USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Aspectos Teóricos dos Métodos Runge-Kutta

Lucía Suárez Suárez

2021/2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Aspectos Teóricos dos Métodos Runge-Kutta

Lucía Suárez Suárez

Xullo, 2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Matemática Aplicada
Título: Aspectos Teóricos dos Métodos Runge-Kutta
Breve descrición do contido
Na materia "Métodos Numéricos en Optimización e Ecuacións Diferenciais" viuse unha breve introdución aos métodos Runge-Kutta. O traballo consistiría en ampliar o estudo teórico destes métodos. Abordaría polo menos os seguintes puntos: A formulación dos conceptos centrais no estudo dos métodos numéricos para EDO (converxencia, consistencia, orde de consistencia, orde de converxencia e estabilidade) no marco xeral dos métodos dun paso. A inclusión dos métodos Runge-Kutta (tanto explícitos como implícitos) en dito marco xeral e a dedución das propiedades da función incremento e as consecuencias desta. A obtención das condicións necesarias e suficientes para que un método Runge-Kutta sexa de orde p, para $p = 1, 2, 3$ e 4. Algunhas consideracións sobre a estabilidade numérica e a función de estabilidade.
Recomendacións
Coñecer os contidos sobre métodos numéricos de resolución de EDOs abordados na materia "Métodos Numéricos en Optimización e Ecuacións Diferenciais".
Outras observacións

Índice

Resumo

Neste traballo presentaremos diferentes propiedades e resultados referidos aos métodos Runge-Kutta.

Comezaremos co estudo dos métodos dun paso, abordando a consistencia, estabilidade, converxencia e orde de devanditos métodos. Tras isto, centrarémonos no estudo destes conceptos nunha familia concreta dos métodos dun paso, que son os chamados métodos Runge-Kutta. Finalmente, estudaremos algúns aspectos teóricos relativos á estabilidade numérica dos métodos Runge-Kutta, que introduciremos en primeiro lugar no caso dos métodos de Euler implícito e explícito, para logo abordar estes aspectos no caso dun método Runge-Kutta xeral.

Abstract

In this project we will present different properties and results concerning Runge-Kutta methods.

We wil start with the study of one-step methods, addressing the consistency, stability, convergence and order of such methods. After this, we will focus on the study of these concepts in a particular family of one-step methods, which are the so-called Runge-Kutta methods. Finally, we will study some theoretical aspects concerning the numerical stability of Runge-Kutta methods, which we will first introduce in the case of implicit and explicit Euler methods, and then address aspects in the case of a general Runge-Kutta method.

Introdución

O obxectivo deste traballo consiste en ampliar algúns coñecementos adquiridos na asignatura do grao "Métodos Numéricos en Optimización e Ecuacións Diferencias". En concreto, os conceptos relacionados cos métodos Runge-Kutta.

Comezaremos co estudo dos métodos dun paso, aportando defincións e algúns resultados para o estudo da súa consistencia, estabilidade, converxencia e orde, para logo abordar unha familia concreta destes métodos. Estes van a ser os métodos Runge-Kutta.

A continuación estudaremos os métodos Runge-Kutta. Primeiro daremos unha definición e comentaremos os diferentes tipos que existen (explícitos, semiimplícitos e implícitos), para logo introducilos no marco xeral dos métodos dun paso.

Para a función incremento destos métodos, veremos algunhas propiedades como son a continuidade, condición de Lipschitz e regularidade que nos proporcionarán a consistencia, estabilidade e converxencia dos métodos Runge-Kutta, e xa por último, abordaremos a orde, iniciando cun método Runge-Kutta explícito de 3 etapas, para logo abordar o caso xeral, que é máis complicado.

Para o estudo dos métodos Runge-Kutta, imos empregar conceptos e resultados doutras asignaturas do grao. Como, por exemplo, de normas de matrices, que foron vistos na asignatura "Análise Numérico Matricial", o Teorema da Función Implícita visto na asignatura "Diferenciación de Funcións de varias Variables Reais" e os desenvolvementos de Taylor vistos na asignatura "Continuidade e Derivabilidade de Funcións dunha Variable Real".

Xa para rematar, comentaremos aspectos teóricos relativos a estabilidade numérica dos métodos Runge-Kutta. Comezaremos este capítulo vendo uns casos particulares de métodos Runge-Kutta, os métodos de Euler implícito e explícito, e introducindo sistemas, para logo poder tratar a estabilidade no caso dun método Runge-Kutta xeral.

Ao longo do traballo introduciranse ademais, diferentes exemplos para axudar a comprensión de certos conceptos.

Capítulo 1

Estudo dos métodos dun paso

Este capítulo está dedicado ao estudo dos métodos dun paso. Defínense os conceptos de consistencia, estabilidade, converxencia e orde de converxencia, e establécense os resultados principais relativos a estes conceptos. A referencia seguida neste capítulo é [3].

1.1. Introdución

Imos considerar o problema de valor inicial

$$\begin{cases} y'(t) = f(t, y(t)) \text{ con } t \in [t_0, t_0 + T], & \text{(1.1a)} \\ y(t_0) = \eta_0 \text{ dado en } \mathbb{R}^m, & \text{(1.1b)} \end{cases}$$

onde t é unha variable independente e y unha variable dependente.

Supoñeremos sempre que f é unha función de $[t_0, t_0+T] \times \mathbb{R}^m$ en $\mathbb{R}^m$, con $m \geq 1$, que cumpre as seguintes hipóteses:

$$\begin{cases} (i)\ f \ \text{ é continua en } [t_0, t_0 + T] \times \mathbb{R}^m \\ (ii)\ f \ \text{ verifica a condición de Lipschitz:} \\ \quad \exists L > 0 \ \text{ tal que } \ \|f(t, y) - f(t, z)\| \leq L\|y - z\|, \\ \quad \forall y, z \in \mathbb{R}^m, \forall t \in [t_0, t_0 + T] \end{cases} \tag{1.2}$$

A constante L denomínase *constante de Lipschitz*.

Baixo estas hipóteses, o problema de valor inicial (1.1) ten unha única solución global $y : [t_0, t_0 + T] \longrightarrow \mathbb{R}^m$. Ademais, $y \in C^1([t_0, t_0 + T]; \mathbb{R}^m)$.

Faremos uso dunha subdivisión $t_0 < t_1 < \cdots < t_{n+1} < t_N \leq t_0 + T$ do intervalo $[t_0, t_0 + T]$

onde $t_j = t_0 + jh,\ j = 1, \dots, N$, $N = \lceil \frac{T}{h} \rceil$ e $h > 0$ será o paso que, neste caso, é constante (para $s \in \mathbb{R}$, $\lceil s \rceil$ denota a parte enteira de s, é dicir, $\lceil s \rceil = \text{máx}\{k \in \mathbb{Z}/k \leq s\}$).

Centrarémonos na resolución aproximada de (1.1) baixo as hipóteses (1.2), mediante métodos dun paso. Podemos escribir ditos métodos da forma:

$$y_{n+1} = y_n + h\phi_f(t_n, y_n; h),\ n = 0, \dots, N-1, \tag{1.3}$$

$$y_0 = \eta_h \text{ dado en } \mathbb{R}^m, \tag{1.4}$$

onde ϕ_f é unha función continua de $[t_0, t_0 + T] \times \mathbb{R}^m \times [0, h^*]$ en $\mathbb{R}^m$, con $h^* > 0$, que depende únicamente da función f, e onde η_h é unha aproximación de η_0.

Por exemplo, o método de Euler explícito,

$$y_{n+1} = y_n + hf(t_n, y_n),\ n = 0, \dots, N-1,$$

é un método dun paso. Este método encaixa na forma xeral (1.3), sendo

$$\phi_f(t, y; h) = f(t, y).$$

Observemos que podemos escribir o esquema (1.3) da forma:

$$\frac{y_{n+1} - y_n}{h} - \phi_f(t_n, y_n; h) = 0;$$

así, imos considerar para $n = 0, \dots, N-1$

$$\tau_{n+1} = \frac{y(t_{n+1}) - y(t_n)}{h} - \phi_f(t_n, y(t_n); h) \tag{1.5}$$

onde $y = y(t)$ é unha solución exacta da ecuación diferencial ordinaria (EDO) $y'(t) = f(t, y)$.

Ademais, imos definir o *erro local de discretización* no instante t_{n+1} mediante

$$T_{n+1} = h\tau_{n+1}. \tag{1.6}$$

Observación 1.1. Si se supón a hipótese de localización, $y_n = y(t_n)$, entón

$$T_{n+1} = y(t_{n-1}) - \tilde{y}_{n+1}, \tag{1.7}$$

onde $\tilde{y}_{n+1}$ denota o valor que se obtería aplicando o método a partir de $y_n = y(t_n)$, é dicir,

$$\tilde{y}_{n+1} = y(t_n) + h\phi_f(t_n, y(t_n); h).$$

En efecto,

$$y(t_{n+1}) - \tilde{y}_{n+1} = y(t_{n+1}) - y(t_n) - h\phi_f(t_n, y_n; h) = h\tau_{n+1} = T_{n+1}.$$

1.2. Estudo da consistencia, estabilidade e converxencia

Cabe destacar, antes de comezar, que dado que en $\mathbb{R}^m$ todas as normas son equivalentes, as definicións que imos introducir ao longo desta sección non dependerán da norma escollida en $\mathbb{R}^m$.

1.2.1. Consistencia

Definición 1.2. O esquema (1.3) dise que é *consistente* coa EDO (1.1a) se:

$$\lim_{h\to 0^+}\left(\max_{1\le n\le N}\|\tau_n\|\right)=0$$

para toda solución exacta $y=y(t)$ da EDO (1.1a).

Lema 1.3. *Unha condición necesaria e suficiente para que o método (1.3) sexa consistente é que:*

$$\phi_f(t,y;0)=f(t,y)\quad \forall t\in[t_0,t_0+T],\ \forall y\in\mathbb{R}^m. \tag{1.8}$$

Demostración. Sexa $y=y(t)$ unha solución exacta da EDO (1.1a). Consideremos o vector τ_{n+1} definido mediante (1.5). Podemos escribir

$$\tau_{n+1}=\frac{y(t_{n+1})-y(t_n)}{h}-y'(t_n)+f(t_n,y(t_n))-\phi_f(t_n,y(t_n);0)+\phi_f(t_n,y(t_n);0)-\phi_f(t_n,y(t_n);h),$$

é dicir,

$$\tau_{n+1}=f(t_n,y(t_n))-\phi_f(t_n,y(t_n);0)+R_n(h)+\tilde{R}_n(h) \tag{1.9}$$

onde $R_n(h)=\frac{y(t_{n+1})-y(t_n)}{h}-y'(t_n)$ e $\tilde{R}_n(h)=\phi_f(t_n,y(t_n);0)-\phi_f(t_n,y(t_n);h)$. Demostraremos que

$$\lim_{h\to 0^+}\max_{0\le n\le N-1}\|R_n(h)\|=0 \tag{1.10}$$

e

$$\lim_{h\to 0^+}\max_{0\le n\le N-1}\|\tilde{R}_n(h)\|=0. \tag{1.11}$$

Debido a regra de Barrow,

$$R_n(h)=\frac{1}{h}\int_{t_n}^{t_{n+1}}y'(s)ds-y'(t_n)=\frac{1}{h}\int_{t_n}^{t_{n+1}}(y'(s)-y'(t_n))ds.$$

Polo tanto:

$$\|R_n(h)\|\le\frac{1}{h}\int_{t_n}^{t_{n+1}}\|(y'(s)-y'(t_n))\|ds\le\max_{s\in[t_n,t_{n+1}]}\|y'(s)-y'(t_n)\|$$

e entón,

$$\max_{0\le n\le N-1}\|R_n(h)\| \le \max_{0\le n\le N-1}\max_{s\in[t_n,t_{n+1}]}\|y'(s)-y'(t_n)\| \le \max_{\substack{s,t\in[t_0,t_0+T]\\ |s-t|\le h}}\|y'(s)-y'(t)\|.$$

E como y' é uniformemente continua en $[t_0,t_0+T]$, xa que y' é continua e $[t_0,t_0+T]$ compacto, verifícase que

$$\forall\varepsilon>0\ \exists\ \delta>0 \text{ tal que } s,\ t\in[t_0,t_0+T],\ |s-t|\le\delta \Rightarrow \|y'(s)-y'(t)\|\le\varepsilon.$$

Se tomamos $0<h\le\delta$, temos que para $s,\ t\in[t_0,t_0+T]$ con $|s-t|\le h$ cúmprese que $\|y'(s)-y'(t)\|\le\varepsilon$ e en consecuencia

$$\max_{0\le n\le N-1}\|R_n(h)\| \le \max_{\substack{s,t\in[t_0,t_0+T]\\ |s-t|\le h}}\|y'(s)-y'(t)\|\le\varepsilon,$$

polo que queda probado (1.10). Para demostrar (1.11) consideramos a función

$$\varphi:(t,h)\in[t_0,t_0+T]\times[0,h^*]\longrightarrow\phi_f(t,y(t);h)\in\mathbb{R}^m,$$

que é continua no compacto $[t_0,t_0+T]\times[0,h^*]$, o que implica que φ é uniformemente continua. Hai que probar que para todo $\varepsilon>0$ existe $\tilde{h}(\varepsilon)>0,\ \tilde{h}(\varepsilon)\le h^*$ tal que para $0<h\le h^*(\varepsilon)$ se verifica que

$$\max_{0\le n\le N-1}\|\phi_f(t_n,y(t_n);0)-\phi_f(t_n,y(t_n);h)\| = \max_{0\le n\le N-1}\|\varphi(t_n,0)-\varphi(t_n,h)\|\le\varepsilon.$$

Como φ é uniformemente continua, para todo $\varepsilon>0$, existe $\delta=\delta(\varepsilon)>0,\ \delta\le h^*$ tal que para $s,\ t\in[t_0,t_0+T]$, $h,\ \tilde{h}\in[0,h^*]$, con $|t-s|\le\delta,\ \|h-\tilde{h}\|\le\delta$ tense que

$$\|\varphi(t,h)-\varphi(s,\tilde{h})\|\le\varepsilon$$

Polo tanto, para todo $h\in(0,\delta]$, e para todo $n=0,\dots,N-1$ tense que

$$\|\phi_f(t_n,y(t_n);h)-\phi_f(t_n,y(t_n);0)\| = \|\varphi(t_n,h)-\varphi(t_n,0)\|\le\varepsilon$$

o que implica que

$$\max_{0\le n\le N-1}\|\phi_f(t_n,y(t_n);h)-\phi_f(t_n,y(t_n);0)\|\le\varepsilon.$$

Así pois, é suficiente tomar $\tilde{h}(\varepsilon)=\delta$.

Suficiencia: Facendo uso de (1.9), (1.10) e (1.11), e dado que

$$f(t,y(t))=\phi_f(t,y(t);0)\ \ \forall t\in[t_0,t_0+T]$$

podemos afirmar que

$$\lim_{h\to0^+}\max_{0\le n\le N-1}\|\tau_{n+1}(h;y)\|=0, \tag{1.12}$$

é dicir, chegamos a que o método é consistente.

Necesidade: Supoñamos agora que o método é consistente.

Consideremos $\bar{t} \in [t_0, t_0 + T]$ e $\bar{y} \in \mathbb{R}^m$ fixos. Logo existe unha única solución $y = y(t)$ para o problema de Cauchy:

$$\begin{cases} y'(t) = f(t, y), \ t \in [t_0, t_0 + T], \\ y(\bar{t}) = \bar{y}, \end{cases}$$

con $y \in C^1([t_0, t_0 + T])$. Esta función é tamén a única solución do problema de Cauchy (1.1) para a condición inicial $\eta_0 = y(t_0)$. Consideramos a resolución numérica deste último problema de Cauchy tomando o paso constante $h > 0$ da forma $h = \frac{\bar{t} - t_0}{n}$, sendo $n \geq 1$ enteiro; entón verifícase (1.9) para $y = y(t)$. Ademais, como o método é consistente, verifícase (1.12).

Así, facendo o límite cando $h \to 0^+$ en (1.9) (nótese que $t_n = \bar{t}$ está fixo), por (1.10), (1.11) e (1.12) deducimos que

$$\phi_f(\bar{t}, \bar{y}; 0) = f(\bar{t}, \bar{y}).$$

Como $(\bar{t}, \bar{y})$ é un punto arbitrario de $[t_0, t_0 + T] \times \mathbb{R}^m$, concluímos que se verifica (1.8), como queríamos demostrar. □

Definición 1.4. Sexa p un enteiro, $p \geq 1$. Dise que o esquema (1.3) é *consistente de polo menos orde p* se para toda función f suficientemente regular e para toda solución $y = y(t)$ da EDO $y'(t) = f(t, y)$ se ten que

$$\max_{1 \leq n \leq N} \|\tau_n(h; y)\| = O(h^p)$$

Dise que o esquema (1.3) é *consistente de orde p*, se é consistente de polo menos orde p e non o é de orde $p + 1$.

