9 y_3,$$

(véase la dinámica de este problema en [53, p.145, Fig.10.1] para comprender el motivo de este cambio) produce un sistema reducido de dimensión tres

$$\begin{aligned}
x_1' &= -7.89\cdot 10^{-10}x_1 - 1.1\cdot 10^{-2}x_1x_3, & x_1(0) &= 1.76\cdot 10^{-3},\\
x_2' &= 0.789x_1 - 1.13x_2x_3, & x_2(0) &= 0,\\
x_3' &= 0.789x_1 + 1.13\cdot 10^{3}(x_2 - x_3) - 1.1\cdot 10^{7}x_1x_3 - 1.13x_2x_3, & x_3(0) &= 0.
\end{aligned} \tag{II.11}$$

Este problema reducido (II.11) posee dos puntos de equilibrio, aunque sólo el punto $(0,0,0)^T$ se ajusta a las A-hipótesis. La matriz Jacobiana en dicho punto tiene autovalores $\lambda_1 = 0,\ \lambda_2 = -7.89\cdot 10^{-10}$ y $\lambda_3 = -1.13\cdot 10^3$. Los autovectores derecho e izquierdo, respectivamente, asociados al autovalor nulo son

$$v = (0,1,1)$$

y

$$p = (1,1,0).$$

En este caso, se satisfacen las A-hipótesis para el valor $\vartheta = -1.13$.

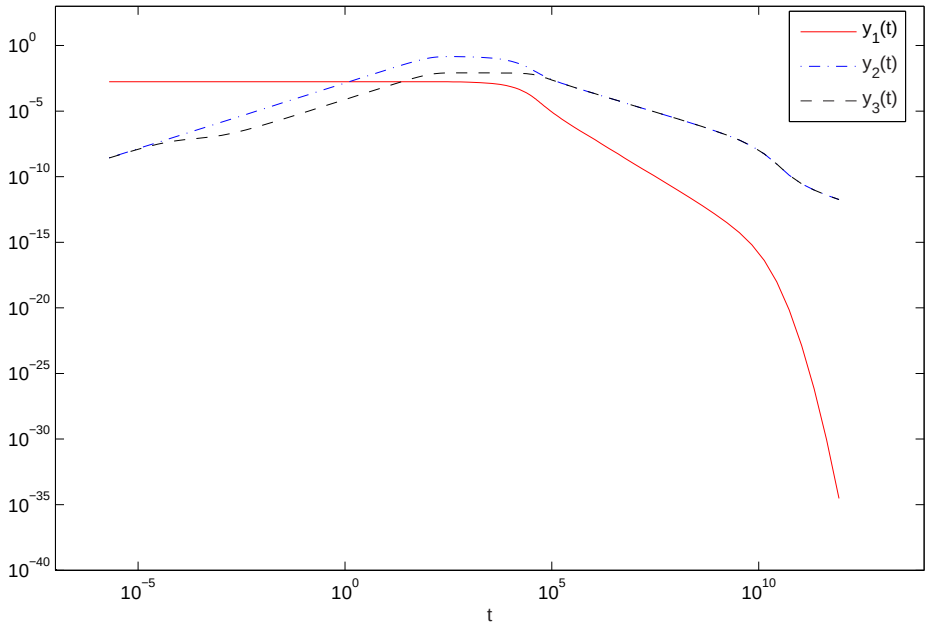

Figura II.3: Ejemplo 2. Componentes de la solución (escala logarítmica).

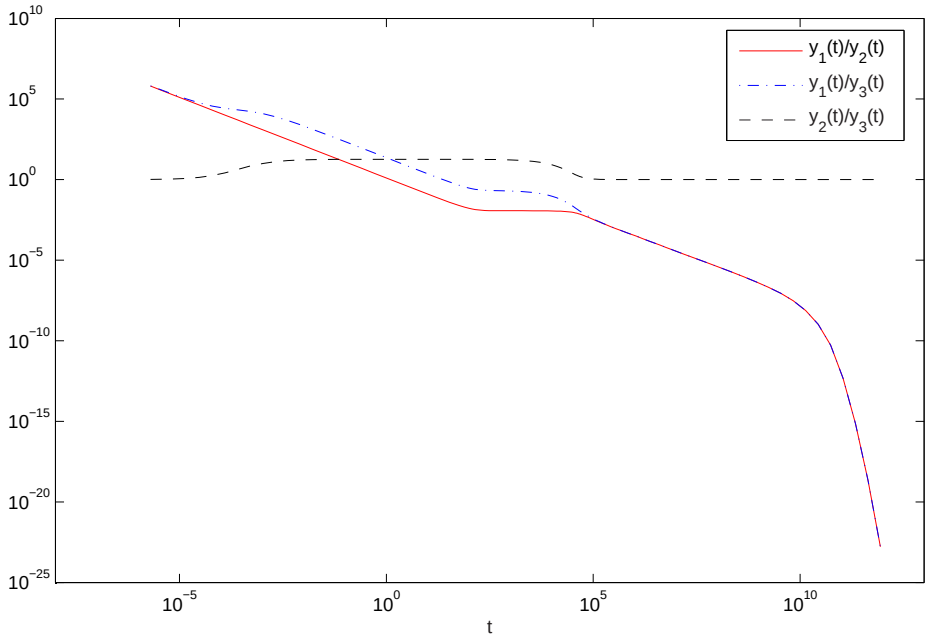

Figura II.4: Ejemplo 2. Pendientes numéricas (escala logarítmica).

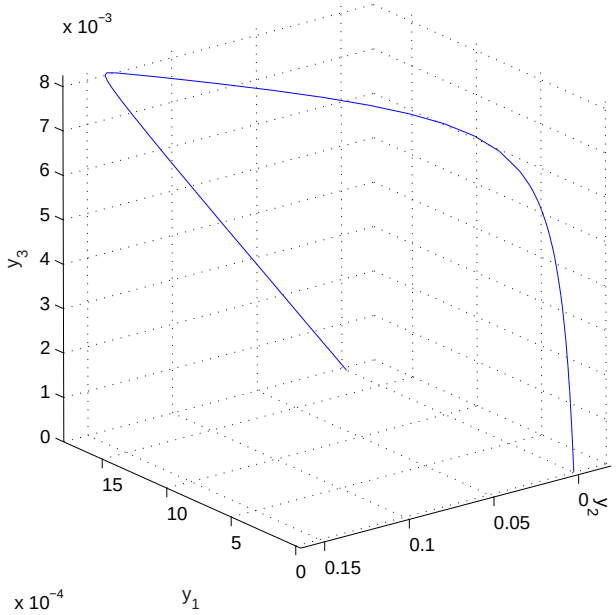

Figura II.5: Ejemplo 2. Órbita numérica.

Ejemplo 3 (Problema *F2*, [38]).

Este problema de dimensión 2 procedente de la Cinética Química viene dado por

$$\begin{aligned} y_1' &= -y_1 - y_1 y_2 + 294 y_2, & y_1(0) &= 1, \\ y_2' &= \frac{y_1(1-y_2)}{98} - 3y_2, & y_2(0) &= 0, \end{aligned} \tag{II.12}$$

tomando $T_f = 10^{16}$. Tal sistema presenta un único punto de equilibrio $(0,0)^T$ que resulta ajustarse a las A-hipótesis. Para este punto la matriz Jacobiana posee autovalores $\lambda_1 = 0$ y $\lambda_2 = -4$, y los autovectores derecho e izquierdo asociados al autovalor nulo son, respectivamente,

$$v = (294, 1)$$

y

$$p = (1, 98).$$

De hecho, la hipótesis $(A2)$ se cumple para el valor $\vartheta = -1.5$.

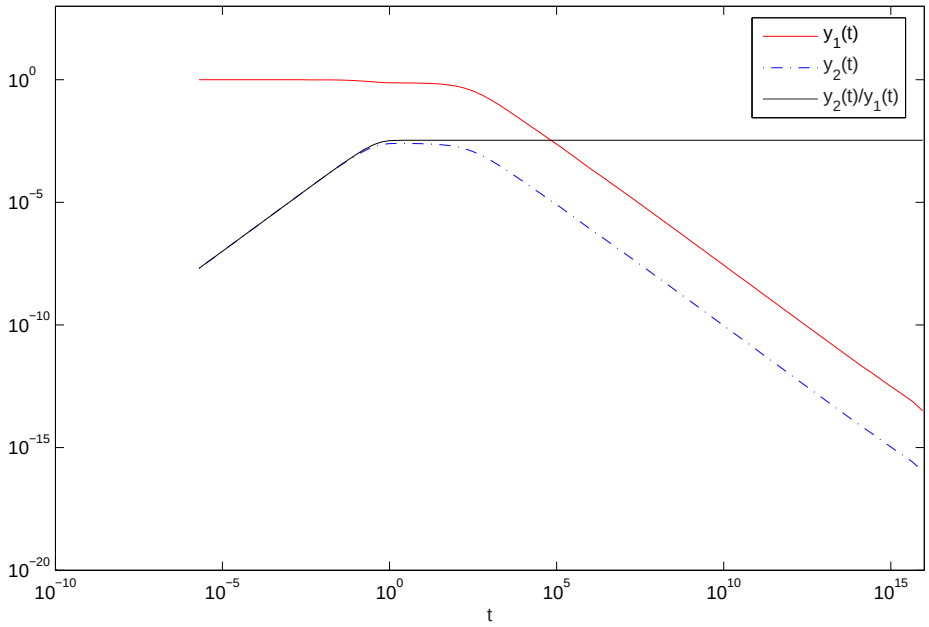

Figura II.6: Ejemplo 3. Componentes de la solución y pendiente numérica (escala logarítmica).

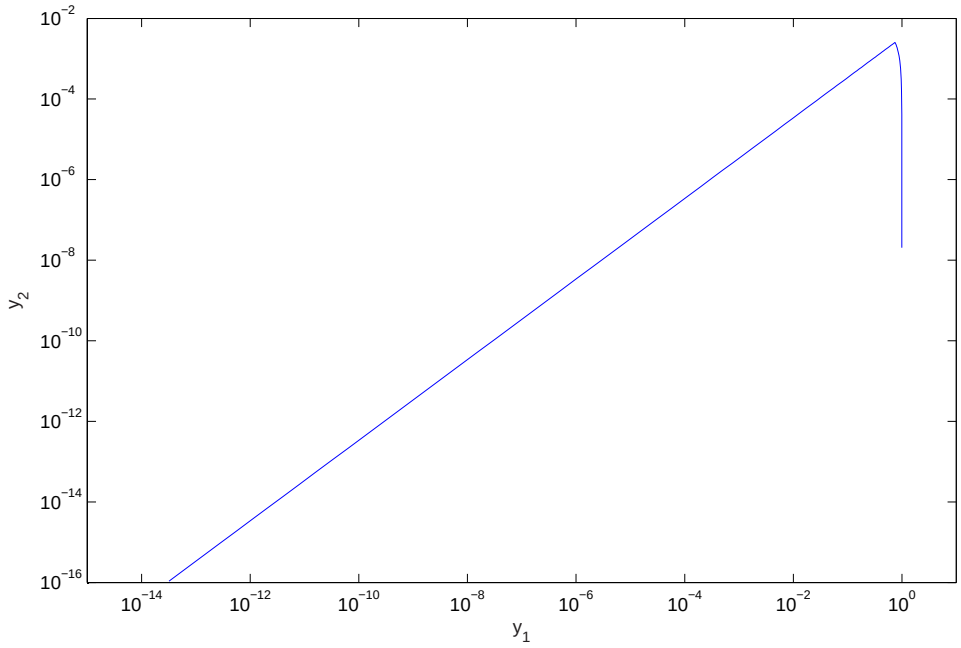

Figura II.7: Ejemplo 3. Órbita numérica (escala logarítmica).

Ejemplo 4 (Problema de Robertson [53, p. 3, 144], [76].)

Este problema proviene del campo de la Cinética Química y modela una reacción en la que intervienen tres productos reactivos de diferente naturaleza y velocidad de reacción.

$$\begin{aligned} y_1' &= -0.04y_1 + 10^4 y_2 y_3, & y_1(0) &= 1, \\ y_2' &= 0.04y_1 - 10^4 y_2 y_3 - 3\cdot 10^7 y_2^2, & y_2(0) &= 0, \\ y_3' &= 3\cdot 10^7 y_2^2, & y_3(0) &= 0, \end{aligned}$$

donde hemos considerado $T_f = 10^{20}$. Para este problema así planteado, el origen de coordenadas es un punto de equilibrio cuya matriz Jacobiana posee autovalor nulo doble, y en consecuencia se escaparía del marco hipotético considerado inicialmente. Más aún el sistema diferencial posee puntos críticos no aislados de la forma $(0, 0, \lambda)^T$ con $\lambda \in \mathbb{R}$ arbitrario. Sin embargo, fruto de la ley de conservación de las masas y de la condición inicial, se deduce para la solución del problema el invariante lineal $y_1 + y_2 + y_3 \equiv 1$. En consecuencia, reemplazando $y_3 = 1 - y_1 - y_2$ y efectuando el cambio de variables lineal

$$x_1 = y_1 + y_2, \quad x_2 = y_2$$

se obtiene el sistema diferencial reducido

$$\begin{aligned} x_1' &= -3\cdot 10^7 x_2^2, & x_1(0) &= 1, \\ x_2' &= 0.04x_1 - (10^4 + 0.04)x_2 + 10^4 x_1 x_2 - 3\cdot 10^7 x_2^2, & x_2(0) &= 0. \end{aligned} \tag{II.13}$$

Este problema reducido (II.13) posee al origen de coordenadas como único punto de equilibrio y la matriz Jacobiana en dicho punto tiene autovalores $\lambda_1 = 0$ y $\lambda_2 = -(10^4 + 0.04)$. En este caso, los autovectores derecho e izquierdo asociados al autovalor nulo son, respectivamente

$$v = (10000.04, 0.04)$$

y

$$p = (1, 0),$$

y es sencillo comprobar que las A-hipótesis se verifican para el valor $\vartheta \simeq -0.48 \cdot 10^{-3}$.

Muchos códigos numéricos no son capaces de concluir de modo aceptable la integración de los problemas arriba presentados si la variable temporal se elige en un intervalo muy amplio ($T_f = 10^{10},\ 10^{15},\ 10^{20}, \dots$). Esto se debe a que la solución numérica no logra ser lo suficientemente estable como para evitar escapar a la región de inestabilidad del problema en cuestión, provocando que la solución numérica tienda a infinito y como consecuencia el desborde numérico en la integración.

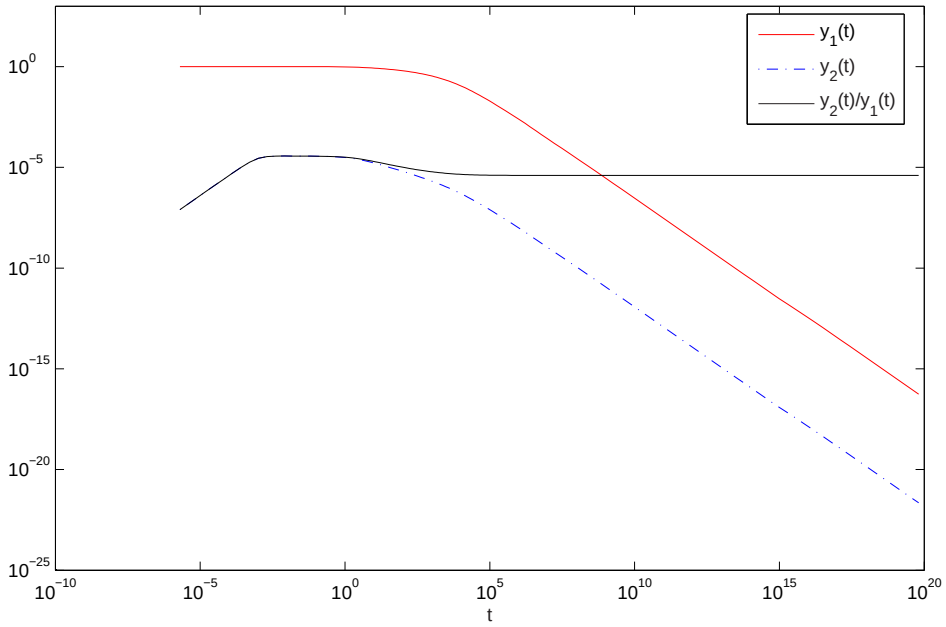

Figura II.8: Ejemplo 4. Componentes de la solución y pendiente numérica (escala logarítmica).

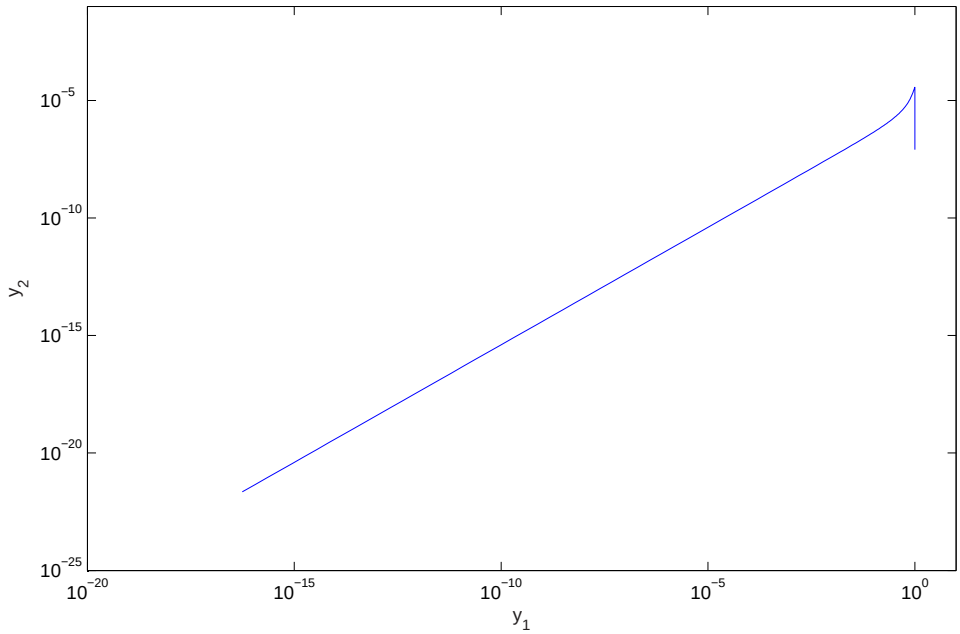

Figura II.9: Ejemplo 4. Órbita numérica (escala logarítmica).

II.3. Dinámica de los sistemas diferenciales bajo las H-hipótesis

La dinámica local alrededor del punto de equilibrio $y = 0$ del sistema diferencial (II.1)-(II.3) bajo las H-hipótesis está determinada por el comportamiento de las órbitas que parten de la variedad centro del punto de equilibrio tal como se comentó en la introducción. No obstante lo anterior, con vistas a profundizar en el estudio de estabilidad de los métodos numéricos aplicados en redes temporales que permitan grandes tamaños de paso, deseamos analizar con algo más de detalle una clase especial de conjuntos invariantes de órbitas estables para (II.1)-(II.3). En otras palabras, consideraremos aquí un tipo especial de conjuntos invariantes bajo el flujo del sistema diferencial que jugarán un papel relevante en el estudio de la estabilidad de las órbitas (positivas), así como en el análisis de estabilidad de los métodos numéricos.

Dado un problema diferencial bajo las H-hipótesis, denotaremos por $v_u^\perp$ al conjunto de vectores unitarios ortogonales al autovector v, esto es,

$$v_u^\perp = \{x \in v^\perp, \ \ \|x\| = 1\}.$$

Definición II.3.1 *Dados $\alpha > 0, \beta > 0$, llamamos* cono *de vértice en el origen y dirección v determinado por α y β al conjunto*

$$\mathcal{C}_{\alpha,\beta} := \{x = \tau(v + \theta w), \ \tau \geq 0, \ \theta \in [0, \alpha], \ \|x\| \leq \beta, \ w \in v_u^\perp\} \tag{II.14}$$

Asimismo introducimos la notación $\mathcal{C}^+_{\alpha,\beta} := \mathcal{C}_{\alpha,\beta} \setminus \{0\}$. En el caso de que $\beta = +\infty$ simplemente denotaremos el cono $\mathcal{C}_{\alpha,\beta}$ por $\mathcal{C}_\alpha$.

Nota II.3.2 Obsérvese que para todo vector $x \in \mathbb{R}^m \setminus v^\perp$ existen escalares únicos $\tau \neq 0$, $\theta > 0$ y un vector $w \in v_u^\perp$ tal que $x = \tau(v + \theta w)$. Si $x \in span\{v\}$ entonces $\theta = 0$ y $w \in v_u^\perp$ no tiene porqué ser único.

Por otro lado, dado $x = \tau(v + \theta w) \in \mathbb{R}^m \setminus \{0\}$, con $\tau > 0$, $\theta > 0$ y $w \in v_u^\perp$, llamaremos *normal exterior* en x al vector

$$n(x) := w - \theta v.$$

También se debe tener presente que el ángulo $\widehat{(x, y)}$ entre dos vectores $x, y \in \mathbb{R}^m$ viene dado por el número real

$$\widehat{(x, y)} := \arccos\left(\frac{\langle x, y\rangle}{\|x\| \cdot \|y\|}\right).$$

Teniendo en cuenta las definiciones previas y dado un vector x de la forma $x = \tau(v + \theta w)$, con $\tau > 0, \ \theta > 0$ y $w \in v_u^\perp$, se tiene que $\langle x, n(x)\rangle = 0$ y que $\widehat{(x, n(x))} = \pi/2$.

Los dos siguientes lemas se introducen con vistas a obtener una cierta perspectiva geométrica en la interpretación de la familia de conos (II.14), así como para resaltar diversas propiedades relativas al interior del cono no acotado $Int(\mathcal{C}_\alpha)$ definido anteriormente.

Lema II.3.3 *Los conjuntos $\mathcal{C}_{\alpha,\beta}$ y $\mathcal{C}_\alpha$ son conjuntos convexos de $\mathbb{R}^m$ para cada $\alpha > 0$, $\beta > 0$.*

Demostración. Pongamos $x = \tau_x(v+\theta_x w_x)$ e $y = \tau_y(v+\theta_y w_y)$, siendo $\tau_x, \tau_y \geq 0$, $\theta_x, \theta_y \in [0, \alpha]$ y $w_x, w_y \in v_u^\perp$. Dado $\lambda \in [0,1]$, tomemos $z = \lambda x + (1-\lambda)y$. Entonces expresando $z = \tau_z(v+w_z)$, con $w_z \in v^\perp$, tenemos que

$$\tau_z = \langle z, v\rangle = \lambda\tau_x + (1-\lambda)\tau_y \geq 0.$$

Por otro lado,

$$\begin{aligned} w_z &= \frac{z - \langle z, v\rangle v}{\tau_z} \\ &= \frac{\lambda\tau_x\theta_x w_x + (1-\lambda)\tau_y\theta_y w_y}{\lambda\tau_x + (1-\lambda)\tau_y}, \end{aligned}$$

y de aquí que $\|w_z\| \leq \alpha$.

Finalmente, si además, $\max\{\|x\|, \|y\|\} \leq \beta$, entonces sigue inmediatamente que

$$\|z\| \leq \lambda \|x\| + (1-\lambda)\|y\| \leq \beta,$$

lo cual concluye la prueba. □

Lema II.3.4 *Sea $x_0 = \tau_0(v + \alpha_0 w_0) \in \mathcal{C}_\alpha^+$.*

1. *$x_0 \in Int(\mathcal{C}_\alpha) \iff \langle x_0, w - \alpha v\rangle < 0,\ \forall w \in v_u^\perp$.*
2. *Si $x_0 \in \partial\mathcal{C}_\alpha$, esto es, $\alpha_0 = \alpha$, se tiene que*
   (a) *$x := x_0 + \epsilon y \notin Int(\mathcal{C}_\alpha),\ \forall \epsilon > 0 \iff \langle y, n(x_0)\rangle \geq 0$.*
   (b) *Existe $\epsilon_0 > 0$ tal que $x := x_0 + \epsilon y \in Int(\mathcal{C}_\alpha), \forall \epsilon \in (0, \epsilon_0]$ si y sólo si $\langle y, n(x_0)\rangle < 0$.*

Demostración. (1) Tenemos que

$$\begin{aligned} x_0 \in Int(\mathcal{C}_\alpha) &\iff \tau_0 > 0, 0 \leq \alpha_0 < \alpha \\ &\iff \langle x_0, w - \alpha v\rangle = \tau_0(-\alpha + \alpha_0\langle w_0, w\rangle) < 0,\ \forall w. \end{aligned}$$

(2a) "⇐" Si existe $\epsilon > 0$ tal que $x := x_0 + \epsilon y \in Int(\mathcal{C}_\alpha)$, entonces, por (1),

$$\langle x_0 + \epsilon y, w_0 - \alpha v\rangle = \epsilon\langle y, n(x_0)\rangle < 0,$$

y por tanto $\langle y, n(x_0)\rangle < 0$.

"$\Rightarrow$" Si $x := x_0 + \epsilon y \notin Int(\mathcal{C}_\alpha), \forall \epsilon > 0$ sigue entonces que $\widehat{(x,v)} \geq \widehat{(x_0,v)}$, esto es, $\cos\widehat{(x,v)} \leq \cos\widehat{(x_0,v)}$. De aquí tenemos que

$$\|x_0\| \langle x, v\rangle \leq \|x\| \langle x_0, v\rangle. \tag{II.15}$$

Insertando las igualdades $\langle x_0, v\rangle = \tau_0$, $\langle x, v\rangle = \tau_0 + \epsilon\langle y, v\rangle$ y el desarrollo asintótico

$$\|x\| = \sqrt{\|x_0\|^2 + 2\epsilon\langle x_0, y\rangle + \epsilon^2 \|y_0\|^2} = \|x_0\| + \epsilon\frac{\langle x_0, y\rangle}{\|x_0\|} + \mathcal{O}(\epsilon^2)$$

en (II.15), sigue tras dividir tal desigualdad por ϵ y multiplicar por $\|x_0\|$ que

$$\|x_0\|^2 \langle y, v\rangle \leq \tau_0\langle y, x_0\rangle + \epsilon, \ \forall \epsilon > 0.$$

De aquí que para $\epsilon \to 0^+$ se obtenga

$$0 \leq \langle y, \tau_0 x_0 - \|x_0\|^2 v\rangle = \alpha\tau_0^2\langle y, n(x_0)\rangle.$$

(2b) De (2a) se deduce que

$$\langle y, n(x_0)\rangle < 0 \iff x_0 + \epsilon_0 y \in Int(\mathcal{C}_\alpha), \text{ para cierto } \epsilon_0 > 0.$$

La prueba concluye teniendo presente que $\mathcal{C}_\alpha$ es un conjunto convexo. □

Seguidamente mostramos que bajo las H-hipótesis existe un cono tal que el origen es el único punto de equilibrio de (II.1) en dicho cono.

Lema II.3.5 *Bajo las H-hipótesis existe un cono $\mathcal{C}_{\alpha,\beta}$ tal que*

$$f(y) = 0 \quad \text{con } y \in \mathcal{C}_{\alpha,\beta} \Longrightarrow y = 0.$$

Demostración. Expresando $y = \tau(v + \theta w),\ w \in v_u^\perp$, $\tau \geq 0$, $\theta \geq 0$ y asumiendo $f(y) = 0$, sigue a partir de las condiciones expresadas en las H-hipótesis que

$$\begin{aligned}
0 &= \langle v, f(y)\rangle \\
&= \langle v, \tau J(v + \theta w) + N(\tau(v + \theta w)) + R(\tau(v + \theta w))\rangle \\
&= \langle v, N(\tau(v + \theta w)) + R(\tau(v + \theta w))\rangle \\
&= \tau^2\langle v, M(v + \theta w, v + \theta w) + \tau(\mathcal{O}(1) + \mathcal{O}(\theta))\rangle \\
&= \tau^2(-\nu + \mathcal{O}(\theta) + \mathcal{O}(\tau)),
\end{aligned}$$

donde las constantes involucradas en $\mathcal{O}(1)$, $\mathcal{O}(\tau)$, $\mathcal{O}(\theta)$ únicamente dependen de las constantes que aparecen en las H-hipótesis. Para τ, θ suficientemente pequeños se deduce entonces que $y = 0$. □

Antes de enunciar el principal resultado teórico relativo al comportamiento de las órbitas alrededor del origen, presentamos el siguiente resultado que nos da una condición práctica que garantiza que las órbitas que parten de cierto cono no escapan del mismo ni cruzan su frontera.

Lema II.3.6 *Para el sistema diferencial (II.1)-(II.3) bajo las H-hipótesis existe un cono $\mathcal{C}_{\alpha,\beta}$ tal que*

$$\langle y, f(y)\rangle \leq -\frac{\nu}{2}\|y\|^3, \quad \forall y \in \mathcal{C}_{\alpha,\beta}. \tag{II.16}$$

Demostración. Pongamos $y = \tau(v + \theta w)$, con $\theta \geq 0$, $\tau \geq 0$ y $w \in v_u^{\perp}$. A partir de las H-hipótesis se deduce entonces que

$$\begin{aligned}
\langle y, f(y)\rangle &= \langle \tau(v+\theta w), \tau J(v+\theta w) + N(\tau(v+\theta w)) + R(\tau(v+\theta w))\rangle \\
&= \tau^2\theta^2\langle w, Jw\rangle + \tau\langle v+\theta w, \tau^2 N(v+\theta w) + R(\tau(v+\theta w))\rangle \\
&= \tau^2\theta^2\langle w, Jw\rangle + \tau\langle v+\theta w, \tau^2(N(v) + \mathcal{O}(\theta) + \mathcal{O}(\tau))\rangle \\
&\leq -\delta_1\tau^2\theta^2 + \tau^3(-\nu + \mathcal{O}(\theta) + \mathcal{O}(\tau)).
\end{aligned}$$

Por lo tanto, podemos elegir constantes positivas α, β suficientemente pequeñas de tal manera que para cada $y \in \mathcal{C}_{\alpha,\beta}$ se tiene

$$\langle y, f(y)\rangle \leq -\frac{\nu}{2}\|y\|^3 < 0.$$

□

Lema II.3.7 *Sea el sistema diferencial (II.1)-(II.3) bajo las H-hipótesis. Para cada $\theta > 0$ fijo, eligiendo*

$$\beta_\theta := \frac{2\delta_1\theta(1+\theta^2)^{-1}}{\delta_2 + \sqrt{\delta_2^2 + 4\delta_1\delta_3\theta(1+\theta^2)^{-1}}}, \tag{II.17}$$

se tiene que

$$\langle n(y), f(y)\rangle < 0, \quad \forall y = \tau(v+\theta w), \text{ siendo } \tau > 0,\ w \in v_u^{\perp},\ \|y\| < \beta_\theta.$$

Demostración. Sea $y = \tau(v + \theta w)$, con $\theta > 0$, $\tau > 0$ y $w \in v_u^{\perp}$. Entonces a partir de las H-hipótesis y la desigualdad de Cauchy-Schwarz se obtiene que

$$\begin{aligned}
\langle n(y), f(y)\rangle &= \langle w - \theta v, \tau J(v + \theta w) + N(\tau(v + \theta w)) + R(\tau(v + \theta w))\rangle \\
&= \tau\theta\langle w, Jw\rangle + \langle w - \theta v, N(\tau(v + \theta w)) + R(\tau(v + \theta w))\rangle \\
&\leq -\delta_1\tau\theta + \|w - \theta v\| \cdot \|N(\tau(v + \theta w)) + R(\tau(v + \theta w))\| \\
&\leq -\delta_1\tau\theta + (1 + \theta^2)^2(\delta_2\tau^2(1 + \theta^2)^{-1/2} + \tau^3\delta_3) \\
&= \tau(1 + \theta^2)^2(-\delta_1\theta(1 + \theta^2)^{-2} + \delta_2\tau(1 + \theta^2)^{-1/2} + \tau^2\delta_3).
\end{aligned}$$

Teniendo en cuenta el desarrollo previo y que $\tau = \|y\|\,(1 + \theta^2)^{-1/2}$ obtenemos que

$$\langle n(y), f(y)\rangle \leq (1 + \theta^2)^{1/2}\,\|y\|\,(-\delta_1\theta(1 + \theta^2)^{-1} + \delta_2\,\|y\| + \delta_3\,\|y\|^2).$$

Teniendo en cuenta que la expresión (II.17) dada para β_θ corresponde a la única raíz positiva de la ecuación de segundo grado en $\|y\|$

$$-\delta_1\theta(1 + \theta^2)^{-1} + \delta_2\,\|y\| + \delta_3\,\|y\|^2 = 0,$$

deducimos entonces que $\langle n(y), f(y)\rangle < 0$ cuando $0 < \|y\| < \beta_\theta$. □

El siguiente teorema describe el comportamiento local de las órbitas alrededor del punto de equilibrio.

Teorema II.3.8 *Bajo las H-hipótesis, existe un cono $\mathcal{C}_{\alpha,\beta}$ tal que para cada $y_0 \in \mathcal{C}^+_{\alpha,\beta}$ se tiene que*

1. $\|y(t; 0, y_0)\| < \|y_0\|\,, \quad \forall t > 0.$
2. $y(t; 0, y_0) \in \mathcal{C}^+_{\alpha,\beta}, \quad \forall t \geq 0.$
3. $\lim_{t\to\infty} y(t; 0, y_0) = 0.$
4. $\lim_{t\to\infty} \widehat{(y(t), v)} = 0$, *siendo* $y(t) \equiv y(t; 0, y_0)$.

Demostración. Consideremos las constantes α y β dadas por el lema II.3.6, y tomemos el cono $\mathcal{C}_{\alpha,\overline{\beta}}$, con $\overline{\beta} = \min\{\beta, \beta_\alpha\}$, donde β_α se define como en (II.17). De este modo se verificará (II.16) en el cono $\mathcal{C}^+_{\alpha,\overline{\beta}}$ y

$$\langle n(y), f(y)\rangle < 0, \text{ si } y = \tau(v + \alpha w), \quad 0 < \|y\| < \overline{\beta},\ w \in v_u^{\perp}. \tag{II.18}$$

Con esto, tomando $y_0 \in \mathcal{C}^+_{\alpha,\overline{\beta}}$ tenemos que la órbita que parte de y_0, $y(t) := y(t;0,y_0)$, debe permanecer completamente en dicho cono, ya que, en otro caso, existiría $t_0 \geq 0$, con $y(t_0) = \tau_0(v+\alpha w)$, $\tau_0 \in (0,\overline{\beta}]$, punto frontera del cono, tal que

$$y(t_0+\epsilon) = y(t_0) + \epsilon f(y(t_0)) + \mathcal{O}(\epsilon^2) \notin Int(\mathcal{C}^+_{\alpha,\overline{\beta}})$$

para $\epsilon \to 0$. De aquí que según el lema (II.3.4) se tendría para $\epsilon > 0$ suficientemente pequeño que

$$\langle n(y(t_0)), f(y(t_0)) + \mathcal{O}(\epsilon)\rangle \geq 0.$$

En consecuencia, debe ser entonces que $\langle n(y(t_0)), f(y(t_0))\rangle \geq 0$, lo cual contradice (II.18).

Por otro lado, la propiedad (II.16) implica que $\phi'(t) \leq -\nu\phi^{3/2}(t)$, siendo $\phi(t) := \|y(t)\|^2$. La prueba de (1), (2) y (3) concluye teniendo en cuenta que

$$\phi(t) \leq \left(\|y_0\|^{-1} + \frac{\nu}{2}t \right)^{-2}, \quad \forall t \geq 0.$$

Para probar (4) tomamos $y(t) = \tau(t)(v+\theta(t)w_t)$, $w_t \in v_u^{\perp}$, donde $\tau(t)$ es una función continua positiva con límite a cero en el infinito. Podemos asumir además que $\theta(t)$ una función continua que toma valores en el intervalo $[0,\alpha]$.

Definiendo $\zeta(t) := \widehat{(y(t),v)}$ tenemos que

$$\cos^2\zeta(t) = \frac{\langle y(t),v\rangle^2}{\|y(t)\|^2} = \frac{1}{1+\theta^2(t)} \in [(1+\alpha^2)^{-1},1].$$

Si $\theta(t)$ oscilara o tuviera un límite en el infinito $\hat{\theta} \neq 0$ entonces, teniendo en cuenta las H-hipótesis, la regla de L'Hôpital permitiría escribir

$$\begin{aligned}
\lim_{t\to\infty} \frac{\langle v,y(t)\rangle^2}{\|y(t)\|^2} &= \lim_{t\to\infty} \frac{\langle v,y(t)\rangle \cdot \langle v,f(y(t))\rangle}{\langle y(t),f(y(t))\rangle} \\
&= \lim_{t\to\infty} \frac{\tau^3(t)(-\nu+\mathcal{O}(\tau(t))+\mathcal{O}(\theta(t)))}{\tau^2(t)\theta^2(t)\langle w_t,Jw_t\rangle + \mathcal{O}(\tau^3(t))} \\
&= 0,
\end{aligned}$$

por cuanto $\langle w_t, Jw_t\rangle \leq -\delta_1 < 0$.

La contradicción surge de suponer que la función $\theta(t)$ no tiene límite nulo. En consecuencia, debe ser que $\zeta(t) \to 0$, si $t \to \infty$, y la prueba queda concluida. □

II.4. Análisis de estabilidad del método de Euler implícito

En esta parte final del capítulo, y bajo el marco de las H-hipótesis, nos interesaremos por el análisis de las órbitas numéricas bajo discretizaciones por métodos de un paso. Consideremos un método de un paso φ_h, que computa soluciones numéricas por medio de una fórmula del tipo

$$y_{n+1} = \varphi_{h_n}(y_n),\ t_0 = 0,\ t_{n+1} = t_n + h_n,\ h_n > 0, \quad n = 0, 1, \dots$$

donde la aplicación φ_h no depende del tiempo t por cuanto estamos considerando problemas autónomos.

Sería deseable analizar qué métodos son *incondicionalmente estables* bajo las *H-hipótesis*, esto es, determinar aquellos métodos cuyas soluciones numéricas imiten el comportamiento de las órbitas estables independientemente de los tamaños de paso h_n que se consideren. Más precisamente, estamos interesados en determinar aquellos métodos para los que existe un cono $\mathcal{C}_{\alpha,\beta}$ (con $\alpha > 0,\ \beta > 0$) tal que,

$$\begin{gathered} y_1 = \varphi_h(y_0) \in \mathcal{C}_{\alpha,\beta} \quad \text{y} \ \|y_1\| < \|y_0\|, \\ \text{siempre que } h > 0,\ y_0 \in \mathcal{C}_{\alpha,\beta}. \end{gathered} \tag{II.19}$$

Esta propiedad implicaría un buen comportamiento de estabilidad sin restricción alguna en la selección del tamaño de paso. Las únicas restricciones posibles en la selección del paso vendrían impuestas por la precisión de las soluciones numéricas, pero en ningún caso por su estabilidad. Esto permitiría al método numérico integrar este tipo de problemas en amplios intervalos temporales en un pequeño número de pasos.

Por otro lado, la propiedad (II.19) resulta ser extremadamente exigente y pocos métodos implícitos de interés, a excepción del método de Euler implícito, la verifican. De hecho otros métodos numéricos con excelentes propiedades de estabilidad, tales como la regla trapezoidal, la regla implícita del punto medio o el método de dos etapas RadauIIA no verifican una tal propiedad, como se verá en la sección final de este capítulo.

Las discretizaciones temporales propuestas por el método de Euler implícito vienen dadas por

$$y_{n+1} = y_n + h_n f(y_{n+1}), \quad t_0 = 0,\ t_{n+1} = t_n + h_n,\ h_n > 0,\ \dots n = 0, 1, 2, \dots$$

La anunciada propiedad de estabilidad incondicional del método de Euler implícito sobre problemas del tipo (II.1)-(II.3) bajo el marco de las H-hipótesis queda reflejada en el siguiente

Teorema II.4.1 *Para el método de Euler implícito y bajo las H-hipótesis existe un cono $\mathcal{C}_{\alpha,\beta}$ tal que*

(1) para cada $h \geq 0$ y cada $y_0 \in \mathcal{C}_{\alpha,\beta}$, el método admite una solución $y_1 \in \mathcal{C}_{\alpha,\beta}$.

(2)
$$\|y_1\| \leq \frac{2}{1+\sqrt{1+2h\nu\,\|y_0\|}}\,\|y_0\|\,;$$

(3) independientemente de la red temporal elegida $t_0 = 0 < t_1 < t_2 \ldots$, se tiene después de n pasos consecutivos que
$$\|y_n\| \leq \frac{2}{1+\sqrt{1+2t_n\nu\,\|y_0\|}}\,\|y_0\|\,.$$

(4)
$$\lim_{h\to\infty} \widehat{(y_1, v)} = 0.$$

Demostración. (1) Por aplicación de los lemas II.3.6 y II.3.7 podemos considerar constantes $\alpha_1 > 0,\ \beta_1 > 0$ de modo que en el cono $\mathcal{C}_{\alpha_1,\beta_1}$ se verifiquen simultáneamente (II.16) y (II.18).

Fijemos ahora $h > 0$ e $y_0 \in Int(\mathcal{C}_{\alpha_1,\beta_1})$, y consideremos la homotopía

$$H(x,t):\ \mathcal{C}_{\alpha_1,\beta_1} \times [0,1] \longrightarrow \mathbb{R}^m \tag{II.20}$$

definida por $H(x,t) := x - y_0 - thf(x)$. Ahora, sigue de (II.16) y (II.18) que

$$H(x,t) \neq 0, \quad \forall t \in [0,1],\ \forall x \in \partial\mathcal{C}_{\alpha_1,\beta_1}, \tag{II.21}$$

donde $\partial\mathcal{C}_{\alpha_1,\beta_1}$ denota la frontera del conjunto $\mathcal{C}_{\alpha_1,\beta_1}$.

En efecto, debe observarse que para $x = 0$ se tiene que $H(0,t) = -y_0 \neq 0,\ \forall t \in [0,1]$. Por otro lado, si $x = \tau(v + \alpha_1 w),\ \tau \in (0, \hat{\beta}_1],\ \hat{\beta}_1 := \beta_1(1+\alpha_1^2)^{-1/2},\ w \in v_u^{\perp}$ y $t \in (0,1]$, sigue, en virtud del apartado 1 del Lema II.3.4, que

$$\langle n(x), H(x,t)\rangle = -\langle n(x), y_0\rangle - th\langle n(x), f(x)\rangle > -\langle n(x), y_0\rangle > 0.$$

Finalmente, si $\|x\| = \beta_1$, entonces sigue que

$$\begin{aligned}
\langle x, H(x,t)\rangle &= \langle x, x - y_0\rangle - ht\langle x, f(x)\rangle \\
&> \langle x, x - y_0\rangle \\
&\geq \|x\|^2 - \|x\|\,\|y_0\| \\
&= (\beta_1 - \|y_0\|)\beta_1 > 0.
\end{aligned}$$

De aquí que $H(x,t) \neq 0,\ \forall x \in \partial\mathcal{C}_{\alpha_1,\beta_1},\ \forall t \in [0,1]$, y se obtiene (II.21).

Por otro lado, veamos que para la norma logarítmica $\mu[\cdot]$ asociada al producto interior existe un cono $\mathcal{C}_{\alpha_2,\beta_2}$ tal que

$$\mu[f'(z)] := \sup_{\|u\|=1} \langle f'(z)u, u\rangle \leq 0, \qquad \forall z \in \mathcal{C}_{\alpha_2,\beta_2}. \tag{II.22}$$

Asumimos, por simplicidad en el cálculo, que $f(x) = Jx + N(x)$ (omitimos los términos de orden mayor o igual a tres en el desarrollo de Taylor de f). Entonces, teniendo en cuenta que

$$\langle f'(z)u, u\rangle = \lim_{\epsilon\to 0} \frac{\langle f(z+\epsilon u) - f(z), u\rangle}{\epsilon},$$

sólo es preciso probar que

$$\langle Ju + 2M(z,u), u\rangle \leq 0, \quad \forall u \in \mathbb{R}^m,\ \|u\| = 1,\ \forall z \in \mathcal{C}^+_{\alpha_2,\beta_2}, \tag{II.23}$$

donde M está definida por (II.7). Con este objetivo en mente tomamos

$$z = \tau(v + \theta w_0), \quad u = xv + yw,\ x^2 + y^2 = 1, \quad \tau \geq 0,\ \theta \geq 0, \quad w, w_0 \in v_u^{\perp}.$$

Teniendo en cuenta las H-hipótesis, un sencillo cálculo permite escribir

$$\begin{aligned}\langle Ju + 2M(z,u), u\rangle &= x^2[2\tau(-\nu + \theta\langle M(w_0,v),v\rangle)] + y^2[\langle Jw, w\rangle + 2\tau\langle M(v+\theta w_0, w), w\rangle] + \\ &\quad xy[2\tau\langle M(v+\theta w_0, w), v\rangle + 2\tau\langle M(v+\theta w_0, v), w\rangle] \\ &\leq Ax^2 + By^2 + C|xy|,\end{aligned}$$

donde

$$A = 2\tau(-\nu + \theta\delta_2), \quad B = -\delta_1 + 2\tau\delta_2\sqrt{1+\theta^2}, \quad C = 4\tau\delta_2\sqrt{1+\theta^2}.$$

Es sencillo comprobar que para τ, θ suficientemente pequeños se tiene

$$A \leq 0, \qquad B \leq 0 \qquad \text{y} \quad C^2 \leq 4AB.$$

Esto determina un cono $\mathcal{C}_{\alpha_2,\beta_2}$ para el que se verifica (II.23) y en consecuencia también (II.22).

Consideremos entonces el cono $\mathcal{C}_{\alpha,\beta}$ definido por $\alpha = \min\{\alpha_1, \alpha_2\}$ y $\beta = \min\{\beta_1, \beta_2\}$, de tal modo que en dicho cono se verifican simultáneamente (II.21) y (II.22).

Ahora, por derivación respecto de la variable h en la ecuación

$$y_1 = y_0 + hf(y_1), \tag{II.24}$$

se obtiene de modo inmediato que

$$(I - hf'(y_1))y_1' = f(y_1), \quad y_1(0) = y_0, \quad y_1 \equiv y_1(h). \tag{II.25}$$

Teniendo en cuenta (II.22), entonces el teorema de Von Neumann [53, p. 168],[87], implica que

$$\left\|(I - hf'(z))^{-1}\right\| \leq \sup_{Re\,\omega \leq 0} \frac{1}{|1-\omega|} = 1, \ \forall z \in \mathcal{C}_{\alpha,\beta}, \quad \forall h \geq 0.$$

En consecuencia, el sistema diferencial (II.25) tiene una única solución $y_1(h)$ continuamente diferenciable para $h \geq 0$ que, en añadidura, no puede atravesar la frontera del cono debido a la condición (II.21). Por lo tanto se verifica que $y_1 = y_1(h) \in \mathcal{C}_{\alpha,\beta}$.

(2) De (II.24) se obtiene directamente que $\|y_1\|^2 = \langle y_1, y_1\rangle = \langle y_0, y_1\rangle + h\langle y_1, f(y_1)\rangle$, y tras considerar la desigualdad de Cauchy-Schwarz y (II.16) sigue que

$$\|y_1\|^2 \leq \|y_0\| \, \|y_1\| - \frac{h}{2}\nu \, \|y_1\|^3 .$$

Esto implica

$$\|y_1\| \leq \|y_0\| - \frac{h}{2}\nu \, \|y_1\|^2 ,$$

y de aquí sigue inmediatamente la afirmación (2).

(3) Si $y_0 = 0$ el resultado es claramente cierto debido a (2). Por tanto, asumamos que $\|y_0\| > 0$ y tomemos cualquier red temporal $0 = t_0 < t_1 < \ldots < t_n < \ldots$, con $h_n = t_{n+1} - t_n, \ n = 0, 1, 2, \ldots$. De (2) se deduce entonces que

$$\|y_{n+1}\| \leq \|y_n\| - \frac{h_n}{2}\nu \, \|y_{n+1}\|^2 , \quad n = 0, 1, 2, \ldots \tag{II.26}$$

Consideremos la sucesión definida por

$$\phi_0 = \|y_0\| , \quad \phi_{n+1} = \frac{2}{1 + \sqrt{1 + 2h_n\nu\phi_n}}\phi_n, \quad n = 0, 1, 2, \ldots ,$$

que satisface

$$0 < \phi_{n+1} < \phi_n \quad \text{y} \quad \phi_{n+1} = \phi_n - \frac{h_n}{2}\nu(\phi_{n+1})^2, \quad n = 0, 1, 2, \ldots \tag{II.27}$$

Teniendo en cuenta (II.26) y (II.27) se deduce que

$$\frac{h_n\nu}{2}(\|y_{n+1}\|^2 - \phi_{n+1}^2) \leq (\|y_n\| - \phi_n) - (\|y_{n+1}\| - \phi_{n+1}).$$

De aquí, un sencillo razonamiento inductivo muestra que

$$\|y_n\| \leq \phi_n, \quad n = 0, 1, \ldots \tag{II.28}$$

Por otro lado, de (II.27) se tiene que

$$\begin{aligned} \phi_n &= \phi_0 - \frac{\nu}{2} \sum_{k=1}^{n} h_{k-1} (\phi_k)^2 \\ &\leq \phi_0 - \frac{\nu}{2} \sum_{k=1}^{n} h_{k-1} (\phi_n)^2 \\ &= \phi_0 - \frac{\nu}{2} t_n (\phi_n)^2. \end{aligned}$$

Por lo tanto sigue que

$$\phi_n \leq \|y_0\| - \frac{\nu}{2} t_n (\phi_n)^2, \quad n = 0, 1, \ldots,$$

lo cual, junto a (II.28) concluye la prueba.

(4) Tomemos

$$y_0 = \tau_0 (v + \theta_0 w_0) \in \mathcal{C}^+_{\alpha,\beta}, \quad w_0 \in v_u^\perp, \quad \theta_0 \geq 0, \ \tau_0 > 0.$$

De (1), para todo $h \geq 0$, tenemos que $y_1 \equiv y_1(h)$ pertenece al cono $\mathcal{C}_{\alpha,\beta}$ y es una función continuamente diferenciable de $h > 0$. En consecuencia podemos escribir

$$y_1 = \tau (v + \theta w) \in \mathcal{C}^+_{\alpha,\beta}, \quad \tau \equiv \tau(h), \ \theta \equiv \theta(h) \geq 0, \quad w \equiv w(h) \in v_u^\perp,$$

siendo $\tau(h)$ una función continua acotada con límite 0 para $h \to \infty$, y $\theta(h)$ una función continua y acotada con valores en el intervalo $[0, \alpha]$.

Insertando esto en (II.24) y considerando (II.2)-(II.3), sigue tras un sencillo cálculo que

$$\tau(v + \theta w) = \tau_0 (v + \theta_0 w_0) + h\tau(\theta J w + \tau N(v + \theta w) + \tau^{-1} R(\tau(v + \theta w))). \tag{II.29}$$

Extrayendo componentes en $span\{v\}$ en la ecuación (II.29), obtenemos que

$$\begin{aligned} \tau &= \tau_0 + h\tau^2 \Big(\langle N(v + \theta w), v \rangle + \tau^{-2} \langle R(\tau(v + \theta w)), v \rangle \Big) \\ &= \tau_0 + h\tau^2 \varsigma_1(h), \end{aligned}$$

siendo

$$\varsigma_1(h) := \langle N(v + \theta w), v \rangle + \tau^{-2} \langle R(\tau(v + \theta w)), v \rangle = -\nu + \mathcal{O}(\theta(h)) + \mathcal{O}(\tau(h))$$

una función acotada para $h \to \infty$ en virtud de $(H4)$ (aquí consideramos sin pérdida de generalidad que α y β son constantes positivas suficientemente pequeñas de modo que $\varsigma_1(h) \geq -\frac{\nu}{2}$, para valores suficientemente grandes de h). Por lo tanto, ya que $\tau_0 \neq 0$, debe ser que la función $h\tau(h)^2$ está acotada para $h \to \infty$ y

$$\lim_{h\to\infty} h\tau(h) = +\infty.$$

Ahora, extrayendo componentes en $v^\perp$ de (II.29), podemos escribir

$$\tau\theta w = \tau_0\theta_0 w_0 + (h\tau\theta)Jw + h\tau^2(N^\perp(v+\theta w) + \tau^{-2}R^\perp(\tau(v+\theta w))),$$

siendo

$$\begin{aligned} N^\perp(v+\theta w) &:= N(v+\theta w) - \langle N(v+\theta w), v\rangle v \\ R^\perp(\tau(v+\theta w)) &:= R(\tau(v+\theta w)) - \langle R(\tau(v+\theta w)), v\rangle v. \end{aligned}$$

Luego

$$\begin{aligned} \tau\theta &= \tau_0\theta_0\langle w_0, w\rangle + (h\tau\theta)\langle Jw, w\rangle + (h\tau^2)\Big(\langle N^\perp(v+\theta w) + \tau^{-2}R^\perp(\tau(v+\theta w)), w\rangle\Big) \\ &= (h\tau\theta)\varsigma_2(h) + \varsigma_3(h), \end{aligned} \tag{II.30}$$

siendo

$$\begin{aligned} \varsigma_2(h) &:= \langle Jw, w\rangle \ (\leq -\delta_1 < 0), \\ \varsigma_3(h) &:= \tau_0\theta_0\langle w_0, w\rangle + (h\tau^2)\Big(\langle N^\perp(v+\theta w) + \tau^{-2}R^\perp(\tau(v+\theta w)), w\rangle\Big), \end{aligned}$$

funciones acotadas de h, para $h \to \infty$.

Puesto que $\lim_{h\to\infty} \tau\theta = 0$, se deduce entonces de (II.30) que la función $h\tau\theta$ está acotada para $h \to \infty$, y en consecuencia debe ser que $\lim_{h\to\infty} \theta = 0$, lo cual concluye la prueba. □

Nota II.4.2 La prueba del enunciado (1) también puede ser deducida de un modo alternativo a partir de la propiedad (II.21) en relación con la teoría del *grado de una aplicación*. Para ello hacemos uso de algunos resultados recogidos en [73, p. 147-165].

Por definición del grado de una aplicación ([73, p. 149-152]) se tiene para la aplicación (II.20) que

$$deg(H(\cdot, 0), Int(\mathcal{C}_{\alpha,\beta}), 0) = 1.$$

Entonces, en virtud de (II.21), el teorema de Invarianza de Homotopía [73, p. 156] garantiza que

$$deg(H(\cdot, t), Int(\mathcal{C}_{\alpha,\beta}), 0) = deg(H(\cdot, 0), Int(\mathcal{C}_{\alpha,\beta}), 0) = 1, \quad \forall t \in [0, 1],$$

y por aplicación del teorema de Kronecker [73, p. 161] podemos asegurar que la ecuación $H(x,1) = 0$ posee al menos una solución $x = y_1 \in \mathcal{C}_{\alpha,\beta}$. Además, esta solución no puede pertenecer a la frontera $\partial\mathcal{C}_{\alpha,\beta}$ debido a (II.21). Esto prueba la existencia de una solución $y_1 \in Int(\mathcal{C}_{\alpha,\beta})$.

Si $y_0 \in \partial\mathcal{C}_{\alpha,\beta}$, entonces considerando argumentos de continuidad respecto del valor inicial y_0 también se puede demostrar además la existencia de solución $y_1 \in \mathcal{C}_{\alpha,\beta}$, así como su continuidad respecto del parámetro h.

II.5. Estabilidad para otros métodos

Como habíamos mencionado anteriormente, existen muy pocos métodos de interés práctico, aparte del método de Euler implícito, que presenten la propiedad de estabilidad incondicional cuando son aplicados a problemas diferenciales bajo las H-hipótesis. Para ilustrar esta afirmación consideramos el sistema (II.5) con $\vartheta = -1$. Ya que este sistema está desacoplado basta considerar únicamente su primera componente, por cuanto la propiedad de A-estabilidad, o estabilidad lineal, de los métodos es suficiente para garantizar integraciones estables sobre la segunda componente

$$\begin{aligned} y'(t) &= -y^2, \\ y(0) &= y_0 > 0, \qquad t \in [0,\infty). \end{aligned} \tag{II.31}$$

Para este problema sabemos que las órbitas estables parten del semieje real positivo y están totalmente contenidas en él.

Si se aplica a este problema la regla implícita del punto medio (son bien conocidas las buenas propiedades de estabilidad lineal y no lineal de este método)

$$y_{n+1} = y_n + h_n f\left(\frac{y_n + y_{n+1}}{2}\right), \quad h_n > 0,\ n = 0, 1, \ldots$$

no es complicado ver que las soluciones propuestas por el método verifican

$$h_n y_{n+1} = \frac{2\left(1 - \dfrac{h_n y_n}{4}\right) h_n y_n}{\sqrt{1 + 2h_n y_n} + 1 + \dfrac{h_n y_n}{2}}.$$

Con esto, aunque las soluciones numéricas estén definidas para todo $h_n > 0$ e $y_n > 0$, se produce inestabilidad, esto es, las soluciones numéricas cruzan al semieje real negativo, para tamaños

de paso verificando $h_n y_n > 4$ ($h_n \equiv t_{n+1} - t_n$). De este modo, resultados como los que se dan en el teorema II.4.1 no son posibles para este método.

El mismo hecho ocurre para la regla trapezoidal

$$y_{n+1} = y_n + \frac{h_n}{2}\left(f(y_n) + f(y_{n+1})\right), \quad h_n > 0,\ n = 0, 1, \ldots$$

En este caso la soluciones numéricas verifican

$$h_n y_{n+1} = \frac{h_n y_n (2 - h_n y_n)}{1 + \sqrt{1 + 2h_n y_n - (h_n y_n)^2}}.$$

De este modo se obtienen integraciones inestables para tamaños de paso tales que $h_n y_n > 2$; más aún, el método no provee solución real y_{n+1} si $h_n y_n > (1 + \sqrt{2})$.

Por otro lado, y como ampliación del caso de la regla trapezoidal, los $\theta-$métodos

$$y_{n+1} = y_n + h_n(\theta f(y_{n+1}) + (1 - \theta) f(y_n)), \tag{II.32}$$

verifican la propiedad de estabilidad lineal para los valores del parámetro $\theta \in [1/2, 1]$. Para $\theta \in [1/2, 1)$ el correspondiente método aplicado al problema (II.31) provee soluciones numéricas verificando

$$h_n y_{n+1} = \frac{2h_n y_n (1 - (1 - \theta) h_n y_n)}{1 + \sqrt{1 + 4\theta h_n y_n - 4\theta(1 - \theta)(h_n y_n)^2}}.$$

En consecuencia, para tamaños de paso h_n tales $h_n y_n > \dfrac{1}{1 - \theta}$ se obtienen soluciones inestables. Además, el método no provee solución real si $h_n y_n > \dfrac{1}{2\sqrt{\theta}(1 - \sqrt{\theta})}$.

Sin embargo para el caso de Euler implícito, esto es, $\theta = 1$, obtenemos

$$h_n y_{n+1} = \frac{2h_n y_n}{1 + \sqrt{1 + 4h_n y_n}},$$

y por lo tanto $0 < y_{n+1} < y_n$ para todo tamaño de paso $h_n > 0$ y valor de arranque $y_n > 0$, lo cual implica la estabilidad incondicional del método sobre este problema, tal como ya conocíamos.

Como último ejemplo ilustrativo, si se considera el método Radau IIA de dos etapas (véase e.g. [53, p.74, Tabla 5.5]) definido por las ecuaciones

$$\begin{aligned} Y &= y_n + h_n\left(\frac{5}{12} f(Y) - \frac{1}{12} f(y_{n+1})\right) \\ y_{n+1} &= y_n + h_n\left(\frac{3}{4} f(Y) + \frac{1}{4} f(y_{n+1})\right) \end{aligned}$$

aplicado al problema (II.31) se obtienen las ecuaciones

$$\begin{aligned} z_n - \frac{5}{12}V^2 + \frac{1}{12}z_{n+1}^2 - V &= 0 \\ \\ z_{n+1} + \frac{3}{4}V^2 + \frac{1}{4}z_{n+1}^2 - z_n &= 0, \end{aligned}$$

siendo $V = h_n Y$, $z_n = h_n y_n$ y $z_{n+1} = h_n y_{n+1}$. Con ayuda de un manipulador algebraico, se concluye que las dos ecuaciones anteriores junto con las condiciones de positividad $z_n > 0$, $z_{n+1} > 0$ determinan para el tamaño de paso h_n la condición $0 < h_n y_n < \frac{27}{4}$. Así el método Radau IIA provee soluciones inestables para $h_n y_n > \frac{27}{4}$. El estudio por medio del método de los multiplicadores de Lagrange del máximo valor de z_n que permite obtener soluciones reales para las dos ecuaciones anteriores definidas por el Radau IIA define una tercera ecuación

$$\frac{2}{3}V z_{n+1} + \frac{5}{6}V + \frac{1}{2}z_{n+1} + 1 = 0,$$

que junto a las dos ecuaciones previas define un valor máximo $z_n = z_{max} \simeq 8.913112943350903$. Luego el método Radau IIA no provee soluciones reales para valores del tamaño de paso h_n tales que $h_n y_n > z_{max}$.

Aunque estos comentarios pudieran parecer algo pesimistas sobre la fiabilidad de los métodos numéricos sobre esta clase de problemas, no se debería concluir a partir de estos hechos que las propiedades de estabilidad lineal y no lineal de los métodos no son suficientes para garantizar que un método con tales propiedades efectúe integraciones satisfactorias (estables) cuando se aplica a problemas diferenciales bajo las H-hipótesis, sino que estas propiedades no bastan para asegurar la estabilidad incondicional de los métodos.

