

**Métodos Numéricos
para las
Ecuaciones Diferenciales.
Apuntes y Ejercicios**

Eliseo Chacón Vera
Departamento de Matemáticas,
Facultad de Matemáticas, Universidad de Murcia

30 de noviembre de 2022

Índice general

1. Introducción a las Ecuaciones Diferenciales	7
1.1. Introducción	7
1.2. Ecuaciones Diferenciales Ordinarias	9
1.3. Problema de Cauchy o de valor inicial	9
1.4. Existencia y unicidad de solución	15
1.5. Estudio cualitativo de las soluciones	19
1.6. Ejemplos (I): Problemas de primer orden	21
1.7. Ejemplos (II): Problemas de orden mayor que uno	28
1.8. Ejemplos (III): Problemas rígidos e inestables	32
1.8.1. Ejemplos en sistemas	33
1.8.2. Ejemplos escalares	37
1.8.3. Problemas inestables	43
1.9. Ejemplos (IV): Problemas oscilatorios	47
1.10. Sistemas, linealización y uso de variable compleja	50
1.10.1. Triangulación de matrices	51
1.11. Conclusión sobre los ejemplos y objetivos	55
1.12. Ejercicios	56
2. Método de Euler explícito	61
2.1. Introducción	61
2.2. Discretización y primeros conceptos	61
2.3. Método de Euler progresivo o explícito	64
2.4. Estimación del orden de convergencia a cero	68
2.4.1. Uso de tablas	70
2.4.2. Recta de pendiente	71
2.4.3. Cuando la solución verdadera no se conoce	72
2.5. Convergencia de Euler explícito	73
2.5.1. Consistencia del esquema numérico	76
2.5.2. Estabilidad del esquema numérico	81
2.5.3. Estimación de error usando la gráfica	84
2.5.4. Estimación de error usando Taylor	87
2.5.5. Observaciones sobre la convergencia	91

2.6. Interpretación de la cota de error	94
2.7. Ejemplos	95
2.7.1. Resumiendo	110
2.8. Problemas rígidos	111
2.9. Euler explícito en sistemas	115
2.9.1. Consistencia	117
2.9.2. Estabilidad	119
2.9.3. Convergencia	121
2.9.4. Estudio del error (<i>¿Si el tiempo lo permite?</i>) [*]	122
2.10. Ejercicios	133
3. Métodos de Euler implícito y de Crank-Nicolson	143
3.1. Mejora en la estabilidad: Euler implícito	144
3.1.1. Error local para Euler implícito	150
3.2. Error global para Euler implícito	151
3.3. Mejora en el orden: Método de Crank-Nicolson	155
3.3.1. Error local para Crank-Nicolson	157
3.4. Error global para Crank-Nicolson	159
3.5. Ejercicios	160
4. Métodos de Runge-Kutta	165
4.1. Introducción	165
4.2. Métodos de Runge-Kutta explícitos	168
4.3. Runge-Kutta implícitos (Si tiempo permite)	169
4.4. Forma general y tablero de Butcher	171
4.5. Estudio del error local	174
4.6. Estudio de la estabilidad: 0-estabilidad	175
4.7. Convergencia de los métodos RKE	177
4.8. Familias RKE de acuerdo al orden	180
4.8.1. Métodos RKE con $S = 1$	180
4.8.2. Métodos RKE con $S = 2$	180
4.8.3. Métodos RKE con $S = 3$	184
4.8.4. Métodos RKE con $S = 4$	187
4.8.5. Resumen	188
4.9. Extensión a sistemas	189
4.9.1. Sobre el orden de convergencia	190
4.10. Estabilidad absoluta: A-estabilidad	192
4.10.1. Efecto de la linealización	195
4.10.2. Factor de amplificación de un esquema	198
4.11. Estabilidad vs Precisión	203
4.11.1. Ejemplos	204
4.12. Control del paso	206

4.12.1. Uso de paso doble	207
4.12.2. Runge-Kutta adaptativos: pares encajados	210
4.12.3. Runge-Kutta-Fehlberg 4(5)	214
4.12.4. Runge-Kutta Dormand-Prince 5(4)	217
4.13. Otras ideas para mejorar precisión	218
4.13.1. Métodos de Taylor	218
4.14. Ejercicios	221
5. Métodos Multipaso	227
5.1. Introducción	227
5.2. Ejemplos	230
5.2.1. Interpolar pendientes $f_j = f(t_j, y_j)$ previas	230
5.2.2. Métodos BDF (Backward Differentiation formulas): interpolando las aproximaciones y_j	232
5.3. Forma general de los métodos multipaso	237
5.3.1. Consecuencias sobre los coeficientes a_j	239
5.3.2. Breve sobre ecuaciones en diferencias	240
5.4. Estudio de la consistencia	247
5.4.1. Cálculo del error local de consistencia	250
5.4.2. Estudio de la estabilidad: 0-estabilidad	253
5.5. Análisis de convergencia	258
5.6. Barreras en el orden de convergencia	259
5.7. Familias de métodos de multipaso	260
5.8. Ejercicios	262
6. Ejercicios computacionales	267
6.1. Práctica Computacional 1	290
6.1.1. Práctica Computacional 1: Soluciones	291
6.2. Práctica Computacional 2	296
6.3. Práctica Computacional 3	299
7. Prácticas computacionales, curso 2021-22	301
7.1. Práctica computacional 1	301
7.2. Práctica computacional 2	312
7.3. Práctica computacional 3	316
7.4. Práctica computacional 4	334
8. Introducción a las Ecuaciones en Derivadas Parciales	337
8.1. Notación	337
8.2. Ecuaciones diferenciales	340
8.3. Leyes de conservación $d = 1$	343
8.3.1. Ejemplos	345

8.4. Leyes de conservación $d > 1$	347
8.5. Condiciones de contorno y dato inicial	349
9. La ecuación de difusión	351
9.1. Decaimiento debido a la difusión	351
9.2. Diferencias finitas	353
9.3. Euler explícito en tiempo y diferencias centradas en espacio (Eex-CE)	355
9.3.1. Uso de matrices	357
9.4. Mejora en estabilidad y orden	357
9.4.1. Euler implícito en tiempo diferencias centradas en espacio (Eim-CE)	359
9.5. Semidiscretización: Método de líneas	360
9.6. Estudio unificado de la estabilidad	364
9.7. Estudio de la consistencia	368
9.7.1. Consistencia para Euler explícito	369
9.7.2. Consistencia para Euler implícito	369
9.7.3. Consistencia para Crank-Nicolson	370
9.8. Estudio unificado de la convergencia	373
9.9. Problemas de contorno estacionario	374
9.9.1. Algunos ejemplos	376
9.10. Ejercicios	379

Capítulo 1

Introducción a las Ecuaciones Diferenciales

Resumen del tema

En este capítulo revisaremos algunos conceptos básicos de la teoría cuantitativa y cualitativa de las ecuaciones diferenciales ordinarias. Veremos ejemplos sencillos que ilustren estos conceptos tanto con ecuaciones de primer, o mayor orden, como con sistemas. Principalmente hay que leer:

- Notación y resultados de teoría básicos, estabilidad de las soluciones con respecto al dato inicial.
- Ejemplos escalares de problemas expansivos, contractivos y oscilatorios.
- Ejemplos básicos de sistemas y ecuaciones de orden superior.

Los ejemplos más complicados se pueden dejar para una segunda lectura. Siendo lo más importante los aspectos cualitativos de los comportamientos de las soluciones asociadas a las ecuaciones.

1.1. Introducción

La mayoría de los modelos matemáticos en cualquier campo científico, técnico, sociológico, etc... estudian la evolución temporal y/o espacial de alguna cantidad de interés. Las derivadas con respecto a las distintas variables de las que depende una magnitud representan la velocidad de cambio de dicha magnitud con respecto a esas variables. Por lo tanto, este tipo de modelos matemáticos se puede describir mediante ecuaciones que involucran derivadas de la función buscada en las variables espaciales y en la variable temporal. Estas ecuaciones se denominan ecuaciones diferenciales.

El ejemplo más cercano nos lo podemos encontrar en **las predicciones meteorológicas** que recibimos todos los días para saber si va a hacer calor, va a llover, etc...es muy frecuente que el meteorólogo nos diga que sus predicciones son las que indican los modelos numéricos que manejan. Estos modelos pronostican la velocidad del aire, la temperatura, la humedad, etc...y su variación tanto temporal como espacial. Otro ejemplo puede ser un modelo matemático que represente la trayectoria y la deformación (dilataciones, contracciones) que sufre una pompa de jabón mientras flota en el aire.

Fácilmente se identifican dos grupos de ecuaciones diferenciales:

- **ecuaciones diferenciales ordinarias:** se estudia sólo la evolución en tiempo de una cantidad
- **ecuaciones diferenciales en derivadas parciales:** se estudia la evolución en tiempo y en espacio de una cantidad

Una de las fuentes más comunes de ecuaciones diferenciales nos las encontramos en la física. Hemos visto que una ecuación diferencial ordinaria nos permite por ejemplo describir la trayectoria de un objeto en términos de una ecuación para la velocidad en la forma

$$r'(t) = f(t, r(t)), \quad r(0) = r_0$$

o para la aceleración en la forma

$$r''(t) = f(t, r(t), r'(t)), \quad r(0) = r_0, \quad r'(0) = v_0.$$

Si ahora permitimos que este objeto se deforme en sus dimensiones espaciales con el transcurso del tiempo entonces las variaciones espaciales también son importantes y pueden influir en la trayectoria final. Nos encontramos con el hecho de que ahora ya no sólo es $r = r(t)$ si no $r = r(t, x, y, z)$ y tenemos que tener en cuenta las variaciones espaciales, esto es las funciones $r_x, r_y, r_z, r_{xx}, r_{yy}, r_{zz}, \dots$

También vamos a introducirnos en el mundo del **modelado matemático** y de la **simulación numérica**. Este campo de la ciencia ha ganado una considerable importancia en las últimas décadas, principalmente se ha revolucionado e impulsado enormemente desde el desarrollo de las capacidades de computación. Existe un ingente esfuerzo en el diseño, análisis y aplicación de técnicas computacionales para obtener soluciones para muchos problemas, entre ellos están los regidos por las ecuaciones diferenciales.

Los avances computacionales forman ya parte fundamental de cualquier aspecto de la vida moderna cotidiana en sus aspectos más dispares como son la predicción meteorológica, la aviación, la navegación, la automoción, medicina, imagen y sonido en medios audiovisuales, etc...

En este curso nos vamos a centrar en el estudio de las técnicas numéricas para aquellos modelos que se pueden describir mediante la **evolución temporal** de una

o varias cantidades. Nos encontramos entonces con lo que se conoce como ecuaciones (o sistemas de ecuaciones) diferenciales ordinarias; reduciremos la denominación de este problema al acrónimo **edo** tanto en el caso escalar como vectorial.

El **modelado matemático** es el arte (o la ciencia dependiendo del punto de vista) de representar (o aproximar) una realidad física en un modelo abstracto accesible al estudio y al cálculo. La **simulación numérica** es el proceso que permite calcular la solución del modelo en un computador y visualizar una aproximación a la realidad física. Este proceso se complementa con el **contraste de los resultados** del modelo con la realidad física representada. Se puede entender el concepto de **matemática aplicada** como **las matemáticas del modelado y de la simulación numérica**. Desde este punto de vista, se encuentra en la intersección de muchas disciplinas científicas de una gran variedad.

1.2. Ecuaciones Diferenciales Ordinarias

Introducimos notación, algunos resultados y múltiples ejemplos que servirán durante el curso como ilustración.

El **problema de Cauchy asociado a una ecuación diferencial ordinaria** plantea la búsqueda de una función de una variable de la que se conoce cómo cambia su derivada, es decir, su razón de cambio. Esta relación se deduce normalmente del modelo que se está estudiando o representando.

Observación 1 *Una edo define una familia de soluciones o curvas. Si queremos una curva en particular tenemos que dar tantos datos extra de información como el orden de derivación que aparece en la ecuación.*

1.3. Problema de Cauchy o de valor inicial

Siendo I un intervalo de la recta real \mathbb{R} , se busca $u : I \subset \mathbb{R} \mapsto \mathbb{R}^m$, $u(t) \in C^1(I)$ tal que

$$(PC) \quad \begin{cases} u'(t) = f(t, u(t)), & t \in I, \\ u(t_0) = u_0, & t_0 \in I \end{cases}$$

De manera explícita, en el caso de un sistema con $m > 1$, tenemos las ecuaciones

$$\begin{cases} u'_1(t) = f_1(t, u_1(t), u_2(t), \dots, u_m(t)), \\ u'_2(t) = f_2(t, u_1(t), u_2(t), \dots, u_m(t)), \\ \vdots & \vdots \\ u'_{m-1}(t) = f_{m-1}(t, u_1(t), u_2(t), \dots, u_m(t)), \\ u'_m(t) = f_m(t, u_1(t), u_2(t), \dots, u_m(t)) \end{cases}$$

y el dato inicial

$$u_1(t_0) = u_{0,1}, \quad u_2(t_0) = u_{0,2}, \dots, \quad u_m(t_0) = u_{0,m}$$

donde $u_0 = (u_{0,1}, u_{0,2}, \dots, u_{0,m})$.

Algunas puntualizaciones son convenientes:

- Cada f_j depende de $(t, u_1, u_2, \dots, u_m)$ por lo que el sistema está **acoplado**.
- Si f_j depende sólo de (t, u_j) nos encontramos con un sistema **desacoplado**. Cada ecuación es independiente de las demás.
- El dato inicial viene dado por m constantes para obtener así solución única y normalmente pondremos $t_0 = 0$ para fijar ideas, aunque $t_0 \neq 0$ en general.
- Suponemos que $f(t, u) : I \times \mathbb{R}^m \mapsto \mathbb{R}^m$ es una función continua en ambas variables siendo I un intervalo real de alguna de las formas

$$I = [t_0, t_0 + T], \quad I = [t_0, t_0 + T), \quad I = [t_0, +\infty).$$

- A la función f se la puede denominar **función pendiente** puesto que es la función que rige el comportamiento de $u'(t)$ que es la pendiente de $u(t)$.
- Por simplicidad, a la variable independiente la denotamos por t y la llamamos **tiempo**. Esto es así puesto que la mayoría de aplicaciones prácticas son aquellas en las que la edo describe un proceso temporal (también puede darse el caso en el que represente una variable espacial, entonces se puede denotar mejor por \mathbf{x}).
- Dejamos fuera de nuestro estudio ecuaciones de la forma $F(t, u(t), u'(t)) = 0$ en donde no siempre es posible despejar $u'(t)$ en términos del resto.
- Cuando $f(t, u) = f(u)$, es decir, f no depende de la variable temporal, entonces hablamos de una ecuación **autónoma**: la variación del **estado** $u(t)$ sólo depende del propio estado $u(t)$. El sistema se comporta de forma intemporal, permanente en tiempo.

Observación 2 Se puede siempre convertir un problema no autónomo en uno autónomo a costa de aumentar en 1 la dimensión del problema. Se usa la nueva variable vectorial $z = (z_1, z_2) = (u, t)$, entonces la derivada es $z' = (u', 1)$. Luego para $F(z) = (f(z), 1)$ el sistema queda como

$$z'(t) = F(z(t)) \iff \begin{cases} z'_1(t) = u'(t) &= f(t, u), \\ z'_2(t) = t' &= 1. \end{cases}$$

Observación 3 Cuando $u(t)$ representa una cantidad,

- $u'(t) < 0$ implica pérdida de materia
- $u'(t) > 0$ implica ganancia de materia

Es por esto por lo que en muchas aplicaciones a la función f se la denomina también **término de reacción**, ya que es el término que genera o destruye u . Podemos tener entonces un **término de reacción lineal** como $f(t, u) = u$ o un **término de reacción no lineal** como $f(t, u) = 1 + u^2$. Es evidente que la forma general del término de reacción puede ser más complicada y depender también de forma explícita del tiempo. Se usa la notación **fuente** ó **sumidero** en un punto (t_*, u_*) cuando $f(t_*, u_*) > 0$ ó $f(t_*, u_*) < 0$ respectivamente.

Campo de velocidades

Una ecuación diferencial de primer orden nos indica en cada punto (t, u) la pendiente de la curva solución que pasa por ese punto. El **campo de velocidades**, o también **campo de pendientes**, nos indica la pendiente de cada trayectoria posible. La importancia del **campo de pendientes** reside en la información cualitativa que aporta sobre la familia de soluciones sin conocerlas explícitamente. Esta información nos permite considerar la solución continua $u(t)$ no como una curva aislada sino como parte de una familia de curvas que se generan variando el dato inicial y que de manera continua (están una contigua a la otra) cubren una parte o todo el plano (t, u) . En cada punto (t_*, u_*) la pendiente de la curva que pasa por (t_*, u_*) es $f(t_*, u_*)$. Luego si consideramos un segmento pequeño en sentido positivo para cada dirección $t_1 > t_*$ y $u_1 > u_*$ entonces

$$\frac{u_1 - u_*}{t_1 - t_*} \approx f(t_*, u_*) \approx \tan(\alpha)$$

donde α es el ángulo formado por la tangente.

Si $f(t_*, u_*) > 0$ tenemos un ángulo entre 0 y $\pi/2$ para la recta tangente a la curva mientras que si $f(t_*, u_*) < 0$ tenemos entonces un ángulo entre $-\pi/2$ y 0 para la recta tangente a la curva.

La ecuación de la recta tangente $r = r(s)$ en el punto (t_*, u_*) viene dada por

$$r(s) = u_* + (s - t_*)f(t_*, u_*)$$

y un vector en (t_*, u_*) con la misma dirección es

$$\vec{u} = (1, f(t_*, u_*)).$$

Si en cada punto (t_*, u_*) dibujamos un vector de componente OX dada por 1 y de componente OY dada por $f(t_*, u_*)$ generaremos el campo de velocidades que representa la edo. En el computador, para construir un **campo de velocidades** se usa una parrilla de puntos en el plano $\{(t_j, u_j)\}_j$ y en cada punto (t_*, u_*) de la parrilla se dibuja el vector normalizado a 1, es decir,

$$\vec{v} = \frac{1}{a}(1, f(t_*, u_*)), \quad a = \sqrt{1^2 + f(t_*, u_*)^2}.$$

Este proceso se puede aproximar de forma manual y sin ser muy preciso, pero dándole más importancia al sentido y la dirección que al valor exacto del vector usando simplemente $(1, f(t_*, u_*))$.

Ejemplo 1 En la Figura 1.1 se representa un sistema no autónomo, esto es, el campo de pendientes cambia con el tiempo. En la Figura 7.1 se ve un sistema autónomo, esto es $(1, f(u))$, ahora el campo de pendientes no cambia con el tiempo. En la Figura 7.3 y en la 1.4 podemos ver las respectivas trayectorias asociadas, o curvas solución, a estos campos de pendientes.

Observación 4 Aparte de usar nuestra propia programación existen algunos recursos web para visualizar un campo de pendientes y las soluciones asociadas. Por ejemplo:

- <http://www.aw-bc.com/ide/>
Interactive differential equations
- <https://www.cs.unm.edu/~joel/dfield/>
Aplicaciones dfield y pplane

Orden de derivación mayor que uno

Las ecuaciones del tipo

$$u^{(q)} = \varphi(t, u, u', u'', \dots, u^{(q-1)}) \quad (q > 1)$$

se pueden reescribir como sistemas de primer orden usando variables auxiliares. Ponemos $Y_j = u^{(j-1)}$ para $j = 1, 2, 3, \dots, q$, es decir, $Y_1 = u$, $Y_2 = u'$, ..., y obtenemos las ecuaciones

$$\begin{cases} Y'_1 &= u', \\ Y'_2 &= u'', \\ \vdots &\vdots \\ Y'_{q-1} &= u^{(q-1)} \\ Y'_q &= \varphi(t, Y_1, Y_2, \dots, Y_{q-1}) \end{cases}$$

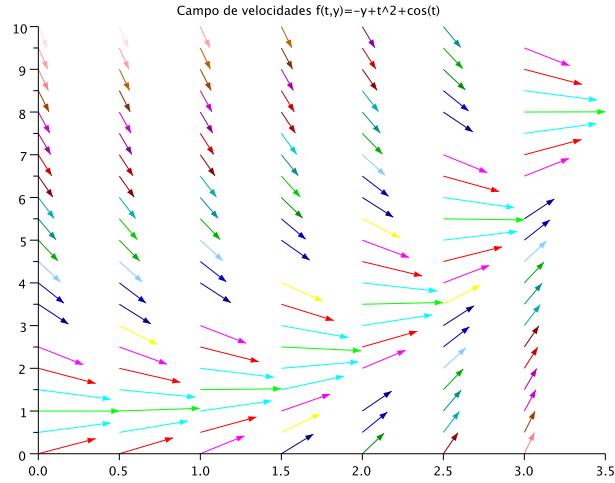


Figura 1.1: Campo de velocidades no autónomo $f(t, y) = -y + t^2 + \cos(t)$. Para cada valor de y (en el eje OY) el campo de velocidades cambia dependiendo del tiempo t (en el eje OX). Observar el comportamiento del campo de acuerdo a las magnitudes de las variables.

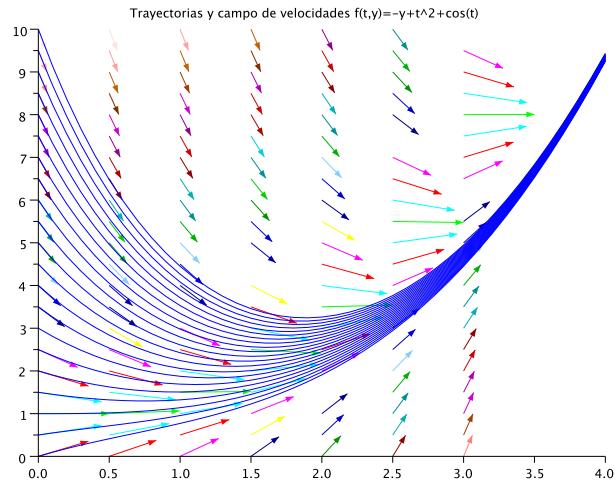


Figura 1.2: Soluciones y campo de velocidades para $f(t, y) = -y + t^2 + \cos(t)$.

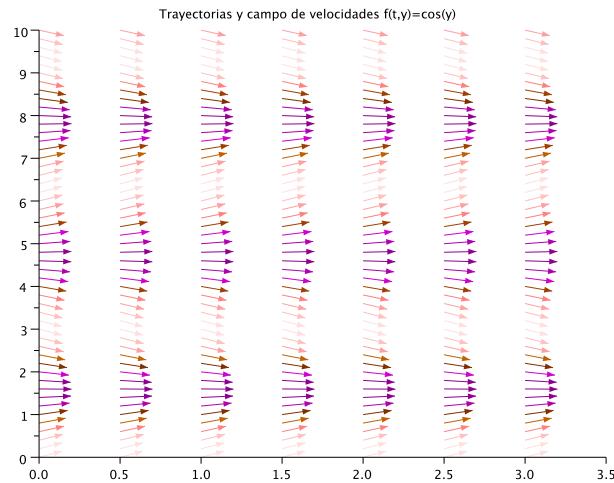


Figura 1.3: Campo de velocidades autónomo $f(t,y) = \cos(y)$. Para cada valor de y (en el eje OY) el campo de velocidades es siempre el mismo, luego no cambia dependiendo del tiempo t (en el eje OX).

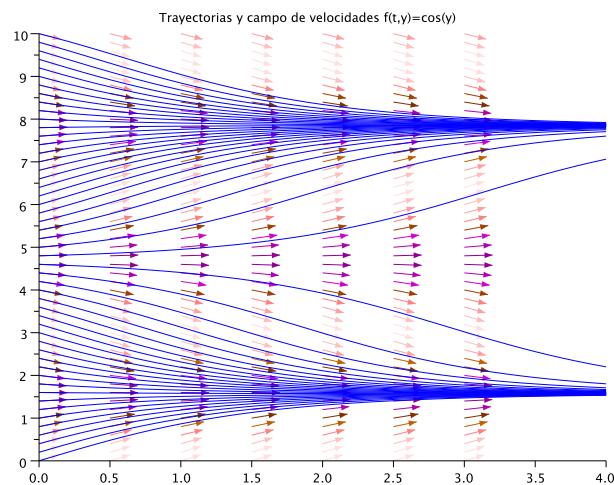


Figura 1.4: Soluciones y campo de velocidades para $f(t,y) = \cos(y)$.

La reducción de un sistema de orden mayor que uno a un sistema de orden uno se suele hacer siempre. En el caso del sistema de segundo orden existe la excepción, voluntaria, tradicional y sobre todo cuando no aparece la primera derivada, de poder trabajarla sin transformarla, ya que históricamente se han diseñado métodos especiales. La comparación sobre lo que conviene más no es clara pero como hay una gran cantidad de software disponible para sistemas de primer orden se suele aplicar la reducción.

Observación 5 Método del factor integrante. *Nos va a ser muy útil recordar esta técnica para resolver ecuaciones lineales: si queremos encontrar la solución de*

$$u'(t) = \lambda u(t) + g(t)$$

ponemos

$$u'(t) - \lambda u(t) = g(t)$$

y multiplicando por $e^{-\lambda t}$ llegamos a

$$(e^{-\lambda t} u(t))' = e^{-\lambda t} g(t).$$

Integrando llegamos a la solución fácilmente:

$$u(t) = e^{\lambda t} u_0 + \int_0^t g(s) e^{-\lambda(s-t)} ds.$$

Ejercicio 2 Acotar la función $u(t)$ usando este proceso

$$u'(t) \leq \lambda u(t) + g(t)$$

1.4. Existencia y unicidad de solución

Buscamos una función $u : \mathbb{R} \mapsto \mathbb{R}$ regular (de clase C^1) solución del problema de Cauchy

$$(PC) \quad \begin{cases} u'(t) &= f(t, u(t)), \quad t \in [0, T], \\ u(0) &= u_0 \end{cases}$$

donde $f(t, u) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$.

Definición 3 Una función $f(t, u)$ cumple la condición de Lipschitz sobre su segunda variable cuando existe una constante $L_f > 0$ tal que

$$|f(t, x) - f(t, y)| \leq L_f |x - y|, \quad \forall x, y \in \mathbb{R}, \quad t \in [0, T].$$

Interpretación geométrica:

Si se cumple la condición de Lipschitz sobre la segunda variable de $f(t, u)$ entonces la función es derivable con respecto a u en casi todo punto y con derivada acotada. Además, en aquellos puntos donde no es derivable las derivadas laterales tienen saltos de discontinuidad finitos. Esto ocurre porque estamos diciendo

$$\frac{|f(t, x) - f(t, y)|}{|x - y|} \leq L_f \quad \forall x, y \in \mathbb{R}, \quad t \in [0, T].$$

Por lo tanto, si hay derivada está acotada y, si no la hay, los límites laterales existen y están acotados aunque no coincidan.

Observación 6 *Es más que continua pero menos que derivable. Aunque es derivable en casi todo punto. Una función de clase C^1 es localmente Lipschitz.*

Ejemplo 4 *Una función en forma de dientes de sierra es Lipschitz y no es derivable en los picos o valles de la sierra.*

Teorema 5 *Si se cumple la condición de Lipschitz*

$$|f(t, x) - f(t, y)| \leq L_f |x - y|, \quad \forall x, y \in \mathbb{R}, \quad t \in [0, T]$$

podemos garantizar la unicidad de la solución.

Dem: Supongamos que para $t > t_0$

$$y'(t) = f(t, y(t)), \quad z'(t) = f(t, z(t))$$

entonces

$$y(t) - z(t) = y(t_0) - z(t_0) + \int_{t_0}^t \{f(s, y(s)) - f(s, z(s))\} ds$$

luego si f es Lipschitz tenemos

$$|y(t) - z(t)| \leq |y(t_0) - z(t_0)| + L_f \int_{t_0}^t |y(s) - z(s)| ds.$$

Poniendo

$$g(t) = |y(t_0) - z(t_0)| + L_f \int_{t_0}^t |y(s) - z(s)| ds$$

entonces tenemos

$$|y(t) - z(t)| \leq g(t) \tag{1.1}$$

y $g(t)$ cumple

$$g(t_0) = |y(t_0) - z(t_0)|, \quad g'(t) = L_f |y(t) - z(t)|.$$

Pero, usando (1.1), tenemos que

$$L_f |y(t) - z(t)| \leq L_f g(t).$$

Entonces $g'(t) \leq L_f g(t)$ y usando el factor integrante

$$\exp(-L_f(t - t_0))$$

obtenemos

$$\frac{d}{dt}(g(t) \exp(-L_f(t - t_0))) \leq 0$$

de donde

$$g(t) \leq g(t_0) \exp(L_f(t - t_0)).$$

Pero como $|y(t) - z(t)| \leq g(t)$ llegamos a

$$|y(t) - z(t)| \leq |y(t_0) - z(t_0)| e^{L_f(t-t_0)}$$

lo que implica la unicidad de soluciones. ■

Observemos que

- El paso de la estimación integral a la puntual se conoce como el **Lema de Gronwall**.
- No sólo hemos comprobado la unicidad, además hemos dado una acotación sobre la separación de las soluciones desde puntos iniciales distintos.
- La acotación se abre exponencialmente conforme crece t debido a la función exponencial que aparece. Esta acotación es precisa en unos casos pero muy pesimista en otros: Por ejemplo, para $a > 0$
 - es precisa si tenemos $y'(t) = a y(t)$
 - pero no lo es en el caso $y'(t) = -a y(t)$.

En ambas situaciones $L_f = a$ pero el primer caso las soluciones se **expanden de forma exponencial** mientras que en el segundo se **contraen de forma exponencial**, ver Figura 1.5.

- Si $f \in C^1$ y usamos el TVM se puede ajustar mejor esta estimación ya que si tenemos dos curvas solución

$$y'(t) = f(t, y(t)), \quad z'(t) = f(t, z(t))$$

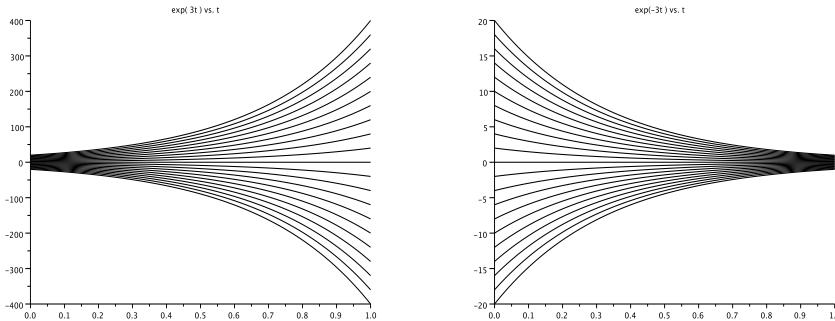


Figura 1.5: Familia de soluciones típicas del problema modelo $y'(t) = a y(t)$ con $a > 0$ y $a < 0$

es inmediato que para $t > t_0$

$$\begin{aligned} y(t) - z(t) &= y(t_0) - z(t_0) + \int_{t_0}^t \{f(s, y(s)) - f(s, z(s))\} ds \\ &= y(t_0) - z(t_0) + \int_{t_0}^t \partial_y f(s, \xi_s) \{(y(s) - z(s))\} ds \\ &= y(t_0) - z(t_0) + \partial_\xi f(t_*, \xi_*) \int_{t_0}^t \{(y(s) - z(s))\} ds \end{aligned}$$

y entonces podemos tener en cuenta el signo de $\partial_\xi f$ (en el caso antes mencionado es $f(t, y) = -a y$ y por lo tanto $\partial_\xi f = -a$).

Vemos entonces que es el comportamiento de $\partial_y f(s, \xi_s)$ en todo el intervalo $[t_0, t]$ lo que influye en la variación entre las curvas. Si $\partial_y f(s, \xi_s) < 0$ hay contracción y si $\partial_y f(s, \xi_s) > 0$ hay expansión, de hecho, si usamos la derivada

$$y'(t) - z'(t) = \partial_y f(t, \xi_t)(y(t) - z(t))$$

y el factor integrante obtenemos

$$y(t) - z(t) = e^{\int_{t_0}^t \partial_y f(s, \xi_s) ds} (y(0) - z(0)).$$

Por lo tanto, si $\partial_y f(s, \xi_s) < 0$ en todo el intervalo podemos concluir

$$|y(t) - z(t)| < |y(0) - z(0)|.$$

Observación 7 *Observar que al acotar*

$$L_f = \max_{t,y} |\partial_y f(t, y)|$$

recuperamos una acotación global, aunque menos precisa en general. Como consecuencia, una condición de Lipschitz local viene a decir que tenemos controlada, en una banda $(t, y) \in (t_0, t_1) \times (y_{\min}, y_{\max})$ la derivada $\partial_y f(t, y)$ por lo que existe un control en el comportamiento de las curvas vecinas a una dada.

Las condiciones que se imponen sobre la función pendiente son

1. $f(t, y) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ es una función continua en ambas variables
2. $f(t, y)$ cumple la condición de Lipschitz sobre la segunda variable

$$|f(t, x) - f(t, y)| \leq L_f |x - y|, \quad \forall x, y \in \mathbb{R}, \quad t \in [0, T]$$

3. Controlamos además el crecimiento de f con respecto a la segunda variable, esto es,

$$|f(t, y)| \leq C_f (1 + |y|), \quad \forall y \in \mathbb{R}, \quad t \in [0, T].$$

Entender esto como $|f(t, y)|$ acotada si $|y| < 1$ y $|f(t, y)| \leq C|y|$ para $|y| > 1$. Esto quiere decir que se puede permitir un crecimiento de la pendiente a lo sumo lineal con respecto a la variable y .

Teorema 6 Teorema de existencia, unicidad. Comportamiento

- *Bajo las condiciones 1 y 2 hay existencia y unicidad de solución local en un subintervalo de $[0, T]$*
- *Si controlamos además el crecimiento de f con respecto a la segunda variable, esto es, el punto 3, entonces tenemos garantizado la existencia y unicidad de solución en todo $[0, T]$, ver por ejemplo el libro de Arnold [2].*
- *Si además $f \in C^p$ entonces $y \in C^{p+1}$.*

1.5. Estudio cualitativo de las soluciones

Observación 8 Punto de equilibrio Dado el problema autónomo $z'(t) = f(z(t))$ si para el valor constante $z = c_\star$ se cumple $f(c_\star) = 0$ entonces $z(t) \equiv c_\star$ es una solución constante que se denomina un **punto de equilibrio** del sistema.

Como $f(c_\star) = 0$, usando el desarrollo de Taylor tenemos que

$$f(z) = f(c_\star) + (z - c_\star)f'(c_\star) + \dots = (z - c_\star)f'(c_\star) + \dots$$

Finalmente, si suponemos para simplificar que $c_\star = 0$ nos encontramos que para z cercano a $c_\star = 0$

$$f(z) \approx az, \quad (a = f'(0))$$

luego la dinámica cerca del punto de equilibrio se puede estudiar viendo el problema modelo

$$z'(t) = a z(t)$$

con solución $z(t) = z_0 e^{at}$ para valores z_0 cercanos a $z = 0$. Si $a > 0$ todas las soluciones se alejan del punto de equilibrio a gran velocidad, tenemos que el **estado estacionario** $z(t) = 0$ es **inestable**. Por otro lado, si $a < 0$ todas las soluciones convergen a gran velocidad hacia $z(t) = 0$ y se tiene que el **estado estacionario es estable**, ver Figura 1.5.

Observación 9 Es posible extraer información sobre las soluciones sin necesidad de conocerlas explícitamente. Por ejemplo, si tenemos el modelo

$$y'(t) = f(y(t))$$

y sabemos sobre f que tiene un comportamiento como sigue

$$\begin{aligned} f(a) &= 0, \quad f(b) = 0, \\ f(y) &> 0 \quad a < y < b, \\ f(y) &< 0 \quad y \in \mathbb{R} \setminus (a, b), \end{aligned}$$

por ejemplo $f(y) = (y-a)(b-y)$. Gracias a que las curvas solución no se pueden cortar podemos deducir fácilmente que hay dos puntos de equilibrio $y_1(t) \equiv a$, $y_2(t) \equiv b$ siendo uno estable y otro inestable.

Observación 10 La dependencia de la función pendiente $f(t, u)$ con respecto a u es la que marca el comportamiento del sistema.

- $\partial_u f > 0$ implica expansión local o **problema inestable localmente**
- $\partial_u f < 0$ implica contracción local o **problema estable localmente**
- $\partial_u f = 0$ sin cambios localmente

Observación 11 Se pueden hacer ejemplos de campos de vectores y obtener información sobre la solución teniendo en cuenta dos hechos fundamentales:

- Dos curvas distintas no se pueden cruzar.
- Si dos curvas se tocan tangencialmente entonces son la misma.

Los ejemplos pueden ser $u' = \pm 3u - 3t$ y también $u' = a(1-u)$ o bien $u' = a(\sin(t)-u)$ para $a > 0$.

Una misma curva $y(t)$ puede vivir en campos de comportamiento muy distinto, ver por ejemplo la Figura 1.6. Esto es muy importante cuando se intenta calcular una aproximación a $y(t)$ ya que, a priori, no se sabe como es el campo. **Será la dificultad que encontraremos al aplicar un método numérico lo que nos diga como es el campo.**

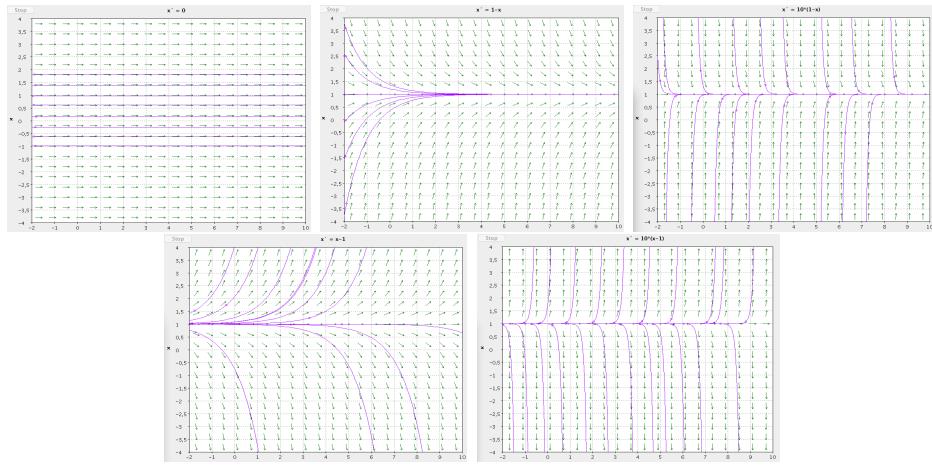


Figura 1.6: La misma curva $y(t) = 1$ puede vivir en campos de comportamiento muy distinto. Puede estar en campos contantes, en campos que se contraigan o que se expandan (ver de izquierda a derecha y de arriba a abajo).

1.6. Ejemplos (I): Problemas de primer orden

Algunos ejemplos que provienen de ciencias experimentales:

Ejemplo 7 Modelo de poblaciones lineal

La ecuación

$$\begin{aligned} \frac{du}{dt}(t) &= \lambda u(t), \quad t > 0, \\ u(0) &= u_0 > 0, \end{aligned}$$

expresa un crecimiento proporcional a la población existente en cada instante y nos lleva a un crecimiento exponencial. Este problema se puede reescribir usando $n(t) = u(t)/u_0$ como

$$\begin{aligned} n'(t) &= \lambda n(t), \quad t > 0, \\ n(0) &= 1. \end{aligned}$$

Ejemplo 8 Modelo de poblaciones no lineal

Es interesante estudiar el comportamiento de estos modelos de población cuando el crecimiento no es proporcional a la población sino a una potencia en torno a uno de la misma. Esto es,

$$\begin{aligned} n'(t) &= n(t)^{1+\epsilon}, \quad t > 0, \\ n(0) &= 1. \end{aligned}$$

para algún $\epsilon \in \mathbb{R}$. En el caso $\epsilon > 0$ existe un tiempo $t_* < +\infty$ tal que $n(t) \rightarrow +\infty$ cuando $t \rightarrow t_*$, se dice que la solución **explota en tiempo finito**. A este tiempo t_* se llama **tiempo finito de explosión**. Aquí es

$$f(t, n) = n^{1+\epsilon}$$

y la condición 3 no se cumple. En particular, las soluciones de la ecuación

$$u'(t) = u(t)^2$$

explotan en tiempo finito positivo y las de

$$u'(t) = -u(t)^2$$

en tiempo finito negativo, dependiendo también de los datos iniciales.