1.2.2. Estabilidade

Definición 1.5. Dise que o método (1.3) é *estable* se existe $h_0 > 0$ e unha constante $M > 0$ independente de h, tal que para todo $0 < h \leq h_0$ e para todas as sucesións y_n, z_n e ε_n, $n = 0, 1, \ldots, N - 1$, que verifican

$$\begin{aligned} y_{n+1} &= y_n + h\phi_f(t_n, y_n; h) \\ z_{n+1} &= z_n + h\phi_f(t_n, z_n; h) + \varepsilon_n \end{aligned} \tag{1.13}$$

temos

$$\max_{0 \leq n \leq N} \|z_n - y_n\| \leq M[\|y_0 - z_0\| + \sum_{n=0}^{N-1} \|\varepsilon_n\|].$$

Lema 1.6. *Sexa $\{\theta_n\}_{n=0}^N$ un conxunto finito de números reais que verifican a desigualdade*

$$\theta_{n+1} \leq \gamma\theta_n + \alpha_{n+1}, \ n = 0, \ldots, N - 1 \tag{1.14}$$

onde $\gamma > 0$ é unha constante independente de n e os $\alpha_{n+1} \in \mathbb{R}$, $n = 0, 1, \ldots, N-1$.
Entón,

$$\theta_n \leq \gamma^n \theta_0 + \sum_{i=1}^{n} \gamma^{n-i} \alpha_i, \ n = 1, \ldots, N. \tag{1.15}$$

Demostración. Imos probar este resultado por indución en n.

- Para $n = 1$, (1.15) redúcese a

  $$\theta_1 \leq \gamma \theta_0 + \alpha_1$$

  que resulta de tomar $n = 0$ en (1.14).

- Imos supoñelo certo para $n = k$ e demostrarémolo para $n = k + 1$. Temos que probar que

  $$\theta_{k+1} \leq \gamma^{k+1} \theta_0 + \sum_{i=1}^{k+1} \gamma^{k+1-i} \alpha_i.$$

  Sabemos que $\theta_k \leq \gamma^k \theta_0 + \sum_{i=1}^{k} \gamma^{k-i} \alpha_i$ e dado que $\gamma > 0$ temos que

  $$\gamma \theta_k \leq \gamma(\gamma^k \theta_0 + \sum_{i=1}^{k} \gamma^{k-i} \alpha_i).$$

  Así,

  $$\theta_{k+1} \leq \gamma \theta_k + \alpha_{k+1} \leq \gamma(\gamma^k \theta_0 + \sum_{i=1}^{k} \gamma^{k-i} \alpha_i) + \alpha_{k+1} =$$

  $$= \gamma^{k+1} \theta_0 + \sum_{i=1}^{k} \gamma^{k+1-i} \alpha_i + \alpha_{k+1} = \gamma^{k+1} \theta_0 + \sum_{i=1}^{k+1} \gamma^{k+1-i} \alpha_i.$$

□

Lema 1.7. *Unha condición suficiente para que o método (1.3) sexa estable é que exista unha constante Λ tal que*

$$\forall t \in [t_0, t_0 + T], \ \forall y, \ z \in \mathbb{R}^m, \ \forall h \in [0, h^*], \|\phi_f(t, y; h) - \phi_f(t, z; h)\| \leq \Lambda \|y - z\|; \tag{1.16}$$

entón $M = e^{\Lambda T}$.

Demostración. De (1.3), (1.13) e (1.16) deducimos que

$$\|y_{n+1} - z_{n+1}\| = \|y_n - z_n + h(\phi_f(t_n, y_n; h) - \phi_f(t_n, z_n; h)) - \varepsilon_n\| \leq$$

$$\leq \|y_n - z_n\| + h\|\phi_f(t_n, y_n; h) - \phi_f(t_n, z_n; h)\| + \|\varepsilon_n\| \leq (1 + h\Lambda)\|y_n - z_n\| + \|\varepsilon_n\|.$$

Se consideramos $\theta_n = \|y_n - z_n\|$, $\gamma = 1 + h\Lambda$ e $\alpha_{n+1} = \|\varepsilon_n\|$ e facemos uso do lema 1.6 tense que:

$$\|y_n - z_n\| \leq (1+h\Lambda)^n \|y_0 - z_0\| + \sum_{i=1}^{n} (1+h\lambda)^{n-i} \|\varepsilon_{i-1}\| \leq (1+h\Lambda)^n [\|y_0 - z_0\| + \sum_{i=1}^{n} \|\varepsilon_{i-1}\|] \leq$$

$$(e^{h\Lambda})^n [\|y_0 - z_0\| + \sum_{i=1}^{n-1} \|\varepsilon_i\|] = e^{h\Lambda n} [\|y_0 - z_0\| + \sum_{i=1}^{n-1} \|\varepsilon_i\|] \leq e^{h\Lambda N} [\|y_0 - z_0\| + \sum_{i=0}^{n-1} \|\varepsilon_i\|] \leq$$

$$e^{\Lambda T} [\|y_0 - z_0\| + \sum_{i=1}^{n-1} \|\varepsilon_i\|],$$

onde a terceira desigualdade se debe a que $1 + x \leq e^x$ para todo $x \in \mathbb{R}$ e a última desigualdade se verifica xa que $N = \lceil \frac{T}{h} \rceil$.
Así, se tomamos $M = e^{\Lambda T}$

$$\|y_n - z_n\| \leq M[\|y_0 - z_0\| + \sum_{i=0}^{n-1} \|\varepsilon_i\|] \ \forall \ n = 0, 1, \ldots, N,$$

e por tanto

$$\max_{0 \leq n \leq N} \|y_n - z_n\| \leq M[\|y_0 - z_0\| + \sum_{i=0}^{N-1} \|\varepsilon_i\|],$$

o que quere dicir que o método é estable. □

1.2.3. Converxencia

Definición 1.8. Diremos que o método (1.3) é *converxente* se baixo a condición

$$\lim_{h \to 0^+} \eta_h = \eta$$

se ten

$$\lim_{h \to 0^+} \max_{0 \leq n \leq N} \|y(t_n) - y_n\| = 0$$

onde $y(.)$ denota a solución de (1.1) e y_n a solución de (1.3), (1.4).

Teorema 1.9. *Se o método dun paso (1.3) é estable e consistente, entón é converxente.*

Demostración. Da definición de τ_{n+1} dada por (1.5), deducimos que

$$y(t_{n+1}) = y(t_n) + h\phi_f(t_n, y(t_n); h) + h\tau_{n+1}, \ n = 0, \ldots, N-1.$$

Esta ecuación é un caso particular de (1.13), con $z_n = y(t_n)$ e $\varepsilon_n = h\tau_{n+1}$.
Como o método é estable, para $h \in (0, h_0]$ temos que

$$0 \leq \max_{0 \leq n \leq N} \|y(t_n) - y_n\| \leq M[\|y_0 - y(t_0)\| + \sum_{n=0}^{N-1} \|\varepsilon_n\|] = M[\|\eta_h - \eta\| + h \sum_{n=0}^{N-1} \|\tau_{n+1}\|] \leq$$

$$\leq M[\|\eta_h - \eta\| + Nh \max_{1 \leq n \leq N} \|\tau_n\|] \leq M[\|\eta_h - \eta\| + T \max_{1 \leq n \leq N} \|\tau_n\|]. \quad (1.17)$$

Ademais, por ser o método (1.3) consistente verifícase que

$$\lim_{h\to 0^+}(\max_{1\leq n\leq N}\|\tau_n(h;y)\|)=0.$$

Polo tanto, dado que $\eta_h \to \eta$ podemos concluír que

$$\lim_{h\to 0^+}(\max_{0\leq n\leq N}\|y(t_n)-y_n\|)=0.$$

□

Teorema 1.10. *Supoñamos que a función ϕ_f verifica (1.8) e a condición de Lipschitz (1.16); entón o método (1.3) é converxente.*

Demostración. Por hipótese ϕ_f satisfai (1.8) polo que, facendo uso do lema 1.3 temos que o método é consistente. Ademais, por verificarse a condición de Lipschitz (1.16), o lema 1.7 garantiza que o método é estable.
Dado que o método é estable e consistente, empregando o teorema 1.9, concluímos que o método (1.3) é converxente. □

Teorema 1.11. *Se o método dun paso é estable e consistente de polo menos orde p e se f é suficientemente regular, temos que*

$$\max_{0\leq n\leq N}\|y(t_n)-y_n\|\leq M[\|\eta_h-\eta\|+O(h^p)]. \tag{1.18}$$

Demostración. Por (1.11) temos que

$$\max_{0\leq n\leq N}\|y(t_n)-y_n\|\leq M[\|\eta_h-\eta\|+T\max_{1\leq n\leq N}\|\tau_n\|].$$

Ademais, como método e de polo menos orde p, verifícase

$$\max_{0\leq n\leq N}\|\tau_n(h;y)\|=O(h^p),$$

o que xunto coa desigualdade anterior implica (1.18). □

1.3. Estudo da orde

Introduzamos as notacións

$$\begin{cases} f^{(0)}(t,y)=f(t,y) \\ f^{(1)}(t,y)=D_tf(t,y)+D_yf(t,y)f(t,y) \\ \dots \\ f^{(k)}(t,y)=D_tf^{k-1}(t,y)+D_yf^{k-1}(t,y)f(t,y) \end{cases} \tag{1.19}$$

Demóstrase por indución que se y é solución de $y'(t) = f(t, y(t))$, tense que

$$\forall\ k \geq 0,\ y^{(k+1)}(t) = f^{(k)}(t, y(t)) = \frac{d^k}{dt^k} f(t, y(t)) \tag{1.20}$$

Teorema 1.12. *Supoñamos que f é unha función p veces continuamente diferenciable en $[t_0, t_0+T]\times\mathbb{R}^m$ e que as funcións ϕ, $\frac{\partial\phi}{\partial h}$, ..., $\frac{\partial^p\phi}{\partial h^p}$ existen e son continuas en $[t_0, t_0+T]\times\mathbb{R}^m\times[0, h^*]$. Entón, unha condición necesaria e suficiente para que o método sexa polo menos de orde p escríbese da seguinte forma: para todo $(t, y) \in [t_0, t_0 + T] \times \mathbb{R}^m$*

$$\begin{cases} \phi(t, y; 0) = f(t, y) \\ \frac{\partial\phi}{\partial h}(t, y; 0) = \frac{1}{2} f^{(1)}(t, y) \\ \ldots \\ \frac{\partial^{p-1}\phi}{\partial h^{p-1}}(t, y; 0) = \frac{1}{p} f^{(p-1)}(t, y) \end{cases} \tag{1.21}$$

Demostración. Consideremos

$$T_{n+1} = h\tau_{n+1} = y(t_{n+1}) - y(t_n) - h\phi_f(t_n, y(t_n); h)$$

e sexa $f \in C^p([t_0, t_0 + T] \times \mathbb{R}^m)$. Isto implica que $y \in C^{p+1}([t_0, t_0 + T])$ e entón podemos escribir

$$y(t_{n+1}) = y(t_n+h) = y(t_n)+hy'(t_n)+\frac{h^2}{2!}y''(t_n)+\ldots+\frac{h^p}{p!}y^{(p)}(t_n)+\frac{1}{p!}\int_{t_n}^{t_{n+1}}(t_{n+1}-s)^p y^{(p+1)}(s)ds,$$

é dicir,

$$y(t_{n+1}) - y(t_n) = \sum_{k=0}^{p-1} \frac{h^{k+1}}{(k+1)!} y^{(k+1)}(t_n) + R_{n,1}(h)$$

onde

$$R_{n,1}(h) = \frac{1}{p!}\int_{t_n}^{t_{n+1}} (t_{n+1} - s)^p y^{(p+1)}(s)ds.$$

Ademais, tense que

$$\phi_f(t_n, y(t_n); h) = \sum_{k=0}^{p-1} \frac{h^k}{k!} \frac{\partial^k\phi}{\partial h^k}(t_n, y(t_n); 0) + R_{n,2}(h)$$

onde

$$R_{n,2}(h) = \frac{1}{(p-1)!}\int_0^h (h - s)^{p-1} \frac{\partial^p\phi}{\partial h^p}(t_n, y(t_n); s)ds.$$

En consecuencia, podemos escribir

$$T_{n+1} = \sum_{k=0}^{p-1} \frac{h^{k+1}}{(k+1)!} y^{(k+1)}(t_n) + R_{n,1}(h) - h\left(\sum_{k=0}^{p-1} \frac{h^k}{k!} \frac{\partial^k\phi}{\partial h^k}(t_n, y(t_n); 0) + R_{n,2}(h)\right) =$$

$$= \sum_{k=0}^{p-1} h^{k+1}\left(\frac{1}{(k+1)!} f^{(k)}(t_n, y(t_n)) - \frac{1}{k!}\frac{\partial^k\phi}{\partial h^k}(t_n, y(t_n); 0)\right) + R_{n,1}(h) - hR_{n,2}(h).$$

onde a última igualdade se debe a (1.20).

Temos que

$$\|R_{n,1}(h)\| \leq \frac{1}{p!}\int_{t_n}^{t_{n+1}} (t_{n+1}-s)^p \|y^{(p+1)}(s)\| ds \leq c_1 h^{p+1},$$

onde $c_1 = \frac{1}{(p+1)!}\,\text{máx}_{s\in[t_0,t_0+T]}\|y^{(p+1)}(s)\|$.

$$\|R_{n,2}(h)\| \leq \frac{1}{(p-1)!}\int_0^h (h-s)^{p-1}\|\frac{\partial^p \phi}{\partial h^p}(t_n, y(t_n); s)\| ds \leq c_2 h^p$$

onde $c_2 = \frac{1}{p!}\,\text{máx}\limits_{\substack{0\leq s\leq h^* \\ t_0\leq t\leq t_0+T}} \|\frac{\partial^p \phi}{\partial h^p}(t, y(t); s)\|$.

Así pois

$$T_{n+1} = \sum_{k=0}^{p-1} h^{k+1}\Psi_k(t_n, y(t_n)) + O(h^{p+1})$$

onde $\Psi_k(t,y) = \frac{1}{(k+1)!}f^k(t,y) - \frac{1}{k!}\frac{\partial^k \phi}{\partial h^k}(t,y;0)$ e o término $O(h^p)$ é uniforme respecto de n.

Suficiencia

As condicións (1.19) equivalen a que para todo $k < p$, $\Psi_k(t,y) \equiv 0$. En consecuencia, se se cumpren as condicións (1.19), entón $T_{n+1} = O(h^{p+1})$ uniformemente en n; polo tanto

$$\text{máx}_{0\leq n\leq N-1}\|\tau_{n+1}\| = O(h^p),$$

o que quere dicir que o método é polo menos de orde p.

Necesidade

Supoñamos que non se cumpren as condicións (1.19). Entón existe un enteiro o máis pequeno posible $\bar{k} < p$ ($\bar{k} \geq 0$) tal que $\Psi_{\bar{k}}(t,y) \not\equiv 0$.

Temos que

$$T_{n+1} = h^{\bar{k}+1}\Psi_{\bar{k}}(t_n, y(t_n)) + O(h^{\bar{k}+2}) + O(h^{p+1}) = h^{\bar{k}+1}\Psi_{\bar{k}}(t_n, y(t_n)) + O(h^{\bar{k}+2}),$$

xa que $\bar{k} \leq p-1$. Por outra parte, como o método é de polo menos orde p, $T_{n+1} = O(h^{p+1})$, o que implica que $T_{n+1} = O(h^{\bar{k}+2})$. Por tanto,

$$h^{\bar{k}+1}\Psi_{\bar{k}}(t_n, y(t_n)) + O(h^{\bar{k}+2}) = O(h^{\bar{k}+2}),$$

do que deducimos que

$$\Psi_{\bar{k}}(t_n, y(t_n)) = O(h).$$

Como $\Psi_{\bar{k}}(t,y) \not\equiv 0$, existe $(\bar{t}, \bar{y}) \in [t_0, t_0+T]\times\mathbb{R}^m$ tal que $\Psi_k(\bar{t},\bar{y}) \neq 0$.

Se consideramos a solución $y = y(t)$ do problema de valor inicial

$$\begin{cases} y' = f(t,y),\ t\in[t_0, t_0+T] \\ y(\bar{t}) = \bar{y} \end{cases}$$

esta función é tamén a única solución do problema de Cauchy (1.1) para a condición inicial $\eta_0 = y(t_0)$. Consideramos a resolución numérica deste último problema de Cauchy tomando o paso constante $h > 0$ con $h = \frac{\bar{t}-t_0}{n}$; dado que $t_n = t_0 + nh = \bar{t}$ permanece fixo, tense que

$$\Psi_{\bar{k}}(\bar{t}, \bar{y}) = \Psi_{\bar{k}}(\bar{t}, y(\bar{t})) = O(h);$$

por tanto, $\Psi_{\bar{k}}(\bar{t}, \bar{y}) = 0$, co cal obtivemos unha contradición. Por conseguinte, $\Psi_k(t, y) \equiv 0$ para todo $k < p$. □

Capítulo 2

Métodos Runge-Kutta

Este capítulo está dedicado ao estudo dunha familia de métodos dun paso, denominados métodos Runge-Kutta. A dificultade principal da análise destos métodos, podemos encontrala ao considerar métodos que poden ser implícitos ou semiimplícitos xa que a función incremento nestes casos non está definida de forma explícita, o que complica o seu estudo.