Otro importante hecho a subrayar es que aunque la propiedad de A-estabilidad (o estabilidad lineal) de los métodos numéricos es suficiente para obtener integraciones estables sobre la clase particular de problemas (II.5) en redes temporales del tipo

$$0 < h_{n+1}/h_n \leq r^*, \quad h_n = t_{n+1} - t_n, \quad n = 0, 1, \ldots \tag{II.33}$$

para determinado $r^* > 1$, no es suficiente para alcanzar integraciones estables sobre sistemas generales bajo las H-hipótesis.

A modo de ilustración consideramos la integración del sistema diferencial

$$\begin{aligned} y_1' &= -\lambda y_1^2 - y_2^2, & \lambda &= 10^{-10} \\ y_2' &= -\eta y_2, & \eta &= 10^6, \qquad t \in [0, 10^{25}], \end{aligned} \tag{II.34}$$

con condiciones iniciales $y_1(0) = y_2(0) = 1$.

El origen de coordenadas $(0,0)^T$ es el único punto de equilibrio para este problema, y de hecho se ajusta a las A/H-hipótesis por cuanto la matriz Jacobiana del sistema en dicho punto posee por espectro a $\{0, -\eta\}$, con autovectores derecho e izquierdo asociados al autovalor nulo dados respectivamente por $v = p = e_1$, donde e_1 denota el primer vector de la base canónica de $\mathbb{R}^2$, y la constante

$$\vartheta := \frac{p^T N(v)}{p^T v} = -\lambda$$

es no nula.

Este problema (II.34) ha sido integrado por medio de los θ-métodos (II.32) con tamaño de paso inicial $h_0 = 10^{-6}$. Tras dos pasos de igual longitud h, el error local en el punto $t_n + 2h$ ha sido estimado por medio de la estrategia usual de extrapolación local (aquí $h \equiv h_n$ es el tamaño de paso propuesto)

$$est = \left\| y^{(2h)}(t_n + 2h) - y^{(h,h)}(t_n + 2h) \right\|_\infty / (2^p - 1), \tag{II.35}$$

donde p denota el orden del método, $y^{(2h)}(t_n + 2h)$ la solución numérica después de un paso desde t_n de tamaño $2h$, e $y^{(h,h)}(t_n + 2h)$ la solución numérica después de dos pasos consecutivos de tamaño h. El tamaño de paso actual $h = h_n$ es aceptado e incrementado de modo estándar cuando la estimación del error est no supera una tolerancia de error $toler$ prefijada, $est \leq toler$,

$$h_{n+1} = \min\{2,\ 0.9 * (toler/est)^{1/(p+1)}\} h_n. \tag{II.36}$$

En caso de rechazo del tamaño de paso por el estimador, $est > toler$, procedemos a una reducción del tipo

$$h_{n+1} = \max\{0.1,\ 0.9 * (toler/est)^{1/(p+1)}\} h_n. \tag{II.37}$$

Hemos de señalar que para la acotación del error hemos considerado un criterio de tolerancia mixta

$$toler = (TOL/1000) * \|y_n\| + TOL, \tag{II.38}$$

que permite obtener una cota de error absoluto igual a TOL, y una cota de error relativo igual a $TOL/1000$, siendo TOL un parámetro a introducir ($TOL = 10^{-1}, 10^{-2}, 10^{-3}, \ldots$). Por otro lado, la ecuación implícita que define al correspondiente $\theta-$método es resuelta aproximadamente por medio de iteraciones tipo Newton simplificado.

En las tablas II.1-II.4 apreciamos que los θ-métodos con $\theta > 1/2$ proveen integraciones satisfactorias en lo que a estabilidad se refiere en todos los casos. No obstante, para el caso

$\theta = 0.5$, esto es, la regla trapezoidal, se obtienen integraciones inestables, en el sentido de que la primera componente de la solución llega a ser negativa en algún momento de la integración (y a partir de este punto la solución numérica escapa al infinito), para todas las tolerancias de error $TOL = 10^{-1}, 10^{-2}, \ldots, 10^{-8}$ (véase la tabla II.5).

En estas tablas NPA representa el número de pasos necesarios para concluir una integración satisfactoria; $NPRE$, el número de pasos rechazados por el estimador; $NPRC$, el número de pasos rechazados debido a fallos en la convergencia de la iteración del método de Newton simplificado; *Rmedio* representa el promedio de las razones de paso a lo largo de toda la integración; T_{inest}, el punto temporal en el que la primera componente de la solución numérica se hizo negativa; T_f, indica que el punto final del intervalo de integración fue alcanzado sin problemas de estabilidad; $Y_j(T_-),\ j = 1, 2$, expresa la solución en el correspondiente punto temporal; mientras que $Y_2(T_f)/Y_1(T_f)$ representa el valor de la pendiente numérica de la solución que provee el método en el punto final del intervalo de integración.

Podemos notar en estas tablas que las elecciones $\theta = 1, 0.7, 0.6, 0.55$ proveen en todo caso integraciones estables (al menos para tolerancias medias de error, $TOL = 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}$), aunque la precisión en la solución numérica es algo baja por cuanto se trata de métodos de orden de consistencia 1. Además únicamente un dígito significativo está garantizado. Apreciamos además que el número de pasos necesarios para concluir la integración disminuye a medida que el parámetro θ se reduce de 1 a 0.5. Sin embargo, se aprecia que la regla trapezoidal no es en absoluto fiable para la integración de este problema, a pesar de que dicho método posee orden de consistencia 2, posee las propiedades de estabilidad lineal y no lineal (A-estabilidad y B-estabilidad), y da menos pasos que los otros métodos para los mismos valores del parámetro TOL que define la tolerancia de error. Además debemos notar que en el caso de integrar con la regla trapezoidal se logra mantener la estabilidad en un intervalo temporal más amplio al hacer más severa la tolerancia de error; aunque incluso para tolerancias exigentes ($TOL = 10^{-9}, 10^{-10}, 10^{-11}, 10^{-12}$) no se evita que el método llegue a ser inestable en valores temporales más allá de 10^{25}.

Por otro lado, y de acuerdo al Teorema II.3.8, la solución exacta para el problema (II.34) debe tender en tiempo infinito al origen de coordenadas como punto de equilibrio del sistema diferencial formando ángulo cero con el eje de abscisas $y_2 = 0$, esto es, la solución exacta en tiempo infinito forma angulo nulo con el autovector derecho asociado al autovalor nulo. En las tablas (II.1)-(II.4) observamos que las soluciones numéricas provistas por los θ-métodos, con

$\theta > 0.5$, satisfacen que sus pendientes numéricas y_2/y_1 se ajustan satisfactoriamente al modelo teórico, siendo aquellos métodos con θ próximo a 1 quienes mejor reproducen este comportamiento. De hecho, para todas las tolerancias de error consideradas notamos que las soluciones numéricas obtenidas por medio del método de Euler implícito verifican que su pendiente coincide exactamente con el valor teórico.

Prácticamente todo método numérico de un paso aplicado al problema lineal modelo $y' = \lambda y$ provee como solución numérica

$$y_1 = R(z)y_0, \quad z = h\lambda$$

donde $R(z)$ denota la denominada *función de estabilidad lineal* del método, siendo ésta usualmente una función racional de la variable z. La función de estabilidad lineal para los $\theta-$métodos viene dada por

$$R(z,\theta) = \frac{1 + z(1-\theta)}{1 - z\theta}$$

con valor en el infinito

$$R(\infty,\theta) = \frac{\theta - 1}{\theta}. \tag{II.39}$$

Así, podemos intuir que aquellos métodos que poseen una función de estabilidad lineal con valor absoluto en el infinito suficientemente alejado de 1, serán los que provean soluciones numéricas con un mejor comportamiento cualitativo.

Esto no es mera casualidad y de hecho probaremos en el siguiente capítulo que cualquier método Runge-Kutta A-estable verificando que $|R(\infty)| < 1$, donde $R(z)$ denota la función de estabilidad lineal del método, esto es, cualquier método fuertemente A-estable, integra de modo estable este tipo de problemas bajo las H-hipótesis si las redes temporales satisfacen (II.33)

Otra conclusión práctica es que el estudio de estabilidad de los métodos numéricos aplicados a (II.5) no es relevante para totalidad de la clase de sistemas diferenciales bajo las H-hipótesis. Véase a modo de ejemplo que la regla trapezoidal es estable para el problema (II.5) sobre redes del tipo (II.33).

Las gráficas II.10–II.14 corresponden a la integración del problema (II.34) por medio de los $\theta-$métodos, para los valores de $\theta = 1, 0.7, 0.6, 0.55, 0.5$, con valor del parámetro $TOL = 10^{-5}$. Observamos que aquellos métodos con un menor valor para $|R(\infty,\theta)|$ (esto es, θ alejado de 0.5) producen un más adecuado amortiguamiento en la segunda componente de la solución numérica, y en consecuencia, reproducen de modo más satisfactorio el comportamiento de la pendiente de la solución en el equilibrio. Para este valor $TOL = 10^{-5}$, la regla trapezoidal

provee soluciones inestables aproximadamente a partir de $T = 4.3{\cdot}10^{14}$ y ello queda reflejado en la gráfica II.14.

Tabla II.1: Método de Euler Implícito ($\theta = 1$).

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-1}	216	0	1	1.4775	$0.1571{\cdot}10^{-14}$	0	0
10^{-2}	248	3	1	1.4229	$0.1565{\cdot}10^{-14}$	0	0
10^{-3}	360	3	1	1.2910	$0.1568{\cdot}10^{-14}$	0	0
10^{-4}	736	3	0	1.1411	$0.1580{\cdot}10^{-14}$	0	0
10^{-5}	1956	3	0	1.0529	$0.1579{\cdot}10^{-14}$	0	0
10^{-6}	5832	3	0	1.0177	$0.1615{\cdot}10^{-14}$	0	0
10^{-7}	18106	4	0	1.0057	$0.1569{\cdot}10^{-14}$	0	0
10^{-8}	56940	4	0	1.0018	$0.1590{\cdot}10^{-14}$	0	0

Tabla II.2: $\theta = 0.7$.

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-1}	212	0	1	1.4854	$0.1180 \cdot 10^{-14}$	$0.1446{\cdot}10^{-81}$	$0.1226{\cdot}10^{-66}$
10^{-2}	246	4	1	1.4243	$0.1171{\cdot}10^{-14}$	$0.4006{\cdot}10^{-95}$	$0.3421{\cdot}10^{-80}$
10^{-3}	318	4	1	1.3357	$0.1165{\cdot}10^{-14}$	$0.4071{\cdot}10^{-107}$	$0.3495{\cdot}10^{-92}$
10^{-4}	534	4	1	1.1971	$0.1192{\cdot}10^{-14}$	$0.1328{\cdot}10^{-150}$	$0.1114{\cdot}10^{-135}$
10^{-5}	1286	3	0	1.0808	$0.1168{\cdot}10^{-14}$	$0.1173{\cdot}10^{-309}$	$0.1004{\cdot}10^{-294}$
10^{-6}	3730	3	0	1.0277	$0.1176{\cdot}10^{-14}$	0	0
10^{-7}	11486	4	0	1.0090	$0.1165{\cdot}10^{-14}$	0	0
10^{-8}	36034	4	0	1.0029	$0.1195{\cdot}10^{-14}$	0	0

Tabla II.3: $\theta = 0.6$.

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-1}	214	0	1	1.4843	$0.1089{\cdot}10^{-14}$	$0.1346{\cdot}10^{-43}$	$0.1236{\cdot}10^{-28}$
10^{-2}	242	4	1	1.4328	$0.1087{\cdot}10^{-14}$	$0.4002{\cdot}10^{-48}$	$0.3681{\cdot}10^{-33}$
10^{-3}	248	4	1	1.4240	$0.1078{\cdot}10^{-14}$	$0.3499{\cdot}10^{-49}$	$0.3247{\cdot}10^{-34}$
10^{-4}	404	3	0	1.2595	$0.1080{\cdot}10^{-14}$	$0.2604{\cdot}10^{-64}$	$0.2412{\cdot}10^{-49}$
10^{-5}	936	3	0	1.1113	$0.1077{\cdot}10^{-14}$	$0.9655{\cdot}10^{-117}$	$0.8964{\cdot}10^{-102}$
10^{-6}	2658	3	0	1.0393	$0.1089{\cdot}10^{-14}$	$0.2453{\cdot}10^{-294}$	$0.2253{\cdot}10^{-279}$
10^{-7}	8134	3	0	1.0127	$0.1076{\cdot}10^{-14}$	$0.4941{\cdot}10^{-323}$	$0.4593{\cdot}10^{-308}$
10^{-8}	25488	4	0	1.0040	$0.1081{\cdot}10^{-14}$	$0.4941{\cdot}10^{-323}$	$0.4572{\cdot}10^{-308}$

Tabla II.4: $\theta = 0.55$

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-1}	212	0	1	1.4878	$0.1037{\cdot}10^{-14}$	$0.7839{\cdot}10^{-24}$	$0.7563{\cdot}10^{-9}$
10^{-2}	228	4	1	1.4590	$0.1036{\cdot}10^{-14}$	$0.1322{\cdot}10^{-25}$	$0.1276{\cdot}10^{-10}$
10^{-3}	242	3	1	1.4355	$0.1043{\cdot}10^{-14}$	$0.3545{\cdot}10^{-27}$	$0.3399{\cdot}10^{-12}$
10^{-4}	322	14	0	1.3343	$0.1036{\cdot}10^{-14}$	$0.1359{\cdot}10^{-30}$	$0.1311{\cdot}10^{-15}$
10^{-5}	680	13	0	1.1554	$0.1037{\cdot}10^{-14}$	$0.2453{\cdot}10^{-49}$	$0.2366{\cdot}10^{-34}$
10^{-6}	1880	14	0	1.0569	$0.1038{\cdot}10^{-14}$	$0.5058{\cdot}10^{-111}$	$0.4874{\cdot}10^{-96}$
10^{-7}	5736	12	0	1.0184	$0.1038{\cdot}10^{-14}$	$0.2522{\cdot}10^{-308}$	$0.2429{\cdot}10^{-293}$
10^{-8}	17992	15	0	1.0058	$0.1037{\cdot}10^{-14}$	$0.9881{\cdot}10^{-323}$	$0.9529{\cdot}10^{-308}$

Tabla II.5: Regla Trapezoidal ($\theta = 0.5$).

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_1(T_{inest})$	$Y_2(T_{inest})$	T_{inest}
10^{-1}	116	0	1	1.4548	$-0.6263{\cdot}10^{-1}$	$0.1549{\cdot}10^{-5}$	$0.12{\cdot}10^{12}$
10^{-2}	130	3	1	1.4218	$-0.6736{\cdot}10^{-2}$	$0.3594{\cdot}10^{-6}$	$0.47{\cdot}10^{12}$
10^{-3}	162	4	1	1.3696	$-0.3208{\cdot}10^{-3}$	$0.1730{\cdot}10^{-7}$	$0.99{\cdot}10^{13}$
10^{-4}	232	4	2	1.2785	$-0.2502{\cdot}10^{-4}$	$0.7476{\cdot}10^{-9}$	$0.25{\cdot}10^{15}$
10^{-5}	364	4	1	1.1781	$-0.1340{\cdot}10^{-5}$	$0.3656{\cdot}10^{-9}$	$0.43{\cdot}10^{15}$
10^{-6}	664	4	1	1.1047	$-0.3295{\cdot}10^{-6}$	$0.1030{\cdot}10^{-10}$	$0.18{\cdot}10^{17}$
10^{-7}	1294	5	0	1.0538	$-0.1001{\cdot}10^{-7}$	$0.3644{\cdot}10^{-11}$	$0.43{\cdot}10^{17}$
10^{-8}	2660	5	1	1.0276	$-0.9506{\cdot}10^{-8}$	$0.2989{\cdot}10^{-12}$	$0.61{\cdot}10^{18}$

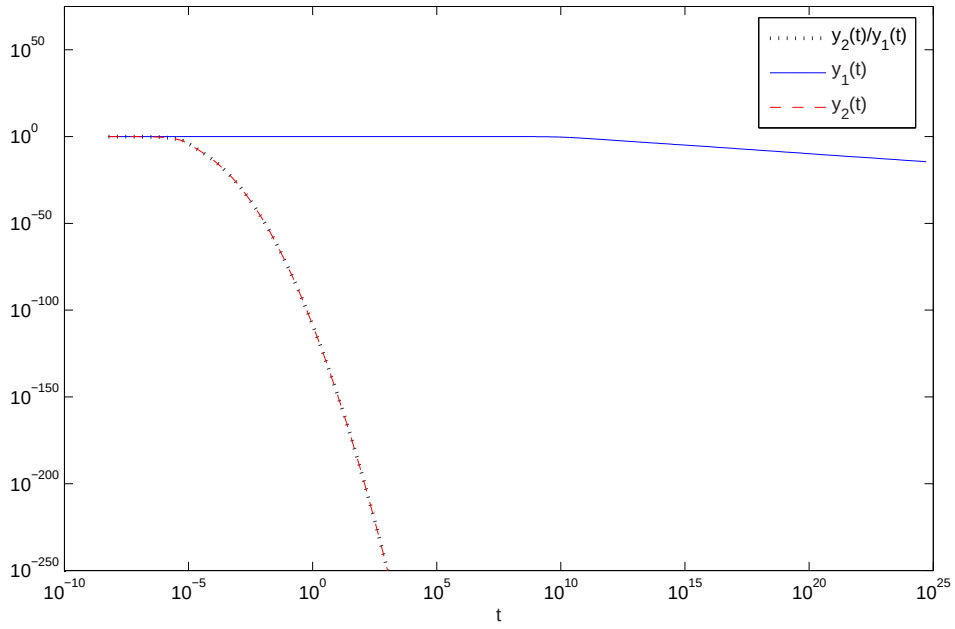

Figura II.10: $\theta = 1$. Componentes de la solución y pendiente numérica (escala logarítmica).

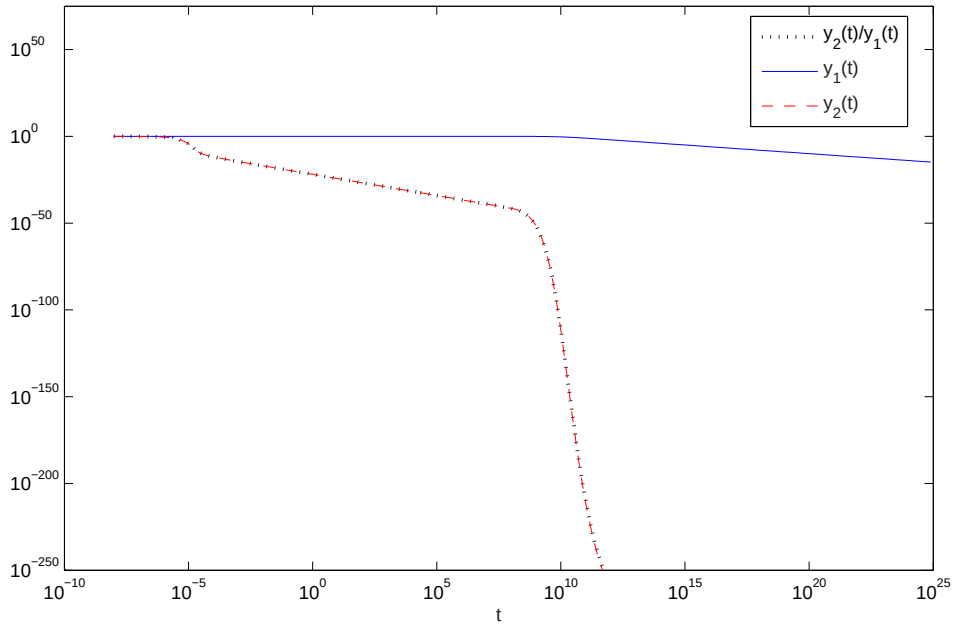

Figura II.11: $\theta = 0.7$. Componentes de la solución y pendiente numérica (escala logarítmica).

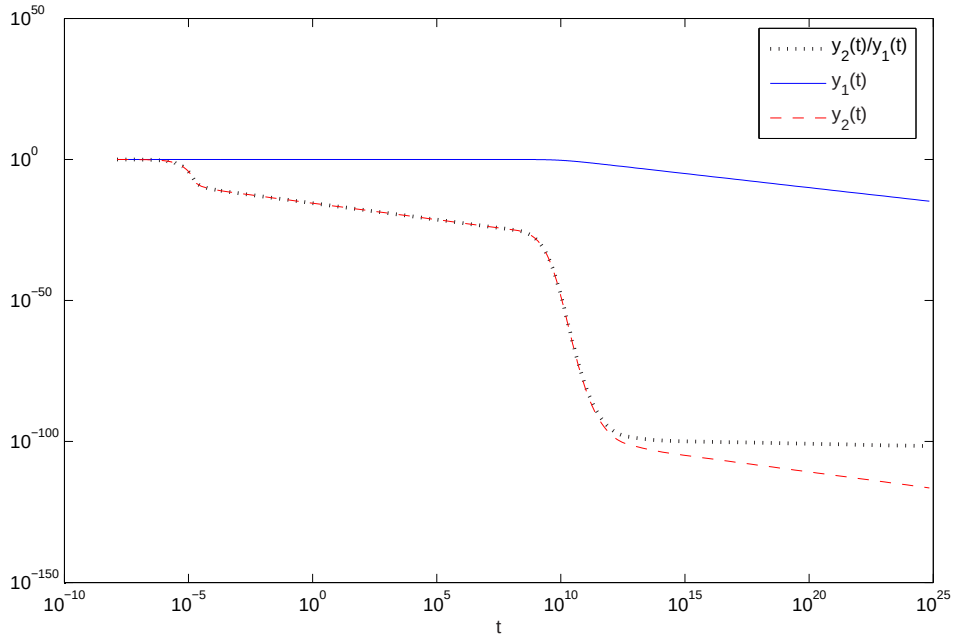

Figura II.12: $\theta = 0.6$. Componentes de la solución y pendiente numérica (escala logarítmica).

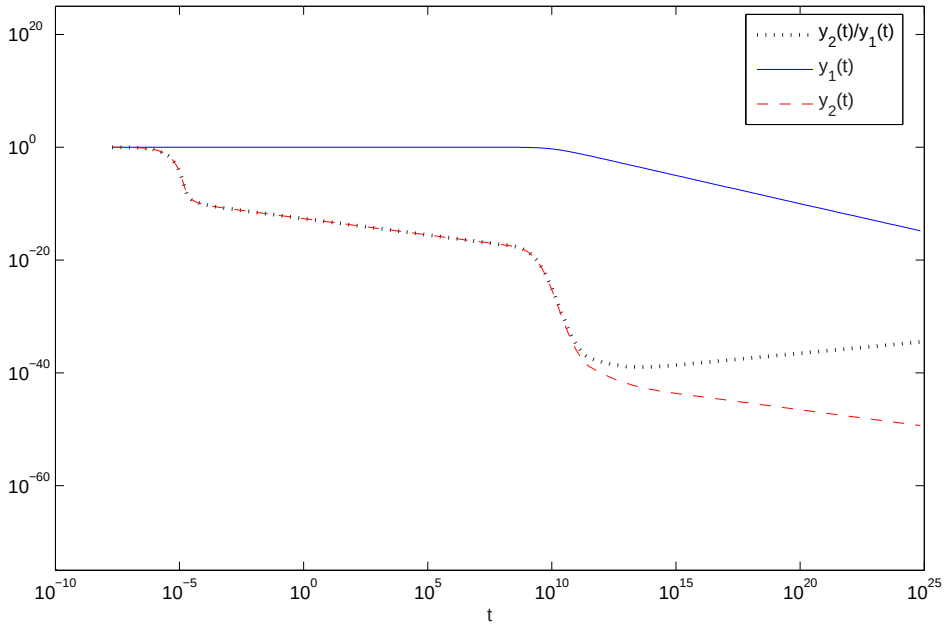

Figura II.13: $\theta = 0.55$. Componentes de la solución y pendiente numérica (escala logarítmica).

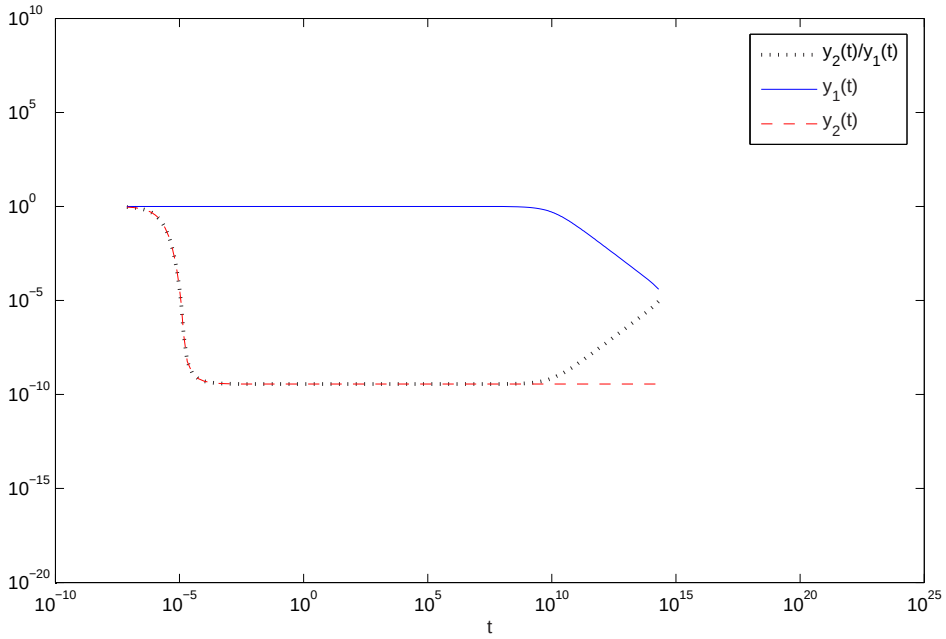

Figura II.14: $\theta = 0.5$. Componentes de la solución y pendiente numérica (escala logarítmica).

Capítulo III

Estabilidad de los métodos Runge-Kutta sobre sistemas diferenciales con equilibrios semiestables.

III.1. Consideraciones preliminares

En el capítulo previo consideramos la integración numérica de sistemas diferenciales (II.1) con un punto de equilibrio *semiestable* $y = 0$, donde el término equilibrio *semiestable* hace referencia a puntos de equilibrio inestables en el sentido de Lyapunov, pero que poseen la propiedad de que en todo entorno de estos puntos coexisten órbitas estables e inestables. Asimismo, mostramos que el método de Euler implícito es de los pocos métodos numéricos de un paso de interés práctico con la propiedad de ser *incondicionalmente estable* para la integración de problemas diferenciales del tipo (II.1) bajo las *H-hipótesis*. También se comentó el hecho de que problemas interesantes en la literatura stiff provenientes de las más diversas aplicaciones, como por ejemplo el problema de Robertson [53, p. 144] o la reacción química *E5* [53, p.145] se encuadran dentro del marco hipotético de las *H-hipótesis*, y de aquí que los resultados de estabilidad enunciados en el capítulo previo puedan aplicarse a este tipo de problemas.

En este capítulo describiremos una manera de solventar la mencionada barrera de *estabilidad incondicional* de cara a obtener estabilidad para un mayor número de métodos. La estrategia que será propuesta podrá ser aplicada a una amplia gama de métodos numéricos, a pesar de

que la mayoría de ellos no verifique ser *incondicionalmente estable*. En cierto sentido, este hecho justificará la razón por la que muchos métodos numéricos integran de modo estable la clase de problemas en consideración. Debemos comentar que los resultados que se mostrarán en este capítulo están parcialmente recogidos en los trabajos [44] y [45].

Aunque los resultados que se obtendrán en el desarrollo posterior pueden ser demostrados bajo las *H-hipótesis*, asumiremos en lo que sigue, por simplicidad en los detalles técnicos en las demostraciones, que los términos de tercer orden en el desarrollo de Taylor de la función f que define (II.1) (véase (II.2)-(II.3)) se anulan; esto es, asumiremos en lo que sigue que $R(y) \equiv 0$.

El resto del capítulo se organiza del modo siguiente. En la sección 2 introduciremos el concepto de *E-estabilidad*, que establece un marco adecuado para solventar la barrera de *estabilidad incondicional* de los métodos numéricos. En la sección 3, se discutirá la resolubilidad de las etapas implícitas que definen a los métodos de tipo Runge-Kutta sobre la clase de problemas bajo las H-hipótesis. En la sección cuarta del capítulo estableceremos una caracterización sencilla de la A-aceptabilidad fuerte de las funciones racionales, y que será de gran utilidad de cara a demostrar el principal resultado relativo a la estabilidad de los métodos, resultado éste que será demostrado en la sección 5. En esta misma sección se darán condiciones para incluir a una amplia clase de métodos Runge-Kutta implícitos. Finalmente, la sección 6 se dedicará a la ilustración numérica de los resultados teóricos establecidos así como de la necesidad de las condiciones impuestas para la obtención de la estabilidad de los métodos.

III.2. E-estabilidad

Consideremos un método de un paso

$$y_{n+1} = \varphi_{h_n}(y_n),\ t_0 = 0,\ h_n = t_{n+1} - t_n > 0, \quad n = 0, 1, \ldots, \tag{III.1}$$

donde se asume que función φ_h satisface $\varphi_h(0) \equiv 0$ ya que $f(0) = 0$. Continuando en la línea del capítulo previo, un método de un paso aplicado a sistemas diferenciales (II.1)-(II.3) bajo las H-hipótesis se dice *incondicionalmente estable* si existe un cono $\mathcal{C}_{\alpha,\beta}$ (II.14) tal que

$$y_1 = \varphi_h(y_0) \in \mathcal{C}^+_{\alpha,\beta} \quad \text{y} \ \|y_1\| < \|y_0\|$$

cuando $h > 0$ e $y_0 \in \mathcal{C}^+_{\alpha,\beta}$.

En el capítulo previo se demostró que el método de Euler implícito es prácticamente el único método de interés con la propiedad de *estabilidad incondicional*. Obsérvese que esta última propiedad garantiza integraciones estables sin restricción alguna sobre el tamaño de paso. Esto resulta de especial importancia por cuanto estamos involucrados en la integración de sistemas diferenciales en intervalos temporales extremadamente amplios, $[0,T],\ T \gg 1$, o inclusive, $T = \infty$.

Con vistas a solventar la *barrera de estabilidad incondicional*, podemos de forma natural preguntarnos por la estabilidad de las integraciones sobre redes $\mathcal{P} = \{t_n,\ n = 0,1,\ldots\}$ satisfaciendo

$$0 < r_n = h_{n+1}/h_n \leq r^*, \qquad h_n = t_{n+1} - t_n, \quad n = 0,1,2,\ldots,$$

para alguna constante $r^* > 1$. La anterior consideración es bastante realista en la medida de que los códigos numéricos usuales en la práctica hacen uso de una condición de este tipo para conseguir integraciones satisfactorias en intervalos temporales amplios. En consecuencia, proponemos la búsqueda de métodos cumpliendo la siguiente propiedad:

Existe un cono $\mathcal{C}_{\alpha,\beta}$ y dos constantes $\gamma > 0,\ r^* > 1$ tales que para todo $y_0 \in \mathcal{C}^+_{\alpha,\beta}$ satisfaciendo $h_0 \|y_0\| \leq \gamma$, se tiene que

$$y_n \in \mathcal{C}^+_{\alpha,\beta},\ \forall n \in \mathbb{N}, \quad \forall \mathcal{P} \in \mathrm{P}_{r^*} := \{\{t_n\}_{n=0}^{\infty} \,:\, r_n = h_{n+1}/h_n \leq r^*\} \tag{III.2}$$

$$\lim_{t_n \to \infty} y_n = 0. \tag{III.3}$$

Nota III.2.1 Dado el problema modelo simple

$$\left\{ \begin{array}{rcl} \eta' & = & -\kappa\eta^2, \quad \kappa > 0 \\ \eta(0) & = & \eta_0 > 0 \end{array} \right.$$

tenemos para su flujo $\phi_t(\eta)$ que

$$\phi_h(\eta_0) = \frac{\eta_0}{1 + \kappa h \eta_0}.$$

Sería deseable para un método numérico del tipo (III.1) respecto a este sencillo problema modelo un comportamiento de la forma

$$0 \leq \varphi_h(\eta_0) \leq \phi_h(\eta_0), \quad \forall h \geq 0,\ \forall \eta_0 > 0.$$

Esto implicaría debido a la monotonía y aditividad del flujo que tras dos pasos consecutivos del método se tuviera

$$\begin{array}{rcl} 0 \leq \varphi_{h_2}(\varphi_{h_1}(\eta_0)) & \leq & \phi_{h_2}(\varphi_{h_1}(\eta_0)) \\ & \leq & \phi_{h_2}(\phi_{h_1}(\eta_0)) \\ & = & \phi_{h_1+h_2}(\eta_0). \end{array}$$

Por lo tanto, tras k pasos consecutivos, si $T = \sum_{i=1}^{k} h_i$, se tendría que

$$0 \leq \varphi_T(\eta_0) \leq \phi_T(\eta_0)$$

y de aquí que $\varphi_T(\eta_0) \to 0$, si $T \to \infty$.

Con vistas a obtener métodos satisfaciendo las propiedades (III.2)-(III.3), y teniendo en cuenta la idea subyacente en la nota anterior, establecemos la definición de *E-estabilidad*, que, en cierto modo, quiere reflejar la estabilidad de los métodos en entornos de los puntos de equilibrio de sistemas diferenciales bajo las H-hipótesis.

Definición III.2.2 *[E-estabilidad] Bajo las H-hipótesis, un método de un paso φ_h se dice E-estable si existen cuatro constantes positivas $\gamma, \kappa, \alpha, \beta$ (únicamente dependientes de las constantes involucradas en las H-hipótesis y de los coeficientes del método) tales que*

$$y_0 \in \mathcal{C}^+_{\alpha,\beta}, \quad y \quad h\,\|y_0\| \leq \gamma,\ h > 0$$

implica

$$y_1 \in \mathcal{C}^+_{\alpha,\beta}, \quad y \quad h\,\|y_1\| \leq h\,\|y_0\|\,(1 - \kappa h\,\|y_0\|),$$

donde $y_1 := \varphi_h(y_0)$.

A partir de la definición previa debe quedar claro que

$$\|y_1\| < \|y_0\| \leq \beta.$$

Seguidamente y como principal resultado de esta sección, enunciamos y demostramos un teorema que garantiza el buen comportamiento de estabilidad para los métodos E-estables en el sentido de la definición anterior.

Teorema III.2.3 *Para un método E-estable, existen un cono $\mathcal{C}_{\alpha,\beta}$ y dos constantes $\gamma > 0,\ r^* > 1$ tales que si $h_0\,\|y_0\| \leq \gamma$ e $y_0 \in \mathcal{C}^+_{\alpha,\beta}$, entonces, para redes temporales arbitrarias $\mathcal{P} \in \mathrm{P}_{r^*}$, se tiene que*

1. *$y_n \in \mathcal{C}^+_{\alpha,\beta},\ n = 1, 2, \ldots$*
2. *$\|y_{n+1}\| < \|y_n\|\,,\ n = 0, 1, 2, \ldots$*
3. *$\|y_n\| \leq z_n,\ n \geq 0$, donde z_n denota la única raíz real de la ecuación*

$$z = \|y_0\| \exp\left(-\kappa t_n z\right). \tag{III.4}$$

Además, z_n está acotada del modo siguiente,

$$\frac{\|y_0\|}{1 + \kappa t_n\,\|y_0\|} \leq z_n \leq \frac{\|y_0\|\,(1 + \ln\left(1 + \kappa t_n\,\|y_0\|\right))}{1 + \kappa t_n\,\|y_0\|}. \tag{III.5}$$

Aquí, las constantes positivas $\alpha, \beta, \gamma, \kappa$ están dadas por la definición III.2.2.

Obsérvese que el enunciado (3) en el Teorema III.2.3 implica que $\lim_{t_n \to \infty} y_n = 0$, independientemente de la red temporal elegida $\mathcal{P} \in \mathrm{P}_{r^*}$.

Demostración. Consideremos el cono $\mathcal{C}_{\alpha,\beta}$ y las constantes $\gamma > 0$, $\kappa > 0$ dadas en la definición III.2.2. Definamos

$$r^* := \begin{cases} 4\kappa\gamma, & \text{si } \gamma \geq (2\kappa)^{-1} \\ (1-\kappa\gamma)^{-1}, & \text{en otro caso.} \end{cases}$$

Notemos que $r^* \in (1,2)$ si $\gamma < (2\kappa)^{-1}$, y que $r^* \geq 2$ si $\gamma \geq (2\kappa)^{-1}$.

Consideremos ahora una red temporal $\mathcal{P} \in \mathrm{P}_{r^*}$. Para demostrar los enunciados (1) y (2), es suficiente con mostrar que

$$h_n \|y_n\| \leq \gamma \quad \Rightarrow \quad h_{n+1} \|y_{n+1}\| \leq \gamma, \quad \text{para } n = 0, 1, \ldots. \tag{III.6}$$

Basta considerar el caso $n = 0$ ya que para $n = 1, 2 \ldots$ podemos proceder de modo análogo por un proceso inductivo. Por definición de E-estabilidad se tiene que

$$\begin{aligned} h_1 \|y_1\| &= r_0 h_0 \|y_1\| \\ &\leq r_0(1 - \kappa h_0 \|y_0\|) h_0 \|y_0\| . \end{aligned} \tag{III.7}$$

Teniendo en cuenta (III.7) y que $(1-\kappa t)t$ es una función creciente de t en $[0, (2\kappa)^{-1}]$ deducimos que si $0 < \gamma < (2\kappa)^{-1}$ entonces

$$\begin{aligned} h_1 \|y_1\| &\leq r_0(1-\kappa\gamma)\gamma \\ &\leq r^*(1-\kappa\gamma)\gamma = \gamma \end{aligned}$$

Del mismo modo, ya que $(1-\kappa t)t$ decrece para $t \geq (2\kappa)^{-1}$ se obtiene que si $\gamma \geq (2\kappa)^{-1}$ entonces

$$\begin{aligned} h_1 \|y_1\| &\leq r_0(1 - \kappa(2\kappa)^{-1})(2\kappa)^{-1} \\ &\leq r^*(4\kappa)^{-1} = \gamma \end{aligned}$$

Con esto concluimos inmediatamente la prueba de los enunciados (1) y (2).

Para demostrar (3), tenemos de (III.6) que

$$h_{n-1} \|y_n\| < h_{n-1} \|y_{n-1}\| \leq \gamma, \quad n = 1, 2, \ldots$$

Además, la E-estabilidad del método implica que

$$\begin{aligned} \|y_n\| &\leq (1 - \kappa h_{n-1} \|y_{n-1}\|) \|y_{n-1}\| \\ &\leq \exp(-\kappa h_{n-1} \|y_{n-1}\|) \|y_{n-1}\| , \quad n = 1, 2, \ldots \end{aligned}$$

Considerando por reiteración la última desigualdad llegamos a que

$$\begin{aligned}\|y_n\| &\leq \exp\left(-\kappa h_{n-1}\|y_{n-1}\|\right)\|y_{n-1}\| \\ &\leq \exp\left(-\kappa h_{n-1}\|y_{n-1}\|\right)\exp\left(-\kappa h_{n-2}\|y_{n-2}\|\right)\|y_{n-2}\| \\ &\leq \exp\left(-\kappa h_{n-1}\|y_{n-1}\|\right)\cdot\ldots\cdot\exp\left(-\kappa h_0\|y_0\|\right)\|y_0\| \\ &\leq \exp\left(-\kappa h_{n-1}\|y_n\|\right)\cdot\ldots\cdot\exp\left(-\kappa h_0\|y_n\|\right)\|y_0\|.\end{aligned}$$

De este modo se obtiene entonces que

$$\|y_n\| \leq \exp\left(-\kappa t_n\|y_n\|\right)\|y_0\|, \quad n = 0, 1, 2, \ldots \tag{III.8}$$

Notemos ahora que la función $e(z) := z - \|y_0\| e^{-\kappa t_n z}$ representa una biyección de la recta real en sí misma, y en consecuencia la ecuación (III.4) posee un único cero $z_n \in \mathbb{R}$. Entonces partiendo de (III.4), (III.8) y el Teorema del Valor Medio, se deduce que

$$\begin{aligned}\|y_n\| - z_n &\leq \|y_0\|\left(\exp\left(-\kappa t_n\|y_n\|\right) - \exp\left(-\kappa t_n z_n\right)\right) \\ &= -\kappa t_n\|y_0\|\exp\left(-\kappa t_n\zeta_n\right)(\|y_n\| - z_n),\end{aligned}$$

donde ζ_n pertenece al intervalo real determinado por $\|y_n\|$ y z_n. Esto implica que $\|y_n\| \leq z_n$.

Con vistas a demostrar (III.5), consideramos $t_n > 0$ y definimos la función estrictamente creciente

$$\phi(z) := \ln\frac{z}{\|y_0\|} + \kappa t_n z, \quad z > 0.$$

Se debe notar que $z = z_n$ es precisamente la única raíz real de ϕ, y, más aún, $\lim_{z\to 0^+}\phi(z) = -\infty$.

Consideremos además las funciones

$$\begin{aligned}e_1(x) &:= \ln\left(\frac{1}{1+x}\right) + \frac{x}{1+x} < 0, \quad \forall x \in (0, +\infty) \\ e_2(x) &:= \ln(1+\ln(1+x)) + \frac{x - \ln(1+x)}{1+x} > 0, \quad \forall x \in (0, +\infty).\end{aligned}$$

Sea también $s_n := \kappa t_n\|y_0\| > 0$. Entonces, teniendo en cuenta que $e_2(x)$ puede ser expresado como

$$e_2(x) = \ln\left(\frac{1+\ln(1+x)}{1+x}\right) + \frac{x}{1+x}(1+\ln(1+x)),$$

es sencillo comprobar que tomando $z_n^* := \dfrac{\|y_0\|}{1+s_n}$ y $\overline{z}_n = \dfrac{\|y_0\|}{1+s_n}(1+\ln(1+s_n))$ se obtiene que

$$\phi(z_n^*) = e_1(s_n) < 0, \qquad \phi(\overline{z}_n) = e_2(s_n) > 0.$$

Esto concluye la prueba de (III.5). □

Nota III.2.4 Teniendo como punto de referencia la definición III.2.2 de E-estabilidad y el teorema III.2.3 dedicaremos nuestra atención en el resto del capítulo a la búsqueda de métodos E-estables de tipo Runge-Kutta.

III.3. Integraciones Runge-Kutta. Existencia, unicidad y acotación de la solución

Consideremos un método de tipo Runge-Kutta $RK(A,b)$ aplicado al sistema diferencial autónomo (II.1)-(II.3) dado por las ecuaciones

$$\begin{aligned} y_{n+1} &= y_n + h_n \sum_{j=1}^{s} b_j f(Y_j), \\ Y_i &= y_n + h_n \sum_{j=1}^{s} a_{ij} f(Y_j), \quad 1 \leq i \leq s, \end{aligned} \tag{III.9}$$

donde $h_n = t_{n+1} - t_n$ denota el tamaño de paso; Y_i, $1 \leq i \leq s$, son las etapas del método; y $A = (a_{ij}) \in \mathbb{R}^{s \times s}$, $b = (b_1, \ldots, b_s)^T \in \mathbb{R}^s$ son matrices de coeficientes que definen al método.

Asumamos que el método Runge-Kutta $RK(A,b)$ está expresado en forma *stiffly accurate*, esto es, que la solución de avance del método coincide con la última de las etapas que definen al método (obsérvese que esto no supone pérdida de generalidad, por cuanto al método original se le puede añadir una etapa *ficticia* adicional que coincida exactamente con la solución). De este modo la ecuación de las etapas en el punto t_n viene dada por

$$Y_i = y_n + h_n \sum_{j=1}^{s} a_{ij} f(Y_j), \quad h_n = t_{n+1} - t_n, \quad 1 \leq i \leq s,$$

mientras que la solución de avance del método coincide con la última etapa, esto es, $y_{n+1} = Y_s$.

Reemplazando y_n por y_0, y h_n por h por simplicidad, nos proponemos demostrar la existencia y unicidad de solución para la ecuación de las etapas internas. Más precisamente, estamos interesados en garantizar que existen dos constantes positivas γ (dependiente exclusivamente de las constantes que aparecen en las H-hipótesis) y p (únicamente dependiente de los coeficientes del método) de tal modo que

$$h \, \|y_0\| \leq \gamma, \quad h \geq 0,$$

implique la existencia y unicidad de solución para la ecuación de las etapas, así como una cota de estabilidad del tipo

$$\|Y_j\| \leq p\,\|y_0\|\,, \quad 1 \leq j \leq s.$$

Tras hacer el cambio de variables $Z_i = hY_i,\ 1 \leq i \leq s, \quad z_0 = hy_0, \quad z_1 = hy_1$, y aplicar el método al sistema diferencial (II.1), obtenemos que

$$\begin{array}{rcl} Z_i & = & z_0 + h\displaystyle\sum_{j=1}^{s} a_{ij} J Z_j + \sum_{j=1}^{s} a_{ij} N(Z_j), \quad 1 \leq i \leq s, \\ z_1 & = & Z_s, \end{array} \tag{III.10}$$

ya que $N(h^{-1}Z_j) = h^{-2}N(Z_j)$.

Con vistas a trabajar con estas ecuaciones de un modo más adecuado, introducimos los siguientes super-vectores de $\mathbb{R}^{ms}$,

$$Z := (Z_1^T, \ldots, Z_s^T)^T \in \mathbb{R}^{ms}, \qquad N(Z) := (N^T(Z_1), \ldots, N^T(Z_s))^T \in \mathbb{R}^{ms},$$

así como la norma en $\mathbb{R}^{ms}$

$$|Z| := \max_{1 \leq j \leq s} \|Z_j\|\,.$$

Entonces, (III.10) puede ser reescrito como

$$Z = \Phi(Z) := (I_s \otimes I_m - (A \otimes hJ))^{-1}\Big((e \otimes I_m)z_0 + (A \otimes I_m)N(Z)\Big), \tag{III.11}$$

con la condición de que

$$\det\,(I_s - \mu A) \neq 0, \forall \mu \in \mathbb{C}^- := \{\mu \in \mathbb{C} \,:\, \mathrm{Re}(\mu) < 0\}.$$

Aquí $e := (1, \ldots, 1)^T \in \mathbb{R}^s$ y $\otimes$ representa el producto de Kronecker de matrices $A \otimes B = (a_{ij}B)$. En lo que sigue, y por simplicidad en la notación, denotaremos I a la matriz identidad de la dimensión apropiada en cada caso.

Considerando entonces las funciones de estabilidad lineal $R_i(\mu)$ de las etapas internas Y_i del método Runge-Kutta y las funciones racionales $Q_{ij}(\mu)$ dadas por

$$\begin{array}{rcll} R_i(\mu) & = & e_i^T (I - \mu A)^{-1} e, & 1 \leq i \leq s, \\ Q_{ij}(\mu) & = & e_i^T (I - \mu A)^{-1} A e_j, & 1 \leq i, j \leq s, \end{array} \tag{III.12}$$

donde e_i denota el i-ésimo vector de la base canónica de $\mathbb{R}^s$, así como las funciones matriciales asociadas

$$\begin{aligned} R_i(hJ) &= (e_i \otimes I)^T (I \otimes I - A \otimes hJ)^{-1} (e \otimes I), \qquad 1 \le i \le s \\ Q_{ij}(hJ) &= (e_i^T \otimes I)(I \otimes I - A \otimes hJ)^{-1} (Ae_j \otimes I), \qquad 1 \le i, j \le s, \end{aligned}$$

se tiene a partir de (III.11) que

$$Z_i = (e_i^T \otimes I)Z = R_i(hJ)z_0 + \sum_{j=1}^{s} Q_{ij}(hJ)N(Z_j), \quad 1 \le i \le s. \tag{III.13}$$

Con vistas a demostrar la existencia y unicidad de solución, asumimos inicialmente que el método satisface la siguiente condición

(M1):

$$\sup_{h>0} \max_{1\le i\le s} \|R_i(hJ)\| \le p < \infty, \qquad \sup_{h>0} \max_{1\le i\le s} \sum_{j=1}^{s} \|Q_{ij}(hJ)\| \le p < \infty.$$

Con todo tenemos el siguiente

Teorema III.3.1 *(a) Para cada z_0 tal que $\|z_0\| < (4p^2\delta_2)^{-1}$, el problema* (III.11) *posee una solución única en la bola abierta*

$$B_{\gamma_0} = \{U \in \mathbb{R}^{ms} : \ |U| < \gamma_0 := (2p\delta_2)^{-1}\}.$$

(b) Si $\|z_0\| \le (4p^2\delta_2)^{-1}$, entonces el problema admite solución en la bola cerrada $\bar{B}_{\gamma_0}$.

Demostración. *(a)* Tomando $\gamma_1 = 2p\,\|z_0\|$, es claro que $\bar{B}_{\gamma_1} \subset B_{\gamma_0}$. Más aún, tenemos que $\Phi(\bar{B}_{\gamma_1}) \subset \bar{B}_{\gamma_1}$. En efecto, tomando $Z \in \bar{B}_{\gamma_1}$, ya que

$$\Phi_i(Z) = R_i(hJ)z_0 + \sum_{j=1}^{s} Q_{ij}(hJ)N(Z_j), \tag{III.14}$$

podemos escribir para cada $i = 1, \ldots, s$,

$$\begin{aligned} \|\Phi_i(Z)\| &\le p\,\|z_0\| + \delta_2 p|Z|^2 \\ &\le p(1 + 4\delta_2 p^2\,\|z_0\|)\,\|z_0\| \\ &\le p(1 + 4\delta_2 p^2 (4p^2\delta_2)^{-1})\,\|z_0\| = \gamma_1. \end{aligned}$$

Con todo, si $X, Y \in \bar{B}_{\gamma_1}$, se tiene entonces de (II.9) y (III.13) que

$$\begin{aligned} \|\Phi_i(X) - \Phi_i(Y)\| &\leq \delta_2 \sum_{j=1}^{s} \|Q_{ij}(hJ)\| \ \|X_j - Y_j\| \ \|X_j + Y_j\| \\ &\leq \delta_2 p |X - Y|(|X| + |Y|) \\ &\leq 2\delta_2 p \gamma_1 |X - Y|, \end{aligned}$$

para todo $1 \leq i \leq s$. Por lo tanto Φ posee en $\bar{B}_{\gamma_1}$ una constante de Lipschitz $L_\Phi := 2\delta_2 p \gamma_1$ menor que uno, y, en consecuencia, el Teorema del Punto Fijo permite establecer el enunciado como conclusión.

(b) El enunciado se deduce (a) usando argumentos de continuidad con respecto a $z_0 \in \bar{B}_{\gamma_0} \setminus B_{\gamma_0}$.

Este resultado puede ser también demostrado aplicando el Teorema del Punto Fijo de Brouwer [73, p.161] en la bola cerrada $\bar{B}_{\gamma_0}$, por cuanto se satisface la condición $\Phi(\bar{B}_{\gamma_0}) \subset \bar{B}_{\gamma_0}$. □

Nota III.3.2 Si $\|z_0\| < (4\delta_2 p^2)^{-1}$ entonces la única solución de la ecuación (III.11) puede ser obtenida por medio del siguiente proceso iterativo

$$Z^{(k+1)} = \Phi(Z^{(k)}), \quad k \geq 0,$$

tomando como valor de arranque $Z^{(0)} = 0$. De hecho, teniendo en cuenta (III.14), es sencillo comprobar que

$$|Z^{(k+1)}| \leq p(\|z_0\| + \delta_2 |Z^{(k)}|^2), \quad k \geq 0,$$

con lo cual se deduce por medio de un proceso inductivo que

$$|Z^{(k)}| \leq 2p \|z_0\| \equiv \gamma_1, \quad k \geq 0.$$

Además, debido a la contractividad de la función Φ en la bola $\bar{B}_{\gamma_1}$, tenemos que

$$\begin{aligned} |Z^{(n+1)} - Z^{(n)}| &\leq (L_\Phi)^n |Z^{(1)} - Z^{(0)}| \\ &\leq p \|z_0\| (L_\Phi)^n, \end{aligned}$$

donde $L_\Phi = 2\delta_2 q \gamma_1 < 1$ es la constante de Lipschitz de Φ en $\bar{B}_{\gamma_1}$. Vemos entonces que $\{Z^{(k)}\}$ es una sucesión de Cauchy en $\bar{B}_{\gamma_1}$, y por lo tanto converge a $Z \in \bar{B}_{\gamma_1}$. Finalmente tenemos que

$$Z = \lim_{k\to\infty} Z^{(k+1)} = \lim_{k\to\infty} \Phi(Z^{(k)}) = \Phi(\lim_{k\to\infty} Z^{(k)}) = \Phi(Z).$$

De aquí se concluye que $\{Z^{(k)}\}$ converge a la solución Z de (III.11), y que ésta satisface

$$|Z| \leq 2p \|z_0\|.$$

□

A partir del siguiente lema, deduciremos una condición *suficiente* sobre los coeficientes del método para obtener la propiedad *(M1)*.

Lema III.3.3 *Asumamos que J satisface (H1),(H2),(H3), y que $R(\mu) := P(\mu)(Q(\mu))^{-1}$ es una función racional sin polos en $\mathbb{C}^- \bigcup\{0\}$. Entonces*

$$\sup_{Re(\mu)<0} |R(\mu)| \leq p < \infty,$$

implica que

$$\sup_{h\geq 0} \|R(hJ)\| \leq p.$$

Demostración. Ya que los autovalores de $Q(hJ)$ vienen dados por $Q(h\lambda_j),\ j = 1,\ldots,m$, donde $\{\lambda_j,\ j = 1,\ldots,m\}$ son los autovalores de la matriz J, entonces tenemos por hipótesis que $\det(Q(hJ)) \neq 0,\ \forall h \geq 0$, y la matriz $R(hJ)$ está bien definida para todo $h \geq 0$.

Por otro lado, si consideramos $J_\epsilon = J - \epsilon I$, $\epsilon > 0$, se deduce que

$$Re(\langle u, hJ_\epsilon u\rangle) \leq -h\epsilon \|u\|^2, \quad \forall u \in \mathbb{C}^m,\ \forall h \geq 0.$$

Teniendo en cuenta el enunciado del Corolario 11.3 en [53, p.169], entonces obtenemos para cada $\epsilon > 0$ que

$$\begin{aligned} \|R(hJ_\epsilon)\| &\leq \sup_{Re(\mu)\leq -h\epsilon} |R(\mu)| \\ &\leq \sup_{Re(\mu)\leq 0} |R(\mu)| \leq p. \end{aligned}$$

La prueba finaliza teniendo presente que

$$\|R(hJ)\| = \lim_{\epsilon\to 0^+} \|R(hJ_\epsilon)\| \leq p.$$

□

Corolario III.3.4 *Supongamos que la matriz J satisface (H1),(H2),(H3), que* $\det(I - \mu A)$ *no se anula en $\mathbb{C}^-$ y que las funciones dadas por* (III.12) *verifican*

$$(P1) \equiv \begin{cases} \max\limits_{1\leq i\leq s}\left(\sup\limits_{Re(\mu)<0} |R_i(\mu)|\right) \ \leq \ p < \infty, \\ \\ \max\limits_{1\leq i\leq s} \sum\limits_{j=1}^{s} \left(\sup\limits_{Re(\mu)<0} |Q_{ij}(\mu)|\right) \ \leq \ p < \infty. \end{cases}$$

Entonces, se tiene que

$$\sup_{h>0} \max_{1\leq i\leq s} \|R_i(hJ)\| \leq p,$$

$$\sup_{h>0} \max_{1\leq i\leq s} \sum_{j=1}^{s} \|Q_{ij}(hJ)\| \leq p.$$

Demostración. La prueba es consecuencia inmediata del corolario previo. □

III.4. A-aceptabilidad fuerte de las funciones racionales

En esta sección estableceremos una caracterización de las funciones racionales *fuertemente A-aceptables* que será de gran utilidad de cara a dar condiciones suficientes para la E-estabilidad de los métodos de tipo Runge-Kutta. Debemos señalar que una función racional compleja $R(z)$, $z \in \mathbb{C}$, es A-aceptable si

$$|R(z)| \leq 1, \quad \forall z \in \mathbb{C}^{-}.$$

Se dirá que la A-aceptabilidad es fuerte si en añadidura se verifica que $|R(\infty)| < 1$.

Notemos que todo método de tipo Runge-Kutta (III.9) aplicado al problema lineal

$$y' = Ay, \quad y \in \mathbb{C}^m, \tag{III.15}$$

da como solución numérica

$$y_1 = R(hA)y_0,$$

en términos del valor inicial y_0, donde $R(z)$ denota la función de estabilidad lineal del método, siendo ésta una función racional de la variable z. En este sentido, un método de tipo Runge-Kutta se dice A-estable (resp. fuertemente A-estable) si su función de estabilidad lineal es A-aceptable (resp. fuertemente A-aceptable).

Con esto, haciendo uso de un bien conocido teorema de Von Neumann [53, p. 168-169], [87], se deduce la siguiente cota para la solución de avance en términos del valor inicial

$$\|y_1\| \leq \phi_R(h\mu[A])\, \|y_0\|\,, \tag{III.16}$$

siendo ϕ_R la *función crecimiento de error* asociada a la función de estabilidad lineal $R(z)$ del método Runge-Kutta

$$\phi_R(x) := \sup_{Re\ z\leq x} |R(z)|, \quad x \in \mathbb{R}, \tag{III.17}$$

y donde $\mu[A]$ representa la *norma logarítmica* [33, Cap. I] de la matriz A asociada a la norma inducida por un determinado producto interior $\langle \cdot, \cdot \rangle$, esto es,

$$\mu[A] := \sup_{u \neq 0} \frac{\mathrm{Re}\langle u, Au \rangle}{\langle u, u \rangle}, \quad \|u\|^2 = \langle u, u \rangle, \quad u \in \mathbb{C}^m.$$

Notemos además que la solución exacta del problema lineal (III.15) verifica

$$\|y(t+h)\| \leq \exp(h\mu[A]) \, \|y(t)\| \, . \tag{III.18}$$

De este modo, comparando (III.16) y (III.18), es importante de cara a la estabilidad o contractividad de los métodos numéricos que la función $\phi_R(x)$ aproxime a la función exponencial $\exp(x)$ para valores reales de x siendo $|x|$ suficientemente pequeño. En este sentido, algunas propiedades de la *función crecimiento de error* de un método Runge-Kutta (o de un método semi-implícito) fueron estudiadas en [1] y [53, Cap. IV.11]. En particular, haciendo uso de la teoría de *estrellas del orden* [51] se demuestra en [53, Cap. IV.11] el siguiente

Teorema III.4.1 *Sea $R(z)$ una aproximación A-aceptable a la función $\exp z$ de orden exacto p, esto es, $R(z) = \exp z + Cz^{p+1} + \mathcal{O}(z^{p+2})$, $z \to 0$, con $C \neq 0$. Si además $|R(iy)| < 1$, para $y \neq 0$, y $|R(\infty)| < 1$, entonces se tiene que*

a) si p es impar

$$\phi_R(x) = \exp x + \mathcal{O}(x^{p+1}), \quad x \to 0;$$

b) si p es par y $(-1)^{p/2} C x < 0$

$$\phi_R(x) = \exp x + \mathcal{O}(x^{p+1}), \quad x \to 0;$$

b) si p es par y $(-1)^{p/2} C x > 0$

$$\phi_R(x) = \exp x + \mathcal{O}(x^{r+1}), \quad x \to 0,$$

para algún número racional positivo $r \leq p/2$.

Como consecuencia inmediata de este resultado se demuestra [1, Corolario 8]

Proposición III.4.2 *Sea $R(z)$ una aproximación racional A-aceptable de orden $p \geq 1$ a la función exponencial $\exp(z)$, de tal modo que $|R(\infty)| < 1$ y $|R(iy)| < 1$, $\forall y \in \mathbb{R} \setminus \{0\}$. Entonces se tiene que*

$$\phi_R(x) = 1 + x + \mathbf{o}(x), \quad x \to 0. \tag{III.19}$$

Debemos señalar que la propiedad (III.19) anteriormente introducida juega un papel muy importante en el estudio de la contractividad y estabilidad de las soluciones numéricas a través de métodos semi-implícitos cuando éstos se aplican no sólo a problemas lineales sino también a otras clases más generales de problemas no lineales, véase [1, p.222].

Por otro lado, en la sección siguiente demostraremos que aquellos métodos de tipo Runge-Kutta cuya función crecimiento de error (III.17) satisface

$$\phi_R(x) \leq \max\{\rho, 1+px\},\ \rho < 1,\ p > 0,\ \forall x \leq 0, \tag{III.20}$$

proveen integraciones estables en intervalos temporales de gran amplitud cuando se aplican a los sistemas diferenciales autónomos que poseen equilibrios semi-estables, en el sentido introducido por las H-hipótesis.

Notemos que el enunciado (III.20) se deduce fácilmente a partir de la Proposición III.4.2. Sin embargo, esta propiedad (III.20) puede ser obtenida sin requerir

$$|R(iy)| < 1,\ \forall y \in \mathbb{R} \setminus \{0\},$$

tal como probaremos en el desarrollo de esta sección. Para ello establecemos a continuación un lema previo al principal resultado de la sección.

Lema III.4.3 *Sean $R(z) = P(z)/Q(z)$ una función racional A-aceptable no constante y $z_0 \in i\mathbb{R}$ tal que $|R(z_0)| = 1$. Entonces, existen constantes positivas r y p tales que*

$$|R(z_0+w)| \leq 1 + pRe\ w,\ \ \forall w:\ |w| \leq r,\ Re\ w \leq 0. \tag{III.21}$$

Demostración. En primer lugar, se debe notar que

$$E(w) + pRe\ w \geq 0,\ \ \forall w:\ |w| \leq r,\ Re\ w \leq 0, \tag{III.22}$$

implica (III.21), siendo $E(w)$ definida por

$$E(w) := |Q(z_0+w)|^2 - |P(z_0+w)|^2,$$

aunque los valores de la constante positiva p que aparece en (III.21) y (III.22) no son necesariamente iguales.