Ejemplo 9 Si ponemos $a \in (0, 1)$

$$\begin{aligned} n'(t) &= n(t)^a, & t > 0, \\ n(0) &= 0. \end{aligned}$$

no posee solución única ya que para cada $c > 0$ tiene por solución

$$n_c(t) = 0, \quad \forall t$$

pero también

$$n_c(t) = \begin{cases} 0, & 0 < t \leq c, \\ (1-a)^{1/(1-a)}(t-c)^{1/(1-a)}, & t > c. \end{cases} \quad (1.2)$$

Aquí lo que falla es que

$$f(n) = n^a \Rightarrow f'(n) = an^{a-1} = a \frac{1}{n^{1-a}}$$

y entonces la función no es lipschitziana en ningún entorno de $n = 0$.

Ejemplo 10 Modelo de Verhulst o ecuación logística

El modelo de poblaciones lineal $n'(t) = \lambda n(t)$ tiene la contrariedad de predecir un crecimiento exponencial ilimitado en tiempo, lo que no es muy realista. Una mejora la constituye el modelo de Verhulst que tiene en cuenta la **capacidad limitada del medio para sustentar la población**.

Supongamos que $n(t)$ es el tamaño de una población en el instante t . En el modelo más sencillo de crecimiento de una población se supone un crecimiento proporcional al tamaño de la población,

$$\frac{d}{dt}n(t) = \varepsilon n(t), \quad t > 0$$

donde la constante $\varepsilon > 0$ es la tasa de crecimiento: la variación neta de población por unidad de tiempo dividida por la población total en ese mismo instante de tiempo. La solución de esta ecuación viene dada por

$$n(t) = n_0 e^{\varepsilon t}, \quad t > 0.$$

Esto da una población que crece de forma exponencial e ilimitada. Sin embargo, el crecimiento de la población puede estar limitado por los recursos del medio, lo cual, a su vez, implica que la rapidez de crecimiento debe de ser una función de la población, es decir, la tasa de crecimiento debe depender de n :

$$\frac{1}{n(t)} n'(t) = F(n(t)).$$

Cuando la población es pequeña, habrá suficientes recursos, de modo que la rapidez de crecimiento será independiente de los recursos y por tanto, podremos decir que, aproximadamente, la tasa de crecimiento será constante, es decir, $F(n) \approx \varepsilon$. Sin embargo, conforme la población crece, los recursos van desapareciendo y se produce un efecto inhibidor sobre el crecimiento de la población. Entonces, la tasa de crecimiento $F(n)$ debe decrecer conforme n crece, de modo que, eventualmente, sea $F(n) < 0$. La función más sencilla que satisface estas condiciones es

$$F(n) = \varepsilon - \sigma n$$

donde $\sigma > 0$. Por lo tanto, consideramos el problema de valor inicial

$$\begin{cases} \frac{d}{dt} n(t) = \varepsilon n(t) - \sigma n(t)^2, & t > 0, \\ n(0) = n_0 > 0, \end{cases} \quad (1.3)$$

donde $\varepsilon, \sigma > 0$ son números positivos dados. El problema (1.3) se conoce en ecología como ecuación logística y surge cuando se hace el análisis del desarrollo de una sola especie que tiene acceso a recursos limitados.

Como hemos notado, si el término no lineal σn^2 no se toma en cuenta en la ecuación (1.3), entonces la población crece ilimitadamente. La situación es completamente diferente si se conserva el término no lineal. Además, si $n = 0$ ó si $n = \varepsilon/\sigma$ el segundo miembro de (1.3) se anula y tenemos dos soluciones constantes o estacionarias

$$n_1(t) \equiv 0, \quad n_2(t) \equiv \varepsilon/\sigma.$$

Estas dos soluciones poseen una importancia particular y se denominan **puntos críticos**, o como $n'(t) \equiv 0$, **puntos de equilibrio**.

Siempre y cuando $n \neq 0$ ó $n \neq \varepsilon/\sigma$ podemos resolver la ecuación diferencial del problema (1.3) de forma explícita usando técnicas de separación de variables y encontramos la expresión

$$n(t) = \frac{\varepsilon n_0}{\sigma n_0 + (\varepsilon - \sigma n_0) e^{-\varepsilon t}}$$

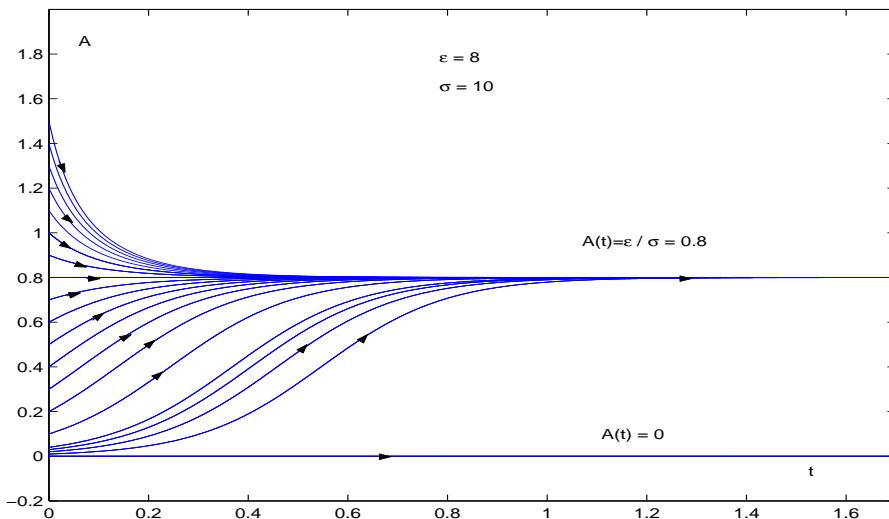


Figura 1.7: $A'(t) = \varepsilon A(t) - \sigma A(t)^2$, $\varepsilon > 0$, $\sigma > 0$. Varias soluciones típicas para distintos valores de A_0 así como las soluciones constantes $A_1(t) \equiv 0$ y $A_2(t) \equiv \varepsilon/\sigma$.

que sí que es válida para todo $n_0 > 0$.

Podemos observar de la Figura 1.7 que para cualquier $n_0 > 0$, independientemente de lo pequeño que sea, al final se tiende siempre a la solución constante $n_1(t) \equiv \varepsilon/\sigma$ cuando $t \rightarrow +\infty$. A la razón ε/σ se le denomina **nivel de saturación**; una población que se inicia por debajo de este nivel no podrá sobrepasarla nunca y si hay sobre población lleva a la eliminación de individuos hasta alcanzar el nivel de saturación.

Si examinamos la solución $n(t) \equiv 0$ resulta que si $n_0 = 0$ es la entrada, dato inicial, de nuestro modelo, tendremos que $n(t) \equiv 0$ es la solución en todo tiempo, pero si cometemos un pequeño error y la entrada es $n_0 > 0$, aunque pequeña, resulta que la solución $n(t)$ se aleja asintóticamente en tiempo hacia ε/σ . Se dice entonces, y de forma natural, que la solución $n(t) \equiv 0$ es una **solución inestable** de (1.3), o que $n = 0$ es un **punto crítico (de equilibrio) inestable**.

Si consideramos ahora el caso de la solución $n(t) \equiv \varepsilon/\sigma$, si la entrada, dato inicial, es $n_0 = \varepsilon/\sigma$, entonces la salida, la solución, será siempre $n(t) = \varepsilon/\sigma$. Si se comete un error en la entrada, la salida se aproximará asintóticamente a ε/σ cuando $t \rightarrow +\infty$. Por lo tanto, decimos que la solución $n(t) \equiv \varepsilon/\sigma$ es una **solución asintóticamente estable** de (1.3), o que $n = \varepsilon/\sigma$ es un **punto crítico (de equilibrio) estable**. Podemos dar a esto una **interpretación biológica**:

- Si queremos realizar **un experimento libre de bacterias**, en un medio en el que el desarrollo de las bacterias viene dado por (1.3). Entonces, a no ser que eliminemos completamente la población de bacterias, no conseguiremos realizar el experimento manteniendo la población de bacterias por debajo de

unos mínimos menores que el valor ε/σ . El experimento no resultará seguro.

- Si por otro lado estamos tratando de realizar **un experimento manteniendo un nivel de bacterias** del orden de ε/σ podemos estar tranquilos de que, aunque de vez en cuando se eliminan bacterias por algún contaminante, la población no desaparecerá y tenderá en tiempo a aproximar el valor ε/σ .

Este **comportamiento cualitativo** es de gran interés aplicado y demuestra desde el punto de vista matemático que, en general, **el valor asintótico de la solución no depende de forma continua de los datos iniciales**.

Observación 12 La información presentada se ha conseguido resolviendo la ecuación, pero para problemas más complicados, para expresiones más elaboradas de (1.3), es significativo que el comportamiento asintótico, es decir, en relación con el comportamiento cuando $t \rightarrow +\infty$, a menudo puede determinarse sin un conocimiento previo de la solución. Por ejemplo, si tenemos el modelo

$$y'(t) = f(y(t))$$

y $f(y) = (y - a)(b - y)$, entonces

$$\begin{aligned} f(a) &= 0, & f(b) &= 0, \\ f(y) &> 0 & a < y < b, \\ f(y) &< 0 & y \in \mathbb{R} \setminus (a, b). \end{aligned}$$

Podemos deducir fácilmente que hay dos puntos de equilibrio $y_1(t) \equiv a$, $y_2(t) \equiv b$ y que el comportamiento de la solución es similar al del modelo logístico y tenemos un gráfico similar a la Figura 1.7. En el ejemplo previo tenemos

$$\begin{cases} y'(t) = y(t)(\varepsilon - \sigma y(t)), \\ y(0) = y_0 > 0, \end{cases} \quad (1.4)$$

luego los puntos de equilibrio son $y = 0$ e $y = \varepsilon/\sigma$ siendo el primero inestable y el segundo estable.

Ejemplo 11 Modelo depredador-presa de Lotka-Volterra

Supongamos que queremos aproximar el comportamiento de dos especies que interactúan entre sí siendo depredador $D(t)$ y presa en un entorno $P(t)$. Por ejemplo, la evolución de una población de insectos con respecto a la de pájaros. Normalmente, el comportamiento puede ser el siguiente:

Si crece la población de presas los depredadores aumentan porque tienen comida. Entonces empieza a decrecer la población de presas y los depredadores van desapareciendo porque no tienen comida, con lo que aumentan las presas... y así volvemos a empezar el ciclo de la vida en el ecosistema.

Por lo tanto, nos encontramos con una evolución oscilatoria e interconectada para cada especie. La **interacción** entre dos especies $P(t)$ y $D(t)$ se representa por el **producto** de ambas cantidades $P(t)D(t)$. Entonces, uno de los modelos más simples puede ser el **modelo de Lotka-Volterra** dado por, ver por ejemplo Golub-Ortega [12] entre muchos otros,

$$\begin{aligned}\frac{d}{dt}P(t) &= aP(t) - bP(t)D(t) \\ \frac{d}{dt}D(t) &= -cD(t) + dP(t)D(t)\end{aligned}$$

donde $D(t)$ es la población de depredadores, $P(t)$ es la población de presas en el instante de tiempo t y los valores a, b, c, d son parámetros positivos del modelo.

Observación 13 Como anécdota, es curioso que también este modelo sirve para una relación entre dos personas donde se modela una dependencias afectivas. A veces, en el contexto de una pareja se le da también el nombre de modelo de Romeo y Julieta.

En forma vectorial, poniendo $X = (P, D)$ y

$$f(P, D) = aP - bPD, \quad q(P, D) = -cD + dPD, \quad F = (f, g)$$

se puede escribir como

$$X'(t) = F(X(t)).$$

Se puede observar como los insectos $P(t)$ tendrían un crecimiento exponencial si no hubiese depredadores mientras que los depredadores $D(t)$ tendrían un decrecimiento exponencial:

$$\begin{aligned}\frac{d}{dt}P(t) &= aP(t) \\ \frac{d}{dt}D(t) &= -cD(t)\end{aligned}$$

La idea presente es que la capacidad reproductiva de las presas es mucho más grande que la de los depredadores, y estos últimos tienen crecimiento si hay presas. El efecto de la interacción es negativo en las presas $-bP(t)D(t)$ y positivo en los depredadores $+dP(t)D(t)$ que es de donde sale el modelo, ver Figura 1.8.

Una herramienta de gran utilidad a la hora de estudiar sistemas dinámicos es lo que se conoce como **plano de fases** (se describe la fase (x, y) del sistema). El comportamiento de las curvas paramétricas $(x(t), y(t))$ nos da información relevante sobre **puntos de equilibrio estables e inestables** así como de **órbitas periódicas y trayectorias no cerradas**.

Se puede observar fácilmente que este modelo de Lotka-Volterra posee dos puntos estacionarios $(P, Q) = (0, 0)$ que es inestable y $(P, Q) = (c/d, a/b)$ que es mucho más interesante.

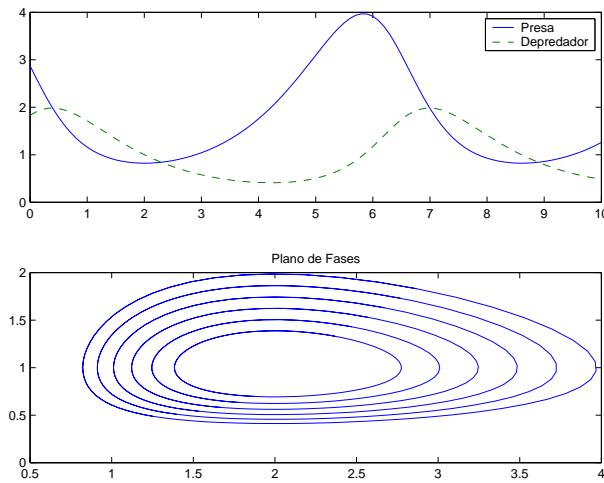


Figura 1.8: Solución típica del modelo de Lotka-Volterra así como algunas de las curvas dentro de su plano de fases

Las ecuaciones de Lotka-Volterra no se pueden resolver de forma exacta y por eso también es útil la versión linealizada (¿Linealizada con respecto a qué punto?) de estos modelos

$$\begin{aligned}\frac{d}{dt}P(t) &= aP(t) - bD(t) \\ \frac{d}{dt}D(t) &= -cD(t) + dP(t)\end{aligned}$$

que sí que se puede resolver de forma exacta recurriendo a los autovectores y autovalores de la matriz de coeficientes $A = [a, -b; -c, d]$.

Observación 14 La consideración de un número genérico de m especies y_1, \dots, y_m y las interacciones mutuas nos lleva con facilidad a un sistema de m ecuaciones diferenciales no lineales.

$$\left\{ \begin{array}{lcl} y'_1(t) & = & f_1(t, y_1(t), y_2(t), \dots, y_m(t)), \\ y'_2(t) & = & f_2(t, y_1(t), y_2(t), \dots, y_m(t)), \\ \vdots & \vdots & \vdots \\ y'_{m-1}(t) & = & f_{m-1}(t, y_1(t), y_2(t), \dots, y_m(t)), \\ y'_m(t) & = & f_m(t, y_1(t), y_2(t), \dots, y_m(t)) \end{array} \right.$$

con dato inicial

$$y_1(t_0) = y_{0,1}, y_2(t_0) = y_{0,2}, \dots, y_m(t_0) = y_{0,m}$$

donde $y_0 = (y_{0,1}, y_{0,2}, \dots, y_{0,m})$. Por ejemplo, si ponemos $y_1 = P$ e $y_2 = Q$ entonces el modelo anterior se escribe como

$$\left\{ \begin{array}{lcl} y'_1(t) & = & a y_1(t) - b y_1(t)y_2(t), \\ y'_2(t) & = & -c y_2(t) + d y_1(t)y_2(t). \end{array} \right.$$

Ejemplo 12 Modelo de Lorenz o el efecto Mariposa

Fue obtenido por Lorenz (1963) en sus estudios sobre meteorología. Consiste en el sistema no lineal

$$\begin{aligned}x'(t) &= -cx(t) + cy(t), \\y'(t) &= ax(t) - y(t) - x(t)z(t), \\z'(t) &= bx(t)y(t) - bz(t),\end{aligned}$$

en donde a, b y c son parámetros y las tres magnitudes $x(t), y(t), z(t)$ son cantidades termodinámicas obtenidas simplificando un modelo meteorológico. A pesar de ser simples, no disponen de una solución analítica en general y es un ejemplo de un sistema caótico en \mathbb{R}^3 .

El efecto mariposa se puede observar cuando se visualizan soluciones calculadas y se observa que se mantienen rotando durante un tiempo en un ala y de repente saltan hacia la otra ala donde también siguen rotando, ver Figura 1.9. El tiempo de salto entre alas es impredecible y soluciones distintas para datos iniciales diferentes generan comportamientos similares pero con tiempos de salto distintos. Por lo tanto, las curvas solución son muy diferentes pero generan la misma figura geométrica de las dos alas. Esto ha contribuido a entender por qué es difícil una predicción meteorológica a largo tiempo. Además es el origen de la famosa frase pronunciada por Lorenz:

el movimiento de las alas de una mariposa en Brasil pueden generar un tornado en Texas dos semanas mas tarde.

1.7. Ejemplos (II): Problemas de orden mayor que uno

Los problemas vistos hasta ahora se consideran de primer orden pero muchas aplicaciones provienen de la **Segunda Ley de Newton** $F = ma$. Aquí F es la fuerza sobre un cuerpo de masa m que produce una aceleración a . La aceleración es la segunda derivada del vector de posición del cuerpo.

Si denotamos por $x(t)$ la posición y ponemos $g = F/m$ nos encontramos con una ecuación donde aparece la segunda derivada, ya que $a(t) = x''(t)$, de la forma

$$x''(t) = g(t, x(t), x'(t)).$$

La transformación a una ecuación de primer orden se hace usando el par de incógnitas $(x(t), v(t))$ donde $v(t) = x'(t)$ y escribiendo entonces el problema como

$$\begin{cases} x'(t) &= v(t), \\ v'(t) &= g(t, x(t), v(t)). \end{cases}$$

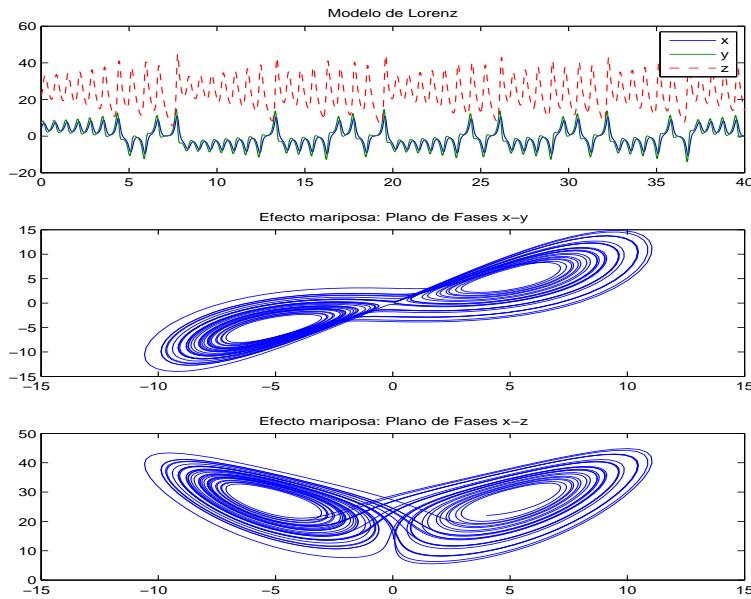


Figura 1.9: Modelo de Lorenz (1963) y el efecto mariposa ($a = 28$, $c = 10$, $b = 8/3$).

Tenemos un sistema $u'(t) = f(t, u(t))$ donde

$$u = (u_1, u_2)^T = (x, v)^T \quad f(t, u) = (u_2, g(t, u_1, u_2))^T$$

Ejemplo 13 Movimiento armónico simple

La variación del ángulo que un péndulo forma con la vertical es lo único necesario para conocer su movimiento. Si se llama este ángulo $\theta(t)$, entonces el modelo se describe por:

$$\begin{cases} \theta''(t) = -\sin(\theta(t)), & t > 0, \\ \theta(0) = \theta_0, \\ \theta'(0) = v_0. \end{cases}$$

y como es una ecuación no lineal, usando Taylor para $f(\theta) = \sin(\theta)$ en torno a $\theta_0 = 0$ tenemos una aproximación a su movimiento desde la posición de reposo dada por

$$\begin{cases} \theta''(t) = -\theta(t), & t > 0, \\ \theta(0) = 0, \\ \theta'(0) = v_0. \end{cases}$$

En términos de un sistema de primer orden con las incógnitas $u(t) = \theta(t)$ y $v(t) = \theta'(t)$ queda como

$$\frac{d}{dt} \begin{pmatrix} u(t) \\ v(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}.$$

y nos encontramos con autovalores complejos puros conjugados para la matriz A del sistema dados por $\sigma(A) = \{i, -i\}$. La solución general es una combinación lineal real de las funciones complejas $\{e^{it}, e^{-it}\}$.

Ejemplo 14 Cálculo de trayectorias (I): satélites

La situación de un satélite viene determinada por su vector de posición $\vec{r}(t) \in \mathbb{R}^3$, su velocidad $\vec{v}(t) \in \mathbb{R}^3$ y su aceleración $\vec{a}(t) \in \mathbb{R}^3$, donde

$$\vec{v}(t) = \frac{d}{dt} \vec{r}(t) = \vec{r}'(t) \in \mathbb{R}^3, \quad \vec{a}(t) = \frac{d}{dt} \vec{v}(t) = \frac{d^2}{dt^2} \vec{r}(t) = \vec{r}''(t) \in \mathbb{R}^3.$$

La relación que liga entre sí estas magnitudes viene dada por la segunda Ley de Newton y si la fuerza se toma dependiente de la posición y de la velocidad llegamos a una ecuación diferencial ordinaria de segundo orden

$$\frac{d^2}{dt^2} \vec{r}(t) = f(t, \vec{r}(t), \vec{r}'(t))$$

que suele venir complementada con la posición inicial $\vec{r}(0) = \vec{r}_0$ y la velocidad inicial $\vec{r}'(0) = \vec{v}_0$. Para simplificar, podemos suponer que $r(t)$ es un escalar, entonces tenemos la ecuación

$$\begin{cases} r(0) &= r_0, \\ r'(0) &= v_0, \\ r''(t) &= f(t, r(t), r'(t)), \quad t > 0. \end{cases}$$

La ecuación se puede reescribir en términos de un sistema de primer orden renombrando la derivada $r'(t)$. Por ejemplo, usamos $r(t)$ y $v(t) = r'(t)$ y queda como

$$\begin{cases} r'(t) &= v(t), \quad t > 0, \\ v'(t) &= f(t, r(t), v(t)), \quad t > 0, \end{cases}$$

con $(r(0), v(0)) = (r_0, v_0)$.

Ejemplo 15 Cálculo de trayectorias (II): balística

Los problemas de balística tienen una larga historia en la computación científica y fueron la principal motivación del desarrollo de los computadores durante las dos grandes guerras mundiales, principalmente la segunda, por la **necesidad de confeccionar tablas de tiro precisas**. Siguiendo el ejemplo descrito en el libro de Golub y Ortega [12] podemos considerar un caso simple del estudio de la trayectoria de un cohete que se lanza con un ángulo de inclinación respecto al suelo, ver Figura 1.10. La trayectoria depende de muchos factores:

- características del cohete y su propulsor
- el arrastre producido por la densidad del aire

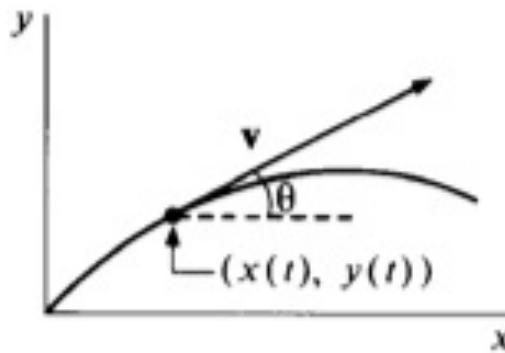


Figura 1.10: Trayectoria típica de un objeto

- la fuerza de la gravedad
- efectos del viento
- etc...

Para poder realizar un modelo de este problema debemos simplificar la realidad. Si suponemos que el cohete se mueve en un plano, es decir, no hay viento cruzado, y su posición viene descrita por las coordenadas $x(t)$ e $y(t)$ nos encontramos con unas ecuaciones obtenidas de la segunda ley de Newton $F = ma$ dadas por

$$\begin{cases} x''(t) = \frac{1}{m(t)}(T(t) - 0.5 c\rho s v(t)^2) \cos(\theta(t)) - \frac{m'(t)}{m(t)}x'(t), \\ y''(t) = \frac{1}{m(t)}(T(t) - 0.5 c\rho s v(t)^2) \sin(\theta(t)) - \frac{m'(t)}{m(t)}y'(t) - g. \end{cases}$$

Aquí tenemos

- $v(t) = (x'(t)^2 + y'(t)^2)^{1/2}$ es el módulo de la velocidad del cohete
- $\theta(t) = \tan^{-1}(y'(t)/x'(t))$ es el ángulo del vector velocidad del cohete con la horizontal, ver Figura 1.10.
- $m(t)$ la masa del cohete, que cambia conforme se consume el combustible
- $T(t)$ el empuje producido por el propulsor
- $0.5c\rho s v(t)^2$ es la fuerza de arrastre sobre el cohete. Aquí ρ es la densidad del aire, s es la sección del cohete y c es un coeficiente de arrastre.
- etc...

Aquí usamos las variables $(x(t), y(t), v(t), \theta(t))$ de acuerdo a las expresiones

$$x'(t) = v(t) \cos(\theta(t)), \quad y'(t) = v(t) \sin(\theta(t))$$

derivando y sustituyendo en las ecuaciones para $x''(t)$ e $y''(t)$ se obtiene un sistema de ecuaciones de primer orden dado por

$$\begin{cases} x'(t) &= v(t) \cos(\theta(t)), \\ y'(t) &= v(t) \sin(\theta(t)), \\ v'(t) &= \frac{1}{m(t)}(T(t) - 0.5 c\rho s v(t)^2) - g \sin(\theta(t)) - \frac{m'(t)}{m(t)}v(t), \\ \theta'(t) &= \frac{-g}{v(t)} \cos(\theta(t)). \end{cases}$$

Como se puede observar, hay una gran dificultad intrínseca en resolver este tipo de ecuaciones. En el caso de un **proyectil lanzado por un cañón** nos encontramos con una simplificación puesto que podemos suponer que la masa del proyectil es constante y que no hay empuje por el propulsor. Entonces las ecuaciones se simplifican como

$$\begin{cases} x''(t) &= \frac{-c\rho s v(t)^2}{2m} \cos(\theta(t)), \\ y''(t) &= \frac{-c\rho s v(t)^2}{2m} \sin(\theta(t)) - g. \end{cases}$$

Siendo posible resolverse en algunas situaciones.

1.8. Ejemplos (III): Problemas rígidos e inestables

Muchos sistemas presentan soluciones o partes de las soluciones (por ejemplo, si la solución es una suma de funciones podemos estar hablando de uno de los sumandos) que tienen un comportamiento muy brusco con respecto al resto de la solución.

- Si la solución, o alguna componente suya, **desaparece o tiende a cero** con mucha más rapidez que el resto, o manifiesta **un cambio brusco de comportamiento** decimos que el problema es **rígido**.
- Si la solución, o alguna componente suya, **crece en magnitud** con más rapidez que el resto y de una forma importante decimos que el problema es **inestable**.

Esto es una descripción, o definición, poco matemática pero aquí no se puede precisar más debido a la subjetividad del fenómeno.

1.8.1. Ejemplos en sistemas

Ejemplo 16 Consideremos el sistema lineal

$$\begin{cases} x'_1(t) &= -298x_1(t) + 99x_2(t), \\ x'_2(t) &= -594x_1(t) + 197x_2(t). \end{cases}$$

La solución exacta es

$$\vec{x}(t) = c_1 e^{-t} \begin{pmatrix} 1 \\ 3 \end{pmatrix} + c_2 e^{-100t} \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

donde c_1, c_2 se fijan con el dato inicial. Esto es, tenemos dos partes en la solución, la correspondiente a e^{-t} y la correspondiente a e^{-100t} y la solución se escribe como

$$\vec{x}(t) = \vec{w}_1 e^{-t} + \vec{w}_2 e^{-100t}$$

para dos vectores $\vec{w}_1, \vec{w}_2 \in \mathbb{R}^2$. La parte con la función coeficiente e^{-100t} decrece mucho más rápido que la parte asociada a e^{-t} .

Podemos describir esta situación en sistemas lineales de forma general. Supongamos que tenemos el sistema

$$x'(t) = Ax(t) + \varphi(t) \in \mathbb{R}^m$$

y A tiene espectro $\sigma(A) = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$ tal que $\operatorname{Re}(\lambda_j) < 0$ para todo $j = 1, 2, \dots, m$. Entonces la solución se puede escribir como

$$\vec{x}(t) = \sum_j \alpha_j e^{\lambda_j t} \vec{c}_j + \vec{\Psi}(t)$$

donde $\vec{\Psi}(t)$ es una solución particular del problema. Si $\operatorname{Re}(\lambda_j) < 0$ se tiene

$$e^{\lambda_j t} \vec{c}_j \rightarrow 0, \quad t \rightarrow +\infty$$

y por lo tanto

$$\vec{x}(t) \sim \vec{\Psi}(t), \quad t \rightarrow +\infty.$$

A cada uno de los términos $e^{\lambda_j t} \vec{c}_j$ se le suele denominar **modo transitorio (lento o rápido)** y al término $\vec{\Psi}(t)$ **modo estacionario**.

Si $\operatorname{Re}(\lambda_j) < 0$ pero además $|\operatorname{Re}(\lambda_j)| \gg 1$ entonces el modo asociado decae muy rápido, es un **modo transitorio rápido** mientras que si $|\operatorname{Re}(\lambda_j)| \sim 1$ el modo asociado decae más lentamente y es un **modo transitorio lento**.

Si $\underline{\lambda}$ es el autovalor con la parte real negativa más pequeña en valor absoluto y $\bar{\lambda}$ es el autovalor con la parte real negativa más grande en valor absoluto, entonces el modo asociado a $\bar{\lambda}$ es el más rápido y el asociado a $\underline{\lambda}$ el más lento. El cociente

$$\frac{\operatorname{Re}(\bar{\lambda})}{\operatorname{Re}(\underline{\lambda})}$$

*se denomina **razón de rigidez** de la matriz A. Las dificultades prácticas surgen cuando esta razón es muy grande puesto que se hace muy complicado seguir y calcular de forma correcta todos los modos al mismo tiempo. Los que decaen más rápido requieren, en general, más esfuerzo que los que decaen más lento, aunque estos sean los que realmente puedan importar.*

A parte de este ejemplo académico, existen muchas situaciones de gran importancia física, como por ejemplo **las reacciones químicas**, donde diferentes compuestos se transforman de acuerdo a diferentes escalas temporales, con frecuencia con distintos órdenes de magnitud, pensemos en explosiones o el encendido de una cerilla junto con la difusión de compuestos químicos restantes en el explosivo.

Estos problemas empezaron a plantearse en la década de los 60 con el empuje en el modelado y cálculo efectivo en fenómenos físicos, químicos como reacciones químicas, dinámica de guiado de misiles, circuitos electrónicos, biomatemáticas, etc....Este empuje fue grandemente potenciado por el desarrollo de los computadores o, incluso se puede pensar que, el desarrollo de las capacidades de computación ayudaron enormemente a este empuje.

Este tipo de problemas se denominan **problemas rígidos** (stiff, en inglés, râides en francés) y son muy importantes por su presencia predominante en cualquier campo de la ciencia. Dos químicos, Curtis y Hirschfeld de la Universidad de Wisconsin publicaron en 1952 un trabajo en donde ya estudiaban este fenómeno desde el punto de vista computacional.

...En torno a 1960 las cosas cambiaron y todo el mundo se dió cuenta
que el mundo estaba lleno de problema rígidos.
(G. Dahlquist 1985)

Una idea clara encontramos en la siguiente descripción

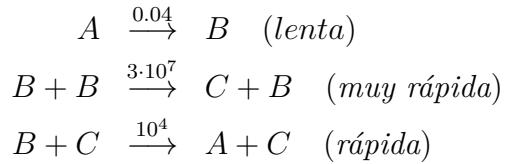
La esencia de la rigidez consiste en que la solución buscada varía despacio pero existen perturbaciones que decaen o cambian rápidamente. La presencia de estas perturbaciones complican el cálculo efectivo de la solución buscada.

Las componentes de la solución que cambian más lentamente se suelen denominar **modos estacionarios o transitivos lentos** en el sentido de que sus derivadas son mucho más pequeñas que las derivadas de los **modos transitivos rápidos**, que son esas componentes de la solución con decaimiento rápido. La dificultad práctica reside en que nos encontramos forzados a describirlos o calcularlos todas al mismo tiempo.

Ejemplo 17 Modelo de Robertson

La cinética química estudia la evolución de la concentración de sustancias químicas sometidas a diversas reacciones. El modelo de Robertson (1966) ha llegado a ser muy

popular en estudios numéricos y describe la reacción química entre tres compuestos con distintas velocidades de reacción



los números encima de las flechas indican la velocidad a la que se produce cada reacción. Este modelo lleva a las ecuaciones para las concentraciones dadas por

$$\begin{aligned} x'(t) &= -k_1 x(t) + k_2 y(t)z(t), \\ y'(t) &= k_1 x(t) - k_2 y(t)z(t) - k_3 y(t)^2, \\ z'(t) &= k_3 y(t)^2. \end{aligned}$$

Aquí $x(t)$, $y(t)$, $z(t)$ son las concentraciones, $k_1 = 0.04$, $k_2 = 10^4$, $k_3 = 3 \cdot 10^7$ son las velocidades de reacción lenta, rápida y muy rápida y los valores iniciales son $x(0) = 1$, $y(0) = 0$, $z(0) = 0$. La función pendiente es

$$f(x, y, z) = (-k_1 x + k_2 y z, k_1 x - k_2 y z - k_3 y^2, k_3 y^2)$$

luego el único punto crítico es $(x(t), y(t), z(t)) \equiv (1, 0, 0)$ ya que si $z(t)$ es constante entonces $y(t)$, $x(t)$ también lo son, esto es, el estado estacionario es $(x', y', z') = 0$ y de acuerdo al dato inicial debe ser $(x(t), y(t), z(t)) \equiv (1, 0, 0)$, $\forall t > 0$.

Por otro lado, se conserva la materia en la reacción química, y eso se ve porque la suma de las tres magnitudes se mantiene siempre constante igual a uno

$$x(t) + y(t) + z(t) = 1$$

y además, se ve también que $x(+\infty) = z(+\infty) = 0$ siendo por lo tanto $y(+\infty) = 1$, ver Figura 1.11. El estado estacionario se alcanza después de un largo periodo de tiempo, como del orden de 10^6 . La concentración $y(t)$ alcanza rápidamente un estado casi estacionario en un entorno de $y'(t) = 0$ el cual en principio, usando que $x(0) = 1$ y $z(0) = 0$ se puede estimar mediante la aproximación $0.04 \approx 3 \cdot 10^7 y(t)^2$ que da $y_2(t) \approx 3.65 \cdot 10^{-5}$, y entonces lentamente va a cero otra vez, ver Figura 1.12.

Como se puede ver por la Figuras 1.11 y 1.12, las magnitudes son muy distintas entre las tres concentraciones y el comportamiento inicial de la componente y presenta un salto brusco desde cero, análogo a la creación de una llama. Nos encontramos entonces con tres modos, uno lento, otro rápido y otro muy rápido.

Es difícil ahora enmarcar todas las posibles situaciones prácticas dentro de una misma definición, pero sí que se pueden hacer intentos de definición que capturan la idea de lo que pasa

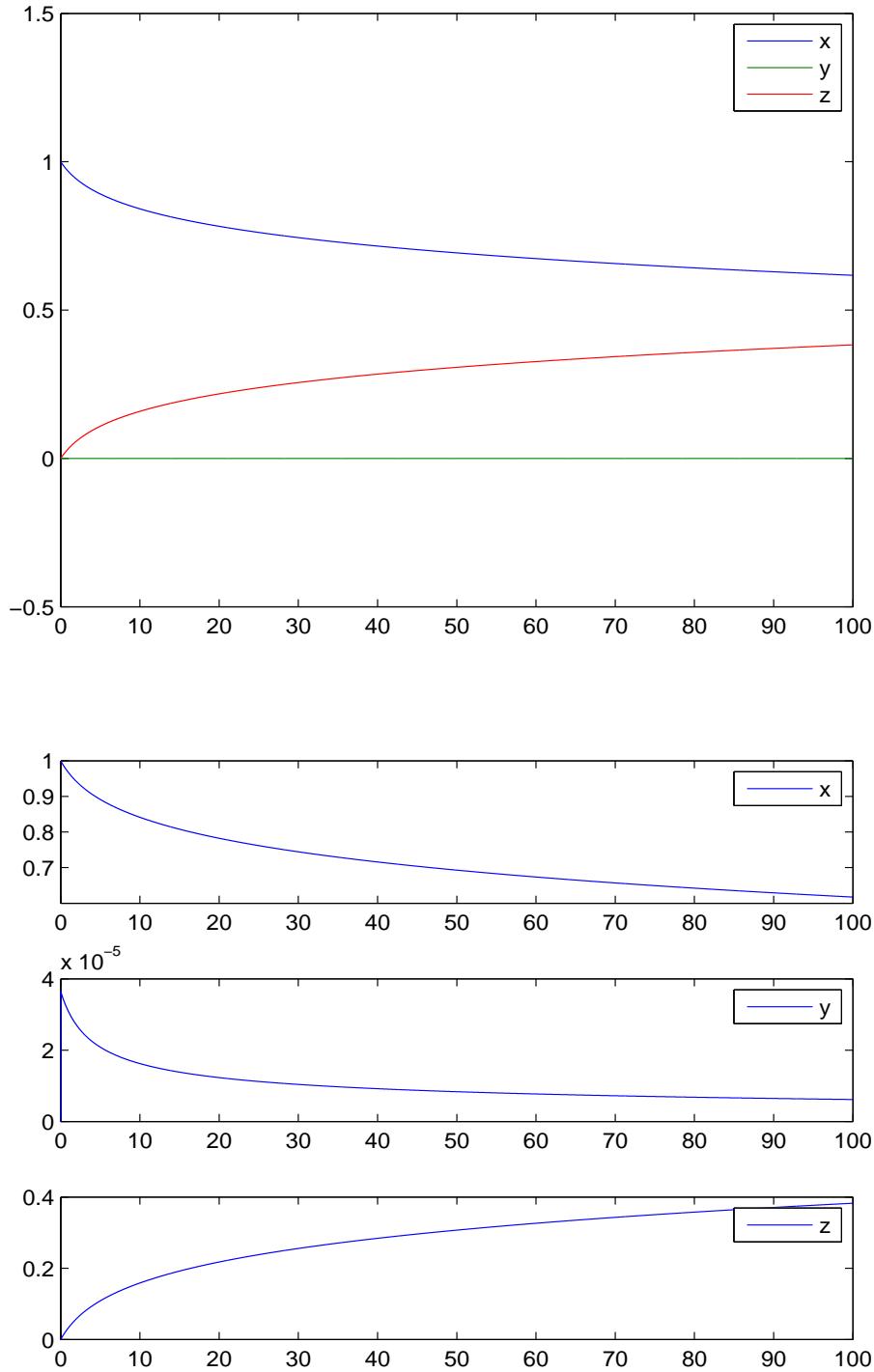


Figura 1.11: Modelo de Robertson (1963) para $T = 100$. Visualización de las tres soluciones juntas y por separado. Observar las diferencias en escala.