Cabe destacar que a sección 2.5, que aborda a orde dos métodos Runge-Kutta explícitos no caso escalar, foi introducida xa que neste caso basta usar técnicas máis sinxelas que as empregadas na sección 2.8, na cal se aborda a teoría xeral que ten maior dificultade, entre outras cousas porque se inclúen os casos implícitos.

As referencias seguidas neste capítulo son [2], [3] e [4].

Neste capítulo supoñeremos que f satisfai a seguinte hipótese, lixeiramente máis forte que (1.2): existe $\delta_1 \in \mathbb{R}$, $\delta_1 > 0$ tal que f é continua de $[t_0 - \delta_1, t_0 + T + \delta_1] \times \mathbb{R}^m$ en $\mathbb{R}^m$ e verifica a condición de Lipschitz

$$\exists L > 0 \text{ tal que } \|f(t,y) - f(t,z)\| \leq L\|y - z\|,\ \forall y, z \in \mathbb{R}^m,\ \forall t \in [t_0 - \delta_1, t_0 + T + \delta_1] \quad (2.1)$$

Na práctica isto non é restritivo.

2.1. Introdución

Definimos un método Runge-Kutta xeral de q etapas para o problema (1.1) como:

$$\begin{cases} y_{n+1} = y_n + h\sum_{i=1}^{q} b_i f(t_{n,i}, y_{n,i}) & \text{(2.2a)} \\ \text{onde} & \\ y_{n,i} = y_n + h\sum_{j=1}^{q} a_{ij} f(t_{n,j}, y_{n,j}),\ i = 1, \dots, q & \text{(2.2b)} \\ t_{n,i} = t_n + hc_i,\ i = 1, \dots, q & \end{cases}$$

Unha forma, alternativa, para escribir o mesmo método Runge-Kutta é:

$$\begin{cases} y_{n+1} = y_n + h\sum_{i=1}^{q} b_i k_{n,i} & \text{(2.3a)} \\ \textit{onde} & \\ k_{n,i} = f\left(t_{n,i}, y_n + h\sum_{j=1}^{q} a_{ij} k_{n,j}\right),\ i = 1, \dots, q & \text{(2.3b)} \\ t_{n,i} = t_n + hc_i,\ i = 1, \dots, q & \end{cases}$$

Lema 2.1. *As formulacións (2.2) e (2.3) son equivalentes.*

Demostración. Vexamos en primeiro lugar que (2.2) implica (2.3). Supoñamos que os $y_{n,i}$, $i = 1, \dots, q$, cumpren (2.2b). Sexa

$$k_{n,i} = f(t_{n,i}, y_{n,i}),\ i = 1, \dots, q. \tag{2.4}$$

Logo, podemos escribir:

$$y_{n,i} = y_n + h\sum_{j=1}^{q} a_{ij} k_{n,j},\ i = 1, \dots, q. \tag{2.5}$$

Substituíndo $y_{n,i}$ en (2.4), obtemos (2.3b). Por outra parte, (2.4) e (2.2a) implican (2.3a).
Agora temos que demostrar que (2.3) implica (2.2). Supoñamos que os $k_{n,i}$, $i = 1, \dots, q$, cumpren (2.3b) e definamos $y_{n,i}$ mediante (2.5). Isto, xunto con (2.3b) implica (2.4). Usando (2.4) para substituír en (2.5) os $k_{n,j}$, obtemos (2.2b). Por último, de (2.3a) e (2.4) dedúcese (2.2a). □

Imos representar os coeficientes de (2.2) e (2.3) mediante a denominada *táboa de Butcher*:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1q} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_q & a_{q1} & a_{q2} & \cdots & a_{qq} \\ \hline & b_1 & b_2 & \cdots & b_q \end{array}$$

que denotaremos de forma simplificada como

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$$

onde definimos os vectores q-dimensionais c e b e a matriz A de dimensión $q \times q$ mediante:

$$c = [c_1, c_2, \dots c_q]^T\,,\ b = [b_1, b_2, \dots b_q]^T\,,\ A = [a_{ij}]\,.$$

É conveniente introducir tamén $e = (1, \dots, 1)^T \in \mathbb{R}^q$ e a matriz diagonal C de dimensión $q \times q$ dada por $c_{ii} = c_i$, $i = 1, \dots, q$ (é dicir, $C = diag(c_1, \dots, c_q)$).

Cabe destacar que moitos métodos Runge-Kutta verifican a *condición de filas*

$$c_i = \sum_{j=1}^{q} a_{ij},\ i = 1, \dots, q,$$

que podemos escribir como $c = Ae$.

2.2. Métodos Runge-Kutta explícitos, implícitos e semiimplícitos

Dise que un método Runge-Kutta (2.2) (ou (2.3)) é *explícito* se verifica que $a_{ij} = 0$ para todo i, $j = 1, \dots, q$ tales que $i \leq j$, é dicir, se a matriz A é unha matriz estritamente triangular inferior. A táboa de Butcher dun método Runge-Kutta explícito é da seguite forma:

$$\begin{array}{c|ccccc} c_1 & 0 & 0 & \cdots & 0 & 0 \\ c_2 & a_{21} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_q & a_{q1} & a_{q2} & \cdots & a_{qq-1} & 0 \\ \hline & b_1 & b_2 & \cdots & b_{q-1} & b_q \end{array}$$

Se un método Runge-Kutta é explícito, os vectores $k_{n,i}$ ou, de forma alternativa, $y_{n,i}$, $i = 1, \dots, q$, obtéñense de forma recursiva e explícita, é dicir, sen necesidade de resolver ningún sistema de ecuacións.

En efecto, para a formulación con $k_{n,i}$ tense:

$$k_{n,1} = f(t_{n,1}, y_n),$$

$$k_{n,2} = f(t_{n,2}, y_n + ha_{21}k_{n,1})$$

e en xeral

$$k_{n,i} = f\left(t_{n,i}, y_n + h\sum_{j=1}^{i-1} a_{ij}k_{n,j}\right),\ i = 2, ..., q.$$

Para a formulación en $y_{n,i}$ tense:

$$y_{n,1} = y_n, \tag{2.6}$$

$$y_{n,2} = y_n + ha_{21}f(t_{n,1}, y_{n,1})$$

e en xeral

$$y_{n,i} = y_n + h\sum_{j=1}^{i-1} a_{ij}f(t_{n,j}, y_{n,j}),\ i = 2, \dots, q. \tag{2.7}$$

Exemplo 2.2 (Método de Euler Explícito)**.** Este método ten táboa de Butcher

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

A súa formulación en $y_{n,i}$ é:

$$y_{n+1} = y_n + hf(t_{n,1}, y_{n,1})$$

onde

$y_{n,1} = y_n$

$t_{n,1} = t_n$

A súa formulación en $k_{n,i}$ é:

$$y_{n+1} = y_n + hk_{n,1}$$

onde

$k_{n,1} = f(x_{n,1}, y_n)$

e $t_{n,1}$ se corresponde co da formulación en $y_{n,i}$.

Exemplo 2.3 (Método Runge-Kutta uniparamétrico de dúas etapas)**.** Este método ten táboa de Butcher

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1 - \frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}$$

onde α é un parámetro distinto de cero.

A súa formulación en $y_{n,i}$ é:

$$y_{n+1} = y_n + h(1 - \frac{1}{2\alpha})f(t_{n,1}, y_{n,1}) + h\frac{1}{2\alpha}f(t_{n,2}, y_{n,2})$$

onde

$y_{n,1} = y_n$

$y_{n,2} = y_n + h\alpha f(t_{n,1}, y_{n,1})$

$t_{n,1} = t_n$

$t_{n,2} = t_n + h\alpha$

A formulación en $k_{n,i}$ é:

$$y_{n+1} = y_n + h(1 - \frac{1}{2\alpha})k_{n,1} + h\frac{1}{2\alpha}k_{n,2}$$

onde

$k_{n,1} = f(t_{n,1}, y_n)$

$k_{n,2} = f(t_{n,2}, y_n + h\alpha k_{n,1})$

e $t_{n,1}$ e $t_{n,2}$ se corresponden cos da formulación en $y_{n,i}$.

Exemplo 2.4 (Método Runge-Kutta clásico de orde 4)**.** Este método ten táboa de Butcher

0	0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
1	0	0	1	0
	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

A formulación en $y_{n,i}$ é:

$$y_{n+1} = y_n + \frac{1}{6}hf(t_{n,1}, y_{n,1}) + \frac{2}{6}hf(t_{n,2}, y_{n,2}) + \frac{2}{6}hf(t_{n,3}, y_{n,3}) + \frac{1}{6}hf(t_{n,4}, y_{n,4})$$

onde

$y_{n,1} = y_n$

$y_{n,2} = y_n + \frac{1}{2}hf(t_{n,1}, y_{n,1})$

$y_{n,3} = y_n + \frac{1}{2}hf(t_{n,2}, y_{n,2})$

$y_{n,4} = y_n + hf(t_{n,3}, y_{n,3})$

$t_{n,1} = t_n$

$t_{n,2} = t_n + \frac{1}{2}h$

$t_{n,3} = t_n + \frac{1}{2}h$

$t_{n,4} = t_n + h$

A formulación en $k_{n,i}$ é:

$$y_{n+1} = y_n + \frac{1}{6}hk_{n,1} + \frac{2}{6}hk_{n,2} + \frac{2}{6}hk_{n,3} + \frac{1}{6}hk_{n,4}$$

onde

$k_{n,1} = f(t_{n,1}, y_n)$

$k_{n,2} = f(t_{n,2}, y_n + \frac{1}{2}hk_{n,1})$

$k_{n,3} = f(t_{n,3}, y_n + \frac{1}{2}hk_{n,2})$
$k_{n,4} = f(t_{n,4}, y_n + hk_{n,3})$
e $t_{n,1}$, $t_{n,2}$, $t_{n,3}$ e $t_{n,4}$ se corresponden cos da formulación en $y_{n,i}$.

O método Runge-Kutta (2.2) (ou (2.3)) dise que é *semiimplícito* se verifica que $a_{ij} = 0$ para todo i, $j = 1, \dots, q$ tales que $i < j$, ou equivalentemente, cando a matriz A é triangular inferior. A táboa de Butcher para este tipo de métodos é da forma:

$$\begin{array}{c|cccc} c_1 & a_{11} & 0 & \cdots & 0 \\ c_2 & a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_q & a_{q1} & a_{q2} & \cdots & a_{qq} \\ \hline & b_1 & b_2 & \cdots & b_q \end{array}$$

Se un método Runge-Kutta é semiimplícito, o cálculo dos vectores $k_{n,i}$ ou, de forma alternativa, $y_{n,i}$, $i = 1, \dots, q$, pode facerse de forma recursiva pero, neste caso, hai que resolver q sistemas de m ecuacións con m incógnitas (unha por cada vector $k_{n,i}$ ou $y_{n,i}$).

Por exemplo, para a formulación en $k_{n,i}$ tense:

$$k_{n,1} = f(t_{n,1}, y_n + ha_{11}k_{n,1})$$

$$k_{n,2} = f(t_{n,2}, y_n + ha_{21}k_{n,1} + ha_{22}k_{n,2})$$

onde $k_{n,1}$ é coñecido. En xeral, para $i = 2, \dots, q$

$$k_{n,i} = f(t_{n,i}, y_n + h\sum_{j=1}^{i-1} a_{ij}k_{n,j} + ha_{ii}k_{n,i}),$$

onde os $k_{n,j}$, $1 \leq j \leq i-1$, xa se coñecen.

Isto é máis ventaxoso dende o punto de vista computacional que a resolución dun sistema $(mq) \times (mq)$, que é o caso dos Runge-Kutta implícitos.

Exemplo 2.5 (Método Runge-Kutta uniparamétrico dunha etapa)**.** Este método ten táboa de Butcher

$$\begin{array}{c|c} \theta & \theta \\ \hline & 1 \end{array}$$

con $\theta \in \mathbb{R}$.
A formulación en $y_{n,i}$ é:

$$y_{n+1} = y_n + hf(t_{n,1}, y_{n,1}) \tag{2.8}$$

onde
$y_{n,1} = y_n + \theta h f(t_{n,1}, y_{n,1})$
$t_{n,1} = t_n + \theta h$
A formulación en $k_{n,i}$ é:

$$y_{n+1} = y_n + hk_{n,1}$$

onde
$k_{n,1} = f(t_{n,1}, y_n + \theta h k_{n,1})$
e $t_{n,1}$ se corresponde co da formulación en $y_{n,i}$.
Por (2.8) podemos escribir

$$y_{n,1} = y_n + \theta(y_{n+1} - y_n)$$

e, por tanto, o método pode reescribirse da forma

$$y_{n+1} = y_n + hf(t_n + \theta h, \theta y_{n+1} + (1-\theta)y_n).$$

Este método é semiimplícito se $\theta \neq 0$. Para $\theta = 0$ obtense o método de Euler explícito e para $\theta = 1$ o de Euler implícito.

Exemplo 2.6 (Método Runge-Kutta diagonalmente implícito (DIRK) de dúas etapas)**.** Este método ten táboa de Butcher

$$\begin{array}{c|cc} \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{3}{4} & \frac{1}{2} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

A formulación en $y_{n,i}$ é:

$$y_{n+1} = y_n + \frac{1}{2}hf(t_{n,1}, y_{n,1}) + \frac{1}{2}hf(t_{n,2}, y_{n,2})$$

onde
$y_{n,1} = y_n + \frac{1}{4}hf(t_{n,1}, y_{n,1})$
$y_{n,2} = y_n + \frac{1}{2}hf(t_{n,1}, y_{n,1}) + \frac{1}{4}hf(t_{n,2}, y_{n,2})$
$t_{n,1} = t_n + \frac{1}{4}h$
$t_{n,2} = t_n + \frac{3}{4}h$
A formulación en $k_{n,i}$ é:

$$y_{n+1} = y_n + \frac{1}{2}hk_{n,1} + \frac{1}{2}hk_{n,2}$$

onde
$k_{n,1} = f(t_{n,1}, y_n + \frac{1}{4}hk_{n,1})$
$k_{n,2} = f(t_{n,2}, y_n + \frac{1}{2}hk_{n,1} + \frac{1}{4}hk_{n,2})$
e $t_{n,1}$ e $t_{n,2}$ se corresponden cos da formulación en $y_{n,i}$.

Por último, diremos que o método Runge-Kutta (2.2) (ou (2.3)) é *implícito* se verifica que $a_{ij} \neq 0$ para algún $j > i$, é dicir, a matriz A non é triangular inferior.

Neste caso a obtención dos $k_{n,i}$ ou, de forma alternativa, $y_{n,i}$, $i = 1, \dots, q$, require a resolución dun sistema de qm ecuacións con qm incógnitas.

Exemplo 2.7 (Método Runge-Kutta Gauss-Legendre)**.** Este método ten táboa de Butcher

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

A formulación en $y_{n,i}$ é:

$$y_{n+1} = y_n + \frac{1}{2} h f(t_{n,1}, y_{n,1}) + \frac{1}{2} h f(t_{n,2}, y_{n,2})$$

onde
$y_{n,1} = y_n + \frac{1}{4} h f(t_{n,1}, y_{n,1}) + \frac{3-2\sqrt{3}}{12} h f(t_{n,2}, y_{n,2})$
$y_{n,2} = y_n + \frac{3+2\sqrt{3}}{12} h f(t_{n,1}, y_{n,1}) + \frac{1}{4} h f(t_{n,2}, y_{n,2})$
$t_{n,1} = t_n + \frac{3-\sqrt{3}}{6} h$
$t_{n,2} = t_n + \frac{3+\sqrt{3}}{6} h$
A formulación en $k_{n,i}$ é:

$$y_{n+1} = y_n + \frac{1}{2} h k_{n,1} + \frac{1}{2} h k_{n,2}$$

onde
$k_{n,1} = f(t_{n,1}, y_n + \frac{1}{4} h k_{n,1} + \frac{3-2\sqrt{3}}{12} h k_{n,2})$
$k_{n,2} = f(t_{n,2}, y_n + \frac{3+2\sqrt{3}}{12} h k_{n,1} + \frac{1}{4} h k_{n,2})$
e $t_{n,1}$ e $t_{n,2}$ se corresponden cos da formulación en $y_{n,i}$.