Asumamos, sin pérdida de generalidad, que $P(z_0) = Q(z_0) = 1$. Entonces para algún entero positivo s se tiene que

$$Q(z_0 + w) = \sum_{j=0}^{s} \alpha_j w^j, \quad P(z_0 + w) = \sum_{j=0}^{s} \beta_j w^j, \quad \alpha_0 = \beta_0 = 1.$$

Tomando $w = i\varepsilon e^{i\theta}$, y denotando $E(\varepsilon, \theta) \equiv E(w)$, sigue de la A-aceptabilidad de la función racional $R(z)$ que

$$E(\varepsilon, \theta) = \sum_{j,k=0}^{s} (\alpha_j \overline{\alpha_k} - \beta_j \overline{\beta_k})(-1)^k i^{j+k} e^{i(j-k)\theta} \varepsilon^{j+k} \geq 0, \quad \forall \varepsilon > 0, \ \forall \theta \in [0, \pi]. \tag{III.23}$$

Considerando (III.23) tenemos entonces que

$$0 \leq \lim_{\varepsilon \to 0^+} \frac{E(\varepsilon, \theta)}{\varepsilon} = i\Big\{(\alpha_1 - \beta_1)e^{i\theta} - (\overline{\alpha_1} - \overline{\beta_1})e^{-i\theta}\Big\}, \ \ 0 \leq \theta \leq \pi. \tag{III.24}$$

Las elecciones particulares $\theta = 0, \pi$, implican que $\alpha_1 - \beta_1 \in \mathbb{R}$. Así, de (III.24) se obtiene que

$$0 \leq 2(\beta_1 - \alpha_1) \sin\theta, \quad \forall \theta \in [0, \pi].$$

Esto implica que $\beta_1 \geq \alpha_1$. Además, se tiene que $\beta_1 > \alpha_1$. En efecto, si asumimos que $\alpha_j = \beta_j$, $\forall j < l$ (para algún $l \geq 2$), entonces partiendo de (III.23) se deduciría que

$$0 \leq \lim_{\varepsilon \to 0^+} \frac{E(\varepsilon, \theta)}{\varepsilon^l} = i^l\Big\{(\alpha_l - \beta_l)e^{il\theta} + (-1)^l(\overline{\alpha_l} - \overline{\beta_l})e^{-il\theta}\Big\}, \ \ 0 \leq \theta \leq \pi.$$

Esto implicaría que $\alpha_l = \beta_l$, y al considerar el mismo argumento para $l+1, l+2, \ldots$ nos llevaría a que $P(z) = Q(z)$, lo cual contradice la hipótesis inicial de que $R(z)$ es una función racional no constante.

Por otro lado, denotando $p := \beta_1 - \alpha_1 > 0$, sigue que

$$\begin{aligned} E(w) + pRe\ w &= E(\varepsilon, \theta) - p\varepsilon \sin\theta \\ &= p\varepsilon \sin\theta + \sum_{\substack{j,k=0 \\ j+k \geq 2}}^{s} (\alpha_j \overline{\alpha_k} - \beta_j \overline{\beta_k})(-1)^k i^{j+k} e^{i(j-k)\theta} \varepsilon^{j+k}. \end{aligned}$$

Además, teniendo en mente la no negatividad de $E(\varepsilon, \theta)$, podemos deducir que

$$E(\varepsilon, \theta) \geq E(\varepsilon, \theta) - \frac{1}{\pi}\Big\{(\pi - \theta)E(\varepsilon, 0) + \theta E(\varepsilon, \pi)\Big\}, \ \ 0 \leq \theta \leq \pi.$$

Con esto se tiene que

$$\begin{aligned} E(\varepsilon,\theta)-p\varepsilon\sin\theta &\geq p\varepsilon\sin\theta \\ &\quad+\sum_{\substack{j,k=0\\ j+k\geq 2}}^{s}(\alpha_j\overline{\alpha_k}-\beta_j\overline{\beta_k})(-1)^k i^{j+k}\left\{e^{i(j-k)\theta}-\frac{\pi-\theta}{\pi}-\frac{\theta}{\pi}e^{i(j-k)\pi}\right\}\varepsilon^{j+k} \\ &= \varepsilon\sin\theta\left(p+\sum_{\substack{j,k=0\\ j+k\geq 2}}^{s}(\alpha_j\overline{\alpha_k}-\beta_j\overline{\beta_k})(-1)^k i^{j+k}\varphi_{j-k}(\theta)\varepsilon^{j+k-1}\right), \end{aligned}$$

donde las funciones $\varphi_l(\theta),\ l\in\mathbb{Z}$, están dadas por

$$\varphi_l(\theta):=\left(e^{il\theta}-\frac{\pi-\theta}{\pi}-\frac{\theta}{\pi}e^{il\pi}\right)/\sin\theta,\quad \theta\in(0,\pi),$$

y para los casos particulares $\theta=0,\ \pi$, consideramos los correspondientes límites.

Ya que las funciones $|\varphi_l(\theta)|,\ l=-s,-s+1,\ldots,s$, están acotadas para $0\leq\theta\leq\pi$, concluimos la prueba eligiendo una constante $r>0$ tal que

$$p+\sum_{\substack{j,k=0\\ j+k\geq 2}}^{s}(\alpha_j\overline{\alpha_k}-\beta_j\overline{\beta_k})(-1)^k i^{j+k}\varphi_{j-k}(\theta)\varepsilon^{j+k-1}\geq 0,\ \forall\varepsilon\in[0,r],\ \forall\theta\in[0,\pi].$$

□

Teorema III.4.4 *Sea $R(z)$ una función racional. Entonces, la propiedad de A-aceptabilidad junto a $|R(\infty)|<1$ es equivalente a (III.20).*

Demostración. El hecho de que (III.20) implica para la función racional R la A-aceptabilidad y $|R(\infty)|<1$ es trivial, sin más que considerar que $\phi_R(0)\leq 1$ y $\phi_R(-\infty)\leq\rho<1$.

Con vistas a probar el recíproco, notamos que, ya que $|R(\infty)|<1$, entonces existen dos constantes $C>0$ y $\rho_1<1$ tales que

$$|R(z)|\leq\rho_1,\ \forall z\in\mathbb{C},\ |Im\ z|\geq C. \tag{III.25}$$

Si no existen puntos z en el segmento del eje imaginario complejo $z\in[-iC,iC]$ cumpliendo $|R(z)|=1$, entonces la prueba se deduce directamente por cuanto se tendría que $\sup_{\mathrm{Re}\ z\leq 0}|R(z)|<1$.

En otro caso, existe a lo sumo un número finito de puntos $z_1,\ldots,z_l\in[-iC,iC]$ tales que

$$|R(z_j)|=1,\ 1\leq j\leq l.$$

Por lo tanto en virtud del Lema III.4.3, existen constantes positivas p_j, r_j tales que

$$|R(z)| \leq 1 + p_j Re\ z,\ \forall z \in C(z_j, r_j), \quad j = 1, \ldots, l,$$

donde $C(\mu, r)$ denota el conjunto dado por

$$C(\mu, r) := \{z \in \mathbb{C} :\ -r < Re\ z \leq 0,\ |Im\ z - Im\ \mu| < r/2\}.$$

Notemos que $C(\mu, r)$ es un conjunto abierto en la topología relativa de

$$\mathbb{C}_0^- := \{z \in \mathbb{C} :\ Re\ z \leq 0\}.$$

Así, tomando

$$p := \min_{1 \leq j \leq l} p_j,$$

se tiene para cada $\mu \in [-iC, iC]$ que existe una constante $r_\mu > 0$ tal que

$$|R(z)| \leq 1 + pRe\ z,\ \ \forall\ z \in C(\mu, r_\mu).$$

En virtud de que $[-iC, iC]$ es un conjunto compacto de $\mathbb{C}_0^-$ y que

$$\{C(\mu, r_\mu),\ \mu \in [-iC, iC]\}$$

constituye un recubrimiento por abiertos del mismo, podemos extraer un número finito de conjuntos, pongamos $\{C(\mu_i, r_{\mu_i})\}_{i=1}^n$. Entonces, tomando

$$r := \min\{r_{\mu_i}\} > 0, \quad i = 1, \ldots, n,$$

y teniendo en cuenta (III.25), se tiene que

$$|R(z)| \leq \max\{\rho_1, 1 + pRe\ z\}, \quad \text{cuando } -r < Re\ z \leq 0. \tag{III.26}$$

Si $Re\ z \leq -r$, el principio del máximo permite afirmar que

$$\phi_R(x) \leq \rho_2 := \max_{Re\ z \leq -r} |R(z)| < 1.$$

Esto junto con (III.26) da que

$$\phi_R(x) \leq \max\{\rho, 1 + px\}, \quad \forall x \leq 0,$$

siendo $\rho := \max\{\rho_1, \rho_2\}$. □

Para métodos de tipo Runge-Kutta A-estables cumpliendo que $|R(\infty)| = 1$ no es posible deducir la acotación (III.20), ya que si $x_0 < 0$ entonces

$$\lim_{y\to\infty} |R(x_0 + iy)| = 1,$$

mientras que

$$\max\{\rho, 1 + px_0\} < 1.$$

Este es el caso de interesantes métodos con buenas propiedades de estabilidad como la regla trapezoidal. En general, para cualquier número de etapas s, los métodos Runge-Kutta Gauss poseen como función de estabilidad lineal al aproximante de Padé $\Pi_{s,s}$ a la exponencial $\exp z$ [53, Cap. IV.5], y de aquí que estos métodos constituyan ejemplos clásicos de métodos A-estables pero no fuertemente A-estables. En consecuencia, no es posible deducir una acotación del tipo (III.20) para la función de estabilidad lineal de esta clase de métodos.

Finalizamos esta sección con la idea de dar ejemplos de métodos de tipo Runge-Kutta fuertemente A-estables verificando que su función de estabilidad lineal toma valores con módulo igual a 1 en diversos puntos aislados del eje imaginario. Para ello establecemos la siguiente

Proposición III.4.5 *Sea $R(z) = P(z)/Q(z)$ una función racional compleja A-aceptable de orden $p \geq \max\{0, 2s-4\}$, siendo $s = \max\{deg(P), deg(Q)\}$. Si $|R(z)|$ es no constante para $z \in i\mathbb{R}$, entonces se tiene que*

$$|R(iy)| = 1, \;\; y \in \mathbb{R} \Rightarrow y = 0.$$

Demostración. Consideremos el polinomio

$$E(y) := |Q(iy)|^2 - |P(iy)|^2, \quad y \in \mathbb{R}. \tag{III.27}$$

En virtud de [53, Proposición 3.4, p. 43] sabemos que $E(y)$ es un polinomio de grado $deg(E) \leq 2s$ que sólo contiene potencias pares en la variable y, y que

$$E(y) = \mathcal{O}(y^{p+1}).$$

Por lo tanto debe ser que E toma la forma $E(y) = \alpha y^{2s-2} + \beta y^{2s}$, y debido a la A-aceptabilidad de R se tiene que $\alpha, \beta \geq 0$.

En consecuencia tenemos que $E(y) = 0, \; \forall y \in \mathbb{R}$, o bien $y = 0$ es la única raíz real del polinomio E. Esto concluye la prueba. □

Nota III.4.6 Como consecuencia del resultado previo tenemos que aquellas funciones racionales A-aceptables $R(z)$ con numerador y denominador de grado menor o igual que s que aproximan a la función exponencial e^z hasta el orden $p \geq \max\{0, 2s-4\}$ cumplen que

$$|R(iy)| = 1, \ \forall\, y \in \mathbb{R},$$

o bien

$$|R(iy)| < 1, \ \forall\, y \in \mathbb{R} \setminus \{0\}.$$

En consecuencia, el orden de un método Runge-Kutta A-estable de s etapas cuya función de estabilidad lineal satisface que $|R(\infty)| < 1$ y $|R(iy)| = 1$ en algunos puntos aislados del eje imaginario complejo tiene que ser $p \leq 2s-5$.

Como ejemplos ilustrativos de esta situación consideramos los métodos *Singly Diagonally Implicit Runge-Kutta*, $SDIRK$, en forma *Stiffly Accurate*, de tal modo que $R(\infty) = 0$. Estos métodos mantienen un carácter implícito, pero la matriz de coeficientes A que los define es triangular inferior y posee elementos diagonales iguales, $a_{11} = \ldots = a_{ss} = \gamma > 0$. De este modo, si se imponen condiciones de orden $p \geq s-1$ entonces la correspondiente función de estabilidad lineal queda unívocamente definida por la fórmula (véase [53, Cap. IV.6])

$$R(z) = \frac{P(z)}{(1-\gamma z)^s},$$

siendo

$$P(z) = (-1)^s \sum_{j=0}^{s-1} L_s^{(s-j)}\left(\frac{1}{\gamma}\right)(\gamma z)^j,$$

donde L_s denota el polinomio de Laguerre de grado s. De este modo tenemos que

$$R(z) - \exp z = Cz^s + \mathcal{O}(z^{s+1}),$$

donde la constante de error C está dada por la fórmula

$$C = (-1)^s L_s\left(\frac{1}{\gamma}\right)\gamma^s.$$

A modo de ejemplo consideramos el caso de los métodos $SDIRK$ de 4 etapas y orden $p \geq 3$. Ya que para estos métodos la función de estabilidad lineal no tiene polos en el semiplano complejo negativo, la condición de A-estabilidad es equivalente a la no negatividad del polinomio

$E(y)$ dado por (III.27), siendo $R(z) = P(z)/Q(z)$ la función de estabilidad lineal del método. En este caso tenemos entonces que (véase [53, p.98])

$$E(y) := |Q(iy)|^2 - |P(iy)|^2 = y^4 p_1(\gamma) + y^6 p_2(\gamma) + y^8 p_3(\gamma) \geq 0, \ \forall y \in \mathbb{R},$$

donde

$$\begin{aligned} p_1(\gamma) &:= \frac{1}{12} - \frac{4}{3}\gamma + 6\gamma^2 - 8\gamma^3 + 2\gamma^4 \\ p_2(\gamma) &:= -\frac{1}{36} + \frac{2}{3}\gamma - 6\gamma^2 + \frac{76}{3}\gamma^3 - 52\gamma^4 + 48\gamma^5 - 12\gamma^6 \\ p_3(\gamma) &:= \gamma^8. \end{aligned}$$

De este modo, estos métodos resultan ser A-(y L-)estables con orden $p \geq 3$ para valores del parámetro $\gamma_1 \leq \gamma \leq \gamma_2$, siendo $\gamma_1 \simeq 0.22364780$ y $\gamma_2 \simeq 0.57281606$.

De hecho, $\gamma = \gamma_1$ es la única solución real de la ecuación

$$\begin{cases} 4p_1(\gamma)p_3(\gamma) = p_2(\gamma)^2 \\ p_2(\gamma) < 0, \end{cases}$$

y entonces para $\gamma := \gamma_1$ el polinomio E toma la expresión

$$E(y) = y^4 \left(\frac{1}{4} \frac{p_2(\gamma)^2}{p_3(\gamma)} + p_2(\gamma) y^2 + p_3(\gamma) y^4 \right),$$

con $p_2(\gamma) < 0$ y $p_3(\gamma) > 0$.

Por lo tanto obtenemos que $E(y) \geq 0$, $\forall y \in \mathbb{R}$, y $E(y) = 0$ para $y = 0, \pm\sqrt{\dfrac{-p_2(\gamma)}{2p_3(\gamma)}}$. En consecuencia, para todos aquellos métodos con $\gamma = \gamma_1$ se tiene que

$$\left| R\left(\pm i\sqrt{\frac{-p_2(\gamma)}{2p_3(\gamma)}} \right) \right| = 1.$$

III.5. E-estabilidad de los métodos Runge-Kutta

En esta sección demostraremos que la propiedad *(P1)* introducida en el Corolario III.3.4 junto con la propiedad *(P2)* que se especifica abajo resultan ser condiciones suficientes para la E-estabilidad de los métodos de tipo Runge-Kutta. Más precisamente, y teniendo en cuenta los resultados de la sección previa, estamos interesados en la siguiente propiedad

(P2): Existen dos constantes positivas ϱ y $\rho < 1$ tales que la función crecimiento de error (III.17) satisface

$$\phi_R(x) := \sup_{Re(\mu)\leq x} |R(\mu)| \leq \max\{\rho, 1+\varrho x\}, \forall x \leq 0,$$

donde $R(\mu)$ denota la función de estabilidad lineal del método Runge-Kutta en cuestión.

Antes de enunciar el principal resultado de este capítulo, en el que se darán condiciones suficientes para garantizar la E-estabilidad de los métodos de tipo Runge-Kutta, necesitamos previamente enunciar y demostrar algunos resultados adicionales que resultarán fundamentales para demostrar la veracidad de este teorema principal (Teorema III.5.3).

Teorema III.5.1 *Sea K una matriz compleja cuadrada de dimensión m que posee al autovalor nulo como autovalor simple y sea v un autovector asociado al autovalor nulo. Sea además $R(\mu)$ una función racional compleja irreducible sin polos en la región del plano complejo $\{\mu \in \mathbb{C} \,/\, Re(\mu) \leq -\delta\}$ tal que $R(0) \neq \infty$. Entonces para todo producto interior complejo $\langle\cdot,\cdot\rangle : \mathbb{C}^m \times \mathbb{C}^m \longmapsto \mathbb{C}$ cumpliendo*

$$\begin{array}{l} \text{(a) } Kv^{\perp} \subset v^{\perp}, \\ \text{(b) } \operatorname{Re}\langle Ky, y\rangle \leq -\delta \|y\|^2, \quad y \in v^{\perp}, \end{array} \tag{III.28}$$

se verifica que

$$|\langle R(K)y_1, y_2\rangle| \leq \phi_R(-\delta)\|y_1\| \cdot \|y_2\|, \quad y_1, y_2 \in v^{\perp},$$

donde $\phi_R(x)$ representa la función crecimiento de error (III.17) asociada a la función racional $R(\mu)$.

Demostración. Es claro que $R(K)$ es una matriz bien definida por cuanto $R(0) \neq \infty$ y los autovalores no nulos de K, pongamos $\{\lambda_2, \ldots, \lambda_m\}$, satisfacen

$$\operatorname{Re}(\lambda_j) \leq -\delta, \quad j = 2, \ldots, m. \tag{III.29}$$

En efecto, consideremos un autovector u_j correspondiente al autovalor λ_j. Entonces, $u_j = a_j v + w_j,\ a_j \in \mathbb{C},\ w_j \neq 0,\ w_j \in v^{\perp}$. Sigue entonces que,

$$\begin{aligned} \lambda_j \|w_j\|^2 &= \langle \lambda_j w_j, w_j\rangle \\ &= \langle \lambda_j u_j, w_j\rangle \\ &= \langle K u_j, w_j\rangle \\ &= \langle K w_j, w_j\rangle. \end{aligned}$$

En consecuencia

$$\|w_j\|^2 \operatorname{Re}(\lambda_j) \leq -\delta \|w_j\|^2,$$

lo cual implica (III.29). Por otro lado, para el producto interior en cuestión existe una matriz no singular P tal que (ver por ejemplo [33, p.22])

$$\langle u_1, u_2\rangle = \langle Pu_1, Pu_2\rangle_2 = u_2^* P^* P u_1, \ \ \forall\, u_1, u_2 \in \mathbb{C}^m,$$

donde $\langle\cdot,\cdot\rangle_2$ denota el producto interior euclídeo estándar. A partir de (III.28)(b) sigue entonces que

$$Re\langle Kw, w\rangle \leq -\delta \left\|Pw\right\|_2^2, \ \forall w \in v^{\perp}. \tag{III.30}$$

Por otro lado, no es difícil ver que

$$y \in (Pv)_2^{\perp} \Longleftrightarrow P^{-1}y \in v^{\perp}, \tag{III.31}$$

donde $(Pv)_2^{\perp}$ es el subespacio ortogonal a Pv con respecto al producto interior euclídeo estándar. Entonces denotando

$$K_1 = PKP^{-1}, \quad y = Pw, \quad u = Pv,$$

sigue de (III.30) y (III.31) que,

$$Re\langle K_1 y, y\rangle_2 \leq -\delta \left\|y\right\|_2^2, \quad \forall y \in (u)_2^{\perp}. \tag{III.32}$$

Además, teniendo en cuenta (III.28)(a) se deduce que

$$K_1(u)_2^{\perp} \subset (u)_2^{\perp}. \tag{III.33}$$

Considerando la descomposición de Schur de la matriz K_1, podemos escribir

$$K_1 = QDQ^*,$$

donde $Q = [v_1, v_2, \ldots, v_m]$ es una matriz unitaria y $D = (d_{ij})$ es una matriz triangular superior de la forma

$$D = \begin{pmatrix} 0 & d^T \\ 0 & \Lambda \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \lambda_2 & \cdots & \cdots \\ \mathbf{0} & \ddots & \cdots \\ 0 & 0^T & \lambda_m \end{pmatrix}, \quad d^T = (d_{12}, \ldots, d_{1m}).$$

Puesto que $DQ^*u = Q^*K_1u = 0$ tenemos que

$$\Lambda[v_2^*u, \ldots, v_m^*u]^T = 0,$$

y ya que Λ es una matriz no singular deducimos que u es ortogonal ,con respecto al producto interior euclídeo usual, a $v_2, \ldots, v_m$. En consecuencia, podemos asumir sin pérdida de generalidad que $v_1 = u$ y $(u)_2^{\perp} = span\{v_2, \ldots, v_m\}$. Además, ya que $K_1 Q = QD$, sigue que

$$K_1 v_j = d_{1j} u + \sum_{l=2}^{j} d_{lj} v_l, \quad j = 2, \ldots, m.$$

De este modo, las condiciones de ortogonalidad y (III.33) implican $d^T = 0$.
Por otro lado, para cada $z \in \mathbb{C}^{m-1}$, el vector asociado $w_z := Q \cdot \begin{pmatrix} 0 \\ z \end{pmatrix}$ pertenece a $(u)_2^{\perp}$ y $\|w_z\|_2 = \|z\|_2$. Por tanto sigue de (III.32) que

$$\begin{aligned} Re\langle \Lambda z, z\rangle_2 &= Re\langle K_1 w_z, w_z\rangle_2 \\ &\leq -\delta \|w_z\|_2^2 \\ &= -\delta \|z\|_2^2 . \end{aligned} \tag{III.34}$$

Finalmente, tenemos para cada $y_1, y_2 \in v^{\perp}$

$$\begin{aligned} |\langle R(K) y_1, y_2\rangle| &= |\langle R(K_1) P y_1, P y_2\rangle_2| \\ &= |\langle R(D) Q^* P y_1, Q^* P y_2\rangle_2| \\ &\leq \|R(D)\|_2 \cdot \|y_1\| \cdot \|y_2\| . \end{aligned}$$

La prueba concluye teniendo en cuenta que $\|R(D)\|_2 = \|R(\Lambda)\|_2$ y considerando el Corolario 11.4 en [53, p.169] para obtener en virtud de (III.34) que

$$\|R(\Lambda)\|_2 \leq \sup_{\text{Re } z \leq -\delta} |R(z)| = \phi_R(-\delta).$$

□

El siguiente corolario es determinante para demostrar el principal resultado de este capítulo (Teorema III.5.3).

Corolario III.5.2 *Sea J una matriz verificando las hipótesis (H1), (H2) y (H3), y sea $R(\mu)$ una función racional compleja sin polos en el semiplano complejo negativo satisfaciendo (P2). Entonces*

$$|\langle R(hJ) w_1, w_2\rangle| \leq \max\{\rho, 1 - h\varrho\delta_1\} \|w_1\| \cdot \|w_2\| , \quad w_1, w_2 \in v^{\perp},\ h \geq 0.$$

Demostración. En virtud de *(H1), (H2), (H3)*, se tiene que $Jv^{\perp} \subset v^{\perp}$, y

$$Re(\langle hJw, w\rangle) \leq -h\delta_1 \|w\|^2, \quad w \in v^{\perp},\ h \geq 0.$$

A partir del Teorema III.5.1 sigue que,

$$|\langle R(hJ)w_1, w_2\rangle| \leq \phi_R(-h\delta_1)\, \|w_1\| \cdot \|w_2\|,\ w_1, w_2 \in v^{\perp},\ h \geq 0.$$

Esta acotación junto con la propiedad *(P2)* completan la prueba. □

Teorema III.5.3 *Sea un método Runge-Kutta $RK(A, b)$ satisfaciendo las propiedades (P1) y (P2). Entonces el método es E-estable.*

Demostración. La constante p que aparece a largo de esta demostración se refiere a aquella constante obtenida por aplicación del Corolario III.3.4.

Denotemos $z_0 = hy_0$ y $Z_i = hY_i$, $1 \leq i \leq s$, siendo $h > 0$, y_0 valor inicial e Y_i las etapas internas del método Runge-Kutta (III.9). Asumimos sin pérdida de generalidad que $Z_s = hy_1 = z_1$, donde y_1 representa la solución de avance del método. Atendiendo a la descomposición $\mathbb{R}^m = span\{v\} \oplus v^{\perp}$ podemos escribir

$$\begin{aligned} z_0 &= \tau_0 v + w_0, \quad \tau_0 > 0,\ \ w_0 \in v^{\perp} \\ Z_i &= \tau_i v + w_i, \quad \tau_i > 0,\ \ w_j \in v^{\perp},\ \ 1 \leq i \leq s. \end{aligned}$$

Entonces, mostrar la E-estabilidad del método (definición III.2.2) es equivalente a mostrar que existen cuatro constantes positivas $\gamma, \kappa, \alpha, \beta$, únicamente dependientes de las constantes involucradas en las *H-hipótesis* y de los coeficientes del método, de tal modo que para cada z_0 y cada h verificando

$$0 < \|z_0\| \leq \gamma, \quad \|z_0\| \leq h\beta \quad \text{y} \quad \|w_0\| \leq \alpha\tau_0$$

tengamos que

$$\|z_1\| \leq \|z_0\| (1 - \kappa \|z_0\|) \tag{III.35}$$

y

$$\|w_s\| \leq \alpha\tau_s. \tag{III.36}$$

De (III.10) sigue que,

$$\tau_i v + w_i = \tau_0 v + w_0 + h \sum_{j=1}^{s} a_{ij} J w_j + \sum_{j=1}^{s} a_{ij} N(\tau_j v + w_j), \quad 1 \leq i \leq s.$$

Entonces, considerando las proyecciones sobre los subespacios $span(v)$ y $v^{\perp}$, obtenemos que

$$\tau_i = \tau_0 + \sum_{j=1}^{s} a_{ij}\langle v, N(\tau_j v + w_j)\rangle, \quad 1 \leq i \leq s, \tag{III.37}$$

y

$$w_i = w_0 + h\sum_{j=1}^{s} a_{ij} J w_j + \sum_{j=1}^{s} a_{ij} N^{\perp}(\tau_j v + w_j), \quad 1 \leq i \leq s,$$

con $N^{\perp}(u) := N(u) - \langle v, N(u)\rangle v$.

Más específicamente, de (III.13) tenemos que

$$w_i = R_i(hJ)w_0 + \sum_{j=1}^{s} Q_{ij}(hJ) N^{\perp}(\tau_j v + w_j), \quad 1 \leq i \leq s.$$

Así, puesto que el método satisface la propiedad $(P1)$, tenemos en virtud del corolario III.3.4 y el teorema III.3.1 que $\|Z_i\| \leq 2p\,\|z_0\|\,,\ 1 \leq i \leq s$. Por lo tanto, podemos deducir la existencia de dos constantes positivas C_1, C_2 únicamente dependientes de δ_2 (definida en *(H4)*) y de los coeficientes del método, de tal manera que

$$|\tau_i - \tau_0| \leq C_1 \|z_0\|^2\,, \quad \forall h \geq 0, \quad 1 \leq i \leq s, \tag{III.38}$$

$$\|w_i - R_i(hJ)w_0\| \leq C_2 \|z_0\|^2\,, \quad \forall h \geq 0, \quad 1 \leq i \leq s. \tag{III.39}$$

Ahora, mostraremos que existen dos constantes positivas α', γ' tales que

$$\tau_0 - \frac{3}{2}\nu\tau_0^2 \leq \tau_s \leq \tau_0 - \frac{1}{2}\nu\tau_0^2, \tag{III.40}$$

siempre que $\|w_0\| \leq \alpha'\tau_0,\ \|z_0\| \leq \gamma'$ y $h > 0$.

A partir de (III.37) podemos escribir, para cualquier índice $i = 1, \ldots, s$, que

$$\tau_i = \tau_0 + \sum_{j=1}^{s} a_{ij}\langle v, N(\tau_0 v)\rangle + \sum_{j=1}^{s} a_{ij}\langle v, N(\tau_j v + w_j) - N(\tau_0 v)\rangle. \tag{III.41}$$

Además, de (III.38), (III.39) y (II.9) sigue que

$$\begin{aligned}
\|N(\tau_j v + w_j) - N(\tau_0 v)\| &\leq \delta_2 \|(\tau_j - \tau_0)v + w_j\| \cdot \|(\tau_j + \tau_0)v + w_j\| \\
&\leq \delta_2(|\tau_j - \tau_0| + \|w_j\|)(\|Z_j\| + \|z_0\|) \\
&\leq \delta_2(|\tau_j - \tau_0| + \|R_j(hJ)w_0\| + C_2\|z_0\|^2)(\|Z_j\| + \|z_0\|) \\
&\leq (1+2p)\delta_2((C_1 + C_2)\|z_0\|^2 + p\|w_0\|)\|z_0\|\,.
\end{aligned}$$

De este modo concluimos que existen dos constantes $\alpha' > 0, \gamma' > 0$ tales que si $\|w_0\| \leq \alpha'\tau_0$, $\|z_0\| \leq \gamma'$, entonces $\|z_0\| \leq \tau_0(1+\alpha'^2)^{1/2}$ y

$$\begin{aligned}\Big|\sum_{j=1}^{s} a_{ij}\langle v, N(\tau_j v + w_j) - N(\tau_0 v)\rangle\Big| &\leq \max_{1\leq i\leq s}\sum_{j=1}^{s}|a_{ij}|\,\|N(\tau_j v + w_j) - N(\tau_0 v)\| \\ &\leq C_3((C_1+C_2)\|z_0\|^2 + p\|w_0\|)\|z_0\| \\ &\leq C_3((C_1+C_2)\gamma'(1+\alpha'^2)^{1/2} + p\alpha')(1+\alpha'^2)^{1/2}\tau_0^2 \\ &\leq \frac{\nu}{2}\tau_0^2,\end{aligned} \tag{III.42}$$

con $C_3 = (1+2p)\delta_2 \max_{1\leq i\leq s}\sum_{j=1}^{s}|a_{ij}|$.

Ahora usando (III.41),(III.42) y *(H5)*, podemos escribir

$$|\tau_i - \tau_0 + \nu c_i \tau_0^2| \leq \frac{\nu}{2}\tau_0^2, \quad 1 \leq i \leq s,$$

siendo $c_i := \sum_{j=1}^{s} a_{ij}$. Esto prueba la afirmación (III.40) por cuanto $c_s = 1$ (el método se considera en forma stiffly accurate, esto es, $Z_s = hy_1$).

Para mostrar (III.35) basta comprobar que

$$\|z_1\|^2 - \|z_0\|^2(1-\kappa\|z_0\|)^2 \leq 0, \quad \text{para algún } \kappa > 0. \tag{III.43}$$

Así, escribiendo $\|w_0\| = \theta\tau_0,\ \tau_0 > 0,\ \theta \geq 0$, tenemos que $\|z_0\|^2 = (1+\theta^2)\tau_0^2$. Además, ya que R_s coincide con la función de estabilidad lineal del método y éste satisface $(P2)$, considerando (III.39) sigue que

$$\|w_s\| \leq \|w_0\| + C_2\|z_0\|^2,$$

lo cual implica que

$$\|w_s\| \leq \tau_0(\theta + C_2(1+\theta^2)\tau_0). \tag{III.44}$$

De (III.40) y (III.44), y tras efectuar un cálculo elemental se obtiene que

$$\begin{aligned}\|z_1\|^2 - \|z_0\|^2(1-\kappa\|z_0\|)^2 &\leq \tau_0^3\left(-\nu + 2\kappa(1+\theta^2)^{3/2} + 2C_2\theta(1+\theta^2)\right) \\ &\quad +\tau_0^4\left((\nu/2)^2 + (C_2^2 - \kappa^2)(1+\theta^2)^2\right).\end{aligned}$$

De aquí concluimos (III.43) para algunas constantes positivas $\kappa, \alpha'', \gamma''$, siempre que $\|w_0\| \leq \alpha''\tau_0$, $\|z_0\| \leq \gamma''$ y $h > 0$.

Finalmente, para mostrar (III.36), tomamos $\alpha := \min\{\alpha',\ \alpha''\}$ y asumimos $\|w_0\| \leq \alpha\tau_0$. Entonces a partir de (III.38) sigue que

$$|\tau_s - \tau_0| \leq C_1(1+\alpha^2)\tau_0^2. \tag{III.45}$$

Por otro lado, si $w_s \neq 0$ entonces de (III.39) se deduce que

$$\begin{aligned} |\langle \|w_s\|^{-1} w_s, w_s - R_s(hJ)w_0\rangle| &\leq \|w_s - R_s(hJ)w_0\| \\ &\leq C_2(1+\alpha^2)\tau_0^2. \end{aligned}$$

Esto implica que

$$\|w_s\| \leq |\langle \|w_s\|^{-1} w_s, R_s(hJ)w_0\rangle| + C_2(1+\alpha^2)\tau_0^2.$$

Ahora, considerando el Corolario III.5.2 obtenemos que

$$\|w_s\| \leq M\|w_0\| + C_2(1+\alpha^2)\tau_0^2,$$

donde

$$M := \max\{\rho, 1 - \varrho\delta_1 h\}.$$

Consecuentemente,

$$\|w_s\| \leq \alpha M\tau_0 + C_2(1+\alpha^2)\tau_0^2. \tag{III.46}$$

La condición (III.46) es claramente cierta también cuando $w_s = 0$.

Entonces, asumiendo que $0 < \tau_0 < (C_1(1+\alpha^2))^{-1}$ sigue de (III.45) y (III.46) que $\tau_s > 0$ y

$$\frac{\|w_s\|}{\tau_s} \leq \alpha\frac{M + C_2(\alpha^{-1}+\alpha)\tau_0}{1 - C_1(1+\alpha^2)\tau_0}. \tag{III.47}$$

De esta manera, considerando $\gamma := \min\{\gamma',\ \gamma'',\ \gamma'''\}$, con

$$\gamma''' := \min\left\{(C_1(1+\alpha^2))^{-1}, \frac{\alpha(1-\rho)}{(1+\alpha^2)(C_1\alpha + C_2)}\right\},$$

y

$$\beta := \frac{\varrho\delta_1\alpha}{(1+\alpha^2)(\alpha C_1 + C_2)},$$

obtenemos (III.36) a partir de (III.47) si $\|z_0\| \leq \min\{\gamma, h\beta\}$. Esto concluye la prueba. □

Nota III.5.4 No es complicado ver que una condición suficiente para obtener la propiedad *(P1)* para un método Runge-Kutta $RK(A,b)$, no necesariamente en forma stiffly accurate,

(y, en consecuencia, para garantizar la resolubilidad de la ecuación de las etapas internas del método) es la siguiente propiedad

$$\sup_{\mathrm{Re}\,\mu<0} \left\|(I-\mu A)^{-1}\right\| < \infty. \tag{III.48}$$

La norma en la ecuación anterior es irrelevante.

Esta propiedad ha sido ampliamente usada tanto en el estudio de la existencia de solución como en las propiedades de estabilidad y convergencia de los métodos Runge-Kutta sobre problemas lineales y semilineales de tipo stiff (ver, por ejemplo, [9, 14, 15]. Por otro lado, en [15] se demuestra que (III.48) es equivalente a que los autovalores no nulos de A tengan parte real positiva y que el autovalor nulo tenga igual multiplicidad algebraica y geométrica, esto es,

$$\lambda \in \sigma[A] \Rightarrow \begin{cases} \mathrm{Re}\,\lambda > 0, \text{ o} \\ \lambda = 0 \text{ y posee la misma multiplicidad algebraica y geométrica.} \end{cases} \tag{III.49}$$

Aquí $\sigma[A]$ representa el espectro de la matriz A.

Como consecuencia del Teorema III.5.3 se deducen los siguientes resultados.

Teorema III.5.5 *Si un método Runge-Kutta A-estable satisface (III.49) y $|R(\infty)| < 1$ entonces es E-estable.*

Demostración. El resultado sigue considerando el Teorema III.5.3, teniendo en cuenta que *(P1)* se obtiene debido a (III.49) y que *(P2)* sigue de la propiedad de A-estabilidad fuerte del método Runge-Kutta en virtud del teorema III.4.4. □

Corolario III.5.6 *(i) Los métodos Runge-Kutta de las familias Radau IA, Radau IIA y Lobatto IIIC son E-estables.*
(ii) Todos los métodos DIRK (Diagonally Implicit Runge-Kutta) y SIRK (Single Implicit Runge-Kutta) A-estables verificando $|R(\infty)| < 1$ son E-estables.

Demostración.

(i) Los métodos de las familias RadauIA y RadauIIA poseen como función de estabilidad lineal al aproximante de Padé $\Pi_{s-1,s}$ a la función exponencial, mientras que los métodos LobattoIIIC poseen al aproximante $\Pi_{s-2,s}$ como función de estabilidad lineal. De aquí que estos métodos sean fuertemente A-estables. Además, las matrices de coeficientes que definen a estos métodos tienen todos sus autovalores con parte real positiva y en consecuencia satisfacen (III.48).

(ii) Los métodos $DIRK$ son aquellos que poseen una matriz de coeficientes triangular inferior, y en caso de A-estabilidad su diagonal debe estar formada por elementos positivos. Por lo tanto estos métodos satisfacen (III.48).

Por otro lado, los métodos $SIRK$ poseen una matriz de coeficientes con un único autovalor real, que en caso de A-estabilidad debe ser positivo. De este modo, estos métodos también verifican (III.48). □

Nota III.5.7 El Teorema III.5.3 puede ser generalizado a métodos Runge-Kutta linealmente implícitos (o métodos de tipo Rosenbrock), incluso en el caso de métodos con matrices Jacobianas exactas pero mantenidas en el tiempo (time-lagged Jacobian matrices). Estas ideas serán discutidas con más detalle en el siguiente capítulo.

III.6. Experimentos numéricos

Aunque la necesidad de la propiedad *(P2)* introducida al comienzo de la sección previa para la garantizar la E-estabilidad de los métodos de tipo Runge-Kutta no ha sido demostrada, los experimentos numéricos presentados en el capítulo anterior usando los $\theta-$métodos (II.32) sugieren que tal propiedad es en cierto modo, y desde un punto de vista práctico, algo necesaria. En particular, los experimentos mostrados en el capítulo previo reflejan que la regla trapezoidal ($\theta = \frac{1}{2}$) integra de modo inestable, mientras que los $\theta-$métodos con $\theta > 1/2$ proceden de forma satisfactoria al integrar el problema test (II.34) cuando se consideran integraciones numéricas en amplios intervalos temporales.

En esta sección consideramos ejemplos de carácter más práctico, no sólo un problema académico. Consideremos la integración en largos tiempos de los ejemplos (II.11) y (II.13). Para estos problemas la solución exacta tiende al punto de equilibrio $x = 0$ cuando $t \to \infty$, y la pendiente de la solución en dicho punto viene determinada de acuerdo al Teorema de la Variedad Centro por cualquier autovector asociado al autovalor nulo de la matriz Jacobiana en el equilibrio $J = f'(0)$. En este sentido, consideraremos la pendiente de las soluciones numéricas que proveen los integradores de tipo Runge-Kutta aplicados a estos problemas como indicador del comportamiento cualitativo y estabilidad de las soluciones numéricas.

En primer lugar, y continuando en la línea de los experimentos numéricos mostrados en el capítulo previo, hemos considerado la integración numérica de los problemas (II.11) y (II.13) a

través de los $\theta-$métodos (II.32) con tamaño de paso inicial $h_0 = 10^{-6}$ siguiendo una estrategia a paso variable en la que el error local se estima por extrapolación según la fórmula estándar (II.35) tras dos pasos consecutivos de igual longitud, y donde el tamaño de paso es actualizado en caso de aceptación o rechazo según las expresiones (II.36) o (II.37), respectivamente. La tolerancia de error en el punto de avance se ha calculado según la fórmula (II.38). Las ilustraciones que se presentan en las figuras III.1-III.6 obedecen a una integración de estas características con un tolerancia de error media $TOL = 10^{-5}$.

La función de estabilidad lineal $R(z, \theta)$ de los $\theta-$métodos toma un valor en el infinito dado por (II.39). De esta manera, de cara a los ejemplos citados arriba hemos considerado un valor del parámetro θ para el que $|R(\infty, \theta)|$ es notablemente menor que 1 para ilustrar que el correspondiente método integra satisfactoriamente, así como un valor que provea un valor $|R(\infty, \theta)|$ tan próximo a 1 que impida una integración eficiente.

En la figuras III.1, III.2 y III.5 observamos que el valor $\theta = 0.7$ es suficiente para obtener integraciones estables en ambos ejemplos (II.11) y (II.13), y que la pendiente de la solución en el punto final se ajusta aceptablemente al valor teórico conocido para la solución exacta de tales problemas. Por otro lado, para ambos problemas observamos que el valor $\theta = 0.51$ provee una integración deficiente. Notamos en las figuras III.3, III.4 y III.6 que la carencia de la propiedad de A-estabilidad fuerte implica un mal comportamiento de la pendiente numérica, lo cual a su vez conlleva a inestabilidades en la integración.

Por otro lado, hemos considerado además la integración de los problemas (II.11) y (II.13) por medio de la familia uniparamétrica de métodos de 2 etapas $SDIRK$ introducidos por Nørsett en [72], y que, en forma de *tabla de Butcher* [52, p. 133], vienen dados por la tabla III.1. Estos métodos poseen orden de consistencia dos en general, excepto para los valores del parámetro $\gamma = 1/2 \pm \sqrt{3}/6$, para los cuales los métodos alcanzan orden 3. Más aún, estos métodos son A-estables (y B-estables) si y sólo si $\gamma \geq \frac{1}{4}$; y son fuertemente A-estables si y sólo si $\gamma \in (1/4, \infty) \setminus \{1/2\}$.

Un sencillo cálculo muestra que para estos métodos la función de estabilidad lineal toma la expresión

$$R(z, \gamma) = \frac{1 + (1 - 2\gamma)z + (\frac{1}{2} - 2\gamma + \gamma^2)z^2}{(1 - \gamma z)^2}$$

y de aquí que

$$R(\infty, \gamma) = \frac{\frac{1}{2} - 2\gamma + \gamma^2}{\gamma^2}.$$

Tabla III.1: Métodos SDIRK de 2 etapas

$$\begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline & 1/2 & 1/2 \end{array}$$

Además, denotando l_2EC la norma l_2 de los coeficientes del *término principal de error* [52, p. 158] del método correspondiente obtenemos que

$$l_2EC(\gamma) = \frac{\sqrt{5}}{2}\left|\gamma^2 - \gamma + \frac{1}{6}\right|.$$

La tabla III.2 muestra los valores de $R(\infty, \gamma)$ y $l_2EC(\gamma)$ para algunos valores particulares del parámetro γ que serán considerados en las integraciones de los problemas anteriormente mencionados.

Tabla III.2: Algunas propiedades de $SDIRK(\gamma)$

γ	$R(\infty, \gamma)$	$l_2EC(\gamma)$
$1 - \frac{\sqrt{2}}{2} \simeq 0.292893218$	0	0.045213422
0.26	0.704142011	0.028770741
0.254	0.875999752	0.025510554
0.2501	0.996802558	0.023348265
0.25	1	0.023292374

Estos métodos han sido implementados usando extrapolación local clásica como estimador del error, de tal modo que la tolerancia de error en el punto t_{2n} se calcula por medio de la fórmula

$$Tol_{2n} = Rtol \cdot \max\{\|y_{2n}\|_\infty, Atol\},\ n \geq 1,$$

donde $Rtol$ representa la tolerancia de error relativo y $Rtol \cdot Atol$ denota la tolerancia de error absoluto. Los resultados que se muestran en las tablas del final de la sección corresponden a los valores $Atol = 10^{-1}$ y $Rtol = TOL$, siendo TOL un parámetro a elegir.

En cada paso las etapas internas del método se obtienen por aplicación de la iteración usual de Newton modificada hasta la convergencia. El código toma en todo caso dos pasos consecutivos de la misma amplitud y el tamaño de paso se reduce a la mitad en caso de rechazo por el estimador de error o bien en caso de divergencia en la iteración que resuelve las etapas. El hecho de reducir el paso a la mitad tiene su sentido en el hecho de que en caso de rechazo entonces se ahorra el cómputo de la solución numérica con paso doble, pues ya se habría calculado en el paso rechazado anterior. En caso de aceptación del paso, el cambio en el tamaño de paso se lleva a cabo de forma estándar (II.36).

En las tablas que van desde la tabla III.3 hasta la tabla III.10, hemos resumido los resultados obtenidos por medio de la integración con los métodos $SDIRK$ dados anteriormente para algunos valores representativos del parámetro γ. En todos los casos presentados el tamaño de paso inicial fue elegido como $h_0 = 10^{-6}$. En estas tablas NPA denota el número de pasos necesarios para concluir la integración; $NPRE$ representa el número de pasos rechazados por el código debido a rechazo en la estimación del error; $NPRC$ denota el número de pasos rechazados debido a divergencia en la iteración tipo Newton que resuelve las etapas del método; $Rmedio$ representa el promedio de las razones de paso a lo largo de toda la integración; $Y_i(T)$ es la i-ésima componente de la solución numérica en el punto temporal T; T_{break} indica que la integración fue abortada cuando el código numérico se vio forzado a tomar un elevado número de pasos para poder proseguir la integración; y T_{inest} indica que la integración fue abortada cuando las soluciones numéricas provistas por el código cruzan a la región de inestabilidad del problema.

Se puede observar en estas tablas que aquellos métodos con $|R(\infty)|$ notablemente menor que 1 fueron capaces de alcanzar el punto final del intervalo de integración T_f en una cantidad razonable de pasos, mientras que para $\gamma = 1/4$ la integración no fue satisfactoria en ninguno de los casos, bien debido a inestabilidad en la solución numérica (ver la tabla III.6 correspondientes a la integración del ejemplo (II.11)), o bien debido a un excesivo número de pasos en la integración (ver la tabla III.10 correspondiente a la integración del ejemplo (II.13)). Además, a medida que $\gamma \to 1/4$ (esto es, $|R(\infty)| \to 1$) la integración se vuelve más inestable y el comportamiento cualitativo de las soluciones, en lo que a la pendiente numérica se refiere, es más deficiente (compárese el comportamiento de la pendiente numérica para los valores de $\gamma = 1 - \sqrt{2}/2$ y $\gamma = 0.26$ con el comportamiento de la pendiente para un valor de γ próximo a 0.25 en las tablas III.5 y III.9). También se debe remarcar que en general el método L-estable ($\gamma = 1 - \sqrt{2}/2$) no provee mejores resultados que, por ejemplo, el caso obtenido para $\gamma = 0.26$

para el cual $R(\infty) \simeq 0.70$ (ver los resultados en las tablas III.3 y III.4, y III.7 y III.8). En estos dos casos, ambos métodos proveen soluciones similares, aunque sin embargo el método L-estable se ve forzado a dar un mayor número de pasos. Esto se debe muy probablemente al hecho de que el segundo método posee un valor más pequeño para la norma de los coeficientes del error l_2EC.

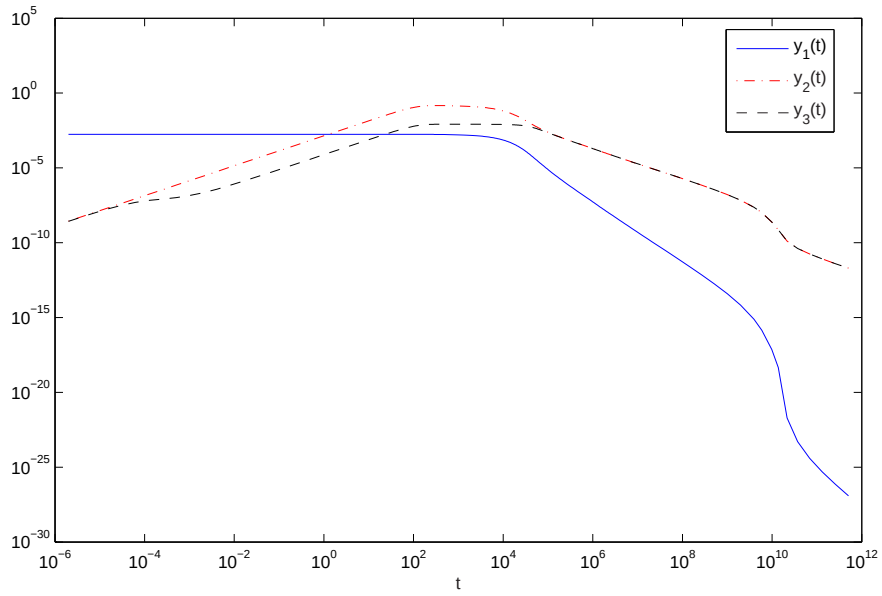

Figura III.1: Ejemplo 1 (II.11), $\theta = 0.7$. Componentes de la solución (escala logarítmica).

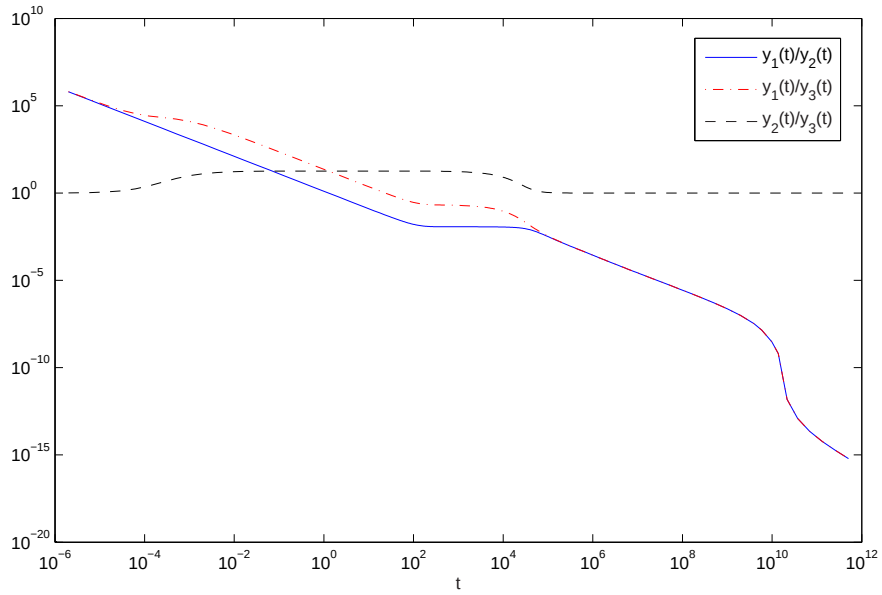

Figura III.2: Ejemplo 1 (II.11), $\theta = 0.7$. Pendientes numéricas (escala logarítmica).

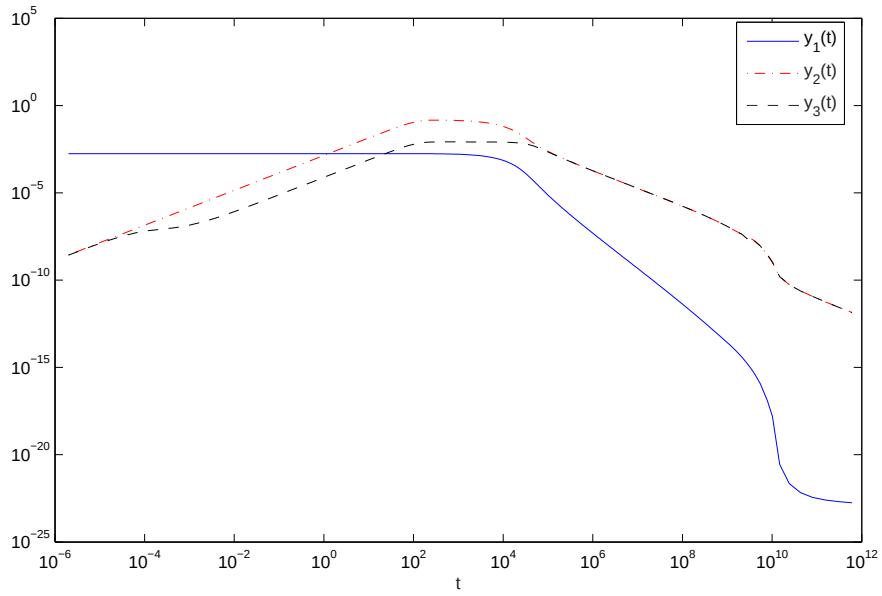

Figura III.3: Ejemplo 1 (II.11), $\theta = 0.51$. Componentes de la solución (escala logarítmica).

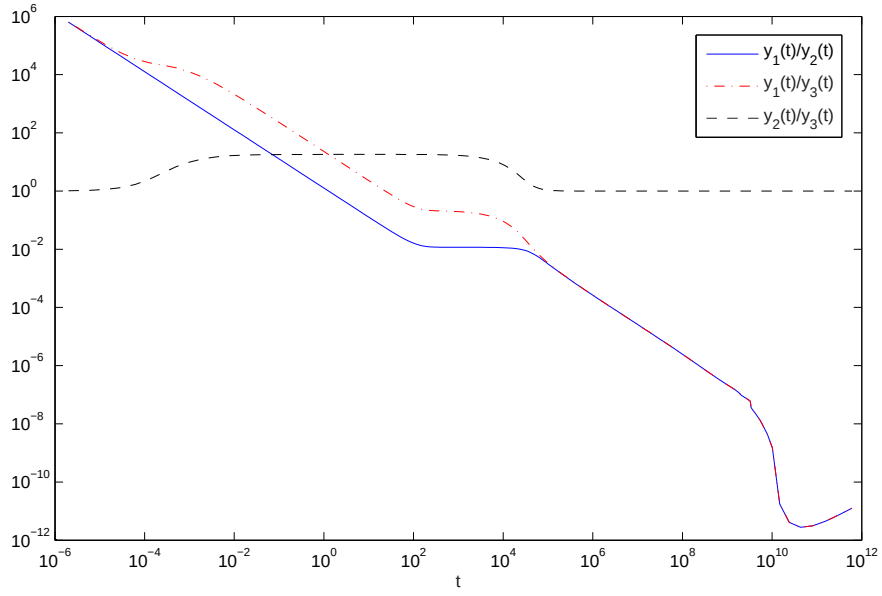

Figura III.4: Ejemplo 1 (II.11), $\theta = 0.51$. Pendientes numéricas (escala logarítmica).

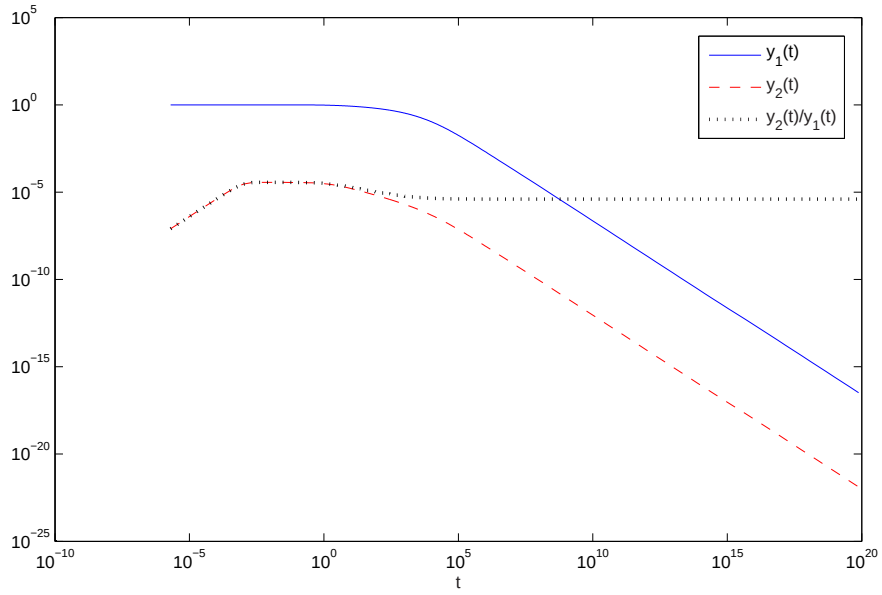

Figura III.5: Ejemplo 2 (II.13), $\theta = 0.7$. Componentes de la solución y pendiente numérica (escala logarítmica).

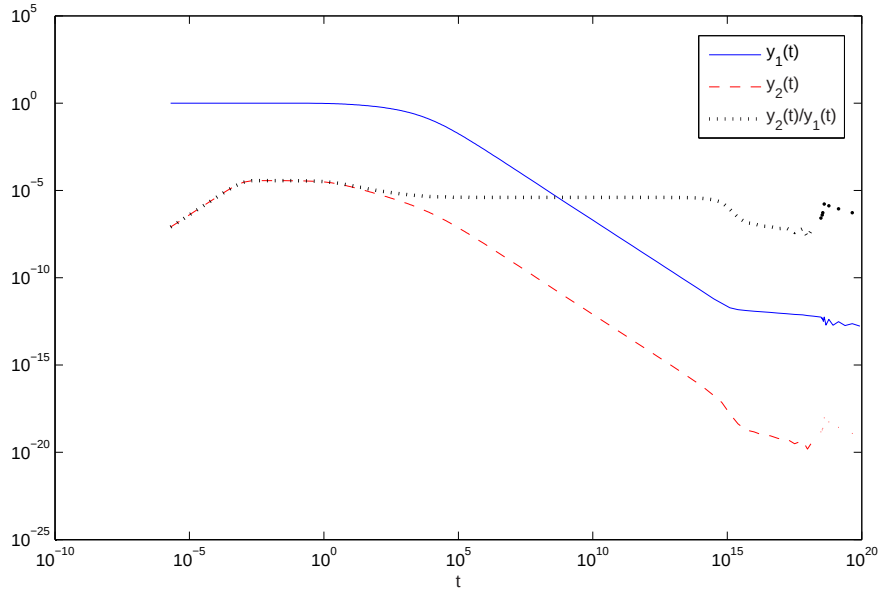

Figura III.6: Ejemplo 2 (II.13), $\theta = 0.51$. Componentes de la solución y pendiente numérica (escala logarítmica).

Tabla III.3: Ejemplo 1 (II.11). $SDIRK$, $\gamma = 1 - \frac{\sqrt{2}}{2}$.

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_3(T_f)$	$Y_1(T_f)/Y_2(T_f)$	$Y_2(T_f)/Y_3(T_f)$
10^{-3}	194	0	32	1.2986	$0.8830 \cdot 10^{-12}$	$0.2087 \cdot 10^{-25}$	$0.10000000 \cdot 10^{1}$
10^{-4}	216	0	20	1.2630	$0.8842 \cdot 10^{-12}$	$0.4759 \cdot 10^{-23}$	$0.10000000 \cdot 10^{1}$
10^{-5}	322	0	15	1.1721	$0.8837 \cdot 10^{-12}$	$0.1296 \cdot 10^{-22}$	$0.10000000 \cdot 10^{1}$
10^{-6}	644	0	11	1.0834	$0.8835 \cdot 10^{-12}$	$0.1879 \cdot 10^{-22}$	$0.10000000 \cdot 10^{1}$
10^{-7}	1568	0	7	1.0328	$0.8815 \cdot 10^{-12}$	$0.2879 \cdot 10^{-23}$	$0.10000000 \cdot 10^{1}$
10^{-8}	4156	0	4	1.0121	$0.8845 \cdot 10^{-12}$	$0.5761 \cdot 10^{-24}$	$0.10000000 \cdot 10^{1}$
10^{-9}	11354	0	0	1.0043	$0.8842 \cdot 10^{-12}$	$0.3818 \cdot 10^{-23}$	$0.10000000 \cdot 10^{1}$

Tabla III.4: Ejemplo 1 (II.11). $SDIRK$, $\gamma = 0.26$.

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_3(T_f)$	$Y_1(T_f)/Y_2(T_f)$	$Y_2(T_f)/Y_3(T_f)$
10^{-3}	194	0	34	1.2977	$0.8890 \cdot 10^{-12}$	$0.1145 \cdot 10^{-17}$	$0.99999999 \cdot 10^{0}$
10^{-4}	204	0	20	1.2774	$0.8876 \cdot 10^{-12}$	$0.2658 \cdot 10^{-18}$	$0.99999997 \cdot 10^{0}$
10^{-5}	266	0	15	1.2088	$0.8890 \cdot 10^{-12}$	$0.1160 \cdot 10^{-17}$	$0.99999999 \cdot 10^{0}$
10^{-6}	464	4	10	1.1176	$0.8883 \cdot 10^{-12}$	$0.4910 \cdot 10^{-20}$	$0.10000000 \cdot 10^{1}$
10^{-7}	1082	6	6	1.0494	$0.8884 \cdot 10^{-12}$	$0.2543 \cdot 10^{-20}$	$0.10000000 \cdot 10^{1}$
10^{-8}	2818	6	3	1.0183	$0.8872 \cdot 10^{-12}$	$0.1405 \cdot 10^{-18}$	$0.10000000 \cdot 10^{1}$
10^{-9}	7492	8	1	1.0066	$0.8877 \cdot 10^{-12}$	$0.5780 \cdot 10^{-19}$	$0.10000000 \cdot 10^{1}$

Tabla III.5: Ejemplo 1 (II.11). $SDIRK$, $\gamma = 0.254$.