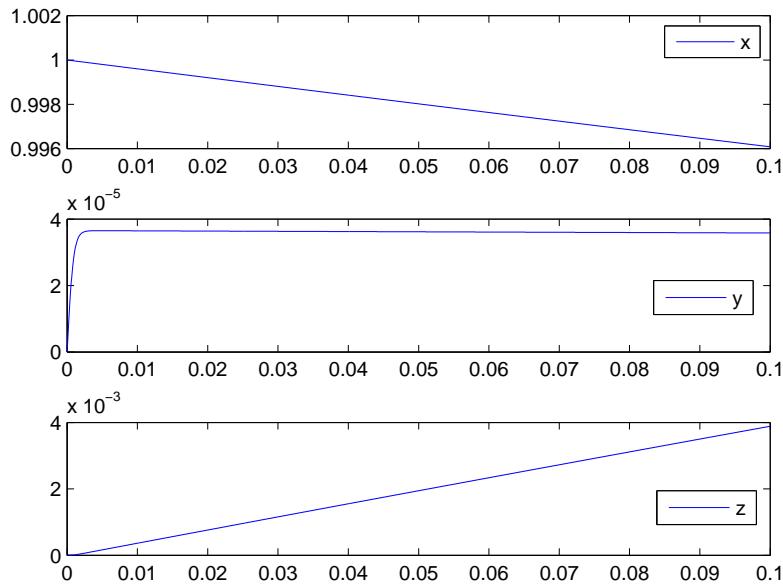


Figura 1.12: Modelo de Robertson (1963). Comportamientos iniciales de las concentraciones.

El sistema homogéneo se dice **rígido** si la razón de rigidez es muy grande.

Muy grande... ¿con respecto a qué?

El sistema no homogéneo se dice **rígido** si la razón de rigidez es muy grande cuando se compara con la razón de cambio del modo estacionario.

Se puede dar otras definiciones que, en el fondo, lo que hacen es caracterizar distintas propiedades que pueden estar presentes o no. Nos quedamos con esta idea:

Definición 18 *El sistema se dice rígido en un intervalo de tiempo si en este intervalo las curvas solución vecinas a la curva buscada se aproximan a esta a una velocidad muy grande en comparación a la velocidad con la que la curva buscada cambia en este intervalo.*

1.8.2. Ejemplos escalares

Los problemas escalares también poseen soluciones con distintos modos.

Ejemplo 19 *La ecuación de Dahlquist-Bjorck (1956 aprox.) es un ejemplo académico del efecto de la rigidez. El modelo es*

$$\begin{cases} y'(t) = 100(\sin(t) - y(t)), & 0 < t \leq 3, \\ y(0) = 0 \end{cases}$$

y se puede ver como un caso particular del problema

$$\begin{cases} y'(t) = a(\sin(t) - y(t)), & t > 0, \\ y(0) = y_0 \end{cases}$$

cuya solución es

$$y(t) = e^{-at}y_0 + \frac{\sin(t) - a^{-1}\cos(t) + a^{-1}e^{-at}}{1 + a^{-2}}.$$

Por lo tanto, para $a \gg 1$ el modo transitorio exponencial asociado con e^{-at} decae muy rápido. Aquí la solución se descompone en

$$\text{modo rápido} \sim e^{-at}y_0 + \frac{a^{-1}e^{-at}}{1 + a^{-2}}.$$

$$\text{modo estacionario o lento} \sim \frac{\sin(t) - a^{-1}\cos(t)}{1 + a^{-2}}.$$

En las Figuras 1.13, 1.14 y 1.15 se ve el efecto de incrementar el valor de a en el decaimiento de las distintas soluciones. Este modo debe ser capturado bien por cualquier proceso de cálculo que se precie de funcionar. Aquí se tiene

$$f(t, y) = -ay + a\sin(t) \Rightarrow \partial_y f(t, y) = -a.$$

Se tiene el mismo resultado cualitativo con el ejemplo más simple

$$\begin{cases} y'(t) = a(1 - y(t)), & 0 < t, \\ y(0) = y_0 \end{cases}$$

donde la solución exacta es $y(t) = e^{-at}(y_0 - 1) + 1$. Esto quiere decir que la dificultad fundamental la provoca el término $-ay$ en la función pendiente, o lo que es lo mismo, tener $\partial_y f(t, y) \ll 0$, módulo grande con signo negativo. Observar que tenemos

$$f(t, y) = -a(\sin(t) - y), \quad f(t, y) = -a(1 - y)$$

luego en ambos casos es

$$\partial_y f(t, y) = -a y.$$

Observación 15 Recordar que una misma curva $y(t) \equiv 1$ puede estar en un campo de vectores muy simple como el dado por la ecuación $y'(t) = 0$ o en otros mucho mas complicados como $y'(t) = a(1 - y(t))$, revisar la Figura 1.6 por su interés.

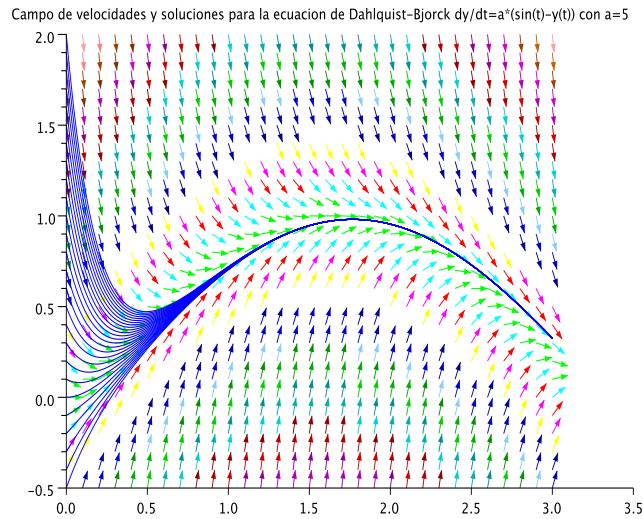


Figura 1.13: Soluciones y campo de velocidades de la ecuación de Dahlquist-Bjorck para $a = 5$. Los valores iniciales en $t_0 = 0$ son $y_0 = -0.5 : 0.1 : 2$ (notación a:h:b={a, a+h,a+2h,...}).

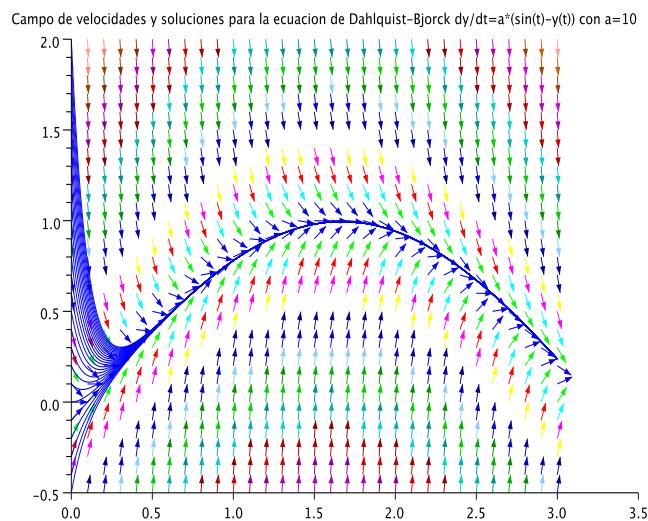


Figura 1.14: Soluciones y campo de velocidades de la ecuación de Dahlquist-Bjorck para $a = 10$. Los valores iniciales en $t_0 = 0$ son $y_0 = -0.5 : 0.1 : 2$.

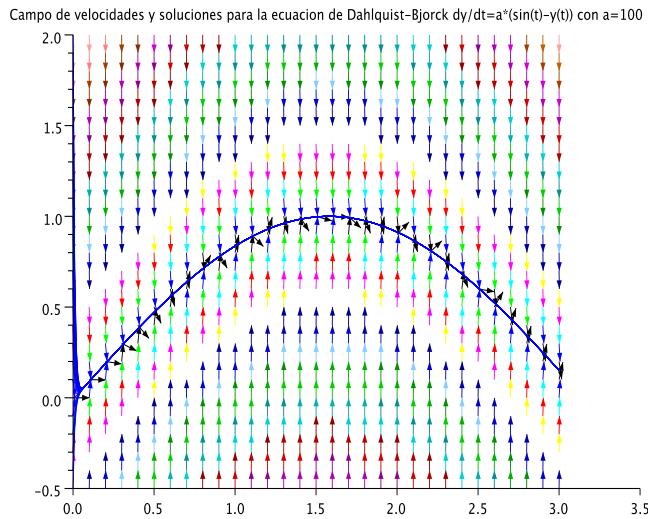


Figura 1.15: Soluciones y campo de velocidades de la ecuación de Dahlquist-Bjorck para $a = 100$. Los valores iniciales en $t_0 = 0$ son $y_0 = -0.5 : 0.1 : 2$.

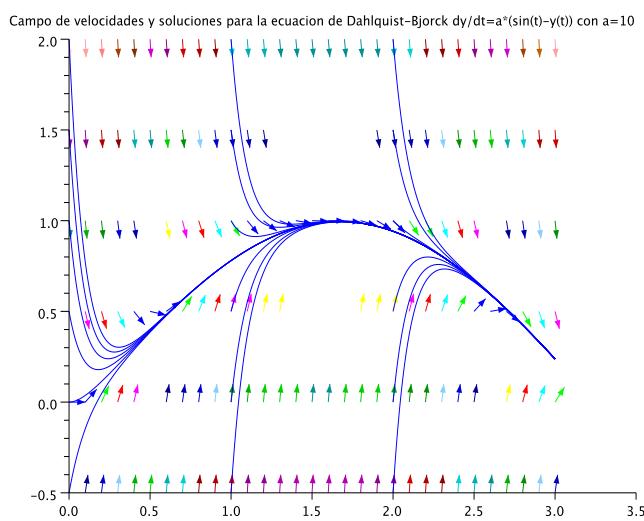


Figura 1.16: Soluciones y campo de velocidades de la ecuación de Dahlquist-Bjorck para $a = 10$ y distintos tiempos iniciales. Los valores iniciales en $t_0 = 0$, $t_0 = 1$ y $t_0 = 2$ son $y_0 = -0.5 : 0.5 : 2$.

Ejemplo 20 *Un caso similar es la ecuación de Prothero-Robinson*

$$\begin{cases} y'(t) &= L(\varphi(t) - y(t)) + \varphi'(t), \quad 0 < t, \\ y(0) &= y_0 \end{cases}$$

con solución exacta

$$y(t) = e^{-Lt}(y_0 - \varphi(0)) + \varphi(t)$$

y otra vez el modo rápido e^{-Lt} para $L \gg 1$ debe ser capturado correctamente. El modo estacionario es $\varphi(t)$ y puede ser una función suave sin cambios bruscos, como por ejemplo $\varphi(t) = \sin(t)$. También se tiene

$$f(t, y) = -Ly + L\varphi(t) + \varphi'(t) \Rightarrow \partial_y f(t, y) = -L.$$

Incluso en el caso en el que $y_0 = \varphi(0)$ donde aparentemente el modo rápido no está en la solución, éste si que se encuentra en todas las soluciones vecinas y se debe también tener en cuenta para conseguir que no aparezca cuando se haga un cálculo aproximado del modo estacionario $\varphi(t)$.

Ejemplo 21 Ecuación de Van der Pol *Este modelo aparece en la teoría de circuitos eléctricos*

$$\begin{cases} x''(t) - \mu(1 - x(t)^2)x'(t) + x(t) &= 0, \quad t > 0, \\ x(0) = 1, \quad x'(0) &= 0. \end{cases}$$

Una vez más, la ecuación se puede reescribir en términos de un sistema de primer orden usando las incógnitas $u(t) = x(t)$ y $v(t) = x'(t)$:

$$\begin{aligned} u'(t) &= v, & t > 0, \\ v'(t) &= \mu(1 - u^2)v - u, & t > 0, \end{aligned}$$

con $(u(0), v(0)) = (1, 0)$. Por ejemplo, en el caso $\mu = 6$ observa la gráfica de $x(t)$ en la Figura 1.17. Aquí hay cambios bruscos también aunque el sistema es más difícil de predecir.

Incluso cuando no podemos comparar dos componentes de la solución puesto que sólo tenemos una, también nos podemos encontrar con rigidez. Por ejemplo,

$$\begin{cases} y'(t) &= -ay(t), \quad 0 < t, \\ y(0) &= 1 \end{cases}$$

también genera un comportamiento rígido si $a \gg 1$. ¿Con respecto a qué comparamos aquí? Pues probablemente con respecto a la solución de la ecuación $y'(t) = 0$ que es la función constante $y(t) \equiv 1$.

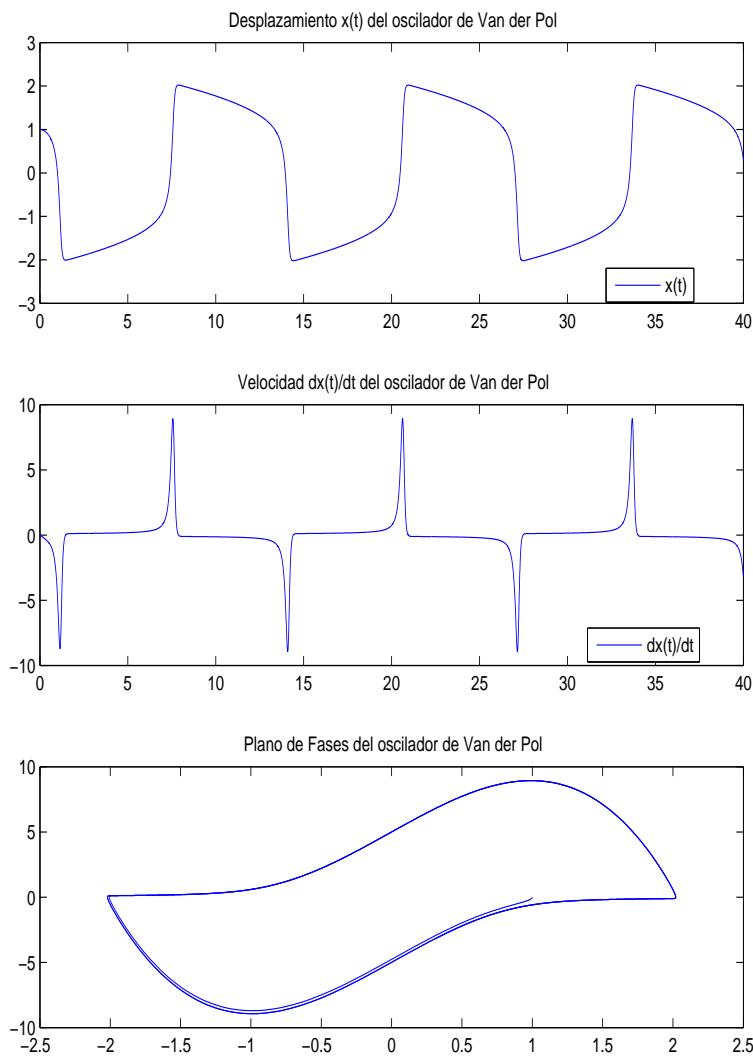


Figura 1.17: Solución ecuación de Van der Pol para $\mu = 6$, $x(0) = 1$ y $x'(0) = 0$.

Influencia de la rigidez en datos iniciales

Aunque no tiene por que siempre ser así, podemos observar en los ejemplos previos que los cambios en los datos iniciales son rápidamente anulados si están afectados por los modos rápidos:

Ejemplo 22 En los problemas

$$\begin{cases} y'(t) = -a y(t), & 0 < t, \\ y(0) = y_0 \end{cases}$$

donde

$$y(t) = y_0 e^{-at}$$

o la ecuación de Dahlquist-Bjorck

$$\begin{cases} y'(t) = a(\sin(t) - y(t)), & t > 0, \\ y(0) = y_0 \end{cases}$$

cuya solución es

$$y(t) = e^{-at} y_0 + \frac{\sin(t) - a^{-1} \cos(t) + a^{-1} e^{-at}}{1 + a^{-2}}$$

o la ecuación de Prothero-Robinson

$$\begin{cases} y'(t) = L(\varphi(t) - y(t)) + \varphi'(t), & 0 < t, \\ y(0) = y_0 \end{cases}$$

con solución exacta

$$y(t) = e^{-Lt}(y_0 - \varphi(0)) + \varphi(t).$$

En todos estos casos se ve el valor inicial está afectado por el modo rápido de decaimiento e^{-Lt} para $L \gg 1$, luego si $y_0 \neq 0$ o $y_0 \neq \varphi(0)$ la contribución del valor inicial y_0 desaparece rápidamente.

1.8.3. Problemas inestables

Como ya hemos dicho, podemos entender por problemas inestables aquellos problemas donde ligeras variaciones iniciales producen una gran diferencia en las soluciones en un corto periodo de tiempo. También se pueden interpretar como problemas donde hay **modos o componentes de la solución que aumentan muy rápido**.

Ejemplo 23 Los ejemplos vistos de Dahlquist-Bjorck o de Prothero-Robinson pasan de ser rígidos a inestables si tomamos los coeficientes a y L negativos. En este caso $\partial_y f \gg 1$ para valores grandes, lo que provoca la inestabilidad.

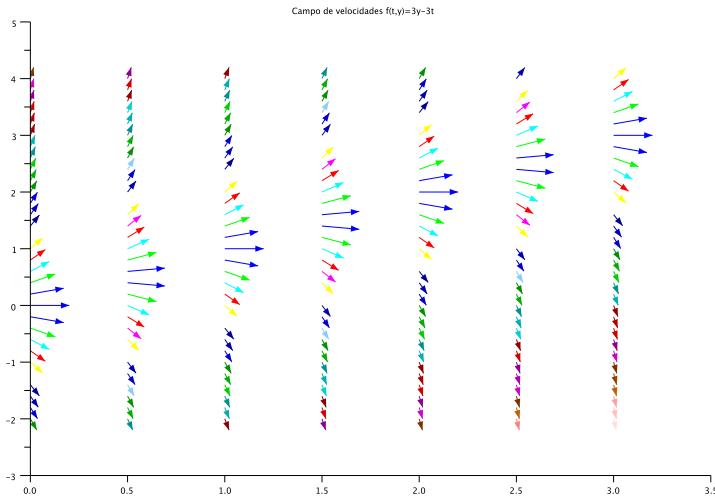


Figura 1.18: Campo repulsivo desde la solución $y_{1/3}(t) = t + 1/3$.

Ejemplo 24 La ecuación

$$\begin{cases} \frac{dy}{dt} = 3y - 3t, & t > 0, \\ y(0) = \alpha, \end{cases}$$

tiene por solución

$$y_\alpha(t) = (\alpha - 1/3)e^{3t} + t + 1/3$$

que se descompone en

$$\text{modo rápido } \sim (\alpha - 1/3)e^{3t}.$$

$$\text{modo estacionario o lento } \sim t + 1/3.$$

En el caso exacto de $\alpha = 1/3$ tenemos que

$$y_{1/3}(t) = t + 1/3$$

algo muy razonable. Pero si $\alpha - 1/3 \neq 0$ y pequeño tendremos

$$y_\alpha(t) = (\alpha - 1/3)e^{3t} + t + 1/3$$

y el factor e^{3t} amplifica de forma muy importante el valor $\alpha - 1/3$, ver Figura 1.18 y Figura 1.19. La situación cambia totalmente si $f(t, y) = -3y - 3t$, aquí el problema pasa de ser inestable a ser rígido:

$$\begin{cases} \frac{dz}{dt} = -3z - 3t, & t > 0, \\ z(0) = \alpha, \end{cases}$$

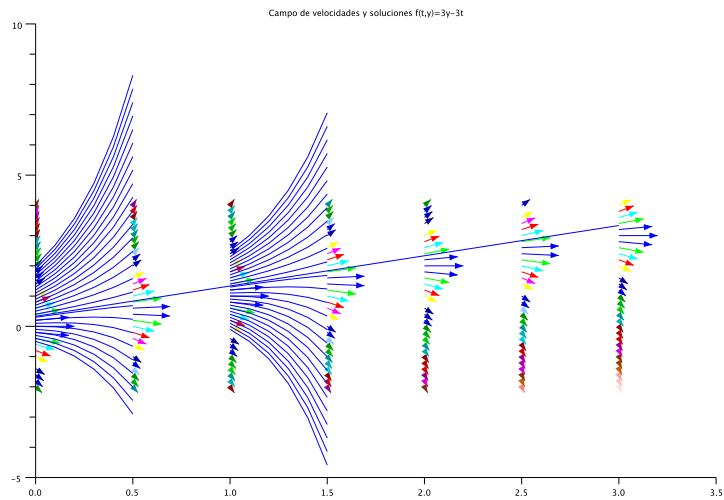


Figura 1.19: Campo repulsivo y soluciones incluyendo la solución $z_{1/3}(t) = t + 1/3$ con datos iniciales empezando en $t = 0$ y en $t = 1$. Se observa como se alejan de $z_{1/3}(t)$.

tiene por solución

$$z_\alpha(t) = (\alpha - 1/3)e^{-3t} - t + 1/3$$

que se descompone en

$$\text{modo rápido } \sim (\alpha - 1/3)e^{-3t}.$$

$$\text{modo estacionario o lento } \sim -t + 1/3.$$

Otra vez, en el caso exacto de $\alpha = 1/3$ tenemos que

$$z_{1/3}(t) = -t + 1/3.$$

Si $\alpha - 1/3 \neq 0$ pequeño o grande tendremos

$$z_\alpha(t) = (\alpha - 1/3)e^{-3t} + t + 1/3$$

y ahora el factor e^{-3t} reduce de forma muy rápida el valor de $\alpha - 1/3$, ver Figura 1.20. En el primer caso la trayectoria

$$y_{1/3}(t) = t + 1/3$$

es **inestable**, repele todas las trayectorias vecinas, mientras que en el segundo caso

$$z_{1/3}(t) = -t + 1/3$$

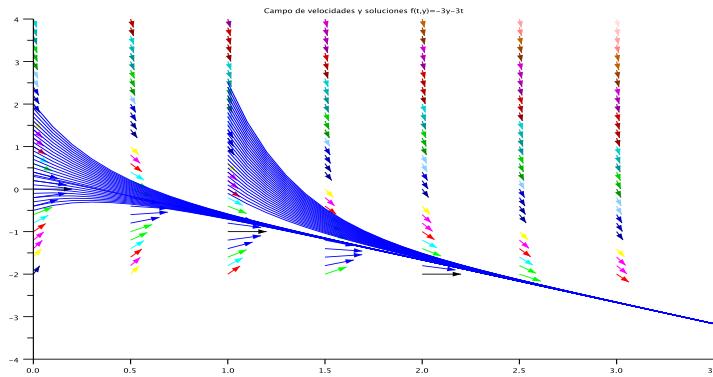


Figura 1.20: Campo contractivo de velocidades y soluciones hacia $z_{1/3}(t) = -t + 1/3$. Usando datos iniciales empezando en $t = 0$ y en $t = 1$ se observa como se contraen hacia de $z_{1/3}(t)$.

es estable o atrae a todas las trayectorias. En estos dos casos tenemos

$$f(t, y) = \pm 3y - 3t$$

y por lo tanto la constante de Lipschitz asociada es $L_f = 3$. Esta constante no distingue entre una situación estable y otra inestable, cosa que si que hace el signo de $\partial_y f(t, y)$.

Observación 16 Se puede interpretar que cuando partes de la solución tienen una estabilidad muy fuerte llegamos a la **rigidez**. También aparece un crecimiento muy fuerte de las derivadas de orden superior de la solución.

Ejemplo 25 La ecuación (Golub-Ortega [12])

$$y'' - 10y' - 11y = 0, \quad y(0) = 1, \quad y'(0) = -1$$

posee solución $y(t) = e^{-t}$. Pero si usamos como dato inicial

$$y(0) = 1 + \epsilon, \quad y'(0) = -1$$

la solución es

$$y_\epsilon(t) = \left(1 + \frac{11}{12}\epsilon\right)e^{-t} + \frac{\epsilon}{12}e^{11t}$$

entonces para cualquier $\epsilon > 0$ independientemente de lo pequeño que sea tendremos un crecimiento muy fuerte de $y_\epsilon(t)$ para $t > 0$. Tenemos aquí dos modos muy dispares en la solución. Como consecuencia, pequeñas perturbaciones generan grandes

cambios a tiempo finito y resulta que $y(t) = e^{-t}$ se puede considerar como una solución inestable.

También podemos escribir este problema en términos de un sistema de dos ecuaciones lineales usando $y_1 = y$ e $y_2 = y'$ para obtener

$$\begin{cases} y'_1 &= y_2, \\ y'_2 &= 11y_1 + 10y_2 \end{cases}$$

o bien en forma matricial

$$\frac{d}{dt}\vec{y}(t) = A\vec{y}(t)$$

donde

$$A = \begin{pmatrix} 0 & 1 \\ 11 & 10 \end{pmatrix}$$

y entonces nos encontramos con autovalores para A dados por $\sigma(A) = \{-1, 11\}$ luego la solución es una combinación lineal de $\{e^{-t}, e^{11t}\}$. Tenemos entonces dos modos, uno dado por e^{-t} que decrece y otro modo dado por e^{11t} que crece. La comparación entre ambos da el modo lento y el modo rápido.

Ejemplo 26 El siguiente ejemplo (Golub-Ortega [12]) también es interesante ya que muestra una inestabilidad más fuerte

$$y' = t y (y - 2), \quad y(0) = y_0.$$

Se tiene que $f(t, y) = ty^2 - 2ty$ luego $f(t, y) \sim ty^2$ para valores grandes de t e y por lo que no se cumple la condición de crecimiento, a lo sumo lineal, de la función pendiente con respecto a y . Aquí $y(t) \equiv 0$ e $y(t) \equiv 2$ son puntos estacionarios y la solución general es

$$y(t) = \frac{2y_0}{y_0 + (2 - y_0)e^{t^2}}.$$

Si $y_0 < 2$ entonces $y(t) \rightarrow 0$ pero si $y_0 > 2$ entonces hay explosión en tiempo finito de la solución, ver Figura 1.21 y Figura 1.22

1.9. Ejemplos (IV): Problemas oscilatorios

Aunque la mayoría de sistemas físicos exhiben algún **nivel de disipación**, esto es, pérdida de masa, información etc...hay otros en donde este efecto no es el importante, o simplemente no es así. Estamos hablando, por ejemplo, de la **transmisión de señales sin pérdida de información, movimiento de planetas, satélites, etc....** Las dificultades principales en este tipo de problemas son dos:

- se debe de calcular en un intervalo de tiempo muy grande, y ...

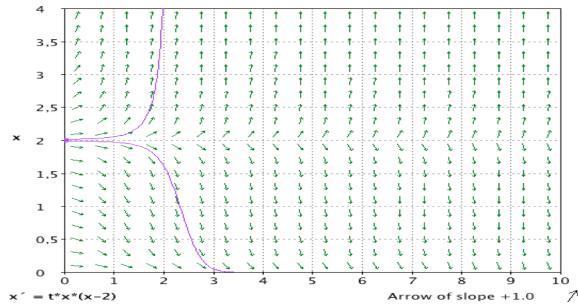


Figura 1.21: Dos soluciones inicialmente muy cercanas, una de ellas tiende a infinito en tiempo finito mientras que la otra se estabiliza en cero.

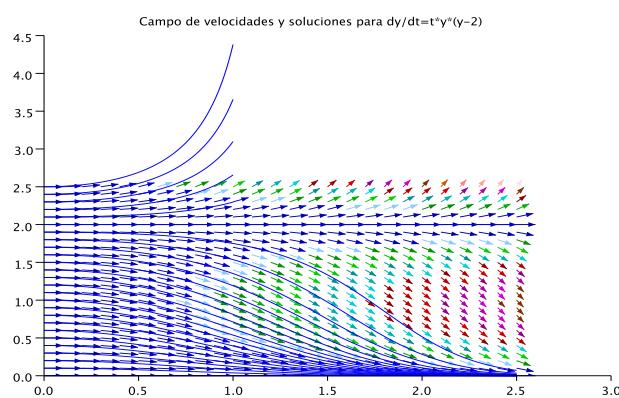


Figura 1.22: Soluciones y campo de velocidades del modelo $dy/dt = t^2 * y * (y - 2)$

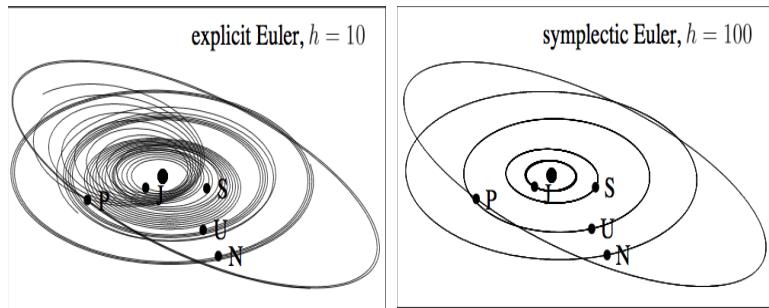


Figura 1.23: Trayectorias calculadas con un método clásico y con un método simpléctico. Las trayectorias cerradas no se calculan bien con el método clásico de Euler.

- se debe conseguir mantener las trayectorias cerradas.

Los métodos que se adaptan bien a estas dificultades se llaman **simplécticos** (simpléctico significa preservar área).

Ejemplo 27 Movimiento armónico simple La ecuación linealizada que rige el movimiento de un péndulo (usando su ángulo con respecto a la vertical) es

$$\begin{cases} \theta''(t) + \theta(t) = 0, & t > 0, \\ \theta(0) = 0, \quad \theta'(0) = \theta_0. \end{cases}$$

La solución general es una combinación lineal de funciones trigonométricas y por lo tanto oscila.

Ejemplo 28 Cálculo de trayectorias (III): Sistema planetario

El problema de N cuerpos es importante en campos tan dispares como astronomía y biología molecular (movimientos de moléculas). Las ecuaciones son

$$y_i'' = G \sum_{j \neq i} m_j \frac{y_j - y_i}{\|y_j - y_i\|^3}$$

donde $y_i(t) \in \mathbb{R}^3$ describe la trayectoria del cuerpo i -ésimo, m_i es su masa y G la constante gravitacional. Introduciendo la velocidad $v_i(t) = y'_i(t)$ como nueva variable se obtiene un sistema de primer orden de dimensión $6N$ siendo N el número de cuerpos. Si cada cuerpo se mueve en un mundo 3d entonces el número de ecuaciones es $18N$ al ser cada una de ellas de tres componentes.

La Figura 1.23 muestra la diferencia de resultados cuando se usa un método clásico no eficiente para resolver el problema (método no simpléctico) y cuando se usa uno bien adaptado a este tipo de problemas (método simpléctico).

En el caso de una partícula que orbita con respecto al origen, tenemos

$$\begin{cases} r(0) = r_0, \\ r'(0) = v_0, \\ r''(t) = -r(t)/|r(t)|^3, \quad t > 0. \end{cases}$$

Si se escribe en términos de las incógnitas $r = (y_1, y_2)$ y $r' = (y_3, y_4)$ (aquí y_3 e y_4 son las velocidades), nos encontramos con el sistema

$$\begin{cases} y'_1 = y_3, \\ y'_2 = y_4, \\ y'_3 = -y_1/(y_1^2 + y_2^2)^{3/2}, \\ y'_4 = -y_2/(y_1^2 + y_2^2)^{3/2}. \end{cases}$$

1.10. Sistemas, linealización y uso de variable compleja

Sabemos que dado el problema escalar $y'(t) = f(t, y(t))$ no lineal una herramienta de gran importancia para conocer el comportamiento de una solución es la linealización que se realiza usando el desarrollo de Taylor sobre $f(t, y)$ y que la parte de la variable t no influye en el comportamiento, es decir, al linealizar $y'(t) = a(t)y(t) + b(t)$ tiene el mismo comportamiento que $y'(t) = a(t)y(t)$ y viene determinado por $a(t) = \partial_y f(t, y)$ localmente.

Esta idea vista anteriormente para la estabilidad de ecuaciones de acuerdo a su naturaleza disipativa o contractiva en el caso $m = 1$ **no se extiende de manera trivial al caso $m > 1$ ya que el Teorema del Valor Medio se debe de aplicar a cada componente en el caso vectorial.** En el caso de un sistema de m ecuaciones diferenciales no lineales tenemos

$$\begin{cases} z'_1(t) = f_1(t, z_1(t), z_2(t), \dots, z_m(t)), \\ z'_2(t) = f_2(t, z_1(t), z_2(t), \dots, z_m(t)), \\ \vdots \quad \vdots \quad \vdots \\ z'_{m-1}(t) = f_{m-1}(t, z_1(t), z_2(t), \dots, z_m(t)), \\ z'_m(t) = f_m(t, z_1(t), z_2(t), \dots, z_m(t)) \end{cases}$$

y se puede escribir en forma vectorial como

$$z'(t) = F(t, z(t)), \quad z(0) = z_0$$

y la linealización con respecto a un punto fijo nos produce la matriz de los gradientes evaluados en puntos distintos para cada ecuación:

$$\nabla F = \begin{pmatrix} \nabla_y f_1(\xi_1) \\ \nabla_y f_2(\xi_2) \\ \vdots \\ \nabla_y f_m(\xi_m) \end{pmatrix}$$

lo que nos lleva a una ecuación lineal de la forma

$$Y'(t) = A Y(t) + b(t), \quad Y(0) = Y_0$$

donde $A = \nabla F \in \mathbb{R}^{m \times m}$. Simplificando la ecuación, si consideramos $y' = A y$ y suponemos que **A es diagonalizable** obtenemos fácilmente la forma de la solución. Si $\Lambda = V^{-1}AV$ con $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_m\}$ y estamos admitiendo que $V \in \mathbb{R}^{m \times m}$; usando $z = V^{-1}y$ llegamos al problema desacoplado

$$\begin{cases} z'(t) &= \Lambda z(t), \quad t > 0 \Leftrightarrow z'_j(t) = \lambda_j z_j(t), \quad j = 1, 2, \dots, m \\ z(0) &= V^{-1} y_0 = z_0 \end{cases}$$

y deshaciendo el cambio de variables

$$\vec{y}(t) = \sum_{j=1}^m c_j e^{\lambda_j t} \vec{v}_j \quad (1.5)$$

siendo los c_j tales que

$$\sum_{j=1}^d c_j \vec{v}_j = y_0 \Leftrightarrow V c = y_0.$$

Por lo tanto, la solución es básicamente una combinación lineal de las funciones $e^{\lambda_j t}$ y está determinada por los autovalores de la matriz A , o lo que es lo mismo, por su espectro. Recordemos los siguientes resultados sobre matrices:

1.10.1. Triangulación de matrices

No todas las matrices son diagonalizables pero siempre se puede encontrar un buen cambio de base que las transforma en matrices triangulares. Teniendo en cuenta el producto escalar euclídeo, vamos a recordar los resultados más importantes, ver Allaire-Kaber [1]:

Definición 29 Una matriz $A \in \mathbb{C}^{n \times n}$ se puede reducir a una forma triangular superior (respectivamente a una forma diagonal) si existe una matriz no singular P una matriz triangular T (respectivamente, una matriz diagonal D) tal que

$$A = PTP^{-1}, \quad (\text{respectivamente } A = PDP^{-1}).$$

Cuando A se reduce a una forma triangular o diagonal, los autovalores de A repetidos con su multiplicidad $\sigma(A) = \{\lambda_1, \dots, \lambda_n\}$ aparecen en la diagonal de A , es decir,

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \quad \text{o} \quad T = \begin{pmatrix} \lambda_1 & \dots & * \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix}$$

Teorema 30 (Factorización de Schur general) *Cualquier matriz $A \in \mathbb{C}^{n \times n}$ (en particular si $A \in \mathbb{R}^{n \times n}$) se puede reducir a una forma triangular superior T mediante una matriz unitaria U , es decir,*

$$T = U^{-1}AU, \quad \text{donde } U^{-1} = U^*.$$

Los elementos de la diagonal de T son los autovalores, reales o complejos, de A

Teorema 31 *$A \in \mathbb{R}^{n \times n}$ es simétrica sí y sólo sí, existe una matriz ortogonal real Q ($Q^{-1} = Q^t$) y valores reales $\lambda_1, \dots, \lambda_n$ tal que si $D = \text{diag}\{\lambda_1, \dots, \lambda_n\}$, entonces $A = QDQ^{-1}$.*

Observación 17 *Las matrices A y T o A y D son semejantes, es decir, corresponden a la misma aplicación lineal pero expresada en dos bases distintas. U es la matriz del cambio de base y es un cambio de base ortogonal.*

Observación 18 *Los autovalores de A , $\{\lambda_1, \dots, \lambda_n\}$ repetidos con su multiplicidad aparecen en la diagonal de D o en la de T , por lo tanto, $\det(A) = \det(D) = \prod_{i=1}^n (\lambda_i)$ o $\det(A) = \det(T) = \prod_{i=1}^n (\lambda_i)$*

Prueba completa en Allaire-Kaber [1] secciones 2.4 and 2.5.

Ejemplo 32 *No todas las diagonalizaciones se pueden conseguir en una base de vectores real ortogonal, por ejemplo*

$$\begin{pmatrix} -1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & \sqrt{2} \end{pmatrix}$$

También podemos tener el caso de una matriz real con descomposición compleja

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix}$$

Observación 19 *Una vez factorizada $A = UTU^{-1}$ la resolución del sistema lineal $y' = Ay$ pasa por $y' = UTU^{-1}y$ de donde $U^{-1}y' = TU^{-1}y$, luego usando $w(t) = U^{-1}y(t)$ podemos simplificar la ecuación diferencial y tener $w' = Tw$ que se puede resolver desde la última ecuación hacia arriba. En el caso más sencillo donde $A = UDU^{-1}$ con D diagonal las ecuaciones están desacopladas, pero nos podemos encontrar con ecuaciones escalares de la forma $y'(t) = \lambda y(t)$ con $\lambda \in \mathbb{C}$ y es aquí donde podemos tener el interés en trabajar con tiempo t real pero con valores complejos.*

Ejemplo 33 Usemos las funciones **eig** y **schur** de MATLAB o de Octave. Consideremos la matriz

$$A = \begin{pmatrix} 1 & 1 & 1 & 3 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ -2 & 1 & 1 & 4 \end{pmatrix}$$

sus autovalores son

$$\text{eig}(A) = \begin{pmatrix} 4.8121 + 0.0000i \\ 1.9202 + 1.4742i \\ 1.9202 - 1.4742i \\ 1.3474 + 0.0000i \end{pmatrix}$$

y si realizamos la descomposición de Schur tenemos

```
[U,T]=schur(A,'complex')
```

U =

```
-0.4916 + 0.0000i -0.2756 - 0.4411i 0.2133 + 0.5699i -0.3428 + 0.0000i
-0.4980 + 0.0000i -0.1012 + 0.2163i -0.1046 + 0.2093i 0.8001 + 0.0000i
-0.6751 + 0.0000i 0.1842 + 0.3860i -0.1867 - 0.3808i -0.4260 + 0.0000i
-0.2337 + 0.0000i 0.2635 - 0.6481i 0.3134 - 0.5448i 0.2466 + 0.0000i
```

T =

```
4.8121 + 0.0000i -0.9697 + 1.0778i -0.5212 + 2.0051i -1.0067 + 0.0000i
0.0000 + 0.0000i 1.9202 + 1.4742i 2.3355 - 0.0000i 0.1117 + 1.6547i
0.0000 + 0.0000i 0.0000 + 0.0000i 1.9202 - 1.4742i 0.8002 + 0.2310i
0.0000 + 0.0000i 0.0000 + 0.0000i 0.0000 + 0.0000i 1.3474 + 0.0000i
```

```
>> [U,T]=schur(A,'real')
```

U =

```
-0.4916 -0.4900 -0.6331 -0.3428
-0.4980 0.2403 -0.2325 0.8001
-0.6751 0.4288 0.4230 -0.4260
-0.2337 -0.7200 0.6052 0.2466
```

T =

```
4.8121 1.1972 -2.2273 -1.0067
0 1.9202 -3.0485 -1.8381
0 0.7129 1.9202 0.2566
0 0 0 1.3474
```

Podemos observar como la descomposición real nos da los bloques de Jordan mientras que la factorización compleja nos facilita directamente los autovalores con su multiplicidad.

De acuerdo a lo anterior, siendo t real y usando la derivada siempre con respecto a la variable real t , tiene sentido considerar $y(t) \in \mathbb{R}$ y $\lambda \in \mathbb{C}$ y trabajar en el cuerpo de los números complejos como se hace en el de los reales. Se puede usar la exponencial compleja que simplifica operaciones algebraicas.

La ecuación $y'(t) = \lambda y(t)$, $y(0) = y_0$ posee solución

$$y(t) = y_0 e^{\lambda t} \in \mathbb{C}, \quad \forall t > 0$$

y equivale a tener $y(t) = y_0 e^{Re\lambda t} e^{iIm\lambda t}$.