2.3. Inclusión dos métodos Runge-Kutta no marco xeral dos métodos dun paso

2.3.1. Métodos Runge-Kutta explícitos

Comezamos considerando un método Runge-Kutta explícito de dúas etapas, que terá como táboa de Butcher:

$$\begin{array}{c|cc} c_1 & 0 & 0 \\ c_2 & a_{21} & 0 \\ \hline & b_1 & b_2 \end{array}$$

Entón, temos que:

$k_{n,1} = f(t_n + c_1 h, y_n)$

$k_{n,2} = f(t_n + c_2 h, y_n + h a_{21} k_{n,1})$

$y_{n+1} = y_n + h(b_1 k_{n,1} + b_2 k_{n,2})$

Imos tratar de escribir o método da forma

$$y_{n+1} = y_n + h\phi_f(t_n, y_n; h).$$

Sabemos que

$$y_{n+1} = y_n + h\left[b_1 f(t_n + c_1 h, y_n) + b_2 f(t_n + c_2 h, y_n + h a_{21} f(t_n + c_1 h, y_n))\right]$$

necesitamos pois asegurar que

$$\phi_f(t_n, y_n; h) = b_1 f(t_n + c_1 h, y_n) + b_2 f(t_n + c_2 h, y_n + h a_{21} f(t_n + c_1 h, y_n)).$$

Polo tanto, consideramos

$$\phi_f(t, y; h) = b_1 f(t + c_1 h, y) + b_2 f(t + c_2 h, y + h a_{21} f(t + c_1 h, y)).$$

Proposición 2.8. *Se un método Runge-Kutta é explícito, a súa función incremento ϕ_f é explícita.*

Demostración. Consideremos un método Runge-Kutta explícito de q etapas; a súa formulación en $y_{n,i}$ está dada por (2.6), (2.7) e (2.2a).
O noso obxectivo é definir unha función ϕ_f explícita de tal xeito que

$$y_{n+1} = y_n + h\phi_f(t_n, y_n; h).$$

Para definir $\phi_f : (t, y; h) \in [t_0, t_0 + T] \times \mathbb{R}^m \times [0, h^*] \to \phi_f(t, y; h) \in \mathbb{R}^m$, onde $h^* > 0$ é tal que $h^* \max_{1 \le j \le q} |c_j| \le \delta_1$ [1], procedemos como sigue.
En primeiro lugar, introducimos os vectores $Y_1, Y_2, \dots, Y_q \in \mathbb{R}^m$ dados por
$Y_1 = y$
$Y_2 = y + h a_{21} f(t + c_1 h, Y_1)$
$\vdots$
$Y_i = y + h \sum_{j=1}^{i-1} a_{ij} f(t + c_j h, Y_j),\ i = 2, \dots, q$
A continuación, definimos ϕ_f mediante

$$\phi_f(t, y; h) := \sum_{j=1}^{q} b_j f(t + c_j h, Y_j).$$

[1] Esta restrición garante que $\forall h \in [0, h^*]$, $\forall j = 1, \dots, q$, $t + c_j h \in [t_0 - \delta_1, t_0 + T + \delta_1]$, polo que $f(t + c_j h, Y_j)$ está ben definido.

Nótese que se $t = t_n$ e $y = y_n$, tense que

$$Y_i = y_{n,i},\ i = 1, \dots, q$$

e por tanto

$$\phi_f(t_n, y_n; h) := \sum_{j=1}^{q} b_j f(t_n + c_j h, y_{n,j}).$$

□

2.3.2. Métodos Runge-Kutta implícitos

[2] En xeral, a formulación en $y_{n,i}$ dun método Runge-Kutta implícito está dada por (2.2). O noso obxectivo e escribir o método na forma xeral dos métodos dun paso,

$$y_{n+1} = y_n + h\phi_f(t_n, y_n; h) \tag{2.9}$$

definindo para eso unha función incremento $\phi_f : [t_0, t_0 + T] \times \mathbb{R}^m \times [0, h^*] \to \mathbb{R}^m$ apropiada, sendo $h^* > 0$ suficientemente pequeno.

Dados $(t, y, h) \in [t_0, t_0 + T] \times \mathbb{R}^m \times [0, h^*]$, introducimos (por analoxía co sistema (2.2b)) os vectores $Y_i \in \mathbb{R}^m$, mediante o sistema de ecuacións

$$Y_i := y + h \sum_{j=1}^{q} a_{ij} f(t + c_j h, Y_j),\ i = 1, \dots, q \tag{2.10}$$

e definimos

$$\phi_f(t, y; h) = \sum_{i=1}^{q} b_i f(t + c_i h, Y_i). \tag{2.11}$$

Veremos, na sección 2.6, que si $h^* > 0$ é suficientemente pequeno, o sistema (2.10) ten solución única e por tanto ϕ_f está ben definida.

Nótese que si $t = t_n$ e $y = y_n$, entón $Y_i = y_{n,i}$, $i = 1, \dots, q$; por conseguinte

$$\phi_f(t_n, y_n; h) = \sum_{i=1}^{q} b_i f(t_n + c_i h, y_{n,i})$$

de maneira que o método xeral (2.9) para esta función incremento ϕ_f coincide co método Runge-Kutta (2.2).

Cabe destacar, que no caso dos Runge-Kutta implícitos, ϕ_f non está dada de forma explícita.

[2] Nesta subsección referímonos tanto aos métodos implícitos como aos semiimplícitos.

2.4. Introducción a consistencia dos métodos Runge-Kutta

Nesta sección, imos considerar que a función incremento ϕ_f é continua en $[t_0, t_0+T] \times \mathbb{R}^m \times [0, h^*]$. Probaremos isto na sección 2.6.

Proposición 2.9. *Unha condición necesaria e suficiente para que un método Runge-Kutta sexa consistente é que*

$$\sum_{i=1}^{q} b_i = 1. \tag{2.12}$$

Demostración. Demostramos no lema 1.3 que unha condición necesaria e suficiente para que o método (1.3) sexa consistente é (1.8).
Recordemos que ϕ_f está ben definida mediante (2.9) e (2.10). Tomando $h = 0$, obtemos

$$Y_i = y,\ i = 1, \dots, q$$

e

$$\phi_f(t, y : 0) = \sum_{i=1}^{q} b_i f(t, Y_i) = \left(\sum_{i=1}^{q} b_i \right) f(t, y).$$

Polo tanto, a condición (1.8) equivale a (2.12). □

2.5. Orde dos métodos Runge-Kutta explícitos para EDOs escalares

Consideremos un método Runge-Kutta explícito de 3 etapas arbitrario que cumpra a condición de filas. Utilizando (2.3), imos escribilo como:

$$\begin{cases} y_{n+1} = y_n + h(b_1 k_1 + b_2 k_2 + b_3 k_3) \\ k_1 = f(t_n, y_n) \\ k_2 = f(t_n + hc_2, y_n + hc_2 k_1) \\ k_3 = f(t_n + hc_3, y_n + h(c_3 - a_{23})k_1 + ha_{32}k_2) \end{cases} \tag{2.13}$$

Nesta sección, para alixeirar a notación, escribiremos k_i en lugar de $k_{n,i}$, para $i = 1, 2, 3$.

Necesitamos obter un desenvolvemento de T_{n+1} en potencias de h. O cálculo de T_{n+1} a partir de (1.5) e (1.6) requeriría a obtención previa da expresión de ϕ_f, o que complicaría os cálculos. Para evitar isto, no caso dos métodos Runge-Kutta explícitos facemos uso da observación 1.1: supoñemos a hipótese de localización $y_n = y(t_n)$, co cal $T_{n+1} = y(t_{n+1}) - \tilde{y}_{n+1}$, e polo tanto o

desenvolvemento requerido de T_{n+1} obtense efectuando un desenvolvemento de $y(t_n+h)=y(t_n+h)$ e un desenvolvemento de $\tilde{y}_{n+1}$.

Imos supoñer que f é suficientemente regular e introducimos a notación

$$f=f(t_n,y(t_n)),\ f_t=\frac{\partial f}{\partial t}(t_n,y(t_n)),\ f_{tt}=\frac{\partial^2 f}{\partial t^2}(t_n,y(t_n)),\ f_{ty}\equiv f_{yt}=\frac{\partial^2 f}{\partial t\partial y}(t_n,y(t_n)).$$

Escribimos o desenvolvemento de Taylor de grado 3 de $y(t_{n+1})$:

$$y(t_{n+1})=y(t_n)+hy'(t_n)+\frac{1}{2}h^2y''(t_n)+\frac{1}{6}h^3y'''(t_n)+O(h^4)$$

Agora,

- $y'(t_n)=f$
- $y''(t_n)=f_t+f_yy'(t_n)=f_t+ff_y=:F$
- $y'''(t_n)=f_{tt}+f_{ty}y'(t_n)+[f_{yy}y'(t_n)+f_yt]y'(t_n)+y''(t_n)f_y=$
  $f_{tt}+ff_{ty}+[f_{yy}f+f_{yt}]f+[f_t+ff_y]f_y=f_{tt}+ff_{ty}+f^2f_{yy}+ff_{yt}+f_yF=$
  $f_{tt}+2ff_{ty}+f^2f_{yy}+f_yF=G+f_yF$

onde $F=f_t+ff_y$ e $G=f_{tt}+2f_{ty}+f^2f_{yy}$. Así, podemos escribir

$$y(t_{n+1})=y(t_n)+hf+\frac{1}{2}h^2F+\frac{1}{6}h^3(G+f_yF)+O(h^4). \tag{2.14}$$

Facendo desenvolvementos de Taylor en (2.12) temos que:

- $k_1=f$
- $k_2=f+c_2hf_t+hc_2k_1f_y+\frac{1}{2}\left[(c_2h)^2f_{tt}+2c_2h(hc_2k_1)f_{ty}+(hc_2k_1)^2f_{yy}\right]+O(h^3)=$
  $f+c_2h[f_t+ff_y]+\frac{1}{2}c_2^2h^2\left[f_{tt}+2ff_{ty}+f^2f_{yy}\right]+O(h^3)=f+c_2hF+\frac{1}{2}c_2^2h^2G+O(h^3)$
- $k_3=f+hc_3f_t+[h(c_3-a_{32})k_1+ha_{32}k_2]\,f_y+$
  $\frac{1}{2}\left\{(c_3h)^2f_{tt}+2(c_3h)\,[h(c_3-a_{32})k_1+ha_{32}k_2]\,f_{ty}+[h(c_3-a_{32})k_1+ha_{32}k_2]^2\,f_{yy}\right\}+O(h^3)=$
  $f+h\left\{c_3f_t+[(c_3-a_{32})k_1+a_{32}k_2]\,f_y\right\}+$
  $\frac{1}{2}h^2\left\{c_3^2f_{tt}+2c_3\,[(c_3-a_{32})k_1+a_{32}k_2]\,f_{ty}+[(c_3-a_{32})k_1+a_{32}k_2]^2\,f_{yy}\right\}+O(h^3).$

Se substituímos k_2 en k_3 obtemos:

$k_3=f+h\left\{c_3f_t+\left[(c_3-a_{32})k_1+a_{32}(f+hc_2F+O(h^2))\right]f_y\right\}+$
$\frac{1}{2}h^2\left\{c_3^2f_{tt}+2c_3\,[(c_3-a_{32})k_1+a_{32}(f+O(h))]\,f_{ty}+[(c_3-a_{32})k_1+a_{32}(f+O(h))]^2\,f_{yy}\right\}+O(h^3)=$
$f+h\left\{c_3(f_t+ff_y)+ha_{32}c_2Ff_y\right\}+\frac{1}{2}h^2c_3^2\left\{f_{tt}+2ff_{ty}+f^2f_{yy}\right\}+O(h^3)=f+hc_3F+h^2\left\{a_{32}c_2Ff_y+\frac{1}{2}c_3^2G\right\}+$

$O(h^3)$,
onde a segunda igualdade se debe a que
$[(c_3 - a_{32})k_1 + a_{32}(f + O(h))]^2 = [c_3 f + O(h)]^2 = c_3^2 f^2 + 2c_3 f O(h) + O(h^2) = c_3^2 f^2 + O(h)$.

Así, substituíndo k_1, k_2 e k_3 na primeira ecuación de (2.13) e usando a hipótese de localización $y_n = y(t_n)$ obtemos:

$$\begin{aligned}
&\tilde{y}_{n+1} = y(t_n) + \\
&h\left[b_1 f + b_2\left(f + c_2 hF + \frac{1}{2}c_2^2 h^2 G + O(h^3)\right) + b_3\left(f + hc_3 F + h^2\left\{a_{32}c_2 F f_y + \frac{1}{2}c_3^2 G\right\}\right) + O(h^3)\right] = \\
&y(t_n) + h\left(b_1 + b_2 + b_3\right) f + h^2\left(b_2 c_2 + b_3 c_3\right) F + \frac{1}{2}h^3\left[2b_3 c_2 a_{32} F f_y + \left(b_2 c_2^2 + b_3 c_3^2\right) G\right] + O(h^4).
\end{aligned} \tag{2.15}$$

Lema 2.10. *Un método Runge-Kutta explícito de dúas etapas que cumpre a condición de filas é de orde* 2 *se, e só se,* $b_1 + b_2 = 1$ *e* $b_2 c_2 = \frac{1}{2}$.

Demostración. Se o método Runge-Kutta é de dúas etapas, entón verifícase que $b_3 = 0$. Logo, podemos escribir:

$$\tilde{y}_{n+1} = y(t_n) + h(b_1 + b_2)f + h^2 b_2 c_2 F + \frac{1}{2}h^3 b_2 c_2^2 G + O(h^4).$$

Por outra parte, o desenvolvemento en serie de Taylor da solución ven dado por (2.14).
Para que o método teña orde de consistencia polo menos dous, debe cumprirse que
$\tau_{n+1} = O(h^2)$, é dicir, $T_{n+1} = O(h^3)$. Agora ben, baixo a hipótese de localización temos

$$T_{n+1} = y(t_{n+1}) - \tilde{y}_{n+1}.$$

Por tanto, que $T_{n+1} = O(h^3)$ equivale a que se verifiquen as condicións do enunciado.
Dado que o término en h^3 dos desenvolvementos de $\tilde{y}_{n+1}$ e $y(t_{n+1})$, en xeral, é distinto, non se alcanza a orde 3. □

Lema 2.11. *Un método Runge-Kutta explícito de tres etapas que cumpre a condición de filas é polo menos de orde* 3 *se, e só se,* $b_1 + b_2 + b_3 = 1$, $b_2 c_2 + b_3 c_3 = \frac{1}{2}$, $b_2 c_2^2 + b_3 c_3^2 = \frac{1}{3}$ *e* $b_3 c_2 a_{32} = \frac{1}{6}$.

Demostración. Para un método explícito de tres etapas temos o desenvolvemento de Taylor (2.15). Por outra parte, o desenvolvemento de Taylor de grado 3 da solución é (2.14). Para que o método teña orde de consistencia polo menos tres debe cumprirse que $\tau_{n+1} = O(h^3)$, é dicir, $T_{n+1} = O(h^4)$, o que equivale a que se verifiquen as condicións do enunciado. □

2.6. Propiedades da función incremento (I). Consistencia e estabilidade dos métodos Runge-Kutta

Consideraremos, a partir de agora, $Y = (Y_1, \dots, Y_q)^T$, onde $Y_i \in \mathbb{R}^m$, $i = 1, \dots, q$ e imos a tomar a norma produto en $(\mathbb{R}^m)^q$ dada por $\|Y\|_\infty = \text{máx}_{1\leq i\leq q}\|Y_i\|$.

Lema 2.12. *Existe $h^* > 0$ tal que $\forall t \in [t_0, t_0 + T]$, $\forall y \in \mathbb{R}^m$, $\forall h \in [0, h^*]$ o sistema (2.9) ten solución única.*

Demostración. Dados y, t e h (tal que $h\,\text{máx}_{1\leq j\leq q}|c_j| \leq \delta_1$) sexa $\theta : (\mathbb{R}^m)^q \to (\mathbb{R}^m)^q$ a aplicación definida por

$$\theta_i(Y) = y + h\sum_{j=1}^{q} a_{ij} f(t + c_j h, Y_j),\ i = 1, \dots, q.$$

A ecuación (2.10) equivale a $Y_i = \theta_i(Y)$, $i = 1, \dots, q$, e polo tanto a $Y = \theta(Y)$.
Probaremos que para $h > 0$ suficientemente pequeno θ é contractiva

$$\forall Y, Z \in (\mathbb{R}^m)^q\ \|\theta_i(Y) - \theta_i(Z)\| = \|h\sum_{j=1}^{q} a_{ij}\left[f(t + hc_j, Y_j) - f(t + hc_j, Z_j)\right]\| \leq$$

$$h\sum_{j=1}^{q}|a_{ij}|\|f(t + hc_j, Y_j) - f(t + hc_j, Z_j)\| \leq hL\sum_{j=1}^{q}|a_{ij}|\|Y_j - Z_j\|$$

$$\leq hL\left(\sum_{j=1}^{q}|a_{ij}|\right)\|Y - Z\|_\infty,\ i = 1, \dots, q,$$

o que implica que

$$\|\theta(Y) - \theta(Z)\|_\infty = \max_{1\leq i\leq q}\|\theta_i(Y) - \theta_i(Z) \leq hL\left(\max_{1\leq i\leq q}\sum_{j=1}^{q}|a_{ij}|\right)\|Y - Z\|_\infty =$$

$$hL\|A\|_\infty\|Y - Z\|_\infty,\ \forall Y, Z \in (\mathbb{R}^m)^q.$$

Se $h > 0$ cumpre que $hL\|A\|_\infty < 1$, entón θ é contractiva en $((\mathbb{R}^m)^q, \|.\|_\infty)$, polo que θ ten un punto fixo e só un, e en consecuencia (2.10) ten solución única. Así pois, basta tomar $h^* > 0$ tal que $h^*L\|A\|_\infty < 1$ e $h^*\,\text{máx}_{1\leq j\leq q}|c_j| \leq \delta_1$. □

Observación 2.13. En virtude do lema 2.12, a función ϕ_f está ben definida en $[t_0, t_0 + T] \times \mathbb{R}^m \times [0, h^*]$.