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_3(T_f)$	$Y_1(T_f)/Y_2(T_f)$	$Y_2(T_f)/Y_3(T_f)$
10^{-3}	190	0	32	1.3039	$0.3123 \cdot 10^{-11}$	$0.3698 \cdot 10^{-16}$	$0.28484620 \cdot 10^{0}$
10^{-4}	202	0	19	1.2797	$0.2939 \cdot 10^{-11}$	$0.1953 \cdot 10^{-16}$	$0.30265399 \cdot 10^{0}$
10^{-5}	266	0	14	1.2085	$0.1715 \cdot 10^{-11}$	$0.9704 \cdot 10^{-17}$	$0.51872954 \cdot 10^{0}$
10^{-6}	416	12	9	1.1320	$0.9238 \cdot 10^{-12}$	$0.3916 \cdot 10^{-18}$	$0.96183220 \cdot 10^{0}$
10^{-7}	932	14	5	1.0564	$0.8891 \cdot 10^{-12}$	$0.2176 \cdot 10^{-17}$	$0.99978898 \cdot 10^{0}$
10^{-8}	2398	10	3	1.0214	$0.8897 \cdot 10^{-12}$	$0.1384 \cdot 10^{-16}$	$0.99999879 \cdot 10^{0}$
10^{-9}	6214	8	0	1.0079	$0.8900 \cdot 10^{-12}$	$0.1977 \cdot 10^{-16}$	$0.99999999 \cdot 10^{0}$

Tabla III.6: Ejemplo 1 (II.11). $SDIRK$, $\gamma = 0.25$.

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_1(T_{inest})$	$Y_2(T_{inest})$	$Y_3(T_{inest})$	T_{inest}
10^{-3}	136	0	20	1.3235	$0.16 \cdot 10^{-10}$	$0.32 \cdot 10^{-5}$	$-0.11 \cdot 10^{-5}$	$0.53 \cdot 10^{8}$
10^{-4}	170	0	24	1.2452	$0.36 \cdot 10^{-10}$	$0.50 \cdot 10^{-5}$	$-0.28 \cdot 10^{-7}$	$0.36 \cdot 10^{8}$
10^{-5}	218	0	2	1.1936	$0.20 \cdot 10^{-11}$	$0.12 \cdot 10^{-5}$	$-0.34 \cdot 10^{-6}$	$0.15 \cdot 10^{9}$
10^{-6}	374	0	0	1.1115	$0.29 \cdot 10^{-12}$	$0.44 \cdot 10^{-6}$	$-0.63 \cdot 10^{-7}$	$0.37 \cdot 10^{9}$
10^{-7}	770	22	10	1.0547	$0.23 \cdot 10^{-13}$	$0.13 \cdot 10^{-6}$	$-0.22 \cdot 10^{-8}$	$0.11 \cdot 10^{10}$
10^{-8}	2104	16	0	1.0209	$0.13 \cdot 10^{-17}$	$0.12 \cdot 10^{-8}$	$-0.96 \cdot 10^{-9}$	$0.11 \cdot 10^{11}$
10^{-9}	5332	22	0	1.0082	$0.20 \cdot 10^{-22}$	$0.89 \cdot 10^{-10}$	$-0.61 \cdot 10^{-10}$	$0.23 \cdot 10^{11}$

Tabla III.7: Ejemplo 2 (II.13). $SDIRK$, $\gamma = 1 - \frac{\sqrt{2}}{2}$.

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-2}	180	0	4	1.4728	$0.2641 \cdot 10^{-16}$	$0.1055 \cdot 10^{-21}$	$0.39999840 \cdot 10^{-5}$
10^{-3}	198	0	6	1.4251	$0.2587 \cdot 10^{-16}$	$0.1035 \cdot 10^{-21}$	$0.39999840 \cdot 10^{-5}$
10^{-4}	240	0	2	1.3397	$0.2552 \cdot 10^{-16}$	$0.1021 \cdot 10^{-21}$	$0.39999840 \cdot 10^{-5}$
10^{-5}	346	0	0	1.2286	$0.2567 \cdot 10^{-16}$	$0.1027 \cdot 10^{-21}$	$0.39999840 \cdot 10^{-5}$
10^{-6}	586	0	0	1.1313	$0.2702 \cdot 10^{-16}$	$0.1081 \cdot 10^{-21}$	$0.39999840 \cdot 10^{-5}$
10^{-7}	1108	0	0	1.0679	$0.2685 \cdot 10^{-16}$	$0.1074 \cdot 10^{-21}$	$0.39999840 \cdot 10^{-5}$
10^{-8}	2238	0	0	1.0331	$0.2581 \cdot 10^{-16}$	$0.1032 \cdot 10^{-21}$	$0.39999840 \cdot 10^{-5}$

Tabla III.8: Ejemplo 2 (II.13). $SDIRK$, $\gamma = 0.26$.

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-2}	176	0	2	1.4815	$0.2568 \cdot 10^{-16}$	$0.1027 \cdot 10^{-21}$	$0.39995416 \cdot 10^{-5}$
10^{-3}	188	0	2	1.4442	$0.2583 \cdot 10^{-16}$	$0.1033 \cdot 10^{-21}$	$0.39998827 \cdot 10^{-5}$
10^{-4}	226	0	1	1.3604	$0.2581 \cdot 10^{-16}$	$0.1032 \cdot 10^{-21}$	$0.39999738 \cdot 10^{-5}$
10^{-5}	318	0	0	1.2485	$0.2667 \cdot 10^{-16}$	$0.1067 \cdot 10^{-21}$	$0.39999831 \cdot 10^{-5}$
10^{-6}	522	0	0	1.1484	$0.2610 \cdot 10^{-16}$	$0.1044 \cdot 10^{-21}$	$0.39999839 \cdot 10^{-5}$
10^{-7}	970	0	0	1.0781	$0.2608 \cdot 10^{-16}$	$0.1043 \cdot 10^{-21}$	$0.39999840 \cdot 10^{-5}$
10^{-8}	1942	0	0	1.0381	$0.2674 \cdot 10^{-16}$	$0.1070 \cdot 10^{-21}$	$0.39999840 \cdot 10^{-5}$

Tabla III.9: Ejemplo 2 (II.13). $SDIRK$, $\gamma = 0.2501$.

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-2}	6040	0	2934	1.0151	$0.3680 \cdot 10^{-12}$	$0.2927 \cdot 10^{-15}$	$0.79541670 \cdot 10^{-3}$
10^{-3}	6022	0	2920	1.0150	$0.3730 \cdot 10^{-12}$	$0.2998 \cdot 10^{-15}$	$0.80377565 \cdot 10^{-3}$
10^{-4}	5934	0	2860	1.0167	$0.3594 \cdot 10^{-12}$	$0.2800 \cdot 10^{-15}$	$0.77895623 \cdot 10^{-3}$
10^{-5}	5576	0	2637	1.0158	$0.3764 \cdot 10^{-12}$	$0.3054 \cdot 10^{-15}$	$0.81129387 \cdot 10^{-3}$
10^{-6}	5024	0	2266	1.0167	$0.4327 \cdot 10^{-12}$	$0.3956 \cdot 10^{-15}$	$0.91426594 \cdot 10^{-3}$
10^{-7}	4556	0	1822	1.0193	$0.3436 \cdot 10^{-12}$	$0.2579 \cdot 10^{-15}$	$0.75063939 \cdot 10^{-3}$
10^{-8}	3866	0	1024	1.0216	$0.3624 \cdot 10^{-12}$	$0.2837 \cdot 10^{-15}$	$0.78278037 \cdot 10^{-3}$

Tabla III.10: Ejemplo 2 (II.13). $SDIRK$, $\gamma = 0.25$.

TOL	NPA	$NPRE$	$NPRC$	$Rmedio$	$Y_1(T_{break})$	$Y_2(T_{break})$	T_{break}
10^{-2}	20002	0	9963	1.0019	$0.8135 \cdot 10^{-6}$	$0.5520 \cdot 10^{-7}$	$0.25 \cdot 10^{10}$
10^{-3}	20002	0	9959	1.0018	$0.1334 \cdot 10^{-5}$	$0.5394 \cdot 10^{-7}$	$0.16 \cdot 10^{10}$
10^{-4}	20002	0	9941	1.0018	$0.6206 \cdot 10^{-6}$	$0.4410 \cdot 10^{-7}$	$0.34 \cdot 10^{10}$
10^{-5}	20002	0	9897	1.0018	$0.2058 \cdot 10^{-6}$	$0.1523 \cdot 10^{-7}$	$0.10 \cdot 10^{11}$
10^{-6}	20002	0	9800	1.0040	$0.2680 \cdot 10^{-7}$	$0.3326 \cdot 10^{-8}$	$0.78 \cdot 10^{11}$
10^{-7}	20002	0	9589	1.0020	$0.3139 \cdot 10^{-8}$	$0.4405 \cdot 10^{-9}$	$0.67 \cdot 10^{12}$
10^{-8}	20002	0	9132	1.0020	$0.2881 \cdot 10^{-9}$	$0.3728 \cdot 10^{-10}$	$0.75 \cdot 10^{13}$

Capítulo IV

Métodos semi-implícitos y sistemas diferenciales con equilibrios semiestables.

IV.1. Consideraciones preliminares

En el capítulo anterior se introdujo el concepto de E-estabilidad en relación con la estabilidad de los métodos numéricos de un paso en entornos de los equilibrios, y se obtuvieron resultados positivos en lo referente a estabilidad para una amplia clase de métodos de tipo Runge-Kutta implícitos al considerar redes temporales cuyas razones de pasos consecutivos se consideran acotadas por alguna constante mayor que uno. Esto justifica porqué en la práctica muchos integradores implícitos abordan satisfactoriamente la integración de estos problemas. En dicho capítulo se demostró que prácticamente cualquier método Runge-Kutta satisfaciendo la propiedad de A-estabilidad fuerte es *E-estable*.

El objetivo de este capítulo es la extensión de los resultados de estabilidad previos a la clase de métodos semi-implícitos (también conocidos como métodos de tipo Rosenbrock), haciendo énfasis en la inclusión del caso de métodos semi-implícitos con matrices Jacobianas mantenidas en el tiempo (time-lagged Jacobian matrices). Esta extensión ni es trivial ni es consecuencia del estudio que se ha hecho en el capítulo tercero para el caso de los métodos de tipo Runge-Kutta. Los resultados que serán expuestos a continuación conforman una revisión del trabajo [43].

Los métodos de tipo Rosenbrock [77] surgen a través de la linealización de la ecuación

de etapas asociada a los métodos de tipo Runge-Kutta *diagonalmente implícitos* [53, p. 102], esto es, métodos de tipo Runge-Kutta cuya matriz de coeficientes es triangular inferior. Con la motivación de reducir el costo computacional que involucra la resolución de las ecuaciones implícitas que definen las etapas internas de los métodos de Runge-Kutta, la idea subyacente en la formulación de los métodos de Rosenbrock consiste en sustituir estas ecuaciones implícitas por la resolución de un determinado número (finito) de sistemas lineales, y es por esto que estos métodos también son conocidos en la literatura como métodos semi-implícitos.

Así pues, los métodos de Rosenbrock requieren en cada paso de la integración la evaluación de la matriz Jacobiana del dato f que define el sistema diferencial (II.1) en la aproximación inicial conocida, así como una descomposición LU de la misma para la resolución práctica de la ecuación de las etapas internas del método. Algunos métodos numéricos con matrices Jacobianas mantenidas en el tiempo han sido propuestos en [3, 81, 86] con vistas a una reducción en el costo computacional de los métodos de Rosenbrock originales sin una pérdida de precisión significativa.

En lo que sigue asumiremos que $f(y)$ es una función al menos tres veces continuamente diferenciable en un entorno del origen con radio $\eta > 0$, cumpliendo que $f(0) = 0$. Entonces, denotando $J = f'(0)$, asumimos que $f(y)$ puede ser expresada como en (II.2)-(II.3) satisfaciendo las H-hipótesis introducidas en el capítulo segundo.

Del mismo modo que en los dos capítulos previos, necesitamos considerar la clase de conjuntos invariantes $\mathcal{C}_{\alpha,\beta}$, denominados conos, que juegan un papel decisivo en los estudios de estabilidad de los métodos numéricos. Para constantes positivas α y β, recordamos que el cono $\mathcal{C}_{\alpha,\beta}$ se define por (II.14).

Del tercer capítulo recordamos que un método de un paso

$$y_{n+1} = \varphi_h(y_n),\ t_0 = 0,\ h_n = t_{n+1} - t_n > 0, \quad n = 0, 1, \ldots,$$

se dice E-estable si existen cuatro constantes positivas $\gamma, \kappa, \alpha, \beta$ (únicamente dependientes de las constantes involucradas en las H-hipótesis y de los coeficientes del método), tales que

$$y_0 \in \mathcal{C}^+_{\alpha,\beta}, \quad \text{y} \quad h\,\|y_0\| \leq \gamma,\ h > 0$$

implica

$$y_1 \in \mathcal{C}^+_{\alpha,\beta}, \quad \text{y} \quad h\,\|y_1\| \leq h\,\|y_0\|\,(1 - \kappa h\,\|y_0\|),$$

siendo y_1 la solución de avance del método, $y_1 := \varphi_h(y_0)$.

El teorema III.2.3 del capítulo tercero garantiza un buen comportamiento de estabilidad para los métodos E-estables sobre redes temporales del tipo

$$\mathcal{P}_r = \{t_0 = 0 < t_1 < \ldots < t_n < \ldots;\ r_n := h_{n+1}/h_n \leq r,\ h_n := t_{n+1} - t_n\}$$

donde $r > 1$ es alguna constante que depende solamente de los coeficientes del método y de las constantes involucradas en las H-hipótesis.

Los resultados que mostraremos en este capítulo pueden ser probados en general bajo el marco de las H-hipótesis. Sin embargo, como en el capítulo previo, asumiremos por simplicidad y claridad en la exposición que los términos del desarrollo de Taylor de $f(y)$ de orden superior a dos se anulan, esto es,

$$f(y) = Jy + N(y). \tag{IV.1}$$

Además, será frecuente el uso de la aplicación bilineal simétrica $M(x, y)$ definida por (II.7), que extiende a la aplicación N en el sentido de que $M(y, y) = N(y)$ y satisface las acotaciones dadas en el lema II.2.6.

El resto del capítulo se organiza del modo siguiente. En la sección 2 se enuncia y demuestra una condición suficiente en términos de la A-estabilidad fuerte para los métodos de tipo Rosenbrock que permite garantizar su E-estabilidad. Posteriormente, en la sección 3, extendemos este resultado al caso de métodos semi-implícitos con matrices Jacobianas mantenidas en el tiempo. Para concluir este capítulo, la necesidad en la práctica de A-estabilidad fuerte será ilustrada con algunos experimentos numéricos en la sección 4.

IV.2. E-estabilidad de los métodos de Rosenbrock

Un método semi-implícito, o de Rosenbrock, de s etapas [53, Cap. IV.7.] aplicado con tamaño de paso $h > 0$ al sistema diferencial (II.1) sobre el punto (t_0, y_0) viene definido a partir de la fórmula de avance

$$y_1 = y_0 + \sum_{i=1}^{s} b_i K_i, \tag{IV.2}$$

donde las etapas internas K_i del método se calculan por

$$K_i = hf(y_0 + \sum_{j=1}^{s} a_{ij} K_j) + hf'(y_0) \sum_{j=1}^{s} c_{ij} K_j, \quad 1 \leq i \leq s, \tag{IV.3}$$

En (IV.3), $A = (a_{ij})$ es una matriz estrictamente triangular inferior de dimensión $s \times s$ y $C = (c_{ij})$ es una matriz triangular inferior de la misma dimensión y con elementos diagonales positivos. Además, asumiremos que el método posee al menos orden de consistencia uno, esto es, $b^T e = 1$, siendo $b^T := (b_1, \ldots, b_s)$ y $e := (1, \ldots, 1)^T \in \mathbb{R}^s$. Diversas relaciones de carácter no lineal entre los coeficientes del método pueden ser impuestas con vistas a asegurar orden de consistencia p para la solución numérica provista por el método en relación a la solución exacta del sistema diferencial que se integra. Detalles adicionales acerca de las condiciones de orden de un método de Rosenbrock pueden ser encontrados, por ejemplo, en [3] o [53, p. 104-108].

Teniendo en cuenta (IV.1) obtenemos de modo inmediato que

$$f'(y_0)y = Jy + 2M(y_0, y), \quad \forall y \in \mathbb{R}^m. \tag{IV.4}$$

De aquí, insertando (IV.1) y (IV.4) en (IV.3) llegamos a que

$$K_i = hJy_0 + hJ\sum_{j=1}^{s} d_{ij}K_j + h\left(N(y_0 + \sum_{j=1}^{s} a_{ij}K_j) + 2M(y_0, \sum_{j=1}^{s} c_{ij}K_j)\right), \quad 1 \le i \le s, \tag{IV.5}$$

donde $D := (d_{ij}) = A + C$.

Considerando el producto de Kronecker de matrices $A \otimes B = (a_{ij}B)$, y agrupando la parte lineal en (IV.5), obtenemos tras un sencillo cálculo que

$$K_i = R_i(hJ)y_0 + h\sum_{j=1}^{s} \varphi_{ij}(hJ)\left(N(y_0 + \sum_{l=1}^{s} a_{jl}K_l) + 2M(y_0, \sum_{l=1}^{s} c_{jl}K_l)\right), \quad 1 \le i \le s, \tag{IV.6}$$

siendo

$$\begin{aligned} R_i(hJ) &:= (e_i^T \otimes I)(I \otimes I - D \otimes hJ)^{-1}(e \otimes hJ), \quad 1 \le i \le s, \\ \varphi_{ij}(hJ) &:= (e_i^T \otimes I)(I \otimes I - D \otimes hJ)^{-1}(e_j \otimes I), \quad 1 \le i, j \le s, \end{aligned}$$

donde e_i denota el i-ésimo vector de la base canónica de $\mathbb{R}^s$ e I la matriz identidad de la dimensión adecuada. Téngase en cuenta que $(I \otimes I - D \otimes hJ)$ es una matriz triangular inferior por bloques no singular, puesto que D es una matriz triangular inferior con elementos diagonales positivos, $d_{ii} = c_{ii} > 0$, $1 \le i \le s$. Más aún, $Re\ \lambda \le 0$ para todo $\lambda \in \sigma[J]$ ($\sigma[J]$ denota el espectro de la matriz J).

Ahora, a partir de (IV.2) sigue que

$$y_1 = R(hJ)y_0 + h\sum_{i=1}^{s} \psi_i(hJ)\left(N(y_0 + \sum_{j=1}^{s} a_{ij}K_j) + 2M(y_0, \sum_{j=1}^{s} c_{ij}K_j)\right), \tag{IV.7}$$

donde

$$\begin{aligned} R(hJ) &:= I + (b^T \otimes I)(I \otimes I - D \otimes hJ)^{-1}(e \otimes hJ), \\ \psi_i(hJ) &:= (b^T \otimes I)(I \otimes I - D \otimes hJ)^{-1}(e_i \otimes I), \quad 1 \leq i \leq s. \end{aligned}$$

El desarrollo anterior nos conduce a considerar, para cada $1 \leq i, j \leq s$ las siguientes funciones auxiliares de variable compleja μ,

$$\begin{aligned} R(\mu) &= 1 + \mu b^T (I - \mu D)^{-1} e, \\ R_i(\mu) &= \mu e_i^T (I - \mu D)^{-1} e, \\ \varphi_{ij}(\mu) &= e_i^T (I - \mu D)^{-1} e_j, \\ \psi_i(\mu) &= b^T (I - \mu D)^{-1} e_i. \end{aligned}$$

Debe quedar claro que

$$R(0) = 1, \quad R_i(0) = 0, \quad \psi_i(0) = b_i, \quad \phi_{ij}(0) = e_i^T e_j, \qquad 1 \leq i, j \leq s.$$

Obsérvese además que $R(\mu)$ corresponde a la función de estabilidad lineal del método semi-implícito (IV.2)-(IV.3), y que ésta, a su vez, coincide con la función de estabilidad lineal del método de tipo Runge-Kutta *diagonalmente implícito* $DIRK(D, b)$ definido por la matriz de coeficientes D y el vector b^T.

Consideremos ahora el siguiente cambio de variables que facilitará el estudio de la E-estabilidad de los métodos

$$z_i = hy_i, \ i = 0, 1, \qquad V_j = hK_j, \ j = 1, \ldots, s.$$

Así, para las nuevas variables, las ecuaciones (IV.6) y (IV.7) se convierten, respectivamente, en

$$V_i = R_i(hJ) z_0 + \sum_{j=1}^{s} \varphi_{ij}(hJ) \left(N(z_0 + \sum_{l=1}^{s} a_{jl} V_l) + 2M(z_0, \sum_{l=1}^{s} c_{jl} V_l) \right), \quad 1 \leq i \leq s, \qquad \text{(IV.8)}$$

y

$$z_1 = R(hJ) z_0 + \sum_{i=1}^{s} \psi_i(hJ) \left(N(z_0 + \sum_{j=1}^{s} a_{ij} V_j) + 2M(z_0, \sum_{j=1}^{s} c_{ij} V_j) \right). \qquad \text{(IV.9)}$$

En virtud de la equivalencia entre (III.48) y (III.49) obtenemos que

$$\sup_{Re(\mu) \leq 0} \left\| (I - \mu D)^{-1} \right\| < \infty.$$

Por lo tanto como consecuencia de un teorema de J. Von Neumann (ver [53, p. 168], [87]), concluimos que existe una constante positiva p, dependiente únicamente de los coeficientes del método tal que

$$\left\{\begin{array}{rcl} \sup\limits_{h>0}\max\limits_{1\leq i\leq s}\|R_i(hJ)\| & \leq & p, \\ \sup\limits_{h>0}\|R(hJ)\| & \leq & p, \\ \sup\limits_{h>0}\max\limits_{1\leq i\leq s}\sum\limits_{j=1}^{s}\|\varphi_{ij}(hJ)\| & \leq & p, \\ \sup\limits_{h>0}\sum\limits_{i=1}^{s}\|\psi_i(hJ)\| & \leq & p. \end{array}\right. \tag{IV.10}$$

Con esto, y usando argumentos similares a los del capítulo tercero, no es complicado comprobar que existe una constante q dependiente únicamente de los coeficientes del método tal que

$$\|z_0\| < q\delta_2^{-1}, \quad h > 0,$$

implica la resolubilidad unívoca de (IV.8) de tal modo que

$$\begin{array}{rcl} \|V_i\| & \leq & 2p\,\|z_0\|\,,\ 1\leq i\leq s, \\ \|z_1\| & \leq & 2p\,\|z_0\|\,. \end{array} \tag{IV.11}$$

Más precisamente tenemos la siguiente

Proposición IV.2.1 *Sea $z_0\in\mathbb{R}^m$. Existe $q>0$ tal que*

i) si $\|z_0\|<q\delta_2^{-1}$, entonces la ecuación (IV.8) admite una única solución $V:=(V_1^T,\ldots,V_s^T)^T$ en

$$B_{\gamma_0}:=\{Y:=(Y_1^T,\ldots,Y_s^T)^T\in\mathbb{R}^{ms}/\,|Y|:=\max_{1\leq i\leq s}\|Y_i\|<\gamma_0\},$$

siendo $\gamma_0:=2pq\delta_2^{-1}$.

ii) si $\|z_0\|\leq q\delta_2^{-1}$, entonces (IV.8) admite solución en $\overline{B}_{\gamma_0}$.

Demostración. *i)* Dadas las matrices A y C que definen (IV.3) y la constante p que aparece en (IV.10), tomemos

$$q:=\min\left\{\frac{1}{(1+2p\,\|A\|_\infty)^2+4p\,\|C\|_\infty},\frac{1}{2p(\|A\|_\infty(1+2p\,\|A\|_\infty)+\|C\|_\infty)}\right\},$$

siendo $\|L\|_\infty=\max\limits_{1\leq i\leq s}\sum\limits_{j=1}^{s}|l_{ij}|$, para $L=(l_{ij})$, y sea $\|z_0\|<q\delta_2^{-1}$.

Sea entonces $\gamma_1 := 2p\,\|z_0\| < \gamma_0$, de modo que $\overline{B}_{\gamma_1} \subset B_{\gamma_0}$. Teniendo en mente la ecuación (IV.8) definimos $\Phi(Y) := (\Phi_1(Y)^T, \ldots, \Phi_s(Y)^T)^T \in \mathbb{R}^{ms}$, para cada $Y \in \mathbb{R}^{ms}$, siendo

$$\Phi_i(Y) := R_i(hJ)z_0 + \sum_{j=1}^{s} \varphi_{ij}(hJ)\left(N(z_0 + \sum_{l=1}^{s} a_{jl}Y_l) + 2M(z_0, \sum_{l=1}^{s} c_{jl}Y_l)\right), \quad 1 \le i \le s.$$

Entonces para $Y \in \overline{B}_{\gamma_1}$ tenemos para cualquier $1 \le i \le s$ que

$$\begin{aligned}
\|\Phi_i(Y)\| &\le p\,\|z_0\| + p\delta_2\Big((\|z_0\| + \|A\|_\infty |Y|)^2 + 2\,\|C\|_\infty \|z_0\|\,|Y|\Big) \\
&\le p\,\|z_0\| + p\delta_2\Big((1 + 2p\,\|A\|_\infty)^2 + 4p\,\|C\|_\infty\Big)\,\|z_0\|^2 \\
&= p\,\|z_0\|\left(1 + \delta_2\,\|z_0\|\,((1 + 2p\,\|A\|_\infty)^2 + 4p\,\|C\|_\infty)\right) \\
&< 2p\,\|z_0\| = \gamma_1.
\end{aligned}$$

Por otro lado, si $X, Y \in \overline{B}_{\gamma_1}$ entonces tenemos que

$$\begin{aligned}
\|\Phi_i(X) - \Phi_i(Y)\| &\le p\Big(\left\|N(z_0 + \sum_{j=1}^{s} a_{ij}X_j) - N(z_0 + \sum_{j=1}^{s} a_{ij}Y_j)\right\| \\
&\quad +2\left\|M(z_0, \sum_{j=1}^{s} c_{ij}X_j) - M(z_0, \sum_{j=1}^{s} c_{ij}Y_j)\right\|\Big) \\
&\le p\delta_2(2\,\|A\|_\infty |X - Y|(\|z_0\| + \gamma_1\,\|A\|_\infty) + 2\,\|z_0\|\,\|C\|_\infty |X - Y|) \\
&= 2p\delta_2(\|A\|_\infty (1 + 2p\,\|A\|_\infty) + \|C\|_\infty)\,\|z_0\|\,|X - Y| \\
&= L_\Phi |X - Y|,
\end{aligned}$$

siendo

$$L_\Phi := 2p\delta_2(\|A\|_\infty (1 + 2p\,\|A\|_\infty) + \|C\|_\infty)\,\|z_0\| < 1.$$

En definitiva, Φ es una función contractiva en $\overline{B}_{\gamma_1}$ y por tanto posee un único punto fijo $Y \in \overline{B}_{\gamma_1} \subset B_{\gamma_0}$ que verifica $|Y| \le 2p\,\|z_0\|$.

ii) Se deduce de *i*) por un argumento de continuidad. □

De modo similar al caso de los métodos de tipo Runge-Kutta, la siguiente condición sobre la función de estabilidad $R(\mu)$ asociada a un método de Rosenbrock resulta ser crucial para demostrar la E-estabilidad del método en consideración.

(P): Existen dos constantes positivas ϱ y $\rho < 1$ tales que la función crecimiento del error (III.17) satisface

$$\phi_R(x) := \sup_{Re(\mu) \le x} |R(\mu)| \le \max\{\rho, 1 + \varrho x\}, \forall x \le 0.$$

En la sección III.4 del capítulo previo se demostró que la propiedad *(P)* es equivalente a

$$|R(\infty)| < 1 \qquad \text{y} \qquad R(\mu) \text{ es } A-\text{aceptable},$$

cuando $R(\mu)$ es una función racional. Con todo lo anterior tenemos entonces el siguiente

Teorema IV.2.2 *Si el método semi-implícito (IV.2)-(IV.3) satisface (P), entonces es E-estable.*

Demostración. Denotemos $z_j = hy_j$, $j = 0, 1$, y $V_i = hK_i$, $1 \leq i \leq s$, siendo $h > 0$, y_0 valor inicial, y_1 la solución de avance del método dada por (IV.2), y K_i, $1 \leq i \leq s$, las etapas internas del método definidas implícitamente por (IV.3). Atendiendo a la descomposición $\mathbb{R}^m = span\{v\} \oplus v^{\perp}$ podemos escribir

$$z_i = \tau_i v + w_i, \quad w_i \in v^{\perp},\ i = 0, 1.$$

Entonces, mostrar la E-estabilidad del método (definición III.2.2) es equivalente a mostrar que existen cuatro constantes positivas $\gamma, \kappa, \alpha, \beta$, únicamente dependientes de las constantes involucradas en las *H-hipótesis* y de los coeficientes del método, de tal modo que para cada z_0 y cada h verificando

$$0 < \|z_0\| \leq \gamma, \quad \|z_0\| \leq h\beta \quad \text{y} \quad \|w_0\| \leq \alpha\tau_0$$

tengamos que

$$\|z_1\| \leq \|z_0\| (1 - \kappa \|z_0\|) \tag{IV.12}$$

y

$$\|w_1\| \leq \alpha\tau_1. \tag{IV.13}$$

Partiendo de (IV.8) y (IV.9), y teniendo en cuenta las propiedades (IV.10) y (IV.11), se deduce la existencia de una constante positiva η, dependiente exclusivamente de los coeficientes del método, tal que para todo $h > 0$

$$\|V_i - R_i(hJ)z_0\| \leq \eta \|z_0\|^2, \quad 1 \leq i \leq s,$$

y

$$\|z_1 - R(hJ)z_0\| \leq \eta \|z_0\|^2.$$

A partir de las dos ecuaciones anteriores, teniendo en cuenta (IV.10) y separando proyecciones sobre los subespacios $span\{v\}$ y $v^{\perp}$, respectivamente, obtenemos que

$$\|V_i\| \leq p \|w_0\| + \eta \|z_0\|^2, \quad 1 \leq i \leq s, \tag{IV.14}$$

$$\|w_1 - R(hJ)w_0\| \leq \eta \|z_0\|^2,$$

y

$$|\tau_1 - \tau_0| \leq \eta \|z_0\|^2. \tag{IV.15}$$

Además,

$$\begin{aligned} \|w_1\|^2 - |\langle w_1, R(hJ)w_0\rangle| &\leq |\langle w_1, w_1 - R(hJ)w_0\rangle| \\ &\leq \eta \|w_1\| \|z_0\|^2. \end{aligned}$$

Por lo tanto, aplicando el corolario III.5.2 se obtiene que

$$\|w_1\| \leq \xi \|w_0\| + \eta \|z_0\|^2, \quad \text{con} \quad \xi := \max\{\rho, 1 - h\varrho\delta_1\} < 1. \tag{IV.16}$$

La siguiente propiedad resultará de gran importancia para concluir la prueba del resultado: *existen dos constantes* $\gamma' > 0$ *y* $\alpha' > 0$ *tales que*

$$\tau_1 \leq \tau_0(1 - \frac{1}{2}\nu\tau_0), \quad \forall \|z_0\| \leq \gamma', \quad \forall \|w_0\| \leq \alpha'\tau_0, \quad \forall h > 0. \tag{IV.17}$$

Para ver esto último, ponemos

$$\begin{aligned} \mathbf{r} &:= |\tau_1 - \tau_0 + \nu\tau_0^2| \\ &= |\tau_1 - \tau_0 - \tau_0^2 \sum_{i=1}^{s} \psi_i(0)\langle v, N(v)\rangle|. \end{aligned}$$

Por otro lado, separando $N(v) = \langle v, N(v)\rangle v + N^{\perp}(v)$, donde $N^{\perp}(v)$ denota la proyección de $N(v)$ sobre $v^{\perp}$, y teniendo en cuenta que $Jv^{\perp} \subseteq v^{\perp}$, deducimos que

$$\begin{aligned} \langle \psi_i(hJ)N(v), v\rangle &= \langle N(v), v\rangle\langle \psi_i(hJ)v, v\rangle \\ &= \langle N(v), v\rangle\langle \psi_i(0)v, v\rangle \\ &= \psi_i(0)\langle N(v), v\rangle. \end{aligned}$$

Con esto sigue que

$$\begin{aligned} \mathbf{r} &= |\langle (\tau_1 - \tau_0)v + w_1 - w_0 - \tau_0^2 \sum_{i=1}^{s} \psi_i(hJ)N(v), v\rangle| \\ &= |\langle z_1 - z_0 - \sum_{i=1}^{s} \psi_i(hJ)N(\tau_0 v), v\rangle|. \end{aligned}$$

Reemplazando arriba z_1 por su expresión en (IV.9) obtenemos que

$$\begin{aligned}
\mathbf{r} &= |\langle (R(hJ)-I)\, z_0 + \sum_{i=1}^{s}\psi_i(hJ)\left(N(z_0+\sum_{j=1}^{s}a_{ij}V_j)-N(\tau_0 v)\right),\ v\rangle + \\
&\quad 2\langle \sum_{i=1}^{s}\psi_i(hJ)M(z_0,\sum_{j=1}^{s}c_{ij}V_j), v\rangle| \\
&= |\langle \sum_{i=1}^{s}\psi_i(hJ)\left((N(z_0+\sum_{j=1}^{s}a_{ij}V_j)-N(\tau_0 v)) + 2M(z_0,\sum_{j=1}^{s}c_{ij}V_j)\right),\ v\rangle| \\
&= |\sum_{i=1}^{s}\psi_i(0)\langle N(z_0+\sum_{j=1}^{s}a_{ij}V_j)-N(\tau_0 v), v\rangle + 2\sum_{i=1}^{s}\psi_i(0)\langle M(z_0,\sum_{j=1}^{s}c_{ij}V_j), v\rangle|.
\end{aligned}$$

De (II.8) y (II.9) obtenemos entonces que

$$\begin{aligned}
\mathbf{r} &\leq \delta_2\left(\sum_{i=1}^{s}|b_i|\right)\max_{1\leq i\leq s}\left(\left\|w_0+\sum_{j=1}^{s}a_{ij}V_j\right\|\cdot\left\|2\tau_0 v + w_0 + \sum_{j=1}^{s}a_{ij}V_j\right\|\right) + \\
&\quad 2\delta_2\left\|\tau_0 v + w_0\right\|\cdot\left\|\sum_{i,j=1}^{s}b_i c_{ij}V_j\right\|.
\end{aligned}$$

Ahora bien, teniendo en cuenta (IV.14) se tiene que

$$\begin{aligned}
\mathbf{r} &\leq \delta_2\left(\sum_{i=1}^{s}|b_i|\right)\left(\|w_0\|+\|A\|_\infty\left(p\|w_0\|+\eta\|z_0\|^2\right)\right)\left(2\tau_0+\|w_0\|+\|A\|_\infty\left(p\|w_0\|+\eta\|z_0\|^2\right)\right) \\
&\quad +2\delta_2\left(\sum_{i=1}^{s}|b_i|\right)\|C\|_\infty\left(p\|w_0\|+\eta\|z_0\|^2\right)\|z_0\| \\
&\leq \frac{\nu}{2}\tau_0^2,
\end{aligned}$$

si $\|z_0\| \leq \gamma'$ y $\|w_0\| \leq \alpha'\tau_0$, para valores de α' y γ' suficientemente pequeños. Esto prueba (IV.17).

Para deducir (IV.12) ponemos $\|w_0\| = \theta\tau_0$. Así, haciendo uso de (IV.16) y (IV.17), un cálculo directo permite mostrar que

$$\begin{aligned}
\|z_1\|^2 - \|z_0\|^2(1-\kappa\|z_0\|)^2 &\leq \tau_0^3\left(-\nu + 2\kappa(1+\theta^2)^{3/2} + 2\eta\theta(1+\theta^2)\right) \\
&\quad +\tau_0^4\left((\nu/2)^2 + (\eta^2-\kappa^2)(1+\theta^2)^2\right).
\end{aligned}$$

Esto permite asegurar la existencia de tres constantes positivas α'', γ'' y κ tales que (IV.12) se verifica cuando $\|w_0\| \leq \alpha''\tau_0$, $\|z_0\| \leq \gamma''$ y $h > 0$.

Para obtener (IV.13), tomamos $\alpha := \min\{\alpha', \alpha''\}$, y ponemos $\|w_0\| \leq \alpha\tau_0$. Definiendo

$$\gamma''' := min\{(\eta(1+\alpha^2))^{-1}, \frac{(1-\rho)\alpha}{(1+\alpha)(1+\alpha^2)\eta}\},$$

$$\gamma := \min\{\gamma', \gamma'', \gamma'''\}$$

y

$$\beta := \frac{\varrho\delta_1\alpha}{(1+\alpha)(1+\alpha^2)\eta}$$

sigue de (IV.15) y (IV.16) que

$$\frac{\|w_1\|}{\tau_1} \leq \frac{\xi\|w_0\| + \eta\|z_0\|^2}{\tau_0(1-\eta(1+\alpha^2)\tau_0)} \leq \alpha\frac{\xi + \eta(\alpha^{-1}+\alpha)\tau_0}{1-\eta(1+\alpha^2)\tau_0} \leq \alpha,$$

siempre que $\|z_0\| \leq \min\{\gamma, \beta h\}$. Esto concluye la prueba. □

IV.3. E-estabilidad de los métodos con matrices Jacobianas mantenidas en el tiempo

En esta sección tratamos la extensión de los resultados previos a la clase de métodos de Rosenbrock con matrices Jacobianas mantenidas en el tiempo. Notemos que la fórmula (IV.2)-(IV.3) requiere en cada paso de la integración la evaluación de la matriz Jacobiana del dato $f(y)$ del problema a integrar en la solución numérica computada en el paso previo. Nuestro objetivo en esta sección consiste en estudiar la E-estabilidad de los métodos de tipo Rosenbrock modificados de tal forma que la matriz Jacobiana que aparece en la ecuación de las etapas del método (IV.3) se mantiene fija durante una determinada cantidad prefijada de pasos temporales. Nos centraremos principalmente en la situación en la que las matrices Jacobianas son actualizadas cada dos pasos temporales. Sin embargo, el análisis de estabilidad permanece válido en el caso más general en el que las matrices Jacobianas son actualizadas tras k_{max} pasos temporales a lo sumo, siendo k_{max} un entero positivo prefijado.

El primer paso temporal dado por el método de Rosenbrock de s etapas (IV.2)-(IV.3) puede ser expresado de la siguiente manera

$$\begin{aligned} K^{(0)} &= h_0 F(e \otimes y_0 + (A \otimes I)K^{(0)}) + h_0(C \otimes J_0)K^{(0)} \\ y_1 &= y_0 + (b^T \otimes I)K^{(0)}, \end{aligned}$$

siendo

$$F(e\otimes z+(A\otimes I)K):=(f(u_1)^T,\ldots,f(u_s)^T)^T\in\mathbb{R}^{ms},$$

$$K:=(K_1^T,\ldots,K_s^T)^T\in\mathbb{R}^{ms},\qquad u_i:=z+\sum_{j=1}^{s}a_{ij}K_j\in\mathbb{R}^m,\quad 1\le i\le s.$$

Fijemos inicialmente $r^*>1$ y consideremos redes temporales de modo que las razones entre pasos consecutivos estén acotadas por $r^*>1$. En consecuencia, un segundo paso de tamaño $h_1=r_0h_0$ $(0<r_0\le r^*)$ viene dado usando la misma fórmula anterior, aunque en este caso consideramos la misma matriz Jacobiana que en el paso previo $J_0:=\dfrac{\partial f}{\partial y}(y_0)$, esto es,

$$\begin{aligned}K^{(1)} &= h_1F(e\otimes y_1+(A\otimes I)K^{(1)})+h_1(C\otimes J_0)K^{(1)}\\ y_2 &= y_1+(b^T\otimes I)K^{(1)}.\end{aligned}$$

Entonces tomando $h=h_0+h_1=(1+r_0)h_0=\dfrac{1+r_0}{r_0}h_1$ y definiendo

$$\begin{aligned}\widetilde{K}^{(0)} &:= (1+r_0)K^{(0)},\\ \widetilde{K}^{(1)} &:= \frac{1+r_0}{r_0}K^{(1)},\end{aligned}$$

resulta que

$$\begin{aligned}\widetilde{K}^{(0)} &= hF(e\otimes y_0+\left(\frac{1}{1+r_0}A\otimes I\right)\widetilde{K}^{(0)})+h\left(\frac{1}{1+r_0}C\otimes J_0\right)\widetilde{K}^{(0)},\\ \widetilde{K}^{(1)} &= hF(e\otimes y_0+\left(\frac{1}{1+r_0}eb^T\otimes I\right)\widetilde{K}^{(0)}+\left(\frac{r_0}{1+r_0}A\otimes I\right)\widetilde{K}^{(1)})+h\left(\frac{r_0}{1+r_0}C\otimes J_0\right)\widetilde{K}^{(1)},\end{aligned}$$

mientras que la solución de avance y_2 toma la expresión

$$y_2=y_0+\left(\frac{1}{1+r_0}b^T\otimes I\right)\widetilde{K}^{(0)}+\left(\frac{r_0}{1+r_0}b^T\otimes I\right)\widetilde{K}^{(1)}.$$

Así pues, el proceso anterior es teóricamente equivalente a considerar el método de tipo Rosenbrock con $2s$ etapas definido por la ecuación de etapas

$$\widetilde{K}=hF(\tilde{e}\otimes y_0+(A(r_0)\otimes I)\widetilde{K})+h(C(r_0)\otimes J_0)\widetilde{K},\tag{IV.18}$$

con solución de avance dada por

$$y_2=y_0+(b(r_0)^T\otimes I)\widetilde{K},\tag{IV.19}$$

siendo

$$\widetilde{K} := (\widetilde{K}^{(0)^T}, \widetilde{K}^{(1)^T})^T \in \mathbb{R}^{2ms},$$

$$\tilde{e} := (e^T, e^T)^T \in \mathbb{R}^{2ms},$$

y donde los coeficientes que definen al método

$$b(r_0)^T := \left(\frac{1}{1+r_0}b^T, \frac{r_0}{1+r_0}b^T\right) \in \mathbb{R}^{2s},$$

y

$$A(r_0) := \begin{pmatrix} \frac{1}{1+r_0}A & O \\ \frac{1}{1+r_0}eb^T & \frac{r_0}{1+r_0}A \end{pmatrix}, \qquad C(r_0) := \begin{pmatrix} \frac{1}{1+r_0}C & O \\ O & \frac{r_0}{1+r_0}C \end{pmatrix},$$

dependen de una cierta razón de paso acotada. La matriz O que aparece en la definición de las matrices $A(r_0)$ y $C(r_0)$ denota la matriz nula de la misma dimensión que las matrices A y C que definen el método original.

En consecuencia, si procedemos como en (IV.5), obtenemos respectivamente la siguiente expresión para las etapas internas y la solución de avance del método

$$\begin{aligned} \tilde{K}_i &= R_i^{r_0}(hJ)y_0 + \\ &\quad h\sum_{j=1}^{2s}\varphi_{ij}^{r_0}(hJ)\left(N(y_0+\sum_{l=1}^{2s}a_{jl}(r_0)\tilde{K}_l) + 2M(y_0, \sum_{l=1}^{2s}c_{jl}(r_0)\tilde{K}_l)\right), \quad 1 \le i \le 2s, \end{aligned} \tag{IV.20}$$

$$\begin{aligned} y_2 &= R^{r_0}(hJ)y_0 + \\ &\quad h\sum_{i=1}^{2s}\psi_i^{r_0}(hJ)\left(N(y_0+\sum_{j=1}^{2s}a_{ij}(r_0)\tilde{K}_j) + 2M(y_0, \sum_{j=1}^{2s}c_{ij}(r_0)\tilde{K}_j)\right), \end{aligned} \tag{IV.21}$$

habiendo considerado arriba las siguientes funciones racionales

$$\begin{aligned} R_i^{r_0}(\mu) &:= \mu e_i^T(I-\mu D(r_0))^{-1}\tilde{e}, \\ \varphi_{ij}^{r_0}(\mu) &:= e_i^T(I-\mu D(r_0))^{-1}e_j, \\ \psi_i^{r_0}(\mu) &:= b(r_0)^T(I-\mu D(r_0))^{-1}e_i, \\ R^{r_0}(\mu) &:= 1+\mu b(r_0)^T(I-\mu D(r_0))^{-1}\tilde{e}, \end{aligned}$$

siendo $D(r_0) := A(r_0) + C(r_0)$ y e_i, $1 \le i \le 2s$, los vectores canónicos de $\mathbb{R}^{2s}$.

Nota IV.3.1 Puesto que $D(r_0)$ es una matriz triangular inferior con elementos diagonales positivos, podemos concluir en virtud de la equivalencia entre (III.48) y (III.49) que para todo $0 < r_0 \leq r^*$ se tiene que

$$\sup_{\text{Re } \mu \leq 0} \left\| (I - \mu D(r_0))^{-1} \right\| < \infty.$$

Además, teniendo en cuenta que la matriz

$$D(0) = \begin{pmatrix} D & O \\ eb^T & O \end{pmatrix}$$

tiene todos sus autovalores positivos, exceptuando al autovalor nulo que posee igual multiplicidad algebraica y geométrica s, obtenemos que

$$\sup_{\text{Re } \mu \leq 0} \left\| (I - \mu D(0))^{-1} \right\| < \infty.$$

En consecuencia, queda asegurada la existencia de una constante positiva p, únicamente dependiente de los coeficientes del método original y de r^*, tal que

$$\sup_{0 < r_0 \leq r^*} \sup_{Re \mu \leq 0} \left\| (I - \mu D(r_0))^{-1} \right\| \leq p.$$

Esta última cota permite asegurar, en virtud de la proposición IV.2.1, la existencia de una constante positiva q tal que si $h \left\| y_0 \right\| \leq q\delta_2^{-1}$ entonces la ecuación de etapas (IV.20) admite solución única verificando

$$\left\| \tilde{K}_i \right\| \leq 2p \left\| y_0 \right\|, \ 1 \leq i \leq 2s,$$

mientras que para la solución de avance (IV.21) se obtiene la cota de estabilidad en función del valor inicial

$$\left\| y_2 \right\| \leq 2p \left\| y_0 \right\|.$$

En otro orden de cosas, es sencillo comprobar que la función de estabilidad lineal $R^{r_0}(z)$ del método compuesto a dos pasos (IV.18)-(IV.19) satisface

$$R^{r_0}(z) = R\left(\frac{z}{1 + r_0} \right) R\left(\frac{r_0 z}{1 + r_0} \right),$$

donde $R(z)$ denota la función de estabilidad lineal del método original. Por lo tanto asumiendo que $R(z)$ satisface *(P)*, obtenemos la siguiente acotación independiente de la razón de paso r_0

$$\sup_{0 < r_0 \leq r^*} \sup_{Re \mu \leq x} |R^{r_0}(\mu)| \leq \max\{\rho, 1 + \frac{\varrho}{1 + r^*} x\}, \ \forall x \leq 0.$$

En virtud del teorema IV.2.2 queda asegurada la existencia de constantes $\gamma, \kappa, \alpha, \beta$, dependientes únicamente de las constantes involucradas en las H-hipótesis, de los coeficientes del método de Rosenbrock original (IV.2)-(IV.3) y de la constante prefijada r^*, que garantizan la E-estabilidad del método de Rosenbrock compuesto (IV.18)-(IV.19).

En resumen, hemos demostrado el siguiente

Teorema IV.3.2 *Si el método semi-implícito (IV.2)-(IV.3) satisface (P), entonces el método compuesto dado por (IV.18)-(IV.19) es E-estable.*

□

IV.4. Experimentos numéricos

En las dos secciones previas hemos mostrado que aquellos métodos de tipo Rosenbrock, incluidos aquellos métodos modificados de modo que las matrices Jacobianas se mantienen constantes un determinado número de pasos, que verifican la propiedad de A-estabilidad fuerte (o, equivalentemente, la propiedad (P) que aparece en la sección segunda de este capítulo) proporcionan integraciones estables sobre la clase de problemas bajo las H-hipótesis.

Presentamos en este punto algunas pruebas de carácter numérico con vistas a complementar los resultados teóricos presentados en las secciones anteriores. De hecho, aunque la necesidad de la suposición *(P)* para la E-estabilidad de los métodos de tipo Rosenbrock no ha sido probada en ninguna de las dos secciones previas, las ilustraciones numéricas que presentamos en esta sección reflejan claramente la influencia del valor de $|R(\infty)|$ cuando se considera la integración de sistemas diferenciales que poseen equilibrios semi-estables a través de métodos de tipo Rosenbrock en amplios intervalos temporales .

De modo análogo al estudio realizado en el capítulo previo dedicado a la estabilidad de las integraciones de los métodos de tipo Runge-Kutta, consideramos aquí la integración numérica en amplios intervalos temporales de los problemas (II.10) y (II.12) por medio de una familia uniparamétrica de métodos de tipo Rosenbrock. La solución exacta de estos problemas tiende al punto de equilibrio $x = 0$ en tiempo infinito, mientras que cualquier autovector asociado al autovalor nulo de la matriz Jacobiana en el equilibrio $J = f'(0)$ determina, en virtud del teorema de la Variedad Centro, la pendiente de la solución exacta en el equilibrio. De este modo, consideraremos la pendiente numérica que proveen los integradores de tipo Rosenbrock sobre estos problemas como indicador de la estabilidad numérica.

Como integradores numéricos hemos considerado una familia uniparamétrica de métodos de Rosenbrock (IV.2)-(IV.3) de 3 etapas, con orden de consistencia 3 y que satisfacen

$$c_{11} = c_{22} = c_{33} = \gamma,$$

siendo γ el parámetro característico de esta familia de métodos. Estos métodos han sido obtenidos requiriendo que $C_{42},\ C_{43}$ se anulen, donde $C_{42},\ C_{43}$ denotan los coeficientes de error asociados a los árboles etiquetados $\tau_{42} \equiv [\tau_0, [\tau_0]],\ \tau_{43} \equiv [[\tau_0, \tau_0]]$, respectivamente (ver [3, 81, 86] o [53, Cap. IV.7] para obtener detalles adicionales).

$\tau_{42} \equiv$  $\tau_{43} \equiv$ 

Imponiendo además las condiciones simplificadoras

$$b_2 = \frac{1}{2}, \qquad b^T(Ae) = \frac{1}{2}, \qquad e_2^T Ae = e_3^T Ae,$$

donde $e_i^T,\ i = 1, 2, 3$, representan los vectores canónicos de $\mathbb{R}^3$, obtenemos una familia uniparamétrica de métodos de Rosenbrock, con γ como parámetro característico, que son A-estables para los valores $\frac{1}{3} \leq \gamma \leq \gamma^* \simeq 1.06857902..$, siendo γ^* la mayor raíz real de la ecuación $\frac{1}{24} - \frac{1}{2}\gamma + \frac{3}{2}\gamma^2 - \gamma^3 = 0$.

En efecto, para estos métodos la función de estabilidad lineal viene dada por

$$R(z, \gamma) = \frac{1 + (1 - 3\gamma)z + (\frac{1}{2} - 3\gamma + 3\gamma^2)z^2 + (\frac{1}{6} - \frac{3}{2}\gamma + 3\gamma^2 - \gamma^3)z^3}{(1 - \gamma z)^3},$$

que toma valor en el infinito

$$R(\infty, \gamma) = \frac{(-\frac{1}{6} + \frac{3}{2}\gamma - 3\gamma^2 + \gamma^3)}{\gamma^3}.$$

Por lo tanto, para $\gamma > 0$, la A-estabilidad de los métodos es equivalente a la condición $E(y) \geq 0$, siendo $E(y)$ el polinomio asociado a la función racional R definido por (III.27). En nuestro caso tenemos que

$$E(y) = p_1(\gamma)y^4 + p_2(\gamma)y^6,$$

donde

$$\begin{aligned} p_1(\gamma) &:= \tfrac{1}{12} - \gamma + 3\gamma^2 - 2\gamma^3, \\ p_2(\gamma) &:= -\tfrac{1}{36} + \tfrac{1}{2}\gamma - \tfrac{13}{4}\gamma^2 + \tfrac{28}{3}\gamma^3 - 12\gamma^4 + 6\gamma^5. \end{aligned}$$

De este modo, la A-estabilidad de los métodos es equivalente a que

$$\begin{cases} p_1(\gamma) \geq 0, \\ \qquad \text{y} \\ p_2(\gamma) \geq 0, \end{cases}$$

esto es, $\gamma \in [\frac{1}{3}, \gamma^*]$. Todos los valores del parámetro en tal intervalo proveen métodos fuertemente A-estables, exceptuando el valor $\gamma = \frac{1}{3}$.

En las tablas IV.1-IV.9 denotaremos por $Rosen3(\gamma)$ a la familia de métodos así obtenida. Teniendo las anteriores consideraciones en cuenta, no es difícil mostrar que la norma l_2 de los coeficientes de error que definen el término principal del error para esta familia de métodos viene dada por

$$l_2EC(\gamma) = \sqrt{C_{41}^2 + C_{44}^2} = \sqrt{\frac{1}{81} + (1 - 12\gamma + 36\gamma^2 - 24\gamma^3)^2},$$

donde C_{41}, C_{44} denotan los coeficientes de error asociados a los árboles etiquetados de cuarto orden $\tau_{41} \equiv [\tau_0, \tau_0, \tau_0]$ y $\tau_{44} \equiv [[[\tau_0]]]$, respectivamente.

$\tau_{41} \equiv$  $\tau_{44} \equiv$ 

Con vistas a reducir el número de evaluaciones de funciones Jacobianas y de descomposiciones LU de las mismas, estos métodos han sido implementados actualizando las matrices Jacobianas después de dos pasos temporales del mismo tamaño. Un sencillo cálculo permite ver que la condición simplificadora $b^T(Ae) = 1/2$ introducida arriba garantiza que el orden de consistencia 3 se mantiene en cada paso de la integración cuando se fija la matriz Jacobiana cada dos pasos.

Así, tras dos pasos consecutivos de la misma amplitud, el error local del método es estimado por extrapolación local (II.35) dando un nuevo paso de tamaño doble, y el tamaño de paso se controla de modo estándar (II.36) en caso de aceptación por el estimador. Sin embargo, en caso de rechazo, el tamaño de paso se reduce a la mitad, para ahorrar el cálculo de la solución con paso doble en el siguiente intento de aceptación del paso. La tolerancia de error en el punto temporal t_{2n} se calcula a partir de la fórmula

$$Tol_{2n} = Rtol \cdot \max\{\|y_{2n}\|_\infty, Atol\},\ n \geq 1,$$

donde $Rtol$ y $Rtol \cdot Atol$ representan las tolerancias de error relativo y absoluto, respectivamente. En los experimentos numéricos que presentamos en esta sección hemos considerado $Rtol = TOL$ como parámetro variable, mientras que $Atol$ se ha elegido constante y dependiente del problema

que se integra, de tal modo que para el problema (II.10) hemos tomado $ATOL = 10^{-1}$, y $ATOL = 10^{-5}$, para el problema (II.12).

En las tablas IV.2-IV.9 hemos dispuesto los resultados obtenidos en la integración de los problemas (II.10) y (II.12) por medio de la familia de métodos $rosen3(\gamma)$ para algunos valores representativos del parámetro γ. En estas tablas, NPA y $NPRE$ denotan el número de pasos aceptados y rechazados por el estimador de error local, respectivamente; *Rmedio* representa un promedio para la razón de paso; $Y_i(T)$ proporciona la $i-$ésima componente de la solución numérica en el punto temporal T; T_f indica que el punto final de la integración fue alcanzado sin problemas de estabilidad; mientras que T_{inest} indica que el código fue abortado debido a que la solución numérica cruza a la región de inestabilidad del problema. Debemos tener en cuenta que si el código requiere la aceptación de NPA y el rechazo de $NPRE$ pasos temporales, respectivamente, entonces el número de sistemas lineales ($LinSist$) que el código debe resolver, el número de descomposiciones LU (Nlu) que debe llevar a cabo, y el número de derivadas (Der) y matrices Jacobianas (Jac) que debe evaluar vienen dados por las fórmulas

$$LinSist = Der = \frac{9}{2}NPA + 3NPRE, \qquad Nlu = NPA + \frac{1}{2}NPRE, \qquad Jac = \frac{1}{2}NPA.$$

Debemos observar en las tablas IV.2-IV.9 que el punto final de la integración de cada uno de los problemas fue alcanzado para todos los valores de γ que proveen métodos fuertemente A-estables. Sin embargo, la integración no fue del todo satisfactoria a medida que el parámetro γ se toma próximo a 1/3; en otras palabras, la integración se vuelve ineficiente a medida que $|R(\infty)|$ tiende a 1. Más aún, en estos casos, la integración llega a ser inestable en el sentido de que alguna de las componentes de la solución llega a ser negativa y la solución numérica cruza a la región de inestabilidad del problema. Por este motivo la integración es abortada si alguna de las componentes de la solución numérica toma valores negativos. En el caso de integraciones satisfactorias, vemos que el punto final fue alcanzado en pocos pasos y con un número razonable de pasos rechazados por el estimador de error. En todos los casos el tamaño inicial de paso considerado fue $h_0 = 10^{-6}$.

Cada una de las tablas, exceptuando las tablas IV.5 y IV.9, se acompañan de los valores de la pendiente numérica que proveen los métodos de la familia $rosen3$ al integrar cada uno de los problemas. Vemos que aquellos métodos con $|R(\infty)|$ suficientemente menor 1 son capaces de reproducir con una precisión bastante aceptable el valor de pendiente de la solución exacta en el equilibrio. Sin embargo, a medida que el parámetro tiende a 1/3 el método correspondiente

Tabla IV.1: $Rosen3(\gamma)$. Parámetros asociados.

γ	$R(\infty,\gamma)$	$l_2EC(\gamma)$
0.435866521...	0	0.631383603
0.35	0.786200194	0.212383330
0.34	0.911798629	0.177408216
0.336	0.964287964	0.164912661
$\frac{1}{3}$	1	0.157134840

provee soluciones numéricas cuya pendiente pierde exactitud con respecto a la pendiente teórica. Esto refleja claramente los problemas de estabilidad en entornos de los equilibrios que surgen cuando el integrador no presenta la propiedad de A-estabilidad fuerte.

Por otro lado, también debemos hacer énfasis en que el método L-estable ($\gamma = \overline{\gamma} \simeq 0.435866521...$, donde $\overline{\gamma}$ satisface $-\frac{1}{6} + \frac{3}{2}\overline{\gamma} - 3\overline{\gamma}^2 + \overline{\gamma}^3 = 0$) no proporciona mejores resultados que aquellos métodos $Rosen3(\gamma)$ para los que $|R(\infty)|$ es moderadamente menor que 1. De esta manera, no parece que el tamaño de $|R(\infty)|$ sea significativo de cara a una integración estable cuando $|R(\infty)|$ es significativamente menor que 1. En tales casos, vemos que aquellos métodos con un menor valor de $l_2EC(\gamma)$ proveen una integración eficiente requiriendo una menor cantidad de pasos para concluir la integración (compárense los resultados en las tablas IV.2-IV.3 y IV.6-IV.7).

Tabla IV.2: Ejemplo 1 (II.10). *Rosen*3, $\gamma \simeq 0.435866521$.

TOL	NPA	$NPRE$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-1}	240	0	1.490	$0.10421279 \cdot 10^{-2}$	$0.41728564 \cdot 10^{-4}$	$0.40041691 \cdot 10^{-1}$
10^{-2}	298	0	1.384	$0.26264748 \cdot 10^{-3}$	$0.10508657 \cdot 10^{-4}$	$0.40010499 \cdot 10^{-1}$
10^{-3}	506	0	1.223	$0.33663718 \cdot 10^{-4}$	$0.13465940 \cdot 10^{-5}$	$0.40001345 \cdot 10^{-1}$
10^{-4}	1126	0	1.098	$0.40119648 \cdot 10^{-5}$	$0.16047923 \cdot 10^{-6}$	$0.40000160 \cdot 10^{-1}$
10^{-5}	2692	0	1.041	$0.49012698 \cdot 10^{-6}$	$0.19605089 \cdot 10^{-7}$	$0.40000020 \cdot 10^{-1}$
10^{-6}	6752	0	1.016	$0.57127383 \cdot 10^{-7}$	$0.22850954 \cdot 10^{-8}$	$0.40000002 \cdot 10^{-1}$
10^{-7}	33908	0	1.003	$0.67668839 \cdot 10^{-8}$	$0.27067536 \cdot 10^{-9}$	$0.40000000 \cdot 10^{-1}$

Tabla IV.3: Ejemplo 1 (II.10). *Rosen*3, $\gamma = 0.35$.

TOL	NPA	$NPRE$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-1}	240	0	1.491	$0.49581279 \cdot 10^{-3}$	$0.19842306 \cdot 10^{-4}$	$0.40019755 \cdot 10^{-1}$
10^{-2}	296	0	1.387	$0.11419142 \cdot 10^{-3}$	$0.45681762 \cdot 10^{-5}$	$0.40004548 \cdot 10^{-1}$
10^{-3}	490	0	1.231	$0.15066284 \cdot 10^{-4}$	$0.60266039 \cdot 10^{-6}$	$0.40000600 \cdot 10^{-1}$
10^{-4}	1058	0	1.105	$0.18537368 \cdot 10^{-5}$	$0.74149610 \cdot 10^{-7}$	$0.40000074 \cdot 10^{-1}$
10^{-5}	2486	0	1.044	$0.23153366 \cdot 10^{-6}$	$0.92613486 \cdot 10^{-8}$	$0.40000009 \cdot 10^{-1}$
10^{-6}	6990	0	1.015	$0.28565507 \cdot 10^{-7}$	$0.11426203 \cdot 10^{-8}$	$0.40000001 \cdot 10^{-1}$
10^{-7}	41820	0	1.003	$0.35670807 \cdot 10^{-8}$	$0.14268323 \cdot 10^{-9}$	$0.40000000 \cdot 10^{-1}$

Tabla IV.4: Ejemplo 1 (II.10). *Rosen*3, $\gamma = 0.336$.