Observación 20 *La diagonalización del sistema no lineal lleva a un sistema lineal caracterizado por los autovalores de la matriz jacobiana y estos pueden ser complejos. Esta es la razón por la que es importante el estudio de la ecuación $y'(t) = \lambda y(t)$ con $\lambda \in \mathbb{C}$.*

Esquemáticamente

1. Si $Re(\lambda) > 0$ tenemos un modo o componente inestable; con oscilaciones si $Im(\lambda) \neq 0$ o sin oscilaciones cuando $Im(\lambda) = 0$.
2. Si $Re(\lambda) < 0$ tenemos un modo o componente estable; con oscilaciones si $Im(\lambda) \neq 0$ o sin oscilaciones cuando $Im(\lambda) = 0$.
3. Si $Re(\lambda) = 0$ tenemos un modo o componente oscilatorio puro.

Ejemplo 34 *El problema del movimiento armónico simple se reescribe muy fácilmente en términos de la variable compleja $z(t) = u(t) + iv(t)$ como*

$$z'(t) = i z(t), \quad z(0) = 1$$

con solución $z(t) = e^{it}$. En general, el modelo

$$z'(t) = i \omega z(t), \quad z(0) = z_0$$

*con solución $z(t) = z_0 e^{i\omega t}$ representa un oscilador con **frecuencia** (número de vueltas por unidad de tiempo) ω . Para $c = a + ib$ y $|c| = (a^2 + b^2)^{1/2}$ tenemos que*

$$|z(t)| = |z_0|$$

luego el movimiento es circular con radio dado por el módulo del dato inicial.

1.11. Conclusión sobre los ejemplos y objetivos

La mayoría de las ecuaciones diferenciales ordinarias no se pueden resolver por las técnicas básicas analíticas que dan una solución explícita. Incluso cuando podemos, a veces resultan expresiones demasiado complicadas como para ser de utilidad, por lo tanto, debemos de deducir información sobre el comportamiento cualitativo y cuantitativo de las soluciones desde la propia ecuación. Se debe además recurrir a procesos computacionales que generen aproximaciones a la solución exacta. Para ello **calculamos valores puntuales** de las curvas, algo así como el esqueleto de las mismas, e intentaremos que sean próximos a los valores exactos de la curva solución.

En general, y como ya hemos visto, nos encontramos con situaciones y comportamientos muy variados en lo que se refiere a la solución continua. **Buscamos aproximar una curva concreta que está dentro de una familia de curvas** y nos podemos encontrar con:

- **Soluciones expansivas...** esto nos lo hace difícil porque nos movemos de una a otra debido a errores de cómputo y la expansión del sistema nos hace alejarnos mucho de la curva solución que buscamos
- **Soluciones contractivas...** esto nos lo puede hacer fácil porque, aunque cambiemos de curva debido a los errores de cómputo, la propia familia de soluciones nos acerca a la curva buscada.
- **Soluciones oscilatorias...** se debe capturar este tipo de movimiento. Esto lleva a lo que se conoce como **métodos simécticos** y es de gran importancia en el seguimiento de órbitas de planetas o satélites.
- **Soluciones con comportamientos expansivos, contractivos u oscilatorios bruscos y repentinos...** pueden necesitar una adaptación del cálculo de forma dinámica, esto nos complica el trabajo bastante
- **Soluciones vectoriales con modos de velocidades muy distintas...** se deben calcular de manera simultánea y hay que evitar perder información de los modos rápidos.
- **Combinaciones de todo lo anterior**

Cuando estudiemos procesos de aproximación numérica usaremos modelos simples con soluciones conocidas y sobre los que observar el comportamiento de estos procesos de aproximación numérica. Serán problemas triviales, pero **cualquier esquema numérico debe hacerlo bien** en estos casos tan simples. Además de que también son linealizaciones de problemas no lineales. El objetivo será trasladar conclusiones a problemas más complicadas. Algunas ecuaciones modelo pueden ser, por ejemplo:

1. $y'(t) = 1, y(0) = 0$: La solución exacta es $y(t) = t$.
2. $y'(t) = 0, y(0) = 0$: La solución exacta es $y(t) = 0$.
3. $y'(t) = \lambda y(t), y(0) = 1, \lambda \in \mathbb{R}$: Usualmente $\lambda < 0$ tal que $y(t) \rightarrow 0, t \rightarrow +\infty$ y modela decaimiento o disipación. Si $\lambda > 0$ se modela expansión ya que $y(t) \rightarrow +\infty, t \rightarrow +\infty$.
4. $y'(t) = i y(t), y(0) = 1, (i^2 = -1)$: Modela fenómenos oscilatorios.

1.12. Ejercicios

1. El sistema lineal

$$\begin{aligned}\frac{d}{dt}x(t) &= a x(t) + b y(t) \\ \frac{d}{dt}y(t) &= c x(t) + d y(t)\end{aligned}$$

aproxima (usando desarrollos de Taylor) el comportamiento de un sistema no lineal

$$\begin{aligned}\frac{d}{dt}x(t) &= f(x(t), y(t)) \\ \frac{d}{dt}y(t) &= g(x(t), y(t))\end{aligned}$$

que tenga un punto de equilibrio en $(0, 0)$ y tal que $f(0, 0) = g(0, 0) = 0$. El comportamiento de las soluciones del problema lineal indica como se comporta el problema no lineal en el entorno del punto de equilibrio $(0, 0)$. Esto se puede caracterizar de acuerdo a los autovalores de la matriz de coeficientes

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \partial_x f(0, 0) & \partial_y f(0, 0) \\ \partial_x g(0, 0) & \partial_y g(0, 0) \end{pmatrix}.$$

Esta matriz representa el primer término no nulo en el desarrollo de Taylor de $F = (f, g)$ en el punto $(0, 0)$.

- a) Para $f(x, y) = x^2 + y$ y $g(x, y) = -x + y^2$ construye el sistema lineal que aproxima la dinámica en el punto estable $(0, 0)$ y comprueba visualmente usando **pplane.jar** la afirmación anterior de que, en efecto, la dinámica lineal aproxima a la dinámica no lineal, ver Figura 9.5.
- b) Para las siguientes matrices usar **pplane.jar** para visualizar el campo de fases

$$A_1 = \begin{pmatrix} .5 & 1 \\ -1 & .5 \end{pmatrix} \quad A_2 = \begin{pmatrix} -.2 & 1 \\ -1 & -.2 \end{pmatrix} \quad A_3 = \begin{pmatrix} -1 & -1 \\ 4 & 1 \end{pmatrix}$$

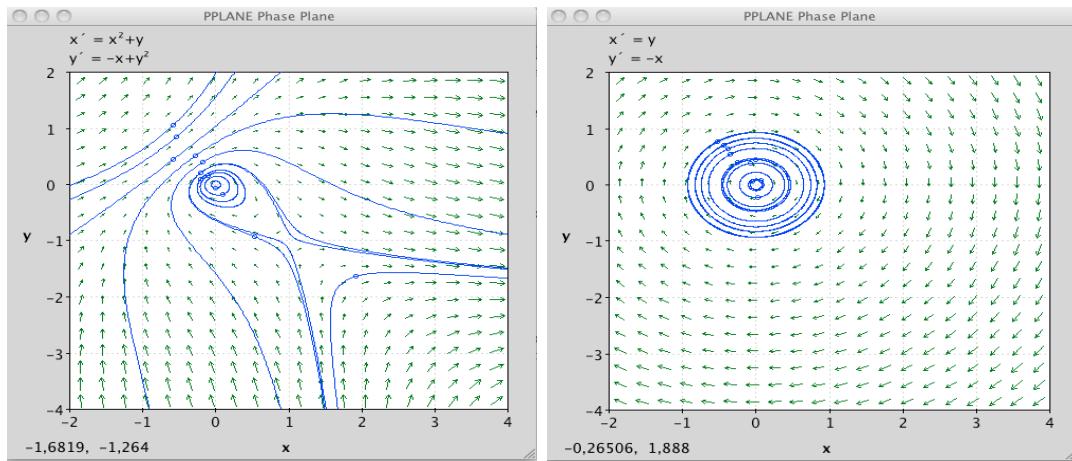


Figura 1.24: Comportamiento no lineal y su linealización en torno a $(0,0)$.

Describir los resultados obtenidos asociandolos con los autovalores de estas matrices. La situación de la matriz A_1 se denomina **fuente en espiral**, la de la matriz A_2 un **sumidero en espiral**, la de A_3 un **equilibrio central**.

2. Usando **dfield.jar**:

- a) Para el sistema $y'(t) = \cos(t) y(t)$ visualizar el campo de pendientes en una cuadrícula de puntos en el plano (t,y) dada por

$$t = -5 : 0.5 : 5, \quad y = -2 : 0.5 : 2.$$

Obtener las soluciones en la misma gráfica donde los valores iniciales estén en la partición $y = -2 : 0.5 : 2$ pero ahora la partición temporal sea más fina $t = -5 : 0.1 : 5$.

- b) Para el sistema $y'(t) = -y + t^2 + \cos(t)$ visualizar el campo de pendientes en la cuadrícula

$$t = -6 : 1 : 6, \quad y = 0 : 4 : 40.$$

Obtener las soluciones en la misma gráfica con valores iniciales en la partición $y = 0 : 4 : 40$ y partición temporal $t = -6 : 0.1 : 6$.

- c) Para el sistema $y'(t) = -10(y - \sin(t)) + \cos(t)$ visualizar el campo de pendientes en una cuadrícula

$$t = 0 : 0.2 : 1, \quad y = -2 : 0.2 : 2.$$

Obtener las soluciones en la misma gráfica con valores iniciales en la partición $y = -2 : 0.2 : 2$. y partición temporal $t = 0 : .01 : 1$.

3. Predecir aproximadamente el comportamiento de las soluciones de los campos de

$$y'(t) = a(\sin(t) - y(t)), \quad t \in (0, 3), \quad y(0) = y_0 \geq 0$$

para valores crecientes de $a = 1, 50, 100$. Explicar razonadamente el comportamiento del campo de velocidades y de las soluciones. Separación de variables nos da la solución

$$y(t) = e^{-at} y_0 + \frac{\sin(t) - a^{-1}\cos(t) + a^{-1}e^{-at}}{1 + a^{-2}}.$$

¿Coincide el comportamiento de las curvas con lo que nos dice la expresión analítica? Usar **dfield.jar** para contrastar.

4. Para cada uno de los problemas de valor inicial siguientes determinar la constante de Lipschitz en el rectángulo indicado usando el Teorema del valor medio:

- a) $y' = y \cdot (1 - y)$, $y(0) = 0.5$, $R = (-1, 1) \times (0, 2)$
b) $y' = y^2$, $y(0) = 1$, $R = (-1, 1) \times (0, 2)$

5. Escribir en forma de sistema de primer orden el problema

$$x''' = \cos(x) + \sin(x') - e^{x''} + t^2, \quad x(0) = 3, \quad x'(0) = 7, \quad x''(0) = 13.$$

6. Escribir en forma de sistema de primer orden el problema

$$\begin{aligned} x'' &= x - y - (3x')^2 + (y')^3 + 6y'' + 2t, \\ y''' &= xy'' - x' + e^x - t, \end{aligned}$$

$$x(1) = 2, \quad x'(1) = -4, \quad y(1) = -2, \quad y'(1) = 7, \quad y''(1) = 6.$$

7. La ecuación diferencial

$$\begin{aligned} x'_1 &= -x_1 \\ x'_k &= (k-1)x_{k-1} - kx_k, \quad k = 2, 3, \dots, 9 \\ x'_{10} &= 9x_9 \end{aligned}$$

describe la evolución de una reacción química. Comprobar que este sistema satisface una ley de conservación lineal. Específicamente, demostrar que la suma de todas las soluciones es constante.

8. El modelo de depredador presa se puede describir como

$$\begin{aligned} x' &= a(x - xy) \\ y' &= -c(y - xy). \end{aligned}$$

Comprobar que este sistema satisface una ley de conservación no lineal. Específicamente, demostrar que

$$G(t, x, y) = x^{-c}y^{-a}e^{cx+ay} = \text{constante}.$$

9. Para constantes a, b, d con $a, b > 0$ considerar la ecuación

$$y'(t) + ay(t) = de^{-bt}.$$

Calcular

$$\lim_{t \rightarrow +\infty} y(t)$$

considerando de forma separada los casos $a \neq b$ y $a = b$.

Capítulo 2

Método de Euler explícito

Resumen del tema

Introducimos notación y algunos resultados generales. Vemos con detalle el método de Euler explícito y los conceptos de consistencia, estabilidad y orden de convergencia.

2.1. Introducción

El método de cálculo aproximado más simple es el método de Euler explícito o progresivo y en este tema lo vamos a presentar con detalle puesto que contiene las ideas germinales de todos los demás métodos.

2.2. Discretización y primeros conceptos

Vamos a introducir unas ideas generales:

- como no podemos calcular indefinidamente, se fija $T > 0$ y nuestro intervalo de cálculo va a ser $[t_0, t_0 + T]$ (con frecuencia será $t_0 = 0$, pero no siempre).
- introducimos una **partición** de $[t_0, t_0 + T]$, que suponemos uniforme para simplificar, $t_n = t_0 + nh$ para $h = (t_0 + T - t_0)/N = T/N$ con $n = 0, 1, 2, \dots, N$ y la llamaremos Π_h , esto es,

$$\Pi_h = \{t_0 < t_1 < \dots < t_N = t_0 + T, \quad t_n = t_n^h = t_0 + nh, \quad h = T/N\}.$$

Queremos comparar una función continua, la solución, en los puntos $u(t_n^h)$ con unos valores discretos u_n^h calculados en los puntos de la discretización t_n^h . Buscamos que sea

$$u_n^h \approx u(t_n^h).$$

Si lo deseamos, podemos reconstruir una curva continua después uniendo los puntos (t_n^h, u_n^h) para formar una curva que aproxime a la curva verdadera $(t, u(t))$.

Definición 35 *Dada una familia de particiones $\{\Pi_h\}_{h>0}$ en el intervalo $[t_0, t_0+T]$, un **método numérico** para aproximar la solución del problema de Cauchy será un procedimiento para generar valores u_n^h sobre cada Π_h que aproximen a $u(t_n^h)$.*

Hipótesis 1 *Fijado un dato inicial tenemos existencia y unicidad de una solución continua y derivable al problema de Cauchy en $[t_0, t_0 + T]$.*

Definición 36 Convergencia: *Diremos que el método es convergente cuando para cualquier solución $u(t)$ del problema de Cauchy se cumple*

$$\max_{0 \leq n \leq N^h} |u(t_n^h) - u_n^h| \rightarrow 0, \quad h \rightarrow 0^+, \quad h = T/N^{(h)}$$

siendo $t_n^h = t_0 + h n$. Observar que

$$h \rightarrow 0^+ \Leftrightarrow N \rightarrow +\infty.$$

Observación 21 *Como es siempre $N h = T$, si hacemos más pequeño h entonces $N = T/h$ crece y las soluciones calculadas con distintos h van cambiando su localización en $[t_0, t_0+T]$, ver la Figura 2.2. Los límites cuando $h \rightarrow 0^+$ de las expresiones que contienen el número de puntos totales N siempre implican el crecimiento $N \rightarrow +\infty$ manteniéndose la razón $N h = T$ constante,*

$$h \rightarrow 0^+ \Leftrightarrow N \rightarrow +\infty, \quad \text{con} \quad N h = T.$$

Nos vamos a fijar normalmente en un punto $t_ = t_0 + n h$ fijo que supondremos que está en todas las particiones Π_h a partir de un cierto h . Introducimos la idea de un proceso de límite simultáneo y estacionario:*

- **límite simultáneo porque $h \rightarrow 0$ y $n \rightarrow +\infty$**
- **límite estacionario porque $t_* = t_0 + n h$ permanece fijo.**

De esta forma, aunque hagamos la partición más fina con $h \rightarrow 0$, nos fijamos siempre en el mismo punto. Los valores (t_n, u_n) generados dependen también de h , por lo tanto sería más correcto poner algo como (t_n^h, u_n^h) . Mantenemos la notación original (t_n, u_n) para simplificar y no sobrecargar, pero no debemos olvidar esto.

Ejemplo 37 Límite estacionario *Se suele usar la expresión*

$$\lim_{h \rightarrow 0} (1 + h)^n$$

pero no está bien escrita ya que se debe de entender que h y n están ligados por la relación $n h = t_\star - t_0$, si no fuese así el límite sería trivial. Se puede precisar en la forma

$$\lim_{\substack{h \rightarrow 0, n \rightarrow +\infty \\ hn=t_\star-t_0}} (1+h)^n$$

aunque resulta excesivamente cargada, o mejor,

$$\lim_{hn=t_\star-t_0} (1+h)^n.$$

Si entendemos la expresión original

$$\lim_{h \rightarrow 0} (1+h)^n$$

en el contexto correcto, hay que buscar el producto nh en la expresión y usar que su valor está fijo, es decir, $nh = t_\star$. Por ejemplo:

$$\begin{aligned} \lim_{h \rightarrow 0} (1+h)^n &= \lim_{h \rightarrow 0} \exp(n \log(1+h)) \\ &= \lim_{h \rightarrow 0} \exp\left(\frac{t_\star - t_0}{h} \log(1+h)\right) = \lim_{h \rightarrow 0} \exp((t_\star - t_0) \frac{\log(1+h)}{h}) \\ &= \exp(t_\star - t_0). \end{aligned}$$

De forma alternativa

$$\lim_{h \rightarrow 0} (1+h)^n = \lim_{h \rightarrow 0} \exp(n \log(1+h)) = \lim_{n \rightarrow +\infty} \exp(n \log(1 + \frac{t_\star - t_0}{n})) = \exp(t_\star - t_0).$$

Observación 22 Los valores u_n se pueden interpolar o extrapolar para aproximar $u(t)$ en puntos t que no están en la partición.

Observación 23 Tradicionalmente se usa el término **integración** para resolver el problema de Cauchy y se usa el término **cuadratura** para aproximar integrales.

Observación 24 De momento usamos un paso constante, pero gran parte de la potencia de los algoritmos modernos consiste en poder **adaptar el paso** conforme se calcula la solución. Esto lo veremos más adelante.

Los métodos numéricos producen los valores $\{u_n\}_n$ de forma iterativa: El valor u_{n+1} se calcula a partir de los valores u_0, u_1, \dots, u_{n-1} obtenidos previamente. Principalmente existen dos familias de métodos

- **Métodos de un paso:** para obtener u_{n+1} sólo usamos la información en t_n .

- **Métodos multipaso:** para obtener u_{n+1} usamos la información en k puntos previos, $t_n, t_{n-1}, \dots, t_{n-k+1}$ y decimos que es un método de k pasos.

Los distintos métodos numéricos existentes se pueden ver como **herramientas** a nuestra disposición para poder calcular una curva que siga un campo de velocidades dado. Cada una de estas herramientas se puede “**sintonizar**” con la curva buscada a través del parámetro $h > 0$ y se describe de acuerdo a las propiedades siguientes

- Orden de convergencia: error $\sim O(h^p)$. Cuanto mayor p mejor
- Restricciones de estabilidad: restricciones sobre el parámetro $h > 0$ necesarias para el cálculo buscado. Normalmente existe $h_* > 0$, que depende del campo de velocidades y del método usado, tal que si $h \geq h_*$ el método produce oscilaciones y divergencia. El valor h_* no se puede conocer en general, pero se puede predecir visualmente en base a los resultados computacionales.

La propiedad del orden suele ser la más importante pero a mayor orden el método es más complicado. Por otro lado, a veces puede compensar un método de menor orden porque tenga una restricción de estabilidad más laxa.

2.3. Método de Euler progresivo o explícito

Buscamos la curva $u(t)$ que cumple la ecuación

$$u'(t) = f(t, u(t)), \quad t \in [t_0, t_0 + T], \quad u(t_0) = \alpha.$$

Nuestros datos son (t_0, u_0) junto con la función $f(t, u)$ que nos evalúa la pendiente. Si $u_0 = \alpha$, o $u_0 \approx \alpha$, podemos construir el valor $f(t_0, u_0)$ que es la aproximación a la pendiente inicial de la curva solución en el punto (t_0, u_0) ; la recta tangente a la curva solución en el punto (t_0, u_0) es

$$z_0(s) \equiv u_0 + (s - t_0) f(t_0, u_0)$$

y tiene sentido entonces reemplazar el valor exacto $u(t_1)$ por el valor $z_0(t_1)$ si la distancia $t_1 - t_0$ no es muy grande. Esto es

$$u(t_1) \approx u_0 + (t_1 - t_0) f(t_0, u_0)$$

Pongamos $u_1 = u_0 + (t_1 - t_0) f(t_0, u_0)$, hemos entonces avanzado al punto (t_1, y_1) y ya tenemos dos puntos que son próximos a puntos de la curva

$$(t_0, u_0), \quad (t_1, u_1).$$

Partimos ahora de (t_1, u_1) , podemos repetir el proceso construyendo la pendiente $f(t_1, u_1)$ y la nueva recta tangente es

$$z_1(s) \equiv u_1 + (s - t_1) f(t_1, u_1)$$

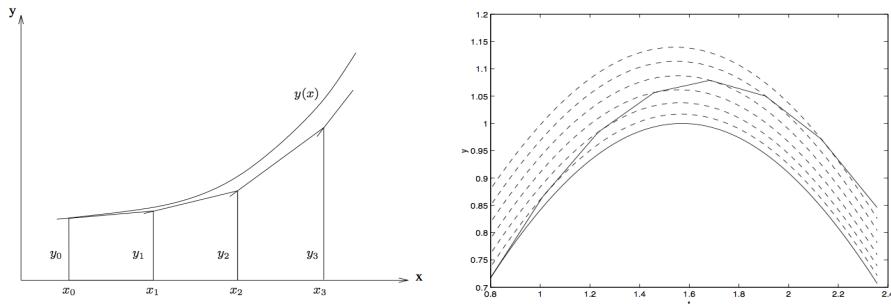


Figura 2.1: Método de Euler progresivo: normalmente terminamos en una trayectoria distinta debido al proceso de cálculo aproximado. El cómputo sólo genera los puntos (t_n, u_n) , pero para visualizar mejor se unen los puntos calculados con un trazo continuo.

y volver a progresar en el cálculo obteniendo una aproximación a $u(t_2)$ para un punto t_2 no muy alejado de t_1 dada por $z_1(t_2)$, esto es

$$y(t_2) \approx u_1 + (t_2 - t_1) f(t_1, u_1)$$

Pongamos $u_2 = u_1 + (t_2 - t_1) f(t_1, u_1)$, hemos entonces avanzado al punto (t_2, u_3) y ya hemos generado tres puntos que sirven para construir una poligonal que va, en principio, ajustándose a la solución buscada

$$(t_0, u_0), \quad (t_1, u_1), \quad (t_2, u_2)$$

y así podemos continuar generando puntos, ver Figuras 2.1 y 2.2. Por lo tanto, el esquema es:

Dado (t_0, u_0) obtener

$$\begin{aligned} t_{n+1} &= t_n + h, \\ u_{n+1} &= u_n + h f(t_n, u_n), \quad n = 0, 1, 2, \dots, N-1, \quad N = T/h \end{aligned}$$

Este esquema fue introducida por Euler en torno a 1770, se denomina **Método de Euler progresivo o explícito** y es el germen de todos los demás métodos numéricos. El proceso de cálculo es muy fácilmente programable y en forma de **pseudo-código** queda como:

Dados $t_0, T, N, y_0, f(t, u)$

Calculamos $h = T/N$

Para $n = 0$ hasta $N - 1$

 calcular $t_{n+1} = t_n + h$

 calcular $u_{n+1} = u_n + h f(t_n, u_n)$

Fin

Dibujar puntos (t_n, u_n)

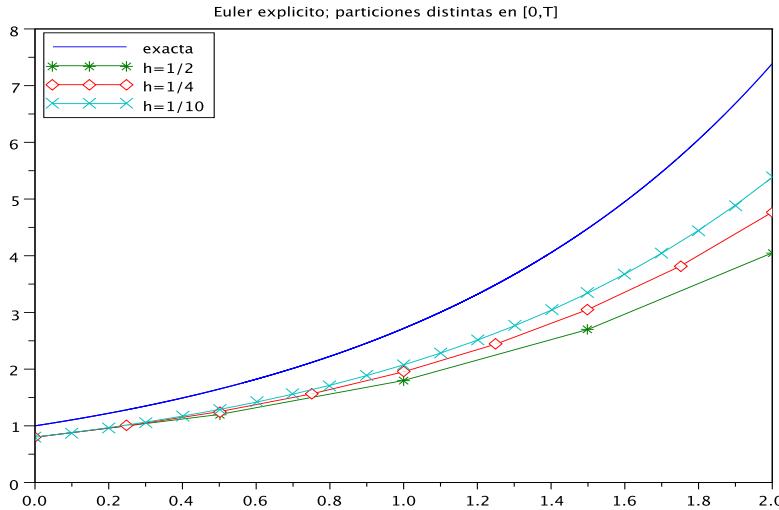


Figura 2.2: Distintas aproximaciones a la solución exacta para distintos h diferenciadas por los símbolos \star , \times , \diamond .

Observación 25 El valor inicial u_0^h no tiene porqué coincidir con $u(t_0) = \alpha$, puede darse el caso donde $u(t_0)$ provenga de un valor experimental, o por ejemplo $u(t_0) = \sqrt{\pi}$ o incluso $u(t_0) = 1/3\dots$. En general, hay un error inicial que no podemos despreciar.

Observación 26 Debido a las imprecisiones o errores terminamos siempre en una curva vecina a la buscada, ver la Figura 2.1.

Observación 27 Aparte de la derivación geométrica que hemos hecho podemos también llegar a este método por otras vías:

- **Fórmula de Taylor:**

Hacemos desarrollo de Taylor en torno al punto t_n :

$$u(t_{n+1}) = u(t_n) + h u'(t_n) + \frac{1}{2} h^2 u''(\xi_n), \quad t_n < \xi_n < t_{n+1},$$

y usando que $u'(t_n) = f(t_n, u(t_n))$ entonces

$$u(t_{n+1}) = u(t_n) + h f(t_n, u(t_n)) + \frac{1}{2} h^2 u''(\xi_n), \quad t_n < \xi_n < t_{n+1}.$$

Si el valor de h es lo suficientemente pequeño podemos despreciar el término que contiene h^2 y aproximar

$$u(t_{n+1}) \sim u(t_n) + h f(t_n, u(t_n))$$

lo que nos lleva al esquema de cálculo

$$u_{n+1} = u_n + h f(t_n, u_n)$$

para algún $u_n \sim u(t_n)$.

Observación 28 El valor $l(u(t_n); h) = \frac{1}{2}h^2 u''(\xi_n)$ se llama **residuo** o **error local de consistencia** sobre la solución $u(t)$. Surge cuando se pretende que la solución continua cumpla el esquema numérico, o equivalentemente, se supone que se puede aplicar el esquema en el valor exacto, $u_n = u(t_n)$

$$l(u(t_n); h) = u(t_{n+1}) - \{u(t_n) + h f(t_n, u(t_n))\}$$

o, lo que es lo mismo,

$$u(t_{n+1}) = u(t_n) + h f(t_n, u(t_n)) + l(t_n; h).$$

■ **Cuadratura numérica:**

Escribimos la edo entre t_n y t_{n+1} en su forma integral y la aproximamos por la fórmula del punto izquierdo.

Ejemplo 38 Consideremos

$$u' + u = 1, \quad y(0) = 0 \quad (2.1)$$

con solución

$$u(t) = 1 - e^{-t}$$

y el problema

$$u' = u, \quad u(0) = 1 \quad (2.2)$$

con solución

$$u(t) = e^t.$$

Podemos entonces realizar una tabla con los **errores de convergencia** obtenidos usando el método de Euler y representando el valor

$$E_N = \max_{0 \leq k \leq N} |u_k - u(t_k)| = \|u^h - u\|_\infty$$

para $N = 1/h$ con $N = 16, 32, 64, \dots$ en el intervalo de tiempo $[0, 1]$. Vemos que el error decae por un factor de 2 cada vez que se divide entre 2 el parámetro h . Luego el método es de orden 1 en h

$N = h^{-1}$	error en (2.1)	error en (2.2)
16	0.0118053	0.0803533
32	0.0058242	0.0412917
64	0.0028929	0.0209369
128	0.0014417	0.0105428
256	0.0007197	0.0052902
512	0.0003595	0.0026498
1024	0.0001797	0.0013261

Veremos en teoría que el método de Euler explícito es de primer orden confirmando los resultados de antes.

2.4. Estimación del orden de convergencia a cero

El estudio general del error sólo se puede hacer desde el punto de vista computacional puesto que no hay una expresión cerrada para el mismo, sólo se puede acotar teóricamente su forma en términos de una constante (que depende de todos los datos del problema y de la solución a la que le aplicamos el método que es desconocida) y del parámetro h . Para conocer este error, una primera clasificación práctica es comprobar el orden del error que se produce sobre problemas donde la solución es conocida.

Necesitamos el concepto de **orden de aproximación** y usamos para ello la notación O (leer O grande): supongamos que $h > 0$ es nuestra variable y tenemos una función $E(h)$ en un intervalo de la forma $[0, h^*)$ que converge a cero cuando h tiende a cero, esto es, $\lim_{h \rightarrow 0} E(h) = 0$. La pregunta que nos interesa resolver es: ¿con qué velocidad decae este error a cero? El concepto fundamental aquí es el de **orden de convergencia**

Definición 39 *Diremos que una función $E(h)$ converge a cero con orden p con respecto a h si converge a cero de forma proporcional a como lo hace h^p , esto es*

$$E(h) = O(h^p), \quad h \rightarrow 0^+$$

o lo que es lo mismo, existe una constante $K > 0$ tal que

$$\lim_{h \rightarrow 0^+} \frac{E(h)}{h^p} = K \neq 0.$$

También equivale a la existencia de un entorno $V(0) = (0, h_0)$ y una constante $K > 0$ tal que

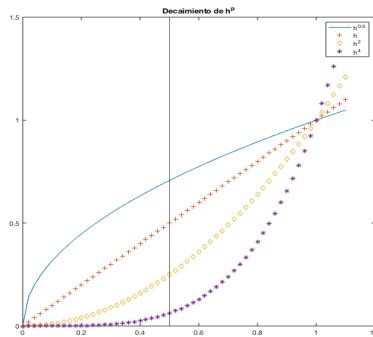
$$|E(h)| \approx K h^p, \quad \forall h \in (0, h_0).$$

En nuestro curso vamos a usar esta definición para controlar los resultados numéricos que produce un cálculo.

Observación 29 *Aun siendo obvio, ver Figura 2.3, recordemos que si $h < 1$ entonces*

$$\dots < h^{p+1} < h^p < \dots < h^4 < h^3 < h^2 < h < 1.$$

Observación 30 *La expresión analítica de $E(h)$ no se conoce y sólo se tienen los valores computacionales que se pueden obtener para casos concretos de h .*

Figura 2.3: Decaimientos a cero de distintas potencias de h .**Ejemplo 40**

$$\begin{aligned} \sin(3h) &= 3h - \frac{(3h)^3}{3!} + \frac{(3h)^5}{5!} + \dots \Rightarrow \sin(3h) = O(h), \quad h \rightarrow 0, \\ \sin(3h) - 3h &= -\frac{(3h)^3}{3!} + \frac{(3h)^5}{5!} \dots \Rightarrow \sin(3h) - 3h = O(h^3), \quad h \rightarrow 0, \\ \sin(3h) - 3h + \frac{(3h)^3}{3!} &= \frac{(3h)^5}{5!} + \dots \Rightarrow \sin(3h) - 3h + \frac{(3h)^3}{3!} = O(h^5), \quad h \rightarrow 0. \end{aligned}$$

Observación 31 En el caso

$$\lim_{h \rightarrow 0^+} \frac{E(h)}{h^p} = 0$$

tenemos que $E(h)$ converge a cero más rápido que h^p . Entonces se escribe $E(h) = o(h^p)$ y se lee o pequeña

$$E(h) = o(h^p), \quad h \rightarrow 0^+.$$

Pero no tiene porqué tender a cero como h^{p+1} ya que puede hacerlo de forma intermedia, por ejemplo podemos tener $E(h) = K h^{p+1} |\log(h)|$ y entonces

$$\frac{E(h)}{h^p} = K h |\log(h)| \rightarrow 0$$

pero

$$\frac{E(h)}{h^{p+1}} = K |\log(h)| \rightarrow +\infty.$$

Definición 41 Convergencia y orden de convergencia: Diremos que un método numérico es convergente cuando para cualquier solución $y(t)$ del problema de Cauchy los valores generados $\{u_n^h\}_n$ a partir del dato $y(t_0)$ cumplen

$$\max_{0 \leq n \leq N^h} |y(t_n^h) - u_n^h| \rightarrow 0, \quad h \rightarrow 0^+, \quad h = T/N^{(h)}$$

siendo $t_n^h = t_0 + h n$. Si además

$$\max_{0 \leq n \leq N^{(h)}} |u_n^{(h)} - y(t_n^h)| = O(h^p), \quad h \rightarrow 0^+, \quad h = T/N^{(h)}$$

se dice que tiene un orden p de convergencia con respecto a h .

Observación 32 Es necesario que el error inicial $|y(t_0^h) - u_0^h|$ tienda a cero con el mismo orden que el orden del método. Esto conviene decirlo ya que el error inicial nos viene dado y el método es nuestra elección.

Observación 33 Como siempre es $h \rightarrow 0^+$ simplificaremos escribiendo $O(h^p)$ sin necesidad de explicitar $h \rightarrow 0^+$.

Observación 34 Este resultado usa el límite $h \rightarrow 0^+$ pero desde el punto de vista de un cómputo puede ocurrir que el valor h sea tan pequeño que resulte impracticable.

Tenemos dos formas de estimar el error, usaremos la segunda por ser mas elegante.

2.4.1. Uso de tablas

En general, si el método es convergente de orden $p > 0$, el error es de la forma $E(h) \approx C h^p$ y entonces si dividimos h entre 2 y volvemos a calcular el error

$$E(h/2) \approx C \left(\frac{h}{2}\right)^p = C \frac{h^p}{2^p} \approx \frac{E(h)}{2^p}$$

por lo tanto, el error obtenido es el que teníamos antes pero dividido por 2^p , esto nos permite calcular p tomando logaritmos (el uso de 2 es arbitrario). La idea es combinar valores distintos de h y estimar el orden de convergencia p . Por ejemplo, debe ser

$$\frac{E(h)}{E(h/2)} \approx 2^p \Rightarrow p \approx \log\left(\frac{E(h)}{E(h/2)}\right)/\log(2).$$

Ejemplo 42 Supongamos que para valores distintos de h se ha obtenido con el computador la siguiente tabla de aproximaciones para las funciones $f(h)$ y $g(h)$:

h	$f(h)$	$g(h)$
0.0625	0.0118053	0.0803533
0.03125	0.0058242	0.0212917
0.015625	0.0028929	0.0059369
0.0078125	0.0014417	0.0015428

Podemos entonces ver que al dividir por 2 el valor de h el valor de $f(h)$ decrece por un factor de 2 mientras que $g(h)$ decrece por un factor 4. Con esta información se puede deducir, en principio, que

$$f(h) = O(h), \quad g(h) = O(h^2), \quad h \rightarrow 0.$$

Observación 35 Cuando los valores de los errores estén cerca del cero del computador, $\sim 10^{-16}$, estos estudios dejan de ser válidos debido a los errores de redondeo o truncatura que realiza la máquina. Por otro lado, si el valor de h no es lo suficientemente pequeño estos patrones tampoco se van a encontrar ya que los cálculos no se aproximan lo suficiente a la solución buscada. Por lo tanto, siempre hay un $h_{\min} < h_{\max}$ tal que si $h \in (h_{\min}, h_{\max})$ el patrón buscado $E(h) \sim O(h^p)$ se encuentra.

2.4.2. Recta de pendiente

Pongamos $E(h) = C h^p$, entonces si hacemos cálculos para distintos valores de h obtenemos los pares $\{h_j, E(h_j)\}_j$ y si tomamos los h_j en el rango adecuado podemos ver claramente el efecto de cada uno de los términos de este error.

Se puede entonces hacer una gráfica con los valores $\{h_j, E(h_j)\}_j$ e intentar deducir el comportamiento de una curva en la forma $y = C x^p$. Pero el rango de escalas puede ser muy amplio (dependiendo del rango de h los valores del error pueden pasar de 10^{10} a 10^{-14} por ejemplo), por eso es mejor usar los logaritmos del error, esto permite visualizar mejor los valores. Tomando entonces logaritmos:

$$E(h) \approx C h^p \Rightarrow \log(E(h)) \approx p \log(h) + \log(C)$$

y el **orden p del error coincide con la pendiente** de la recta sobre la que caen estos datos.

Observación 36 Teniendo en cuenta que

$$\log(10) \approx 2.3 \Rightarrow \log(10^{\pm M}) \approx \pm M * 2.3$$

la magnitud de los errores (su logaritmo) se puede deducir de los valores en el eje OY. En particular, tenemos que

$$\log(10^{-15}) \approx -35$$

luego cuando usando una escala logarítmica en el eje OY alcanzamos aproximadamente el valor negativo -35 el error es del orden de 10^{-15} y estamos cerca del cero del computador.

Como $h = T N^{-1}$ también se puede usar $E(N) \approx C N^{-p}$ y obtener

$$\log(E(N)) \approx -p \log(N).$$

Es decir, los datos $\{\log(h_j), \log(E(h_j))\}_j$ o los datos $\{\log(N_j), \log(E(N_j))\}_j$ visualmente deben de parecerse a una recta de pendiente p o $-p$ respectivamente.

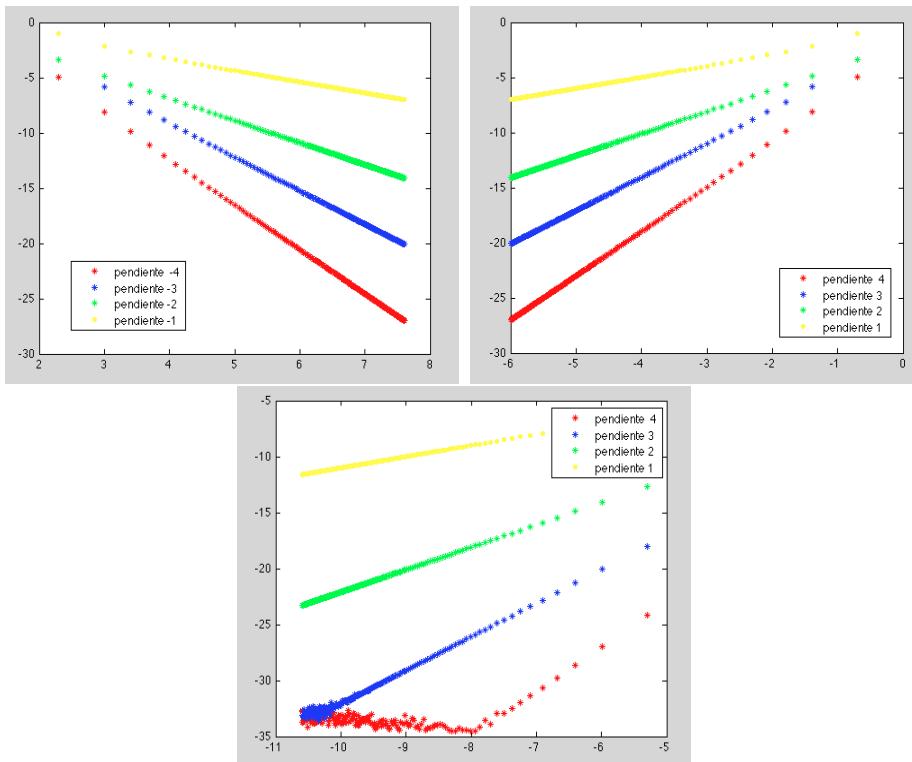


Figura 2.4: Rectas de pendientes para métodos de orden 1,2,3 y 4 visualizadas sobre OX positivo o negativo. En la tercera gráfica se ve el efecto de trabajar con errores por debajo de la precisión del computador. Se ve como las rectas para orden 3 y 4 pierden la pendiente antes que las de orden 1 y 2.