Supoñeremos nesta sección, a partir de agora, que $h^*L\|A\|_\infty < 1$ e $h^*\,\text{máx}_{1\leq j\leq q}|c_j| \leq \delta_1$.

Lema 2.14. *A función ϕ_f definida por (2.10) e (2.11) é globalmente lipschitziana respecto de y, uniformemente en t e h.*

Demostración. Sexa

$$\begin{cases} g : [t_0, t_0 + T] \times \mathbb{R}^m \times [0, h^*] \to \mathbb{R}^m \times \overbrace{\ldots}^{q} \times \mathbb{R}^m \\ (t, y, h) \mapsto Y = (Y_1, \ldots, Y_q)^T \end{cases} \tag{2.16}$$

onde Y é solución única de (2.10). Dado que

$$\phi_f(t, y; h) = \sum_{i=1}^{q} b_i f(t + c_i h, g_i(t, y, h)) \tag{2.17}$$

e f é globalmente lipschitziana na segunda variable, uniformemente na primeira, basta demostrar que g é lipschitziana respecto de y, uniformemente en t e h.
Dados $t \in [t_0, t_0 + T]$, $h \in [0, h^*]$ e $y, z \in \mathbb{R}^m$ sexan $Y = (Y_1, \ldots, Y_q)^T$ e $Z = (Z_1, \ldots, Z_q)^T$ as solucións respectivas de

$$Y_i = y + h \sum_{j=1}^{q} a_{ij} f(t + c_j h, Y_j),\ i = 1, \ldots, q$$

$$Z_i = z + h \sum_{j=1}^{q} a_{ij} f(t + c_j h, Z_j),\ i = 1, \ldots, q$$

(é dicir, $Y = g(t, y, h)$ e $Z = g(t, z, h)$).
Así,

$$Y_i - Z_i = y - zh \sum_{j=1}^{q} a_{ij} \left[f(t + c_j h, Y_j) - f(t + c_j h, Z_j) \right],\ i = 1, \ldots, q.$$

Entón,

$$\|Y_i - Z_i\| \leq \|y - z\| + hL \sum_{j=1}^{q} |a_{ij}| \|Y_j - Z_j\| \leq \|y - z\| + hL \left(\sum_{j=1}^{q} |a_{ij}| \right) \|Y - Z\|_\infty$$

o que implica que

$$\|Y - Z\|_\infty = \max_{1 \leq i \leq q} \|Y_i - Z_i\| \leq \|y - z\| + hL \left(\max_{1 \leq i \leq q} \sum_{j=1}^{q} |a_{ij}| \right) \|Y - Z\|_\infty \leq$$

$$\|y - z\| + h^* L \|A\|_\infty \|Y - Z\|_\infty.$$

Polo tanto, obtemos que

$$(1 - h^* L \|A\|_\infty) \|Y - Z\|_\infty \leq \|y - z\|$$

e como $1 - h^* L \|A\|_\infty > 0$

$$\|g(t, y, h) - g(t, z, h)\|_\infty = \|Y - Z\|_\infty \leq \frac{1}{1 - h^* L \|A\|_\infty} \|y - z\|,\ \forall y, z \in \mathbb{R}^m.$$

□

Lema 2.15. *A función ϕ_f definida por (2.10) e (2.11) é continua.*

Demostración. Dado que ϕ_f cumpre (2.17) e que f é continua, para probar a continuidade de ϕ_f é suficiente con probar que a función g dada por (2.16) é continua.
Definamos

$$\begin{cases} \beta : (t, y, h, Y) \in [t_0, t_0 + T] \times \mathbb{R}^m \times [0, h^*] \times (\mathbb{R}^m)^q \to \beta(t, y, h, Y) \in (\mathbb{R}^m)^q, \\ \beta_i(t, y, h, Y) := y + h \sum_{j=1}^{q} a_{ij} f(t + c_j h, Y_j),\ i = 1, \dots, q \end{cases} \tag{2.18}$$

Pola definición de g, a ecuación $Y = g(t, y, h)$ equivale a $Y = \beta(t, y, h, Y)$. O razoamento efectuado na demostración do lema 2.12 proba que de feito β cumpre a condición de Lipschitz

$$\|\beta(t, y, h, Y) - \beta(t, y, h, Z)\| \leq h^* L \|A\|_\infty \|Y - Z\|_\infty,\ \forall Y, Z \in (\mathbb{R}^m)^q,\ \forall (t, y, h) \in [t_0, t_0 + T] \times \mathbb{R}^m \times [0, h^*]$$

onde $h^* > 0$ é tal que $h^* L \|A\|_\infty < 1$.
Entón, β é unha contracción respecto de Y uniforme en (t, y, h); ademais β é continua xa que f é continua.
Sexa $(\bar{t}, \bar{y}, \bar{h}) \in [t_0, t_0 + T] \times \mathbb{R}^m \times [0, h^*]$. Vexamos que g é continua no devandito punto, para iso consideremos outro punto $(t, y, h) \in [t_0, t_0 + T] \times \mathbb{R}^m \times [0, h^*]$. Denotemos por $\bar{Y} = g(\bar{t}, \bar{y}, \bar{h})$ e $Y = g(t, y, h)$, en consecuencia tense que $\bar{Y} = \beta(\bar{t}, \bar{y}, \bar{h}, \bar{Y})$ e $Y = \beta(t, y, h, Y)$. Logo,

$$Y - \bar{Y} = \beta(t, y, h, Y) - \beta(\bar{t}, \bar{y}, \bar{h}, \bar{Y}) = \beta(t, y, h, Y) - \beta(t, y, h, \bar{Y}) + \beta(t, y, h, \bar{Y}) - \beta(\bar{t}, \bar{y}, \bar{h}, \bar{Y}).$$

Así,

$$\|Y - \bar{Y}\|_\infty \leq \|\beta(t, y, h, Y) - \beta(t, y, h, \bar{Y})\|_\infty + \|\beta(t, y, h, \bar{Y}) - \beta(\bar{t}, \bar{y}, \bar{h}, \bar{Y})\|_\infty \leq$$

$$h^* L \|A\|_\infty \|Y - \bar{Y}\|_\infty + \|\beta(t, y, h, \bar{Y}) - \beta(\bar{t}, \bar{y}, \bar{h}, \bar{Y})\|_\infty,$$

o que implica que

$$(1 - h^* L \|A\|_\infty) \|Y - \bar{Y}\|_\infty \leq \|\beta(t, y, h, \bar{Y}) - \beta(\bar{t}, \bar{y}, \bar{h}, \bar{Y})\|_\infty$$

e dado que $1 - h^* L \|A\|_\infty > 0$,

$$\|Y - \bar{Y}\|_\infty \leq \frac{1}{1 - h^* L \|A\|_\infty} \|\beta(t, y, h, \bar{Y}) - \beta(\bar{t}, \bar{y}, \bar{h}, \bar{Y})\|_\infty. \tag{2.19}$$

Como β é continua en $(\bar{t}, \bar{y}, \bar{h}, \bar{Y})$, dado calquera $\varepsilon > 0$ existe $\delta > 0$ tal que si

$$|t - \bar{t}| \leq \delta,\ \|y - \bar{y}\|_\infty \leq \delta,\ |h - \bar{h}| \leq \delta$$

entón

$$\|\beta(t, y, h, \bar{Y}) - \beta(\bar{t}, \bar{y}, \bar{h}, \bar{Y})\|_\infty < \varepsilon \left(1 - h^* L \|A\|_\infty\right),$$

e tendo en conta a desigualdade (2.19)

$$\|g(t,y,h) - g(\bar{t},\bar{y},\bar{h})\|_\infty = \|Y - \bar{Y}\|_\infty \leq \varepsilon;$$

por tanto g é continua no punto $(\bar{t},\bar{y},\bar{h})$. Dado que $(\bar{t},\bar{y},\bar{h})$ é un punto arbitrario de $[t_0,t_0+T]\times\mathbb{R}^m\times[0,h^*]$, g é continua. □

Teorema 2.16. *Todos os métodos Runge-Kutta son estables.*

Demostración. En virtude do lema 2.14 e do lema 2.15 sabemos que a función ϕ_f definida por (2.10) e (2.11) é continua e globalmente lipschitziana respecto de y, uniformemente en t e h.
En consecuencia, verifícanse as hipóteses do lema 1.7, polo que podemos afirmar que todos os métodos Runge-Kutta son estables. □

Corolario 2.17. *Todos os métodos Runge-Kutta consistentes son convexentes.*

Demostración. Polo teorema 2.16 sabemos que todos os métodos Runge-Kutta son estables. En consecuencia, empregando o teorema 1.9 chegamos a conclusión de que todo método Runge-Kutta consistente é convexente. □

2.7. Propiedades da función incremento (II). Regularidade da función incremento

Nesta sección supoñeremos $h^*L\|A\|_\infty < 1$ e $2h^*\,\text{máx}_{1\leq j\leq q}|c_j| \leq \delta_1$. Isto permite definir $\beta : [t_0-h^*, t_0+T+h^*]\times\mathbb{R}^m\times[-h^*,h^*]\times(\mathbb{R}^m)^q \to (\mathbb{R}^m)^q$ por medio da ecuación (2.18).

Lema 2.18. *Se $f\in C^k([t_0-\delta_1,t_0+T+\delta_1]\times\mathbb{R}^m;\mathbb{R}^m)$ entón a función ϕ_f definida por (2.10) e (2.11) verifica que $\phi_f\in C^k([t_0,t_0+T]\times\mathbb{R}^m\times[0,h^*);\mathbb{R}^m)$.*

Demostración. En virtude de (2.17) para probar que $\phi_f\in C^k([t_0,t_0+T]\times\mathbb{R}^m\times[0,h^*];\mathbb{R}^m)$ é suficiente probar que a función g dada por (2.16) verifica que $g\in C^k([t_0,t_0+T]\times\mathbb{R}^m\times[0,h^*];(\mathbb{R}^m)^q)$.
Pola definición de g e da función β, dada por (2.18), $Y = g(t,y,h)$ equivale a que

$$Y - \beta(t,y,h,Y) = 0.$$

Imos considerar $F(t,y,h,Y) = Y - \beta(t,y,h,Y)$.
Dado que $f\in C^k([t_0-\delta_1,t_0+T+\delta_1]\times\mathbb{R}^m;\mathbb{R}^m)$, tense que $\beta\in C^k([t_0-h^*,t_0+T+h^*]\times\mathbb{R}^m\times[-h^*,h^*]\times(\mathbb{R}^m)^q;(\mathbb{R}^m)^q)$ e por tanto $F\in C^k([t_0,t_0+T]\times\mathbb{R}^m\times[-h^*,h^*]\times(\mathbb{R}^m)^q;(\mathbb{R}^m)^q)$.
Consideremos agora $\bar{t}\in[t_0,t_0+T]$, $\bar{y}\in\mathbb{R}^m$, $\bar{h}\in[0,h^*)$ e $\bar{Y} = g(\bar{t},\bar{y},\bar{h})$. Vexamos que

$D_yF(\bar{t},\bar{y},\bar{h},\bar{Y}) \in M_{mq\times mq}(\mathbb{R})$ é invertible.
Por definición de F, $D_YF(\bar{t},\bar{y},\bar{h},\bar{Y}) = I_{mq} - D_Y\beta(\bar{t},\bar{y},\bar{h},\bar{Y}) \in M_{mq\times mq}(\mathbb{R})$. Esta última matriz ten unha estrutura por bloques de tamaño $m \times m$, onde os bloques están dados por

$$\left[D_y\beta(\bar{t},\bar{y},\bar{h},\bar{Y})\right]_{il} = \bar{h}a_{il}\frac{\partial f}{\partial y}(\bar{t}+c_l\bar{h},\bar{Y}_l),\ 1 \le i,l \le q.$$

Entón, en virtude dos lemas A.1 e A.2

$$\|D_y\beta(\bar{t},\bar{y},\bar{h},\bar{Y})\|_\infty \le \max_{1\le i\le q}\sum_{l=1}^{q}\|\left[D_y\beta(\bar{t},\bar{y},\bar{h},\bar{Y})\right]_{il}\| \le \bar{h}\max_{1\le i\le q}\sum_{l=1}^{q}|a_{il}|\|\frac{\partial f}{\partial y}(\bar{t}+c_l\bar{h},\bar{Y}_l)\|$$

$$\le \bar{h}L\max_{1\le i\le q}\sum_{l=1}^{q}|a_{il}| = L\bar{h}\|A\|_\infty < 1$$

dado que $\bar{h} \in [0,h^*)$ e $h^*L\|A\|_\infty < 1$, donde usamos a notación $\|.\|$ (respectivamente $\|.\|_\infty$) para denotar a norma en $M_{m\times m}(\mathbb{R})$ (respectivamente $M_{mq\times mq}(\mathbb{R})$) subordinada á norma $\|.\|$ de $\mathbb{R}^m$ (respectivamente á norma $\|.\|_\infty$ de $(R^m)^q$ introducida na sección 2.6). En consecuencia, a matriz $D_yF(\bar{t},\bar{y},\bar{h},\bar{Y})$ é non singular e facendo uso do Teorema da Función Implícita podemos dicir que

$$g \in C^k([t_0,t_0+T]\times\mathbb{R}^m\times[0,h^*);(\mathbb{R}^m)^q).$$

□

2.8. Condicións de orde dos métodos Runge-Kutta

Teorema 2.19. *Unha condición necesaria e suficiente para que un método Runge-Kutta sexa polo menos de orde* 1 *é* $b^Te = 1$. *Unha condición necesaria e suficiente para que un método Runge-Kutta sexa polo menos de orde* 2 *é* $b^Te = 1$ *e* $b^TCe = b^TAe = \frac{1}{2}$.

Demostración. Por (2.10) e (2.11) podemos escribir $\phi_f(t,y;0) = \sum_{j=1}^{q} b_jf(t,y)$. O teorema 1.12 dinos que un método é de polo menos orde 1 se, e só se, $\phi_f(t,y;0) = f(t,y)\ \forall t \in [0,T],\ \forall y \in \mathbb{R}^m$ o que equivale a que $\sum_{j=1}^{q} b_j = b^Te = 1$.
Ademais, deducimos de (2.10) e (2.11) as relacións

$$\frac{\partial Y_i}{\partial h} = \sum_{j=1}^{q} a_{ij}f(t+c_jh,Y_j) + h\sum_{j=1}^{q} a_{ij}\frac{\partial}{\partial h}\left[f(t+c_jh,Y_j)\right]$$

$$\frac{\partial \phi}{\partial h}(t,y;h) = \sum_{i=1}^{q} b_ic_iD_tf(t+c_ih,Y_i) + \sum_{i=1}^{q} b_iD_yf(t+c_ih,Y_i)\frac{\partial Y_i}{\partial h}$$

e por tanto

$$\frac{\partial Y_i}{\partial h}_{|h=0} = \sum_{j=0}^{q} a_{ij}f(t,y)$$

$$\frac{\partial \phi}{\partial h}_{|h=0} = \sum_{i=1}^{q} b_i c_i D_t f(t,y) + \sum_{i=1}^{q} \left(b_i \sum_{j=1}^{q} a_{ij} \right) D_y f(t,y) f(t,y).$$

Temos que

$$\frac{\partial \phi}{\partial h}(t,y;0) = \frac{1}{2} f^{(1)}(t,y) = \frac{1}{2} D_t f(t,y) + \frac{1}{2} D_y(t,y) f(t,y)$$

para calquera función $f(t,y)$ continuamente derivable se, e só se, $\sum_{i=1}^{q} b_i c_i = b^T C e = \frac{1}{2}$ e $\sum_{i=1}^{q} \left(b_i \sum_{j=1}^{q} a_{ij} \right) = b^T A e = \frac{1}{2}$. □

Observación 2.20. A condición necesaria e suficiente para que un método Runge-Kutta sexa polo menos de orde 1 coincide coa condición necesaria e suficiente para que sexa consistente. En consecuencia, un método Runge-Kutta é consistente se, e só se, é polo menos de orde 1.