TOL	NPA	$NPRE$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-1}	240	0	1.492	$0.32684585 \cdot 10^{-3}$	$0.12372014 \cdot 10^{-4}$	$0.37852749 \cdot 10^{-1}$
10^{-2}	296	0	1.386	$0.62289681 \cdot 10^{-4}$	$0.23343344 \cdot 10^{-5}$	$0.37475460 \cdot 10^{-1}$
10^{-3}	486	0	1.232	$0.12893622 \cdot 10^{-4}$	$0.51235629 \cdot 10^{-6}$	$0.39737190 \cdot 10^{-1}$
10^{-4}	1048	0	1.106	$0.16792062 \cdot 10^{-5}$	$0.67139203 \cdot 10^{-7}$	$0.39982703 \cdot 10^{-1}$
10^{-5}	2460	0	1.044	$0.20883661 \cdot 10^{-6}$	$0.83531064 \cdot 10^{-8}$	$0.39998285 \cdot 10^{-1}$
10^{-6}	6928	84	1.021	$0.26370975 \cdot 10^{-7}$	$0.10548343 \cdot 10^{-8}$	$0.39999824 \cdot 10^{-1}$
10^{-7}	30836	10914	1.002	$0.33176460 \cdot 10^{-8}$	$0.13270578 \cdot 10^{-9}$	$0.39999982 \cdot 10^{-1}$

Tabla IV.5: Ejemplo 1 (II.10). *Rosen*3, $\gamma = 1/3$.

TOL	NPA	$NPRE$	$Rmedio$	$Y_1(T_{inest})$	$Y_2(T_{inest})$	T_{inest}
10^{-1}	112	0	1.482	$0.85534559 \cdot 10^{-2}$	$-0.17631518 \cdot 10^{-3}$	$0.51 \cdot 10^{11}$
10^{-2}	116	0	1.290	$0.15583916 \cdot 10^{-1}$	$-0.79418921 \cdot 10^{-4}$	$0.29 \cdot 10^{6}$
10^{-3}	176	0	1.168	$0.37549890 \cdot 10^{-1}$	$-0.23104285 \cdot 10^{-4}$	$0.31 \cdot 10^{5}$
10^{-4}	280	0	1.096	$0.69086310 \cdot 10^{-1}$	$-0.11806355 \cdot 10^{-3}$	$0.12 \cdot 10^{5}$
10^{-5}	1346	24	1.024	$0.62686840 \cdot 10^{-1}$	$-0.47720160 \cdot 10^{-5}$	$0.17 \cdot 10^{5}$
10^{-6}	7820	106	1.010	$0.98299479 \cdot 10^{-3}$	$-0.10597383 \cdot 10^{-5}$	$0.15 \cdot 10^{7}$
10^{-7}	12782	198	1.009	$0.23573702 \cdot 10^{-2}$	$-0.81701884 \cdot 10^{-6}$	$0.65 \cdot 10^{6}$

Tabla IV.6: Ejemplo 2 (II.12). *Rosen*3, $\gamma \simeq 0.435866521$.

TOL	NPA	$NPRE$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-4}	254	0	1.261	$0.17764700 \cdot 10^{-13}$	$0.60424149 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$
10^{-5}	374	0	1.171	$0.14912662 \cdot 10^{-13}$	$0.50723341 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$
10^{-6}	590	0	1.105	$0.17837687 \cdot 10^{-13}$	$0.60672405 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$
10^{-7}	980	0	1.061	$0.16060200 \cdot 10^{-13}$	$0.54626529 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$
10^{-8}	1688	0	1.035	$0.16699702 \cdot 10^{-13}$	$0.56801707 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$
10^{-9}	2998	0	1.019	$0.15144950 \cdot 10^{-13}$	$0.51513436 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$

Tabla IV.7: Ejemplo 2 (II.12). *Rosen*3, $\gamma = 0.35$.

TOL	NPA	$NPRE$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-4}	220	0	1.303	$0.16039233 \cdot 10^{-13}$	$0.54555193 \cdot 10^{-16}$	$0.34013592 \cdot 10^{-2}$
10^{-5}	306	0	1.211	$0.17644479 \cdot 10^{-13}$	$0.60015236 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$
10^{-6}	462	0	1.136	$0.15887955 \cdot 10^{-13}$	$0.54040663 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$
10^{-7}	740	0	1.083	$0.18423643 \cdot 10^{-13}$	$0.62665452 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$
10^{-8}	1242	2	1.048	$0.16386717 \cdot 10^{-13}$	$0.55737133 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$
10^{-9}	2176	50	1.028	$0.16690551 \cdot 10^{-13}$	$0.56770583 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$

Tabla IV.8: Ejemplo 2 (II.12). *Rosen*3, $\gamma = 0.34$.

TOL	NPA	$NPRE$	$Rmedio$	$Y_1(T_f)$	$Y_2(T_f)$	$Y_2(T_f)/Y_1(T_f)$
10^{-4}	212	0	1.314	$0.30062286 \cdot 10^{-10}$	$0.25931415 \cdot 10^{-13}$	$0.86258959 \cdot 10^{-3}$
10^{-5}	292	0	1.224	$0.33023126 \cdot 10^{-13}$	$0.50747738 \cdot 10^{-16}$	$0.15367333 \cdot 10^{-2}$
10^{-6}	438	0	1.144	$0.14526511 \cdot 10^{-13}$	$0.49406638 \cdot 10^{-16}$	$0.34011360 \cdot 10^{-2}$
10^{-7}	694	0	1.089	$0.15260104 \cdot 10^{-13}$	$0.51905090 \cdot 10^{-16}$	$0.34013588 \cdot 10^{-2}$
10^{-8}	1180	50	1.052	$0.17662938 \cdot 10^{-13}$	$0.60078020 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$
10^{-9}	2080	178	1.029	$0.16005291 \cdot 10^{-13}$	$0.54439764 \cdot 10^{-16}$	$0.34013605 \cdot 10^{-2}$

Tabla IV.9: Ejemplo 2 (II.12). *Rosen*3, $\gamma = 1/3$.

TOL	NPA	$NPRE$	$Rmedio$	$Y_1(T_{inest})$	$Y_2(T_{inest})$	T_{inest}
10^{-4}	182	4	1.199	$0.31997042 \cdot 10^{-4}$	$-0.14187144 \cdot 10^{-8}$	$0.10 \cdot 10^{8}$
10^{-5}	420	6	1.085	$0.12150190 \cdot 10^{-4}$	$-0.29079735 \cdot 10^{-11}$	$0.26 \cdot 10^{8}$
10^{-6}	756	4	1.129	$0.66387115 \cdot 10^{-5}$	$-0.84510477 \cdot 10^{-11}$	$0.57 \cdot 10^{9}$
10^{-7}	1164	2	1.032	$0.15308684 \cdot 10^{-6}$	$-0.23467970 \cdot 10^{-12}$	$0.21 \cdot 10^{10}$
10^{-8}	1504	18	1.027	$0.96496615 \cdot 10^{-8}$	$-0.92143576 \cdot 10^{-13}$	$0.33 \cdot 10^{11}$
10^{-9}	2196	178	1.022	$0.57715771 \cdot 10^{-9}$	$-0.26261478 \cdot 10^{-14}$	$0.56 \cdot 10^{12}$

Capítulo V

Sobre la contractividad y convergencia de los Métodos Lineales Generales.

V.1. Consideraciones preliminares

Consideremos sistemas diferenciales complejos

$$y' = f(t,y), \quad y(0) = y_0, \qquad f : [0,\infty) \times \mathbb{C}^m \to \mathbb{C}^m, \tag{V.1}$$

siendo f una función continuamente diferenciable en un dominio apropiado y además satisface localmente una condición de Lipschitz lateral alrededor de la solución exacta con respecto a un determinado producto interior en $\mathbb{C}^m$, esto es,

$$\mathrm{Re}\langle f(t,y) - f(t,z), y - z\rangle_X \leq \nu \left\|y - z\right\|_X^2, \quad \forall t \in [0,\infty),\ \forall y, z, \tag{V.2}$$

donde ν es una constante. El producto interior viene definido por

$$\langle u, v\rangle_X := \sum_{i,j=1}^{m} x_{ij}\overline{v}_j u_i; \quad u = (u_j)_{j=1}^m,\ v = (v_j)_{j=1}^m, \tag{V.3}$$

donde la matriz $X = (x_{ij})_{i,j=1}^m \in \mathbb{R}^{m,m}$ es simétrica y definida positiva, y $\bar{z}$ denota el conjugado del número complejo z. En el resto del capítulo, usaremos la notación $\left\|\cdot\right\|_X$ para denotar la norma asociada al producto interior.

Es bien conocido que, bajo las suposiciones anteriores, la diferencia entre dos soluciones del sistema (V.1) satisface para valores iniciales arbitrarios y_0, z_0 (véase, por ejemplo, [33, Cap. I])

$$\left\|y(t;0,y_0) - y(t;0,z_0)\right\|_X \leq \exp(\nu t)\left\|y_0 - z_0\right\|_X, \quad \forall t \geq 0.$$

De esta manera, cuando la constante ν es negativa (caso de soluciones estrictamente disipativas), además de contractividad estricta para las soluciones, tenemos estabilidad asintótica (en sentido de Lyapunov) para cualquier solución definida en el intervalo $[0, +\infty)$. Sería deseable entonces que en tal situación los métodos numéricos aplicados a (V.1) preservaran la propiedad de contractividad estricta en intervalos finitos, así como la estabilidad asintótica en intervalos semi-infinitos.

Nos proponemos como principal objetivo en este capítulo estudiar la contractividad estricta y la convergencia (B-convergencia) de los *Métodos Lineales Generales* en intervalos semi-infinitos para la clase de sistemas diferenciales estrictamente disipativos, esto es, $\nu < 0$ en (V.2). Debemos señalar que algunos de los resultados de B-convergencia que presentamos en este capítulo son similares a los que derivan Huang et al. [57] para el caso de intervalos temporales finitos y problemas disipativos ($\nu = 0$ en (V.2)). Sin embargo, nuestros resultados también se aplican a intervalos semi-infinitos mientras que aquellos en [57] no. El análisis de contractividad estricta que llevaremos a cabo en este capítulo extiende las ideas dadas por Hairer y Zennaro para métodos Runge-Kutta implícitos [54]. De esta manera, explotando el carácter *super-exponencial* (véase, por ejemplo, [54, Definición 4.1]) de la función de estabilidad lineal asociada a los Métodos Lineales Generales, extenderemos muchos de los resultados que se presentan en [54] para el caso de métodos de tipo Runge-Kutta al caso de Métodos Lineales Generales y, en particular, al caso de Métodos Runge-Kutta Multipaso. Por otro lado, los resultados de convergencia, que están parcialmente inspirados en el trabajo de Hundsdorfer [58], se posponen a la penúltima sección del capítulo. Daremos además aplicaciones de estos resultados de contractividad y convergencia a algunas clases de Métodos Lineales Generales que aparecen en la literatura. Debemos añadir que los resultados que se expondrán en este capítulo pueden ser encontrados en el trabajo [41].

El resto del capítulo queda organizado del modo siguiente. En la sección segunda de este capítulo establecemos algunas ideas preliminares acerca de la clase de Métodos Lineales Generales y de las normas matriciales a considerar, así como diversas herramientas para probar los principales resultados que se dan en este capítulo. En la sección 3 llevamos a cabo el análisis de contractividad estricta; mientras que en la sección 4 desarrollamos los estudios de convergencia. Finalmente, dedicamos la sección 5 de este capítulo a establecer aplicaciones numéricas de los resultados principales.

V.2. Resultados previos

Un Método Lineal General de k pasos y s etapas aplicado a (V.1) viene dado por

$$\begin{aligned} V^{(n)} &= h(\tilde{A}\otimes I_m)F(t_n,h,V^{(n)}) + (\tilde{B}\otimes I_m)Y^{(n)}, \\ Y^{(n+1)} &= h(A\otimes I_m)F(t_n,h,V^{(n)}) + (B\otimes I_m)Y^{(n)}, \qquad n\geq 0, \end{aligned} \tag{V.4}$$

donde $h>0$ denota el tamaño de paso, $V^{(n)} := ({v_1^{(n)}}^T, {v_2^{(n)}}^T, \ldots, {v_s^{(n)}}^T)^T \in \mathbb{C}^{ms}$ denota el super-vector de etapas del método, $Y^{(n+1)} := ({y_1^{(n+1)}}^T, {y_2^{(n+1)}}^T, \ldots, {y_k^{(n+1)}}^T)^T \in \mathbb{C}^{mk}$ es la solución de avance, $F(t_n,h,V^{(n)}) := (f(t_n+c_1h,v_1^{(n)})^T, f(t_n+c_2h,v_2^{(n)})^T, \ldots, f(t_n+c_sh,v_s^{(n)})^T)^T$, e $Y^{(n)}$ denota una determinada información acerca de la solución exacta que se conoce desde el paso previo. El método posee como parámetros libres las matrices $\tilde{A}\in\mathbb{R}^{s,s}$, $\tilde{B}\in\mathbb{R}^{s,k}$, $A\in\mathbb{R}^{k,s}$, $B\in\mathbb{R}^{k,k}$, y el vector $c := (c_1,\ldots,c_s)^T \in \mathbb{R}^s$, y éstos pueden ser elegidos convenientemente con vistas a ganar estabilidad y orden de error global y orden de etapa. Representamos por $\otimes$ el producto de Kronecker de matrices, $A\otimes B = (a_{ij}B)$, e I_m denota la matriz identidad de dimensión m.

La existencia de solución para la ecuación de las etapas de un Método Lineal General se deduce bajo las mismas ideas que en el caso de las etapas de los métodos de tipo Runge-Kutta implícitos, véase, por ejemplo, [28], [33, Cap. 5] y [53, Cap. IV.14]. De este modo el siguiente resultado puede ser probado siguiendo las ideas presentadas en los trabajos anteriormente citados. En particular, siguiendo las pruebas de los teoremas 14.2 y 14.3 en [53, Cap. IV.14] se tiene el siguiente

Teorema V.2.1 *Para la clase de sistemas diferenciales (V.1)-(V.2), la ecuación de etapas asociada a un Método Lineal General (V.4) con matriz $\tilde{A}$ no singular posee solución única si se verifica $\nu h < \alpha_0(\tilde{A}^{-1})$.*

Nota V.2.2 Debemos recordar que para una matriz $C\in\mathbb{R}^{s,s}$ y una matriz diagonal definida positiva $D\in\mathbb{R}^{s,s}$ $(D>0)$ se define

$$\alpha_D(C) := \inf_{u\neq 0} \frac{\langle Cu,u\rangle_D}{\langle u,u\rangle_D} = \frac{1}{2}\lambda_{min}(D^{1/2}CD^{-1/2} + D^{-1/2}C^TD^{1/2}),$$

mientras que

$$\alpha_0(C) := \sup_{D>0} \alpha_D(C),$$

siendo $\langle u,v\rangle_D := v^TDu$.

Nota V.2.3 La condición $\nu h < \alpha_0(\tilde{A}^{-1})$ que aparece en el teorema previo es *esencialmente óptima* tal como ha sido indicado por Kraaijevanger y Schneid [63, Teorema 2.12] para el caso particular de los métodos de tipo Runge-Kutta implícitos.

En otro orden de cosas, el análisis de estabilidad lineal de los Métodos Lineales Generales requiere considerar el problema test simple $y' = \lambda y$. En tal caso, la solución de avance satisface

$$Y^{(n+1)} = R(z)Y^{(n)}, \quad z = \lambda h, \quad n \geq 0,$$

donde $R(z)$ es la denominada *matriz de estabilidad* del método

$$R(z) := B + zA(I - z\tilde{A})^{-1}\tilde{B}. \tag{V.5}$$

Entonces, el método (V.4) se dice A-estable si los autovalores de $R(z)$, para cada $z \in \mathbb{C}^- := \{z \in \mathbb{C}, \operatorname{Re} z \leq 0\}$, satisfacen la *condición de las raíces*, esto es, los autovalores de $R(z)$ poseen módulo menor estricto que uno, o bien si poseen módulo igual a uno entonces son autovalores simples.

Por otro lado, las propiedades de estabilidad no lineal de los métodos se estudian en sistemas no lineales disipativos, esto, es problemas (V.1) satisfaciendo (V.2), con $\nu = 0$. Así, un Método Lineal General (V.4) se dice G-estable si existe una matriz simétrica y definida positiva $G = (g_{ij})_{i,j=1}^{k} \in \mathbb{R}^{k,k}$ tal que para dos soluciones de avance del método arbitrarias se tiene que

$$\left\| \hat{Y}^{(n+1)} - Y^{(n+1)} \right\|_{G\otimes X} \leq \left\| \hat{Y}^{(n)} - Y^{(n)} \right\|_{G\otimes X},$$

donde

$$\left\| Y^{(n)} \right\|_{G\otimes X} := \sum_{i,j=1}^{k} g_{ij} \langle Y_i^{(n)}, Y_j^{(n)} \rangle_X.$$

Es un hecho bien conocido que la A-estabilidad es una condición necesaria para la G-estabilidad (véase, por ejemplo, [53, Cap. V.9.]). Notemos que la diferencia $\Delta Y^{(n+1)} := Y^{(n+1)} - \hat{Y}^{(n+1)}$ entre dos soluciones numéricas del método (V.4) al ser aplicado a (V.1) con paso fijo satisface la relación recurrente

$$\Delta Y^{(n+1)} = M(Z^{(n)})\Delta Y^{(n)}, \quad n \geq 0, \quad Z^{(n)} = hJ^{(n)}, \tag{V.6}$$

con

$$J^{(n+1)} := \operatorname{BlockDiag}(J_1, \ldots, J_s) \in \mathbb{C}^{ms,ms}, \tag{V.7}$$

y donde $J_i \in \mathbb{C}^{m,m}$, $1 \leq i \leq s$, son matrices definidas por

$$J_i^{(n)} := \int_0^1 \frac{\partial f}{\partial y}(t_n + c_i h, v_i^{(n)} + \theta \Delta v_i^{(n)}) d\theta. \tag{V.8}$$

Además, $M(Z) \in \mathbb{C}^{mk,mk}$ es la super-matriz dada por

$$M(Z) = B \otimes I_m + (A \otimes I_m) Z (I_{sm} - (\tilde{A} \otimes I_m) Z)^{-1} (\tilde{B} \otimes I_m). \tag{V.9}$$

Burrage y Butcher en [8] han probado que la *estabilidad algebraica* de los métodos es suficiente para obtener G-estabilidad. Recordamos que un Método Lineal General se dice *algebraicamente estable* si existe una matriz diagonal definida no negativa $D = Diag(d_1, \ldots, d_s) \in \mathbb{R}^{s,s}$ y una matriz definida positiva $G \in \mathbb{R}^{k,k}$ tal que

$$N = \begin{pmatrix} G - B^T G B & \tilde{B}^T D - B^T G A \\ D\tilde{B} - A^T G B & D\tilde{A} + \tilde{A}^T D - A^T G A \end{pmatrix} \in \mathbb{R}^{k+s,k+s} \tag{V.10}$$

es definida no negativa. Además, ambos conceptos de G-estabilidad y estabilidad algebraica resultan ser equivalentes para la mayoría de Métodos Lineales Generales de interés. Más concretamente, la G-estabilidad es equivalente a la estabilidad algebraica [11, 12] (véase también [53, 357-359]) para aquellos Métodos Lineales Generales *no* confluentes (esto es, $c_i \neq c_j$, $\forall i \neq j$) que son preconsistentes, esto es, existe un vector $\xi_0 \in \mathbb{R}^k$ tal que

$$B\xi_0 = \xi_0 \qquad \tilde{B}\xi_0 = e := (1, \ldots, 1)^T \in \mathbb{R}^s. \tag{V.11}$$

Con todo lo anterior, podemos establecer un resultado de existencia de solución para la ecuación de etapas asociada a un Método Lineal General algebraicamente estable e irreducible aplicado a sistemas diferenciales estrictamente disipativos. Notemos previamente que un Método Lineal General es irreducible si no posee etapas redundantes (véase, por ejemplo, [57, p. 24] para una definición precisa de irreducibilidad).

Teorema V.2.4 *La ecuación de etapas asociada a un Método Lineal General irreducible, preconsistente y algebraicamente estable, con matriz de coeficientes $\tilde{A}$ no singular, posee solución única, cuando el método se aplica a sistemas diferenciales estrictamente disipativos ($\nu < 0$ en (V.2)).*

Demostración. Debido a la estabilidad algebraica del método, la matriz N en (V.10) es definida no negativa ($N \geq 0$). Además, la matriz diagonal D que aparece en la definición de N es definida positiva [58, Lema 4.1]. Por tanto, ya que $N \geq 0$, sigue que la matriz $D\tilde{A}+\tilde{A}^T D \geq 0$, lo cual implica que $D\tilde{A}^{-1} + \tilde{A}^{-T} D \geq 0$. Por consiguiente, $\alpha_D(\tilde{A}^{-1}) \geq 0$. La prueba concluye en virtud del teorema V.2.1. □

Nota V.2.5 Para los casos en los que la matriz $\tilde{A}$ es singular, se debe llevar a cabo un estudio particular más cuidadoso. Así, en casos donde alguna etapa sea explícita, los resultados anteriores pueden ser aún aplicados tras eliminar la correspondiente fila y columna de la matriz $\tilde{A}$. Este es, por ejemplo, el caso de los métodos Runge-Kutta Lobatto IIIA.

A continuación presentamos algunos resultados preliminares básicos en los que nos apoyaremos de cara a obtener nuevos resultados de contractividad en la sección siguiente. En lo referente a las normas matriciales que consideraremos, recordemos que dada una norma vectorial $\|\cdot\|$ en $\mathbb{C}^m$, se define la norma matricial inducida por $\|J\| := \sup_{v\neq 0} \|Jv\| / \|v\|$, $J \in \mathbb{C}^{m,m}$, mientras que la norma logarítmica asociada se define de modo usual como ([33, p. 27])

$$\mu[J] := \lim_{\epsilon \to 0^+} \frac{\|I_m + \epsilon J\| - 1}{\epsilon}.$$

Debemos notar que para las matrices $J_i^{(n)}$ dadas en (V.8) se tiene que

$$\mu_X[J_i^{(n)}] \leq \nu, \qquad 1 \leq i \leq s,$$

en virtud de (V.2).
Frecuentemente consideraremos normas asociadas al producto interior en (V.3). Para este caso, considerando la descomposición $X = Y^T Y$, donde $Y \in \mathbb{R}^{m,m}$ es una matriz regular (la descomposición no es única), tenemos que

$$\|J\|_X := \left\|YJY^{-1}\right\|_2, \quad \mu_X[J] = \mu_2[YJY^{-1}], \quad \forall J \in \mathbb{C}^{m,m}.$$

Haremos uso además de normas en el superespacio $\mathbb{C}^{mk}$ para algún $k \in \mathbb{N}$. De esta manera, dada una matriz simétrica y definida positiva $G = (g_{ij})_{i,j=1}^k \in \mathbb{R}^{k,k}$, podemos extender el producto interior en (V.3) a $\mathbb{C}^{mk}$ de modo estándar

$$\langle u, v\rangle_{G\otimes X} := \sum_{i,j=1}^{k} g_{ij}\langle u_i, v_j\rangle_X, \;\; u = (u_j)_{j=1}^k, \; v = (v_j)_{j=1}^k, \; u_j, v_j \in \mathbb{C}^m, \; 1 \leq j \leq k. \qquad \text{(V.12)}$$

Descomponiendo la matriz $G = L^T L$, con $L \in \mathbb{R}^{k,k}$ regular (la descomposición no es única), no es complicado mostrar que

$$\langle u, v\rangle_{G\otimes X} = \langle (L\otimes Y)u, (L\otimes Y)v\rangle_2. \tag{V.13}$$

De aquí que para la norma asociada al producto interior en (V.12) y para cualquier $K \in \mathbb{C}^{mk,mk}$ tengamos que

$$\|K\|_{G\otimes X} := \sup_{v\neq 0} \|Kv\|_{G\otimes X} / \|v\|_{G\otimes X} = \max_{\substack{\|u\|_{G\otimes X}=1\\ \|v\|_{G\otimes X}=1}} |u^*(G\otimes X)Kv|. \tag{V.14}$$

Tengamos en cuenta que para cada norma l_p con $p \geq 1$ y $q^{-1} + p^{-1} = 1$, se tiene que

$$\|v\|_p = \max_{\|u\|_q=1} |u^* v|,\ \forall v \in \mathbb{C}^n, \quad \|J\|_p = \max_{\substack{\|u\|_q=1\\ \|v\|_p=1}} |u^* Jv|,\ \forall J \in \mathbb{C}^{n,n},\ n \in \mathbb{N}. \tag{V.15}$$

Esta última propiedad (V.15) se deduce a partir de la dualidad entre los espacios l_p y l_q, con $p \geq 1$ y $q^{-1} + p^{-1} = 1$ (véase, por ejemplo, [62, Sec. 19]). Así, para $p = 2$ en (V.15), se deduce (V.14) como consecuencia de (V.13). Aquí denotamos u^* el vector traspuesto conjugado del vector u. Con estos preliminares y denotando respectivamente por $\mu_X[\cdot]$ y $\mu_p[\cdot]$ las normas logarítmicas asociadas a las normas $\|\cdot\|_X$ y $\|\cdot\|_p$ tenemos el siguiente

Teorema V.2.6 *Sea $M(Z_1, \ldots, Z_s) := (m_{ij}(Z_1, \ldots, Z_s))_{i,j=1}^k \in \mathbb{C}^{mk,mk}$ una aplicación actuando sobre el conjunto de matrices, $\{(Z_1, \ldots, Z_s),\ Z_j \in \mathbb{C}^{m,m},\ j = 1, \ldots, s\}$, tal que $m_{ij}(Z_1 + zI_m, \ldots, Z_s + zI_m),\ 1 \leq i, j \leq k$, es una matriz cuadrada de dimensión m con componentes analíticas en el semiplano complejo negativo Re $z \leq 0$, siempre que las matrices Z_j verifican respectivamente*
(a) $\mu_X[Z_j] \leq 0,\ (j = 1, \ldots, s)$ o
(b) $\mu_p[Z_j] \leq 0,\ (j = 1, \ldots, s)$ (para alguna norma l_p, $p \geq 1$).
Entonces, tenemos respectivamente para cada caso que
(a) la función real

$$\varphi_{G,M}(x) := \sup_{\substack{\mu_X[Z_j]\leq x\\ 1\leq j\leq s}} \|M(Z_1, \ldots, Z_s)\|_{G\otimes X} \tag{V.16}$$

es no decreciente para $x \leq 0$ y satisface la siguiente propiedad

$$\varphi(x)\varphi(y) \leq \varphi(0)\varphi(x + y), \quad \text{para todos } x, y \text{ no positivos}, \tag{V.17}$$

siempre que $G \in \mathbb{R}^{k,k}$ sea simétrica y definida positiva;
(b) la función real

$$\varphi_{p,M}(x) := \sup_{\substack{\mu_X[Z_j]\leq x\\ 1\leq j\leq s}} \|M(Z_1, \ldots, Z_s)\|_p$$

es no decreciente para $x \leq 0$ y satisface (V.17).

Nota V.2.7 Una función $\varphi:\ I \subseteq \mathbb{R} \to \mathbb{R}$, con $0 \in I$, satisfaciendo $\varphi(0) = 1$ y (V.17) se dice superexponencial (véase, por ejemplo, [54, Definición 4.1]).

Demostración. La prueba se basa en las ideas expuestas en la prueba del teorema [54, Teorema 4.3]. Sin embargo, aquí señalamos algunas modificaciones relacionadas al caso matricial.
(a) Tomemos $x < 0$ e $y < 0$, y asumamos $\varphi_{G,M}(0) < \infty$. Consideremos matrices complejas cuadradas de dimensión m $A_1, \ldots, A_s, B_1, \ldots, B_s$ satisfaciendo

$$\mu_X[A_i] \leq x + y, \quad \mu_X[B_i] \leq 0, \quad 1 \leq i \leq s,$$

y vectores $u_i, v_i \in \mathbb{C}^{mk}$, con $\|u_i\|_{G\otimes X} = \|v_i\|_{G\otimes X} = 1$, $(i = 1, 2)$. Definamos para cada $z \in \mathbb{C}$,

$$S(z) = \Big(u_1^*(G \otimes X)M(A_1 - zI_m, \ldots, A_s - zI_m)v_1\Big)\Big(u_2^*(G \otimes X)M(B_1 + zI_m, \ldots, B_s + zI_m)v_2\Big).$$

Así, teniendo en cuenta que en general la norma logarítmica satisface ([33, p. 31])

$$\mu[J + zI_m] = \mu[J] + Rez, \quad \text{para toda } J \in \mathbb{C}^{m,m},$$

deducimos que $S(z)$ es una función analítica en la banda

$$\mathcal{S}(0, x + y) := \{z \in \mathbb{C}:\ x + y \leq \text{Re } z \leq 0\}.$$

Entonces, según el Principio del Módulo Máximo y (V.14), sigue que

$$|S(z)| \leq \sup_{\substack{Re\ z=0,\\ Re\ z=x+y}} |S(z)| \leq \varphi_{G,M}(0)\varphi_{G,M}(x + y), \quad \forall z \in \mathcal{S}(0, x + y). \tag{V.18}$$

Ahora, tomando $z = y$, y considerando supremos sobre $\|u_2\|_{G\otimes X} = \|v_2\|_{G\otimes X} = 1$ y sobre $\mu_X[B_j] \leq 0,\ (j = 1, \ldots, s)$, sigue a partir de (V.18) que

$$|u_1^* M(A_1 - yI_m, \ldots, A_s - yI_m)v_1|\ \varphi_{G,M}(y) \leq \varphi_{G,M}(0)\varphi_{G,M}(x + y).$$

La prueba concluye tomando supremos sobre $\|u_1\|_{G\otimes X} = \|v_1\|_{G\otimes X} = 1$, y sobre $\mu_X[A_i] \leq x + y,\ (i = 1, \ldots, s)$.

(b) La prueba está basada en las mismas ideas de la prueba del apartado (a), aunque en este caso definimos

$$S(z) := \Big(u_1^* M(A_1 - zI_m, \ldots, A_s - zI_m)v_1\Big)\Big(u_2^* M(B_1 + zI_m, \ldots, B_s + zI_m)v_2\Big),$$

y tomamos $\|u_i\|_q = \|v_i\|_p = 1,\ (i = 1, 2),$ con $q^{-1} + p^{-1} = 1$. □

Corolario V.2.8 *Sea* $M(z_1,\ldots,z_s) := (m_{ij}(z_1,\ldots,z_s))_{i,j=1}^{k} \in \mathbb{C}^{k,k}$, *donde* m_{ij} *son funciones analíticas respecto de cada variable compleja* z_j *para* $Re\ z_j \leq 0$, $(j = 1,\ldots,s)$. *Entonces, las funciones crecientes* $\varphi_{G,M}(x)$ *y* $\varphi_{p,M}(x)$ *dadas respectivamente por*

$$\varphi_{G,M}(x) = \sup_{\substack{Re\ z_j \leq x \\ 1\leq j\leq s}} \|M(z_1,\ldots,z_s)\|_G$$

$$\varphi_{p,M}(x) = \sup_{\substack{Re\ z_j \leq x \\ 1\leq j\leq s}} \|M(z_1,\ldots,z_s)\|_p$$

cumplen (V.17), para toda matriz simétrica y definida positiva $G \in \mathbb{R}^{k,k}$ *y cada norma* l_p *(*$p \geq 1$*).*

□

Nota V.2.9 El enunciado del corolario V.2.8 también es cierto, cuando x e y tienen el mismo signo y satisfacen $x + y \leq x_0$ para algún $x_0 \geq 0$, y las funciones $m_{ij}(z_1,\ldots,z_s)$, $1 \leq i,j \leq s$, son analíticas en $Re\ z_k \leq x_0$, $(k = 1,\ldots,s)$. Del mismo modo, los enunciados del teorema V.2.6 también se verifican bajo análogas consideraciones. □

Nota V.2.10 En relación al enunciado del teorema V.2.6, recordemos que dada una matriz simétrica y definida positiva $G \in \mathbb{R}^{k,k}$, un Método Lineal General de k pasos y s etapas (V.4) será G-estable si la función $\varphi_{G,M}$ (V.16) verifica $\varphi_{G,M}(0) \leq 1$, donde M es la matriz definida en (V.9).

Nota V.2.11 Aunque los resultados presentados en esta sección están relacionados tanto a normas de tipo euclídeo como a normas l_p, las principales aplicaciones que obtendremos en este capítulo en el estudio de los Métodos Lineales Generales surgirán al considerar normas de tipo euclídeo.

V.3. Contractividad estricta para Métodos Lineales Generales

En relación a los resultados preliminares enunciados en la sección 2, presentamos ahora un lema simple que describe el comportamiento en entornos del origen de aquellas funciones que verifican la propiedad (V.17), bajo cierta suposición acerca del comportamiento en el infinito. Este resultado es un punto clave para probar los principales resultados relacionados con contractividad que se presentarán en esta sección.

Lema V.3.1 *Sea* $\varphi : (-\infty, 0] \to [0, +\infty)$ *una función tal que*

i) $\lim_{x\to-\infty} \varphi(x) = \zeta < 1$,

ii) $\varphi(x)\varphi(y) \le \varphi(0)\varphi(x+y), \quad \forall x, y \in (-\infty, 0]$.

Entonces, existe una constante positiva σ *tal que*

$$\varphi(x) \le \max\left\{\frac{1+\zeta}{2}, Ke^{\sigma x}\right\}, \ \ \forall x \le 0,$$

donde $K = \max\{1, \varphi(0)\}$.

Demostración. En virtud de $i)$, existe $x_0 < 0$ tal que

$$\varphi(x) \le \delta := \frac{1+\zeta}{2}, \quad \forall x \le x_0.$$

Tomemos $\sigma := \dfrac{\ln \delta}{2x_0} > 0$. Así, si $x \in [2x_0, x_0]$, sigue que

$$\varphi(x) \le \delta = e^{\sigma(2x_0)} \le e^{\sigma x}.$$

Por otro lado, si $x \in [2x_0, x_0]$ y $l \in \mathbb{N}$, entonces en virtud de $ii)$

$$\varphi\Big(\frac{x}{l}\Big)^l \le \varphi(0)^{l-1} e^{\sigma x}.$$

Esto implica

$$\varphi\Big(\frac{x}{l}\Big) \le Ke^{\sigma(\frac{x}{l})}, \quad K = \max\{1, \varphi(0)\}.$$

Consideremos $x \in (x_0, 0)$. Entonces, existe $n \in \mathbb{N}$ tal que $x \in [\frac{x_0}{n}, \frac{x_0}{n+1})$. Por lo tanto, $y := (n+1)x \in [\frac{n+1}{n}x_0, x_0) \subseteq [2x_0, x_0)$, y

$$\varphi(x) = \varphi\Big(\frac{y}{n+1}\Big) \le Ke^{\sigma(\frac{y}{n+1})} = Ke^{\sigma x}.$$

Esto concluye la prueba. □

El siguiente teorema establece una cota acerca del comportamiento en el infinito de la función $\varphi_{G,M}$ (V.16). Debemos observar que para un Método Lineal General (V.4) la función de estabilidad lineal $R(z)$ (V.5) toma valor en el infinito $R(\infty) = B - A\tilde{A}^{-1}\tilde{B}$ siempre que la matriz $\tilde{A}$ sea no singular.

Teorema V.3.2 *Sea $M(Z)$, con $Z = \mathrm{BlockDiag}(Z_1, \ldots, Z_s)$, dada por (V.9), con $\tilde{A}$ no singular, y sea $G = L^T L$ una matriz simétrica y definida positiva, con $L \in \mathbb{R}^{k,k}$ no singular. Entonces, tenemos para $\varphi_{G,M}$ dada por (V.16) que*

$$\varphi_{G,M}(x) \leq \left\| B - A\tilde{A}^{-1}\tilde{B} \right\|_G + \left\| LA\tilde{A}^{-1} \right\|_2 \left\| \tilde{B}L^{-1} \right\|_2 \frac{\left\| \tilde{A}^{-1} \right\|_2}{|x| - \left\| \tilde{A}^{-1} \right\|_2}, \quad \forall x < -\left\| \tilde{A}^{-1} \right\|_2 .$$

[Notemos que la cota superior que aparece en este teorema no depende de la descomposición de la matriz $G = L^T L$.]

Demostración. Ya que $\tilde{A}$ es una matriz no singular, podemos escribir partiendo de (V.9) que

$$M(Z) = (B - A\tilde{A}^{-1}\tilde{B}) \otimes I_m + (A\tilde{A}^{-1} \otimes I_m)(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}(\tilde{B} \otimes I_m).$$

De este modo, teniendo en cuenta que $X = Y^T Y$, $Y \in \mathbb{R}^{m,m}$, y (V.13), sigue que

$$\begin{aligned} \|M(Z)\|_{G\otimes X} &= \|(L \otimes Y)M(Z)(L^{-1} \otimes Y^{-1})\|_2 \\ &\leq \left\| B - A\tilde{A}^{-1}\tilde{B} \right\|_G + \left\| LA\tilde{A}^{-1} \right\|_2 \left\| \tilde{B}L^{-1} \right\|_2 \left\| (I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1} \right\|_{I_s \otimes Y} . \end{aligned} \tag{V.19}$$

Tomemos $u \in \mathbb{C}^{ms}$, $\|u\|_2 = 1$, y definamos

$$v := (I_s \otimes Y)(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}(I_s \otimes Y^{-1})u.$$

nuestro objetivo es, entonces, obtener una cota para $\|v\|_2$. De la ecuación anterior obtenemos que

$$(\tilde{A}^{-1} \otimes I_m)u = (\tilde{A}^{-1} \otimes I_m)v - (I_s \otimes Y)Z(I_s \otimes Y^{-1})v, \tag{V.20}$$

y de aquí que

$$\mathrm{Re}\langle v, (\tilde{A}^{-1} \otimes I_m)u\rangle_2 = \mathrm{Re}\langle v, (\tilde{A}^{-1} \otimes I_m)v\rangle_2 - \mathrm{Re}\langle v, (I_s \otimes Y)Z(I_s \otimes Y^{-1})v\rangle_2. \tag{V.21}$$

Por tanto, teniendo en cuenta que

$$(I_s \otimes Y)Z(I_s \otimes Y^{-1}) = \mathrm{BlockDiag}(YZ_1Y^{-1}, \ldots, YZ_sY^{-1}),$$

y denotando $v = (v_1^T, \ldots, v_s^T)^T$, con $v_i \in \mathbb{C}^m$, $1 \leq i \leq s$, sigue que

$$\begin{aligned} \mathrm{Re}\langle v, (I_s \otimes Y)Z(I_s \otimes Y^{-1})v\rangle_2 &= \sum_{i=1}^{s} \mathrm{Re}\langle v_i, YZ_iY^{-1}v_i\rangle_2 \\ &\leq \sum_{i=1}^{s} \mu_2[YZ_iY^{-1}] \, \|v_i\|_2^2 \\ &= \sum_{i=1}^{s} \mu_X[Z_i] \, \|v_i\|_2^2 \\ &\leq x \, \|v\|_2^2 , \end{aligned}$$

con $x = \max_{1\leq i\leq s} \mu_X[Z_i]$. Más aún, es directo comprobar que

$$\mathrm{Re}\langle v, (\tilde{A}^{-1} \otimes I_m)v\rangle_2 \geq -\left\|\tilde{A}^{-1}\right\|_2 \|v\|_2^2$$

y

$$\mathrm{Re}\langle v, (\tilde{A}^{-1} \otimes I_m)u\rangle_2 \leq \left\|\tilde{A}^{-1}\right\|_2 \|v\|_2$$

Entonces, partiendo de (V.21) concluimos que

$$(-x - \left\|\tilde{A}^{-1}\right\|_2) \|v\|_2^2 \leq \left\|\tilde{A}^{-1}\right\|_2 \|v\|_2 .$$

La prueba concluye teniendo en cuenta (V.19). □

Presentamos ahora un resultado importante de cara a la contractividad de los Métodos Lineales Generales.

Teorema V.3.3 *Asumamos que el Método Lineal General (V.4) posee una matriz $\tilde{A}$ no singular, que es G-estable y que su matriz de estabilidad lineal (V.5) satisface $\|R(\infty)\|_G < 1$. Entonces, el método es estrictamente contractivo en la clase de sistemas diferenciales (V.1) estrictamente disipativos ($\nu < 0$ en (V.2)), esto es, existen dos constantes positivas σ y $\gamma < 1$ únicamente dependientes de los coeficientes del método tales que*

$$\left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} \leq \max\{\gamma, e^{h\sigma\nu}\} \left\|\Delta Y^{(n)}\right\|_{G\otimes X}, \quad \forall h > 0, \quad n \geq 0.$$

Demostración. En virtud del teorema V.2.6 y de la G-estabilidad del método, tenemos que la función $\varphi_{G,M}(x)$ (V.16) verifica (V.17) y $\varphi_{G,M}(0) \leq 1$.

Por otro lado, ya que $\|R(\infty)\|_G < 1$, deducimos del Teorema V.3.2 y del Lema V.3.1 la existencia de dos constantes positivas σ y $\gamma < 1$ tales que $\varphi_{G,M}(h\nu) \leq \max\{\gamma, e^{h\sigma\nu}\}$. La prueba concluye teniendo en cuenta (V.6). □

Nota V.3.4 El principal inconveniente para aplicar el teorema previo a los métodos de tipo Runge-Kutta Multipaso surge del hecho de que muchos de estos métodos cumplen que $\|R(\infty)\|_G = 1$, donde $\|\cdot\|_G$ denota la norma de tipo euclídeo que hace G-estable al método correspondiente. Por ejemplo, este es el caso del método BDF de dos pasos o de los Métodos Runge-Kutta Gauss Multipaso introducidos por Burrage [6, 7], tal como veremos en la última sección del capítulo. De esta manera, para obtener contractividad estricta para muchos métodos G-estables parece necesario componer los métodos varios pasos consecutivos, pongamos l pasos consecutivos. A menudo, para métodos de tipo Runge-Kutta multipaso, l coincide con el número de pasos (k) en el que se basa el método. Todo esto motiva el estudio del *método compuesto a l pasos* asociado a un Método Lineal General (V.4).

Un sencillo cálculo permite ver que la solución numérica provista por (V.4) tras l pasos consecutivos del mismo tamaño $h > 0$ está dada por

$$Y^{(n+l)} = h\sum_{i=0}^{l-1}(B^{l-1-i}A \otimes I_m)F(t_{n+i}, h, V^{(n+i)}) + (B^l \otimes I_m)Y^{(n)}, \quad l \geq 1,$$

donde las etapas $V^{(n+j)}$ se calculan a partir del sistema

$$\begin{aligned} V^{(n+j)} &= h\sum_{i=0}^{j-1}(\tilde{B}B^{j-1-i}A \otimes I_m)F(t_{n+i}, h, V^{(n+i)}) \\ &\quad + h(\tilde{A} \otimes I_m)F(t_{n+j}, h, V^{(n+j)}) + (\tilde{B}B^j \otimes I_m)Y^{(n)}, \quad 0 \leq j \leq l-1. \end{aligned}$$

Así, el método compuesto a l pasos puede ser visto como un nuevo Método Lineal General de k pasos y ls etapas de la forma

$$\begin{aligned} \overline{V}^{(n+l-1)} &= h(\tilde{\alpha} \otimes I_m)\overline{F}(t_n, h, \overline{V}^{(n+l-1)}) + (\tilde{\beta} \otimes I_m)Y^{(n)}, \\ Y^{(n+l)} &= h(\alpha \otimes I_m)\overline{F}(t_n, h, \overline{V}^{(n+l-1)}) + (\beta \otimes I_m)Y^{(n)}, \end{aligned} \tag{V.22}$$

donde

$$\begin{aligned} \overline{V}^{(n+l-1)} &:= ({V^{(n)}}^T, {V^{(n+1)}}^T, \ldots, {V^{(n+l-1)}}^T)^T \in \mathbb{C}^{mls}, \\ \overline{F}(t_n, h, \overline{V}^{(n+l-1)}) &:= (F(t_n, h, V^{(n)})^T, F(t_{n+1}, h, V^{(n+1)})^T, \ldots, F(t_{n+l-1}, h, V^{(n+l-1)})^T)^T \in \mathbb{C}^{mls}, \end{aligned}$$

y

$$\tilde{c} = \begin{pmatrix} c \\ c+e \\ \vdots \\ c+(l-1)e \end{pmatrix} \in \mathbb{R}^{ls}, \qquad \alpha = [B^{l-1}A,\ B^{l-2}A,\ \ldots\ BA,\ A] \in \mathbb{R}^{k,ls}, \quad \beta = B^l \in \mathbb{R}^{k,k},$$

$$\tilde{\alpha} = \begin{pmatrix} \tilde{A} & O & O & \ldots & O & O \\ \tilde{B}A & \tilde{A} & O & \ldots & O & O \\ \tilde{B}BA & \tilde{B}A & \tilde{A} & \ldots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \tilde{B}B^{l-2}A, & \tilde{B}B^{l-3}A, & \ldots & \ldots, & \tilde{B}A, & \tilde{A} \end{pmatrix} \in \mathbb{R}^{ls,ls}, \quad \tilde{\beta} = \begin{pmatrix} \tilde{B} \\ \tilde{B}B \\ \vdots \\ \tilde{B}B^{l-1} \end{pmatrix} \in \mathbb{R}^{ls,k}.$$

En consecuencia, para estudiar la A-estabilidad del método (V.22) debemos considerar la matriz

$$R_l(z) = \beta + z\alpha(I - z\tilde{\alpha})^{-1}\tilde{\beta}.$$

Ya que para problemas lineales $y' = \lambda y,\ z = \lambda h$, el método compuesto a l pasos da $Y^{(n+l)} = R_l(z)Y^{(n)}$ y también $Y^{(n+l)} = R(z)^l Y^{(n)}$, donde $R(z)$ denota la matriz de estabilidad (V.5) del método (V.4), entonces sigue que

$$R_l(z) = R(z)^l, \quad \forall z \in \mathbb{C}.$$

Por otro lado, es claro que si $\tilde{A}$ es regular, entonces también lo es $\tilde{\alpha}$, y en tal caso

$$R_l(\infty) = \beta - \alpha\tilde{\alpha}^{-1}\tilde{\beta} = (B - A\tilde{A}^{-1}\tilde{B})^l = R(\infty)^l.$$

Para estudiar la contractividad estricta del método compuesto a l pasos sobre problemas no lineales en general, si procedemos del mismo modo que en (V.6) tenemos que

$$\Delta Y^{(n+l)} = M_l(h\overline{J})\Delta Y^{(n)}, \quad n = 0, 1, 2, \ldots,$$

donde

$$\overline{J} = \text{BlockDiag}(J^{(n+1)}, \ldots, J^{(n+l)}) \in \mathbb{C}^{mls,mls},$$

siendo las matrices $J^{(n+j)},\ (j = 1, \ldots, l)$ definidas como en (V.7), y la super-matriz $M_l(\cdot)$ dada por,

$$M_l(\overline{Z}) := \beta \otimes I_m + (\alpha \otimes I_m)\overline{Z}(I_{mls} - (\tilde{\alpha} \otimes I_m)\overline{Z})^{-1}(\tilde{\beta} \otimes I_m) \in \mathbb{C}^{mk,mk}.$$

Por supuesto, si el método original (V.4) es G-estable, entonces también lo es el método compuesto a l pasos (V.22). En consecuencia, podemos aplicar el teorema V.3.3 al Método Lineal General compuesto a l pasos para obtener el siguiente

Teorema V.3.5 *Sea un Método Lineal General G-estable dado por (V.4). Supongamos que la matriz $\tilde{A}$ es no singular y que existe $l \in \mathbb{N}$ tal que $\left\|R(\infty)^l\right\|_G < 1$, con $R(z)$ dado por (V.5). Entonces, existen dos constantes positivas σ y $\gamma < 1$, dependientes únicamente de los coeficientes del método, tales que*

$$\left\|\Delta Y^{(n+l)}\right\|_{G\otimes X} \leq \max\{\gamma, e^{h\sigma\nu}\} \left\|\Delta Y^{(n)}\right\|_{G\otimes X}, \quad \forall h > 0,\ \ \forall n \geq 0,$$

se verifica en la clase de sistemas diferenciales (V.1) estrictamente disipativos ($\nu < 0$ en (V.2)). En particular, el Método Lineal General compuesto a l pasos (V.22) es estrictamente contractivo.

Otra consecuencia es el siguiente

Teorema V.3.6 *Sea un Método Lineal General G-estable dado por (V.4). Supongamos que la matriz $\tilde{A}$ es no singular y que $\rho(R(\infty)) < 1$. Entonces, existe un entero positivo l y dos constantes positivas σ y $\gamma < 1$, dependientes únicamente de los coeficientes del método, tales que*

$$\left\|\Delta Y^{(n+l)}\right\|_{G\otimes X} \leq \max\{\gamma, e^{h\sigma\nu}\} \left\|\Delta Y^{(n)}\right\|_{G\otimes X}, \quad \forall h > 0, \ \ \forall n \geq 0,$$

se verifica en la clase de sistemas diferenciales (V.1) estrictamente disipativos ($\nu < 0$ en (V.2)). En particular, el Método Lineal General compuesto a l pasos (V.22) es estrictamente contractivo.

Demostración. La demostración es consecuencia inmediata del corolario previo. Basta tener en cuenta que

$$\rho(R(\infty)) = \lim_{l\to\infty} \left\|R(\infty)^l\right\|_G^{1/l} < 1.$$

□

Nota V.3.7 Aún es posible deducir resultados de contractividad si la matriz $\tilde{A}$ del Método Lineal General original es singular y el método es algebraicamente estable. Para ello asumiremos en lo que resta de sección que el método (V.4) es preconsistente (ver (V.11)) e irreducible. Además, asumiendo que el método es algebraicamente estable, entonces la matriz $D = Diag(d_1, d_2, \ldots, d_s)$ es definida positiva ($D > 0$), véase [57, Lema 3.2] o [58, Lema 4.1].

En virtud de [53, Lema 9.5, p. 359], tenemos para la matriz diagonal D y la matriz G que definen la estabilidad algebraica del método (véase (V.10)) que que

$$De = A^T G\xi_0, \qquad G\xi_0 = B^T G\xi_0. \tag{V.23}$$

Por lo tanto, deducimos el siguiente

Lema V.3.8 *Sea (V.4) un Método Lineal General preconsistente, irreducible y algebraicamente estable. Entonces, el siguiente límite existe*

$$\lim_{\varepsilon\to 0} A(\tilde{A} + \varepsilon I)^{-1} = \tilde{A}_0.$$

Demostración. Si la matriz $\tilde{A}$ es regular, entonces el resultado sigue con $\tilde{A}_0 = A\tilde{A}^{-1}$.

Supongamos ahora que $\tilde{A}$ es no regular. Entonces, probaremos en primer lugar que $\lambda = 0$ debe ser un autovalor simple de $\tilde{A}$. Para ello tomamos $u \in \mathbb{R}^s$ tal que $\tilde{A}u = 0$, y supongamos que existe $w \in \mathbb{R}^s$ tal que $\tilde{A}w = u$. Definamos entonces $u_0 := (\xi_0^T, u^T)^T \in \mathbb{R}^{k+s}$. Ya que N, dada por (V.10), es definida no negativa, deducimos a partir de (V.11) y (V.23) que

$$\begin{aligned} 0 &\leq u_0^T N u_0 \\ &= \xi_0^T (G - B^T G B)\xi_0 + 2u^T (D\tilde{B} - A^T G B)\xi_0 + u^T (D\tilde{A} + \tilde{A}^T D - A^T G A)u \\ &= -(Au)^T G (Au). \end{aligned}$$

Entonces $u_0^T N u_0 = (Au)^T G(Au) = 0$, lo cual implica $Au = 0$, ya que G es definida positiva. Con todo lo anterior, ya que N es definida no negativa, obtenemos que

$$0 \leq (\lambda u_0 + w_0)^T N(\lambda u_0 + w_0) = 2\lambda u_0^T N w_0 + w_0^T N w_0, \quad \forall \lambda \in \mathbb{R}.$$

Esto implica $u_0^T N w_0 = 0$ para todo $w_0 \in \mathbb{R}^{s+k}$. Entonces, tomando $w_0 = \begin{pmatrix} 0 \\ w \end{pmatrix} \in \mathbb{R}^{s+k}$, obtenemos usando (V.23) que

$$\begin{aligned} 0 &= u_0^T N w_0 \\ &= \xi_0^T(\tilde{B}^T D - B^T G A)w + u^T D u + u^T \tilde{A}^T D w - (Au)^T G(Aw) \\ &= u^T D u, \end{aligned}$$

y por lo tanto $u = 0$ ya que D es definida positiva.

Consideremos ahora el desarrollo $(\tilde{A} + \varepsilon I)^{-1} = \varepsilon^{-1}\tilde{A}_{-1} + \tilde{A}_0 + \varepsilon\tilde{A}_1 + \ldots$, con matrices $\tilde{A}_i \in \mathbb{R}^{s,s}$, $i = -1, 0, 1, \ldots$ Ya que $I = (\tilde{A} + \varepsilon I)(\tilde{A} + \varepsilon I)^{-1} = (\tilde{A} + \varepsilon I)^{-1}(\tilde{A} + \varepsilon I)$, sigue que $\tilde{A}\tilde{A}_{-1} = 0 = \tilde{A}_{-1}\tilde{A}$. Por otro lado, la matriz

$$\mathcal{N} = \begin{pmatrix} O & O \\ O & \tilde{A}_{-1}^T \end{pmatrix} N \begin{pmatrix} O & O \\ O & \tilde{A}_{-1} \end{pmatrix}$$

es definida no negativa (aquí O denota la correspondiente matriz nula de $\mathbb{R}^{k,k}$, $\mathbb{R}^{k,l}$ y $\mathbb{R}^{l,k}$). Sin embargo, un cálculo directo muestra que

$$\mathcal{N} = \begin{pmatrix} O & O \\ O & \tilde{A}_{-1}^T(D\tilde{A} + \tilde{A}^T D - A^T G A)\tilde{A}_{-1} \end{pmatrix} = \begin{pmatrix} O & O \\ O & -(A\tilde{A}_{-1})^T G(A\tilde{A}_{-1}) \end{pmatrix},$$

que es definida no positiva. Así, $\mathcal{N} = 0$, y $A\tilde{A}_{-1} = 0$, ya que G es definida positiva. □

Como consecuencia del lema previo, tenemos el siguiente teorema, que puede ser interpretado como una generalización del teorema [54, Teorema 3.1] al caso de Métodos Lineales Generales.

Teorema V.3.9 *Sea (V.4) un Método Lineal General preconsistente y algebraicamente estable para las matrices G y D, con matriz diagonal D definida positiva (ver (V.10)). Entonces la diferencia $\Delta Y^{(j)} = Y^{(j)} - \hat{Y}^{(j)}$, $j \geq 0$, entre dos soluciones numéricas del método aplicado al sistema diferencial disipativo (V.1) ($\nu \leq 0$ en (V.2)) satisface*

$$\left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} \leq \chi(h\nu) \left\|\Delta Y^{(n)}\right\|_{G\otimes X}, \; n \geq 0,$$

donde

$$\chi(x) := \frac{\sqrt{1 - 2x\gamma(1-\zeta^2)} - 2x\gamma\zeta}{1 - 2x\gamma}, \tag{V.24}$$

y

$$\zeta = \|R(\infty)\|_G\,,$$
$$\gamma = \left(\lim_{\varepsilon\to 0}\left\|LA(\tilde{A}+\varepsilon I)^{-1}D^{-1/2}\right\|_2\right)^{-2}, \quad \text{con } G = L^TL. \tag{V.25}$$

Nota V.3.10 Debemos notar que, de acuerdo con [54, p. 214], la función $\chi(x)$ resulta ser superexponencial (véase (V.2.7)), aunque la prueba es algo tediosa. Por otro lado es inmediato comprobar que esta función verifica $\chi(x) \geq \zeta$, $\forall x \leq 0$, siempre que $\zeta \leq 1$.

Demostración. Debido a la estabilidad algebraica del método, sigue de [53, Lema 9.2, p. 357] (ver también [8]) que

$$\left\|\Delta Y^{(n+1)}\right\|^2_{G\otimes X} - \left\|\Delta Y^{(n)}\right\|^2_{G\otimes X} \leq 2x\sum_{i=1}^{s} d_i\left\|\Delta v_i^{(n)}\right\|^2_X, \quad x := h\nu \leq 0, \tag{V.26}$$

donde $\Delta v_i^{(n)} := \hat{v}_i^{(n)} - v_i^{(n)}$, y $d_i > 0$, $1 \leq i \leq s$.

Definamos $\tilde{A}_\varepsilon := \tilde{A} + \varepsilon I$, y consideremos el Método Lineal General definido como en (V.4) reemplazando $\tilde{A}$ por $\tilde{A}_\varepsilon$. Entonces

$$\begin{aligned}\Delta Y_\varepsilon^{(n+1)} &= h(A\otimes I_m)\Delta F(t_n,h,V_\varepsilon^{(n)}) + (B\otimes I_m)\Delta Y^{(n)},\\ \Delta V_\varepsilon^{(n)} &= h(\tilde{A}_\varepsilon\otimes I_m)\Delta F(t_n,h,V_\varepsilon^{(n)}) + (\tilde{B}\otimes I_m)\Delta Y^{(n)}.\end{aligned}$$

Ya que $h\Delta F(t_n,h,V_\varepsilon^{(n)}) = (\tilde{A}_\varepsilon^{-1}\otimes I_m)\left(\Delta V_\varepsilon^{(n)} - (\tilde{B}\otimes I_m)\Delta Y^{(n)}\right)$, obtenemos que

$$\Delta Y_\varepsilon^{(n+1)} = ((B - A\tilde{A}_\varepsilon^{-1}\tilde{B})\otimes I_m)\Delta Y^{(n)} + (A\tilde{A}_\varepsilon^{-1}\otimes I_m)\Delta V_\varepsilon^{(n)}. \tag{V.27}$$

Ya que D es regular tenemos que

$$\begin{aligned}\left\|(A\tilde{A}_\varepsilon^{-1}\otimes I_m)\Delta V_\varepsilon^{(n)}\right\|_{G\otimes X} &= \left\|(LA\tilde{A}_\varepsilon^{-1}D^{-1/2}\otimes I_m)(D^{1/2}\otimes Y)\Delta V_\varepsilon^{(n)}\right\|_2\\ &\leq \left\|LA\tilde{A}_\varepsilon^{-1}D^{-1/2}\right\|_2\left(\sum_{i=1}^{s} d_i\left\|\Delta v_{i,\varepsilon}^{(n)}\right\|^2_X\right)^{1/2}.\end{aligned}$$

En virtud del lema previo, tenemos que el límite $\tilde{A}_0 := \lim_{\varepsilon\to 0} A\tilde{A}_\varepsilon^{-1}$ existe. En consecuencia, tomando límites en (V.27), deducimos que

$$\left\|\Delta Y^{(n+1)} - (R(\infty)\otimes I_m)\Delta Y^{(n)}\right\|_{G\otimes X} \leq \sigma\left(\sum_{i=1}^{s} d_i\left\|\Delta v_i^{(n)}\right\|^2_X\right)^{1/2},$$

con $\sigma = \left\| L\tilde{A}_0 D^{-1/2} \right\|_2$.

Así, si $\left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} \geq \zeta \left\|\Delta Y^{(n)}\right\|_{G\otimes X}$, donde $\zeta := \|R(\infty)\|_G$, entonces

$$\begin{aligned}\left| \left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} - \zeta \left\|\Delta Y^{(n)}\right\|_{G\otimes X} \right| &= \left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} - \|R(\infty)\otimes I_m\|_{G\otimes X} \left\|\Delta Y^{(n)}\right\|_{G\otimes X} \\ &\leq \left\|\Delta Y^{(n+1)} - (R(\infty)\otimes I_m)\Delta Y^{(n)}\right\|_{G\otimes X} \\ &\leq \sigma\left(\sum_{i=1}^{s} d_i \left\|\Delta v_i^{(n)}\right\|_X^2\right)^{1/2}.\end{aligned}$$

Consecuentemente, deducimos de aquí que

$$\sum_{i=1}^{s} d_i \left\|\Delta v_i^{(n)}\right\|_X^2 \geq \gamma\left(\left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} - \zeta \left\|\Delta Y^{(n)}\right\|_{G\otimes X} \right)^2, \tag{V.28}$$

con $\gamma := \sigma^{-2}$. Insertando esta última desigualdad (V.28) en (V.26), obtenemos la desigualdad

$$(1-2x\gamma)\left\|\Delta Y^{(n+1)}\right\|_{G\otimes X}^2 + 4x\gamma\zeta \left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} \left\|\Delta Y^{(n)}\right\|_{G\otimes X} - (1+2x\gamma\zeta^2)\left\|\Delta Y^{(n)}\right\|_{G\otimes X}^2 \leq 0.$$

Con esto, un cálculo directo permite mostrar que

$$\left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} \leq \chi(x) \left\|\Delta Y^{(n)}\right\|_{G\otimes X}, \quad x = h\nu,$$

donde la función $\chi(x)$ está dada por (V.24). Esto concluye la prueba. □

Nota V.3.11 Aunque los teoremas V.3.3 y V.3.9 permiten deducir conclusiones análogas en lo referente a la contractividad de los métodos, debemos notar que la prueba del teorema V.3.9 hace uso de la expresión de la solución de avance del Método Lineal General en términos de las etapas internas, mientras que la prueba del teorema V.3.3 se basa exclusivamente en la expresión de la matriz $M(Z_1,\ldots,Z_s)$ dada por (V.9), y por tanto posee un rango más amplio de aplicación y podría ser considerado, junto al teorema V.3.2, en situaciones más generales.