Ejemplo 43 En la Figura 2.4 se ven las rectas de pendientes para métodos de orden 1,2,3 y 4 visualizadas sobre OX positivo o negativo, dependiendo de escribir $E(h) = Ch^p$ o $E(N) = CN^{-p}$, y tomando logaritmos. En la tercera gráfica se observa el efecto de trabajar con errores por debajo de la precisión del computador. Se ve como las rectas para orden 3 y 4 pierden la pendiente antes que las de orden 1 y 2. Como sabemos que $\log(10^{-15}) \approx -35$ cuando en el eje OY alcanzamos aproximadamente el valor negativo -35 el error es del orden de 10^{-15} y estamos cerca del cero del computador.

2.4.3. Cuando la solución verdadera no se conoce

En este caso, podemos tener

$$u^h = u + Ch^p + O(h^{p+1})$$

donde $u(t)$ es la solución desconocida y u^h es la calculada mediante un método numérico. El orden se puede determinar a partir de 3 soluciones con valores $h, h/2$

y $h/4$ como sigue:

$$\begin{aligned} u^h &= u + Ch^p + O(h^{p+1}) \\ u^{\frac{h}{2}} &= u + C\left(\frac{h}{2}\right)^p + O\left(\left(\frac{h}{2}\right)^{p+1}\right) \\ u^{\frac{h}{4}} &= u + C\left(\frac{h}{4}\right)^p + O\left(\left(\frac{h}{4}\right)^{p+1}\right) \end{aligned}$$

entonces, dividiendo por h^p en la razón siguiente,

$$\begin{aligned} \frac{u^h - u^{\frac{h}{2}}}{u^{\frac{h}{2}} - u^{\frac{h}{4}}} &= \frac{Ch^p + O(h^{p+1}) - C\left(\frac{h}{2}\right)^p + O\left(\left(\frac{h}{2}\right)^{p+1}\right)}{C\left(\frac{h}{2}\right)^p + O\left(\left(\frac{h}{2}\right)^{p+1}\right) - C\left(\frac{h}{4}\right)^p + O\left(\left(\frac{h}{4}\right)^{p+1}\right)} \\ &= \frac{1 - 2^{-p} + O(h)}{2^{-p} - 2^{-2p} + O(h)} = 2^p + O(h) \end{aligned}$$

procedemos como antes para obtener el orden p del método. Evidentemente, hay que tener cuidado con las comparaciones y hacerlas todas en los mismos nodos, los comunes a todas las particiones, es decir, los de la partición de talla h . Quedaría como

$$\frac{\|u^h - u^{\frac{h}{2}}\|_h}{\|u^{\frac{h}{2}} - u^{\frac{h}{4}}\|_h} = 2^p + O(h)$$

donde

$$\|u^h - u^{\frac{h}{2}}\|_h = \max_n |u^h(t_n) - u^{\frac{h}{2}}(t_n)|, \quad t_n = nh, \quad n = 0, 1, 2, \dots, N = T/h.$$

Por último, cuando no conocemos la solución exacta y queremos plantear un cómputo con una tolerancia dada podemos usar un cálculo u^{h_\star} para h_\star lo suficientemente pequeño como para considerar que u^{h_\star} es la solución exacta y computar

$$\frac{\|u^h - u^{h_\star}\|_h}{\|u^{h_\star}\|_h}$$

como medida del error relativo de nuestra solución aproximada u^h .

2.5. Convergencia de Euler explícito

Definición 44 Si $\{u_n^h\}_n$ es la solución obtenida por el esquema entenderemos por error global en el tiempo t_n^h como la diferencia

$$e_n^h = y(t_n^h) - u_n^h.$$

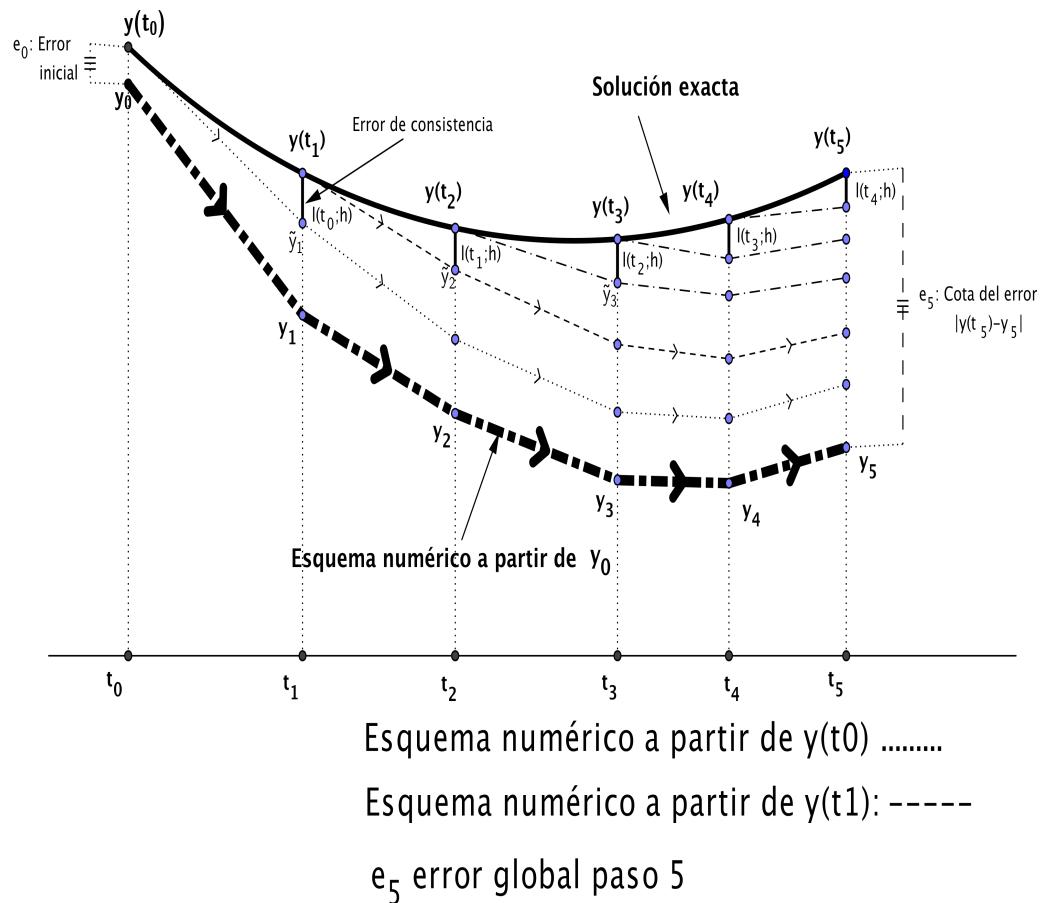


Figura 2.5: Interpretación gráfica del método de Euler explícito. Los trazos rectos se construyen con Euler mientras que la curva continua es la solución exacta. En cada vertical t_n se ve el error local en el trazo grueso y el resto depende de aplicar el esquema en puntos originales distintos. Esta figura se conoce como el abanico de Lady Windermere, nombre dado por Gerhard Wanner [14] en los años 1980 inspirado en la obra homónima de Oscar Wilde.

Vamos a caracterizar el **error global** cometido en términos del parámetro h que define la partición del intervalo temporal. Nuestro objetivo es conseguir un error que dependa de h y que tienda a cero con h . La rapidez con la que lo haga indica el orden del método con respecto a h .

No es evidente como estimar este error pero si nos fijamos en la Figura 2.5 podemos tener una idea de como proceder. Cada **error global** en un instante t_n se puede representar por **el segmento vertical que une u_n con $y(t_n)$** . Este segmento se puede descomponer en:

1. El **trazo grueso vertical** es el error de aplicar el esquema numérico en el punto exacto $y(t_n)$. Esto se denomina **el error local** que genera el esquema en un paso si no hay error inicial en ese paso. Se controla **usando la consistencia del esquema numérico**.
2. El **resto de segmentos de trazo fino** surgen por aplicar el esquema numérico en puntos previos distintos. Todos estos segmentos se pueden controlar **usando la estabilidad del esquema numérico**.

Esta segunda parte es el error generado por el esquema cuando se empezó en puntos distintos en algun paso previo. Por ejemplo, en la vertical t_3 queremos estimar el error $y(t_3) - u_3$. Entonces, el primer trazo grueso es el error cometido por lanzar el esquema desde $y(t_2)$ mientras que los dos segmentos restantes son la suma de errores cometidos por realizaciones del esquema en puntos iniciales distintos.

Por lo tanto, la calidad de la aproximación que hemos realizado depende de dos aspectos:

1. **Estabilidad:** La propagación por el esquema numérico de los errores que se cometen en cada paso de cálculo (la amplitud de esta propagación depende del método numérico usado y de la longitud h usada).
2. **Consistencia:** La talla de los errores locales cometidos en cada paso (se computan suponiendo que $y(t_n)$ coincide con u_n , esto se llama **hipótesis de localización**, y dependen también del método numérico usado).

Estas dos nociones son fundamentales en todos los métodos numéricos para ecuaciones diferenciales:

Teorema 45 Teorema de Lax: *Un método numérico es convergente sí y sólo si es estable y consistente.*

Vamos a describir estas ideas en términos matemáticos.

2.5.1. Consistencia del esquema numérico

Tenemos una partición del intervalo de tiempo t_0, t_1, \dots, t_N y buscamos valores u_0, u_1, \dots, u_N que aproximen a los valores exactos $y(t_0), y(t_1), \dots, y(t_N)$. Una forma de verlo es que vamos a reemplazar el operador diferencial por un cociente incremental. Esto se llama discretización del operador, se conoce como **método de diferencias finitas**. La idea fundamental proviene de la misma definición de derivada:

$$y'(t) = \lim_{h \rightarrow 0} \frac{y(t+h) - y(t)}{h}$$

por lo que si h es lo suficientemente pequeño entonces

$$y'(t) \sim \frac{y(t+h) - y(t)}{h}.$$

El error que se comete al hacer esta aproximación se puede estimar usando los desarrollos de Taylor. Por lo tanto, el error está intimamente relacionado con la regularidad que podamos tener sobre $y(t)$. Por ejemplo, cuando $y \in C^2$ tenemos las **diferencias progresivas**

$$y'(t) \sim \frac{y(t+h) - y(t)}{h}$$

y las **diferencias regresivas**

$$y'(t) \sim \frac{y(t) - y(t-h)}{h}$$

donde el error es proporcional a h mientras que si $y \in C^3$ la **diferencia central**

$$y'(t) \sim \frac{y(t+h) - y(t-h)}{2h}$$

tiene un error proporcional a h^2 .

Observación 37 *La estimación de error cuando se usa un cociente incremental o diferencia finita depende de la regularidad de la función que se está approximando.*

Tenemos el operador diferencial

$$\mathcal{L}(y(t)) = y'(t) - f(t, y(t))$$

y sabemos que la solución que buscamos cumple

$$\mathcal{L}(y(t)) = 0.$$

Para $h > 0$, usando el desarrollo de Taylor, podemos escribir

$$\mathcal{L}(y(t)) = y'(t) - f(t, y(t)) = \frac{y(t+h) - y(t)}{h} - f(t, y(t)) + \frac{h}{2}y''(\xi).$$

Si definimos

$$\mathcal{D}(y(t); h) = \frac{y(t+h) - y(t)}{h} - f(t, y(t)), \quad \mathcal{T}(y(t); h) = -\frac{h}{2}y''(\xi)$$

tenemos entonces que

$$\mathcal{L}(y(t)) = \mathcal{D}(y(t); h) - \mathcal{T}(y(t); h).$$

- La expresión $\mathcal{D}(y(t); h)$ no contiene derivadas y es una **discretización de la ecuación diferencial**.
- La parte $\mathcal{T}(y(t); h)$ es el error que se comete al reemplazar la ecuación diferencial por su discretización y se llama **error de truncatura**. Resultado de truncar la derivada con el cociente incremental. Su propio nombre refleja que sólo una parte del desarrollo de Taylor se usa en su construcción.

Observación 38 *Podemos usar cualquier otra expresión para el operador en diferencias $\mathcal{D}(y(t); h)$, evidentemente nos lleva a una expresión consecuentemente distinta para el error de truncatura $\mathcal{T}(y(t); h)$.*

Cuando y es solución del problema, se tiene que $\mathcal{L}(y(t)) = 0$ pero como

$$\mathcal{L}(y(t)) = \mathcal{D}(y(t); h) - \mathcal{T}(y(t); h)$$

entonces

$$\mathcal{L}(y(t)) = 0 \Leftrightarrow \mathcal{D}(y(t); h) = \mathcal{T}(y(t); h).$$

Entonces, la idea en diferencias finitas es reemplazar los valores nodales que se usan en $\mathcal{D}(y(t); h)$ por unos valores aproximados $\{v_j\}$ y resolver

$$\mathcal{D}(\{v_j\}; h) = 0.$$

Formalmente, tendremos las ecuaciones

$$\begin{aligned}\mathcal{D}(\{v_j\}; h) &= 0 \\ \mathcal{D}(y(t); h) &= \mathcal{T}(y(t); h)\end{aligned}$$

y podremos obtener la estimación de error

$$\mathcal{D}(\{v_j\}, h) - \mathcal{D}(\{y(t_j)\}; h) = \mathcal{T}(y(t); h).$$

A partir de aquí se deduce la convergencia en alguna norma ya que por la estabilidad del método se tiene una acotación de la forma (en alguna norma conveniente)

$$\|\{v_j\} - \{w_j\}\| \leq C(\|v_0 - w_0\| + \|\mathcal{D}(\{v_j\}; h) - \mathcal{D}(\{w_j\}; h)\|)$$

para C independiente de h . Si se supone que no hay errores en la realización del método de diferencias finitas, entonces

$$\mathcal{D}(\{v_j\}; h) = \mathcal{D}(\{w_j\}; h) = 0.$$

Cuando los valores $v_j = y_j$ son el resultado de aplicar el esquema del método numérico y se usa la solución exacta $y(t_j) = w_j$, entonces aparece el error de truncatura y se obtiene la convergencia usando la acotación sobre este error de truncatura. Supongamos que

$$\|\mathcal{T}(y(t); h)\| \leq O(h^p), \quad \|y_0 - y(t_0)\| \leq O(h^p)$$

entonces

$$\begin{aligned} \|\{y_j\} - \{y(t_j)\}\| &\leq C(\|y_0 - y(t_0)\| + \|\mathcal{D}(\{y_j\}; h) - \mathcal{D}(\{y(t_j)\}; h)\|) \\ &= C(\|y_0 - y(t_0)\| + \|0 - \mathcal{D}(\{y(t_j)\}; h)\|) \\ &= C(\|y_0 - y(t_0)\| + \|\mathcal{T}(y(t); h)\|) \leq O(h^p). \end{aligned}$$

En general, se puede hablar del máximo de los errores locales de truncatura en todo el intervalo $[t_0, t_0 + T]$ que podemos llamarlo **error de truncatura global sobre la función $y(t)$**

$$\mathcal{T}(y; h) = \max_t \mathcal{T}(y(t); h).$$

Es razonable para cualquier discretización de una ecuación diferencial que el error de truncatura tienda a cero con h

Definición 46 *El método numérico se dice consistente cuando el error de truncatura global sobre cualquier solución $y(t)$ de la ecuación diferencial lo suficientemente derivable tiende a cero con h , es decir,*

$$\mathcal{T}(y; h) \rightarrow 0, \quad h \rightarrow 0 \quad \forall y(t) \text{ si } \mathcal{L}(y(t)) = y'(t) - f(t, y(t)) = 0.$$

*A esta propiedad se le llama **consistencia**. Si además se cumple para algún $p \geq 0$ que sobre cualquier solución de la ecuación diferencial lo suficientemente derivable*

$$\mathcal{T}(y; h) = O(h^p), \quad h \rightarrow 0 \quad \text{siendo } \mathcal{L}(y(t)) = y'(t) - f(t, y(t)) = 0$$

*entonces se dice que el método tiene **orden de consistencia p** .*

Por otro lado,

$$\mathcal{T}(y(t); h) = \frac{y(t+h) - y(t)}{h} - f(t, y(t))$$

entonces

$$y(t+h) = y(t) + hf(t, y(t)) + h\mathcal{T}(y(t); h)$$

y a la expresión

$$l(y(t); h) = h\mathcal{T}(y(t); h)$$

se le llama **error local de consistencia en el punto t con paso h para la solución $y(t)$** y se puede interpretar como

$$l(y(t); h) = y(t+h) - \tilde{y}$$

donde $\tilde{y} = y(t) + hf(t, y(t))$ es el resultado de suponer que se puede usar el valor exacto $y(t)$ para aplicar el esquema de cálculo un paso (esto se suele llamar **hipótesis de localización**). También podemos hablar del máximo de los errores locales de consistencia en todo el intervalo $[t_0, t_0 + T]$ para una solución $y(t)$ como

$$l(y; h) = \max_t l(y(t); h).$$

Observación 39 *El error de truncatura local $\mathcal{T}(y(t); h)$ para la solución $y(t)$ en un punto t generado por el esquema numérico con paso h es el error que se comete al reemplazar la ecuación diferencial por la ecuación en diferencias en el punto $(t, y(t))$ y con paso h .*

Definición 47 *El error de consistencia local $l(y(t); h)$ para la solución $y(t)$ en un punto t generado por el esquema numérico con paso h es el error que se comete al aplicar el esquema sobre la solución exacta, $(t, y(t))$, en el punto t , para aproximar el valor $y(t+h)$.*

En nuestro caso

$$l(y(t); h) = y(t+h) - \{y(t) + h f(t, y(t))\} = h\mathcal{T}(y(t); h).$$

Observación 40 *El error de consistencia local se puede entender como lo que le falta al esquema para que la solución exacta lo cumpla cuando se parte del valor exacto $(t, y(t))$.*

Observación 41 *Gracias al desarrollo de Taylor tenemos*

$$l(y(t); h) = \frac{h^2}{2} y''(\xi^h), \quad \xi^h \in (t, t+h).$$

Pero $y''(t)$ es la segunda derivada de la función que buscamos y , por lo tanto, es desconocida. Podemos solventar esta dificultad escribiendo y'' en términos de f y sus derivadas:

$$y''(t) = f_t(t, y(t)) + f(t, y(t))f_y(t, y(t))$$

Como estamos tomando $f \in C^1$ si suponemos que la solución $(t, y(t))$ para $t \in [t_0, t_0 + T]$ se encuentra en una región acotada $R \subset \mathbb{R}^2$ del plano (t, y) podemos determinar una constante

$$M = \max_{(t,y) \in R} |f_t(t, y) + f(t, y)f_y(t, y)|$$

y entonces escribir para cualquier valor de t y para cualquier curva solución del problema de Cauchy con $f(t, y)$ que

$$l(y; h) \leq \frac{h^2}{2} M = O(h^2).$$

Ejemplo 48 Se puede acotar una función solución del problema de Cauchy y sus derivadas sin conocerla explícitamente. Por ejemplo

$$y'(t) = \sin(e^{y(t)}), \quad t \in [0, 1], \quad y(0) = 0.$$

ya nos dice que $|y'(t)| \leq 1$. Además, como

$$y(t) = 0 + \int_0^t y'(s)ds$$

entonces $|y(t)| \leq t \leq 1$. Además, al derivar la ecuación

$$y''(t) = \cos(e^{y(t)})e^{y(t)}\sin(e^{y(t)})$$

de donde $|y''(t)| \leq e$.

Observación 42 El valor preciso de esta acotación no interesa desde el punto de vista práctico ni computacional. Sólo interesa la potencia de h que acompaña a la constante puesto que es la que marca el orden de convergencia del método.

Observación 43 Sabemos que **un esquema numéricico es consistente** cuando para cada partición, caracterizada por h , el error de truncatura global sobre cada solución $y(t)$ del problema de valor inicial tiende a cero,

$$\mathcal{T}(y; h) \rightarrow 0, \quad h \rightarrow 0.$$

Esto es lo mismo que decir que la suma de los errores locales tiende a cero:

$$\sum_{j=0}^{N-1} l(y(t_j); h) \rightarrow 0, \quad h \rightarrow 0, \quad (N = T/h, t_j = t_0 + j h)$$

ya que si trabajamos con el error local máximo $l(y; h)$ entonces

$$\sum_{j=0}^{N-1} l(y(t_j); h) \leq \frac{T}{h} l(y; h) \leq T\mathcal{T}(y; h) \rightarrow 0, \quad h \rightarrow 0.$$

La consistencia de un esquema obliga a tener

$$\mathcal{T}(y; h) = O(h^p), \quad h \rightarrow 0$$

para algún valor $p > 0$, o lo que es lo mismo,

$$l(y; h) = O(h^{p+1}), \quad h \rightarrow 0.$$

Normalmente este p es un entero y en nuestro caso es $p = 1$ y garantiza que la suma de los errores locales tiende a cero. Como aplicamos el esquema en cada punto t_n de la partición y tenemos N puntos es razonable pensar que tendremos un error parecido al producto de N por $l(y; h)$.

2.5.2. Estabilidad del esquema numérico

Para estimar el resto de errores usamos la estabilidad del esquema numérico. Datos de partida distintos en un esquema numérico de cálculo generarán valores finales distintos y debemos controlar esta desviación. Esto es lo que se denomina **el estudio de la estabilidad del esquema numérico**, estabilidad a cero, o simplemente **0-estabilidad** que es la estabilidad que se obtiene cuando $h \rightarrow 0^+$.

Supongamos que damos un paso con dos valores iniciales distintos

$$\begin{aligned} u_{n+1} &= u_n + h f(t_n, u_n) \\ v_{n+1} &= v_n + h f(t_n, v_n) \end{aligned}$$

entonces para ξ_n entre u_n y v_n tenemos

$$u_{n+1} - v_{n+1} = u_n - v_n + h[f(t_n, u_n) - f(t_n, v_n)] = (1 + h\partial_y f(t_n, \xi_n))(u_n - v_n).$$

El factor que relaciona la diferencia en el paso n con la diferencia en el paso $n + 1$ se llama **factor de amplificación** y es

$$|1 + h\partial_y f(t_n, \xi_n)|.$$

Observación 44 Supongamos que $\partial_y f(t_n, \xi_n)$ no es muy grande. Entonces

$$\partial_y f(t_n, \xi_n) > 0 \Rightarrow 1 + h\partial_y f(t_n, \xi_n) > 1.$$

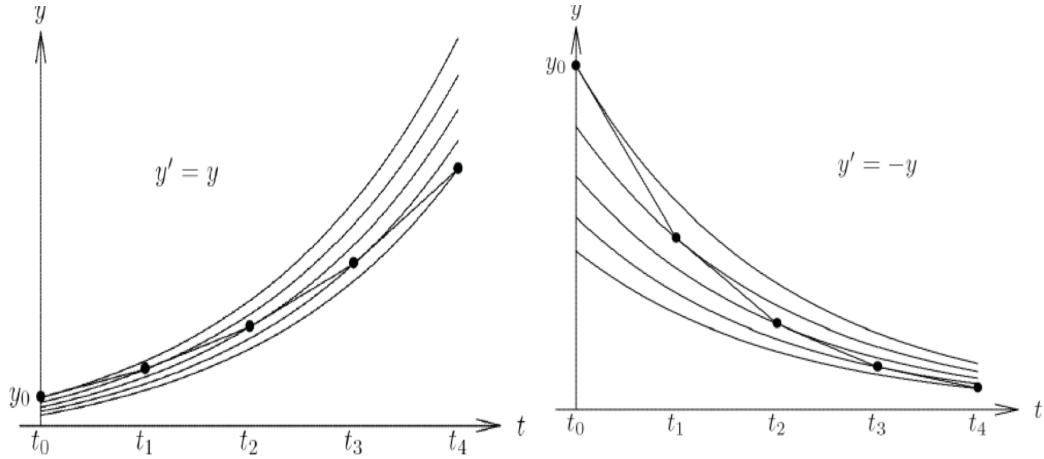


Figura 2.6: Idea del comportamiento y de su desviación del método de Euler progresivo cuando nos encontramos en sistemas contractivos o no.

Luego $\partial_y f > 0$ es en principio malo puesto que amplifica el error previo mientras que

$$\partial_y f(t_n, \xi_n) < 0 \Rightarrow 1 + h\partial_y f(t_n, \xi_n) < 1$$

y tener $\partial_y f < 0$ y tomar h tal que $|1 + h\partial_y f(t_n, \xi_n)| < 1$ es en principio bueno ya que ayuda a reducir errores. Esta restricción sobre h se denomina **restricción de estabilidad**.

Definición 49 Podemos decir que el carácter local de la ecuación diferencial es

- **expansivo** cuando $\partial_y f > 0$
- **contractivo** cuando $\partial_y f < 0$

y como nuestro esquema numéricico depende de f de una forma tan directa, podemos reproducir esta definición para el esquema, ver Figura 2.6

Definición 50 El esquema de Euler progresivo tiene un carácter

- **expansivo** cuando $\partial_y f > 0$
- **contractivo** cuando $\partial_y f < 0$

Vamos a ver la variación final en términos de la inicial. Si no tenemos mucho control sobre $\partial_y f$ podemos simplemente acotarla y usar una desigualdad de tipo Lipschitz

$$|u_{n+1} - v_{n+1}| \leq |u_n - v_n| + h L_f |u_n - v_n| = (1 + h L_f) |u_n - v_n|.$$

Observación 45 A la desigualdad

$$|u_{n+1} - v_{n+1}| \leq (1 + hL_f) |u_n - v_n|$$

se le denomina **estimación de estabilidad** y al factor

$$1 + hL_f$$

se le llama **factor de amplificación** (es una versión más ruda y general del factor $1 + h\partial_y f(t_n, \xi_n)$ que ya hemos visto).

Iterando recursivamente llegamos a

$$|u_n - v_n| \leq (1 + hL_f)^n |u_0 - v_0|.$$

Como siempre es $1 + hL_f > 1$, no podemos hacer más que intentar compensar n y h usando que $t_n = nh \leq T$. Por lo tanto, como $1 + x \leq e^x$, $x \geq 0$ podemos concluir

$$|u_n - v_n| \leq e^{(t_n - t_0)\lambda} |u_0 - v_0| \leq e^{T\lambda} |u_0 - v_0| \quad n = 0, 1, 2, \dots, N \quad (2.3)$$

que es la relación de **estabilidad** clásica y básica para el método de Euler explícito. Si $|u_0^h - v_0^h| \rightarrow 0$ entonces $|u_n^h - v_n^h| \rightarrow 0$ para todo $n = 0, 1, \dots, N = T/h$.

Observación 46 Aquí no nos hemos fijado en el signo de $\partial_y f$. Luego es una estimación válida en cualquier caso pero también puede estar muy lejos del valor real que se quiere acotar si $\partial_y f$ es negativa.

Recordemos, se fija $T > 0$ y nuestro intervalo de cálculo va a ser $[t_0, t_0 + T]$. Introducimos una **partición** de $[t_0, t_0 + T]$ que suponemos uniforme, para simplificar, $t_n = t_0 + nh$ para $h = T/N$ con $n = 0, 1, 2, \dots, N$ y la llamaremos Π_h , esto es

$$\Pi_h = \{t_0 < t_1 < \dots < t_N = t_0 + T, \quad t_n = t_n^h = t_0 + nh, \quad h = T/N\}.$$

Vamos a comparar una función continua, la solución, en los puntos $y(t_n^h)$ con unos valores discretos calculados en los puntos de la discretización u_n^h y buscamos que sea

$$u_n^h \approx y(t_n^h).$$

Estamos en disposición de estimar el error global en cada paso dado por

$$e_n = y(t_n) - u_n, \quad n = 0, 1, 2, \dots, N.$$

Para simplificar, definimos el **error global** $E(h) = \max_n |e_n|$ y vamos a estudiar su comportamiento cuando reducimos el valor de h , o lo que es lo mismo, aumentamos N , pero manteniendo siempre $hN = T$. Recordemos la definición de convergencia

Definición 51 Convergencia y orden: Diremos que el método es convergente cuando

$$\max_{1 \leq n \leq N^h} |y(t_n^h) - u_n^h| \rightarrow 0, \quad h \rightarrow 0^+, \quad h = T/N^{(h)}$$

siendo $t_n^h = t_0 + h n$. Cuando además para $p \geq 1$

$$\max_{0 \leq n \leq N^{(h)}} |y(t_n^h) - u_n^h| = \mathcal{O}(h^p), \quad h \rightarrow 0^+$$

se dice que converge con orden p .

2.5.3. Estimación de error usando la gráfica

Vamos a considerar una solución cualquiera $y(t)$ de la edo $u'(t) = f(t, u(t))$ y estimar el error que cometemos sobre esta solución. De acuerdo a la Figura 2.5 vamos a acotar el error $y(t_{n+1}^h) - u_{n+1}^h$ o, simplificando la notación, $y(t_{n+1}) - u_{n+1}$. El primer error generado en el paso t_{n+1} aparece cuando suponemos que $u_0 \neq y(0)$ y es el error producido por el esquema numérico al empezar en los puntos distintos u_0 e $y(0)$ (este error se puede visualizar como el primer segmento de trazo fino empezando desde la parte inferior hacia arriba). De acuerdo a la estabilidad del esquema numérico se acota por

$$(1 + hL_f)^{n+1} |u_0 - y(t_0)|$$

A continuación, el siguiente segmento vertical proviene de aplicar el esquema numérico en los dos primeros (de arriba a abajo) puntos iniciales en la vertical t_1 , su distancia es el error local $l(y(t_1); h)$. Por lo tanto, el segundo segmento de trazo fino (de abajo a arriba) viene a estar controlado por

$$(1 + hL_f)^n |l(y(t_1); h)| = (1 + hL_f)^n h |\mathcal{T}(y(t_1); h)|.$$

De forma análoga, el tercero por

$$(1 + hL_f)^{n-1} |l(y(t_2); h)| = (1 + hL_f)^{n-1} h |\mathcal{T}(y(t_2); h)|.$$

y así sucesivamente. El último segmento, el de trazo grueso es simplemente el error local $l(y(t_n); h) = h \mathcal{T}(y(t_n); h)$. Por lo tanto, el error global $y(t_{n+1}) - u_{n+1}$ está acotado por

$$|y(t_{n+1}) - u_{n+1}| \leq (1 + hL_f)^{n+1} |u_0 - y(t_0)| + h \sum_{j=1}^{n+1} (1 + hL_f)^{n-j+1} |\mathcal{T}(y(t_j); h)|$$

donde se ve claramente que el error total en el paso $n + 1$ está controlado por

- la propagación por parte del esquema numérico del error inicial

$$(1 + hL_f)^{n+1} |u_0 - y(t_0)|$$

observar que el número de pasos donde este error aparece coincide con la potencia $n + 1$ del factor $1 + hL_f$. Es decir, con el número de pasos que lleva este error introducido en el esquema.

- La propagación por parte del esquema numérico de los errores locales y su suma es

$$\sum_{j=1}^{n+1} (1 + hL_f)^{n-j+1} h |\mathcal{T}(y(t_j); h)|,$$

observar también que el número de pasos donde este error aparece coincide con la potencia $n - j + 1$ del factor $1 + hL_f$. Es decir, con el número de pasos que lleva este error introducido en el cálculo del esquema. Tomando el valor máximo $|\mathcal{T}(y; h)| = \max_j |\mathcal{T}(y(t_j); h)|$ podemos escribir

$$\sum_{j=1}^{n+1} (1 + hL_f)^{n-j+1} h |\mathcal{T}(y(t_j); h)| \leq |\mathcal{T}(y; h)| h \sum_{j=1}^{n+1} (1 + hL_f)^{n-j+1}$$

La estabilidad del esquema numérico equivale a decir que todos los coeficientes $(1 + hL_f)^j$ para $j = 1, 2, \dots, N = T/h$ están uniformemente acotados de una forma independiente de h , esto es,

$$(1 + hL_f)^j \leq e^{jhL_f} \leq e^{TL_f}.$$

Por otro lado, su acumulación N veces también debe de estarlo (recordemos que $hN = T$) y para eso se usa que $\mathcal{T}(y; h) = l(y; h)/h = O(h)$. Entonces

$$\begin{aligned} \sum_{j=1}^{n+1} (1 + hL_f)^{n-j+1} h |\mathcal{T}(y(t_j); h)| &\leq |\mathcal{T}(y; h)| h \sum_{j=1}^{n+1} e^{TL_f} \leq |\mathcal{T}(y; h)| h N e^{TL_f} \\ &= Te^{TL_f} |\mathcal{T}(y; h)| = O(h) \rightarrow 0, \quad h \rightarrow 0. \end{aligned}$$

Observación 47 Tener que $\mathcal{T}(y(t); h) = O(h^p)$ con $p \geq 1$ es coherente con el hecho de que se debe aproximar a una curva tanto en sus valores puntuales como en sus valores de la derivada, es decir,

$$\mathcal{T}(y(t); h) = \frac{l(y(t); h)}{h} = \frac{y(t+h) - y(t)}{h} - f(t, y(t))$$

y debe de ser

$$\mathcal{T}(y(t); h) \rightarrow 0, \quad h \rightarrow 0$$

es decir

$$\frac{y(t+h) - y(t)}{h} \rightarrow f(t, y(t)) = y'(t), \quad h \rightarrow 0.$$

Debemos usar la **consistencia del esquema numérico** para controlar los errores locales de una forma eficiente y que permita acotar su suma. Usando el máximo de los errores locales tomamos

$$\max_t \mathcal{T}(y(t); h) = \mathcal{T}(y; h)$$

y en nuestro caso de Euler explícito sabemos que tenemos

$$\mathcal{T}(y; h) = O(h) = C h$$

donde $C = C(f, T)$ o $C = C(y'', T)$. Entonces

$$|y(t_{n+1}) - y_{n+1}| \leq (1 + hL_f)^{n+1} |y_0 - y(t_0)| + h\mathcal{T}(y; h) \sum_{j=1}^{n+1} (1 + hL_f)^{n-j+1}.$$

Hacemos la suma geométrica del segundo miembro, y usamos que $1 + hL_f \leq e^{hL_f}$ para poder controlar h y n en términos del producto nh :

$$\begin{aligned} |y(t_{n+1}) - y_{n+1}| &\leq (1 + hL_f)^{n+1} |y_0 - y(t_0)| + \frac{(1 + hL_f)^{n+1} - 1}{1 + hL_f - 1} h\mathcal{T}(y; h) \\ &\leq (1 + hL_f)^{n+1} |y_0 - y(t_0)| + \frac{(1 + hL_f)^{n+1} - 1}{hL_f} h\mathcal{T}(y; h) \\ &\leq e^{h(n+1)L_f} |y_0 - y(t_0)| + \frac{e^{h(n+1)L_f} - 1}{L_f} \mathcal{T}(y; h) \end{aligned}$$

Observamos entonces que tenemos que poder controlar el valor $\mathcal{T}(y; h) = h^{-1}l(y; h)$, es decir, el valor $Nl(y; h)$, las acumulaciones $N = T/h = O(h^{-1})$ veces del error de consistencia local. Como en el caso de Euler explícito esto si ocurre así, tenemos

$$\begin{aligned} |y(t_{n+1}) - y_{n+1}| &\leq e^{(t_{n+1}-t_0)L_f} |y_0 - y(t_0)| + \frac{e^{(t_{n+1}-t_0)L_f} - 1}{L_f} h C \\ &\leq e^{TL_f} |y_0 - y(t_0)| + \frac{e^{TL_f} - 1}{L_f} h C \\ &= C \{ |y_0 - y(t_0)| + h \} \end{aligned}$$

para una constante C distinta de las de antes e independiente de h dada por

$$C = C(T, f, L_f) = e^{TL_f} + \frac{e^{TL_f} - 1}{L_f} C$$

y donde hemos usado $nh = t_n - t_0 \leq T$ y sirve para $n = 0, 1, 2, \dots, N - 1$.

La **acumulación de errores de locales** ha generado un coeficiente delante del máximo de ellos $l(y; h)$ en la forma

$$\frac{e^{t_{n+1}L_f} - 1}{hL_f} \leq \frac{e^{TL_f} - 1}{hL_f} = O(h^{-1}), \quad h \rightarrow 0^+$$

esto es razonable puesto que tomamos el mayor coeficiente y lo sumamos N veces, pero $N \approx h^{-1}$. Así que nos viene imprescindible y necesario el hecho de que el error local máximo sea de la forma $O(h^{p+1})$ con $p > 0$ para poder así controlar este coeficiente. Podemos entonces concluir que el método de Euler explícito converge con orden 1 con respecto al parámetro h . Hemos probado

Teorema 52 Convergencia del método de Euler explícito *Supongamos que f es globalmente Lipschitz con $f \in C^1$, luego $y \in C^2$. Para toda función $y(t)$ solución de $u'(t) = f(t, u(t))$ con $y(t_0) = \alpha$, si el error inicial $e_0 = |u_0 - \alpha|$ cumple $e_0 = O(h)$ entonces*

$$|u_n - y(t_n)| = O(h), \quad n = 1, 2, \dots, N.$$

donde $y(t)$ es la curva solución del problema de Cauchy con dato inicial $y(t_0) = \alpha$. Lo que es lo mismo, existe una constante $C > 0$ tal que

$$|u_n - y(t_n)| \leq Ch, \quad (h \rightarrow 0^+), \quad n = 1, 2, \dots, N$$

para $N = T/h$ por lo que el método de Euler explícito converge con orden 1 con respecto al parámetro h . La constante C depende de la solución continua $y(t)$ que se está aproximando y es proporcional a e^{TL_f} por lo que es extremadamente grande para valores moderados del producto TL_f , aunque puede ser precisa.

Observación 48 Como

$$\mathcal{T}(y; h) = \frac{h}{2}y''(\xi^h), \quad \xi^h \in (t, t+h)$$

va a ser imposible que el error del método de Euler sea más pequeño que $O(h)$, es decir, para toda curva solución la acotación de error es de la forma $\approx Ch$. Sólo cuando $y''(t) \equiv 0$, es decir, si la solución buscada es de la forma $y(t) = at + b$. Entonces el método de Euler explícito es exacto en cada paso y los errores vendrán de los datos iniciales o de las imprecisiones que se cometan en cada paso del cálculo.

2.5.4. Estimación de error usando Taylor

Se obtiene el mismo resultado sin recurrir al esquema gráfico usando el desarrollo de Taylor. De esta forma se ve menos lo que está ocurriendo, pero el resultado es el mismo.

Teorema 53 *El método de Euler explícito converge con orden 1 con respecto a h.*

Dem: Tomemos una curva cualquiera dentro del campo de velocidades, $w = w(t)$, esto es, la curva $w(t)$ cumple

$$w'(t) = f(t, w(t)), \quad t > t_0, \quad w(t_0) = \alpha \quad \text{dado.}$$

El esquema es:

Dado $y_0 \approx w(t_0)$ obtener

$$y_{n+1} = y_n + h f(t_n, y_n), \quad n = 0, 1, 2, \dots, N-1.$$

Sabemos que $w \in C^2([t_0, t_0 + T])$ ya que suponemos $f \in C^1$ al menos. De acuerdo al desarrollo de Taylor tenemos

$$w(t_{n+1}) = w(t_n) + h w'(t_n) + \frac{1}{2} h^2 w''(\xi_n), \quad t_n < \xi_n < t_{n+1},$$

es decir,

$$w(t_{n+1}) = w(t_n) + h f(t_n, w(t_n)) + \frac{1}{2} h^2 w''(\xi_n), \quad t_n < \xi_n < t_{n+1}.$$

El valor $l(w(t_n); h) = \frac{1}{2} h^2 w''(\xi_n)$ es **el residuo**, o **el error local**, que surge cuando se pretende que la solución exacta cumpla el esquema numérico

$$w(t_{n+1}) = w(t_n) + h f(t_n, w(t_n)) + l(t_n; h)$$

Si ponemos $\tilde{w}_{n+1}^h = w(t_n) + h f(t_n, w(t_n))$ entonces $w(t_{n+1}) = \tilde{w}_{n+1}^h + l(w(t_n); h)$. En cada iteración tenemos

$$e_{n+1} := w(t_{n+1}) - y_{n+1} = w(t_n) + h f(t_n, w(t_n)) + \frac{1}{2} h^2 w''(\xi_n) - y_n - h f(t_n, y_n).$$

Ya se puede agrupar como

$$e_{n+1} := w(t_{n+1}) - y_{n+1} = w(t_n) - y_n + h \{ f(t_n, w(t_n)) - f(t_n, y_n) \} + \frac{1}{2} h^2 w''(\xi_n)$$

y acotar. Pero también se puede escribir, una interpretación geométrica más clara,

$$e_{n+1} := w(t_{n+1}) - y_{n+1} = (w(t_{n+1}) - \tilde{w}_{n+1}^h) + (\tilde{w}_{n+1}^h - y_{n+1})$$

de donde

- $w(t_{n+1}) - \tilde{w}_{n+1}^h$ representa el error producido por el esquema numérico en un sólo paso

- $\tilde{w}_{n+1}^h - y_{n+1}$ representa la propagación del error acumulado en el paso t_n por el esquema numérico de todos los errores previos generados por los errores de consistencia locales.