Exemplo 2.21. Consideremos o método Runge-Kutta uniparamétrico de dúas etapas xa visto no exemplo 2.3. Da súa táboa de Butcher obtemos que

$$A = \begin{pmatrix} 0 & 0 \\ \alpha & 0 \end{pmatrix}, \; b^T = \begin{pmatrix} 1 - \frac{1}{2\alpha} & \frac{1}{2\alpha} \end{pmatrix}, \; c^T = \begin{pmatrix} 0 & \alpha \end{pmatrix} \text{ e } C = \begin{pmatrix} 0 & 0 \\ 0 & \alpha \end{pmatrix}$$

Vexamos se se cumpren as condicións de orde dadas polo teorema 2.19

$b^T e = \begin{pmatrix} 1 - \frac{1}{2\alpha} & \frac{1}{2\alpha} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1$

$b^T C e = \begin{pmatrix} 1 - \frac{1}{2\alpha} & \frac{1}{2\alpha} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & \alpha \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{2}$

$b^T A e = \begin{pmatrix} 1 - \frac{1}{2\alpha} & \frac{1}{2\alpha} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ \alpha & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{2}$

Polo tanto, en virtude do teorema 2.19, podemos dicir que este método é polo menos de orde 2 para todo $\alpha \in \mathbb{R} \setminus \{0\}$.

Outra forma de ver isto (pero só para EDOs escalares), podería ser partir de que estamos ante un método explícito de dúas etapas que cumpre a condición de filas, xa que

$$Ae = \begin{pmatrix} 0 & 0 \\ \alpha & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \alpha \end{pmatrix} = c$$

e como se cumpren as condicións de orde dadas polo lema 2.10

$b_1 + b_2 = 1 - \frac{1}{2\alpha} + \frac{1}{2\alpha} = 1$

$b_2 c_2 = \frac{1}{2\alpha} \alpha = \frac{1}{2}$

podemos dicir que o método é de orde 2 para EDOs escalares, para todo $\alpha \in \mathbb{R} \setminus \{0\}$.

Teorema 2.22. *Supoñamos que Ae=Ce; unha condición necesaria e suficiente para que un método Runge-Kutta sexa de polo menos orde* 3 *escríbese* $b^T e = 1$, $b^T C e = \frac{1}{2}$, $b^T C^2 e = \frac{1}{3}$, $b^T A C e = \frac{1}{6}$. *Para que o método sexa de polo menos orde* 4 *é necesario e suficiente ter ademais que* $b^T C^3 e = \frac{1}{4}$,$b^T A C^2 e = \frac{1}{12}$,$b^T A^2 C e = \frac{1}{24}$ *e* $b^T C A C e = \frac{1}{8}$.

Demostración. Facemos a demostración para a orde 4. Sexa

$$T_{n+1} = y(t_{n+1}) - y(t_n) - h\phi_f(t_n, y(t_n); h).$$

Polas condicións (1.19), para que un método sexa polo menos de orde p, é necesario e suficiente que para calquera solución $y(t)$ de $y'(t) = f(t, y(y, t))$ teñamos

$$T_{n+1} = O(h^{p+1}).$$

Sexa $z_{n,i},\ i = 1, \dots, q$, a solución do sistema

$$z_{n,i} = y(t_n) + h\sum_{j=1}^{q} a_{ij} f(t_{n,j}, z_{n,j}),\ i = 1, \dots, q; \tag{2.20}$$

(esta solución existe para $h \in [0, h^*]$); temos a partir de (2.9) e (2.10)

$$T_{n+1} = y(t_{n+1}) - y(t_n) - h\sum_{j=1}^{q} b_j f(t_{n,j}, z_{n,j}) \tag{2.21}$$

Por outro lado, según a fórmula de Taylor, temos

$$y(t_{n+1}) = y(t_n + h) = y(t_n) + hy'(t_n) + \frac{1}{2!}h^2 y''(t_n) + \frac{1}{3!}h^3 y'''(t_n) + \frac{1}{4!}h^4 y^{(4)}(t_n) + O(h^5)$$

$$y'(t_{n,j}) = y'(t_n + c_j h) = y'(t_n) + hc_j y''(t_n) + \frac{1}{2!}h^2 c_j^2 y'''(t_n) + \frac{1}{3!}h^3 c_j^3 y^{(4)}(t_n) + O(h^4)$$

entón

$$y(t_{n+1}) - y(t_n) - h\sum_{j=1}^{q} b_j y'(t_{n,j}) = h\left(1 - \sum_{j=1}^{q} b_j\right) y'(t_n) + h^2\left(\frac{1}{2} - \sum_{j=1}^{q} b_j c_j\right) y''(t_n) +$$

$$\frac{h^3}{2}\left(\frac{1}{3} - \sum_{j=1}^{q} b_j c_j^2\right) y'''(t_n) + \frac{h^4}{6}\left(\frac{1}{4} - \sum_{j=1}^{q} b_j c_j^3\right) y^{(4)}(t_n) + O(h^5) \tag{2.22}$$

Sumando e restando $h\sum_{j=1}^{q} b_j y'(t_{n,j})$ no segundo membro de (2.21), e usando (2.22) e

$$y'(t_{n,j}) = f(t_{n,j}, y(t_{n,j})) \tag{2.23}$$

dedúcese que

$$T_{n+1} = h\sum_{j=1}^{q} b_j \left[f(t_{n+j}, y(t_{n+j})) - f(t_{n+j}, z_{n+j})\right] + h\left(1 - b^T e\right) y'(t_n) + h^2\left(\frac{1}{2} - b^T C e\right) y''(t_n) +$$

$$\frac{h^3}{2}\left(\frac{1}{3} - b^T C^2 e\right) y'''(t_n) + \frac{h^4}{6}\left(\frac{1}{4} - b^T C^3 e\right) y^{(4)}(t_n) + O(h^5). \tag{2.24}$$

Vamos a obter algunhas das condicións necesarias para que un método sexa de orde 4. Para iso, tomamos sucesivamente $f(t, y) = kt^{k-1}$ e $y(t) = t^k,\ k = 1, 2, 3, 4$.

Para $k=1$, $T_{n+1}=h(1-b^Te)+O(h^5)$; por tanto $T_{n+1}=O(h^5)$ se, e só se, $b^Te=1$.
Supoñamos que se cumpre esta condición. Entón, para $k=2$, $T_{n+1}=2h^2\left(\frac{1}{2}-b^TCe\right)+O(h^3)$; por tanto $T_{n+1}=O(h^5)$ se, e só se, $b^TCe=\frac{1}{2}$.
Supoñamos que se cumpren as condicións $b^Te=1$ e $b^TCe=\frac{1}{2}$. Entón, para $k=3$,
$T_{n+1}=6\frac{h^3}{2}\left(\frac{1}{3}-b^TC^2e\right)+O(h^4)$, logo $T_{n+1}=O(h^5)$ se, e só se, $b^TC^2e=\frac{1}{3}$.
Supoñamos que se cumpren $b^Te=1$, $b^TCe=\frac{1}{2}$ e $b^TC^2e=\frac{1}{3}$. Entón, para $k=4$,
$T_{n+1}=24\frac{h^4}{6}\left(\frac{1}{4}-b^TC^3e\right)+O(h^5)$; por tanto, $T_{n+1}=O(h^5)$ se, e só se, $b^TC^3e=\frac{1}{4}$.
Así, obtemos que unhas condicións necesarias para que o método sexa polo menos de orde 4 é que teñamos

$$1=b^Te,\ \frac{1}{2}=b^TCe,\ \frac{1}{3}=b^TC^2e,\ \frac{1}{4}=b^TC^3e. \tag{2.25}$$

Supoñendo que se cumpren estas condicións, a ecuación (2.24) redúcese a

$$T_{n+1}=h\sum_{j=1}^{q}b_j\left[f(t_{n,j},y(t_{n,j}))-f(t_{n,j},z_{n,j})\right]+O(h^5). \tag{2.26}$$

Utilizando de novo a fórmula de Taylor temos
$y(t_{n,i})=y(t_n+c_ih)=y(t_n)+hc_iy'(t_n)+\frac{1}{2}h^2c_i^2y''(t_n)+\frac{1}{6}h^3c_i^3y'''(t_n)+O(h^4)$
$y'(t_{n,j})=y'(t_n+c_jh)=y'(t_n)+hc_jy''(t_n)+\frac{1}{2}h^2c_j^2y'''(t_n)+O(h^3)$
$\sum_{j=1}^{q}a_{ij}y'(t_{n,j})=\left(\sum_{j=1}^{q}a_{ij}\right)y'(t_n)+h\left(\sum_{j=1}^{q}a_{ij}c_j\right)y''(t_n)+\frac{h^2}{2}\left(\sum_{j=1}^{q}a_{ij}c_j^2\right)y'''(t_n)+O(h^3)$
polo que, dado que Ae=Ce
$y(t_{n,i})=y(t_n)+h\left\{\sum_{j=1}^{q}a_{ij}y'(t_{n,j})-h\left(\sum_{j=1}^{q}a_{ij}c_j\right)y''(t_n)-\frac{h^2}{2}\left(\sum_{j=1}^{q}a_{ij}c_j^2\right)y'''(t_n)+O(h^3)\right\}+$
$\frac{1}{2}h^2c_i^2y''(t_n)+\frac{1}{6}h^3c_i^3y'''(t_n)+O(h^4)=y(t_n)+h\sum_{j=1}^{q}a_{ij}y'(t_{n,j})+\frac{h^2}{2}\left(c_i^2-2\sum_{j=1}^{q}a_{ij}c_j\right)y''(t_n)+$
$\frac{h^3}{6}\left(c_i^3-3\sum_{j=1}^{q}a_{ij}c_j^2\right)y'''(t_n)+O(h^4)$.
Entón, restando (2.20) e usando (2.23), resulta

$$y(t_{n,i})-z_{n,i}=h\sum_{j=1}^{q}a_{ij}\left[f(t_{n,j},y(t_{n,j}))-f(t_{n,j},z_{n,j})\right]+\frac{h^2}{2}\left(c_i^2-2\sum_{j=1}^{q}a_{ij}c_j\right)y''(t_n)+$$

$$\frac{h^3}{6}\left(c_i^3-3\sum_{j=1}^{q}a_{ij}c_j^2\right)y'''(t_n)+O(h^4). \tag{2.27}$$

Utilizando a desigualdade (que é consecuencia da lipschitzianidade (2.1))

$$\|f(t_{n,j},y(t_{n,j}))-f(t_{n,j},z_{n,j})\|\leq L\|y(t_{n,j})-z_{n,j}\|, \tag{2.28}$$

obtemos

$$\|y(t_{n,i})-z_{n,i}\|\leq hL\left(\sum_{j=1}^{q}|a_{ij}|\|y(t_{n,j})-z_{n,j}\|\right)+O(h^2)$$

entón,

$$\max_{1\leq i\leq q}\|y(t_{n,i})-z_{n,i}\|\leq Lh\|A\|_\infty\left(\max_{1\leq j\leq q}\|y(t_{n,j}-z_{n,j})\|\right)+O(h^2)$$

o que implica que

$$(1 - Lh\|A\|_\infty) \max_{1\le i\le q} \|y(t_{n,i}) - z_{n,i}\| \le O(h^2)$$

e en consecuencia

$$\max_{1\le i\le q} \|y(t_{n,i}) - z_{n,i}\| \le \frac{1}{1 - Lh\|A\|_\infty} O(h^2) \le \frac{1}{1 - Lh^*\|A\|_\infty} O(h^2) = O(h^2) \tag{2.29}$$

xa que $0 \le h \le h^*$ e $Lh^*\|A\|_\infty < 1$. Usando agora (2.28) e (2.29) en (2.27), resulta

$$y(t_{n,i}) - z_{n,i} = \frac{h^2}{2} \left(c_i^2 - 2\sum_{j=1}^{q} a_{ij}c_j \right) y''(t_n) + O(h^3),\ i = 1, \dots, q; \tag{2.30}$$

denotando $g(t) = D_y f(t, y(t))$ e utilizando a fórmula de Taylor,

$$f(t_{n,j}, z_{n,j}) = f(t_{n,j}, y(t_{n,j})) + D_y(t_{n,j}, y(t_{n,j}))(z_{n,j} - y(t_{n,j})) + O(\|y(t_{n,j}) - z_{n,j}\|^2)$$

así, debido a (2.30)

$$f(t_{n,j}, y(t_{n,j})) - f(t_{n,j}, z_{n,j}) = g(t_{n,j})(y(t_{n,j}) - z_{n,j}) + O(h^4) =$$
$$g(t_n + O(h))(y(t_{n,j}) - z_{n,j}) + O(h^4) = g(t_n)(y(t_{n,j}) - z_{n,j}) + O(h^3).$$

Insertando isto en (2.27), obtemos

$$y(t_{n,i}) - z_{n,i} = h\sum_{j=1}^{q} a_{ij} \left[g(t_n)(y(t_{n,j}) - z_{n,j}) + O(h^3) \right] + \frac{h^2}{2} \left(c_i^2 - 2\sum_{j=1}^{q} a_{ij}c_j \right) y''(t_n) +$$
$$\frac{h^3}{6} \left(c_i^3 - 3\sum_{j=1}^{q} a_{ij}c_j^2 \right) y'''(t_n) + O(h^4)$$

e usando agora (2.25) para substituír os vectores $y(t_{n,j})$ e $z_{n,j}$ que aparecen no segundo membro, resulta

$$y(t_{n,i}) - z_{n,i} = \frac{h^3}{2} \sum_{j=1}^{q} a_{ij} g(t_n) \left(c_j^2 - 2\sum_{l=1}^{q} a_{lj}c_l \right) y''(t_n) + \frac{h^2}{2} \left(c_i^2 - 2\sum_{j=1}^{q} a_{ij}c_j \right) y''(t_n) +$$
$$\frac{h^3}{6} \left(c_i^3 - 3\sum_{j=1}^{q} a_{ij}c_j^2 \right) y'''(t_n) + O(h^4) \tag{2.31}$$

Volvendo a (2.26) obtemos

$$T_{n+1} = h\sum_{j=1}^{q} b_j \left[D_y f(t_{n,j}, y(t_{n,j}))(y(t_{n,j}) - z_{n,j}) + O(\|y(t_{n,j}), z_{n,j}\|^2) \right] =$$
$$h\sum_{j=1}^{q} b_j g(t_{n,j})(y(t_{n,j}) - z_{n,j}) + O(h^5)$$

donde a última igualdade se debe a (2.30). Utilizando agora a fórmula de Taylor

$$g(t_{n,j}) = g(t_n + c_j h) = g(t_n) + hc_j g'(t_n) + O(h^2)$$

temos que

$$T_{n+1} = h\sum_{j=1}^{q} b_j g(t_n)(y(t_{n,j}) - z_{n,j}) + h^2 \sum_{j=1}^{q} b_j c_j g'(t_n)(y(t_{n,j}) - z_{n,j}) + O(h^5).$$

Utilizando (2.31) na primeira suma e (2.30) na segunda deducimos que

$$T_{n+1} = \frac{h^3}{2} g(t_n) y''(t_n)(b^T C^2 e - 2b^T ACe) + \frac{h^4}{2}(g(t_n))^2 y''(t_n)(b^T AC^2 e - 2b^T A^2 Ce) +$$

$$\frac{h^4}{6} g(t_n) y'''(t_n)(b^T C^3 e - 3b^T AC^2 e) + \frac{h^4}{2} g'(t_n) y''(t_n)(b^T C^3 e - 2b^T CACe) + O(h^5).$$

Por tanto, para que o método sexa polo menos de orde 4, é suficiente que

$$\begin{cases} b^T C^2 e = 2b^T ACe,\ b^T AC^2 e = 2b^T A^2 Ce \\ b^T C^3 e = 3b^T AC^2 e,\ b^T C^3 e = 2b^T CACe \end{cases} \tag{2.32}$$

Vexamos que estas condicións son necesarias.
Tomando $f(t,y) = y + 2t - t^2$ e $y(t) = t^2$, temos que $g(t) = 1$, polo que

$$T_{n+1} = h^3(b^T C^2 e - 2b^T ACe) + h^4(b^T AC^2 e - 2b^T A^2 Ce) + O(h^5);$$

e para que o método sexa polo menos de orde 4 debemos ter

$$b^T C^2 e = 2b^T ACe \text{ e } b^T AC^2 e = 2b^T A^2 Ce. \tag{2.33}$$

Supoñamos que estas condicións se cumpren.
Tomando $f(t,y) = y + 3t^2 - t^3$ e $y(t) = t^3$, temos $g(t) = 1$, polo que

$$T_{n+1} = h^4(b^T C^3 e - 3b^T AC^2 e) + O(h^5);$$

para que o método sexa polo menos de orde 4 é necesario, polo tanto, ter ademais

$$b^T C^3 e = 3b^T AC^2 e.$$

Supoñamos que (2.33) e esta condición se cumpren.
Tomando agora $f(t,y) = ty + 2t - t^3$ e $y(t) = t^2$, de onde $g'(t) = 1$, polo que

$$T_{n+1} = h^4(b^T C^3 e - 2b^T CACe) + O(h^5)$$

entón, para que o método sexa de polo menos orde 4, debemos ter

$$b^T C^3 e = 2b^T CACe.$$

Todas as condicións (2.32) son por tanto necesarias para que o método sexa de orde 4.
Concluímos que as condicións (2.25) e (2.32) son condicións necesarias e suficientes para que o método sexa de polo menos orde 4, co que queda probado o lema. □

Exemplo 2.23. Consideremos o método Runge-Kutta clásico visto no exemplo 2.4. Da súa táboa de Butcher obtemos que

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, b^T = \begin{pmatrix} \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{pmatrix}, c^T = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} \text{ e } C = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

É fácil comprobar que se cumpre a condición de filas e as condicións $b^T e = 1$, $b^T Ce = \frac{1}{2}$, $b^T C^2 e = \frac{1}{3}$, $b^T ACe = \frac{1}{6}$, $b^T C^3 e = \frac{1}{4}$, $b^T AC^2 e = \frac{1}{12}$, $b^T A^2 Ce = \frac{1}{24}$ e $b^T CACe = \frac{1}{8}$. Polo tanto, en virtude do teorema 2.22 podemos dicir que o método é polo menos de orde 4.