Corolario V.3.12 *Cualquier Método Lineal General (V.4) preconsistente y algebraicamente estable, con matriz diagonal D definida positiva (V.10), verificando* $\|R(\infty)\|_G < 1$ *es estrictamente contractivo en la clase de sistemas diferenciales estrictamente disipativos. Además, se tiene que*

$$\left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} \leq \chi(h\nu) \left\|\Delta Y^{(n)}\right\|_{G\otimes X}, \quad h>0,\ n\geq 0,$$

donde $\chi(\cdot)$ *está dada por (V.24).*

Demostración. La prueba se obtiene a partir del teorema V.3.9 teniendo en cuenta que si $\zeta := \|R(\infty)\|_G < 1$, entonces la función $\chi(x)$ satisface $\chi(x) < 1$, $\forall x < 0$. □

Nota V.3.13 Si la matriz de estabilidad $R(z)$ del Método Lineal General satisface $\|R(\infty)\|_G = 1$, entonces aún se pueden deducir resultados de contractividad estricta si $\rho(R(\infty)) < 1$. Supongamos entonces que el Método Lineal General (V.4) es preconsistente (para cierto vector $\xi_0 \in \mathbb{R}^k$) y algebraicamente estable para las matrices G y D, ambas definidas positivas. Esto implica que tras $l \geq 1$ pasos consecutivos de igual longitud, el método compuesto (V.22) es también preconsistente (para $\xi_0 \in \mathbb{R}^k$) y algebraicamente estable (véase [57, Lema 3.1]) para la matriz G y la matriz diagonal $D_l \in \mathbb{R}^{ls,ls}$ dada por

$$D_l e = \alpha^T G \xi_0 = ((De)^T, (De)^T, \ldots, (De)^T)^T. \tag{V.29}$$

Como consecuencia de esta nota previa y del corolario V.3.12, deducimos el siguiente

Corolario V.3.14 *Sea (V.4) un Método Lineal General preconsistente y algebraicamente estable con matriz diagonal D definida positiva, tal que $\rho(R(\infty)) < 1$. Entonces, existe un entero $l \geq 1$, dependiente únicamente de los coeficientes del método, tal que el método compuesto a l pasos (V.22) es estrictamente contractivo en la clase de sistemas diferenciales estrictamente disipativos ($\nu < 0$ en (V.2)). En particular, se verifica que*

$$\left\|\Delta Y^{(n+l)}\right\|_{G\otimes X} \leq \chi_l(h\nu) \left\|\Delta Y^{(n)}\right\|_{G\otimes X}, \quad h > 0,\ n \geq 0,$$

donde $\chi_l(\cdot)$ está dada como (V.24), con $\zeta := \left\|R(\infty)^l\right\|_G$, y reemplazando en (V.25) A por α, $\tilde{A}$ por $\tilde{\alpha}$ (ver (V.22)) y D por D_l (ver V.29), respectivamente.

V.4. Resultados de convergencia

Comenzamos esta sección notando que los errores de discretización local asociados a un Método Lineal General (V.4) vienen definidos por el par de vectores $(\xi(t_n), \eta(t_n))$ definidos por (véase, por ejemplo, [58, p. 365])

$$\begin{aligned} \xi(t_n) &:= Y(t_n + h) - (B \otimes I_m)Y(t_n) - h(A \otimes I_m)V'(t_n), \\ \eta(t_n) &:= \quad V(t_n) - (\tilde{B} \otimes I_m)Y(t_n) - h(\tilde{A} \otimes I_m)V'(t_n), \end{aligned}$$

donde $V(t_n) = (y(t_n + c_j h))_{j=1}^s$, $V'(t_n) = (y'(t_n + c_j h))_{j=1}^s \in \mathbb{C}^{ms}$, e $Y(t_n) = (Y_j(t_n))_{j=1}^k \in \mathbb{C}^{mk}$. Aquí $Y(t)$ representa la solución exacta de avance que debe ser aproximada por el método, mientras que $y(t)$ denota la solución exacta del problema de valor inicial (V.1). No es difícil comprobar que si el método posee orden de etapa q, entonces

$$\begin{aligned} \xi(t_n) &= h^{q+1}(d_1^{(q+1)} \otimes I_m)y^{(q+1)}(t_n) + \mathcal{O}(h^{q+2}), \\ \eta(t_n) &= h^{q+1}(d_2^{(q+1)} \otimes I_m)y^{(q+1)}(t_n) + \mathcal{O}(h^{q+2}), \end{aligned} \tag{V.30}$$

donde los vectores $d_1^{(q+1)} \in \mathbb{R}^k$ y $d_2^{(q+1)} \in \mathbb{R}^s$ dependen únicamente de los coeficientes del método (véase, por ejemplo, [58, p. 366]).

Un cálculo directo nos permite ver las siguientes cotas superiores para los errores de discretización local

$$\begin{aligned} \|\xi(t_n)\|_{I_k\otimes X} &\leq h^q C_1 \int_{t_n}^{t_{n+1}} \left\|y^{(q+1)}(t)\right\|_X dt, \\ \|\eta(t_n)\|_{I_s\otimes X} &\leq h^q C_2 \int_{t_n}^{t_{n+1}} \left\|y^{(q+1)}(t)\right\|_X dt, \end{aligned} \tag{V.31}$$

donde las constantes C_j, $j = 1, 2$, dependen únicamente de los coeficientes del método. Por otro lado, también tenemos que

$$\begin{aligned} \left\|\xi(t_n) - h^{q+1}(d_1^{(q+1)} \otimes I_m)y^{(q+1)}(t_n)\right\|_{I_k\otimes X} &\leq h^{q+1} C_1' \int_{t_n}^{t_{n+1}} \left\|y^{(q+2)}(t)\right\|_X dt, \\ \left\|\eta(t_n) - h^{q+1}(d_2^{(q+1)} \otimes I_m)y^{(q+1)}(t_n)\right\|_{I_s\otimes X} &\leq h^{q+1} C_2' \int_{t_n}^{t_{n+1}} \left\|y^{(q+2)}(t)\right\|_X dt, \end{aligned} \tag{V.32}$$

donde, de nuevo, las constantes $C_j', (j = 1, 2)$ dependen exclusivamente de los coeficientes del método.

El estudio de los errores globales $\epsilon_n := Y(t_n) - Y^{(n)},\ n = 1, 2, \ldots,$ se puede llevar a cabo, por ejemplo, siguiendo las ideas en [58, Sec. 2]. Así, se demuestra tras un desarrollo que los errores verifican, para $n \geq 0$ la ley de recurrencia

$$\begin{aligned} \epsilon_{n+1} &= M(Z^{(n)})\epsilon_n + \tau_n, \\ \tau_n &:= \xi(t_n) + \omega(Z^{(n)})\eta(t_n), \\ \omega(Z) &:= (A \otimes I_m)Z(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}, \end{aligned} \tag{V.33}$$

donde $M(Z)$ es la función de estabilidad dada en (V.9), con $Z^{(n)} = hJ^{(n)}$ y $J^{(n)}$ dada por (V.8) y (V.7).

Por tanto, debemos poner interés en la obtención de cotas superiores para

$$\left\|\prod_{j=0}^{q} M(Z^{(n-j)})\right\|_{G\otimes X}, \qquad q \leq n,$$

así como para

$$\sup_{\|u\|_{I_s\otimes X}=1} \left\|\omega(Z^{(n)})u\right\|_{G\otimes X}.$$

Con esta idea en mente presentamos el siguiente

Teorema V.4.1 *Supongamos que $\tilde{A} \in \mathbb{R}^{s,s}$ es no singular y que existe una matriz diagonal definida positiva $\tilde{D} = Diag(\tilde{d}_1, \ldots, \tilde{d}_s)$ tal que*

$$\tilde{\alpha} := \frac{1}{2}\lambda_{min}(\tilde{D}\tilde{A}^{-1} + \tilde{A}^{-T}\tilde{D}) \geq 0.$$

Entonces, para cualquier matriz simétrica y definida positiva $G \in \mathbb{R}^{k,k}$, y cualquier $Z = \text{BlockDiag}(Z_1, \ldots, Z_s)$, con $Z_j \in \mathbb{C}^m$, $1 \leq j \leq s$, se tiene que

$$\sup_{\substack{\mu_X[Z_j] \leq x,\ 1 \leq j \leq s, \\ \|u\|_{I_s \otimes X} = 1}} \|\omega(Z)u\|_{G \otimes X} \leq \kappa \left(1 + \frac{\left\|\tilde{D}\tilde{A}^{-1}\right\|_2}{\tilde{\alpha} - x\delta}\right), \quad \forall x < 0,$$

donde

$$\kappa = \sqrt{\lambda_{max}((A\tilde{A}^{-1})^T G A\tilde{A}^{-1})}, \quad \text{y} \quad \delta = \min_{1 \leq j \leq s} \tilde{d}_j. \tag{V.34}$$

Además, si $\tilde{\alpha} > 0$ entonces el enunciado es válido para $x = 0$.

Demostración. Pongamos $G = L^T L$ y $X = Y^T Y$, y tomemos cualquier $u \in \mathbb{C}^{ms}$ satisfaciendo $\|u\|_{I_s \otimes X} = \|(I_s \otimes Y)u\|_2 = 1$. Definiendo $y = \omega(Z)u$ sigue que

$$\begin{aligned}
\|y\|_{G \otimes X} &= \left\|((A\tilde{A}^{-1}) \otimes I_m)((I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}u - u)\right\|_{G \otimes X} \\
&\leq \left\|((A\tilde{A}^{-1}) \otimes I_m)(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}u\right\|_{G \otimes X} + \left\|((A\tilde{A}^{-1}) \otimes I_m)u\right\|_{G \otimes X} \\
&\leq \left\|LA\tilde{A}^{-1}\right\|_2 \left(\left\|(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}\right\|_{I_s \otimes X} + 1\right).
\end{aligned}$$

Ahora, para acotar $\left\|(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}\right\|_{I_s \otimes X}$, tomamos $w \in \mathbb{C}^{ms}$ arbitrario con $\|w\|_2 = 1$ y ponemos

$$v = (I_s \otimes Y)(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}(I_s \otimes Y^{-1})w.$$

Del mismo modo que en la prueba del teorema V.3.2 y pre-multiplicando por $v^*(\tilde{D} \otimes I_m)$ en ambos lados de la fórmula (V.20) obtenemos que

$$v^*(\tilde{D}\tilde{A}^{-1} \otimes I_m)w = v^*(\tilde{D}\tilde{A}^{-1} \otimes I_m)v - v^*(I_s \otimes Y)(\tilde{D} \otimes I_m)Z(I_s \otimes Y^{-1})v. \tag{V.35}$$

Tomando partes reales en la ecuación previa y haciendo uso de las desigualdades

$$\text{Re}(v^*(I_s \otimes Y)(\tilde{D} \otimes I_m)Z(I_s \otimes Y^{-1})v) \leq \delta x \|v\|_2^2, \quad \forall x < 0,$$

y

$$\text{Re}(v^*((\tilde{D}\tilde{A}^{-1}) \otimes I_m)v) = 2^{-1}v^*((\tilde{D}\tilde{A}^{-1} + \tilde{A}^{-T}\tilde{D}) \otimes I_m)v \geq \tilde{\alpha}\|v\|_2^2,$$

se deduce de (V.35) que

$$(\tilde{\alpha} - \delta x) \|v\|_2^2 \leq \mathrm{Re}(v^*(\tilde{D}\tilde{A}^{-1} \otimes I_m)w) \leq \|v\|_2 \cdot \left\|\tilde{D}\tilde{A}^{-1}\right\|_2 \cdot \|w\|_2 .$$

En consecuencia, deducimos que

$$\left\|(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}\right\|_{I_s \otimes X} \leq \frac{\left\|\tilde{D}\tilde{A}^{-1}\right\|_2}{\tilde{\alpha} - \delta x}.$$

Esto completa la prueba. □

Seguidamente, enunciamos y demostramos el principal resultado de convergencia.

Teorema V.4.2 *Consideremos un Método Lineal General (V.4) irreducible y algebraicamente estable (para las matrices G y D diagonal), que posee orden de etapa q y cuya matriz de coeficientes $\tilde{A}$ es no singular. Si $\rho(R(\infty)) < 1$, entonces tomando cualquier matriz diagonal definida positiva $\tilde{D}$ tal que $\tilde{\alpha} = \alpha_{\tilde{D}}(\tilde{A}^{-1}) \geq 0$, se tiene que los errores globales $\epsilon_n = Y(t_n) - Y^{(n)}$ sobre problemas estrictamente disipativos ($\nu < 0$) satisfacen para $n \geq 0$ y $h > 0$ que*

$$\|\epsilon_n\|_{G\otimes X} \leq \varrho^{[l^{-1}n]} \|\epsilon_0\|_{G\otimes X} + h^{q+1} C_3 l \frac{1 - \varrho^{1+l^{-1}h^{-1}t_{n-1}}}{1 - \varrho} \mathcal{M}_{q+1}(t_n),$$

donde $[x]$ denota la parte entera del número real x y

$$C_3 = C_1\sqrt{\rho(G)} + \kappa C_2 \left(1 + \frac{\left\|\tilde{D}\tilde{A}^{-1}\right\|_2}{\tilde{\alpha} + |\nu|\delta h}\right). \tag{V.36}$$

Aquí, l es el primer entero positivo verificando $\left\|R^l(\infty)\right\|_G < 1$, $\varrho = \max\{\gamma, e^{h\sigma\nu}\} < 1$ (siendo $0 < \gamma < 1$ y $\sigma > 0$ las dos constantes dadas en el teorema V.3.6), C_1 y C_2 son las constantes dadas en (V.31), κ y δ están definidas por (V.34), mientras que

$$\mathcal{M}_j(t_n) := \max_{t\in[0,t_n]} \left\|y^{(j)}(t)\right\|_X . \tag{V.37}$$

[Todas las constantes que aparecen en el enunciado (a excepción de ϱ) dependen exclusivamente de los coeficientes del método. Notemos además que la matriz $\widetilde{D}$ depende exclusivamente de la matriz de coeficientes $\widetilde{A}$.]

Demostración. Tomemos $\varrho = \max\{\gamma, e^{h\sigma\nu}\} < 1$, donde $0 < \gamma < 1$ y $\sigma > 0$ son las dos constantes dadas en el teorema V.3.6. Entonces, según el teorema V.3.6 se verifica que

$$\left\|\prod_{k=1}^{j} M(Z^{(n-k)})\right\| \leq \varrho^{[j/l]}, \quad \forall\, j \leq n. \tag{V.38}$$

Partiendo de (V.33) obtenemos para los errores globales que

$$\|\epsilon_n\| \leq \|\tau_{n-1}\| + \sum_{j=1}^{n-1} \left\| \prod_{k=1}^{j} M(Z^{(n-k)}) \right\| \cdot \|\tau_{n-j-1}\| + \left\| \prod_{k=1}^{n} M(Z^{(n-k)}) \right\| \cdot \|\epsilon_0\|, \tag{V.39}$$

donde, por simplicidad, ponemos $\|\cdot\| \equiv \|\cdot\|_{G\otimes X}$.

Así, para acotar $\|\tau_j\|$, de (V.31) y del teorema V.4.1 deducimos que

$$\|\tau_j\| \leq C_3 h^{q+1} \mathcal{M}_{q+1}(t_{j+1}), \quad h > 0,\ j \geq 0, \tag{V.40}$$

con C_3 definido por (V.36). Insertando (V.38) y (V.40) en (V.39), sigue que

$$\|\epsilon_n\| \leq h^{q+1} C_3 \mathcal{M}_{q+1}(t_n) \sum_{j=0}^{n-1} \varrho^{[j/l]} + \varrho^{[n/l]} \|\epsilon_0\|, \quad h > 0,\ n \geq 1.$$

La prueba concluye teniendo en cuenta que, poniendo $n-1 = pl+q$, con p y q enteros, $0 \leq q < l$, se deduce

$$\sum_{j=0}^{n-1} \varrho^{[j/l]} = l \sum_{k=0}^{p-1} \varrho^k + (q+1)\varrho^p \leq l \sum_{k=0}^{p} \varrho^k = l \frac{1-\varrho^{p+1}}{1-\varrho} = l \frac{1-\varrho^{1+[l^{-1}(n-1)]}}{1-\varrho},$$

mientras que $\varrho^{1+[l^{-1}(n-1)]} \geq \varrho^{1+l^{-1}(n-1)}$. □

Como consecuencia del teorema previo deducimos el orden de convergencia (y B-convergencia) de Métodos Lineales Generales algebraicamente estables en intervalos finitos y semi-infinitos. Veremos que estos resultados de convergencia extienden los resultados pioneros de B-convergencia dados por Frank et al. [39] (véase además [53, Teorema 15.3]) para métodos de tipo Runge-Kutta al caso de Métodos Lineales Generales. Nuestros resultados en intervalos temporales finitos son similares a los que se establecen en [57]. No obstante lo anterior, los resultados sobre intervalos semi-infinitos son, a nuestro entender, completamente nuevos.

Teorema V.4.3 *Bajo las supuestos del teorema V.4.2 para un Método Lineal General, tenemos para problemas disipativos ($\nu \leq 0$) que sus errores globales $\epsilon_n := Y(t_n) - Y^{(n)}$ satisfacen que*

1. *si $\tilde{\alpha} > 0$ y $\nu < 0$, entonces*

$$\|\epsilon_n\|_{G\otimes X} \leq h^q \frac{K}{|\nu|} \mathcal{M}_{q+1}(t_n), \quad h > 0, \quad n \geq 0;$$

2. *si $\tilde{\alpha} > 0$ y $\nu \leq 0$, entonces*

$$\|\epsilon_n\|_{G\otimes X} \leq h^q K' t_n \mathcal{M}_{q+1}(t_n), \quad h > 0, \quad n \geq 0;$$

3. si $\tilde{\alpha} = 0$ *y* $\nu < 0$, *entonces*

$$\|\epsilon_n\|_{G\otimes X} \leq h^{q-1}\frac{K}{|\nu|^2}\mathcal{M}_{q+1}(t_n), \quad h > 0, \quad n \geq 0;$$

y

$$\|\epsilon_n\|_{G\otimes X} \leq h^{q-1}\frac{K'}{|\nu|}t_n\mathcal{M}_{q+1}(t_n), \quad h > 0, \quad n \geq 0,$$

donde $\mathcal{M}_{q+1}(t_n)$ *está dado como en (V.37). En todo caso, las constantes* K *y* K' *dependen únicamente de los coeficientes del método.*

Demostración. La prueba es consecuencia inmediata del teorema V.4.2. Para ello debemos tener en cuenta que si $\nu < 0$ entonces

$$l\frac{1-e^{\sigma\nu h(1+\frac{n-1}{l})}}{1-e^{\sigma\nu h}} \leq \frac{l}{1-e^{\sigma\nu h}} = \frac{l}{\sigma|\nu|h} + \mathcal{O}(1),\ h \to 0;$$

mientras que si $\nu \to 0^-$ entonces se tiene que

$$l\frac{1-e^{\sigma\nu h(1+\frac{n-1}{l})}}{1-e^{\sigma\nu h}} \leq \frac{t_n}{h} + l - 1,$$

puesto que la función $(1-\exp(x))^{-1}(1-\exp(ax))$ es creciente en x si $a > 1$. □

Aún es posible ganar un orden de convergencia adicional para un Método Lineal General en el caso de que

$$d_1^{(q+1)} = (I_k - B)x, \quad d_2^{(q+1)} = -\tilde{B}x, \tag{V.41}$$

se verifique para un determinado vector $x \in \mathbb{R}^k$. Aquí, los vectores $d_1^{(q+1)}$ y $d_2^{(q+1)}$ están dados por (V.30).

Teorema V.4.4 *Bajo los supuestos del teorema V.4.2 para un Método Lineal General que, además, satisface (V.41), se tiene para problemas disipativos (*$\nu \leq 0$*) que sus errores globales satisfacen la conclusión del enunciado del teorema V.4.3 con* q *reemplazado por* $q+1$.

Demostración. La prueba sigue en la misma línea que en la prueba del teorema V.4.2, pero teniendo presente las siguientes modificaciones, que se basan en las ideas establecidas en [58, Sec. 4]. Así, definimos los errores globales modificados

$$\bar{\epsilon}_n := \epsilon_n - (x \otimes I_m)h^{q+1}y^{(q+1)}(t_n),$$

que satisfacen la ley de recurrencia

$$\bar{\epsilon}_n = M(Z^{(n)})\bar{\epsilon}_{n-1} + \bar{\tau}_{n-1}, \quad n \geq 1,$$

donde

$$\|\bar{\tau}_{n-1}\|_{G\otimes X} \leq h^{q+2} C_3' \mathcal{M}_{q+2}(t_n),$$

en virtud de (V.32) y (V.41), con C_3' dado como en (V.36) sustituyendo respectivamente las constantes C_i, $i = 1, 2$, por C_i', $i = 1, 2$, dadas en (V.32). Véase, además, la prueba del teorema 4.2 en [58] para más detalles. Con esto, la prueba concluye de modo directo análogamente a como se hace en la prueba del teorema V.4.2. □

Aunque según el teorema previo se pueda ganar un orden de convergencia adicional al orden de etapa, muy pocos métodos de interés cumplen (V.41), tal como ha sido indicado por Hundsdorfer [58, p. 378]. Además, esta condición parece ser necesaria para poder alcanzar un orden de convergencia adicional al orden de etapa [58, Teorema 4.2]. Discutiremos algo más esta posibilidad de ganancia de orden en la siguiente sección considerando algunos ejemplos interesantes que aparecen en la literatura.

V.5. Ilustraciones numéricas

En esta sección final del capítulo, presentamos aplicaciones de los resultados establecidos en las secciones previas a Métodos Lineales Generales de interés práctico. Tal como comentamos en la nota V.3.4, muchos Métodos Lineales Generales G-estables de k pasos, con $k \geq 2$, poseen una matriz de estabilidad $R(z)$ (V.5) que verifica $\|R(\infty)\|_G = 1$. De este modo, no podemos deducir directamente contractividad estricta para estos métodos. Sin embargo, en virtud de los teoremas V.3.6 y V.4.3, aún podemos obtener cotas de contractividad y convergencia para estos métodos si el radio espectral de la matriz de estabilidad es menor que uno. A modo de ilustración consideramos en este punto algunos ejemplos en la clase de métodos Runge-Kutta multipaso, que toman la forma de Métodos Lineales Generales (V.4) con

$$A = \begin{pmatrix} \gamma_1 & \dots & \gamma_s \\ & O & \end{pmatrix}, \qquad B = \begin{pmatrix} \alpha_1 & \dots & \alpha_{k-1} & \alpha_k \\ & I_{k-1} & & O \end{pmatrix},$$

donde $\gamma_1, \dots, \gamma_s, \alpha_1, \dots, \alpha_k \in \mathbb{R}$. Además, debido a la preconsistencia de los métodos, asumiremos que $\alpha_1 + \ldots + \alpha_k = 1$.

El método BDF2

El método a dos pasos clásico $BDF2$ con tamaño de paso fijo $h > 0$, definido como

$$\frac{3}{2}y_{n+1} - 2y_n + \frac{1}{2}y_{n-1} = hf(t_{n+1}, y_{n+1}), \quad n \geq 1,$$

toma la forma de un Método Lineal General de dos pasos y una etapa (V.4), con $c_1 = 1$, y coeficientes matriciales dados por

$$\tilde{A} = \begin{pmatrix} \frac{2}{3} \end{pmatrix} \quad \tilde{B} = \begin{pmatrix} \frac{4}{3} & -\frac{1}{3} \end{pmatrix} \quad A = \begin{pmatrix} \frac{2}{3} \\ 0 \end{pmatrix} \quad B = \begin{pmatrix} \frac{4}{3} & -\frac{1}{3} \\ 1 & 0 \end{pmatrix}.$$

Este método es G-estable (véase, por ejemplo, [31], [53, Ejemplo 6.5, p. 308-309]) con matriz real, simétrica y definida positiva

$$G := \begin{pmatrix} \frac{5}{4} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{4} \end{pmatrix}.$$

En particular, la matriz de estabilidad (V.5) para el $BDF2$, esto es,

$$R(z) = \begin{pmatrix} \frac{4}{3}r(z) & -\frac{1}{3}r(z) \\ 1 & 0 \end{pmatrix}, \quad \text{con } r(z) = \frac{1}{1 - \frac{2}{3}z},$$

satisface $\|R(z)\|_G \leq 1,\ \forall z \in \mathbb{C}^-$, y $\|R(\infty)\|_G = 1$. En consecuencia, no podemos aplicar el teorema V.3.3 para deducir deducir contractividad estricta para el método $BDF2$, a pesar de que el método es estrictamente contractivo como veremos a continuación. Para ello, consideramos el método compuesto tras dos pasos consecutivos del método $BFD2$. Podemos comprobar fácilmente que $\|R(z)^2\|_G \leq 1,\ \forall z \in \mathbb{C}^-$, y que $\|R(\infty)^2\|_G = 0$. Entonces, de acuerdo al teorema V.3.6, el método compuesto definido por dos pasos consecutivos de la fórmula $BDF2$ define un método estrictamente contractivo en la clase de sistemas diferenciales (V.1)-(V.2). Además, considerando $D = \tilde{A} = (2/3)$, obtenemos de modo inmediato que $\alpha_D(\tilde{A}^{-1}) = 3/2$. Más aún, a partir del teorema V.4.3, este método alcanza orden de convergencia 2 en intervalos semi-infinitos para la clase de problemas estrictamente disipativos ($\nu < 0$), así como orden de convergencia 2 en la clase de problemas disipativos en intervalos finitos.

Métodos RK-Gauss multipaso.

Por otro lado, K. Burrage [6, 7] introdujo el siguiente conjunto de condiciones simplificadoras con vistas a la determinación de métodos Runge-Kutta multipaso algebraicamente estables de

alto orden

$$\begin{aligned}
B(p): \quad \bar{b}_p &:= q\sum_{j=1}^{s}\gamma_j c_j^{q-1} + \sum_{j=1}^{k}\alpha_j(1-j)^q - 1 = 0, \qquad 1\le q\le p,\\
C(p): \quad \bar{c}_p &:= q\sum_{j=1}^{s}\tilde{a}_{ij} c_j^{q-1} + \sum_{j=1}^{k}\tilde{b}_{ij}(1-j)^q - c_i^q = 0, \quad 1\le q\le p,\ \forall i,\\
D(p): \quad \bar{d}_p &:= q\sum_{i=1}^{s}\gamma_i c_i^{q-1}\tilde{b}_{ij} - \alpha_j(1-(1-j)^q) = 0, \qquad 1\le q\le p,\ \forall j,\\
E(p): \quad \bar{e}_p &:= q\sum_{i=1}^{s}\gamma_i c_i^{q-1}\tilde{a}_{ij} - \gamma_j(1-c_j^q) = 0, \qquad 1\le q\le p,\ \forall j.
\end{aligned} \tag{V.42}$$

De este modo, la familia de métodos Runge-Kutta-Gauss de k pasos y s etapas, con orden de etapa $q = s$ y orden de consistencia $2s$, se obtiene imponiendo las condiciones $B(s)$, $C(s)$, $D(s)$ y $E(s)$. Esta familia de métodos, con parámetros $\alpha_1 \ge 0, \ldots, \alpha_{k-1} \ge 0$, $\alpha_k > 0$, $\alpha_1 + \ldots \alpha_k = 1$, provee métodos G-estables, siendo $G := Diag(1, \alpha_2 + \ldots \alpha_k, \ldots, \alpha_{k-1} + \alpha_k, \alpha_k)$ (véase, por ejemplo, el teorema 9.15 [53, p. 367]). Si $R(z)$ denota la matriz de estabilidad (V.5) del correspondiente método Runge-Kutta-Gauss de k pasos y s etapas, entonces, teniendo en cuenta las condiciones simplificadoras $D(s)$ y $E(s)$, no es complicado ver que

$$R(\infty) = \begin{pmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_{k-1} & \alpha_k \\ 1 & 0 & \ldots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} - \begin{pmatrix} 1 & \ldots & 1 \\ 0 & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} 1-c_1 & \ldots & 1-c_s \\ \vdots & & \vdots \\ 1-c_1^s & \ldots & 1-c_s^s \end{pmatrix}^{-1} \Psi(\alpha_1, \ldots, \alpha_k), \tag{V.43}$$

donde $\Psi(\alpha_1, \ldots, \alpha_k) = (\psi_{ij})$ posee por coeficientes $\psi_{ij} := (1-(1-j)^i)\alpha_j$, $1 \le i \le s$, $1 \le j \le k$, mientras que los nodos $c_1, \ldots, c_s$ son reales, dependen de $k-1$ parametros, y pertenecen al intervalo $(1-k, 1)$. De hecho, estos nodos $c_1, \ldots, c_s$ están unívocamente determinados en términos de los k_1 parámetros por la condición $B(2s)$ (véase, por ejemplo, [53, Lema 9.11]), esto es,

$$\sum_{j=1}^{k}\alpha_j \int_{1-j}^{1} \pi(x) x^{q-1} dx = 0, \quad 1 \le q \le s,$$

donde $\pi(x) = (x-c_1)\cdot \ldots \cdot (x-c_s)$ representa el polinomio nodal.

En particular, para la familia uniparamétrica de métodos Runge-Kutta-Gauss de dos pasos y dos etapas, con parámetro $\alpha_2 \equiv \alpha \in (0, 1]$ ($\alpha_1 = 1-\alpha$), los nodos c_1 y c_2 se calculan a partir

de la fórmula

$$c_1 = \frac{3-3\alpha^2-\sqrt{3}\sqrt{1+40\alpha+942\alpha^2+40\alpha^3+\alpha^4}}{6+84\alpha+6\alpha^2},$$

$$c_2 = \frac{3-3\alpha^2+\sqrt{3}\sqrt{1+40\alpha+942\alpha^2+40\alpha^3+\alpha^4}}{6+84\alpha+6\alpha^2}.$$

A partir de $B(2)$, $D(2)$ y $C(2)$ obtenemos respectivamente que

$$\gamma_1 = \frac{(1-2c_2)-\alpha(1+2c_2)}{2(c_1-c_2)} \qquad \gamma_2 = \frac{(2c_1-1)+\alpha(1+2c_1)}{2(c_1-c_2)}, \tag{V.44}$$

$$\widetilde{B} = D^{-1}\begin{pmatrix}1 & 1\\ 2c_1 & 2c_2\end{pmatrix}^{-1}\begin{pmatrix}\alpha_1 & 2\alpha_2\\ \alpha_1 & 0\end{pmatrix}, \tag{V.45}$$

$$\widetilde{A} = \left\{\begin{pmatrix}c_1 & c_2\\ c_1^2 & c_2^2\end{pmatrix} - \widetilde{B}\begin{pmatrix}0 & 0\\ -1 & 1\end{pmatrix}\right\}\begin{pmatrix}1 & 2c_1\\ 1 & 2c_2\end{pmatrix}^{-1}, \tag{V.46}$$

donde $D := Diag(\gamma_1, \gamma_2)$. Para cada $\alpha \in (0,1]$, tales métodos son G-estables y algebraicamente estables con matriz $G = Diag(1,\alpha)$ y matriz diagonal $D = Diag(\gamma_1,\gamma_2)$. Notemos que si $\alpha \to 0^+$ entonces se obtiene el método clásico de Gauss de un paso para avanzar la integración desde t_n a t_{n+1} con tamaño de paso fijo $h > 0$; mientras que para $\alpha \to 1^-$ se deduce el método clásico de Gauss de un paso para avanzar la integración desde t_{n-1} a t_{n+1} con tamaño de paso fijo $2h$.

Partiendo de (V.43), se obtiene que el valor $\rho(R(\infty)) = \sqrt{2}-1 = 0.4142\ldots$ es mínimo para $\hat{\alpha} = 17-12\sqrt{2} = 0.0294\ldots$. Más aún, se comprueba fácilmente que $\|R(z)\|_G \leq 1,\ \forall z \in \mathbb{C}^-$, y que $\|R(\infty)\|_G = 1,\ \forall \alpha \in (0,1]$. No obstante lo anterior, un cálculo sencillo muestra que $\|R(\infty)^2\|_G < 1$ se verifica para cada $\alpha \in (0,1)$, mientras que el valor $\|R(\infty)^2\|_G = 3-2\sqrt{2} = 0.1715\ldots$ es mínimo para $\hat{\alpha} = 17-12\sqrt{2}$. Por lo tanto, en virtud del teorema V.3.6, la composición en dos pasos consecutivos del correspondiente método RK-Gauss de dos pasos y dos etapas con $\alpha \in (0,1)$, provee métodos estrictamente contractivos.

Cálculos numéricos con la ayuda de Mathematica [90] permiten mostrar que, para $D = Diag(\gamma_1,\gamma_2)$, tenemos que $\alpha_D(\tilde{A}^{-1}) > 0$, cuando $\alpha \in (0,1)$. Por otro lado, para estos métodos la condición (V.41) es equivalente a

$$3\widetilde{A}c^2 + \widetilde{B}\begin{pmatrix}0\\ -1\end{pmatrix}^3 - c^3 = p\cdot e, \tag{V.47}$$

para algún $p \in \mathbb{R},\ p \neq 0$, y esto último no puede ser satisfecho si $\alpha \in (0, 1]$. De esta manera, se deduce orden de convergencia global 2 a partir del teorema V.4.3.

Cálculos numéricos adicionales revelan cotas de contractividad estricta y orden de convergencia global 3 para la familia uniparamétrica de métodos RK-Gauss de dos pasos y tres etapas, con parámetro $\alpha \in (0, 1)$.

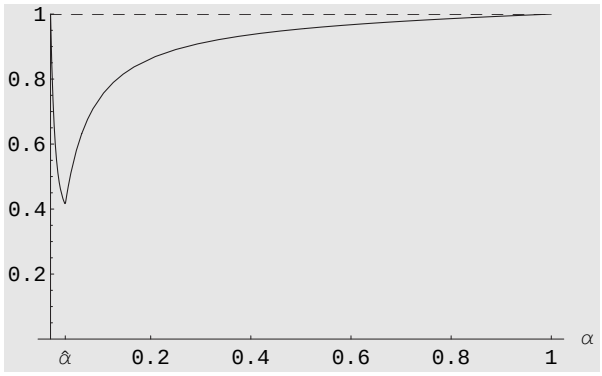

Figura V.1: Métodos RK-Gauss ($k = s = 2$): $\rho(R(\infty))$ (línea continua), $\|R(\infty)\|_G$ (línea discontinua).

Considerando las mismas condiciones simplificadoras introducidas por Burrage para la construcción de Métodos Lineales Generales algebraicamente estables de alto orden, S. Li [71] ha determinado seis clases de métodos Runge-Kutta multipaso de alto orden algebraicamente estables y B-convergentes Runge-Kutta. En particular, aquellos métodos que pertenecen a las clases 1-4 (véase [71, p. 1491]) pueden ser considerados respectivamente como generalizaciones de los métodos clásicos de un paso Runge-Kutta Gauss, RadauIA, RadauIIA y LobattoIIIC.

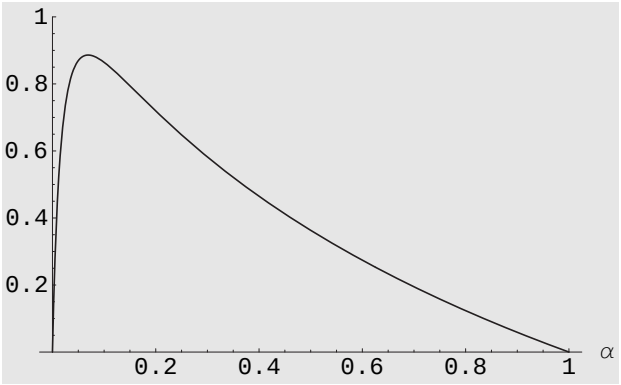

Figura V.2: Métodos RK-Gauss ($k = s = 2$): $\alpha_D(\tilde{A}^{-1})$.

Aquellos métodos que pertenecen a la clase 2 [71] poseen orden de etapa $q = s$ y orden de consistencia $p = 2s - 1$, se obtienen asumiendo $B(s)$, $C(s)$, $D(s)$ y $E(s-1)$, y la estabilidad algebraica se deduce en el caso de que $\bar{b}_{2s} > 0$ (véase (V.42)). Por otro lado, los métodos que pertenecen a la clase 3 [71] poseen orden de etapa $q = s-1$ y orden de consistencia $p = 2s-1$, se construyen asumiendo $B(s)$, $C(s-1)$, $D(s)$ y $E(s)$, y la estabilidad algebraica requiere que $\bar{b}_{2s} < 0$; mientras que aquellos métodos algebraicamente estables que pertenecen a la clase 4 [71] poseen orden de etapa $q = s-1$ y orden de consistencia $p = \max\{2s-2, s\}$ y se obtienen imponiendo $B(s)$, $C(s-1)$, $D(s)$ y $E(s-1)$.

Para el caso $k = 2$ y $s = 2$ hemos comprobado que aquellos resultados teóricos dados en los teoremas V.3.6 y V.4.3 también se aplican a determinadas subclases de métodos que a continuación especificamos. Notemos que para los métodos pertenecientes a las clases 1-4, los pesos γ_1 y γ_2 están unívocamente determinados por $B(2)$ en términos de los nodos c_1 y c_2 y del parámetro α; de este modo, la fórmula (V.44) se aplica no sólo a los métodos RK-Gauss sino también a los métodos pertenecientes a las clases 2-4. Más aún, la matriz $\widetilde{B}$ dada por (V.45) también se corresponde con la de los métodos pertenecientes a estas tres clases 2-4 en virtud de la condición $D(2)$. Por supuesto, asumimos en la derivación de los métodos que $c_1 \neq c_2$ y $\gamma_i \neq 0$, $i = 1, 2$.

Métodos tipo RK-RadauIIA multipaso.

En virtud de $B(3)$, para aquellos métodos de dos pasos y dos etapas que pertenecen a la clase 2 se tiene que

$$\sum_{j=1}^{2} \alpha_j \int_{1-j}^{1} \pi(x)dx = 0.$$

Por este motivo, uno de los nodos es libre y esta clase de métodos depende de dos parámetros $\alpha \in (0, 1]$ y u, que representa uno de los nodos. Imponiendo $c_1 \neq c_2$ entonces se tiene que la matriz de coeficientes $\widetilde{A}$ está unívocamente determinada por (V.46) en virtud de $C(2)$.

Fijando el nodo $c_2 = 1$ obtenemos una subclase uniparamétrica de métodos pertenecientes a la clase 2, con parámetro $\alpha \in [0, 1]$ donde $c_1 = \frac{(1-5\alpha)}{3(1+3\alpha)}$, y los pesos $\gamma_1 = \frac{3(1+3\alpha)^2}{4(1+7\alpha)}$, $\gamma_2 = \frac{1+14\alpha+\alpha^2}{4(1+7\alpha)}$ son positivos. Ya que

$$\bar{b}_4 = 4(\gamma_1 c_1^3 + \gamma_2 c_2^3) + \alpha - 1 = \frac{1 + 50\alpha + 13\alpha^2}{9(1+3\alpha)} > 0$$

obtenemos estabilidad algebraica, con $G = Diag(1, \alpha)$ y $D = Diag(\gamma_1, \gamma_2)$ para cada $\alpha \in (0, 1]$. En particular, para $\alpha \to 0$ se deduce el método RK-RadauIIA clásico de un paso con paso

$h > 0$ para avanzar la integración desde t_n a t_{n+1}; mientras que si $\alpha = 1$, entonces se obtiene el método RK-RadauIIA con paso $2h$ para avanzar al integración desde t_{n-1} a t_{n+1}.

La matriz de estabilidad (V.5) en el infinito para esta subclase de métodos posee un radio espectral menor que uno, y por tanto se deduce contractividad estricta para cada $\alpha \in (0, 1]$ en virtud del teorema V.3.6. Otros cálculos numéricos también muestran que, considerando $D = Diag(\gamma_1, \gamma_2)$, $\alpha_D(\tilde{A}^{-1}) > 0$, para cada valor del parámetro $\alpha \in (0, 1]$. Esto implica orden global de convergencia $q = 2$ para esta clase de métodos. Por otro lado, no existen métodos en esta familia verificando (V.47) y por tanto la condición (V.41) no se cumple y el teorema V.4.4 no puede ser aplicado para obtener un orden de convergencia adicional.

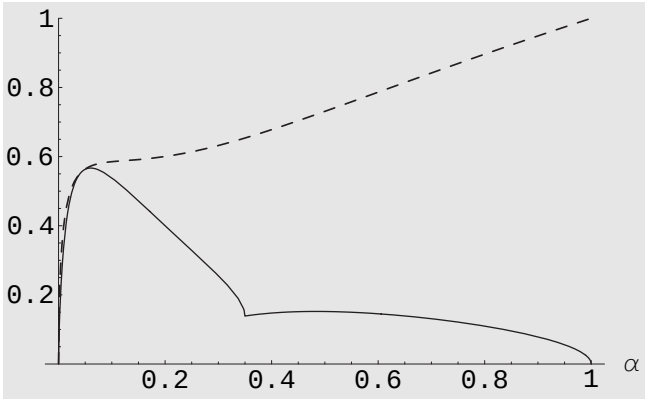

Figura V.3: Métodos RadauIIA ($k = s = 2$): $\rho(R(\infty))$ (línea continua), $\|R(\infty)\|_G$ (línea discontinua).

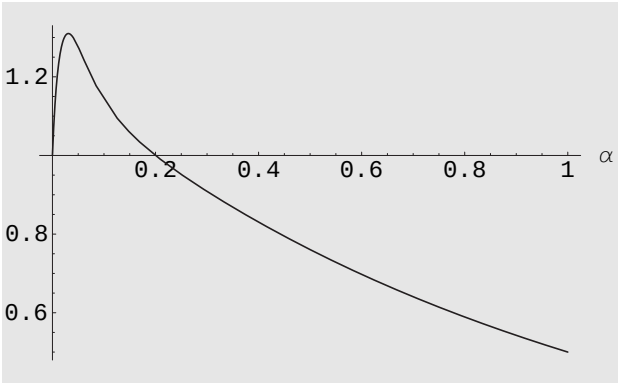

Figura V.4: Métodos RadauIIA ($k = s = 2$): $\alpha_D(\tilde{A}^{-1})$.

Métodos tipo RK-RadauIA multipaso.

Para construir los métodos pertenecientes a la clase biparamétrica 3 hacen consideraciones análogas al caso de los métodos tipo RadauIIA. Los pesos y la matriz de coeficientes $\widetilde{B}$ están unívocamente determinados en términos de los nodos c_1, c_2, y el parámetro α en virtud de $B(2)$ y $D(2)$, si se asume que $\gamma_i \neq 0,\ i = 1, 2$, y $c_1 \neq c_2$. En este caso, la matriz de coeficientes $\widetilde{A}$ está unívocamente determinada por $E(2)$, esto es,

$$\widetilde{A} = D^{-1} \begin{pmatrix} 1 & 1 \\ 2c_1 & 2c_2 \end{pmatrix}^{-1} \begin{pmatrix} 1-c_1 & 1-c_2 \\ 1-c_1^2 & 1-c_2^2 \end{pmatrix} D,$$

donde $D = Diag(\gamma_1, \gamma_2)$. Por otro lado, como en el caso de los métodos tipo RadauIIA multipaso, un nodo permanece libre en virtud de $B(3)$. De este modo, fijando el nodo $c_1 = -1$, obtenemos una subclase uniparamétrica de métodos pertenecientes a la clase 3 [71], con parámetro $\alpha \in [0, 1]$ donde $c_2 = \frac{(5-\alpha)}{3(3+\alpha)}$, y los pesos $\gamma_1 = \frac{1+14\alpha+\alpha^2}{4(1+7\alpha)},\ \gamma_2 = \frac{3(3+\alpha)^2}{4(1+7\alpha)}$ son positivos. Ya que

$$\bar{b}_4 = 4(\gamma_1 c_1^3 + \gamma_2 c_2^3) + \alpha - 1 = -\frac{13 + 50\alpha + \alpha^2}{9(3+\alpha)} < 0$$

obtenemos estabilidad algebraica, con $G = Diag(1, \alpha)$ y $D = Diag(\gamma_1, \gamma_2)$ para cada $\alpha \in (0, 1]$. En particular, para $\alpha \to 1$ se deduce el método clásico de un paso RadauIA con tamaño de paso $2h$ para avanzar la integración desde t_{n-1} a t_{n+1}.

La matriz de estabilidad (V.5) en el infinito para esta subclase de métodos posee radio espectral menor que uno y por tanto se deduce contractividad estricta para todo $\alpha \in (0, 1]$ en virtud del teorema V.3.6. Además, considerando $D = Diag(\gamma_1, \gamma_2)$, tenemos que $\alpha_D(\tilde{A}^{-1}) > 0$, para cada valor del parámetro $\alpha \in (0, 1]$. Esto implica orden de convergencia global $q = 1$ para esta subclase de métodos. Por otro lado, no es complicado ver que no existen métodos en esta subfamilia verificando la condición (V.41), y por tanto el teorema V.4.4 no puede ser aplicado para ganar un orden de convergencia adicional. Notemos que, en este caso, la condición (V.41) es equivalente a

$$2\widetilde{A}c + \widetilde{B} \begin{pmatrix} 0 \\ -1 \end{pmatrix}^2 - c^2 = p \cdot e. \tag{V.48}$$

Métodos tipo RK-LobattoIIIC multipaso.

Los métodos de tipo LobattoIIIC se deducen imponiendo las condiciones $B(2)$, $C(1)$, $D(2)$ y $E(1)$. En general, estas condiciones proveen una familia de métodos dependientes de cuatro parámetros: $\alpha \equiv \alpha_2$, los nodos c_1 y c_2, y una componente de la matriz $\widetilde{A}$. Fijando los nodos

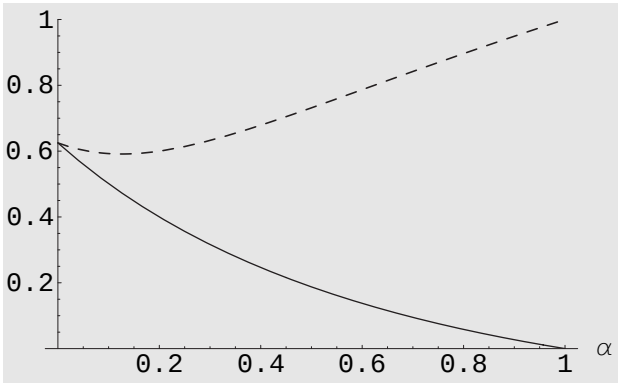

Figura V.5: Métodos RadauIA ($k = s = 2$): $\rho(R(\infty))$ (línea continua), $\|R(\infty)\|_G$ (línea discontinua).

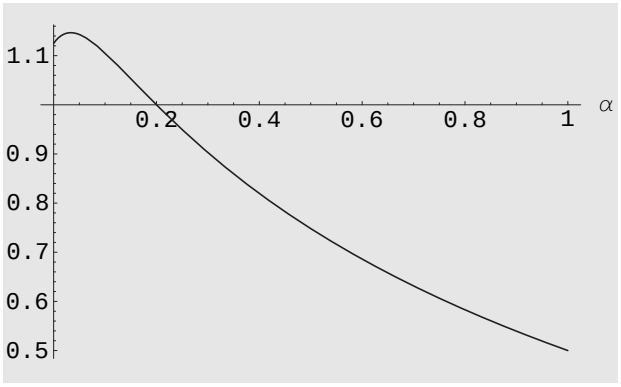

Figura V.6: Métodos RadauIA ($k = s = 2$): $\alpha_D(\tilde{A}^{-1})$.

$c_1 = -1$ y $c_2 = 1$ obtenemos de $B(2)$ que $\gamma_1 = \frac{1+3\alpha}{4}$, $\gamma_2 = \frac{3+\alpha}{4}$. Además, $\widetilde{B}$ está unívocamente determinada en términos de α a consecuencia de $D(2)$. Entonces, partiendo de $C(1)$ y $E(1)$ obtenemos para $\widetilde{A} = (\widetilde{a}_{ij})$ que

$$\begin{aligned} \widetilde{a}_{11} &= \tfrac{(\alpha-1)+3\alpha\widetilde{a}_{22}}{1+3\alpha}, \\ \widetilde{a}_{21} &= -\tfrac{3+\alpha}{1+3\alpha}\widetilde{a}_{22}, \\ \widetilde{a}_{21} &= \tfrac{3+5\alpha}{3+\alpha} - \widetilde{a}_{22}. \end{aligned}$$

Notemos que para $\alpha = 1$ y $\widetilde{a}_{22} = 1$ se obtiene el método de un paso clásico LobattoIIIC con tamaño de paso $2h$ para avanzar la integración de t_{n-1} a t_{n+1}. Fijemos entonces $\widetilde{a}_{22} = 1$ y consideremos la subclase uniparamétrica así obtenida de métodos pertenecientes a la clase 4, con paramétro $\alpha \in [0, 1]$.

Para esta subclase de métodos se obtiene estabilidad algebraica con las matrices $G = Diag(1, \alpha)$ y $D = Diag(\gamma_1, \gamma_2)$, para todos los valores del parámetro $\alpha \in (0, 1]$, puesto que

la matriz (V.10) posee dos autovalores nulos y otros dos autovalores no negativos dependientes de α. De nuevo, la matriz de estabilidad para estos métodos (V.5) evaluada en el infinito posee radio espectral menor que uno y se deduce contractividad estricta para esta subclase uniparamétrica de métodos para cada $\alpha \in (0, 1]$ a consecuencia del teorema V.3.6. Por otro lado, diversos cálculos numéricos muestran que considerando $D = Diag(\gamma_1, \gamma_2)$ entonces $\alpha_D(\tilde{A}^{-1}) > 0$, para cada valor del parámetro $\alpha \in (0, 1]$. Esto implica orden de convergencia global $q = 1$ para estos métodos. Más aún, no existen métodos en esta subclase satisfaciendo la condición (V.41), que es equivalente a (V.48), y por tanto el teorema V.4.4 no puede ser aplicado con vistas a ganar un orden de convergencia adicional.

Finalmente debemos comentar que los resultados de contractividad y convergencia se aplican directamente a los métodos multipaso RadauIIA, RadauIA y LobattoIIIC originales pertenecientes a las subclases arriba indicadas, y que no hay necesidad de considerar la composición de los métodos en pasos consecutivos de cara a obtener contractividad. Esto sigue del hecho de que $\|R(\infty)\|_G < 1$, $\alpha \in (0, 1)$, siendo R la matriz de estabilidad (V.5). Un resumen acerca de la determinación de las familias de métodos consideradas en esta sección, así como de sus propiedades de contractividad y convergencia en intervalos semi-infinitos puede verse en la tabla V.1.

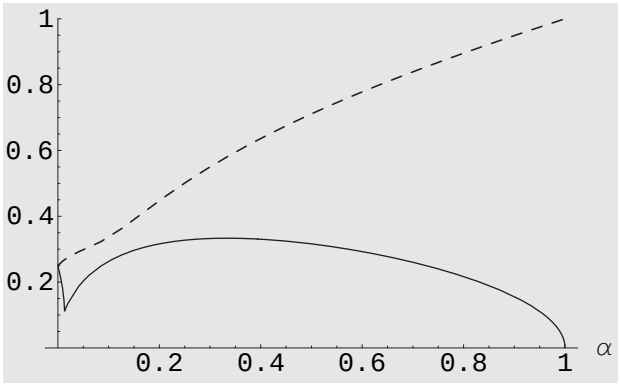

Figura V.7: Métodos LobattoIIIC ($k = s = 2$): $\rho(R(\infty))$ (línea continua), $\|R(\infty)\|_G$ (línea discontinua).

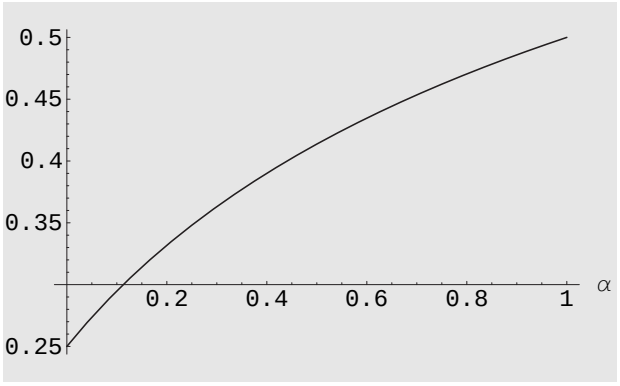

Figura V.8: Métodos LobattoIIIC ($k = s = 2$): $\alpha_D(\tilde{A}^{-1})$.

	$Gauss(\alpha)$
$\|R(\infty)\|_G < 1$	Nunca
$\|R(\infty)^2\|_G < 1$	$0 < \alpha < 1$
$\alpha_D(\widetilde{A}^{-1}) > 0$	$0 < \alpha < 1$
$\rho(R(\infty)) = 1,\ \alpha_D(\widetilde{A}^{-1}) = 0$	$\alpha = 1$
$Condiciones\ simplificadoras$	$B(2),\ C(2),\ D(2),\ E(2)$
$Convergencia\ en\ [0,\infty)$	$q = 2$ (solamente para $0 < \alpha < 1$)
	$RadauIIA(\alpha)$
$\|R(\infty)\|_G < 1$	$0 < \alpha < 1$
$\|R(\infty)^2\|_G < 1$	$0 < \alpha \leq 1$
$\alpha_D(\widetilde{A}^{-1}) > 0$	$0 < \alpha \leq 1$
$Condiciones\ simplificadoras$	$B(2),\ C(2),\ D(2),\ E(1),\ c_2 = 1$
$Convergencia\ en\ [0,\infty)$	$q = 2\ (0 < \alpha \leq 1)$
	$RadauIA(\alpha)$
$\|R(\infty)\|_G < 1$	$0 < \alpha < 1$
$\|R(\infty)^2\|_G < 1$	$0 < \alpha \leq 1$
$\alpha_D(\widetilde{A}^{-1}) > 0$	$0 < \alpha \leq 1$
$Condiciones\ simplificadoras$	$B(2),\ C(1),\ D(2),\ E(2),\ c_1 = -1$
$Convergencia\ en\ [0,\infty)$	$q = 1\ (0 < \alpha \leq 1)$
	$LobattoIIIC(\alpha)$
$\|R(\infty)\|_G < 1$	$0 < \alpha < 1$
$\|R(\infty)^2\|_G < 1$	$0 < \alpha \leq 1$
$\alpha_D(\widetilde{A}^{-1}) > 0$	$0 < \alpha \leq 1$
$Condiciones\ simplificadoras$	$B(2),\ C(1),\ D(2),\ E(1),\ c_{1,2} = -1, 1,\ \tilde{a}_{22} = 1$
$Convergencia\ en\ [0,\infty)$	$q = 1\ (0 < \alpha \leq 1)$

Tabla V.1: Algunas características de los métodos Runge-Kutta multipaso algebraicamente estables de dos pasos y dos etapas denotados como $Gauss(\alpha)$, $RadauIIA(\alpha)$, $RadauIA(\alpha)$ y $LobattoIIIC(\alpha)$, $0 < \alpha \leq 1$.

Capítulo VI

Conservación de invariantes por medio de métodos Runge-Kutta explícitos.

VI.1. Consideraciones preliminares

Las integrales primeras e invariantes de sistemas diferenciales autónomos

$$y' = f(y), \tag{VI.1}$$

donde $f : \mathcal{D} \subset \mathbb{R}^m \to \mathbb{R}^m$ es una función suficientemente regular, juegan un papel muy importante en el estudio tanto cualitativo como cuantitativo del flujo del problema diferencial (VI.1). Estos invariantes permiten describir la geometría de las órbitas, así como medir la precisión de los integradores numéricos de la correspondiente ecuación diferencial.

En este sentido, una función $G = G(y) : \widehat{\mathcal{D}} \subset \mathbb{R}^m \to \mathbb{R}^l$ de clase $\mathcal{C}^1(\widehat{\mathcal{D}})$, $\widehat{\mathcal{D}} \subset \mathcal{D}$, se dice un *sistema l-invariante* de (VI.1) en $\widehat{\mathcal{D}}$ (véase [89, p. 61]) si G verifica las dos siguientes condiciones

i) existe algún $y^0 \in \widehat{\mathcal{D}}$ tal que $G(y^0) = 0$;

ii) para toda solución $y = y(t)$ de (VI.1), o bien $G(y(t)) = 0$ se cumple para cada t en el intervalo de definición de $y(t)$, o bien no se cumple para ningún valor de t en dicho intervalo.

En tal caso, podemos considerar el flujo definido por (VI.1) restringido a la variedad $(m-l)$-dimensional $\mathcal{M} = \{y \in \widehat{\mathcal{D}} / \ G(y) = 0\}$ para obtener una ecuación diferencial sobre la variedad

$\mathcal{M}$. En el caso $l = 1$, el sistema invariante se denomina una *relación invariante*. En relación con el concepto anterior, merece la pena notar que para $l > 1$ las l relaciones escalares que constituyen el sistema invariante no necesariamente tienen porqué ser relaciones invariantes en el sentido previamente definido. Un sistema l-invariante se suele denominar también *invariante débil* (ver [50]), y no es difícil ver que $G(y)$ es un invariante débil de (VI.1) si y sólo si existe algún $y^0 \in \widehat{\mathcal{D}}$ tal que $G(y^0) = 0$, y $\nabla G(y)^T f(y) = 0$, para todo $y \in \mathcal{M}$.

Por otro lado, una función escalar $F \subset \widehat{\mathcal{D}} : \mathbb{R}^m \to \mathbb{R}$ de clase $C^1(\widehat{\mathcal{D}}), \widehat{\mathcal{D}} \subset \mathcal{D}$, es una *integral primera* (o una *cantidad conservada*) de (VI.1) si $F(y(t))$ es una constante a lo largo de cualquier solución $y = y(t)$ de (VI.1). En tal caso, ya que

$$0 = \frac{dF(y(t))}{dt} = \nabla F(y(t))^T \cdot f(y(t)), \qquad \text{para cada } y(t),$$

sigue que $\nabla F(y)^T \cdot f(y) = 0$, para todo $y \in \widehat{\mathcal{D}}$. Además, para todo $y^0 \in \widehat{\mathcal{D}}$, la solución $y = y(t)$ de (VI.1) tal que $y(0) = y^0$ está contenida en la hipersuperficie $\mathcal{M}_{y^0} = \{y \in R^m \ / \ F(y) = F(y^0)\}$. Las integrales primeras también son conocidas en la literatura como *invariantes fuertes* (véase [50, 60]), y en particular, las integrales cuadráticas $G(y) = y^T S y$, donde $S \in \mathbb{R}^{m \times m}$ es una matriz simétrica constante, son invariantes fuertes de (VI.1) si y sólo si $y^T \ S \ f(y) = 0$, para cada $y \in \mathcal{D}$.

Para simplificar la presentación, usaremos en el desarrollo de este capítulo el término *invariante* del sistema diferencial (VI.1) indistintamente para referirnos a una integral primera o a un sistema (o una relación) invariante de (VI.1).

Puesto que gran parte de los integradores numéricos generales no conservan por sí mismos las integrales primeras o invariantes de los sistemas diferenciales a los que son aplicados, estas funciones han sido frecuentemente consideradas como herramientas útiles de cara a medir la precisión y comportamiento en amplios intervalos temporales de los integradores numéricos. De hecho, en las últimas dos décadas ha surgido un interés creciente por los integradores numéricos que preservan, en intervalos temporales lo más amplios posibles, las propiedades cualitativas de los sistemas diferenciales en consideración. En particular, un alternativa natural de proceder consiste en considerar para la integración de sistemas diferenciales (VI.1) aquellos métodos numéricos de tipo Runge-Kutta que preserven integrales primeras o invariantes.

En este sentido, Cooper [27] ha demostrado que todo método de tipo Runge-Kutta conserva invariantes lineales; sin embargo, una condición necesaria y suficiente para que un método

Runge-Kutta irreducible conserve todos los invariantes cuadráticos $G(y) = y^T Sy + \mu^T y + \nu$ con constantes $S \in \mathbb{R}^{m\times m}, \mu \in \mathbb{R}^m, \nu \in \mathbb{R}$ es que sus coeficientes ($A \in \mathbb{R}^{s\times s}, b \in \mathbb{R}^s$) satisfagan $m_{ij} \equiv b_i a_{ij} + b_j a_{ji} - b_i b_j = 0$, $1 \leq i \leq j \leq s$. Estas condiciones, que también implican el método Runge-Kutta $RK(A, b)$ es un método *simpléctico*, imponen una condición tan exigente sobre los coeficientes de un método Runge-Kutta que tan sólo algunos pocos métodos de carácter implícito son capaces de satisfacerla. Más aún, se demuestra que (véase, por ejemplo, [50, Teorema 3.3, p. 102]) que para $n \geq 3$ ningún método de tipo Runge-Kutta puede preservar todos los invariantes polinomiales de grado n. Diversos estudios adicionales sobre la conservación de invariantes algebraicos por medio de métodos de Runge-Kutta han sido llevados a cabo por Iserles y Zanna [60].