Necesitamos la convergencia a cero de ambos sumandos. Observemos que

$$e_{n+1} = e_n + h\{f(t_n, w(t_n)) - f(t_n, y_n)\} + l(w(t_n); h)$$

y usando simplemente la propiedad de Lipschitz sobre f

$$|e_{n+1}| \leq (1 + hL_f) |e_n| + l(w; h) = (1 + hL_f) |e_n| + h\mathcal{T}(w; h)$$

donde $\mathcal{T}(w; h) = \max_{t_0 < t_n < t_0 + T} \mathcal{T}(w(t_n); h) \leq \max_{t_0 < t < t_0 + T} |w''(t)|h/2 = O(h)$.

La desigualdad clave entonces es

$$|e_{n+1}| \leq (1 + hL_f) |e_n| + h\mathcal{T}(w; h), \quad n \geq 0. \quad (2.4)$$

o bien

$$|w(t_{n+1}) - y_{n+1}| \leq (1 + hL_f) |w(t_n) - y_n| + h\mathcal{T}(w; h), \quad n \geq 0.$$

Iterando en $n \geq 1$

$$|e_1| \leq (1 + hL_f) |e_0| + h\mathcal{T}(w; h)$$

luego

$$|e_2| \leq (1 + hL_f) |e_1| + h\mathcal{T}(w; h) \leq (1 + hL_f)^2 |e_0| + ((1 + hL_f) + 1)h\mathcal{T}(w; h)$$

y recursivamente llegamos a

$$|e_{n+1}| \leq (1 + hL_f)^{n+1} |e_0| + [1 + (1 + hL_f) + \dots + (1 + hL_f)^n]h\mathcal{T}(w; h).$$

de donde fácilmente, usando la suma geométrica

$$\begin{aligned} |e_{n+1}| &\leq (1 + hL_f)^{n+1} |e_0| + \frac{(1 + hL_f)^{n+1} - 1}{(1 + hL_f) - 1} h\mathcal{T}(w; h) \\ &\leq (1 + hL_f)^{n+1} |e_0| + \frac{(1 + hL_f)^{n+1} - 1}{L_f} \mathcal{T}(w; h) \\ &\leq e^{L_f t_{n+1}} |e_0| + \frac{e^{L_f t_{n+1}} - 1}{L_f} \mathcal{T}(w; h) \end{aligned}$$

que es una cota creciente en cada tiempo de cálculo t_{n+1} . Tomando la mayor de las cotas posibles, es decir, usando que $t_{n+1} \leq t_0 + T$ nos encontramos con que para $0 \leq n \leq N - 1$ se tiene

$$|e_{n+1}| \leq e^{L_f T} |e_0| + \frac{e^{L_f T} - 1}{L_f} \mathcal{T}(w; h)$$

es decir

$$|e_{n+1}| \leq C\{|e_0| + \mathcal{T}(w; h)\}, \quad n \geq 0. \quad (2.5)$$

o bien

$$|w(t_{n+1}) - y_{n+1}| \leq C\{|w(t_0) - y_0| + \mathcal{T}(w; h)\}, \quad n \geq 0.$$

Concretando, usando lo que se sabe sobre $\mathcal{T}(w; h)$, tenemos

$$|e_{n+1}| \leq e^{L_f(t_{n+1}-t_0)} |e_0| + \frac{e^{L_f(t_{n+1}-t_0)} - 1}{2L} \max_{t_0 < t < t_0 + T} |w''(t)| h$$

o simplemente

$$|e_{n+1}| \leq C\{|e_0| + h\}$$

donde $C = C(f, w, T)$ es una constante que depende de todos los datos del problema incluida la curva continua w que se intenta aproximar, o si uno desea, simplemente las primeras derivadas de f .

Observar que **hasta aquí no se ha planteado ninguna restricción sobre h puesto que se va a considerar el límite $h \rightarrow 0$** . Para tener convergencia necesitamos entonces $|e_0| \rightarrow 0$ para $h \rightarrow 0$. Y si además $|e_0| \sim C_0 h$, generamos un error también proporcional a h . Así podemos respetar la potencia h que sale del segundo término y concluir

$$|e_n| \leq C h, \quad 0 \leq n \leq N = T/h$$

luego el método es de orden uno. La constante $C \sim e^{TL}$ puede sobreestimar el error en varios ordenes de magnitud y no debe usarse en la práctica ya que puede ser demasiado pesimista. Pero desde el punto de vista teórico ayuda a indicar el orden de convergencia. La convergencia es hacia la única curva solución del problema de Cauchy que tiene como dato inicial $w(t_0) = \alpha$, es decir, hacia la curva $w(t)$ que cumple

$$w'(t) = f(t, w(t)), \quad w(t_0) = \alpha \quad \text{dado.}$$



Observación 49 Las acotaciones (2.4) y (2.5) son comunes en todos los métodos numéricos que vamos a estudiar y se pueden conseguir gracias a la estabilidad del método. Esta estabilidad garantiza una acumulación controlada de los errores que se generan en cada paso.

Las dos fuentes de error que dependen de nosotros (una vez fijado el método numérico, en este caso Euler Explícito) son

1. el error inicial $|e_0|$,
2. el error de truncatura del método $\mathcal{T}(h)$ graduado por el valor de h y que depende del método.

Normalmente es el valor de h el que se fija e intentamos que el error sea descrito en términos de h . Por lo tanto, vamos a suponer que $|e_0|$ también depende de h y, observando el segundo término, parece razonable pedir, para alguna constante $C > 0$ fija que sea

$$|e_0| \leq C h.$$

Esto es, que el error que cometamos esté controlado por h . Esto se abrevia usando

$$e_0 = O(h), \quad h \rightarrow 0^+$$

o diciendo que e_0 converge a cero con orden 1 con respecto a h . Observar que si no se respeta la potencia de h en el error inicial, por ejemplo, $|e_0| \sim C_0 h^{1/2}$, no conseguimos que el error global sea $O(h)$ puesto que tendremos algo como $C_0 h^{1/2} + C_1 h = O(h^{1/2})$.

Corolario 54 Si $e_0 = O(h)$, $h \rightarrow 0^+$ entonces el **método de Euler es de primer orden** $p = 1$ con respecto al parámetro $h > 0$.

2.5.5. Observaciones sobre la convergencia

Varios puntos interesantes para la comprensión del resultado se pueden destacar.

Observación 50 El uso de la constante de Lipschitz es la forma más usual de presentar el resultado de convergencia del método de Euler explícito. En el caso de un problema expansivo está ajustada a los cálculos mientras que en un problema disipativo no distingue el decaimiento, la **constante de Lipschitz es insensible al signo**. Realmente se usa una cota global de $\partial_y f$ y esta función puede variar mucho a lo largo de toda la región de cálculo (en el plano (t,y)).

Cuando $\partial_y f < 0$ se puede usar $0 < 1 + h \partial_y f < 1$ para h lo suficientemente pequeño y mejorar la estimación de error. Aunque no merece la pena el esfuerzo teórico y se sigue usando $|1+h \partial_y f| \leq 1+hL_f$ obteniendo también convergencia en el límite $h \rightarrow 0^+$. Esto da una cota de error que puede denominar, de forma despectiva,

como perteneciente a la **era pre-computacional** por los valores impracticables que se pueden generar en las constantes que aparecen. En todo caso, el camino es simple y, de forma coloquial podemos decir que, como se trabaja poco, se recibe poco.

Si, por ejemplo, $L_f = 10$ y $T = 3$ el valor $e^{30} \approx 10^{13}$ no es manejable desde el punto de vista práctico. Además, si $e_0 \approx 10^{-4}$ (que puede ser razonable) resulta que la amplificación sólo de este error es $e^{30}|e_0| \approx 10^9$. Por esto, se dice que es una estimación de la **era precomputacional**: Si $h \rightarrow 0^+$ el método converge, pero de acuerdo a la estimación del error obtenido, el valor necesario de h para realizar una computación efectiva no es practicable.

Si por ejemplo, suponemos $|e_0| \approx h$ tenemos para $n = 0, 1, 2, \dots, N - 1$

$$|e_{n+1}| \leq e^{TL_f} h + M \frac{e^{TL_f} - 1}{L_f} h = O(h) \quad h \rightarrow 0^+$$

luego de forma aproximada se tiene $|e_n| \approx e^{TL_f} h$. Luego, para un error total menor que, por ejemplo, 10^{-3} con los datos anteriores debe ser $h < 10^{-3} 10^{-13} = 10^{-16}$ que es prácticamente el cero del computador.

Observación 51 Hemos perdido una potencia de h en el error local máximo debido a la acumulación de errores locales. Esto tiene sentido puesto que cada error local es proporcional a h^2 y necesitamos tomar desde $t = t_0$ a $t = t_0 + T$ un número de pasos igual a $N = T/h$ cada uno de ellos de longitud h . Entonces la acumulación de N valores, cada uno de ellos proporcional a h^2 , esto es, $O(h^2)$, genera un error global $O(Nh^2)$ y como $N = T/h$ nos quedamos con una contribución final al error de la forma $O(h)$. Esta acotación se consigue gracias a la estimación de estabilidad que tenemos a nuestra disposición ya que permite mantener las constantes de amplificación bajo control, aunque sean muy grandes.

Observación 52 Observar que si $\partial_y f \leq 0$ entonces tenemos una mejor estimación ya que

$$e_{n+1} = (1 + h\{\partial_y f\})e_n + h\mathcal{T}(w; h)$$

y si $|1 + h\partial_y f| < 1$, es decir,

$$h < 2 / \max\{|1 + h\partial_y f|\} \tag{2.6}$$

entonces usando que $h\mathcal{T}(w; h) = l(w; h)$,

$$|e_{n+1}| \leq |e_n| + l(w; h)$$

$$|e_{n+1}| \leq |e_0| + \frac{1}{2} \max_{t_0 < t < t_0 + T} |w''(t)| h.$$

La restricción sobre h (2.6) se denomina de **restricción de estabilidad**.

Al ser el error de la forma $E(h) \approx Ch$ si, por ejemplo, dividimos h entre 2 y calculamos en una partición más fina entonces para el error tenemos

$$E\left(\frac{h}{2}\right) \approx C \frac{h}{2} \approx \frac{E(h)}{2}.$$

Por lo tanto, el error obtenido es el previo pero dividido por 2.

Observación 53 *Hay dos partes en la acotación del error:*

- El primer sumando contiene el efecto del error inicial. Este error se amplifica debido a que se hace un cálculo iterativo

$$(1 + hL_f)^{n+1} |w(t_0) - y_0^h| \leq e^{(t_{n+1}-t_0)L_f} |w(t_0) - y_0^h| \leq e^{TL_f} |w(t_0) - y_0^h|.$$

- El segundo término contiene el efecto del residuo en cada paso. Estos residuos se acumulan y van apareciendo debido a que la solución exacta buscada, $w(t)$ no cumple con el esquema numérico planteado (le sobra el residuo)

$$\frac{(1 + hL_f)^{n+1} - 1}{2L_f} \max_{t_0 < t < t_0 + T} |w''(t)| h \sim e^{TL_f} h$$

Se ve claramente que este término no aparece si $w'' \equiv 0$ y que si buscamos una curva solución $z(t)$ distinta a la curva $w(t)$ el valor constante delante de h cambia puesto que aparecerá $\max_{t_0 < t < t_0 + T} |z''(t)|$ en vez de $\max_{t_0 < t < t_0 + T} |w''(t)|$. Aunque esto no cambia el orden uno de convergencia.

- *En términos de potencias de h el equilibrio óptimo se da cuando*

$$|w(t_0) - y_0^h| = O(h)$$

puesto que el segundo sumando no se puede mejorar y es $O(h)$. Esto quiere decir que el esfuerzo en aproximar el error inicial es suficiente con que sea igual al del orden del método usado, en este caso $O(h)$. Podriamos exigir $|w(0) - y_0|$ muy pequeño en relación con h , por ejemplo $O(h^2)$, pero sería inútil cuando aparece el otro término de error, ya que no podemos mejorar la potencia uno sobre h , y será esta la que mande, ya que $O(h^2) + O(h) = O(h)$.

Observación 54 *Cuando $w''(t) \equiv 0$ ($w(t) = a + bt$) entonces la solución particular buscada $w(t)$ cumple el esquema de forma exacta, esto es,*

$$w(t_{n+1}) = w(t_n) + h f(t_n, w(t_n)), \quad \forall n \geq 0.$$

Por lo tanto, si tomamos $y_0 = w(t_0)$ entonces $y_{n+1} = w(t_{n+1})$ para todo $n \geq 0$. En cambio, si $y_0 \neq w(t_0)$ sólo nos queda la primera parte de la estimación de error y tenemos la acotación

$$|w(t_{n+1}) - y_{n+1}| \leq (1 + hL_f)^{n+1} |w(t_0) - y_0| \leq e^{(t_{n+1}-t_0)L_f} |w(t_0) - y_0|.$$

Esto equivale a lanzar el esquema desde dos puntos iniciales distintos $w(t_0)$ e y_0 y estimar cuánto se alejan entre sí los valores obtenidos.

Observación 55 Si el error inicial es cero, $|w(t_0) - y_0| = 0$, tendremos

$$|e_{n+1}| \leq C_{n+1} h$$

donde

$$C_{n+1} \leq \frac{e^{(t_{n+1}-t_0)L_f} - 1}{2L_f} \max_{t_0 < t < t_0 + T} |w''(t)|$$

crece conforme se avanza en tiempo, puede tener un valor máximo dado por (usando $1 + x \leq e^x$)

$$C = \max_{t_0 < t < t_0 + T} |w''(t)| \frac{e^{L_f T} - 1}{2L_f}.$$

Como se observa, la constante C depende de

- del campo de velocidades presente a través de L_f cota superior de $\partial_y f(t, y)$
- de la curva solución buscada dentro de este campo mediante $w''(t)$
- de la longitud del intervalo de tiempo pedido T
- y finalmente del esquema numérico usado que lo amalgama todo generando el error local en cada punto.

Observación 56 El error de truncatura: Un error de truncatura local $O(h)$ implica error global $O(h)$. Más adelante buscaremos esquemas donde este error de truncatura local sea más pequeño, esto es, tenga una potencia de h más grande. En general, el error de truncatura local tendrá la forma

$$\mathcal{T}(w; h) \sim C h^p$$

donde C es una constante que dependerá de derivadas de alto orden de la función solución exacta buscada y el exponente p marcará el orden p de convergencia del método.

2.6. Interpretación de la cota de error

Que un método sea convergente con orden p quiere decir que su **error global es siempre menor o igual que $C h^p$ para todas las soluciones regulares de cualquier campo y alguna constante C que depende del problema**. Esta estimación no se puede disminuir, es decir, no se tiene $C h^q$ con $q > p$ para **toda solución regular de cualquier campo** ya que siempre existe alguna solución donde el error es igual a $C h^p$.

Es decir, fijado un método de orden p vamos a:

1. suponer que tenemos cualquier función $w(t)$ solución de un campo de velocidades con la regularidad suficiente.
 2. denotar por $E_{w(t)}(h)$ el error cometido al aplicar el método,
- entonces se cumple siempre

$$E_{w(t)}(h) \leq C h^p, \quad \forall w(t) \text{ solución suficientemente regular.}$$

Además, existe al menos una solución regular $\tilde{w}(t)$ y $C \neq 0$ tal que

$$E_{\tilde{w}(t)}(h) = C h^p \Rightarrow \frac{E_{\tilde{w}(t)}(h)}{h^{p+1}} = \frac{C}{h} \rightarrow +\infty$$

es decir donde no se puede hacer más pequeño el error (más grande la potencia p). Normalmente, esto se puede verificar con el simple problema test $y'(t) = y(t)$ donde $f(t, y) = y$ y la mayoría de las derivadas de f se hacen cero, lo que simplifica la expresión del error local de consistencia.

Esto no implica que en algún caso particular sí que se pueda tener un error más pequeño o incluso cero, es decir, pueden existir u solución de un problema de Cauchy y $r > p$ tal que

$$E_{u(t)}(h) = C h^r < C h^p$$

o incluso

$$E_{u(t)}(h) = 0 < C h^p.$$

Esto ocurrirá cuando se anulen términos en el desarrollo de Taylor de forma particular y debido a la forma de $u(t)$ y de $f(t, u)$. Pero esta anulación de términos no siempre ocurrirá y tendremos en el caso general soluciones como $\tilde{w}(t)$ donde el error no va a mejorar en orden.

2.7. Ejemplos

Vamos a necesitar el siguiente resultado técnico:

Lema 1 *Se cumple*

$$\frac{\log(1+x)}{x} = 1 - \frac{x}{2} + O(x^2), \quad x \rightarrow 0.$$

Esto se ve fácilmente mediante un desarrollo de Taylor de $\log(1+x)$ en $x = 0$

Tenemos que

$$e^{T L_f} h \leq \epsilon \Rightarrow h < e^{-T L_f} \epsilon$$

luego conseguir que esta estimación de error sea de utilidad práctica para obtener errores pequeños implica tener h extremada y exponencialmente pequeño. Esto lleva a valores de n muy grandes que pueden ser innecesarios. Vamos a ver algunos ejemplos que ilustren esta idea

Ejemplo 55 Caso donde la estimación de error SI es ajustada: Sea $f(t, y) = \lambda y$ con $\lambda > 0$. Entonces $L_f = \partial_y f(t_n, \xi_n) = \lambda$. La solución exacta es $y(t) = e^{\lambda t}$ si usamos $y(0) = 1$; el esquema de Euler explícito en un intervalo $[0, T]$ usando N puntos y tomando $h = T/N$ es

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

para $n = 0, 1, 2, \dots, N - 1$, o lo que es lo mismo,

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

recursivamente queda como

$$y_n = (1 + h\lambda)^n, \quad 0 \leq n \leq N.$$

Podemos calcular el error en $t = nh$

$$\begin{aligned} e^{\lambda t} - (1 + h\lambda)^n &= e^{\lambda t} - e^{n \log(1+h\lambda)} = e^{\lambda t} - e^{nh\lambda \log(1+h\lambda)/(h\lambda)} \\ &= e^{\lambda t} - e^{t\lambda \log(1+h\lambda)/(h\lambda)}. \end{aligned}$$

pero sabemos que

$$\frac{\log(1+x)}{x} = 1 - \frac{x}{2} + O(x^2), \quad x \rightarrow 0.$$

de donde para h lo suficientemente pequeño, tenemos

$$\begin{aligned} e^{\lambda t} - (1 + h\lambda)^n &= e^{\lambda t} - e^{t\lambda(1-h\lambda/2+O(h^2\lambda^2))} = e^{\lambda t} - e^{t\lambda - th\lambda^2/2+tO(h^2\lambda^3)} \\ &= e^{\lambda t}(1 - e^{-th\lambda^2/2+O(h^2\lambda^3)}) \approx e^{\lambda t}(th\lambda^2/2) \end{aligned}$$

es decir, para $t = nh$

$$e^{\lambda t} - (1 + h\lambda)^n \approx \frac{1}{2}e^{\lambda t} t\lambda^2 h = K h, \quad h \rightarrow 0^+, \quad (K = \frac{1}{2}e^{\lambda t} t\lambda^2).$$

Luego en el caso $f(t, y) = \lambda y > 0$ con $\lambda > 0$ resulta que la estimación de error es **ajustada y la constante, así como el error, ambos crecen de forma exponencial, siendo esto correcto.**

Si nos fijamos en un intervalo $[0, T]$ tendremos si $t = nh$ que

$$e^{\lambda t} - (1 + h\lambda)^n \leq \frac{1}{2}e^{\lambda T} T\lambda^2 h, \quad \forall n = 0, 1, 2, \dots, N = T/h$$

luego para t fijo necesitamos h exponencialmente pequeño (luego n exponencialmente grande) si queremos tener un error pequeño. Así que, incluso cuando es correcta, puede ser inútil desde el punto de vista práctico ya que el error que se comete con el método de Euler explícito en estos casos puede ser muy grande.

Observamos también aquí que el error de Euler explícito es exactamente como Ch en este ejemplo con C exponencialmente grande, $C \approx e^{L_f T}$.

Ejemplo 56 Caso donde la estimación NO es ajustada: Sea $f(t, y) = -\lambda y$ con $\lambda > 0$. Entonces $L_f = \lambda$, pero $\partial_y f(t_n, \xi_n) = -\lambda$. La solución exacta es $y(t) = e^{-\lambda t}$; el esquema de Euler explícito en un intervalo $[0, T]$ usando N puntos y tomando $h = T/N$ es

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

para $n = 0, 1, 2, \dots, N-1$, o lo que es lo mismo,

$$y_{n+1} = (1 - h\lambda) y_n$$

recursivamente queda como

$$y_n = (1 - h\lambda)^n, \quad 0 \leq n \leq N.$$

La estimación de error en $t = nh$ nos dice

$$|e^{-\lambda t} - (1 - h\lambda)^n| \approx C e^{\lambda t} h, \quad h \rightarrow 0^+.$$

Pero podemos calcular el error en $t = nh$ ($y_0 = 1$): Supongamos $h < 1/\lambda$ para garantizar $0 < 1 - h\lambda < 1$. Entonces

$$\begin{aligned} e^{-\lambda t} - (1 - h\lambda)^n &= e^{-\lambda t} - e^{n \log(1-h\lambda)} = e^{-\lambda t} - e^{-nh\lambda \log(1-h\lambda)/(-h\lambda)} \\ &= e^{-\lambda t} - e^{-t\lambda \log(1-h\lambda)/(-h\lambda)}. \end{aligned}$$

de donde para h lo suficientemente pequeño, tenemos

$$\begin{aligned} e^{-\lambda t} - (1 - h\lambda)^n &\approx e^{-\lambda t} - e^{-t\lambda(1-h\lambda/2)} = e^{-\lambda t} - e^{-t\lambda + th\lambda^2/2} \\ &= e^{-\lambda t}(1 - e^{th\lambda^2/2}) \approx e^{-\lambda t}(-th\lambda^2/2) \end{aligned}$$

es decir,

$$e^{-\lambda t} - (1 - h\lambda)^n \approx -\frac{1}{2}e^{-\lambda t} t\lambda^2 h, \quad h \rightarrow 0^+.$$

Luego en el caso $f(t, y) = -\lambda y > 0$ con $\lambda > 0$ resulta que la constante en el error decrece de forma exponencial con el tiempo mientras que la estimación de error obtenida usando la constante de Lipschitz crece de forma exponencial. Por lo tanto, aquí la estimación de error es totalmente inservible. Observamos también aquí que el error de Euler explícito es exactamente como Ch en este ejemplo con C exponencialmente pequeña, $C \approx e^{-L_f T}$.

Observación 57 Si para $h > 0$ fijo se quiere que el cálculo $y_n = (1 - h\lambda)^n$ se parezca a $y(t_n) = e^{-\lambda t_n}$, en el sentido de que decaiga puesto que así lo hace $e^{-\lambda t_n}$, tenemos que exigir $h < 2/\lambda$ y esto es la **restricción de estabilidad** sobre h .

Observación 58 En el caso $y(t_n) = e^{\lambda t_n}$ no hace falta imponer ninguna restricción sobre h puesto que los valores, tanto continuos $e^{\lambda t_n}$, como discretos $(1+h\lambda)^n$, crecen de forma exponencial.

Observación 59 Estos dos ejemplos se pueden comprobar computacionalmente con el siguiente **pseudo-código** cambiando el valor de λ de acuerdo a la situación buscada

```
Dados  $t_0, T, N, y_0 = 1, f(t, y) = \lambda y$ 
Calculamos  $h = T/N$ 
 $t = 0 : h : T;$ 
Para  $n = 0$  hasta  $N - 1$ 
    calcular  $y_{n+1} = y_n + h \lambda y_n$ 
Fin
//Constante de error en cada  $t(n)$  usando
cte = max(exp(t) - y)/h;
Dibujar puntos  $(t, cte)$ 
```

Ejemplo 57 En teoría buscamos el límite $h \rightarrow 0$, pero en la práctica se debe calcular con h fijo y positivo. Para este método no vale cualquier valor de h . Es interesante observar el comportamiento del esquema numérico para un valor fijo de h puesto que sabemos que hay convergencia cuando $h \rightarrow 0$. Pero esto es la teoría y necesitamos el aspecto práctico de la implementación en el computador del esquema. Volvamos al problema donde sabemos que la solución exacta

$$y(t) = e^{-\lambda t}, \quad t > 0$$

tiene un comportamiento decreciente de manera uniforme y rápida, cuanto más grande λ más rápido. Sabemos que

$$y_n = (1 - h\lambda)^n, \quad n \geq 0$$

Los valores calculados sólo reproducen el comportamiento cuantitativamente si $0 < h < 2/\lambda$ puesto que entonces

$$|1 - h\lambda| < 1$$

y

$$y_n = (1 - h\lambda)^n, \quad n \geq 0$$

decae, aunque se pueden producir oscilaciones decrecientes que no es lo que hace la solución exacta. Si pedimos además que sea $0 < h < 1/\lambda$ sí que obtenemos el comportamiento cuantitativo y cualitativo de decaimiento sin oscilaciones puesto que

$$0 < 1 - h\lambda < 1$$

y se garantiza un decaimiento uniforme de un valor inicial positivo, tal y como hace el problema continuo, ver Figura 2.7. Evidentemente, en el caso $h > 2/\lambda$ los valores calculados divergen de los buscados. Resumiendo, tenemos tres situaciones

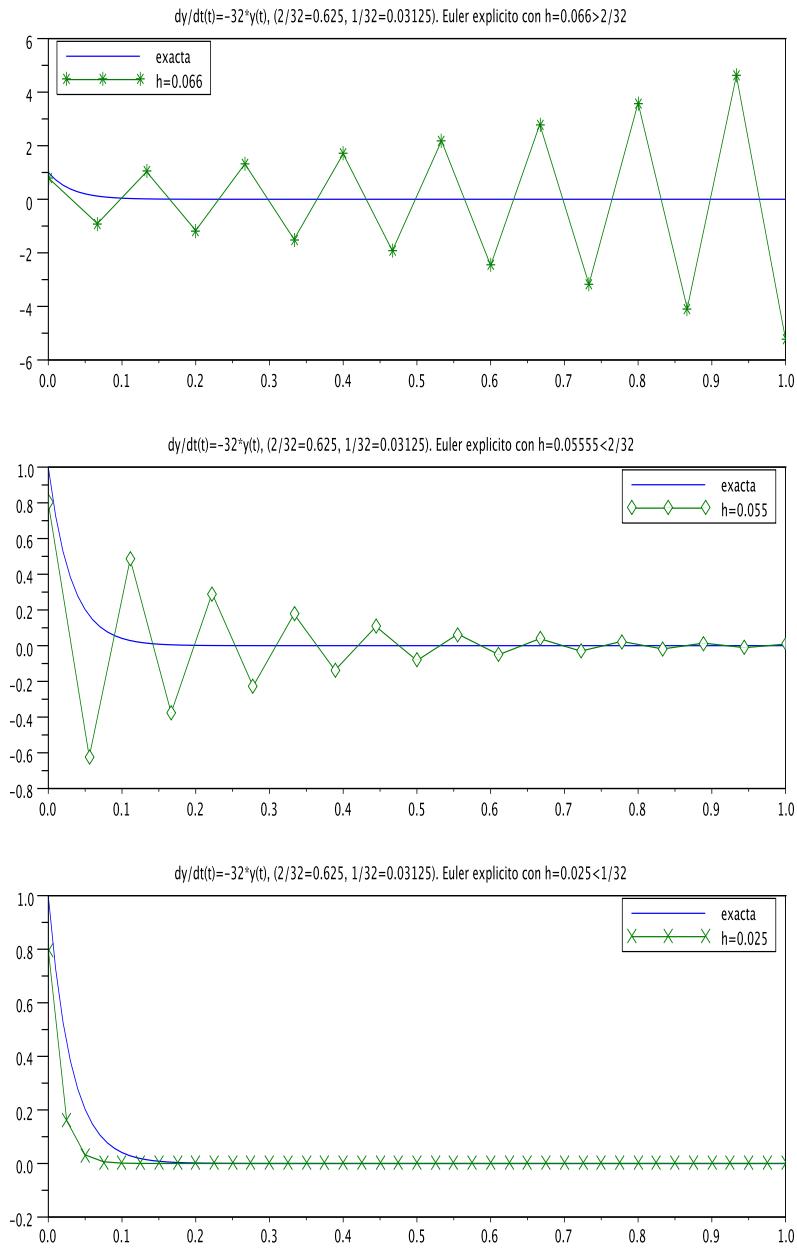


Figura 2.7: Euler explícito para $y' = -\lambda y$ con $\lambda = 32$. Aproximaciones generadas de acuerdo a $h > 2/\lambda$, $0 < h < 2/\lambda$ ó $0 < h < 1/\lambda$. Por defecto y para visualizar mejor se unen los puntos calculados con un trazo continuo, pero el cómputo sólo genera los puntos.

1. $h > 2/\lambda$ los valores calculados divergen aunque su crecimiento está controlado por Ch donde $C \sim e^{L_f T}$, cota nada útil pero matemáticamente cierta.
2. $0 < h < 2/\lambda$ los valores calculados reproducen el comportamiento cuantitativamente pero hay oscilaciones,
3. $0 < h < 1/\lambda$ los valores calculados reproducen el comportamiento cuantitativa y cualitativamente.

Ejemplo 58 Euler explícito no es $O(h^2)$ Para comprobar que el error de Euler explícito no es más pequeño que Ch en general, basta encontrar una curva solución de un problema de Cauchy donde se cumpla de forma estricta que el error es $O(h)$. Tendremos entonces que **no hay convergencia como h^2 a cero**. La aplicación que ya hemos visto en los ejemplos 55 y 56 para $y' = \lambda y$ nos sirve y también podemos encontrar algunos otros: veamos el siguiente ejemplo, aquí la segunda derivada es no nula y los cálculos se hacen a mano de forma simple:

$$w'(t) = t, \quad w(0) = 0$$

con solución

$$w(t) = t^2/2.$$

Si tomamos como valor de inicio $y_0 = 0$ y con $t_n = nh$ entonces se tiene que

$$\begin{aligned} y_1 &= y_0 + ht_0 = 0, \\ y_2 &= y_1 + ht_1 = h^2, \\ y_3 &= y_2 + ht_2 = h^2(1+2), \\ &\vdots \\ y_n &= h^2(1+2+\dots+(n-1)), \end{aligned}$$

es decir,

$$y_n = t_n^2/2 - \frac{t_n}{2}h$$

de donde

$$y(t_n) - y_n = \frac{t_n}{2}h.$$

Para ver entonces la convergencia en un punto $t = t_n$ fijo se realiza el límite para $h \rightarrow 0$ y $n \rightarrow +\infty$ tal que $hn = t$ fijo, esto se representa por el límite estacionario $\lim_{hn=t}$. Entonces resulta que

$$\lim_{hn=t} (y(t_n) - y_n^h) = \lim_{hn=t} t \frac{h}{2} = \frac{t}{2} \lim_{hn=t} h = 0.$$

pero evidentemente

$$\lim_{hn=t} \frac{y(t_n) - y_n^h}{h} = \frac{t}{2} \neq 0$$

luego $y(t_n) - y_n^h = O(h)$ pero $y(t_n) - y_n^h \neq O(h^2)$. Es decir, tenemos un ejemplo donde el error global es de orden uno con respecto a h y no de orden 2. Por lo que la estimación de error global obtenida en teoría no se puede mejorar.

Observación 60 Sólo si $w''(t) \equiv 0$ el orden lo marca el error inicial: En el método de Euler explícito el error local es la segunda derivada. Por lo tanto, el orden se mejora si $w''(t) \equiv 0$ y sólo en este caso. Pero no sólo es que se mejora... sino que en este caso es exactamente cero y la solución buscada cumple el esquema de forma exacta si se empieza en el valor exacto. Si $w''(t) \equiv 0$ entonces

$$w(t_{n+1}) = w(t_n) + h w'(t_n) + \frac{h^2}{2} w''(t_n) + \dots = w(t_n) + h w'(t_n) = w(t_n) + h f(t_n, w(t_n))$$

y si empezamos en $y_0 = w(t_0)$ el esquema es exacto ya que

$$y_1 = y_0 + h f(t_0, y_0) = w(t_0) + h f(t_0, w(t_0)) = w(t_1)$$

entonces

$$y_2 = y_1 + h f(t_1, y_1) = w(t_1) + h f(t_1, w(t_1)) = w(t_2)$$

y así sucesivamente. Luego de hecho el error es cero (y cero es $O(h^p)$ para cualquier $p \geq 0$). Tomemos el caso

$$w'(t) = 1, \quad w(0) = 0$$

donde $f(t, y) \equiv 1$, luego $L_f = 0$, y con solución

$$w(t) = t.$$

Si iniciamos con $y_0 = 0$ entonces el esquema es $y_{n+1} = y_n + h$ y fácilmente se tiene que

$$y_1 = 0 + h 1 = h = t_1, \quad y_2 = h + h 1 = 2h = t_2, \quad \dots, \quad y_n = nh = t_n$$

y en un punto concreto, si ponemos $t = t_n$ fijo resulta que

$$w(t) - y_n = 0.$$

Si ahora tomamos como valor de inicio $y_0 = \alpha \neq 0$ entonces fácilmente se tiene que

$$y_n = \alpha + nh = \alpha + t_n$$

y en un punto concreto, si ponemos $t = t_n$ fijo resulta que

$$w(t) - y_n = \alpha.$$

Por lo tanto, el error inicial, si existe, debe de ser del mismo orden que el del método. Por ejemplo, si $\alpha = \sqrt{h}$ se tiene que $y_0 = \sqrt{h} \rightarrow 0$ pero no lo hace con la rapidez que nos marca el método de Euler y la aproximación usada no es buena ya que se estropea, o se contamina, el orden de convergencia del esquema. Se tiene

$$E(h) = O(h^{1/2}) \neq O(h).$$

Ejemplo 59 Experimento numérico: Vamos a comprobar computacionalmente que el método de Euler es de primer orden con un par de problemas donde la segunda derivada no se anula. Consideremos

$$y'(t) = -y(t) + 1, \quad y(0) = 0 \quad (2.7)$$

con solución

$$y(t) = 1 - e^{-t}$$

y el problema

$$y'(t) = y(t), \quad y(0) = 1 \quad (2.8)$$

con solución

$$y(t) = e^t.$$

Podemos entonces realizar una tabla con los errores obtenidos

$$E_n = \max_{k \leq n} |y(t_k) - y_k|$$

para $n = 1/h$ con $n = 16, 32, 64, \dots$ (duplicándose) en el intervalo de tiempo $[0, 1]$. Vemos que los errores decaen por un factor de 2, confirmando numéricamente el orden 1 en h .

$n = h^{-1}$	error en (2.9)	p	error en (2.10)	p
16	0.0118053	---	0.0803533	---
32	0.0058242	1.019	0.0412917	0.9605
64	0.0028929	1.0095	0.0209369	0.97974
128	0.0014417	1.0047	0.0105428	0.98978
256	0.0007197	1.0023	0.0052902	0.9948
512	0.0003595	1.0014	0.0026498	0.9974
1024	0.0001797	1.0004	0.0013261	0.9987

Lema 2 Para $h \rightarrow 0$ y $n \rightarrow +\infty$ tal que $hn = t$ fijo se cumple

- $(1 - hL)^n \sim e^{-tL}$, es decir, $\lim_{hn=t} (1 - hL)^n = e^{-tL}$
- $(1 + hL)^n \sim e^{tL}$, es decir, $\lim_{hn=t} (1 + hL)^n = e^{tL}$

donde $\lim_{hn=t}$ representa $h \rightarrow 0$ y $n \rightarrow +\infty$ tal que $hn = t$ fijo.

Dem: Ejercicio. ■

Ejemplo 60 Cálculo explícito: Para $f(t, y) = 1000(1 - y) = -1000y + 1000$ tenemos

$$y_{n+1} = y_n + h 1000(1 - y_n) = (1 - h1000)y_n + 1000h, \quad n \geq 1$$

Entonces si $y_0 = 1$ se reproduce la solución $w(t) \equiv 1$. ¿Si $y_0 \neq 1$ qué ocurre? Viendo la recurrencia formada

$$y_n = (1 - h1000)^n y_0 + \frac{(1 - 1000h)^n - 1}{-1000h} 1000h$$

es decir

$$y_n = (1 - h1000)^n y_0 - (1 - 1000h)^n + 1 = (1 - h1000)^n (y_0 - 1) + 1$$

luego la diferencia entre la solución constante $w \equiv 1$ y la calculada es

$$y_n^h - 1 = (1 - h1000)^n (y_0 - 1)$$

y vemos como se amplia, en este caso, se reduce, el error inicial. Entonces para cada $t = nh$ tendremos en el límite

$$\lim_{hn=t} (y_n^h - 1) = e^{-1000t} (y_0 - 1).$$

Luego cuanto mayor sea $t = hn$ más rápido se aproxima el cálculo hacia la solución estacionaria en el límite $\lim_{hn=t}$, esto es, cuando disminuimos h aumentando n tal que $t = nh$ está fijo. Esto reproduce lo que el sistema continuo hace en el límite pero también lo hace fijado h tal que $h < 2/1000$ puesto que en este caso es

$$|1 - h1000| < 1$$

y el cálculo discreto decae a 1 de forma rápida, esta es la **restricción de estabilidad en este caso**.

Por otro lado, los campos $f(t, y) = \pm 1000(1 - y)$ poseen una solución de equilibrio $w(t) \equiv 1$ y el campo de vectores es muy contractivo o expansivo dependiendo del signo. En todo caso, al usar Euler explícito con $y_0 = 1$ se reproduce de forma exacta la solución $w(t) \equiv 1$.

Ejemplo 61 Cálculo explícito: Si $f(t, y) = -1000(1 - y) = 1000y - 1000$ tenemos

$$y_{n+1} = y_n + h (-1000(1 - y_n)) = (1 + h1000)y_n - 1000h, \quad n \geq 1.$$

otra vez no hay problema si $y_0 = 1$ pero si $y_0 \neq 1$ tenemos

$$y_n = (1 + h1000)^n y_0 - \frac{(1 + 1000h)^n - 1}{1000h} 1000h$$

es decir

$$y_n = (1 + h1000)^n (y_0 - 1) + 1$$

y vemos como se amplia, en este caso sí, el error inicial. Entonces para cada $t = nh$ tendremos en el límite

$$\lim_{hn=t} (y_n^h - 1) = e^{1000t} (y_0 - 1)$$

con lo que la solución buscada que empieza en $y_0 \neq 1$ se aleja bruscamente de la estacionaria conforme avanza el tiempo por muy cerca que esté y_0 de 1. Esto reproduce lo que el sistema continuo hace. Por otro lado, para cualquier valor de $h > 0$ fijo tenemos que $1 + h1000 > 1$ por lo que también reproduce en discreto el efecto continuo.

Ejemplo 62 Cálculo explícito: Un caso muy sencillo es

$$w'(t) = 0, \quad w(0) = 1$$

con solución

$$w(t) \equiv 1.$$

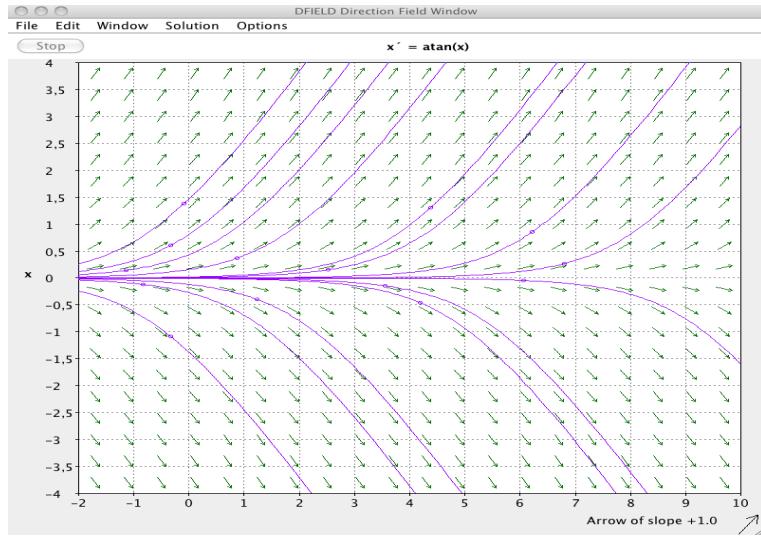
Aquí se tiene que $f(t, y) = 0$ y en cada punto (t, y) hay una pendiente cero. Si tomamos como valor de inicio $y_0 = 1$ entonces fácilmente $y_n = 1 = w(t_n)$ para cualquier n . En este caso la aplicación de Euler explícito es trivial

$$y_{n+1} = y_n + h0 = y_n.$$

Los ejemplos anteriores muestran que una misma curva puede estar en campos de velocidades totalmente distintos y que esto condiciona el rendimiento del método numérico.