Capítulo 3

Estabilidade numérica

Neste capítulo imos abordar algúns aspectos teóricos relativos a estabilidade numérica dos métodos Runge-Kutta. Esencialmente, o estudo da estabilidade numérica dun método consiste en analizar se o método xera solucións numéricas acotadas cando se aplica a problemas con solución exacta acotada.

Limitámonos a estudar problemas escalares da forma $y' = \lambda y$, con $\lambda \in \mathbb{C}$, e problemas vectoriais da forma $y' = By$, con $B \in M_{m\times m}(\mathbb{R})$ diagonalizable. Isto é o que se coñece como o estudo da estabilidade numérica lineal.

Para maior claridade imos considerar os métodos de Euler explícito e implícito antes de abordar os métodos Runge-Kutta.

As referencias usadas ao longo deste capítulo foron [4] e [1].

3.1. Introdución a estabilidade numérica lineal

Consideramos o problema de valor inicial escalar

$$\begin{cases} y'(t) = \lambda y, \text{ con } \lambda \in \mathbb{C}, \\ y(0) = \eta \in \mathbb{C}, \end{cases} \tag{3.1a}$$

onde (3.1a) é a denominada ecuación de Dahlquist. O problema (3.1) ten solución exacta

$$y(t) = \eta e^{\lambda t}, \tag{3.2}$$

por tanto,

$$|y(t)| = |\eta| e^{(Re\lambda)t}. \tag{3.3}$$

Supoñamos que $\eta \neq 0$, en caso contrario estaríamos ante un caso trivial. Imos distinguir tres casos:

1. Se $Re\lambda > 0$ temos que $|y(t)| \nearrow +\infty$ exponencialmente cando $t \to +\infty$. Polo tanto, estamos ante un problema inestable.

2. Se $Re\lambda = 0$ entón $|y(t)| = \eta$ para todo $t \in [0, \infty]$. Neste caso, podemos dicir que o problema é estable.

3. Se $Re\lambda < 0$, entón $|y(t)| \searrow 0$ exponencialmente cando $t \to +\infty$. Aquí, diremos que existe estabilidade asintótica.

Obsérvese que se $Re\lambda \leq 0$ a solución (3.2) está acotada en $[0, \infty)$. O que queremos, e que ao aplicar un método numérico a EDO (3.1a), a solución numérica $\{y_n\}_{n=0}^{\infty}$ esté acotada.

3.1.1. Método de Euler explícito

Ao aplicar o método de Euler Explícito

$$y_{n+1} = y_n + hf(t_n, y_n)$$

á EDO (3.1a) obtemos que

$$y_{n+1} = y_n + h\lambda y_n = (1 + h\lambda)y_n, \ n \geq 0. \tag{3.4}$$

A expresión explícita de y_n é

$$y_n = (1 + h\lambda)^n y_0, \ \forall n \geq 0,$$

e en consecuencia

$$|y_n| = |1 + h\lambda|^n |y_0|.$$

Supoñamos que $y_0 \neq 0$, en caso contrario estaríamos ante un caso trivial. Entón, a solución numérica $\{y_n\}_{n=0}^{\infty}$ está acotada se, e só se, $|1 + h\lambda| \leq 1$.

En xeral, ao aplicar un método dun paso a ecuación (3.1a), obtense unha relación da forma

$$y_{n+1} = R(h\lambda)y_n, \ h \geq 0, \tag{3.5}$$

onde $R = R(z)$ é unha función da variable complexa $z = h\lambda$, que depende do método concreto e que se denomina *función de estabilidade* (ou *factor de amplificación*) do método.

De (3.5) dedúcese que

$$y_n = (R(h\lambda))^n\, y_0,\ n \geq 0.$$

Por tanto, supoñendo $y_0 \neq 0$, a solución numérica $\{y_n\}_{n=0}^{\infty}$ está acotada se, e só se, $h\lambda \in \mathscr{A}$, onde

$$\mathscr{A} = \{z \in \mathbb{C} / |R(z)| \leq 1\}.$$

O conxunto $\mathscr{A}$ chámase *rexión de estabilidade absoluta* do método.

Nótese que de (3.4) dedúcese que a función de estabilidade do método de Euler explícito, R_{EE}, está dada por

$$R_{EE}(z) = 1 + z.$$

A rexión de estabilidade absoluta do método de Euler explícito é o conxunto

$$\mathscr{A}_{EE} = \{z \in \mathbb{C} / |1 + z| \leq 1\} = \overline{D(-1, 1)}$$

.

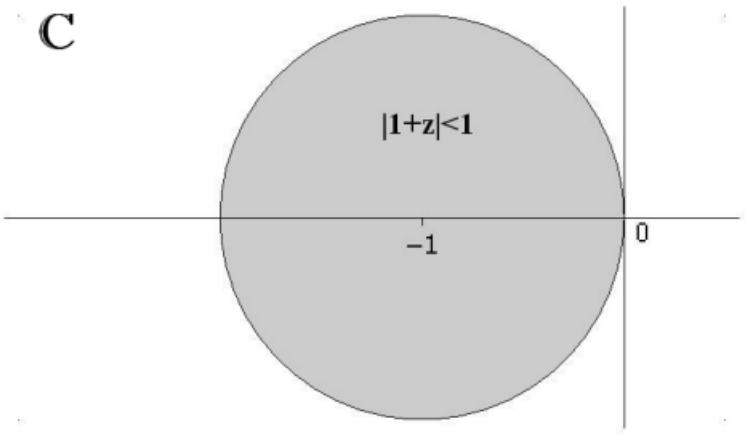

Figura 3.1: Rexión de estabilidade absoluta para o método de Euler explícito. Recuperado de `http://wmatem.eis.uva.es/ jesroj/matem1/Curso/Cap08b_Esquema.pdf`

Vexamos agora uns casos particulares, para ver para que valores de h o método de Euler explícito xera solucións numéricas acotadas.

CASO 1: $\lambda \in \mathbb{R}$, $\lambda < 0$.

Consideremos o intervalo de estabilidade absoluta $I = \mathscr{A} \cap \mathbb{R}$. A solución numérica $\{y_n\}_{n=0}^{\infty}$ estará acotada se, e só se, $h\lambda \in I$, o que equivale a que $-2 \leq h\lambda \leq 0$.

A segunda desigualdade é certa, xa que $h > 0$ e $\lambda < 0$. Impoñamos a primeira

$$-2 \leq h\lambda \Leftrightarrow 2 \geq -h\lambda = h|\lambda| \Leftrightarrow h \leq \frac{2}{|\lambda|}.$$

Por tanto, $\{y_n\}_{n=0}^{\infty}$ está acotada se, e só se, $h \leq \frac{2}{|\lambda|}$. Esta restrición obriga a tomar h pequeno cando λ é grande.

CASO 2: $\lambda \in \mathbb{C}$ con $Re\lambda < 0$.

Neste caso tense que

$$|1 + h\lambda| \leq 1 \Leftrightarrow (1 + hRe\lambda)^2 + (hIm\lambda)^2 \leq 1 \Leftrightarrow 2hRe\lambda + h^2|\lambda|^2 \leq 0 \Leftrightarrow h \leq \frac{-2Re\lambda}{|\lambda|^2}.$$

Esta restrición pode obrigar a tomar h pequeno, por exemplo, se a parte imaxinaria de λ é grande.

Observemos que hai que restrinxirse a h pequeno para que haxa acotación da solución numérica.

3.1.2. Método de Euler Implícito

Ao aplicar o método de Euler implícito

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}),$$

á EDO (3.1a) resulta

$$y_{n+1} = y_n + h\lambda y_{n+1},\ n \geq 0$$

e polo tanto,

$$y_{n+1} = \frac{1}{1 - h\lambda} y_n.$$

Entón, a función de estabilidade do método de Euler implícito é

$$R_{EI}(z) = \frac{1}{1 - z}.$$

A solución numérica $\{y_n\}_{n=0}^{\infty}$ estará acotada se, e só se, $z \in \mathscr{A}_{EI}$ onde

$$\mathscr{A}_{EI} = \{z \in \mathbb{C} / |R_{EI}(z)| \leq 1\} = \left\{ z \in \mathbb{C} / |\frac{1}{1+z}| \leq 1 \right\} = \{z \in \mathbb{C} / 1 \leq |1 - z|\} = \mathbb{C} \setminus D(1, 1)$$

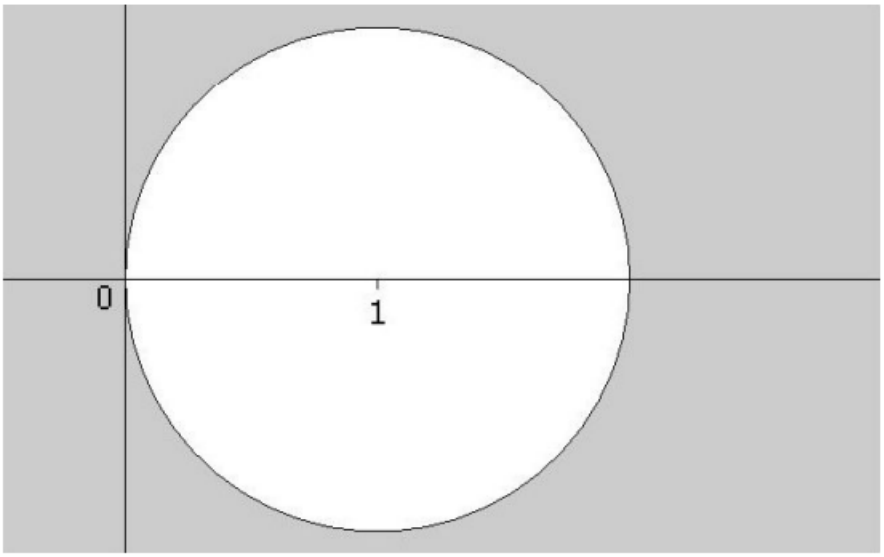

Figura 3.2: Rexión de estabilidade absoluta para o método de Euler implícito. Recuperado de `http://wmatem.eis.uva.es/ jesroj/matem1/Curso/Cap08b_Esquema.pdf`

Definición 3.1. Un método numérico é $\mathscr{A}$-*estable* se a súa rexión de estabilidade absoluta contén todo o semiplano esquerdo do plano complexo, é dicir, se

$$C^- := \{z \in \mathbb{C}/Rez \leq 0\} \subset \mathscr{A}.$$

Dado que $C^- \subset \mathscr{A}_{EI}$ o método de Euler implícito é $\mathscr{A}$-estable.

Se un método é $\mathscr{A}$-estable, entón para todo $\lambda \in \mathbb{C}$ con $Re\lambda \leq 0$ tense que $h\lambda \in \mathscr{A}$ para calquera $h > 0$. Polo tanto, un método $\mathscr{A}$-estable aplicado á ecuación (3.1a) con $Re\lambda \leq 0$, xera sempre solucións numéricas acotadas, sen necesidade de que o paso h sexa pequeno.

Isto sucede, en particular, para o método de Euler implícito.

3.2. Sistemas lineais

3.2.1. Problema continuo

Consideramos o sistema lineal

$$\begin{cases} y'(t) = By, & \text{(3.6a)} \\ y(0) = \eta \in \mathbb{R}^m, & \text{(3.6b)} \end{cases}$$

con $B \in M_{m\times m}(\mathbb{R})$ diagonalizable. Por tanto, existe $P \in M_{m\times m}(\mathbb{C})$ non singular e existe unha matriz diagonal $\Lambda = diag(\lambda_1, ..., \lambda_n)$ tal que $P^{-1}BP = \Lambda$, ou equivalentemente, $B = P\Lambda P^{-1}$.

Entón, podemos escribir (3.6a) como

$$y' = P\Lambda P^{-1}y,$$

ou, equivalentemente,

$$P^{-1}y' = \Lambda P^{-1}y.$$

Como P é constante, é dicir, non depende do tempo

$$\left(P^{-1}y\right)' = \Lambda P^{-1}y.$$

Se consideramos $w = P^{-1}y$, podemos escribir o sistema lineal

$$\begin{cases} w' = \Lambda w \\ w(0) = P^{-1}\eta \end{cases} \tag{3.7}$$

e dado que Λ é unha matriz diagonal, chegamos a m problemas escalares que involucran á ecuación de Dahlquist:

$$\begin{cases} w_i' = \lambda_i w_i \\ w_i(0) = \left[P^{-1}\eta\right]_i \end{cases} \tag{3.8}$$

onde $1 \leq i \leq m$.

Polo tanto, $y(t)$ estará acotada en $[0, \infty)$ se, e só se, $w(t)$ está acotada en $[0, \infty)$, o que equivale a que $w_i(t)$ está acotada para todo $i = 1, \dots, m$.

En consecuencia, $y(t)$ estará acotada en $[0, \infty)$ para toda condición inicial $\eta \in \mathbb{R}^m$ se, e só se, $Re(\lambda_i) \leq 0$ para todo $1 \leq i \leq m$.

3.2.2. Problema discreto

En primeiro lugar, queremos probar é que existe conmutatividade entre aplicarlle o método de Euler explícito ao problema diagonalizado (3.7) e aplicar primeiro dito método ao problema (3.6) e a continuación diagonalizar o esquema numérico.

Cabe destacar, que aínda que só o imos facer para o método de Euler explícito, sucedería o mesmo no caso de Euler implícito, sendo o proceso análogo.

Se a (3.6) lle aplicamos o método de Euler explícito obtemos

$$\begin{cases} \frac{y_{n+1}-y_n}{h} = By_n, \\ y_0 = \eta \in \mathbb{R}^m, \end{cases} \tag{3.9}$$

Agora, diagonalizamos (3.9) introducindo os vectores $w_n = P^{-1}y_n$ e obtemos

$$\begin{cases} \frac{w_{n+1}-w_n}{h} = \Lambda w_n, \\ w_0 = P^{-1}\eta, \end{cases} \tag{3.10}$$

que coincide co resultado de aplicar a (3.7) o método de Euler explícito.

Se desacoplamos (3.10) obtemos

$$\begin{cases} \frac{(w_{n+1})_i-(w_n)_i}{h} = \lambda_i (w_n)_i, \\ (w_0)_i = [P^{-1}\eta]_i, \end{cases} \tag{3.11}$$

para todo $i = 1, \dots, n$. Isto tamén se pode obter aplicándolle a (3.8) o método de Euler explícito.

3.3. Métodos Runge-Kutta

Imos aplicar á ecuación de Dahlquist (3.1a) o método Runge-Kutta xeral (2.2) (que se pode escribir de forma equivalente como (2.3)) para obter

$$y_{n+1} = R_{RK}(h\lambda) y_n.$$

Vexamos isto, antes de comezar, para un método Runge-Kutta concreto.