Una opción alternativa para preservar invariantes en las integraciones numéricas ha sido propuesta por Schropp [80]. Este autor modifica el flujo original de (VI.1) por medio de una técnica de estabilización de la variedad invariante $\mathcal{M}_{y^0} = \{y \in \mathbb{R}^m \,/\, G(y) = G(y^0)\}$, y considera el problema modificado

$$y' = f(y) - \nabla G(y)\; a(y)(G(y) - G(y^0)),$$

donde $a(y)$ es una función regular adecuadamente elegida de tal modo que la variedad $\mathcal{M}_{y^0}$ resulte ser exponencialmente atractora para las órbitas del problema modificado. Con esta idea en mente, se espera que las soluciones numéricas generadas por un método numérico dado se aproximen a este conjunto atractor; sin embargo, desde una perspectiva práctica, el problema modificado puede adoptar una stiffness severa según la elección de la función $a(y)$.

A la vista de las limitaciones mencionadas se han propuesto diversas técnicas para la conservación de invariantes con métodos numéricos prácticos. De este modo, Del Buono y Mastroserio en [34], partiendo de un método Runge-Kutta explícito de s etapas, con coeficientes dados por la matriz $A = (a_{ij})$ y el vector $b = (b_i)$, cuya solución de avance se escribe como

$$\varphi_h(y_n) = y_n + h \sum_{j=1}^{s} b_j f(Y_{n,j}), \tag{VI.2}$$

y donde las etapas se calculan explícitamente a través de una fórmula del tipo

$$Y_{n,1} = y_n, \quad Y_{n,i} = y_n + h \sum_{j=1}^{i-1} a_{ij} f(Y_{n,j}), \quad 2 \leq i \leq s, \tag{VI.3}$$

modifican la solución de avance (VI.2) del método por medio de la introducción en cada paso temporal de un parámetro escalar $\gamma_n = \gamma_n(y_n, h)$ que depende de la solución numérica computada en el paso temporal t_n y del tamaño de paso propuesto h, de tal modo que el método resultante $\widehat{\varphi}_h$ definido por (VI.3) y

$$\widehat{\varphi}_h(y_n) = y_n + h\gamma_n \sum_{j=1}^{s} b_j f(Y_{n,j}), \tag{VI.4}$$

conserve un invariante dado de (VI.1). Claramente, esto implica que γ_n será un escalar dependiente del problema tal que $\gamma_n \to 1$ cuando $h \to 0$ y que puede variar de un paso temporal a otro. Estos autores [34] han demostrado que para todos los sistemas diferenciales (VI.1) que poseen un invariante cuadrático $G(y) = y^T S y$ con matriz simétrica S, existen métodos Runge-Kutta explícitos de cuarto orden y cuatro etapas φ_h tales que los métodos modificados $\widehat{\varphi}_h$ dados por (VI.3) y (VI.4) conservan el invariante cuadrático y poseen orden tres para una elección apropiada en cada paso del parámetro γ_n. Además $\widehat{\varphi}_h$ alcanza orden de consistencia cuatro en la red temporal $t_n + h\gamma_n$, y en consecuencia $\widehat{\varphi}_h$ obtiene orden global de convergencia cuatro en la red temporal no uniforme $\{\widehat{t}_n\}_{n=0}^N$ con $\widehat{t}_0 = t_0$ y $\widehat{t}_{n+1} = \widehat{t}_n + h\gamma_n$.

Para el cómputo de γ_n en cada paso, Del Buono y Mastroserio [34] consideran la fórmula $\gamma_n = 1 - (\delta_n/\eta_n)$ con

$$\eta_n = \sum_{i,j=1}^{4} b_i b_j \langle f(Y_{n,i}), f(Y_{n,j})\rangle, \quad \delta_n = \eta_n - 2\sum_{i=2}^{4}\sum_{j=1}^{i-1} b_i a_{ij} \langle f(Y_{n,i}), f(Y_{n,j})\rangle,$$

donde $\langle u, v\rangle \equiv u^T S v$ para $u, v \in \mathbb{R}^m$, es la forma bilineal asociada al invariante $G(y) = y^T S y$. Esto significa que el costo computacional de $\widehat{\varphi}_h$ en cada paso incluye $s(s-1)/2 = 6$ productos escalar $\langle f(Y_{n,i}), f(Y_{n,j})\rangle$, aparte de las evaluaciones de funciones involucradas en (VI.3). Notemos que $\widehat{\varphi}_h(y_n) = \varphi_h(y_n) + (\gamma_n - 1)(\varphi_h(y_n) - y_n)$ puede ser considerada como la proyección de $\varphi_h(y_n)$ sobre la variedad invariante $\mathcal{M}_{y_n} = \{y \,/\, G(y) = G(y_n)\}$ a lo largo de la dirección $\varphi_h(y_n) - y_n$. En lo que resta de este capítulo esta técnica se denominará *técnica de dirección incremental*.

Una técnica alternativa (véase [50, p. 106]), conocida como *método de proyección estándar*, consiste en combinar un método de tipo Runge-Kutta estándar φ_h (o cualquier otro método de un paso) junto con una proyección ortogonal P sobre la variedad invariante $\mathcal{M}_{y_n}$ en cada paso $(t_n, y_n) \to (t_{n+1} = t_n + h, y_{n+1})$. Para el caso de la norma euclídea, el método estaría definido

por $y_{n+1} = (P \cdot \varphi_h)(y_n)$, siendo y_{n+1} definido por $y_{n+1} = \varphi_h(y_n) + \nabla G(y_{n+1})\ \lambda_n$, donde λ_n es un escalar real convenientemente elegido de modo que $G(y_{n+1}) = G(y_n)$. Debido a su carácter implícito, este sistema se suele reemplazar por (véase [50])

$$y_{n+1} = \varphi_h(y_n) + \nabla G(\varphi_h(y_n))\ \lambda_n, \quad \text{con} \quad G(y_{n+1}) = G(y_n), \tag{VI.5}$$

y el cálculo de λ_n en cada paso se lleva a cabo por medio de iteraciones de Newton simplificadas para la resolución aproximada de la correspondiente ecuación no lineal. Tal como se indica en [50], ya que λ_n es pequeño, la elección de $\lambda_{n,0} = 0$ como valor de arranque conlleva normalmente que una iteración del método de Newton simplificado sea suficiente en la práctica para obtener una precisión aceptable. De esta manera, $\widehat{\varphi}_h(y_n)$ se obtiene proyectando $\varphi_h(y_n)$ a lo largo de la dirección ortogonal a $G(y)$ en el punto $\varphi_h(y_n)$, y tal proyección mantiene el mismo orden de consistencia que el método original.

Tal como se infiere de (VI.4) y (VI.5), las técnicas de proyección mencionadas anteriormente pueden ser incluidas dentro de la forma general

$$\widehat{\varphi}_h(y_n) = \varphi_h(y_n) - \lambda_n \Phi_n, \tag{VI.6}$$

donde el vector $\Phi_n \in \mathbb{R}^m$ define la dirección de proyección y λ_n es un escalar convenientemente elegido de tal manera que $\widehat{\varphi}_h(y_n)$ pertenezca a la variedad invariante del problema diferencial en cuestión. En el caso de (VI.4), la dirección $\Phi_n = \varphi_h(y_n) - y_n$ viene definida a partir del método Runge-Kutta original y un método de orden cero, mientras que en el caso de la proyección (VI.5), Φ_n es ortogonal al invariante en el punto $\varphi_h(y_n)$. Claramente otras direcciones posibles Φ_n pueden ser consideradas, en particular, teniendo en cuenta que para muchos métodos explícitos se dispone normalmente, sin costo computacional adicional, de varias aproximaciones encajadas de orden menor del tipo

$$\widetilde{\varphi}_h(y_n) = y_n + h \sum_{j=1}^{s} \widetilde{b}_j f(Y_{n,j}). \tag{VI.7}$$

De esta manera, una posible vía de proceder consiste en considerar direcciones de la forma $\Phi_n = \varphi_h(y_n) - \widetilde{\varphi}_h(y_n)$, y por lo tanto el método proyectado (VI.6) toma la forma de un método de Runge-Kutta con pesos variables $\widehat{b}_j = (1 - \lambda_n) b_j + \lambda_n \widetilde{b}_j$,

$$\widehat{\varphi}_h(y_n) = y_n + h \sum_{j=1}^{s} ((1 - \lambda_n) b_j + \lambda_n \widetilde{b}_j) f(Y_{n,j}), \tag{VI.8}$$

y además retiene algunas propiedades de estos métodos explícitos, tales como un bajo costo computacional y la conservación invariantes lineales.

Nuestro objetivo en este capítulo será el estudio de los métodos de Runge-Kutta explícitos proyectados del tipo (VI.8) que conservan invariantes cuadráticos o generales. Hemos de añadir que el contenido de este capítulo se corresponde parcialmente con los resultados que se dan en los trabajos [17, 18]. En la sección segunda del capítulo llevaremos a cabo una revisión del caso $\widetilde{\varphi}_h(y_n) = y_n$, en el que el método encajado posee orden cero, con el propósito de generalizar los resultados dados en [34], así como para mostrar un modo más eficiente a la hora de computar el parámetro λ_n que define la proyección. En la sección 3, desarrollaremos un estudio general para invariantes cuadráticos y métodos encajados $\widetilde{\varphi}_h$ (VI.7) de orden q $(1 \leq q < p)$. A continuación, en la sección 4, generalizaremos este estudio a la conservación de invariantes unidimensionales generales, así como al caso de invariantes escalares múltiples. Trataremos además la manera de incluir la estrategia de conservación propuesta en códigos Runge-Kutta adaptativos. Concluiremos este capítulo en la sección 5 presentando algunos experimentos basados en el par encajado clásico de Dordmand y Prince de métodos Runge-Kutta explícitos de órdenes 5 y 4, DoPri5(4) [36], con la idea de dejar clara la fiabilidad de nuestra técnica en la integración de diversos problemas test.

VI.2. La técnica de dirección incremental revisada

En [34], Del Buono y Mastroserio proponen avanzar la integración $t_n \to t_{n+1} = t_n + h$ con el método Runge-Kutta $\widehat{\varphi}_h$ dado por (VI.4), donde γ_n se elige adecuadamente de tal manera que el método preserve un invariante cuadrático prefijado $G(y) = y^T\ S\ y = \langle y, y \rangle$. Aquí escribiremos $\widehat{\varphi}_h$ en la forma equivalente

$$\widehat{\varphi}_h(y_n) = \varphi_h(y_n) - \lambda_n(\varphi_h(y_n) - y_n) = (1 - \lambda_n)\varphi_h(y_n) + \lambda_n y_n, \tag{VI.9}$$

de tal modo que el parámetro escalar $\lambda_n = 1 - \gamma_n$ será una cantidad pequeña en cada paso de la integración.

Seguidamente mostraremos que el método proyectado $\widehat{\varphi}_h$ posee orden $p - 1$, para todo método Runge-Kutta φ_h con orden $p \geq 2$. Este resultado generaliza la Proposición 4 de [34] en dos sentidos: en primer lugar, se extiende el enunciado relativo a métodos de cuarto orden

a cualquier método de orden $p \geq 2$; en segundo lugar, en contraste con la situación en [34], no existen restricciones sobre los coeficientes del método φ_h original. Finalmente, mostraremos que para cualquier orden p, λ_n (o γ_n) puede ser calculado tan sólo necesitando dos evaluaciones de formas bilineales $\langle \cdot, \cdot \rangle$ por paso.

Con la notación (VI.9) y poniendo $\varphi_h(y_n) - y_n = h\Delta_n$, la conservación del invariante cuadrático $G(y) = y^T S y = \langle y, y \rangle$ se obtiene si y sólo si $\lambda_n \neq 1$ satisface la ecuación

$$2\langle y_n, h\Delta_n \rangle + (1 - \lambda_n)\langle h\Delta_n, h\Delta_n \rangle = 0. \tag{VI.10}$$

Si $\langle y_n, h\Delta_n \rangle = \langle h\Delta_n, h\Delta_n \rangle = 0$ entonces $\langle \widehat{\varphi}_h(y_n), \widehat{\varphi}_h(y_n) \rangle = \langle y_n, y_n \rangle$ para todo λ_n, y por lo tanto bastaría con elegir $\lambda_n = 0$ (esto es, elegir el método original φ_h) para conservar el invariante en el paso actual.

Si lo que se tiene es $\langle y_n, h\Delta_n \rangle = 0$, o bien $\langle h\Delta_n, h\Delta_n \rangle = 0$, entonces la ecuación (VI.10) no puede ser satisfecha por ningún $\lambda_n \neq 1$ (notemos que la elección $\lambda_n = 1$ nos conduce al método constante $\widehat{\varphi}_h(y_n) = y_n$), y en consecuencia no sería posible preservar el invariante. De aquí que la siguiente condición

$$\langle y_n, \Delta_n \rangle \, \langle \Delta_n, \Delta_n \rangle \neq 0, \tag{VI.11}$$

deba ser satisfecha de modo que (VI.10) sea resoluble. Observemos que si φ_h es el método de Euler explícito, entonces la condición (VI.11) no se verificaría por cuanto $\Delta_n = f(y_n)$. De esta manera, consideraremos en este apartado únicamente métodos con orden de consistencia $p \geq 2$.

Teorema VI.2.1 *Sea φ_h un método Runge-Kutta explícito con orden $p \geq 2$ y supongamos que (VI.11) se verifica.*

i) El método proyectado $\widehat{\varphi}_h$ dado por (VI.9) conserva el invariante $G(y) = \langle y, y \rangle$ a condición de que λ_n se elija como

$$\lambda_n = \frac{\langle 2y_n + h\Delta_n, h\Delta_n \rangle}{\langle h\Delta_n, h\Delta_n \rangle} = \frac{\langle \varphi_h(y_n), \varphi_h(y_n) \rangle - \langle y_n, y_n \rangle}{\langle \varphi_h(y_n) - y_n, \varphi_h(y_n) - y_n \rangle}. \tag{VI.12}$$

ii) Si $\langle f(y_n), f(y_n) \rangle \neq 0$ entonces el método proyectado $\widehat{\varphi}_h$ alcanza orden $p - 1$.

Demostración. Bajo la suposición (VI.11), la ecuación (VI.10) posee la solución λ_n dada por (VI.12), y por lo tanto el método $\widehat{\varphi}_h$ conserva el invariante cuadrático. Por otro lado, ya que φ_h tiene orden p se deduce que

$$\varphi_h(y_n) = y_n + h\Delta_n = y(t_n + h) + \epsilon_{n+1},$$

donde $\epsilon_{n+1} = C_{p+1}h^{p+1} + \mathcal{O}(h^{p+2})$ es el error local del método φ_h en el punto y_n con tamaño de paso h, e $y(t_n + t)$ denota la solución de (VI.1) en (t_n, y_n). En consecuencia, dicha solución satisface

$$\langle y(t_n + h), y(t_n + h)\rangle = \langle y(t_n), y(t_n)\rangle = \langle y_n, y_n\rangle,$$

para todo h. De aquí que para el numerador y denominador de λ_n en (VI.12) tenemos que

$$\begin{aligned} \langle \varphi_h(y_n), \varphi_h(y_n)\rangle - \langle y_n, y_n\rangle &= 2\langle y_n, C_{p+1}\rangle h^{p+1} + \mathcal{O}(h^{p+2}), \\ \langle h\Delta_n, h\Delta_n\rangle &= h^2\langle y'(t_n), y'(t_n)\rangle + \mathcal{O}(h^3), \end{aligned}$$

y en consecuencia λ_n se comporta como $O(h^{p-1})$ a condición de que

$$\langle y'(t_n), y'(t_n)\rangle = \langle f(y_n), f(y_n)\rangle \neq 0.$$

De acuerdo con (VI.9), esto implica que $\widehat{\varphi}_h$ es un método de orden $p - 1$. □

Para finalizar esta sección, observamos a partir de (VI.12) que el cómputo de λ_n requiere únicamente dos evaluaciones de formas bilineales por paso temporal en contraste con la situación en [34], que requiere de 6 formas bilineales por paso para el caso $s = 4$. Además, ya que las dos formas bilineales requeridas son $\langle \varphi_h(y_n), \varphi_h(y_n)\rangle, \langle \varphi_h(y_n), y_n\rangle$, únicamente se necesita un producto de matriz por vector adicional en cada paso.

VI.3. Una técnica de proyección general para la conservación de invariantes cuadráticos

Sea φ_h un método Runge-Kutta $RK(A, b)$ explícito de s etapas con orden p definido por las ecuaciones (VI.2)–(VI.3), y sea además $\widetilde{\varphi}_h$ un método encajado (VI.7) (esto es, un método explícito con la misma matriz de coeficientes que el método de partida) con orden q $(1 \leq q < p)$ definido por los pesos $\widetilde{b}_i,\ i = 1, \ldots, s$. Tal como se comentó anteriormente, para gran parte de los métodos de tipo Runge-Kutta explícitos de interés práctico existe la posibilidad de encajar otros métodos de orden q menor al del original $(1 \leq q \leq p - 1)$. De esta manera, tenemos bastante libertad en la elección del método encajado $\widetilde{\varphi}_h$, y, más aún, el cómputo de la solución de avance de este método no conlleva evaluaciones adicionales de funciones.

Supongamos ahora que (VI.1) posee un invariante cuadrático del tipo $G(y) = y^T Sy = \langle y, y\rangle$. Entonces, proponemos completar el paso en la integración de t_n a $t_{n+1} = t_n + h$ proyectando

la solución de avance del método original $\varphi_h(y_n)$ sobre la variedad definida por el invariante $\mathcal{M}_{y_n} = \{y \in \mathbb{R}^m / \langle y, y\rangle = \langle y_n, y_n\rangle\}$ a lo largo del vector unitario definido por el vector diferencia entre el método original y el método encajado

$$w_n = w(y_n, h) = \frac{\varphi_h(y_n) - \widetilde{\varphi}_h(y_n)}{\|\varphi_h(y_n) - \widetilde{\varphi}_h(y_n)\|_2}, \tag{VI.13}$$

donde $\|\cdot\|_2$ representa la norma euclídea de $\mathbb{R}^m$. Por tanto, el método proyectado $\widehat{\varphi}_h$ toma la expresión

$$\widehat{\varphi}_h(y_n) = \varphi_h(y_n) - \lambda_n w_n, \tag{VI.14}$$

donde el escalar real $\lambda_n = \lambda_n(h, y_n)$ debe ser elegido de modo que se satisfaga la ecuación de conservación

$$G(\widehat{\varphi}_h(y_n)) - G(y_n) \equiv \langle \widehat{\varphi}_h(y_n), \widehat{\varphi}_h(y_n)\rangle - \langle y_n, y_n\rangle = 0. \tag{VI.15}$$

De esta manera, al sustituir (VI.14) en (VI.15), observamos que λ_n debe satisfacer

$$\alpha_n \lambda_n^2 - 2\beta_n \lambda_n + \delta_n = 0, \tag{VI.16}$$

siendo

$$\begin{aligned} \alpha_n = \alpha(y_n, h) &= \langle w_n, w_n\rangle, \\ \beta_n = \beta(y_n, h) &= \langle \varphi_h(y_n), w_n\rangle, \\ \delta_n = \delta(y_n, h) &= \langle \varphi_h(y_n), \varphi_h(y_n)\rangle - \langle y_n, y_n\rangle. \end{aligned} \tag{VI.17}$$

Ahora bien, dado $h > 0$, el método original preserva el invariante cuadrático si $\delta_n = 0$, y en tal caso tomaríamos $\lambda_n = 0$. Si por el contrario lo que se tiene es que $\delta_n \neq 0$, entonces debemos requerir que $\beta_n^2 - \alpha_n \delta_n \geq 0$ de modo que podamos asegurar la existencia de solución real para la ecuación (VI.16). En este caso, elegiremos entonces λ_n como la raíz real de (VI.16) más próxima a $\lambda = 0$, esto es,

$$\lambda_n = \begin{cases} \dfrac{\delta_n}{\beta_n + \mathrm{sign}\beta_n \sqrt{\beta_n^2 - \alpha_n \delta_n}} = \dfrac{\delta_n}{\beta_n}\left(1 + \sqrt{1 - \dfrac{\alpha_n \delta_n}{\beta_n^2}}\right)^{-1}, & \text{si} \quad \beta_n \neq 0, \\ (-\delta_n/\alpha_n)^{1/2}, & \text{si} \quad \beta_n = 0. \end{cases} \tag{VI.18}$$

Para estudiar el orden de consistencia del método proyectado (VI.14)-(VI.15) observemos que

$$\widetilde{\varphi}_h(y_n) = y_n(t_n + h) + \widetilde{\epsilon}_{n+1}, \tag{VI.19}$$

donde $y_n(t)$ denota la solución local en el punto (t_n, y_n), mientras que $\widetilde{\epsilon}_{n+1}$ representa el error local del método encajado $\widetilde{\varphi}_h$, y que satisface

$$\widetilde{\epsilon}_{n+1} = C_{q+1}(y_n)h^{q+1} + \mathcal{O}(h^{q+2}),$$

para determinada constante C_{q+1}. En consecuencia, el comportamiento asintótico para $h \to 0$ del vector w_n y de las magnitudes $\alpha_n, \delta_n, \beta_n$ viene dado por

$$w_n = \frac{C_{q+1}}{\|C_{q+1}\|_2} + \mathcal{O}(h),$$

$$\begin{aligned} \alpha(y_n, h) &= \mathcal{O}(1) \\ \delta(y_n, h) &= \mathcal{O}(h^{p+1}) \\ \beta(y_n, h) &= \langle y_n, \frac{C_{q+1}}{\|C_{q+1}\|_2} \rangle + \mathcal{O}(h). \end{aligned} \tag{VI.20}$$

Por lo tanto tenemos el siguiente

Teorema VI.3.1 *Sean φ_h y $\widetilde{\varphi}_h$ dos métodos de Runge-Kutta explícitos encajados con órdenes p y q $(1 \leq q \leq p-1)$, respectivamente.*

i) Si $\langle y_n, C_{q+1}(y_n) \rangle \neq 0$ entonces existe $h^* > 0$ tal que el método proyectado $\widehat{\varphi}_h$ definido por (VI.14), con $\lambda_n = \lambda(y_n, h)$ dado por (VI.18), tiene orden de consistencia $\widehat{p} \geq p$, para todo $h \in (0, h^*]$.

ii) Si β_n definido por (VI.17) satisface $\beta_n = B(y_n)h^r + \mathcal{O}(h^{r+1})$, con $B(y_n) \neq 0$, $r \geq 1$ y $p - 2r + 1 > 0$, entonces el método proyectado $\widehat{\varphi}_h$ alcanza orden de consistencia $\widehat{p} \geq p - r$, para cada $h \in (0, h^*]$.

Demostración. *i)* Puesto que $\beta(y_n, 0) \neq 0$, y teniendo en cuenta (VI.20), deducimos la existencia de $h^* > 0$ tal que $\beta_n^2 - \alpha_n \delta_n > 0$ para $h \in (0, h^*]$. En consecuencia, λ_n está bien definido por (VI.18) y $\lambda_n = \mathcal{O}(h^{p+1})$. De este modo, teniendo en cuenta que

$$\widehat{\varphi}_h(y_n) - y(t_n + h) = (\varphi_h(y_n) - y(t_n + h)) - \lambda_n w_n,$$

deducimos que el método proyectado (VI.14) tiene orden de consistencia $\widehat{p} \geq p$.

ii) En este caso, ya que $\beta_n = \mathcal{O}(h^r)$, $\alpha_n = \mathcal{O}(1)$, $\delta_n = \mathcal{O}(h^{p+1})$ y $2r < p + 1$, entonces $\beta_n^2 - \alpha_n \delta_n > 0$ también se verifica para $h \in (0, h^*]$. Además deducimos de (VI.18) que $\lambda_n = \mathcal{O}(h^{p+1-r})$, y por lo tanto, según (VI.14), el método proyectado posee orden $\widehat{p} \geq p - r$. □

En la aplicación práctica de (VI.14), dado un método Runge-Kutta explícito φ_h de orden p, normalmente disponemos de una amplia libertad para elegir métodos encajados $\widetilde{\varphi}_h$ con

distintos órdenes de consistencia ($1 \leq q \leq p-1$). Entonces una cuestión de interés consiste en determinar cómo elegir convenientemente tal método encajado $\widetilde{\varphi}_h$. Es sencillo comprobar que la opción más simple, esto es, la opción de encajar el método de Euler explícito, provee en general un método proyectado de orden mayor o igual a p. De hecho, para tal elección $\widetilde{\varphi}_h(y_n) = y_n + hf(y_n)$, tenemos que $q = 1$ y $\widetilde{\varepsilon}_{n+1} = (1/2)y''(t_n)h^2 + \mathcal{O}(h^3)$ (véase (VI.19)). En consecuencia, $C_2 = \frac{1}{2}y''(t_n) = \frac{1}{2}f'(f(y_n))$, y la condición

$$\beta(y_n, 0) = \langle y_n, \frac{f'(f(y_n))}{\|f'(f(y_n))\|_2} \rangle \neq 0, \tag{VI.21}$$

implica que el método proyectado posee al menos orden de consistencia p.

Finalmente, observemos que la identidad

$$\langle y_n(t+h), y_n(t+h) \rangle - \langle y_n, y_n \rangle = 0,$$

se verifica a lo largo de la solución local $y_n(t)$ en el punto (t_n, y_n) para valores pequeños de $h > 0$, y que esto implica que $\langle y_n, f'(f(y_n)) \rangle = -\langle f(y_n), f(y_n) \rangle$. Así la condición (VI.21) puede ser fácilmente estudiada en la práctica.

Nota VI.3.2 La teoría presentada arriba se extiende de modo inmediato a invariantes más generales que incluyan tanto términos cuadráticos como lineales, esto es, invariantes del tipo $G(y) = y^T S y + d^T y$, donde S es una matriz simétrica constante y d un vector constante. De hecho, para el método proyectado $\widehat{\varphi}_h$ definido por (VI.14), tenemos que (VI.15) se verifica si y sólo si λ_n satisface la ecuación cuadrática

$$\alpha_n \lambda_n^2 - 2\beta_n' \lambda_n + \delta_n' = 0, \tag{VI.22}$$

donde $\beta_n' = \beta_n + \frac{1}{2}d^T w_n$ y $\delta_n' = \delta_n + d^T(\varphi_h(y_n) - y_n)$. Si $\delta_n' = 0$ para un valor dado h, entonces el método original φ_h conserva el invariante y tomaríamos $\lambda_n = 0$. En otro caso, la existencia de una solución real para la ecuación cuadrática (VI.22) depende del signo de la cantidad $(\beta_n')^2 - \alpha_n \delta_n'$. Sin embargo, podemos establecer condiciones suficientes análogas a las que se dan en el teorema VI.3.1 con vistas a garantizar la resolubilidad de la ecuación (VI.22).

VI.4. El método de proyección direccional para la conservación de invariantes generales

Nuestro propósito en esta sección es la extensión de la técnica de proyección direccional para la conservación de invariantes cuadráticos al caso de invariantes escalares generales así como al caso de múltiples invariantes escalares.

Para un invariante escalar general dado por una función regular $G = G(y) : \widehat{\mathcal{D}} \subset \mathbb{R}^m \to \mathbb{R}$ la técnica de proyección que proponemos se reduce a la resolución de la siguiente ecuación escalar no lineal

$$G\Big(\varphi_h(y_n) - \lambda_n w_n\Big) - G(y_n) = 0, \tag{VI.23}$$

donde $w_n = w_n(y_n, h)$ viene dado por (VI.13). En este caso, la existencia de solución real λ_n queda garantizada bajo las condiciones del siguiente

Teorema VI.4.1 *Sea φ_h y $\widetilde{\varphi}_h$ un par encajado de métodos de Runge-Kutta con órdenes p y q $(1 \leq q \leq p-1)$, respectivamente, y sea $C_{q+1}(y)h^{q+1}$ el término principal del error local (VI.19) del método $\widetilde{\varphi}_h(y)$.*

i) Si $\nabla G(y_n)^T C_{q+1}(y_n) \neq 0$, entonces existe $h^* > 0$ tal que la ecuación (VI.23) define una única función $\lambda_n = \lambda(y_n, h)$ para todo $h \in (0, h^*)$, que además verifica $\lambda(y_n, h) = \mathcal{O}(h^{p+1})$, $h \to 0$. Así, el método proyectado correspondiente $\widehat{\varphi}_h$ posee orden $\hat{p} \geq p$.

ii) Si $\nabla G(\varphi_h(y_n))^T w_n = B(y_n)h^r + \mathcal{O}(h^{r+1})$ con $r \geq 1$, $B(y_n) \neq 0$ y $2r \leq p$, entonces existe $h^* > 0$ tal que la ecuación (VI.23) define una única función $\lambda_n = \lambda(y_n, h)$ para cada $h \in (0, h^*)$, que además verifica $\lambda(y_n, h) = \mathcal{O}(h^{p+1-r})$, $h \to 0$. De este modo, el método proyectado asociado $\widehat{\varphi}_h$ posee orden $\hat{p} \geq p - r$.

Demostración.

i) Consideremos (λ, h) en un entorno del origen $(0, 0) \in \mathbb{R}^2$, y definamos la función real

$$g(\lambda, h) \equiv G\Big(\varphi_h(y_n) - \lambda w_n\Big) - G(y_n).$$

Teniendo en cuenta la regularidad asumida para la función G y que

$$g(0, 0) = G(y_n) - G(y_n) = 0,$$

$$\frac{\partial g}{\partial \lambda}(0, 0) = -\nabla G(y_n)^T \frac{C_{q+1}(y_n)}{\|C_{q+1}(y_n)\|_2} \neq 0,$$

entonces el teorema de la función implícita permite asegurar la existencia de un entorno $[0, h^*]$ y una única función regular $\lambda_n = \lambda_n(h)$ verificando $\lambda_n(0) = 0$ y $g(\lambda_n(h), h) = 0$, para cada $h \in (0, h^*)$.

Más aún, podemos considerar el desarrollo

$$g(\lambda_n, h) = g(0, h) + \frac{\partial g}{\partial \lambda_n}(0, h)\lambda_n + \mathcal{O}(\lambda_n^2),$$

con

$$\begin{aligned} g(0,h) &= G(\varphi_h(y_n)) - G(y_n) = \mathcal{O}(h^{p+1}), \\ \frac{\partial g}{\partial \lambda_n}(0,h) &= \frac{\partial g}{\partial \lambda_n}(0,0) + \mathcal{O}(h). \end{aligned}$$

Por lo tanto, deducimos para λ_n el comportamiento asintótico $\lambda_n = \lambda_n(h) = \mathcal{O}(h^{p+1})$, y de aquí que el método proyectado tenga orden de consistencia $\hat{p} \geq p$.

ii) En este caso consideramos la función real $z(\mu, h)$ definida en un entorno del origen por

$$z(\mu,h) = h^{-2r}\left[G(\varphi_h(y_n) - \mu h^r w_n) - G(y_n)\right], \quad \text{para } \ h \neq 0 \tag{VI.24}$$

donde $z(\mu, 0)$ está dada por el correspondiente límite

$$\lim_{h\to 0} z(\mu,h) = -B(y_n)\mu + \frac{1}{2\|C_{q+1}(y_n)\|_2^2} C_{q+1}(y_n)^T \frac{\partial^2 G}{\partial y^2}(y_n) C_{q+1}(y_n)\mu^2. \tag{VI.25}$$

Claramente z es una función continua para cada $h > 0$. Para estudiar la continuidad de z en $h = 0$ notemos que, en virtud del teorema del valor medio, se tiene que

$$G(\varphi_h(y_n) - \mu h^r w_n) = G(\varphi_h(y_n)) - \mu h^r \int_0^1 \nabla G(\varphi_h(y_n) - \theta \mu h^r w_n)^T w_n \mathrm{d}\theta.$$

De este modo, para $h \neq 0$ tenemos que

$$z(\mu,h) = -\mu h^{-r} \int_0^1 \nabla G(\varphi_h(y_n) - \theta\mu h^r w_n)^T w_n \mathrm{d}\theta + h^{-2r}\big(G(\varphi_h(y_n)) - G(y_n)\big).$$

Consideremos ahora la función regular definida por $\varsigma(y) := \nabla G(y)^T w_n$. Considerando de nuevo el teorema del valor medio aplicado a ς podemos escribir

$$\varsigma(\varphi_h(y_n) - \theta\mu h^r w_n) = \varsigma(\varphi_h(y_n)) - \theta\mu h^r \int_0^1 \nabla\varsigma(\varphi_h(y_n) - \rho\theta\mu h^r w_n)^T w_n \mathrm{d}\rho.$$

En consecuencia, tenemos que

$$\begin{aligned} z(\mu,h) &= h^{-2r}(G(\varphi_h(y_n)) - G(y_n)) - \mu h^{-r}\varsigma(\varphi_h(y_n)) + \\ &\quad \mu^2 \int_0^1 \int_0^1 \theta \nabla\varsigma(\varphi_h(y_n) - \rho\theta\mu h^r w_n)^T w_n \mathrm{d}\rho\mathrm{d}\theta. \end{aligned} \tag{VI.26}$$

Ya que $p + 1 > 2r$, entonces deducimos (VI.25) a partir de (VI.26), y por lo tanto $z(\mu, h)$ es una función continua en $h = 0$.

Además $\partial z/\partial \mu$ es también continua en un entorno del origen y $\partial z/\partial \mu(0,0) = -B(y_n) \neq 0$. Por consiguiente, en virtud del teorema de la función implícita, existe $h^* > 0$ y una función continua $\mu = \mu(h)$ tal que $z(\mu(h), h) = 0$ para cada $h \in (0, h^*)$.

Teniendo en cuenta (VI.24), $\lambda_n(h)$ dada por $\lambda_n(h) = h^r \mu(h)$, para todo $h \in (0, h^*)$, es la única solución de $G(\varphi_h(y_n) - \lambda_n w_n) - G(y_n) = 0$.

Finalmente, considerando el desarrollo

$$z(\mu, h) = z(0, h) + \frac{\partial z}{\partial \mu}(0, h)\mu + \mathcal{O}(\mu^2),$$

con

$$\begin{aligned} z(0,h) &= h^{-2r}\big(G(\varphi_h(y_n)) - G(y_n)\big) = \mathcal{O}(h^{p+1-2r}), \\ \frac{\partial z}{\partial \mu}(0,h) &= \frac{\partial z}{\partial \mu}(0,0) + \mathcal{O}(h) = -B(y_n) + \mathcal{O}(h), \end{aligned}$$

se deduce que $\mu = \mathcal{O}(h^{p+1-2r})$, y por tanto $\lambda_n = \mathcal{O}(h^{p+1-r})$, lo cual implica que el método proyectado posee orden $\hat{p} \geq p - r$. □

Si el método de Euler explícito $\widetilde{\varphi}_h(y_n) = y_n + hf(y_n)$ se considera como método encajado, entonces obtenemos el siguiente

Corolario VI.4.2 *Sean φ_h un método Runge-Kutta con orden p y $\widetilde{\varphi}_h$ el método de Euler explícito, esto es, $\widetilde{\varphi}_h(y_n) = y_n + hf(y_n)$. Si $\nabla G(y_n)^T y''(t_n) \neq 0$ entonces existe $h^* > 0$ tal que para todo $h \in (0, h^*)$ la ecuación (VI.23) define una única función $\lambda_n = \lambda_n(h, y_n)$ que satisface $\lambda_n(h, y_n) = \mathcal{O}(h^{p-1})$, $h \to 0$. En particular, el correspondiente método proyectado posee orden p.*

Demostración. Basta con tener en cuenta que el término principal de error para el método de Euler explícito viene dado por $C_2(y_n)h^2 = \frac{1}{2}y''(t_n)h^2$, y que, por hipótesis, $\nabla G(y_n)^T C_2(y_n) = \frac{1}{2}\nabla G(y_n)y''(t_n) \neq 0$. Por tanto, el enunciado se obtiene por aplicación directa del apartado $i)$ del teorema VI.4.1. □

Nota VI.4.3

1. Notemos que para invariantes cuadráticos $G(y) = \langle y, y\rangle$ la condición $ii)$ en el teorema VI.4.1 se reduce a $\nabla G(\varphi_h(y_n))^T w_n = 2\langle \varphi_h(y_n), w_n\rangle = \mathcal{O}(h^r)$, que resulta equivalente a la condición $\beta_n = \mathcal{O}(h^r)$ dada en el teorema VI.3.1.

2. Ya que w_n es un vector unitario en la dirección del error local del método encajado $\widetilde{\varphi}_h(y_n)$ y $w_n = C_{q+1}(y_n)/\left\|C_{q+1}(y_n)\right\|_2 + \mathcal{O}(h)$, con $C_{q+1}(y_n) \neq 0$, entonces la condición $\nabla G(y_n)^T C_{q+1}(y_n) \neq 0$ que aparece en el apartado $i)$ del teorema VI.4.1 es equivalente a que el gradiente de la función $G(y)$ no se anule en la dirección del término principal del error $C_{q+1}(y_n)$ del método encajado $\widetilde{\varphi}_h$ en el punto y_n. Desde una perspectiva

geométrica esto nos dice que el vector $C_{q+1}(y_n)$ no es tangente a la superficie de nivel $G(y) - G(y_n) = 0$ en el punto y_n. Además, si el término principal de $\nabla G(\varphi_h(y_n))^T w_n$, dado por $\nabla G(y_n)^T C_{q+1}(y_n)$ se anula en algún punto y_n pero $\nabla G(\varphi_h(y_n))^T w_n = \mathcal{O}(h^r)$, siendo $1 \leq r \leq p/2$, entonces la estrategia propuesta permite asegurar la conservación del invariante a costa de una reducción de orden de consistencia ($\hat{p} \geq p - r$) de la solución numérica proyectada.

3. El teorema VI.4.1 también puede ser aplicado para demostrar que la técnica de proyección ortogonal estándar de un método de tipo Runge-Kutta de orden p conserva el orden de consistencia. De hecho, la proyección ortogonal (VI.5) toma la forma de proyección direccional (VI.14) para la elección $w_n = -\nabla G(\varphi_h(y_n))$, y en dicho caso, ya que $\nabla G(y_n)^T w_n = \|\nabla G(y_n)\|_2^2 + \mathcal{O}(h)$, el método de proyección ortogonal (VI.5) posee al menos orden p en virtud del apartado i) del teorema VI.4.1. Por otro lado, si $\nabla G(y_n) = 0$ entonces la proyección ortogonal estándar no es aplicable; pero si en tal caso se tuviera $\nabla G(\varphi_h(y_n))^T w_n = \mathcal{O}(h^r)$, para determinada elección de w_n, entonces nuestra técnica de proyección sí puede ser aplicada. Notemos además que la aplicación de la técnica de proyección ortogonal requiere disponer de una expresión analítica para el gradiente $\nabla G(y)$, mientras que nuestra técnica sólo precisa de resolver una ecuación no lineal, que puede ser resuelta por métodos clásicos que no requieren la evaluación del gradiente.

4. Tengamos en cuenta que el flujo de un sistema diferencial autónomo es invariante por transformaciones afines; esto es, para cualquier transformación afín $z \to y = Pz + q$, con $P \in \mathbb{R}^{m\times m}$ y $q \in \mathbb{R}^m$ matriz y vector constantes, si se denota por $\psi_{f,t}$ al flujo del problema diferencial $y' = f(y)$ y $\psi_{\tilde{f},t}$ al flujo correspondiente al sistema transformado $z' = \tilde{f}(z) = P^{-1} f(Pz+q)$, entonces se tiene invariancia afín en el sentido de que $\psi_{f,t}(y_0) = P\psi_{\tilde{f},t}(z_0) + q$, siempre que $y_0 = Pz_0 + q$. Es sencillo comprobar que esta propiedad de conservación se verifica para cualquier método de tipo Runge-Kutta, así como para los métodos de proyección direccional propuestos (VI.14). Sin embargo, la técnica de proyección ortogonal verifica la invariancia afín únicamente en el caso de matrices ortogonales P.

VI.4.1. Implementación en códigos Runge-Kutta adaptativos

Dado un par encajado de métodos de tipo Runge-Kutta con órdenes de consistencia p y $p-1$, el tamaño de paso h_n para avanzar la integración desde t_n a $t_{n+1} = t_n + h_n$ se elige de tal modo que la norma del estimador de error local $EST = y_{n+1} - y_{n+1}^{(p-1)}$, definido por la diferencia entre las soluciones y_{n+1} e $y_{n+1}^{(p-1)}$ de órdenes p y $p-1$, respectivamente, sea menor que una tolerancia de error TOL dada.

Para la nueva solución proyectada $\widehat{y}_{n+1} = y_{n+1} - \lambda_n w_n$ podría pensarse en considerar como nuevo estimador de error $\widehat{\mathrm{EST}} = \widehat{y}_{n+1} - y_{n+1}^{(p-1)}$; sin embargo, $\widehat{y}_{n+1}$ podría perder (en algunos puntos aislados) uno o más órdenes de consistencia respecto a la solución exacta y por tanto no proveería una estimación fiable. En consecuencia, proponemos una variante más conservativa en la que se estudia tanto el estimador EST y la diferencia entre el método proyectado y el método original de alto orden $\widehat{y}_{n+1} - y_{n+1} = \|\lambda_n w_n\| = |\lambda_n|$, y se aceptará un paso siempre que $\max\{\|EST\|, |\lambda_n|\} \leq TOL/2$.

Observemos que $\|EST\| = \mathcal{O}(h^p)$ y si la solución proyectada no presenta reducción de orden entonces $|\lambda_n| = \mathcal{O}(h^{p+1})$. De aquí que el anterior criterio de control del tamaño de paso esté determinado normalmente por la cantidad EST. Por lo tanto, tanto en caso de un paso rechazado o de un paso aceptado, proponemos avanzar la integración seleccionando un nuevo tamaño de paso según la fórmula estándar modificada

$$fac * \left(\frac{TOL}{2\max\{\|EST\|, |\lambda_n|\}} \right)^{1/p} * h_n,$$

siendo fac el factor de seguridad que se provee con el código original.

Por otro lado, notemos que para un par encajado general $(\varphi_h, \widetilde{\varphi_h})$ de métodos Runge-Kutta con órdenes p y q, respectivamente, el método proyectado $\widehat{y}_{n+1} = \widehat{\varphi}_h(y_n)$ se puede reescribir en la forma

$$\widehat{y}_{n+1} = y_{n+1} - \lambda_n w_n = y_{n+1} - \frac{\lambda_n}{\|y_{n+1} - \widetilde{y}_{n+1}\|} (y_{n+1} - \widetilde{y}_{n+1}),$$

y en consecuencia adopta la forma de un método de tipo Runge-Kutta siempre que la cantidad $\lambda_n \|y_{n+1} - \widetilde{y}_{n+1}\|^{-1}$ permanezca acotada. Ya que $\lambda_n = \mathcal{O}(h^{p+1-r})$, con $r \geq 0$, y $\|y_{n+1} - \widetilde{y}_{n+1}\| = \mathcal{O}(h^{q+1})$, entonces el cociente anterior tiene un comportamiento asintótico del tipo $\mathcal{O}(h^{p-q-r})$. De este modo, para el caso de una severa reducción de orden, esto es, $r > p-q$, el método proyectado perdería su fiabilidad. En consecuencia, para prevenir al código de eventuales reducciones severas de orden proponemos incluir un control adicional de la forma $|\lambda_n| < \|y_{n+1} - \widetilde{y}_{n+1}\|$. En el caso de que esta condición no sea satisfecha en un determinado paso temporal, entonces elegiríamos otro método a encajar $\widetilde{\varphi}_h$ que cambiaría la dirección que define al método proyectado. Sin embargo, este caso parece ser poco común y de hecho en los experimentos numéricos que mostramos en la sección final del capítulo no hemos detectado casos de reducción severa de orden que haga necesario un control de este tipo.

Finalmente debemos comentar que en la aplicación práctica de nuestra técnica con un invariante escalar no cuadrático debemos resolver en cada paso la ecuación (no cuadrática) $g(\lambda, h) =$

0. En este sentido, se han propuesto muchos métodos iterativos para resolver ecuaciones no lineales; así, en métodos iterativos tipo Newton necesitamos calcular, ya sea analíticamente o numéricamente, el gradiente de la función G que define el invariante una o más veces por paso temporal dependiendo de la convergencia de la iteración. En nuestros experimentos numéricos llevados a cabo considerando el método de Newton modificado se alcanzó la convergencia con 1 o 2 iteraciones por paso. Sin embargo, también pueden ser considerados otros esquemas iterativos que únicamente requieren la evaluación de la función G. Así, hallamos que usando el método de la secante con valores de arranque $\lambda = 0$ y el valor λ calculado en el paso previo se obtiene convergencia al nivel de error de redondeo de la máquina con 1 o 2 iteraciones para todos los problemas test considerados. En consecuencia, consideramos preferible este esquema iterativo por cuanto posee una convergencia similar a la del método de Newton modificado, tiene un costo computacional menor y además no precisa la estimación numérica del gradiente.

VI.4.2. Conservación de invariantes múltiples

Supongamos ahora que el sistema diferencial (VI.1) posee $l \geq 1$ invariantes regulares $G_1(y), \ldots, G_l(y)$ definidos en $\mathcal{D} \subset \mathbb{R}^m$ y pongamos $G = (G_1, \ldots, G_l)^T : \mathcal{D} \subset \mathbb{R}^m \to \mathbb{R}^l$. Junto con el método de arranque φ_h de orden p, consideramos l métodos encajados linealmente independientes $\widetilde{\varphi}_h^{(1)}, \ldots, \widetilde{\varphi}_h^{(l)}$ con órdenes $q_1 \leq q_2 \leq \ldots \leq q_l < p$. Entonces, generalizando la proyección propuesta en (VI.14), el método proyectado $\widehat{\varphi}_h$ se definirá en este caso como

$$\widehat{\varphi}_h(y_n) = \varphi_h(y_n) - \sum_{i=1}^{l} \lambda_n^{(i)} w_n^{(i)}, \tag{VI.27}$$

donde

$$w_n^{(i)} = \frac{\varphi_h(y_n) - \widetilde{\varphi}_h^{(i)}(y_n)}{\left\| \varphi_h(y_n) - \widetilde{\varphi}_h^{(i)}(y_n) \right\|_2}$$

y $\lambda_n^{(i)} = \lambda^{(i)}(h, y_n)$, $i = 1, \ldots, l$, son escalares dependientes del problema que se integra y que serán elegidos de tal modo que

$$G\Big(\varphi_h(y_n) - \sum_{i=1}^{l} \lambda_n^{(i)} w_n^{(i)}\Big) = G(y_n). \tag{VI.28}$$

Introducimos además la matriz T_n de dimensión $m \times l$ definida por sus columnas como

$$T_n = \left[w_n^{(1)} | \ldots | w_n^{(l)} \right],$$

así que

$$\left(\nabla G(\varphi_h(y_n))\right)^T \; T_n = M(y_n) + \mathcal{O}(h), \tag{VI.29}$$

con $M(y_n) \in \mathbb{R}^{l\times l}$. Con estas notaciones tenemos el siguiente

Teorema VI.4.4 *Si (VI.29) se verifica, siendo $M(y_n)$ una matriz regular, entonces existe $h^* > 0$ y unos únicos escalares reales $\lambda^{(1)}(y_n,h),\ldots,\lambda^{(l)}(y_n,h))$ tal que el método proyectado $\widehat{\varphi}_h$ definido por (VI.27) conserva los l invariantes $G_i, i = 1,\ldots,l$, para cada $h \in (0,h^*)$. Además, el método proyectado posee orden $\hat{p} \geq p$.*

Demostración. Consideremos $\lambda = (\lambda^{(1)},\ldots,\lambda^{(l)})^T \in \mathbb{R}^l$ y $g(\lambda,h) : \mathbb{R}^l \times \mathbb{R} \to \mathbb{R}$ definida por

$$g(\lambda,h) = G\left(\varphi_h(y_n) - \sum_{i=1}^{l} \lambda^{(i)} w_n^{(i)}\right) - G(y_n). \tag{VI.30}$$

Análogamente a la prueba del apartado i) del teorema VI.4.1, y en virtud de la regularidad de G y φ_h, se tiene que

$$g(0,0) = G(y_n) - G(y_n) = 0,$$

$$\frac{\partial g}{\partial \lambda}(0,0) = -M(y_n).$$

Teniendo en cuenta la regularidad de $M(y_n)$ el teorema de la función implícita garantiza la existencia de un número real positivo h^* y de una única función regular $\lambda_n = \lambda_n(h) = (\lambda_n^{(1)}(h),\ldots,\lambda_n^{(l)}(h))$ verificando $\lambda_n(0) = 0$ y $g(\lambda_n(h),h) = 0$, para cada $h \in (0,h^*)$.

Por otro lado, ya que

$$g(\lambda_n,h) = g(0,h) + \frac{\partial g}{\partial \lambda_n}(0,h)\lambda_n + \mathcal{O}(\|\lambda_n\|_2^2),$$

siendo

$$\begin{aligned} g(0,h) &= G(\varphi_h(y_n)) - G(y_n) = \mathcal{O}(h^{p+1}),\ h \to 0, \\ \frac{\partial g}{\partial \lambda_n}(0,h) &= -M(y_n) + \mathcal{O}(h),\ h \to 0, \end{aligned}$$

deducimos que $\lambda_n(h) = \mathcal{O}(h^{p+1}),\ h \to 0$, y en consecuencia $\hat{p} \geq p$. □

Nota VI.4.5 Un desarrollo análogo al efectuado en la prueba del apartado ii) del teorema VI.4.1 permite concluir que si en lugar de la condición (VI.29) lo que se tiene es que

$$\left(\nabla G(\varphi_h(y_n))\right)^T \; T_n = M(y_n)h^r + \mathcal{O}(h^{r+1}), \tag{VI.31}$$

siendo $M(y_n) \in \mathbb{R}^{l\times l}$ una matriz regular y $2r \leq p$, entonces queda garantizada la resolubilidad de la ecuación (VI.28) para valores de h en cierto intervalo $(0, h^*)$. Además la solución presenta un comportamiento asintótico del tipo $\lambda(y_n, h) = \mathcal{O}(h^{p+1-r}),\ h \to 0$, con lo que el método proyectado conservaría el invariante múltiple G a costa de una reducción de orden $\hat{p} = p - r$.

Sin embargo, la condición (VI.31) sólo tiene sentido si cada uno de los productos

$$\left(\nabla G(\varphi_h(y_n))\right)^T w_n^{(i)}, \quad 1 \leq i \leq s,$$

presentan un comportamiento asintótico de la misma potencia en h cuando $h \to 0$. Esto hace que la condición (VI.31) tenga poca relevancia desde una perspectiva práctica. Así, en la práctica, si la condición (VI.29) no se satisface entonces optaremos por una elección alternativa de los métodos encajados $\widetilde{\varphi}_h^{(1)}, \ldots, \widetilde{\varphi}_h^{(l)}$.

Para finalizar esta sección enunciamos el siguiente resultado, que se establece de cara a asegurar que las iteraciones tipo Newton aplicadas a las ecuaciones (VI.23) y (VI.28) proveen sucesiones numéricas que convergen en entornos del origen. Más precisamente, tenemos el siguiente

Corolario VI.4.6 *Bajo las mismas condiciones que en el Teorema VI.4.4, existe $h^* > 0$ tal que para todo $0 < h < h^*$ la sucesión $\{\lambda^{(k)}\}$ definida por el método de Newton aplicado a la ecuación (VI.28), con $\lambda^{(0)} := 0$, converge al único cero de esta ecuación en un entorno del origen $\lambda = 0$ suficientemente pequeño.*

Demostración. Considerando, como en el Teorema VI.4.4, la función $g(\lambda, h)$ dada por (VI.30), obtenemos que

$$\begin{aligned}
\gamma &:= \sup_{\substack{\eta,\mu\in\Lambda_0\\ \eta\neq\mu}} \frac{\|\nabla_\lambda g(\eta, h) - \nabla_\lambda g(\mu, h)\|_2}{\|\eta - \mu\|_2} = \mathcal{O}(1),\\
\alpha &:= \|\nabla_\lambda g(0, h)^{-1} g(0, h)\|_2 = \mathcal{O}(h^{p+1}),\\
\beta &:= \|\nabla_\lambda g(0, h)^{-1}\|_2 = \mathcal{O}(1),
\end{aligned}$$

donde $\Lambda_0 \subset \mathbb{R}^l$ representa un entorno de $\lambda = 0$ suficientemente pequeño. En consecuencia, existe $h^* > 0$ tal que para cada $0 < h < h^*$ se verifica la condición

$$\tau := \gamma\alpha\beta = \mathcal{O}(h^{p+1}) \leq \frac{1}{2}.$$

Por lo tanto, el enunciado sigue en virtud del teorema de Newton-Kantorovich (véase por ejemplo [84, Teorema 5.3.6]). □

Nota VI.4.7 En lo referente a la implementación de la técnica de proyección direccional para la conservación de invariantes múltiples debemos comentar que la extensión del control del tamaño de paso explicado en la subsección VI.4.1 al caso de varios invariantes es análoga

VI.5. Experimentos numéricos

En este punto presentamos diversas pruebas numéricas con el objetivo de mostrar algunas características de la puesta en práctica de la nueva técnica de proyección propuesta. En primer lugar, debemos dejar claro que esta nueva técnica de proyección puede ser incluida de modo sencillo en cualquier código de tipo Runge-Kutta adaptativo para lograr la conservación de invariantes a costa de un ligero incremento en el costo computacional de tal modo que el código numérico resultante mantenga su eficiencia y fiabilidad. A efectos de comparación hemos incluido también en los experimentos los resultados obtenidos a partir de la integración considerando la proyección ortogonal estándar. Como comentamos en las secciones previas, esta técnica mantiene el orden de consistencia del integrador original pero requiere el conocimiento de una expresión explícita para el gradiente ∇G, lo que implica que en general tenga un mayor coste computacional.

Para una comparación uniforme de la técnica de proyección direccional propuesta con la proyección ortogonal clásica, consideramos aquí que las ecuaciones de conservación que definen ambas proyecciones son resueltas por medio de iteraciones del método de Newton simplificado hasta la precisión de la máquina. Sin embargo, tal como ya se ha comentado en la sección previa, nuestra técnica no requiere el conocimiento explícito del gradiente de la función que define el invariante del problema a integrar, y por lo tanto las ecuaciones de conservación pueden ser resueltas en la práctica de modo eficiente por medio de otros métodos iterativos que únicamente requieran la evaluación del invariante del problema, como puede ser el método de la secante.

En los experimentos que se muestran abajo, hemos considerado el par clásico de Dormand y Prince (1980) $DoPri5(4)$ (véase, por ejemplo, [36] para detalles en los coeficientes del par de métodos) como integrador base. Para cada ejemplo a integrar elegiremos tantos métodos de orden uno como el número de invariantes que se desean conservar. Así tras avanzar la integración con el par $DoPri5(4)$ aplicaremos la proyección direccional $DoPri5(4) + PD$ definida por el método de orden 5 del par $DoPri5(4)$ y los métodos de orden uno que se fijan en cada problema a integrar. El método de orden 4 que conforma el $DoPri5(4)$ se considerará para la estimación del error local de la solución proyectada obtenida. En caso de rechazo del tamaño de paso actual por el estimador de error local entonces el tamaño de paso se reduce de forma estándar como en (II.37). En caso de rechazo del tamaño de paso actual en la resolución de la ecuación de conservación entonces el tamaño de paso se reduce a la mitad. Finalmente, tras un paso

temporal aceptado, se propone avanzar la integración de modo usual como en (II.36).

Presentamos a continuación los sistemas diferenciales a integrar a efectos de ilustración de la ejecución de la técnica de proyección direccional. Para los problemas que consideramos o bien se conoce la solución exacta en todo punto, o bien se conoce el periodo de su solución, lo cual permite conocer exactamente el error global y de conservación que cometen los integradores al finalizar la integración. De este modo, para cada problema presentamos gráficas de eficiencia en las que se muestra el comportamiento del error global y de los errores de conservación con respecto al número de evaluaciones de funciones derivadas ($NFCN$) que requiere el método para concluir la integración. En dichas gráficas, $DoPri5(4)$ representa el par clásico original de Dormand y Prince, $DoPri5(4)+PO$ denota la proyección ortogonal estándar del par original, mientras $DoPri5(4)+PD$ simboliza la proyección direccional en base métodos encajados de orden uno que se eligen en función del problema a integrar.

***E*jemplo 1:** *Un problema lineal.*

$$f(y) := (y_2 - y_3, -y_1 + y_3, y_1 - y_2)^T, \tag{VI.32}$$

con $y = (y_1, y_2, y_3)^T$. Ya que $y'(t)$ es ortogonal a la solución exacta $y(t)$, queda claro que $G(y) = y^T y$ es un invariante del problema. Más aún, la cantidad $y_1 + y_2 + y_3$ se conserva a lo largo de cualquier solución. Para el valor inicial $y_0 = (3, 0, 0)^T$ obtenemos la solución exacta dada por

$$y(t) = \left(1 + 2\cos(\sqrt{3}t), 1 - \cos(\sqrt{3}t) - \sqrt{3}\sin(\sqrt{3}t), 1 - \cos(\sqrt{3}t) + \sqrt{3}\sin(\sqrt{3}t)\right)^T.$$

Este problema será integrado hasta el tiempo $T_f = 5T_1$, siendo T_1 el periodo minimal de la solución, $T_1 = \frac{2\pi}{\sqrt{3}}$.

Consideramos aquí la proyección direccional definida por el método de Euler explícito y la proyección ortogonal para conservar únicamente el invariante cuadrático. El objetivo es ver que la naturaleza de la proyección direccional permite conservar el invariante cuadrático sin perturbar la conservación del invariante lineal del problema, mientras que esto no ocurre con la proyección ortogonal. Vemos en la figura VI.3 como el método original y su proyección direccional ($DoPri5(4)+PD$) conservan el invariante lineal, mientras que la proyección ortogonal ($DoPri5(4)+PO$) no tiene en cuenta en su definición la conservación de invariantes lineales. Las gráficas de eficiencia VI.1 y VI.2 muestran, respectivamente, cómo las distintas proyecciones mantienen el orden de consistencia del método original y conservan el invariante cuadrático

del problema. En este ejemplo, la proyección direccional provee un error global menor que la proyección ortogonal como fruto de la conservación del invariante lineal.

***Ejemplo* 2:** *Un problema de micromagnetismo*, (véase, por ejemplo, [82]).

$$f(y) := H_{eff} \times y + \lambda y \times (H_{eff} \times y), \tag{VI.33}$$

siendo $\lambda = 1/20.1$ y $H_{eff}, y \in \mathbb{R}^3$. Hemos elegido como valor inicial $y(0) = (y_0[1], y_0[2], y_0[3]) = (\sin\theta_0 \cos\varphi_0, -\sin\theta_0 \sin\varphi_0, \cos\theta_0)^T$, donde $\varphi_0 = \pi/4$ y $\theta_0 = \pi/3$.

Este tipo de ecuaciones hace aparición en problemas de micromagnetismo al resolver la ecuación de Landau-Lifshitz-Gilbert (ver, por ejemplo, [70, 82, 83]). Como en el ejemplo previo, ya que $y'(t)$ es ortogonal a la solución exacta del problema $y(t)$, se tiene que $G(y) = y^T y$ es un invariante del problema. Por otro lado, para la elección $H_{eff} := (1, 0, 0)^T$ entonces la solución exacta adopta la expresión

$$y(t) = \left(\frac{a(t)}{b(t)}, \frac{2}{b(t)}(y_0[2]\cos t - y_0[3]\sin t), \frac{2}{b(t)}(y_0[2]\sin t + y_0[3]\cos t) \right)^T,$$

con

$$a(t) = e^{\lambda t}(1 + y_0[1]) - e^{-\lambda t}(1 - y_0[1]), \quad b(t) = e^{\lambda t}(1 + y_0[1]) + e^{-\lambda t}(1 - y_0[1]).$$

Este problema será integrado hasta el tiempo $T_f = 16\pi$. Considerando la proyección direccional definida por el método de Euler explícito observamos en las figuras VI.4 y VI.5 que ambas proyecciones mantienen el orden de consistencia del método original y además preservan el invariante cuadrático del problema. En este ejemplo la proyección direccional también provee un error global ligeramente mejor respecto al que provee la proyección ortogonal.

***Ejemplo* 3:** *Ecuación de Duffing no forzada*, (véase, por ejemplo, [49, p. 82-91]).

$$f(y) := (y_2, -(1 - 2K)y_1 - 2Ky_1^3)^T, \tag{VI.34}$$

con $y = (y_1, y_2)^T$. Este problema así planteado posee un invariante polinomial de grado 4 definido por $G(y) := \frac{1-2K}{2}y_1^2 + \frac{1}{2}y_2^2 + \frac{K}{2}y_1^4$.

Para el valor inicial $y_0 = (1, 0)^T$ y la constante $K = 0.51$ la primera componente de la solución exacta viene dada por $y_1(t) = cn(t, K)$, siendo cn la segunda función elíptica de Jacobi. Notemos que este problema admite una formulación como problema diferencial ordinario de segundo orden, y ésta es la razón por la cual únicamente mediremos el error global de las soluciones numéricas respecto a la componente y_1.

El problema será integrado hasta $T_f = 50$. Vemos entonces en las figuras VI.6 y VI.7 que la proyección direccional definida por el método de Euler explícito y la proyección ortogonal muestran un comportamiento similar tanto en el error global que proveen como en la conservación del invariante.