Ejemplo 63 Un problema estable: Consideremos $y'(t) = \text{atan}(y(t))$, $y(0) = y_0$, ver Figura 2.8, entonces $f(t, y) = f(y) = \text{atan}(y)$ luego $\partial_y f(y) = (1 + y^2)^{-1} \leq 1$ y $L_f = 1$. Podemos además estimar $y''(t)$ derivando la ecuación y obtenemos $\|y''\|_\infty \leq \pi/2$. La aproximación de la solución del problema mediante el método de Euler explícito genera una estimación de error en un intervalo de tiempo $[0, T]$ dada por

$$|e_n| \leq e^T |e_0| + \frac{\pi}{4} (e^T - 1) h.$$

Figura 2.8: Soluciones para $y'(t) = \tan(y(t))$.

Si $T = 1$ y queremos limitar el error a un valor 0.001 tendremos que tomar h y el error inicial $|e_0|$ tales que

$$e^1|e_0| + \frac{\pi}{4}(e^1 - 1)h \leq 0.001.$$

por ejemplo, $e_0 = 0$ nos da $h < 0.00074$. Si fijamos un intervalo de cálculo más grande, la restricción sobre el parámetro h empeorará sustancialmente. Pero no podemos mejorar esta cota y deberíamos buscar un método con un orden mayor.

Ejemplo 64 Un problema muy inestable: Consideremos

$$\begin{cases} \frac{dy}{dt} = 10y + 11t - 5t^2 - 1, & 0 < t \leq 3, \\ y(0) = 0, \end{cases}$$

que tiene por solución $y(t) = t^2/2 - t$ como se comprueba fácilmente. Las solución general cuando $y(0) = y_0$ es

$$y(t) = y_0 e^{10t} + t^2/2 - t.$$

que incluye $y(t) = t^2/2 - t$ si $y_0 = 0$. En la Figura 2.9 se ven las soluciones y cómo la solución para $y_0 = 0$ queda oculta por la gran variación que producen aquellas en las que $y_0 \neq 0$.

Aquí tenemos que $f(t, y) = 10y + 11t - 5t^2 - 1$ por lo que la constante de Lipschitz para $f(t, y)$ con respecto a la segunda variable es $L = 10$. Por otro lado,

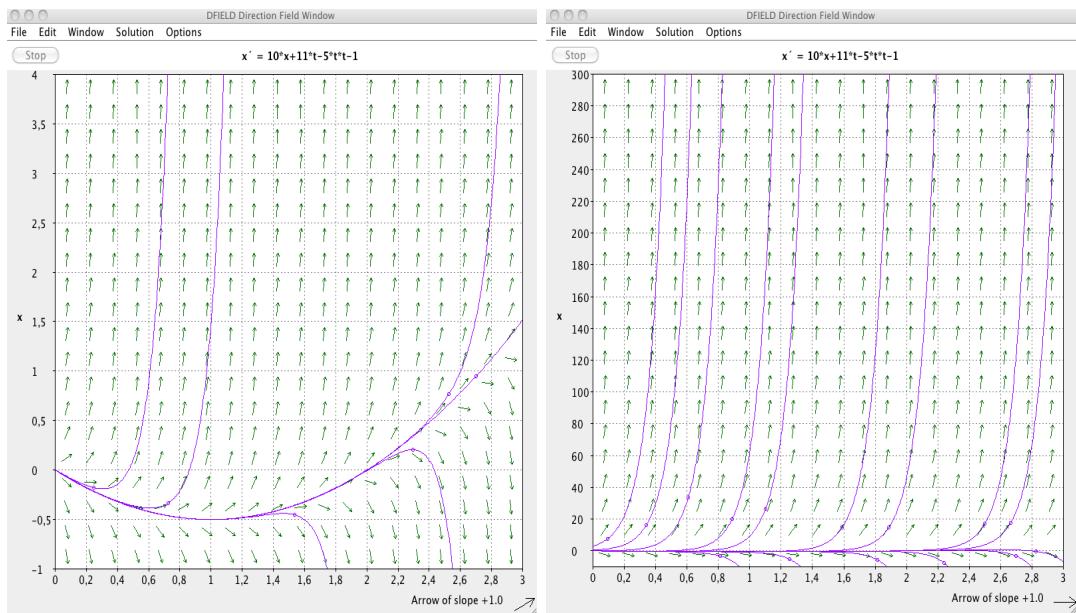


Figura 2.9: Campo de soluciones para $y'(t) = 10y + 11t - 5t^2 - 1$. Las soluciones son $y(t) = y_0 e^{10t} + t^2/2 - t$. Incluye $y(t) = t^2/2 - t$ pero no se ve por la gran variación producida en las otras soluciones.

como $y(t) = t^2/2 - t$ resulta que $y''(t) = 1$. Por último, si tomamos $y_0 = 0$ de tal forma que el error inicial es cero, tenemos la estimación de error

$$|y(t_n) - y_n^{(h)}| \leq \frac{h}{2} \frac{e^{30} - 1}{10}, \quad n = 1, 2, \dots, N^{(h)}$$

donde recordemos que $h N^{(h)} = T = 3$ y $t_n = nh$ para $n = 1, 2, \dots, N^{(h)}$. Entonces si queremos que sea, por ejemplo,

$$|y(t_n) - y_n^{(h)}| \leq 0.01, \quad n = 1, 2, \dots, N^{(h)}$$

estamos forzados a plantear

$$\frac{h}{2} \frac{e^{30} - 1}{10} \leq 0.01 \Rightarrow h < \frac{1}{5(e^{30} - 1)} \approx 1.87 \dots \cdot 10^{-14}.$$

Pero esto genera una restricción inabordable, **prácticamente** $h \approx 0$. Rebajar el intervalo de cálculo tampoco ayuda mucho. Por ejemplo si $T = 1$, tenemos entonces

$$\frac{h}{2} \frac{e^{10} - 1}{10} \leq 0.01 \Rightarrow h < \frac{1}{5(e^{10} - 1)} \approx 9.08 \dots \cdot 10^{-6}.$$

Ahora resulta que esto plantea una restricción $h \approx 10^{-5}$. Entonces tenemos que tomar entorno a $10^5 = 100.000$ puntos en la partición del intervalo $[0, 1]$ y esto es todavía muy costoso en términos de memoria de computador.

Pequeñas variaciones en el dato inicial pueden ser también muy importantes en el cálculo. Si aplicamos el método de Euler progresivo pero ahora hemos tomado $y_0^{(h)} = \epsilon > 0$ para ϵ pequeño, quizás por accidente. Entonces tenemos la estimación de error

$$|y(t_n) - y_n^{(h)}| \leq e^{30}\epsilon + \frac{h}{2} \frac{e^{30} - 1}{10}, \quad n = 1, 2, \dots, N^{(h)}$$

donde recordemos que $hN^{(h)} = T = 3$ y $t_n = nh$ para $n = 1, 2, \dots, N^{(h)}$. Ahora trabaja en nuestra contra también el término

$$e^{30}\epsilon \approx 10^{13}\epsilon.$$

Entonces, hacer pequeño este término sólo será posible para una elección de ϵ prácticamente igual al cero de la máquina

$$10^{13}\epsilon \leq 0.01 \Leftrightarrow \epsilon \leq 10^{-15}.$$

Observación 61 Una alternativa puede ser trabajar con **un método de mayor orden**. En este caso el error local máximo y el error de convergencia son menores. Supongamos que tenemos un método de cuarto orden y que la estimación de error es similar, entonces para $T = 1$ y si, por ejemplo, hemos obtenido un error local máximo $l(h) = h^{4+1}$ ocurre

$$h^4 \frac{e^{10} - 1}{10} \leq 0.01 \Rightarrow h < \left(\frac{10}{100(e^{10} - 1)} \right)^{1/4} = 0.046\dots$$

luego tenemos un valor de h más razonable gracias a la raíz cuarta.

En cuanto al error en el dato inicial, sólo podremos mejorar si tenemos el mismo orden en el error inicial.

Ejemplo 65 Problema rígido: La estimación de error nos dice lo mismo para el problema

$$\begin{cases} \frac{d}{dt}y(t) = -10y + 11t - 5t^2 - 1, & 0 < t \leq 3, \\ y(0) = 0, \end{cases}$$

pero el aspecto práctico computacional es totalmente distinto. El campo de soluciones se ve en la Figura 2.10 y todas las trayectorias se contraen en una.

Observación 62 Como consecuencia de estos ejemplos, un resultado teórico no se debe usar para estimaciones prácticas del error sin pensar si es ajustado o no. En lenguaje de la “Matemática Pura”, si $L T$ está acotado también lo está e^{LT} . Pero cuando tenemos que dar números y hacer cálculos... por ejemplo 50 no es un número grande, pero no parece muy correcto decir que e^{50} esté acotado. Algunos números son

$$e^{10} \sim 22026.466, \quad e^{50} \sim 5.185 \cdot 10^{21}$$

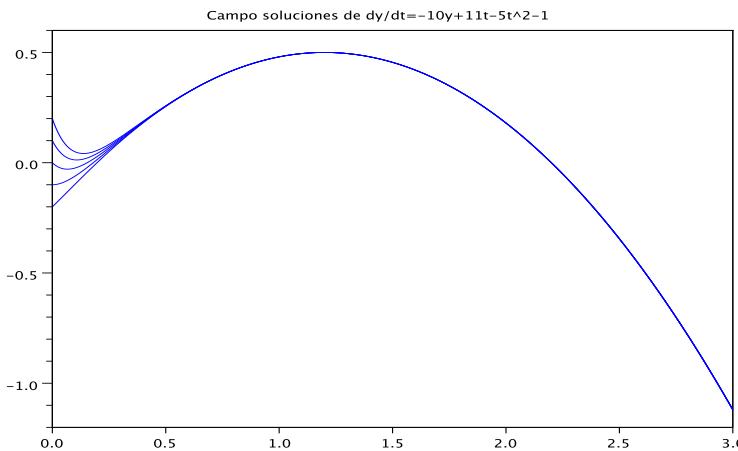


Figura 2.10: Campo de soluciones para $dy/dt = -10y + 11t - 5t^2 - 1$. Las soluciones ahora son $y(t) = (y_0 + 11/50)e^{-10t} - t^2/2 + 6t/5 - 11/50$. Incluye $y(t) = -t^2/2 + 6t/5$.

Vamos a terminar mejorando el resultado de convergencia visto para el caso de Euler explícito en este tipo de problemas:

Teorema 66 *Convergencia Euler explícito en el caso contractivo $w' = -aw$ con $a > 0$.*

Dem: Al igual que en la prueba normal, llegamos a

$$e_{n+1} = e_n + h\{f(t_n, w(t_n)) - f(t_n, y_n)\} + l(w(t_n); h)$$

pero $f(t, y) = -ay$ implica que $\partial_y f(t, y) = -a$ luego

$$e_{n+1} = (1 - ha)e_n + l(w; h)$$

y si tomamos $h < 1/a$ entonces $1 - ha > 0$ y tenemos en particular que

$$|e_{n+1}| < |e_n| + l(w; h).$$

Esto ya mejora la estimación entre errores globales en el caso general que sería tomando la constante de Lipschitz como $L = |\partial_y f(t, y)| = a$ sin observar el signo negativo delante de a y haciendo

$$|e_{n+1}| < (1 + ha)|e_n| + l(w; h).$$

Luego aquí el signo importa. Usando con detalle esta situación tenemos, para $h < 1/a$

$$|e_{n+1}| < (1 - ha)|e_n| + l(w; h)$$

donde $1 > 1 - ha > 0$ y siendo

$$l(w; h) = \max_{t_0 < t_n < t_0 + T} |l(w(t_n); h)| \leq \max_{t_0 < t < t_0 + T} |w''(t)| h^2 / 2$$

y recursivamente llegamos a

$$|e_{n+1}| \leq (1 - ha)^{n+1} |e_0| + \frac{(1 - ha)^{n+1} - 1}{(1 - ha) - 1} l(w; h),$$

es decir,

$$|e_{n+1}| \leq (1 - ha)^{n+1} |e_0| + \frac{1 - (1 - ha)^{n+1}}{a} \max_{t_0 < t < t_0 + T} |w''(t)| h,$$

Entonces, como $(1 - ha)^n \approx e^{-ta}$ que se puede terminar de acotar como

$$|e_n^h| \leq e^{-at_n^h} |e_0^h| + \frac{1}{a} \max_{t_0 < t < t_0 + T} |w''(t)| h.$$

En este caso, se tiene también orden uno en potencia de h si el error inicial es $O(h)$, pero se observa que todas las constantes involucradas decaen exponencialmente o están acotadas.

$$|e_n| \leq C h, \quad 0 \leq n \leq N = T/h$$

luego el método es de orden uno pero en este caso la constante mayor que se encuentra no crece exponencialmente sino que está acotada por 1. De hecho, en cada paso la constante de error es más pequeña y ayuda a que el error global acumulado vaya decreciendo. Tenemos una demostración matemática **útil y coherente con los resultados prácticos**. ■

Ejemplo 67 Si volvemos a considerar los problemas

$$y' + y = 1, \quad y(0) = 0 \tag{2.9}$$

con solución

$$y(t) = 1 - e^{-t}$$

y

$$y' = y, \quad y(0) = 1 \tag{2.10}$$

con solución

$$y(t) = e^t.$$

La Figura 2.11 muestra los errores en $[0, 2]$ calculados con $h = 2/2048 = 1/1024$ para ambos ejemplos. Podemos ver como para el problema (2.10) de solución $y(t) = e^t$, el menor error es justo al principio, de hecho el peor error es siempre el previo y luego crece, el crecimiento del error parece ser exponencial. Por otro lado, para el problema (2.9) de solución $y(t) = 1 - e^{-t}$, el peor error ocurre cerca del principio y luego decae. Estos resultados concuerdan con el análisis de convergencia más detallado que hemos visto.

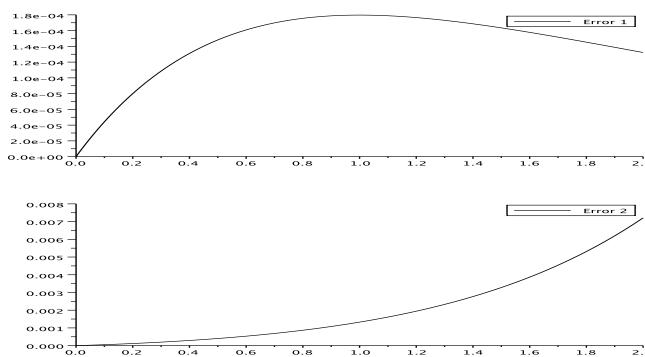


Figura 2.11: Comportamiento del error con el método de Euler en $[0, 2]$ para $h = 2/2048$ para los problemas (2.9) y (2.10) respectivamente.

Observación 63 En los ejemplos ya vistos

$$\begin{cases} y'(t) = -10y + 11t - 5t^2 - 1, \\ y(0) = 0, \end{cases} \quad \begin{cases} y'(t) = -100y, \\ y(0) = 1, \end{cases}$$

tiene sentido el uso del TVM puesto que se captura el signo negativo que tiene $\partial_y f(t, y)$. Pero debemos de observar que en los problemas

$$\begin{cases} y'(t) = 10y + 11t - 5t^2 - 1, \\ y(0) = 0, \end{cases} \quad \begin{cases} y'(t) = 100y, \\ y(0) = 1, \end{cases}$$

el resultado que se obtiene es el mismo que si se usa la condición de Lipschitz. Por lo tanto, aunque la estimación tenga esos factores e^{TL} , realmente es la situación práctica que aparece, el error crece muy rápido por que el sistema es expansivo, ver Figura 2.6. Podemos estimar los errores para estos problemas y ver como el aspecto práctico concuerda con el teórico

Como conclusión podemos decir que es en el caso $\partial_y f(t, y) << 0$ donde es realmente interesante usar el TVM puesto que se es capaz de capturar un comportamiento decreciente de la solución y esto no se consigue con la constante de Lipschitz. Este tipo de problemas son **disipativos, rígidos o contractivos** y su característica principal es que las trayectorias del campo de velocidades se contraen en tiempo.

Suelen aparecer con frecuencia en aplicaciones físicas debido a la presencia del rozamiento o viscosidad de los medios físicos que hacen que la energía se disipe y las trayectorias decaigan a cero o se contraigan. Ejemplos como los visualizados en la Figura 2.6 son representativos de estas situaciones.

2.7.1. Resumiendo

Observaciones sobre Euler explícito:

- Cota de error en un intervalo de longitud T de la forma $K h$ con $K \approx e^{TL_f}$ siendo $L_f = \max_{t,y} |\partial_y f(t, y)|$. Por lo tanto, si $h \rightarrow 0$ el error tiende a cero (aunque le costará llegar por culpa de la gran magnitud de K).
- En los casos donde $\partial_y f(t, y) < 0$ podemos obtener cotas con constantes que decaen en tiempo pero necesitamos restringir el valor de h en una forma que depende del problema en cuestión. Esta es una restricción conocida por **restricción de estabilidad**.

2.8. Problemas rígidos

Existen situaciones muy frecuentes donde la solución del problema representa procesos que se desarrollan a velocidades muy distintas: hay **escalas temporales muy diferentes**.

Ejemplo 68 La solución ya vista del problema $y'(t) = -10y + 11t - 5t^2 - 1$ es

$$y(t) = (y_0 + 11/50)e^{-10t} - t^2/2 + 6t/5 - 11/50$$

mientras que la solución de $z' = -10z$ es

$$z(t) = z_0 e^{-10t}.$$

En ambos problemas tenemos la misma restricción $h < 2/10$ para tener cálculos estables. A partir de ahí, si $h \rightarrow 0$ dan lugar a convergencia.

Si nos fijamos en la solución $z(t) = z_0 e^{-10t}$ como patrón, vemos que lo que hace falta en la solución ligeramente más complicada

$$y(t) = z_0 e^{-10t} - t^2/2 + 6t/5 - 11/50, \quad (z_0 = y_0 + 11/50)$$

es la capacidad de calcular bien la parte que corresponde a e^{-10t} . Escribamos

$$y(t) = (y_0 + 11/50)e^{-10t} - t^2/2 + 6t/5 - 11/50 = F(t) + S(t)$$

donde

$$F(t) = (y_0 + 11/50)e^{-10t}, \quad S(t) = -t^2/2 + 6t/5 - 11/50.$$

La parte $F(t)$ decrece muy rápidamente mientras que $S(t)$ evoluciona a una velocidad más lenta. Podemos decir que:

- $F(t)$ es el **modo transitorio rápido** y
- $S(t)$ el **modo transitorio lento o estacionario**.

Vemos por tanto que si el problema es rígido

en el caso particular del esquema de Euler explícito es la necesidad de calcular bien el modo rápido lo que impone la restricción sobre h . Esta restricción se mantiene durante todo el intervalo incluso cuando ya el valor del modo rápido casi ha desaparecido

Ejemplo 69 Ecuación de Dahlquist-Bjorck. *Este ejemplo fue propuesto en torno a 1974 para ilustrar estos fenómenos. El problema es*

$$\begin{cases} y'(t) = 100(\sin(t) - y(t)), & 0 < t \leq 3, \\ y(0) = 0 \end{cases}$$

y posee como solución

$$y(t) = e^{-100t}y_0 + \frac{\sin(t) - 100^{-1}\cos(t) + 100^{-1}e^{-100t}}{1 + 100^{-2}}.$$

El modo transitorio exponencial e^{-100t} decae muy rápido. Este modo debe ser capturado bien por el método explícito y esto nos lleva a un valor de h excesivamente pequeño. Se puede ver en las Figuras 2.12 y 2.13 el comportamiento del método de Euler explícito en comparación con el método de Euler implícito que veremos en el tema siguiente. Aquí es $\partial_y f = -100$ y debemos de usar $h < 2/100$ en Euler explícito, luego si trabajamos en $[0, 3]$ debe ser

$$N > 3 * 100/2 = 150$$

para que Euler explícito funcione bien.

Ejemplo 70 Ecuación de Prothero-Robinson: parecido al ejemplo anterior, también es muy conocido e ilustra los mismos conceptos:

$$\begin{cases} y'(t) = L(\varphi(t) - y(t)) + \varphi'(t), & 0 < t, \\ y(0) = y_0 \end{cases}$$

la solución exacta es

$$y(t) = e^{-Lt}(y_0 - \varphi(0)) + \varphi(t)$$

y otra vez el modo rápido e^{-Lt} para $L \gg 1$ debe ser capturado correctamente. Aquí, si $\varphi(t) \equiv y(t)$ tenemos un problema trivial. La diferencia $\varphi(t) - y(t)$ es lo que se introduce para perturbar el problema. La función $\varphi(t)$, es el modo estacionario, y puede ser una función suave sin cambios bruscos.

Incluso en el caso en el que $y_0 = \varphi(0)$, la solución es $\varphi(t) \equiv y(t)$ y aparentemente el modo rápido no está presente en la solución, sí que se encuentra en todas las soluciones vecinas y se debe también capturar como si estuviese presente, ver Figura 2.14. Esto se debe a que el campo de soluciones lo describe la función

$$f(t, y) = L(\varphi(t) - y) + \varphi'(t)$$

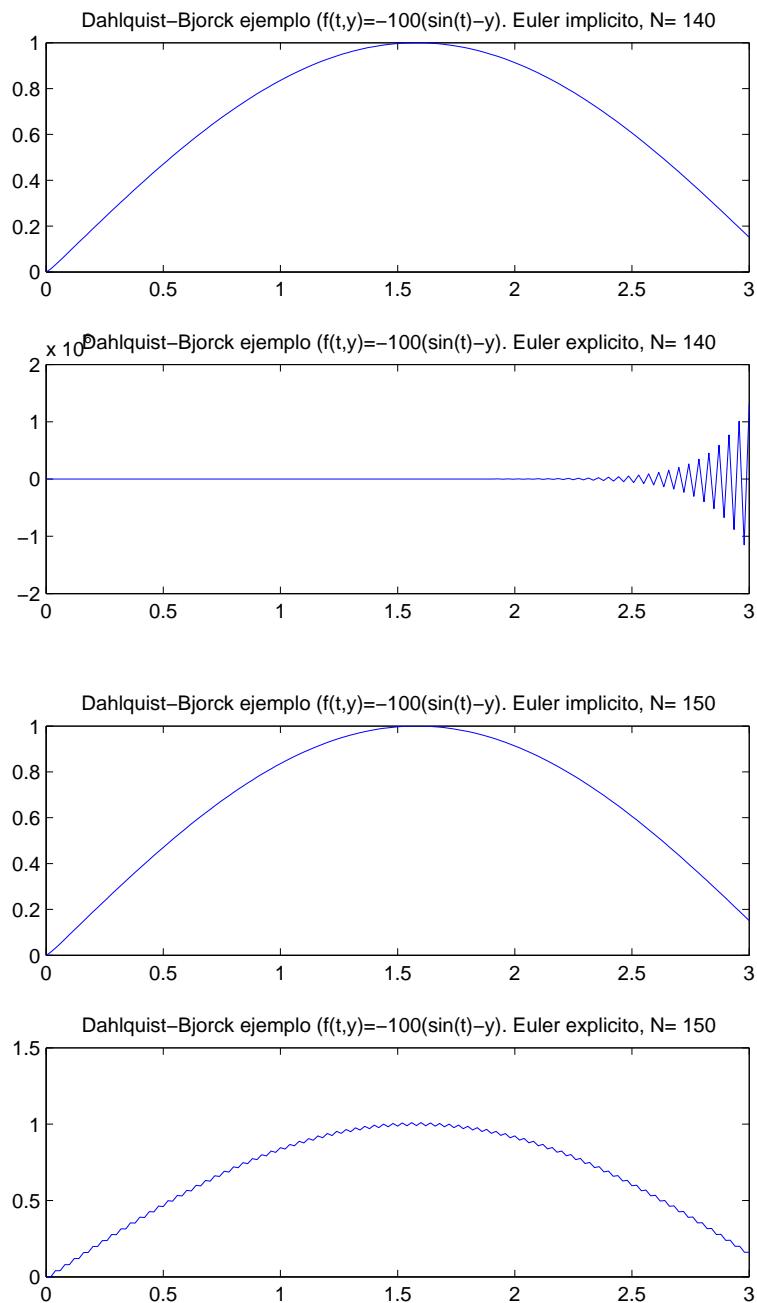


Figura 2.12: Cálculo de la solución del problema de Dahlquist-Bjorck usando Euler explícito y Euler implícito con $N = 140$ y $N = 150$

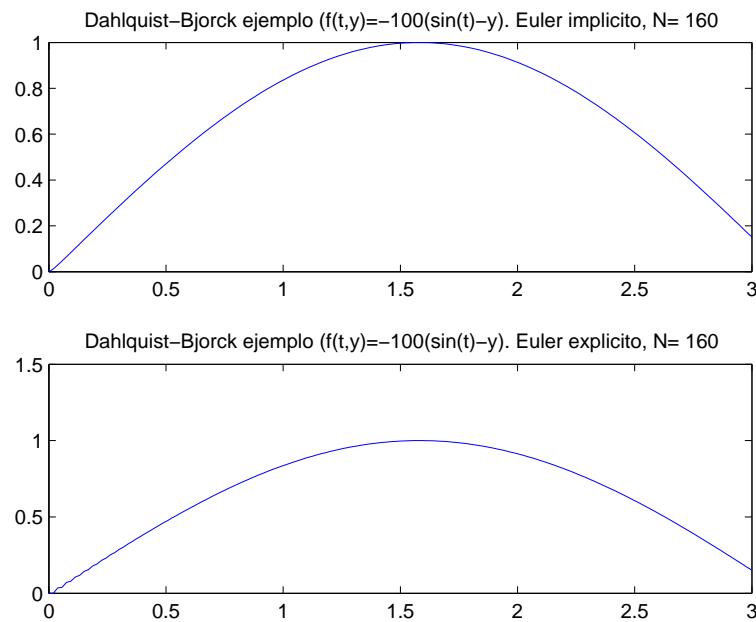


Figura 2.13: Cálculo de la solución del problema de Dahlquist-Bjorck usando Euler explícito y Euler implícito con $N = 160$

y tenemos que

$$\partial_y f(t, y) = -L, \quad (L \gg 1).$$

Luego con $L \gg 1$ las pendientes que aparecen en el campo son muy grandes. Cualquier mínima diferencia con la solución exacta que se busca es amplificada por la ecuación continua y lanzada fuera de la curva buscada. Para la Figura 2.15 el dato en $t = 0$ para las curvas que caen sobre la fase transitoria está muy lejos de $y_0 = 1$. De hecho, esta “lluvia de curvas” se ha obtenido usando tiempos iniciales $t_0 = 0.5, 1, 1.5$ y valores $y(t_0)$ cercanos al valor de la fase transitoria en estos tiempos.

Observación 64 Es como marchar por una senda en una cumbre muy estrecha y con mucha pendiente, cualquier paso en falso te lanza muy abajo ($L \gg 0$). Al contrario que estar en una senda en un valle muy estrecho, por mucho que te des contra la pared no te sales del desfiladero ($L \ll 0$).

Observación 65 Podemos ver como la rigidez es también un concepto de eficiencia. Con un simple método de Euler explícito podemos obtener la solución pero necesitamos reducir mucho el parámetro de discretización y por consiguiente aumentamos mucho la carga de trabajo computacional. Por lo tanto, no se es eficiente.

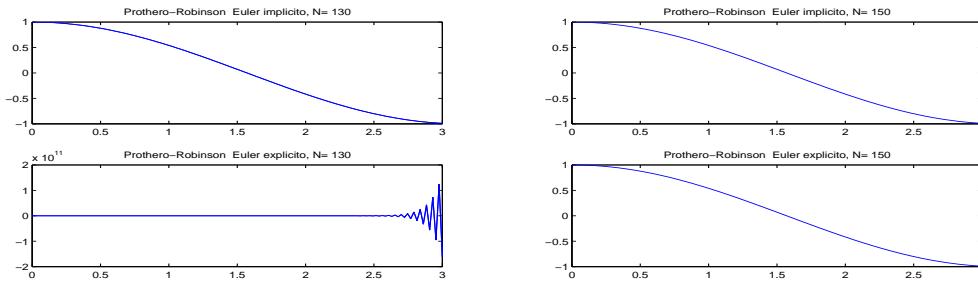


Figura 2.14: Euler Explícito e implícito sobre la ecuación de Prothero-Robinson

2.9. Euler explícito en sistemas

Vamos a generalizar al caso vectorial muchos de los conceptos ya vistos. En este caso la solución representa tantas curvas como ecuaciones y estas curvas evolucionan en tiempo de una forma relacionada entre sí. Podemos aplicar el método de Euler progresivo de una manera natural.

Tenemos $y'(t) = f(t, y(t)) \in \mathbb{R}^m$ con $m > 1$, para $t > 0$ y con $y(0) = y_0 \in \mathbb{R}^m$. De manera explícita tenemos las ecuaciones

$$\begin{cases} y'_1(t) &= f_1(t, y_1(t), y_2(t), \dots, y_m(t)), \\ y'_2(t) &= f_2(t, y_1(t), y_2(t), \dots, y_m(t)), \\ \vdots &\vdots \\ y'_{m-1}(t) &= f_{m-1}(t, y_1(t), y_2(t), \dots, y_m(t)), \\ y'_m(t) &= f_m(t, y_1(t), y_2(t), \dots, y_m(t)), \end{cases}$$

y puesto que cada f_j depende de $(t, y_1, y_2, \dots, y_m)$ el sistema está **acoplado**. Vamos a suponer $f \in C^1$, entonces acotando el Jacobiano de f con respecto a la variable vectorial y , nos encontramos con la condición de Lipschitz, que vamos a suponer global,

$$\|f(t, y) - f(t, z)\| \leq L \|y - z\|$$

donde $\|\cdot\|$ es una norma cualquiera en \mathbb{R}^m . Esto permite tener garantizado existencia y unicidad siendo además la solución y de clase C^2 . El método de Euler progresivo para aproximar la solución en el intervalo $[t_0, t_0 + T]$, o bien $[0, T]$, es simplemente

$$y_{n+1} = y_n + h f(t_n, y_n), \quad n = 0, 1, 2, \dots, N-1 \quad (N = T/h)$$

y se describe explícitamente para $y_n = (y_{1,n}, y_{2,n}, \dots, y_{m,n})$ y $f = (f_1, f_2, \dots, f_m)$ como:

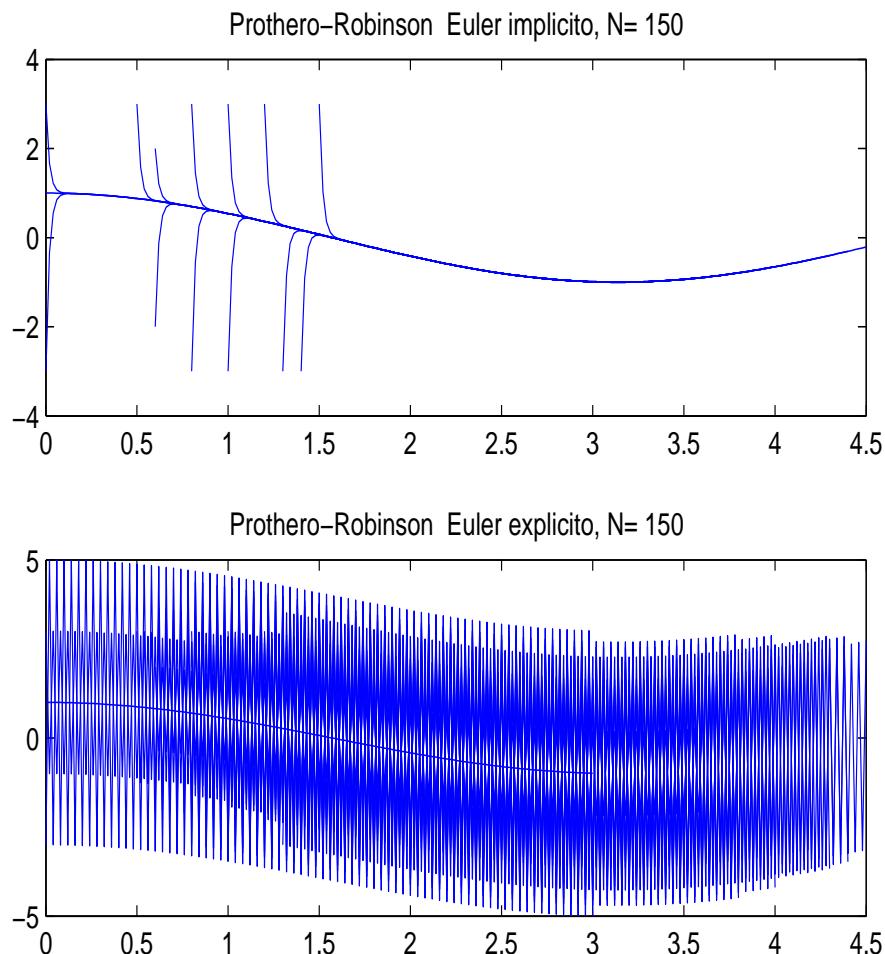


Figura 2.15: Comportamiento de las curvas vecinas a la fase transitoria en la ecuación de Prothero-Robinson. Cálculo posible con Euler implícito pero imposible con Euler explícito. La “lluvia de curvas” se ha obtenido usando tiempo inicial $t_0 = 0.5, 1, 1.5$ y valores $y(t_0)$ cercanos al valor de la fase transitoria en estos tiempos.

Dado $y_0 = (y_{0,1}, y_{0,2}, \dots, y_{0,m}) \in \mathbb{R}^m$ calcular y_{n+1} para $n \geq 0$ como

$$\begin{aligned} y_{1,n+1} &= y_{1,n} + h f_1(t_n, y_{1,n}, y_{2,n}, \dots, y_{m,n}), \\ y_{2,n+1} &= y_{2,n} + h f_2(t_n, y_{1,n}, y_{2,n}, \dots, y_{m,n}), \\ &\vdots \\ y_{m,n+1} &= y_{m,n} + h f_m(t_n, y_{1,n}, y_{2,n}, \dots, y_{m,n}). \end{aligned}$$

Observación 66 Usar un valor de h distinto para cada ecuación no nos sirve ya que generaría valores en puntos distintos para cada ecuación y tendríamos que hacer procesos de interpolación. Esto llevaría a más errores acumulados en el cálculo.

A continuación, prácticamente repetimos los pasos dados en el caso escalar. Definimos el **error global** $e = \max_n \|e_n\|$ siendo

$$e_n = y(t_n) - y_n, \quad n = 0, 1, 2, \dots, N$$

y vamos a estudiar su comportamiento cuando reducimos el valor de h , o lo que es lo mismo, aumentamos N , pero manteniendo siempre $hN = T$.

Definición 71 Convergencia: Diremos que el método es convergente cuando

$$\max_{0 \leq n \leq N^h} \|y(t_n^h) - y_n^h\| \rightarrow 0, \quad h \rightarrow 0^+, \quad h = T/N^{(h)}$$

siendo $t_n^h = t_0 + hn$.

Observación 67 Podemos tomar una norma cualquiera $\|\cdot\|$ en \mathbb{R}^m puesto que la dimensión m está fija y todas son equivalentes en \mathbb{R}^m . En todo caso nos será más útil la norma 2 dada por $\|y\|_2 = \sqrt{\sum_{j=1}^m y_j^2}$.

2.9.1. Consistencia

El **Método de Euler explícito o progresivo (Forward Euler)** consiste en construir la secuencia de valores

$$\begin{aligned} \text{dado } y_0^h &\in \mathbb{R}^m, \\ y_{n+1}^h &= y_n^h + h f(t_n^h, y_n^h) \in \mathbb{R}^m, \quad n = 0, 1, \dots, N^h - 1, \end{aligned}$$

o con una notación más ligera

$$\begin{aligned} \text{dado } y_0 &\in \mathbb{R}^m, \\ y_{n+1} &= y_n + h f(t_n, y_n) \in \mathbb{R}^m, \quad n = 0, 1, \dots, N - 1. \end{aligned} \quad (2.11)$$

Vamos a suponer que el error inicial $\|y(t_0) - y_0^h\|$ tiende a cero cuando $h \rightarrow 0^+$.

El esquema en diferencias finitas que aproxima a la ecuación diferencial se escribe como

$$\mathcal{D}(y(t); h) = \frac{1}{h}(y(t+h) - y(t)) - f(t, y(t))$$

y entonces el error de truncatura es

$$\mathcal{T}(y(t); h) = \frac{h}{2}y''(\bar{\xi}_h)$$

donde $y''(\bar{\xi}_h) = (y_1''(\bar{\xi}_h^1), y_2''(\bar{\xi}_h^2), \dots, y_m''(\bar{\xi}_h^m))$ denota los distintos restos de los desarrollos de Taylor para cada una de las componentes y_j . Los puntos intermedios son distintos en general de una componente a otra.

La consistencia del esquema numérico, por decirlo de una forma coloquial, es lo que le falta a la solución verdadera para cumplir el esquema numérico. Se obtiene insertando la solución continua dentro del esquema numérico. Mide el grado en el que la solución exacta cumple con el esquema numérico usado en un punto determinado t con paso h , o también, es el error que se comete cuando se aplica el método en un paso de longitud h empezando con el valor exacto $y(t)$.

Definición 72 Para cualquier $t \in [t_0, t_0 + T]$ fijo y $h > 0$, se llama **error local, o de consistencia local, en $t + h$ sobre la función $w(t)$ solución del problema de Cauchy a la cantidad**

$$l(w(t); h) = w(t+h) - w(t) - h f(t, w(t)) = h \mathcal{T}(w(t); h). \quad (2.12)$$

En el caso $m = 1$ ya hemos visto, usando el desarrollo de Taylor, que tenemos

$$y(t+h) = y(t) + hy'(t) + \frac{1}{2}h^2y''(\xi), \quad t < \xi < t+h$$

de donde usando la ecuación diferencial

$$y(t+h) = y(t) + hf(t, y(t)) + \frac{1}{2}h^2y''(\xi).$$

En el caso $m > 1$ debemos de trabajar cada componente por separado. Si $y = (y_1, y_2, \dots, y_m)$ para $m > 1$ tenemos para cada $j = 1, 2, \dots, m$ la existencia de valores $t < \xi_j < t+h$ tales que

$$y_j(t+h) = y_j(t) + h(y_j)'(t) + \frac{1}{2}h^2(y_j)''(\xi_j),$$

de donde

$$y_j(t+h) = y_j(t) + hf_j(t, y_1(t), y_2(t), \dots, y_m(t)) + \frac{1}{2}h^2(y_j)''(\xi_j), \quad j = 1, 2, 3, \dots, m$$

y ahora podemos trabajar acotando el error en cada componente. Vamos a usar la notación compacta

$$y(t+h) = y(t) + h f(t, y(t)) + \frac{1}{2} h^2 \bar{y}''(\xi)$$

donde se debe de entender que $\bar{y}''(\xi)$ es el vector de componentes $(y_j)''(\xi_j)$, esto es,

$$\bar{y}''(\xi) = ((y_1)''(\xi_1), \dots, (y_j)''(\xi_j), \dots, (y_m)''(\xi_m))$$

2.9.2. Estabilidad

La consistencia no es suficiente para garantizar la convergencia, necesitamos también el concepto de estabilidad. En el método de Euler la estabilidad está garantizada por construcción puesto que, cuando menos, siempre disponemos de la constante de Lipschitz para obtener una desigualdad de estabilidad que relacione una desviación del cálculo debido a una variación en los datos.