Exemplo 3.2. Consideremos o método de Euler modificado, cuxa táboa de Butcher é

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

Temos que

$y_{n,1} = y_n$

$y_{n,2} = y_n + \frac{1}{2} h f(t_{n,1}, y_{n,1})$

$y_{n+1} = y_n + h f(t_{n,2}, y_{n,2})$

Comparando (1.1) e (3.1) temos que $f(t, y) = \lambda y$ polo que nos queda

$y_{n,1} = y_n$

$y_{n,2} = y_n + \frac{1}{2} h\lambda y_{n,1} = \left(1 + \frac{1}{2} h\lambda\right) y_n$

$y_{n+1} = y_n + h\lambda y_{n,2} = y_n + h\lambda \left(1 + \frac{1}{2} h\lambda\right) y_n = \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right) y_n$

Polo tanto, a función de estabilidade do método de Euler modificado é

$$R_{EM}(z) = 1 + z + \frac{z^2}{2}.$$

Abordamos agora a obtención da función de estabilidade para o método Runge-Kutta xeral (3.1). Aplicando o método á ecuación (3.1a) obtemos

$$\begin{cases} y_{n+1} = y_n + h\sum_{i=1}^{q} b_i \lambda y_{n,i} \\ \text{onde} \\ y_{n,i} = y_n + h\sum_{i=1}^{q} a_{ij}\lambda y_{n,j},\ i = 1, \dots, q \end{cases} \tag{3.12}$$

Poñendo $z = h\lambda \in \mathbb{C}$, podemos escribir (3.15) como

$$\begin{cases} y_{n+1} = y_n + z\sum_{i=1}^{q} b_i y_{n,i} \\ \text{onde} \\ y_{n,i} = y_n + z\sum_{j=1}^{q} a_{ij} y_{n,j},\ i = 1, \dots, q \end{cases} \tag{3.13}$$

Introducindo $Y_n = (y_{n,1}, \dots, y_{n,q})^T \in \mathbb{R}^q$ e recordando a notación $e = (1, \dots, 1)^T \in \mathbb{R}^q$, podemos escribir (3.16) da forma

$$\begin{cases} y_{n+1} = y_n + zb^T Y_n \\ \text{onde } Y_n \text{ satisfai} \\ Y_n = y_n e + zAY_n \end{cases} \tag{3.14}$$

de onde se obtén, dado que $Y_n - zAY_n = y_n e$, é dicir, $Y_n = (I - zA)^{-1} y_n e$ que

$$y_{n+1} = y_n \left[1 + zb^T (I - zA)^{-1} e\right], \tag{3.15}$$

onde I é a matriz identidade de dimensión $q \times q$. A función de estabilidade está dada polo tanto por

$$R_{RK}(z) = 1 + zb^T (I - zA)^{-1} e.$$

Mostraremos agora unha forma alternativa da función de estabilidade. Imos desenvolver a súa dedución para o caso $q = 2$, pero pode extenderse ao caso xeral de calquera valor de q.

Escribimos (3.16) como

$$\begin{pmatrix} 1 - za_{11} & -za_{12} & 0 \\ -za_{12} & 1 - za_{22} & 0 \\ -zb_1 & -zb_2 & 1 \end{pmatrix} \begin{pmatrix} y_{n,1} \\ y_{n,2} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} y_n \\ y_n \\ y_n \end{pmatrix}$$

Imos calcular a solución para y_{n+1} aplicando a regra de Cramer. Consideramos para iso

$$N = det\begin{pmatrix} 1-za_{11} & -za_{12} & y_n \\ -za_{21} & 1-za_{12} & y_n \\ -zb_1 & -zb_2 & y_n \end{pmatrix},\ D = det\begin{pmatrix} 1-za_{11} & -za_{12} & 0 \\ -za_{12} & 1-za_{22} & 0 \\ -zb_1 & -zb_2 & 1 \end{pmatrix}$$

Facendo contas elementais sobre N chégase a que

$$N = det\begin{pmatrix} 1-za_{11}+zb_1 & -za_{12}+zb_2 & 0 \\ -za_{12}+zb_1 & 1-za_{22}+zb_b & 0 \\ -zb_1 & -zb_2 & y_n \end{pmatrix}.$$

Entón, podemos escribir

$$D = det(I - zA) \text{ e } N = det(I - zA + zeb^T)y_n$$

Por tanto, pola regra de Cramer

$$y_{n+1} = \frac{N}{D} = \frac{det(I - zA + zeb^T)}{det(I - zA)}y_n.$$

Obtemos así, que a función de estabilidade do método Runge-Kutta xeral (2.1) é

$$R_{RK}(z) = \frac{det(I - zA + zeb^T)}{det(I - zA)}. \tag{3.16}$$

No caso dun método Runge-Kutta explícito de q etapas, por ser A matriz triangular inferior, temos por unha parte que $det(I - zA) = 1$. Por outra parte, tense que $det(I - zA + zeb^T)$ non é constante e é un polinomio de grado $\leq q$.

En consecuencia,

$$\lim_{\substack{|z|\to+\infty \\ z\in\mathbb{C}}} |R_{ERK}(z)| = +\infty,$$

onde R_{ERK} denota a función de estabilidade.

Entón, para $M > 0$ suficientemente grande, tense que

$$|z| > M \Rightarrow |R_{ERK}(z)| > 1,$$

é dicir, $\mathbb{C} \setminus \overline{D(0,M)} \subset \mathbb{C} \setminus \mathscr{A}_{ERK}$, onde $\mathscr{A}_{ERK}$ denota a rexión de estabilidade do método Runge-Kutta explícito. Posto que para todo $M > 0$, $\mathbb{C}^- \cap \left(\mathbb{C} \setminus \overline{D(0,M)}\right) \neq \emptyset$, temos que $\mathbb{C}^- \cap (\mathbb{C} \setminus \mathscr{A}_{ERK}) \neq \emptyset$, é dicir, $\mathbb{C}^- \not\subset \mathscr{A}_{ERK}$. Probamos así que ningún método Runge-Kutta explícito é $\mathscr{A}$-estable.

Consideremos agora un método Runge-Kutta explícito de orde p. Recordemos que entón

$$T_n = O(h^{p+1}).$$

Ademais pola observación 1.1, baixo a hipótese de localización $y(t_{n-1}) = y_{n-1}$ tense que

$$T_n = y(t_n) - \tilde{y}_n. \tag{3.17}$$

Entón, considerando (3.1) e pola definición de función de estabilidade

$$\tilde{y}_n = (R_{ERK}(h\lambda))\, y_{n-1}. \tag{3.18}$$

Dado que

$$y(t_n) = \eta e^{\lambda t_n} = \eta e^{\lambda(t_{n-1}+h)} = \eta e^{\lambda h} e^{\lambda t_{n-1}} = e^{\lambda h} y(t_{n-1}), \tag{3.19}$$

substituíndo (3.18) e (3.19) en (3.17)

$$T_n = e^{h\lambda} y(t_{n-1}) - R_{ERK}(h\lambda) y(t_{n-1}) = \left(e^{h\lambda} - R_{ERK}(h\lambda)\right) y(t_{n-1}) = O(h^{p+1}).$$

Entón, como $y(t_{n-1})$ está fixo, tense que para todo $\lambda \in \mathbb{C}$,

$$e^{h\lambda} - R_{ERK}(h\lambda) = O(h^{p+1}). \tag{3.20}$$

Como $R_{ERK}(z)$ é un cociente de dous polinomios temos unha función racional ben definida nun entorno de cero xa que $z = 0$ non é raíz do polinomio do denominador. Polo tanto $e^z - R_{ERK}(z)$ é unha función holomorfa nun entorno de cero e facendo un desenvolvemento de Taylor

$$e^z - R_{ERK}(z) = a_0 + a_1 z + a_2 z^2 + ... + a_p z^p + O(z^{p+1}). \tag{3.21}$$

Se tomamos $\lambda \in \mathbb{C}$ fixo, $z = \lambda h$ e comparamos (3.20) e (3.21) obtemos que $a_0 = 0$, $a_i \lambda^i = 0$ para todo $i = 1, ..., p$. E considerando $\lambda \neq 0$ necesariamente $a_0 = a_1 = ... = a_p = 0$.

Así, obtemos que

$$e^z - R_{ERK}(z) = O(z^{p+1})$$

e en consecuencia

$$R_{ERK}(z) = e^z + O(z^{p+1}) \tag{3.22}$$

Finalmente, se consideramos un método Runge-Kutta explícito de q etapas e orde q a súa función de estabilidade, dada por (3.16), vai quedar polinómica de grado q. Entón,

$$R_{ERK}(z) = e^z + O(|z|^{q+1}),$$

e necesariamente

$$R_{ERK}(z) = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^q}{q!} \tag{3.23}$$

que é o polinomio de Taylor de grado q da función e^z en torno á orixe.

Sábese que só existen métodos Runge-Kutta explícitos con orde igual a q, ou número de etapas, para $1 \leq q \leq 4$ (véxase, por exemplo, [4]). Así pois, fixado $q \in \{1, 2, 3, 4\}$, todos os métodos Runge-Kutta explícitos de q etapas con orde q teñen a mesma función de estabilidade, que é a dada por (3.23) e, por tanto, a mesma rexión de estabilidade absoluta.

Exemplo 3.3. Imos considerar o método DIRK xa visto no exemplo 2.6. Aplicando a formulación en $y_{n,i}$ do método á EDO (3.1a), obtemos o seguinte:
en primeiro lugar,

$$y_{n,1} = y_n + \frac{1}{4}h\lambda y_{n,1},$$

por tanto

$$y_{n,1} = \frac{4}{4 - h\lambda} y_n.$$

Seguidamente,

$$y_{n,2} = y_n + \frac{1}{2}h\lambda y_{n,1} + \frac{1}{4}h\lambda y_{n,2};$$

facendo contas e substituíndo $y_{n,1}$, obtemos que

$$y_{n,2} = \frac{16 + 4h\lambda}{(4 - h\lambda)^2} y_n.$$

Por último, temos que

$$y_{n+1} = y_n + \frac{1}{2}h\lambda \left(y_{n,1} + y_{n,2}\right),$$

se substituímos nesta ecuación $y_{n,1}$ e $y_{n,2}$ chégase a que

$$y_{n+1} = \left(\frac{4 + h\lambda}{4 - h\lambda}\right)^2 y_n.$$

Por tanto, a función de estabilidade absoluta do método DIRK é

$$R(z) = \left(\frac{4 + z}{4 - z}\right).$$

Calculemos agora a rexión de estabilidade absoluta.

$$|R(z)| \leq 1 \Leftrightarrow |\frac{4 + z}{4 - z}|^2 \leq 1$$

Considerando $z = a + bi$ e facendo unha serie de cálculos obtense que a desigualdade anterior equivale a que $a \leq 0$. Polo tanto, a rexión de estabilidade absoluta do método é

$$\mathscr{A}(z) = \{z \in \mathbb{C} / Re(z) \leq 0\}.$$

Como $C^- \subset \mathscr{A}(z)$ podemos dicir que o método DIRK é $\mathscr{A}$-estable.

3.3.1. Aplicación dos métodos Runge-Kutta a un sistema lineal

Retomamos a notación e as hipóteses da sección 3.2.

Imos probar agora que se aplicamos un método Runge-Kutta ao problema (3.6) e despois diagonalizamos o esquema numérico, obtemos o esquema que resulta de aplicar o método Runge-Kutta ao problema (3.7). Consideremos, en primeiro lugar, o método Runge-Kutta xeral (2.3). Aplicándolle o método á ecuación (3.6) resulta

$$\begin{cases} y_{n+1} = y_n + h\sum_{i=1}^{q} b_i k_{n,i} \\ \text{onde} \\ k_{n,i} = B\left(y_n + h\sum_{j=1}^{q} a_{ij}k_{n,j}\right),\ i = 1,\dots,q \end{cases} \tag{3.24}$$

Definimos w_n e $l_{n,i}$ mediante

$$y_n = Pw_n \text{ e } k_{n,i} = Pl_{n,i},\ i = 1,\dots,q.$$

Substituíndo y_n e $k_{n,i}$ en (3.24) e multiplicando por P^{-1} temos

$$\begin{cases} w_{n+1} = w_n + h\sum_{i=1}^{q} b_i l_{n,i} \\ \text{onde} \\ l_{n,i} = \Lambda\left[w_n + h\sum_{j=1}^{q} a_{ij}l_{n,j}\right],\ i = 1,\dots,q \end{cases} \tag{3.25}$$

o cal é exactamente o resultado que obteríamos ao aplicar o método (2.3) ao sistema (3.7). En consecuencia, o estudo da estabilidade numérica dun método Runge-Kutta aplicado ao sistema (3.6) redúcese ao da estabilidade numérica de dito método aplicando a cada unha das EDOs escalares (3.8) (ao igual que vimos no caso do método de Euler explícito).

Anexo A

Algúns resultados auxiliares

Lema A.1. *Sexa $\|.\|$ unha norma en $\mathbb{R}^m$. Denótase tamén con $\|.\|$ a correspondente norma subordinada nas matrices $m \times m$. Sexa $\|.\|_\infty$ a norma infinito producto en $(\mathbb{R}^m)^q$. Denótase tamén por $\|.\|_\infty$ a correspondente norma subordinada en $M_{(mq)\times(mq)}(\mathbb{R})$.*
Sexa $B \in M_{(mq)\times(mq)}(\mathbb{R})$ da forma

$$B = \begin{pmatrix} B_{11} & \cdots & B_{1q} \\ \vdots & \ddots & \vdots \\ B_{q1} & \cdots & B_{qq} \end{pmatrix}$$

onde $B_{ij} \in M_{m\times m}(\mathbb{R})$ para todo $i, j = 1, \ldots, q$. Tense que

$$\|B\|_\infty \leq \max_{1\leq i\leq q} \sum_{j=1}^{q} \|B_{ij}\|. \tag{A.1}$$

Demostración. Por definición de norma subordinada

$$\|B\|_\infty = \sup_{\substack{Y\in(\mathbb{R}^m)^q \\ \|Y\|_\infty=1}} \|BY\|_\infty$$

Denotemos $Y \in (\mathbb{R}^m)^q$ por $Y = (Y_1, \ldots, Y_q)^T$ onde $Y_j \in \mathbb{R}^m$ para todo $j = 1, \ldots, q$. Entón, podemos escribir, para todo $Y \in (R^m)^q$ con $\|Y\|_\infty = 1$

$$\|BY\|_\infty \leq \max_{1\leq i\leq q} \left(\sum_{j=1}^{q} \|B_{ij}Y_j\| \right) \leq \max_{1\leq i\leq q} \sum_{j=1}^{q} \|B_{ij}\|\|Y_j\| = \left(\max_{1\leq i\leq q} \sum_{j=1}^{q} \|B_{ij}\| \right) \left(\max_{1\leq i\leq q} \|Y_j\| \right) =$$

$$\left(\max_{1\leq i\leq q} \sum_{j=1}^{q} \|B_{ij}\| \right) \|Y\|_\infty = \max_{1\leq i\leq q} \sum_{j=1}^{q} \|B_{ij}\|,$$

chegamos así a que se verifica (A.1). □

Lema A.2. *Sexa $\|.\|$ unha norma en $\mathbb{R}^m$. Denotamos como $\|.\|$ a correspondente norma subordinada nas matrices.*
Sexa $\Psi : \mathbb{R}^m \to \mathbb{R}^m$ unha función diferenciable que satisfai a condición de Lipschitz

$$\|\Psi(y) - \Psi(y^*)\| \leq L\|y - y^*\|, \ \forall y, y^* \in \mathbb{R}^m; \tag{A.2}$$

entón $\|D\Psi(y)\| \leq L$ para todo $y \in \mathbb{R}^m$.

Demostración. Sexa $y \in \mathbb{R}^m$, entón

$$\|D\Psi(y)\| = \sup_{\substack{e \in \mathbb{R}^m \\ \|e\|_\infty = 1}} \|D\Psi(y)e\|.$$

Tense que,

$$\|D\Psi(y)e\| = \|\lim_{h \to 0} \frac{\Psi(y + he) - \Psi(y)}{h}\| = \lim_{h \to 0} \|\frac{\Psi(y + he) - \Psi(y)}{h}\|.$$

Por (A.2)

$$\lim_{h \to 0} \|\frac{\Psi(y + he) - \Psi(y)}{h}\| \leq \frac{L|h|}{|h|} = L.$$

En consecuencia,

$$\|D\Psi(y)\| = \sup_{\substack{e \in \mathbb{R}^m \\ \|e\|_\infty = 1}} \|D\Psi(y)e\| \leq L.$$

□

Bibliografía

[1] U.M. Ascher, L.R. Petzold (1997). *Computer methods for ordinary differential equations and differential-algebraic equations*, Philadelphia (Pennsylvania).

[2] M. Crouzeix, A.L. Mignot (1989). *Analyse Numérique des Équations Différentielles*, Masson.

[3] M. Crouzeix, A.L. Mignot (1992). *Analyse numérique des équations différentielles*, 2eme. éd. révisée et augmentée, 2e tirage, Springer-Verlag, Paris, Masson.

[4] J.D. Lambert (1991). *Numerical Methods for Ordinary Differential Systems, The Initial Value Problems*, Chichester : John Wiley and Sons.