***Ejemplo* 4:** *El problema restringido de los tres cuerpos*, (véase, por ejemplo, [52, p. 127-129]).

En este caso $F : \mathbb{R}^4 \to \mathbb{R}^4$ esta dada por

$$f(y) := \left(y_3, y_4, y_1 + 2y_4 - \bar{\mu}\frac{y_1+\mu}{D_1^{3/2}} - \mu\frac{y_1-\bar{\mu}}{D_2^{3/2}}, y_2 - 2y_3 - \bar{\mu}\frac{y_2}{D_1^{3/2}} - \mu\frac{y_2}{D_2^{3/2}} \right)^T, \qquad \text{(VI.35)}$$

con $D_1 := (y_1+\mu)^2 + y_2^2$, $D_2 := (y_1-\bar{\mu})^2 + y_2^2$, $y = (y_1, y_2, y_3, y_4)^T$ e

$$y(0) = (0.994, 0, 0, -2.00158510637908252240537862224)^T,$$

donde $\bar{\mu} = 1 - \mu$ y $\mu = 0.012277471$. Para esta elección de las condiciones iniciales se obtiene una solución periódica con periodo minimal dado por $T_2 \simeq 17.065216560157962558891720624 9$. La energía total

$$G(y) = \frac{1}{2}(y_3^2 + y_4^2 - y_1^2 - y_2^2) - \bar{\mu}D_1^{-1/2} - \mu D_2^{-1/2}$$

constituye una integral primera del sistema. El problema será integrado hasta $T_f = 3T_2$.

En este ejemplo, hemos considerado dos alternativas en la elección del método encajado que define la proyección direccional. Hemos considerado como primera opción el método de Euler explícito $(DoPri5(4) + PD1)$ y, en segundo lugar $(DoPri5(4) + PD2)$ el método de orden uno que posee la misma matriz de coeficientes A que el $Dopri5(4)$, pero con vector de pesos $b^T = (\frac{1}{3}, \frac{2}{3}, 0, 0, 0, 0, 0)$. Las figuras VI.8 y VI.9 muestran respectivamente la órbita numérica que proveen el método original y las proyecciones direccional y ortogonal tras considerar tres veces el periodo de la solución. En la figura VI.10 observamos que los métodos de proyección involucrados mantienen el orden de consistencia del método original. Vemos además que la proyección direccional $DoPri5(4)+PD1$ provee un error global similar a la proyección ortogonal, mientras que la proyección direccional $DoPri5(4) + PD2$ muestra claramente un comportamiento más eficiente. Aunque teóricamente el comportamiento del error global que proveen las técnicas de proyección ortogonal y direccional es similar, vemos que en la práctica nuestra técnica es más flexible por cuanto permite cambiar el método a encajar de modo que se pueda obtener una ligera ganancia en el error global conservando al mismo tiempo el invariante del problema. Sin embargo, la elección óptima del método a encajar de cara a maximizar la eficiencia de la proyección direccional es un problema complicado de abordar por cuanto presenta dependencia del

problema que se integra. Finalmente, notamos en la figura VI.11 que la conservación del invariante es en este ejemplo algo más defectuosa. Ello es debido a la presencia de singularidades en el invariante, lo cual produce inexactitudes en la resolución de las ecuaciones de conservación.

***Ejemplo* 5:** *Ecuaciones de Euler*, (véase, por ejemplo, [34] y [50, p. 95–96]).

$$f(y) := \big((\alpha-\beta)y_2y_3,\ (1-\alpha)y_3y_1,\ (\beta-1)y_1y_2\big)^T, \tag{VI.36}$$

con $y=(y_1,y_2,y_3)^T$ e $y(0)=(0,1,1)^T$, siendo $\alpha = 1+\frac{1}{\sqrt{1.51}}$ y $\beta = 1-\frac{0.51}{\sqrt{1.51}}$. En este caso, la solución exacta viene dada por

$$y(t) = \left(\sqrt{1.51}sn(t,0.51), cn(t,0.51), dn(t,0.51)\right)^T,$$

donde sn, cn, dn denotan las funciones elípticas de Jacobi. En general, las órbitas de este sistema diferencial verifican la conservación de dos invariantes cuadráticos dados por

$$(i)\ \ G_1(y) = y_1^2+y_2^2+y_3^2, \qquad (ii)\ \ G_2(y) = y_1^2+\beta y_2^2+\alpha y_3^2.$$

Este problema ha sido integrado hasta $T_f = 100$. En caso de la conservación de invariantes dobles, la proyección direccional queda definida a partir de dos métodos encajados al método de orden 5 que conforma el par $DoPri5(4)$. Una vez más, el método de orden 4 será útil en la estimación del error local que provee el método de proyección direccional.

Aunque no disponemos de un criterio que nos asegure de antemano la resolubilidad de la ecuación de conservación para una elección de métodos encajados dada, en la práctica la clase de métodos de orden menor o igual a dos resulta ser lo suficientemente amplia como para hallar métodos que aseguren la resolubilidad. Las figuras VI.12 y VI.13 corresponden a la integración de este ejemplo considerando la proyección direccional $(DoPri5(4)+PD)$ del $DoPri5(4)$ por medio los métodos encajados de orden uno definidos por los pesos $b_1^T = (\frac{1}{3},\frac{2}{3},0,0,0,0,0)$ y $b_2^T = (\frac{1}{3},0,\frac{2}{3},0,0,0,0)$. Vemos entonces en dichas figuras que la proyección direccional presenta un comportamiento satisfactorio tanto en la conservación del invariante doble como en el mantenimiento del orden de consistencia 5.

***Ejemplo* 6:** *Ecuación del péndulo en coordenadas cartesianas* (véase, por ejemplo, [50, p. 106]).

$$f(y) := \big(y_3, y_4, -y_1\frac{y_3^2+y_4^2-y_2}{y_1^2+y_2^2}, -1-y_2\frac{y_3^2+y_4^2-y_2}{y_1^2+y_2^2}\big)^T, \tag{VI.37}$$

con $y = (y_1, y_2, y_3, y_4)^T$ e $y(0) = (1, 0, 0, 1)^T$. En este caso, la solución exacta es periódica con periodo minimal dado por

$$T_3 = 4 \int_0^{\pi/3} \frac{dx}{\sqrt{2\cos x - 1}} \simeq 8.626062589998573.$$

En la integración de este problema tomaremos $T_f = 4T_3$ y consideraremos los invariantes

$$(i)\ \ G_1(y) \equiv y_1 y_3 + y_2 y_4 = 0, \qquad (ii)\ \ G_2(y) \equiv y_1^2 + y_2^2 = 1.$$

En este caso hemos optado por implementar la proyección direccional definida por el método de Euler explícito y el método de orden uno que posee por vector de pesos $b_1^T = (\frac{1}{3}, \frac{2}{3}, 0, 0, 0, 0, 0)$. Los resultados correspondientes a la integración numérica de este ejemplo se ilustran en las figuras VI.14 y VI.15. En efecto comprobamos que la proyección direccional constituye una estrategia tan eficiente como la proyección ortogonal en el mantenimiento del orden de consistencia y de la conservación del invariante doble del problema.

***E*jemplo 7:** *Problema de los dos cuerpos* (ver, por ejemplo, [52, p. 236]).

$$f(y) := \big(y_3, y_4, \frac{-y_1}{(y_1^2 + y_2^2)^{3/2}}, \frac{-y_2}{(y_1^2 + y_2^2)^{3/2}}\big)^T, \tag{VI.38}$$

con $y = (y_1, y_2, y_3, y_4)^T$ e $y(0) = (0.5, 0, 0, \sqrt{3})^T$. En este caso, la solución exacta es periódica con periodo minimal dado por $T_4 = 2\pi$. En la integración de este problema tomaremos $T_f = 4T_4$ y consideraremos los invariantes definidos por el momento angular y la energía total del sistema, dados respectivamente por

$$(i)\ \ G_1(y) \equiv y_1 y_4 + y_2 y_3, \qquad (ii)\ \ G_2(y) \equiv \frac{1}{2}(y_3^2 + y_4^2) - \frac{1}{\sqrt{y_1^2 + y_2^2}}.$$

En este caso hemos considerado dos alternativas a la hora de implementar la proyección direccional. Por un lado tendremos la proyección direccional $DoPri5(4) + PD1$ del $DoPri5(4)$ definida por los métodos encajados de orden uno que poseen por vector de pesos $b_1^T = (0, 0, 0, 0, 0, 1, 0)$ y $b_2^T = (\frac{1}{3}, \frac{2}{3}, 0, 0, 0, 0, 0)$; y por otro consideraremos la proyección direccional $DoPri5(4) + PD2$ definida por los métodos de orden uno con pesos $b_1^T = (\frac{1}{2}, 0, \frac{1}{2}, 0, 0, 0, 0)$ y $b_2^T = (0, 0, 0, \frac{1}{3}, 0, \frac{2}{3}, 0)$.

Vemos en la figura VI.17 como ambas proyecciones conservan satisfactoriamente los invariantes del problema; sin embargo, notamos en la figura VI.16 que, para ambas elecciones de la proyección direccional, la proyección ortogonal provee un error global ligeramente inferior a ambas proyecciones direccionales. No obstante lo anterior, las proyecciones direccionales mantienen el orden de consistencia del método original.

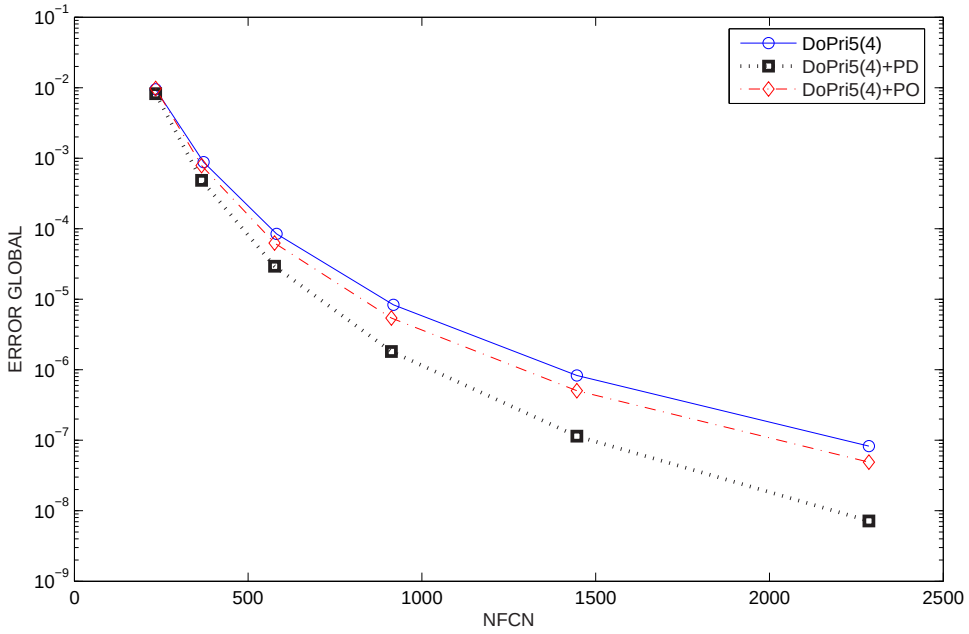

Figura VI.1: Ejemplo 1 (VI.32). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Global.

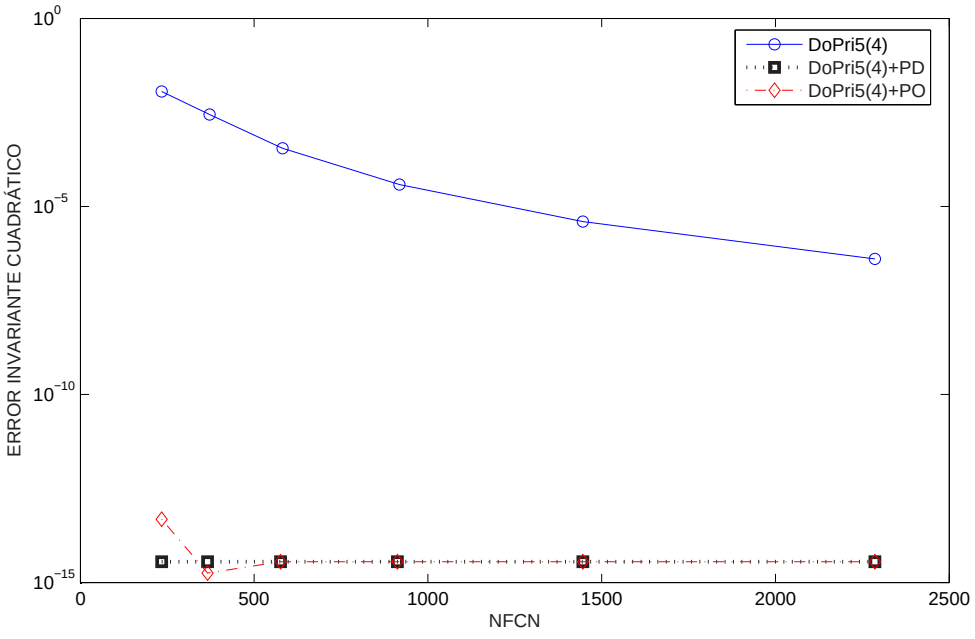

Figura VI.2: Ejemplo 1 (VI.32). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Invariante Cuadrático.

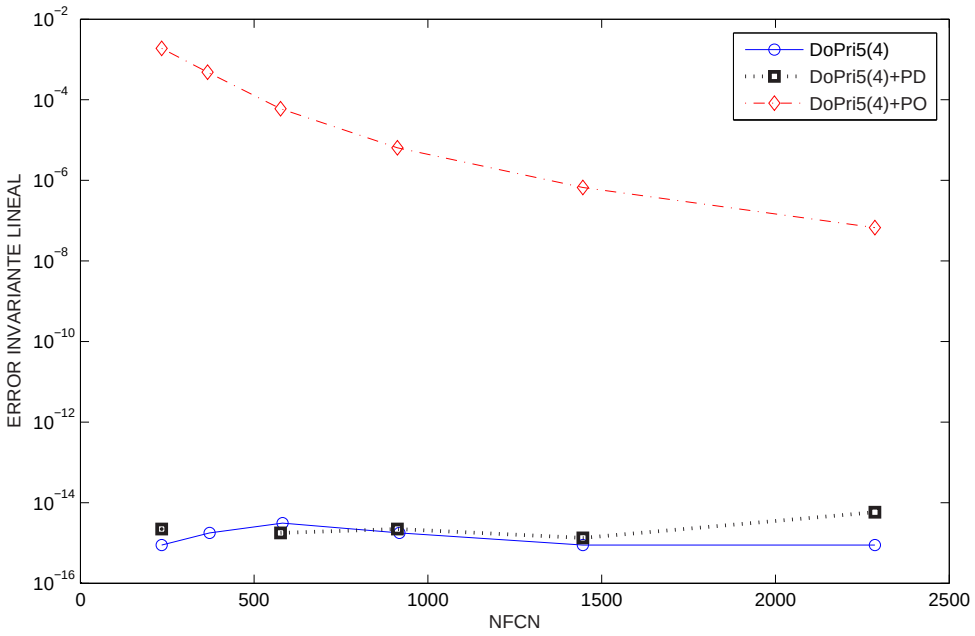

Figura VI.3: Ejemplo 1 (VI.32). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Invariante Lineal.

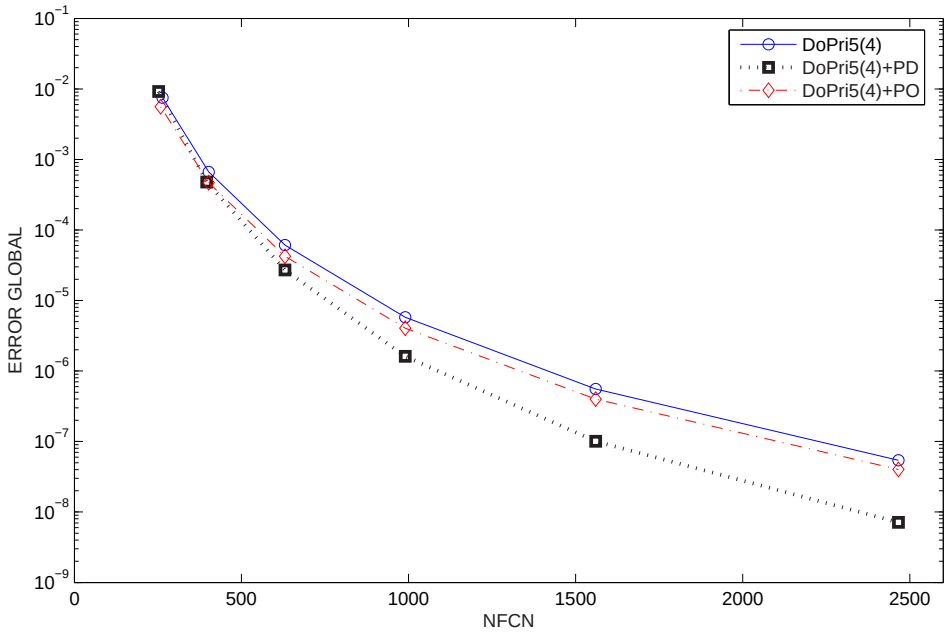

Figura VI.4: Ejemplo 2 (VI.33). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Global.

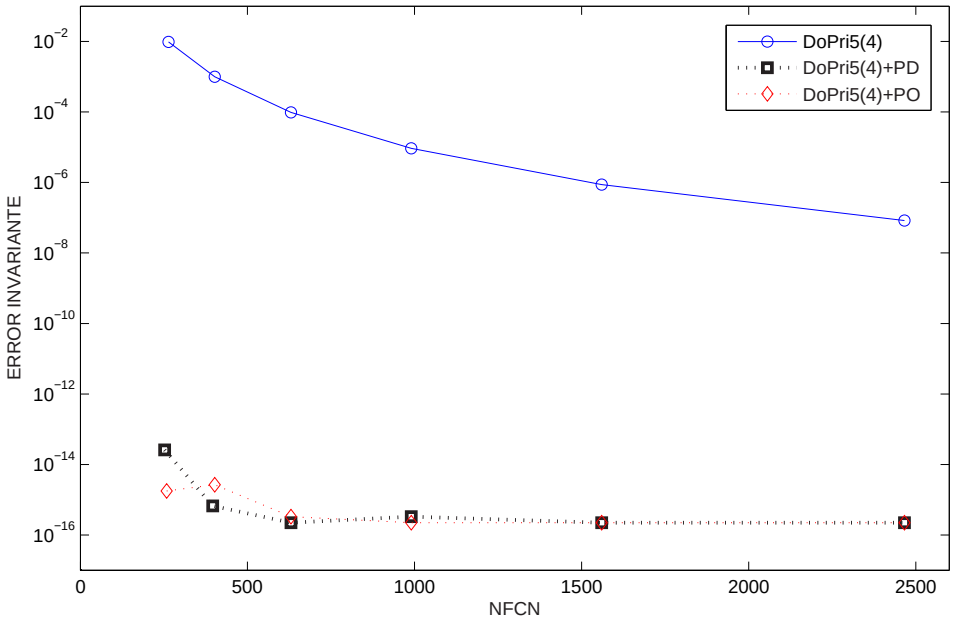

Figura VI.5: Ejemplo 2 (VI.33). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Invariante.

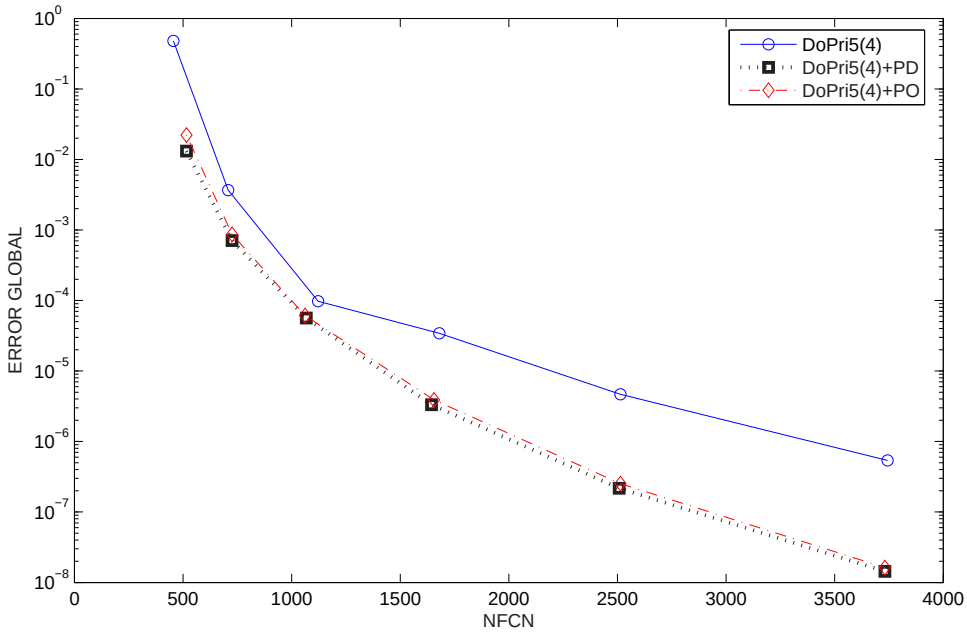

Figura VI.6: Ejemplo 3 (VI.34). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Global.

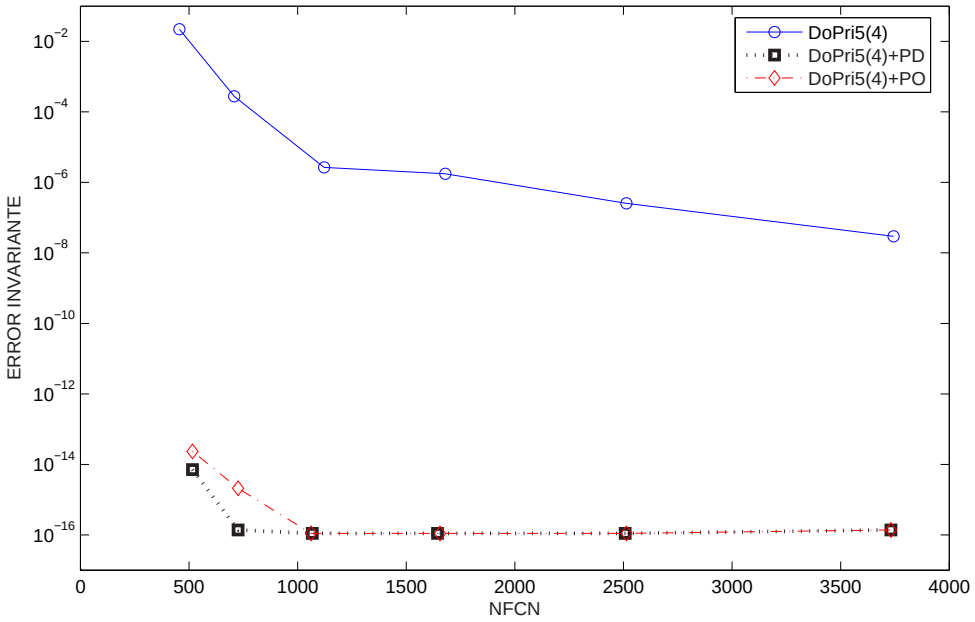

Figura VI.7: Ejemplo 3 (VI.34). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Invariante.

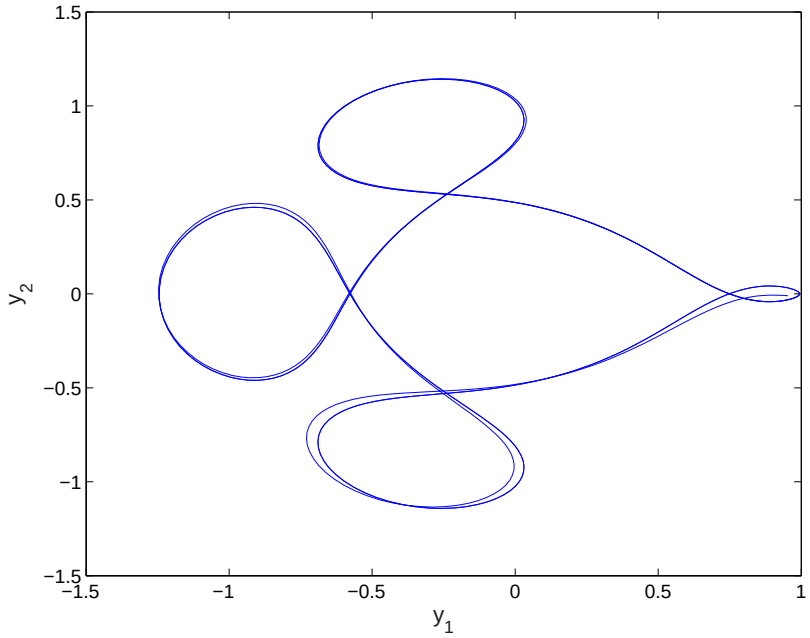

Figura VI.8: Ejemplo 4 (VI.35). DoPri5(4): Órbita Numérica.

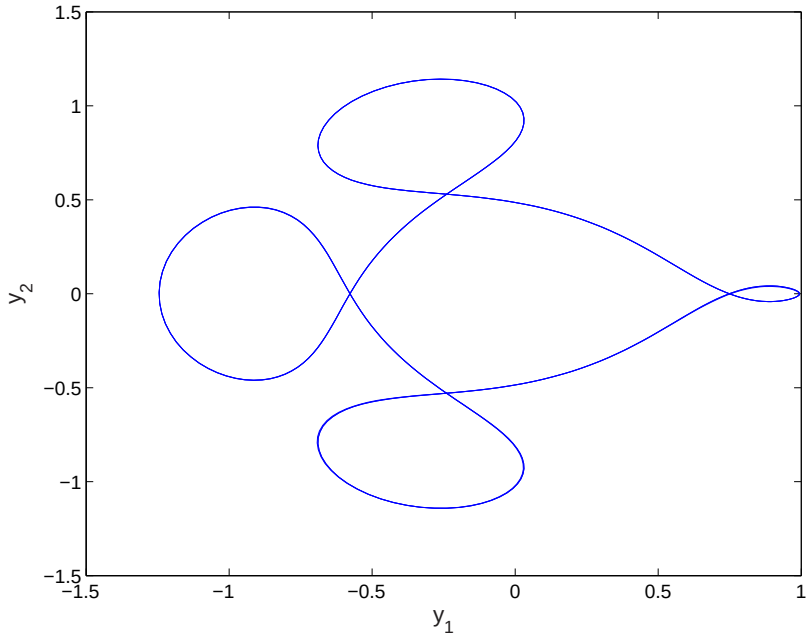

Figura VI.9: Ejemplo 4 (VI.35). DoPri5(4)+PD, DoPri5(4)+PO: Órbita Numérica.

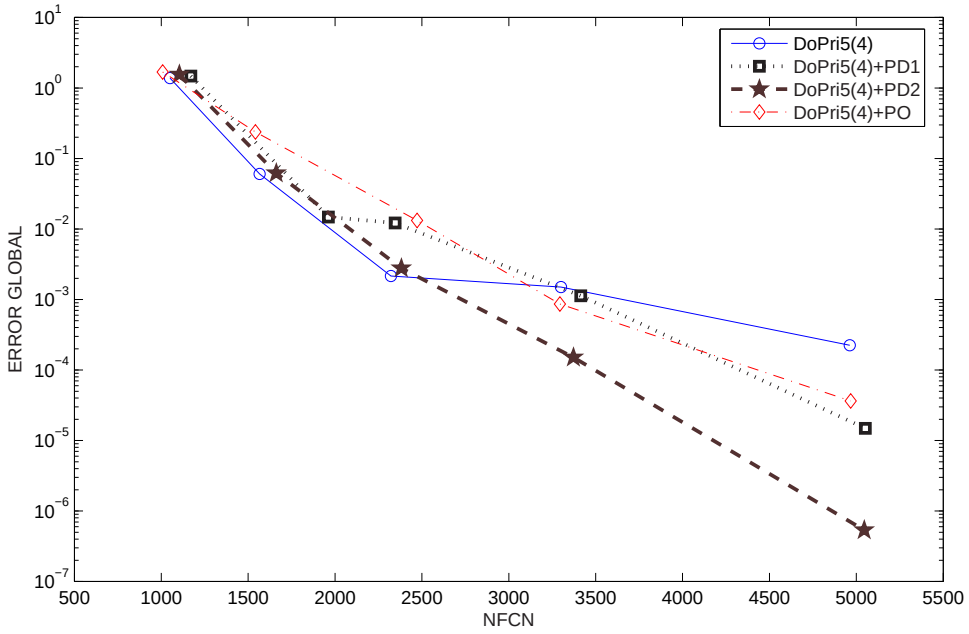

Figura VI.10: Ejemplo 4 (VI.35). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Global.

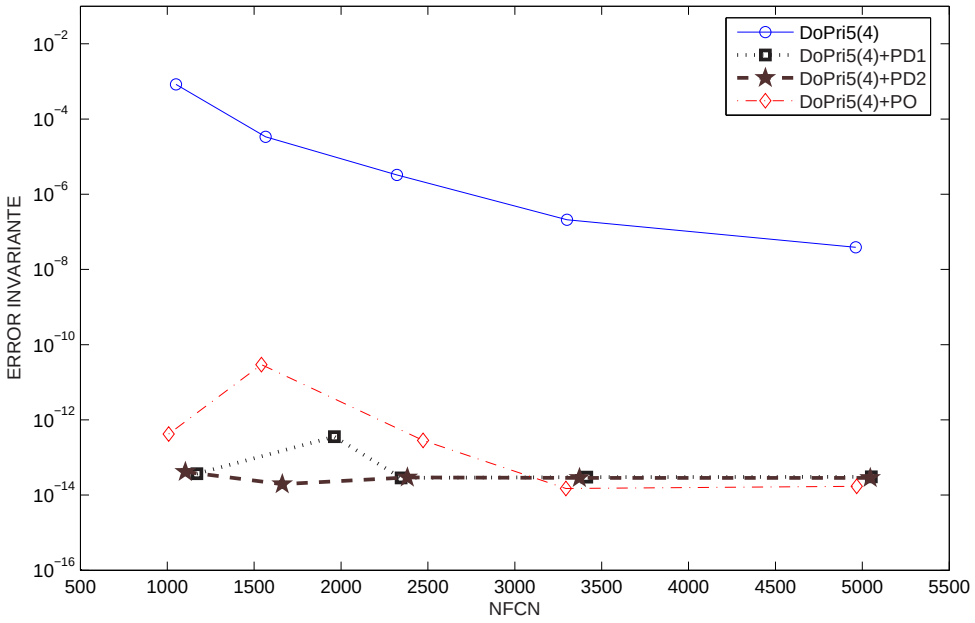

Figura VI.11: Ejemplo 4 (VI.35). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Invariante.

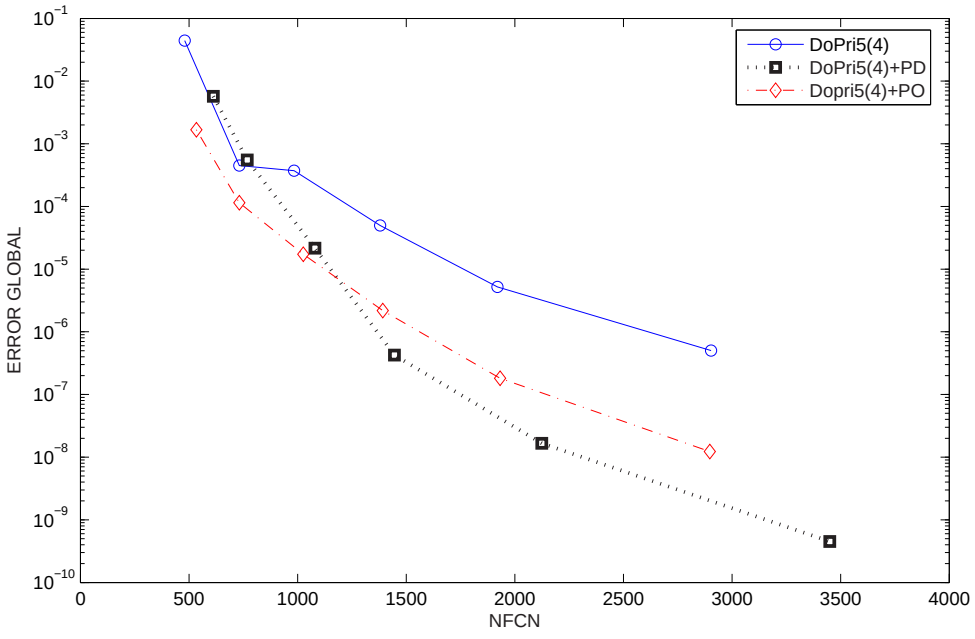

Figura VI.12: Ejemplo 5 (VI.36). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Global.

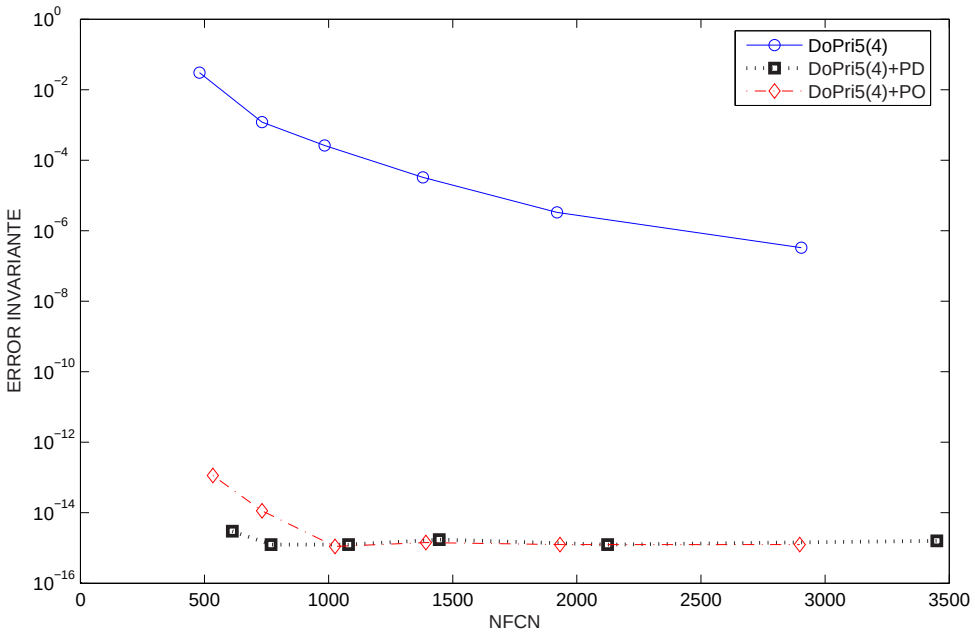

Figura VI.13: Ejemplo 5 (VI.36). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Invariante.

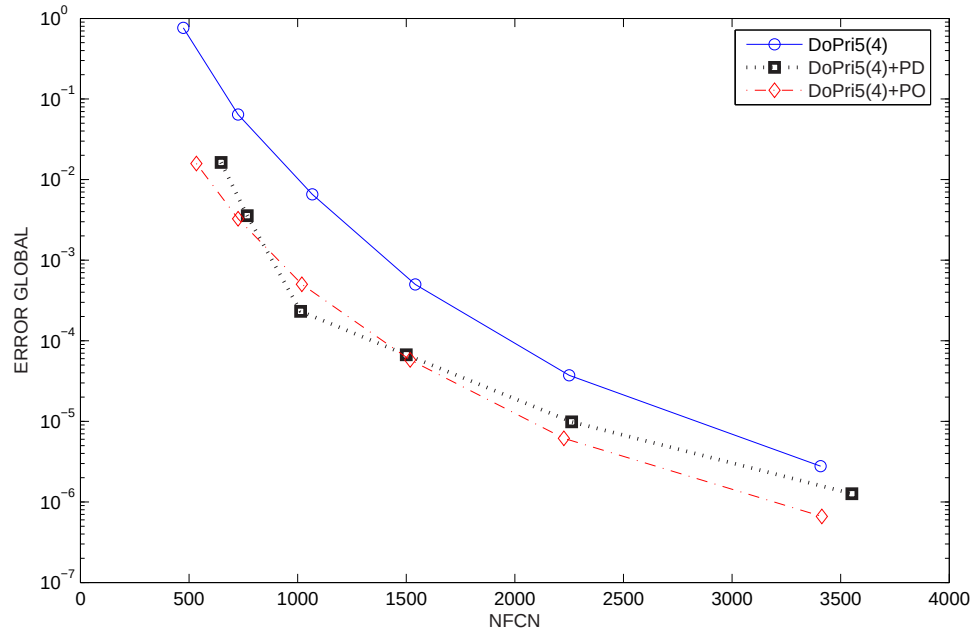

Figura VI.14: Ejemplo 6 (VI.37). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Global.

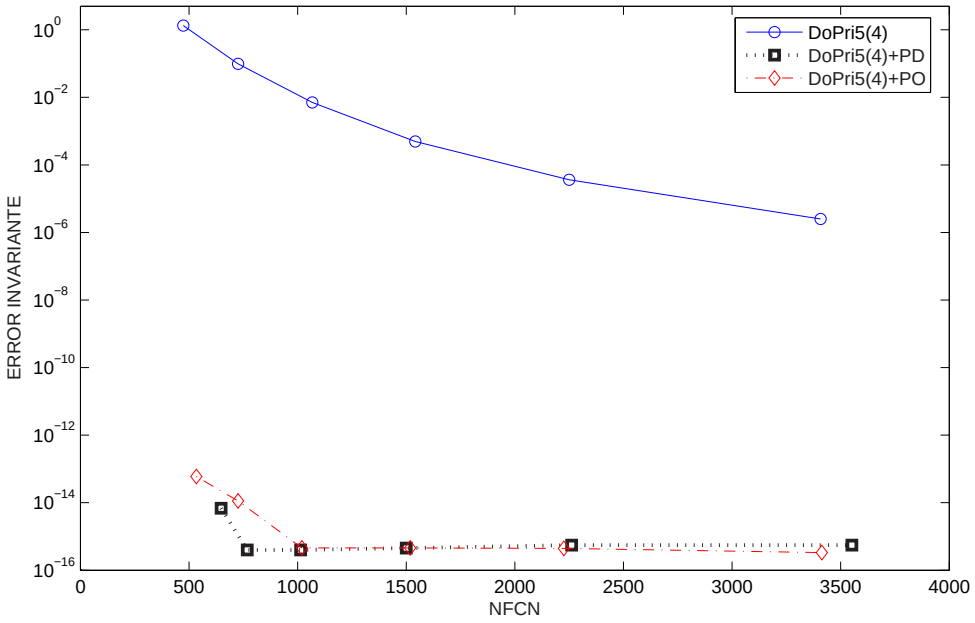

Figura VI.15: Ejemplo 6 (VI.37). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Invariante.

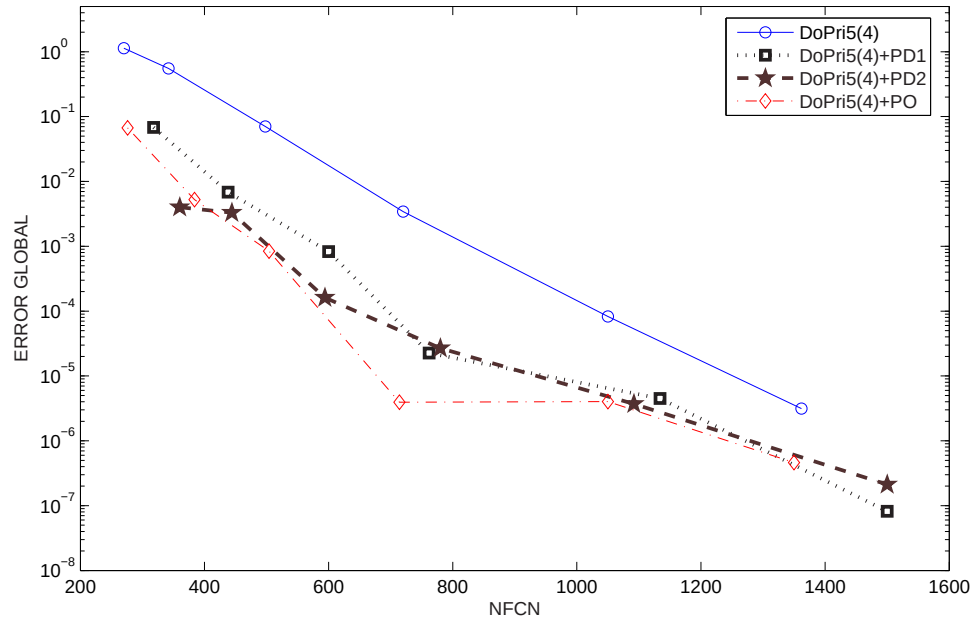

Figura VI.16: Ejemplo 7 (VI.38). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Global.

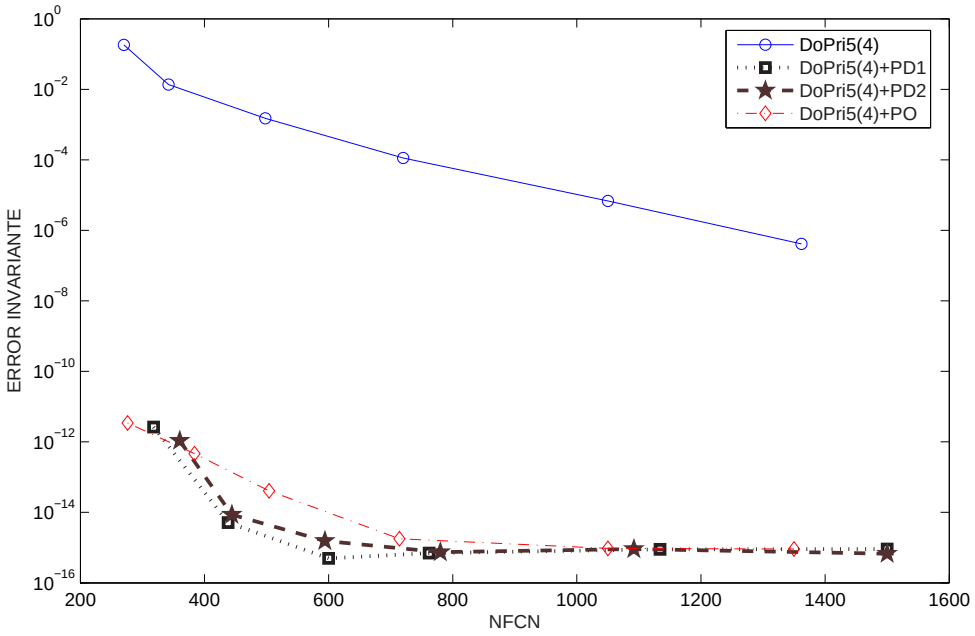

Figura VI.17: Ejemplo 7 (VI.38). DoPri5(4), DoPri5(4)+PD y DoPri5(4)+PO: Error Invariante.

Apéndice A

Conclusiones e investigación futura.

Teniendo presente los resultados teóricos derivados de la investigación y la experimentación numérica presentada en esta memoria podemos concluir lo siguiente:

1) Pocos métodos de interés práctico, a excepción del método de Euler implícito, son capaces de efectuar una integración numérica con la propiedad de estabilidad incondicional sobre la clase de sistemas diferenciales con equilibrios semiestables. Así muchos métodos numéricos con excelentes propiedades de estabilidad lineal y no lineal imponen restricciones al tamaño de paso de cara a obtener la estabilidad de las soluciones numéricas. No obstante lo anterior, y a pesar de que la estabilidad incondicional de los métodos sea una propiedad deseable desde un punto de vista teórico, esta condición no juega un papel relevante en la ejecución de los métodos numéricos de modo que no se necesitan condiciones tan exigentes para garantizar la estabilidad.

2) Considerando la integración numérica de los sistemas diferenciales con equilibrios semiestables correspondientes al caso de variedades centro unidimensionales sobre redes temporales tales que las razones de tamaños de paso consecutivos estén acotadas por una determinada constante mayor que la unidad, podemos establecer resultados de estabilidad para muchos métodos implícitos de tipo Runge-Kutta a condición de que éstos sean fuertemente A-estables. De esta manera, deducimos la estabilidad de las órbitas numéricas en entornos de los equilibrios semiestables para aquellos métodos de Runge-Kutta con la propiedad de estabilidad lineal cuya función de estabilidad lineal posee módulo menor que la unidad en el infinito. La experimentación numérica refleja claramente que la condición

de A-estabilidad fuerte para los métodos de tipo Runge-Kutta es necesaria desde una perspectiva práctica.

3) Continuando el estudio de estabilidad de los métodos numéricos de un paso sobre la clase de sistemas diferenciales con equilibrios semiestables correspondientes al caso de variedades centro unidimensionales hallamos que la condición de A-estabilidad fuerte para los métodos de tipo Rosenbrock constituye también una condición suficiente de cara a la estabilidad de las órbitas numéricas. Asimismo la experimentación numérica llevada a cabo refleja que esta condición es necesaria en la práctica. Esto permite concluir el estudio de estabilidad de los principales métodos de un paso cuando son aplicados a los sistemas diferenciales con equilibrios semiestables correspondientes al caso de variedades centro unidimensionales.

4) En el capítulo quinto de esta memoria establecemos condiciones prácticas que garantizan la contractividad y convergencia de las soluciones que proveen los Métodos Lineales Generales sobre la amplia clase de sistemas diferenciales stiff con constantes de Lipschitz laterales no positivas. Más precisamente, obtenemos resultados interesantes y completamente nuevos de convergencia en intervalos temporales semi-infinitos para aquellos Métodos Lineales Generales irreducibles y algebraicamente estables que poseen una matriz de estabilidad con radio espectral menor que uno en el infinito. Estos resultados de convergencia están basados en un estudio preliminar de la contractividad de las soluciones numéricas que proveen los Métodos Lineales Generales aplicados a sistemas diferenciales stiff con constantes de Lipschitz laterales no positivas. Estos resultados de contractividad y convergencia se aplican a diversas familias de Métodos Lineales Generales, que generalizan los métodos de un paso clásicos Runge-Kutta Gauss, RadauIIA, RadauIA y LobattoIIIC.

5) Finalizamos esta memoria estableciendo una nueva alternativa para la conservación de invariantes conocidos para sistemas diferenciales autónomos. La técnica propuesta está basada en la proyección de métodos Runge-Kutta explícitos de modo que la dirección de proyección viene definida por medio de métodos explícitos encajados. De este modo, obtenemos una técnica de proyección, alternativa a la técnica clásica de proyección ortogonal, que conserva los invariantes del problema a integrar y respeta el orden de consistencia del método original. La técnica propuesta puede ser incluida de modo simple en los códigos

Runge-Kutta adaptativos usuales y posee la virtud de conservar por propia definición los invariantes afines del problema a integrar.

Problemas abiertos

1) Un estudio natural a llevar a cabo en épocas venideras debe ser el tratamiento de los principales métodos de un paso para la integración de sistemas diferenciales que posean equilibrios semiestables para los que la variedad centro posea múltiples dimensiones. Así, en primer lugar, se debe prestar especial atención a establecer el marco hipotético que defina a la clase de problemas a tratar, para posteriormente analizar el comportamiento de los métodos sobre dicha clase de problemas.

2) Una vez estudiada la contractividad y convergencia de los Métodos Lineales Generales aplicados a sistemas diferenciales con constantes de Lipschitz laterales negativas, es necesario estudiar la estabilidad de estos métodos sobre la clase particular de sistemas diferenciales con equilibrios semiestables correspondientes al caso de variedades centro unidimensionales. No obstante lo anterior, debemos notar que, en el caso de integraciones a paso variable, los métodos numéricos vienen definidos por coeficientes dependientes de las razones de paso, lo cual dificulta el análisis. Hasta la fecha, sólo se han podido obtener resultados parciales relativos a la $A(\alpha)-$estabilidad de los métodos.

3) Tras proponer una estrategia para la conservación de invariantes basada en pares encajados de métodos de tipo Runge-Kutta explícitos, debemos ahondar en esta cuestión y analizar una estrategia semejante para la conservación de invariantes a través de métodos multipaso.

Referencias

[1] G. Bader, E. Hairer, Ch. Lubich, *On the stability of semi-implicit methods for ordinary differential equations*, BIT **22** (1982), 211-232.

[2] W.J. Beyn, J. Lorenz, *Center manifolds of dynamical systems under discretization*, Numer. Funct. Anal. Optim. **9** (1987), 381-414.

[3] J.G. Blom, M. Louter-Nool, S. Scholz, J.G. Verwer, *A class of Runge-Kutta methods for solving stiff differential equations*, ZAMM **63** (1983), 13-20.

[4] R.K. Brayton, C.C. Conley, *Some results on the stability and instability of the backward differentiation methods with non-uniform time steps*, IBM Res. Report RC-3964, IBM Watson Research Center, Yorktown Heights, 1972.

[5] K.E. Brenan, S.L. Campbell, L.R. Petzold, *The Numerical Solution of Initial Value Problems in Ordinary Differential-Algebraic Equations*, North Holland Publishing Co., 1989.

[6] K. Burrage, *High order algebraically stable multistep Runge-Kutta methods*, SIAM J. Numer. Anal. **24** (1987), 106-115.

[7] K. Burrage, *Order properties of implicit multivalue methods for ordinary differential equations*, IMA J. Numer. Anal. **8** (1988), 43-69.

[8] K. Burrage, J.C. Butcher, *Non-linear stability of a general class of differential equation methods*, BIT **20** (1980), 185-203.

[9] K. Burrage, W.H. Hundsdorfer, J.G. Verwer, *A study of B-convergence of Runge-Kutta methods*, Computing **36** (1986), 17–34.

[10] J.C. Butcher, *A stability property of implicit Runge-Kutta methods*, BIT **15** (1975), 358-361.

[11] J.C. Butcher, *Linear and non-linear stability for general linear methods*, BIT **27** (1987), 182-189.

[12] J.C. Butcher, *The equivalence of algebraic stability and AN-stability*, BIT **27** (1987), 510-533.

[13] J.C. Butcher, A.D. Heard, *Stability of numerical methods for ordinary differential equations*, Numerical Algorithms **31** (2002), 59-73.

[14] M. Calvo, S. González-Pinto, J.I. Montijano *On the convergence of Runge-Kutta methods for stiff nonlinear differential equations*, Numer. Math. **81** (1998), 31-51.

[15] M. Calvo, S. González-Pinto, J.I. Montijano, *Runge-Kutta methods for the numerical solution of stiff semilinear systems*, BIT **40** (2000), 611-639.

[16] M. Calvo, T. Grande, R.D. Grigorieff, *On the zero stability of the variable order, variable stepsize BDF-formulas*, Numer. Math. **57** (1990), 39-50.

[17] M. Calvo, D. Hernández-Abreu, J.I. Montijano, L. Rández, *Explicit Runge-Kutta methods for the preservation of invariants*, 2004. Informe técnico. Departamento Matemática Aplicada. Universidad Zaragoza.

[18] M. Calvo, D. Hernández-Abreu, J.I. Montijano, L. Rández, *On the preservation of invariants by explicit Runge-Kutta methods*, aceptado para publicación en SIAM J. Sci. Comp., diciembre de 2005.

[19] M. Calvo, D. Hernández-Abreu, J.I. Montijano, L. Rández, *Preservation of quadratic invariants by means of explicit Runge-Kutta methods*, Proceedings of the International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE-2004, University of Uppsala, 34-38.

[20] M. Calvo, F. Lisbona, J.I. Montijano, *On the stability of variable stepsize Nordsieck BDF methods*, SIAM J. Numer. Anal. **24** (1987), 844-854.

[21] M. Calvo, J.I. Montijano, L. Rández, *A_0-stability of variable stepsize BDF methods*, J. Comput. Appl. Math. **45** (1993), 29-39.

[22] M.P. Calvo, M.A. López-Marcos, J.M. Sanz-Serna, *Variable step implementation of geometric integrators*, Appl. Numer. Math. **28** (1998), 1-16.

[23] M.P. Calvo, J. Sanz-Serna, *Numerical Hamiltonian Problems*, Chapman & Hall (1994).

[24] J.K. Carr, *Applications of center manifolds theory*, Springer-Verlag New York, 1980.

[25] R.J. Charron, M. Hu, *A-contractivity of linearly implicit multistep methods*, SIAM J. Numer. Anal. **32** (1995), 285-295.

[26] E.A. Coddington, N. Levinson, *Theory of ordinary differential equations*, McGraw-Hill Book Company, Inc. New York, 1955.

[27] G.J. Cooper, *Stability of Runge–Kutta methods for trajectory problems*, IMA J. Numer. Anal. **7** (1987), 1–13.

[28] M. Crouzeix, W.H. Hundsdorfer, M.N. Spijker, *On the existence of solutions to the algebraic equations in implicit Runge-Kutta methods*, BIT **23** (1983), 84-91.

[29] C.F. Curtiss, J.O. Hirschfelder, *Integration of stiff equations*, Proc. Nat. Acad. Sci. **38** (1952), 235-243.

[30] G. Dahlquist, *A special stability property for linear multistep methods*, BIT **3** (1963), 27-43.

[31] G. Dahlquist, *Error analysis for a class of methods for stiff nonlinear initial value problems*, Numerical Analysis, Lecture Notes in Math., Nº 506, 60-74, Dundee 1975.

[32] M. De Guzmán, *Ecuaciones diferenciales ordinarias. Teoría de estabilidad y control*, Ed. Alhambra, 1975.

[33] K. Dekker, J.G. Verwer, *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, North-Holland, Amsterdam-New York-Oxford, 1984.

[34] N. Del Buono, C. Mastroserio, *Explicit methods based on a class of four stage fourth order Runge–Kutta methods for preserving quadratic laws*, J. Comput. Appl. Math. **140** (2002), 231–243.

[35] A. Dhooge, W. Govaerts, Y.A. Kuznetsov, *MATCONT: A MATLAB package for numerical bifurcation analysis of ODEs*, ACM Trans. Math. Software **29** (2003), 141-164.

[36] J.R. Dormand, P.J. Prince, *A family of embedded Runge-Kutta formulae*, J. Comput. Appl. Math. **6** (1980), 19-26.

[37] W.H. Enright, T.E. Hull, B. Lindberg, B., *Comparing numerical methods for stiff systems of ODEs*, BIT **15** (1975), 10-48.

[38] W.H. Enright, J.D. Pryce, *Two FORTRAN packages for Assessing Initial Value Problems*, ACM Trans. Math. Software **13** (1987), 1-27.

[39] R. Frank, J. Schneid, C.W. Ueberhuber, *Order results for implicit Runge-Kutta methods applied to stiff systems*, SIAM J. Numer. Anal. **22** (1985), 515-534.

[40] S. González-Pinto, *Differential systems with semi-stable equilibria and numerical methods*, Numer. Math. **96** (2003), 253-268.

[41] S. González-Pinto, D. Hernández-Abreu, *On the contractivity and convergence of general linear methods*, enviado a SIAM J. Numer. Anal., abril 2006.

[42] S. González-Pinto, D. Hernández-Abreu, *On the contractivity of the matrices related to general linear methods*, Numerical Analysis Report NA/222, 21*st* Biennial Conference on Numerical Analysis, University of Dundee, p. 23.

[43] S. González-Pinto, D. Hernández-Abreu, *Semi-implicit methods for differential systems with semi-stable equilibria*, Appl. Num. Math. **56** (2006), 210-221.

[44] S. González-Pinto, D. Hernández-Abreu, *Stable Runge-Kutta integrations for differential systems with semi-stable equilibria*, Numer. Math. **97** (2004), 473-491.

[45] S. González-Pinto, D. Hernández-Abreu, *Strong A-Acceptability for rational functions*, BIT **43** (2003), 555-561.

[46] W. Govaerts, *Numerical bifurcation analysis of ODEs*, J. Comput. Appl. Math. **125** (2000), 57-68.

[47] W. Govaerts, Y.A. Kuznetsov, B. Sijnave, *Numerical methods for the generalized Hopf bifurcation*, SIAM J. Numer. Anal. **38** (2000), 329-346.

[48] R.D. Grigorieff, *Stability of multistep-methods on variable grids*, Numer. Math. **42** (1983), 359-377.

[49] J. Guckenheimer, P. Holmes, *Nonlinear oscillations, dynamical systems and bifurcations of vectors fields*, Springer-Verlag New York, 2^{nd} ed., 1986.

[50] E. Hairer, Ch. Lubich, G. Wanner, *Geometric numerical integration: structure preserving algorithms for ordinary differential equations*, Springer Verlag Berlin, 2002.

[51] E. Hairer, S.P. Nørsett, G. Wanner, *Order stars and stability theorems*, BIT **18** (1978), 475-498.

[52] E. Hairer, S.P. Nørsett, G. Wanner, *Solving ordinary differential equations I. Nonstiff problems*, Springer–Verlag, 1980.

[53] E. Hairer, G. Wanner, *Solving ordinary differential equations II. Stiff and differential algebraic problems*, Springer–Verlag, 2^{nd} ed., 1996.

[54] E. Hairer, M. Zennaro, *On error growth functions of Runge-Kutta methods*, Appl. Num. Math. **22** (1996), 205-216.

[55] J. Hale, H. Koçak, *Dynamics and bifurcations*, Springer-Verlag New York Inc., 1991.

[56] D.J. Higham, N.J. Higham, *Matlab guide*, SIAM Philadelphia, 2*nd* ed., 2005.

[57] C. Huang, Q. Chang, A. Xiao, *B-convergence of general linear methods for stiff problems*, Appl. Numer. Math. **47** (2003), 31-44.

[58] W. Hundsdorfer, *On the error of General Linear Methods for stiff dissipative differential equations*, IMA J. Numer. Anal. **14** (1994), 363-379.

[59] W. Hundsdorfer, J.G. Verwer, *Numerical solution of time-dependant advection-diffusion reaction equations*, Springer 2003.

[60] A. Iserles, A. Zanna, *Preserving algebraic invariants with Runge-Kutta methods*, J. Comput. Appl. Math. **125** (2000), 69–81.

[61] P.E. Kloeden, J. Lorenz, *Stable attracting sets in dynamical systems and in their one-step discretizations*, SIAM J. Numer. Anal. **23** (1986), 986-995.

[62] A.N. Kolmogorov, S.V. Fomin, *Introductory Real Analysis*, Dover Publications, Inc. New York. Revised English Edition, 1975.

[63] J.F.B.M. Kraaijevanger and J. Schneid, *On the unique solvability of the Runge-Kutta equations*, Numer. Math. **59** (1991), 129-157.

[64] Y.A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 3*rd* ed., 2004.

[65] Y.A. Kuznetsov, *Numerical normalization techniques for all codim 2 bifurcations of equilibria in ODEs*, SIAM J. Numer. Anal. **36** (1999), 1104-1124.

[66] Y.A. Kuznetsov, *Practical computation of normal forms on center manifolds at degenerate Bogdanov-Takens bifurcations*, Int. J. Bifurcation & Chaos **15** (2005), 3535-3546.

[67] Y.A. Kuznetsov, H.G.E. Meijer, *Numerical normal forms for codim 2 bifurcations of fixed points with at most two critical eigenvalues*, SIAM J. Sci. Comp. **26** (2005), 1932-1954.

[68] J.D. Lambert, *Numerical methods for ordinary differential equations*, Wiley, 1991.

[69] S. Larsson, V. Thomée, *Partial differential equations with numerical methods*, Springer, 2003.

[70] D. Lewis, N. Nigam, *Geometric integration on spheres and some interesting applications*, J. Comput. Appl. Math. **151** (2003), 141–170.

[71] S. Li, *Stability and B-convergence properties of multistep Runge-Kutta methods*, Math. Comp. **69** (2000), 1481-1504.

[72] S.P. Nørsett, *Semi-explicit Runge-Kutta methods*, Report Mathematics and Computation No. 6/74, Dept. Mathematics, University of Trondheim, Norway.

[73] J.M. Ortega, W.C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York, 1970.

[74] L. Perko, *Differential equations and dynamical systems*, Springer-Verlag New York, 2^{nd} ed., 1993.

[75] H. Poincaré, *Mémoire sur les cóurbes définies par les équations différentielles I-VI*, Ouevre I, Gauthier-Villar, Paris, 1880-1890.

[76] H.H. Robertson, *The solution of a set of reaction rate equations*, J. Walsh ed.: Numer. Anal., an Introduction, Academic Press (1966), 178-182.

[77] H.H. Rosenbrock, *Some general implicit processes for the numerical solution of differential equations*, Computer J. **5** (1962-63), 329-330.

[78] D. Ruelle, *Elements of differentiable dynamics and bifurcation theory*, Academic Press, New York, 1989.

[79] J. Sand, *A_0-contractivity of the variable-step BDF's of order less than four*, BIT **25** (1985), 391-398.

[80] J. Schropp, *Conserving first integrals under discretization with variable step size integration procedures*, J. Comput. Appl. Math. **115** (2000), 503–517.

[81] S. Scholz, J.G. Verwer, *Rosenbrock methods and time-lagged Jacobian matrices*, Beiträge zur Numer. Math. **11** (1983), 173-183.

[82] M. Slodička, I. Cimrák, *An iterative approximation scheme for the Landau-Lifshitz-Gilbert equation*, J. Comput. Appl. Math. **169** (2004), 17–32.

[83] M. Slodička, I. Cimrák, *Numerical study of nonlinear ferromagnetic materials*, Appl. Numer. Math. **46** (2003), 95–111.

[84] J. Stoer, R. Bulirsch, *Introduction to numerical analysis*, Springer Verlag New York, 1983.

[85] A.M. Stuart, A.R. Humphries, *Dynamical systems and numerical analysis*. Cambridge University Press, Cambridge Monographs on Applied and Computational Mathematics, New York, 1996.

[86] J.G. Verwer, *On generalized Runge-Kutta methods using an exact Jacobian at a non-step point*, ZAMM **60** (1980), 263-265.

[87] J. Von Neumann, *Eine Spektraltheorie für allgemeine Operatoren eines unitärem Raumes.* Math. Nachrichten **4** (1951), 258-281.

[88] S. Wiggins, *Introduction to applied nonlinear dynamical systems and chaos*, Springer-Verlag New York, 1990.

[89] A. Wintner, *The analytical foundations of celestial mechanics*, Princeton University Press, 1941.

[90] S. Wolfram, *The Mathematica book*, Wolfram Media, 5*th* ed., 2003.