Podemos repetir exactamente el proceso en el caso de una ecuación ($m = 1$). Supongamos que tenemos dos realizaciones del método de Euler debidas a puntos iniciales distintos, esto es:

dado y_0 ,

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

y si variamos el comienzo,

dado $z_0 \neq y_0$,

$$z_{n+1} = z_n + h f(t_n, z_n), \quad n = 0, 1, \dots, N-1.$$

La condición de Lipschitz sobre f garantiza la **estabilidad** del esquema numérico puesto que:

$$y_{n+1} - z_{n+1} = y_n - z_n + h [f(t_n, y_n) - f(t_n, z_n)], \quad n = 0, 1, \dots, N-1.$$

Entonces, para $n = 0, 1, \dots, N-1$ y una norma en \mathbb{R}^m cualquiera

$$\|y_{n+1} - z_{n+1}\| \leq (1 + h L_f) \|y_n - z_n\|.$$

Tenemos entonces

$$\begin{aligned} \|y_1 - z_1\| &\leq (1 + h L_f) \|y_0 - z_0\| \\ \|y_2 - z_2\| &\leq (1 + h L_f) \|y_1 - z_1\| \\ &\vdots \\ \|y_N - z_N\| &\leq (1 + h L_f) \|y_{N-1} - z_{N-1}\| \end{aligned}$$

de donde podemos aplicar una recurrencia y obtener

$$\begin{aligned}\|y_{n+1} - z_{n+1}\| &\leq (1 + h L_f) \|y_n - z_n\| \\ &\leq (1 + h L_f)^2 \|y_{n-1} - z_{n-1}\| \leq \dots \leq (1 + h L_f)^{n+1} \|y_0 - z_0\|\end{aligned}$$

y usando $(1 + x) \leq e^x$ junto con el hecho de que $(n + 1)h \leq T$ obtenemos para $n = 0, 1, \dots, N - 1$

$$\|y_{n+1} - z_{n+1}\| \leq e^{T L_f} \|y_0 - z_0\|.$$

Lo hemos usado pero no le hemos prestado la suficiente atención, de hecho aquí todo depende de h y hemos obtenido para $n = 1, \dots, N^{(h)}$

$$\|y_n^{(h)} - z_n^{(h)}\| \leq e^{T L_f} \|y_0^{(h)} - z_0^{(h)}\|.$$

De donde si $h \rightarrow 0^+$ y $\|y_0^{(h)} - z_0^{(h)}\| \rightarrow 0$ resulta que

$$\max_{0 \leq n \leq N^{(h)}} \{\|y_n^{(h)} - z_n^{(h)}\|\} \rightarrow 0.$$

Esto es la estabilidad del esquema numérico. Como estamos hablando del comportamiento cuando $h \rightarrow 0^+$, vamos a concretar mejor la definición.

Definición 73 Un esquema se dice **0-estable** (leer como cero-estable) si existe una constante $C > 0$ y un valor $h_* > 0$ tal que para todo $h \in (0, h_*)$ y cualquier $n = 0, 1, 2, \dots, N^{(h)}$, $(N^{(h)})h = T$)

$$|y_n^{(h)} - z_n^{(h)}| \leq C |y_0^{(h)} - z_0^{(h)}|.$$

Por lo tanto, si

$$\lim_{h \rightarrow 0^+} |y_0 - z_0| = |y_0^h - z_0^h| \rightarrow 0 \Rightarrow \lim_{h \rightarrow 0^+} \max_n |y_n^h - z_n^h| \rightarrow 0.$$

Hemos visto que el **método de Euler explícito es 0-estable** tanto en el caso vectorial como escalar. El nombre 0-estable viene de que se permite $h \rightarrow 0^+$.

Observación 68 Volvemos a encontrarnos con cotas de estabilidad de tipo exponencial que no tienen porque ser las más ajustadas. Por otro lado parece que hacer $h \rightarrow 0^+$ puede anular el conjunto de la acotación superior con cualquier constante por muy grande que esta sea, pero esto es teoría y en la práctica se tienen que computar valores y_n con un $h > 0$ fijo que no puede ser tan pequeño como uno quiera. Esto ocasiona graves dificultades como ya sabemos. Como consecuencia, se introducen también otros conceptos distintos para la estabilidad con respecto a h que se preocupan más de cuando h está fijo.

2.9.3. Convergencia

En esta sección repetimos formalmente el caso $m = 1$. La única diferencia que usamos una norma $\|\cdot\|$ en vez del valor absoluto. Suponemos $f \in C^1$ y que posee una constante de Lipschitz global L_f , o lo que es lo mismo, una acotación global para la matriz de primeras derivadas de f con respecto a las variables y_1, \dots, y_m , es decir, su Jacobiano $J_y f$ con respecto a la variable vectorial \vec{y} . Tenemos entonces

1. solución única al problema de Cauchy con regularidad C^2 ,
2. el método de Euler explícito es **consistente** y
3. el método de Euler explícito es **0-estable**.

Veamos que esto nos basta para probar la convergencia. Tenemos la ecuación

$$y_{n+1} = y_n + h f(t_n, y_n) \in \mathbb{R}^m, \quad n = 0, 1, \dots, N - 1.$$

y para $t_\star = t_n$ existen $\xi_n^j \in (t_n, t_n + h)$ para $j = 1, 2, \dots, m$ tal que

$$y(t_\star + h) = y(t_\star) + h f(t_\star, y(t_\star)) + \frac{1}{2} h^2 \bar{y}''(\xi_n).$$

Entonces

$$e_{n+1} = e_n + h[f(t_n, y(t_n)) - f(t_n, y_n)] + \frac{1}{2} h^2 \bar{y}''(\xi_n). \quad (2.13)$$

de donde usando la condición de Lipschitz o el Teorema del Valor Medio

$$\|e_{n+1}\| \leq (1 + hL_f) \|e_n\| + h \mathcal{T}(y(t_n); h), \quad n = 0, 1, \dots, N - 1.$$

A partir de aquí todo es como en el caso $m = 1$: Por inducción y usando que $(1 + x) \leq e^x$ llegamos a

$$\begin{aligned} \|e_{n+1}\| &\leq (1 + hL_f)^{n+1} \|e_0\| + h \mathcal{T}(y; h) \frac{(1 + hL_f)^{n+1} - 1}{(1 + hL_f) - 1} \\ &\leq e^{(n+1)hL_f} \|e_0\| + \mathcal{T}(y; h) \frac{e^{(n+1)hL_f} - 1}{L}, \quad n = 0, 1, 2, \dots, N - 1 \end{aligned}$$

siendo siempre $(n + 1)h \leq T$. Si queremos uniformizar los coeficientes delante de $\|e_0\|$ y de $\mathcal{T}(y; h)$ podemos usar

$$\|e_{n+1}^h\| \leq e^{TL} \|e_0^h\| + \mathcal{T}(y; h) \frac{e^{TL} - 1}{L}, \quad n = 0, 1, \dots, N^h - 1.$$

por lo que si $\|e_0^h\| \rightarrow 0$ y $\mathcal{T}(y; h) \rightarrow 0$ cuando $h \rightarrow 0+$ garantizamos que $\|e_{n+1}^h\| \rightarrow 0$ cuando $h \rightarrow 0+$ y que el método es convergente. Otra vez

- gracias a la 0-estabilidad, o a que f es globalmente Lipschitz que es lo mismo en este caso, podemos controlar los errores de acumulación que se generan en cada paso debidos al error inicial $\|e_0^h\|$, aunque sea en términos de la constante e^{TL} .
- nos va a hacer falta tener consistencia para usar $\mathcal{T}(y; h) \rightarrow 0$
- y aunque sea obvio, necesitamos que el error inicial tienda a cero $\|e_0^{(h)}\| \rightarrow 0$

Observación 69 En este caso, sabemos cual es el valor de $\mathcal{T}(y; h)$ y podemos escribir

$$\|e_{n+1}\| \leq e^{TL} \|e_0\| + \|\tilde{y}''\|_\infty \frac{e^{TL} - 1}{2L} h, \quad n = 0, 1, \dots, N-1.$$

Hemos probado

Teorema 74 Convergencia del método de Euler explícito. Caso general $m \geq 1$ con f Lipschitz: Supongamos que $f \in C^1$ y cumple una condición de Lipschitz global. Entonces el método de Euler explícito es 0-estable, consistente y si el error inicial $e_0^{(h)}$ cumple $\|e_0^{(h)}\| \leq C_0 h$, ($h \rightarrow 0^+$) para alguna constante $C_0 > 0$ entonces existe una constante $C > 0$ tal que

$$\|e_n^{(h)}\| \leq Ch, \quad (h \rightarrow 0^+), \quad n = 1, 2, \dots, N^{(h)}$$

para $N^{(h)} = T/h$, esto es, $h \rightarrow 0^+$, $N^{(h)} \rightarrow +\infty$, $hN^{(h)} = T$. Como consecuencia, el **método de Euler es de primer orden** con respecto al parámetro $h > 0$.

Observación 70 Nos hemos apoyado en el hecho de que el método de Euler es 0-estable y además consistente

0-estabilidad + consistencia \Rightarrow convergencia

El recíproco es también cierto.

Observación 71 La constante C es proporcional a e^{TL} por lo que es extremadamente grande para valores moderados del producto TL .

2.9.4. Estudio del error (¿Si el tiempo lo permite?)*

En el caso de un sistema lineal y usando la norma 2 se puede extraer información sobre la convergencia en términos de los autovalores de la matriz del sistema.

Ya sabemos que la constante de Lipschitz es insensible al signo. En esta sección vamos a dar una estimación del error para el método de Euler explícito que tenga en cuenta este hecho y sea más ajustada a los resultados numéricos de la práctica cuando el sistema sea disipativo.

Debemos de modificar la noción de desaparición de modos rápidos y llegada a una curva transitoria, típica del caso escalar en los problemas disipativos, y ahora hablaremos de **contracción de las curvas entre sí**, esto es,

Definición 75 *Un sistema es disipativo cuando todas las curvas tienden a estar cada vez más cercanas. En caso contrario se llama sistema expansivo.*

Vamos a seguir la idea vista en el tema anterior para la estabilidad de ecuaciones y sistemas de acuerdo a su naturaleza disipativa o contractiva. Lo hemos hecho en el caso $m = 1$, pero **la idea usada para $m = 1$ no se extiende de manera trivial al caso $m > 1$ ya que el Teorema del Valor Medio se debe de aplicar a cada componente en el caso vectorial.** Tenemos que trabajar un poco más duro ahora. Por eso sólo lo vamos a ver en el caso de un sistema lineal. Esto ya introduce algunas ideas nuevas al respecto.

Supongamos que $y' = A y$ donde $A \in \mathbb{R}^{m \times m}$ es una matriz constante. Entonces $f(t, y) = A y$, esto es

$$\begin{aligned} f_1(t, y_1, y_2, \dots, y_m) &= a_{11}y_1 + a_{12}y_2 + \dots + a_{1m}y_m \\ f_2(t, y_1, y_2, \dots, y_m) &= a_{21}y_1 + a_{22}y_2 + \dots + a_{2m}y_m \\ &\vdots \\ f_m(t, y_1, y_2, \dots, y_m) &= a_{m1}y_1 + a_{m2}y_2 + \dots + a_{mm}y_m. \end{aligned}$$

El estudio de **sistemas lineales de coeficientes constantes** es inmediato cuando la matriz es diagonalizable. Supongamos que A es una matriz cuadrada $m \times m$ y tenemos el sistema lineal

$$\begin{cases} y'(t) = A y(t), & t > 0, \\ y(0) = y_0 \in \mathbb{R}^m \end{cases}$$

Aquí la solución para $t \geq 0$ es $y(t) = e^{At}y_0$ donde e^{At} es la matriz exponencial. Para simplificar vamos a suponer que A es **diagonalizable en $\mathbb{R}^{m \times m}$** (ocurre igual si aparecen autovalores complejos y tenemos que diagonalizar en $\mathbb{C}^{m \times m}$ pero intentamos simplificar la presentación) tendremos $\Lambda = V^{-1}AV$ con $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_m\}$ y estamos admitiendo que $V \in \mathbb{R}^{m \times m}$; usando $z = V^{-1}y$ llegamos al problema desacoplado

$$\begin{cases} z'(t) = \Lambda z(t), & t > 0 \Leftrightarrow z'_j(t) = \lambda_j z_j(t), \quad j = 1, 2, \dots, m \\ z(0) = V^{-1}y_0 = z_0 \end{cases}$$

y deshaciendo el cambio de variables

$$\vec{y}(t) = \sum_{j=1}^m c_j e^{\lambda_j t} \vec{v}_j \tag{2.14}$$

siendo los c_j tales que

$$\sum_{j=1}^d c_j \vec{v}_j = y_0 \Leftrightarrow V c = y_0.$$

Por lo tanto, la solución es básicamente una combinación lineal de las funciones $e^{\lambda_j t}$ y está determinada por los autovalores de la matriz A , o lo que es lo mismo, por su espectro.

Análisis del error para sistemas

Si denotamos por $A(j, :)$ la fila j de la matriz A y suponemos que y lo tomamos como vector columna, entonces también se puede describir cada entrada del sistema como

$$(y_j)'(t) = A(j, :) \cdot y(t) = \sum_{s=1}^m a_{j,s} y_s(t), \quad j = 1, 2, \dots, m.$$

Por otro lado, tenemos que si A^r es la potencia r de A entonces

$$y'(t) = A \cdot y(t) \Rightarrow y'(t) = A \cdot y' = A^2 y(t) \Rightarrow y^{(r)}(t) = A^r \cdot y(t)$$

de donde para cada componente y_j

$$(y_j)^{(r)}(t) = A^r(j, :) y(t)$$

siendo $A^r(j, :)$ la fila j de la matriz potencia A^r . Entonces el desarrollo de Taylor para una componente j cualquiera tiene la forma

$$y_j(t+h) = y_j(t) + hA(j, :) \cdot y(t) + \frac{h^2}{2} A^2(j, :) \cdot y(t) + \dots + \frac{h^r}{r!} A^r(j, :) \cdot y(\xi^j)$$

donde $\xi^j \in (t, t+h)$. En el caso de aplicar el método de Euler nos interesa el caso $r = 2$ y tenemos entonces

$$y_j(t+h) = y_j(t) + hA(j, :) \cdot y(t) + \frac{h^2}{2} A^2(j, :) \cdot y(\xi^j)$$

donde $\xi^j \in (t, t+h)$. Unificamos esto en notación vectorial y matricial como

$$y(t+h) = y(t) + hA \cdot y(t) + \frac{h^2}{2} A^2 : \bar{y}(\xi)$$

donde entendemos que $A^2 : \bar{y}(\xi)$ es un vector cuyas componentes vienen dadas por

$$(A^2 : \bar{y}(\xi))^j = A^2(j, :) \cdot \bar{y}(\xi).$$

Después de estas observaciones volvamos a la aplicación del método de Euler explícito. Se describe para $n = 0, 1, \dots, N^{(h)} - 1$ como

$$y_{n+1}^h = y_n^h + hA y_n^h = (I_m + hA) y_n^h$$

donde $I_m \in \mathbb{R}^{m \times m}$ es la matriz identidad. Como $y(t_{n+1}^h) = y(t_n^h + h)$, tenemos

$$y(t_n^h + h) = y(t_n^h) + hA \cdot y(t_n^h) + \frac{h^2}{2} A^2 : \bar{y}(\xi_n^h)$$

luego

$$y(t_{n+1}^h) = (I_m + hA)y(t_n^h) + \frac{h^2}{2} A^2 : \bar{y}(\xi_n^h)$$

de donde el error es (quitamos h para no cargar notación)

$$y(t_{n+1}) - y_{n+1} = (I_m + hA)(y(t_n) - y_n) + \frac{h^2}{2} A^2 : \bar{y}(\xi_n)$$

y entonces

$$\|y(t_{n+1}) - y_{n+1}^h\| \leq \|I_m + hA\| \|y(t_n^h) - y_n^h\| + \frac{h^2}{2} \|A^2\| \|\bar{y}(\xi_n)\|.$$

Sistema lineal simétrico y disipativo

El caso más interesante y frecuente en las aplicaciones es cuando A es **simétrica con autovalores reales negativos**, extensión vectorial natural de la situación escalar $f(t, y) = -\lambda y$ con $\lambda > 0$. Por la expresión (2.14) sabemos que

$$\|y(t)\| \rightarrow 0, \quad t \rightarrow +\infty.$$

Como A es diagonalizable vía una matriz ortogonal, $A = V^T \Lambda V$, y se tiene $\|A\|_2 = \|\Lambda\|_2$ siendo $\Lambda = \text{diag}(-\lambda_1, -\lambda_2, \dots, -\lambda_m)$ la matriz con los autovalores de A en la diagonal. Por lo tanto, la norma $\|\cdot\|_2$ es la adecuada para tratar con esta situación.

Tenemos que

$$\|y(t_{n+1}) - y_{n+1}\|_2 \leq \|I_m + hA\|_2 \|y(t_n) - y_n\|_2 + \frac{h^2}{2} \|A^2\|_2 \|y(\xi_n)\|_2$$

de donde usando que V es unitaria podemos poner

$$\|I_m + hA\|_2 = \|VV^T + hVAV^T\|_2 = \|I_m + h\Lambda\|_2$$

y conseguir

$$\|y(t_{n+1}) - y_{n+1}\|_2 \leq \|I_m + h\Lambda\|_2 \|y(t_n) - y_n\|_2 + \frac{h^2}{2} \|\Lambda^2\|_2 \|y(\xi_n)\|_2.$$

Como $\|\Lambda\|_2 = \max_j |\lambda_j|$ cuando $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$, tenemos fácilmente que

$$\|y(t_{n+1}) - y_{n+1}^h\|_2 \leq \max_j |1 - h\lambda_j| \|y(t_n^h) - y_n^h\|_2 + \frac{h^2}{2} \max_j \lambda_j^2 \|y(\xi_n^h)\|_2.$$

Por lo tanto, si tenemos

$$-\lambda_1 < -\lambda_2 < \dots < -\lambda_m < 0.$$

Usando $h < 2/\lambda_1$, entonces $-1 < 1 - h\lambda_1$ de donde

$$-1 < 1 - h\lambda_1 < 1 - h\lambda_2 < \dots < 1 - h\lambda_m < 1$$

y además, si $h < 1/\lambda_1$ entonces

$$0 < 1 - h\lambda_1 < 1 - h\lambda_2 < \dots < 1 - h\lambda_m < 1$$

y se tiene

$$0 < \max_j |1 - h\lambda_j| = 1 - h\lambda_m < 1.$$

Por lo tanto, si $h < 1/\lambda_1$

$$\|y(t_{n+1}) - y_{n+1}\|_2 \leq (1 - h\lambda_m) \|y(t_n) - y_n\|_2 + \frac{h^2}{2} \lambda_1^2 \max_t \|y(t)\|_2.$$

Vemos el papel fundamental de los autovalores de una matriz en el control del esquema numérico. Iterando cuando $h < 1/\lambda_1$ tenemos

$$\|y(t_{n+1}) - y_{n+1}\|_2 \leq (1 - h\lambda_m) \|y(t_n) - y_n\|_2 + \frac{h^2}{2} \lambda_1^2 \max_t \|y(t)\|_2$$

de donde para $1 - h\lambda_m \in (0, 1)$

$$\|y(t_{n+1}) - y_{n+1}\|_2 \leq (1 - h\lambda_m)^{n+1} \|y(t_0) - y_0\|_2 + \frac{1 - (1 - h\lambda_m)^{n+1}}{1 - (1 - h\lambda_m)} \frac{h^2}{2} \lambda_1^2 \max_t \|y(t)\|_2$$

es decir

$$\|y(t_{n+1}) - y_{n+1}\|_2 \leq (1 - h\lambda_m)^{n+1} \|y(t_0) - y_0\|_2 + \frac{1 - (1 + h\lambda_m)^{n+1}}{h\lambda_m} \frac{h^2}{2} \lambda_1^2 \max_t \|y(t)\|_2$$

o bien

$$\|y(t_{n+1}) - y_{n+1}\|_2 \leq (1 - h\lambda_m)^{n+1} \|y(t_0) - y_0^h\|_2 + \frac{\lambda_1}{\lambda_m} \lambda_1 \frac{\max_t \|y(t)\|_2}{2} h.$$

se puede observar como la **razón de rigidez** de la matriz A

$$\frac{\lambda_1}{\lambda_m}$$

también aparece en la estimación de convergencia. Sabemos que si

$$\frac{\lambda_1}{\lambda_m} \gg 1$$

tenemos modos en la solución con velocidades muy distintas. Hemos demostrado el siguiente resultado de convergencia más detallado:

Teorema 76 Convergencia del método de Euler explícito para $y' = Ay$ con A simétrica definida negativa: Supongamos que los autovalores reales de A son $-\lambda_1 < -\lambda_2 < \dots < -\lambda_m < 0$ y pongamos

$$R_A = \frac{\lambda_1}{\lambda_m} > 1$$

la razón de rigidez de la matriz A . Entonces para $h \in (0, 1/\lambda_1)$ se tiene la estimación de error

$$\|e_{n+1}\|_2 \leq (1 - h\lambda_m)^n \|e_0\|_2 + \lambda_1 R_A \frac{\max_t \|y(t)\|_2}{2} h.$$

para $n = 0, 1, 2, \dots, N^h - 1$. Por lo tanto, otra vez tenemos convergencia incluso si el error inicial no tiende a cero. Además, vemos explícitamente la restricción sobre h y la influencia de la rigidez del sistema R_A en la convergencia del esquema.

Si el error inicial es $O(h)$ el esquema es $O(h)$ y además las constantes decaen ayudando a la convergencia cuando se alcanza el régimen $h < h_\star$ para h_\star dependiendo de los autovalores de A .

También se ve que el esquema numérico reproduce el comportamiento continuo ya que si $h < 2/\lambda_1$

$$\|y_{n+1}\|_2 \leq \max_j \{|1 - h\lambda_j|\} \|y_n\|_2 \leq (\max_j \{|1 - h\lambda_j|\})^n \|y_0\|_2 \rightarrow 0.$$

puesto que $0 < \max_j \{|1 - h\lambda_j|\} < 1$.

Observación 72 Como normalmente no se conoce λ_1 puede estimar un valor para λ_1 usando el método de Euler y viendo el h más grande que garantiza el decaimiento en esta norma $\|\cdot\|_2$.

El estudio del caso vectorial $m > 1$ para un problema no-lineal necesita herramientas teóricas más potentes.

Observación 73 En el caso en donde A no sea diagonalizable, tenemos entonces que admitir la existencia de autovalores repetidos y complejos y sabemos que la solución es una combinación lineal de funciones de la forma

$$t^r e^{ta} \cos(bt), \quad t^r e^{ta} \sin(bt), \quad (a = \operatorname{Re}(\lambda_j), b = \operatorname{Im}(\lambda_j)),$$

donde r es la multiplicidad de $a + ib$. El comportamiento de la solución se plantea en términos de las partes reales de los autovalores λ_j . En general tenemos para las soluciones

- convergencia a cero si $\max_j \{\operatorname{Re}(\lambda_j)\} < 0$

- acotación si $\max_j \{Re(\lambda_j)\} \leq 0$ y aquellos autovalores con $Re(\lambda_j) = 0$ son simples.

Observación 74 Cuando A es simétrica positiva definida, SPD es la abreviatura más común y en inglés, entonces A es diagonalizable $\Lambda = V^{-1}AV$ con $\Lambda = diag\{\lambda_1, \dots, \lambda_m\}$ con $V \in \mathbb{R}^{m \times m}$ y todos los autovalores son reales positivos. Por lo tanto, si $-A$ es simétrica positiva definida, todos los autovalores de A son negativos y tenemos un sistema disipativo $y' = Ay$.

Esto es interesante puesto que muchas veces se suele escribir el problema en la forma

$$y' + By = 0$$

y si B es SPD el sistema es disipativo (aquí B juega el papel de $-A$).

Ejemplo 77 Para $t > 0$ consideramos el sistema lineal de edos

$$\begin{cases} x'_1(t) &= -298x_1(t) + 99x_2(t), \\ x'_2(t) &= -594x_1(t) + 197x_2(t). \end{cases}$$

con dato inicial $x_1(0) = -1/2$ y $x_2(0) = 1/2$ y sobre el intervalo temporal $[0, 3]$. La matriz tiene autovalores -1 y -100 y se ven en la Figura 2.16 los cálculos usando Euler explícito y Euler implícito (se verá en el tema siguiente). El valor óptimo para Euler explícito es $h < 0.02$ o bien $N > 150$ para poder obtener convergencia pero con oscilaciones. Si usamos la restricción $h < 0.01$, o bien $N > 300$, lo que indica un autovalor máximo negativo en torno a -100 .

Ejemplo 78 Euler progresivo en $x'(t) = Dx(t)$, $(t > 0)$, $x(0) = (1, 1, 1)'$ donde

$$D = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -100 \end{pmatrix}$$

genera los resultados que se pueden ver en la Figura 2.17. Para cada experimento numérico se presenta el cálculo global en $[0, 10]$ y una visualización local de los resultados en $[0, 4]$ o en $[0, 2]$ para observar su comportamiento inicial. El valor de h óptimo (convergencia sin oscilaciones) para Euler explícito es $h < 0.01$, o bien $N > 1000$, lo que indica un autovalor máximo negativo en torno a -100 .

Normalmente, las matrices no se presentan en formato diagonal. Si modificamos aleatoriamente la matriz D usando una transformación que preserve los autovalores para así esconderlos, esto es ponemos $A = X^{-1}DX$ para X una matriz 3×3 no singular, nos encontramos con resultados similares. Tomemos

$$A = \begin{pmatrix} 3771.9348 & 1783.3507 & 5811.5749 \\ -753.93982 & -359.78508 & -1160.3972 \\ -2280.9485 & -1076.2151 & -3515.1497 \end{pmatrix}.$$

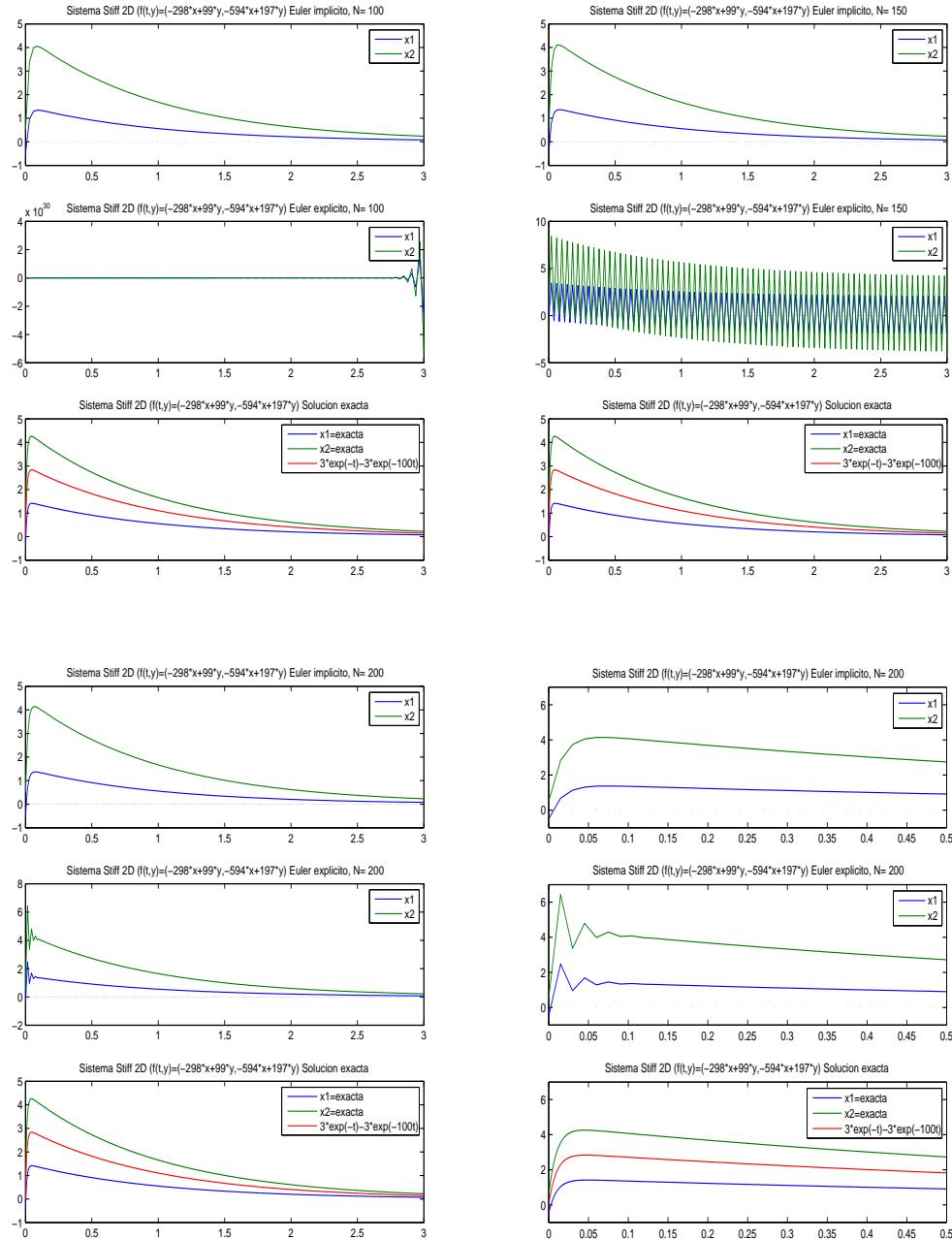


Figura 2.16: Cálculos sobre el intervalo temporal $[0, 3]$ y visualizaciones cercanas al origen de los resultados. Valor óptimo para Euler explícito es $h < 0.02$ o bien $N > 150$ para poder obtener convergencia pero con oscilaciones. Si usamos la restricción $h < 0.01$ entonces necesitamos $N > 300$ y obtenemos convergencia sin oscilaciones.

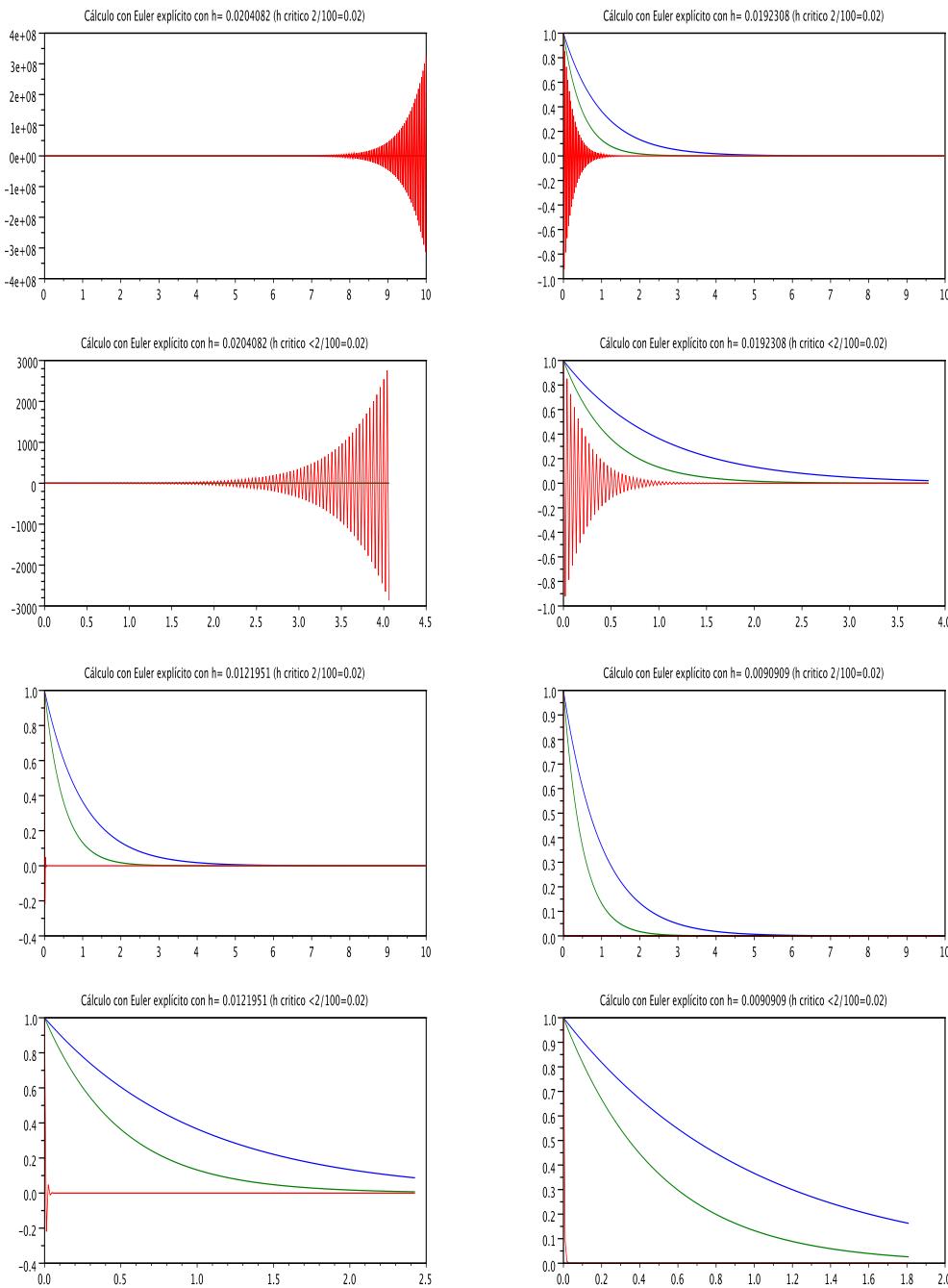


Figura 2.17: Euler explícito para sistema lineal diagonal 3×3 con autovalores $-100, -2, -1$. El valor de h óptimo (convergencia sin oscilaciones) para Euler explícito es $h < 0.01$ ($N > 1000$ al ser $T = 10$), lo que indica un autovalor máximo en torno a -100 . Se presentan los cálculos globales en $[0, 10]$ y una visualización local en $[0, 4]$ o en $[0, 2]$ para ver el comportamiento inicial.

cuyo espectro es $-99.997625, -1.0088944, -1.9934603$. Los resultados se pueden ver en la Figura 2.18. También aquí para cada experimento numérico se presenta el cálculo global en $[0, 10]$ y una visualización local de los resultados en $[0, 4]$ o en $[0, 2]$ para observar su comportamiento inicial. El valor de h óptimo (convergencia sin oscilaciones) para Euler explícito es $h < 0.01$ ($N > 1000$ al ser $T = 10$).

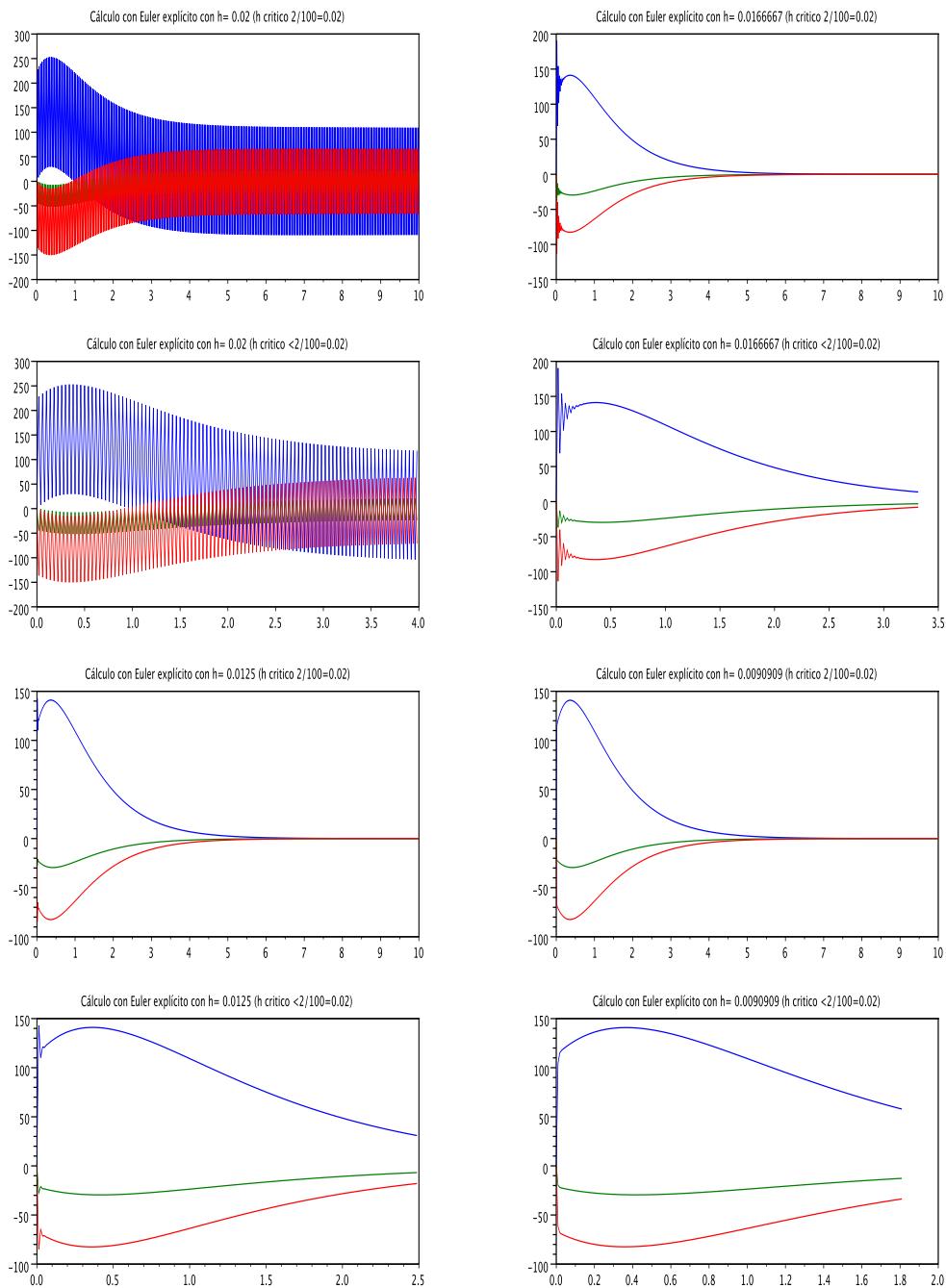


Figura 2.18: Euler explícito para sistema lineal 3×3 con autovalores $-99.997625, -1.0088944, -1.9934603$. El valor de h óptimo (convergencia sin oscilaciones) para Euler explícito es $h < 0.01$ o bien $N > 1000$. Se presentan los cálculos globales en $[0, 10]$ y una visualización local en $[0, 4]$ o en $[0, 2]$ para ver el comportamiento inicial.

2.10. Ejercicios

1. Pondremos $f(x) = O(1)$, ($x \rightarrow x_0$) si $|f(x)| \leq k$ para alguna constante $k > 0$ en un entorno de x_0 . Comprobar los siguientes órdenes de magnitud

a) $x^2 + x = O(x^2)$, ($x \rightarrow \pm\infty$) y $x^2 + x = O(x)$, ($x \rightarrow 0$)

b) $\frac{1}{3+2x^2} = O(1)$ ($x \rightarrow 0$)

c) $\frac{1}{3+2x^2} = O(x^{-2})$ ($x \rightarrow \pm\infty$)

d) Si $f(x) = x \sin(1/x)$ entonces $f(x) = O(x)$, ($x \rightarrow 0$) pero $f(x) \neq o(x)$, ($x \rightarrow 0$).

e) $\log(x) \rightarrow +\infty$ cuando $x \rightarrow +\infty$ pero lo hace más despacio que cualquier potencia positiva de x , esto es,

$$a > 0 \Rightarrow \log(x) = o(x^a), (x \rightarrow +\infty)$$

f) $e^x \rightarrow +\infty$ cuando $x \rightarrow +\infty$ pero lo hace más rápido que cualquier potencia positiva de x , esto es,

$$a > 0 \Rightarrow x^a = o(e^x), (x \rightarrow +\infty)$$

g) $\frac{1}{4-3x-2e^{-x}} = O(1/x)$ ($x \rightarrow +\infty$)

h) $\frac{1}{4-3x-2e^{-x}} = O(e^x)$ ($x \rightarrow -\infty$)

Fijamos la asintótica $x \rightarrow 0$ para simplificar

a) $\sin(x) = O(x)$, $\sin(x^2) = O(x^2)$, $\sin(7x) = O(x)$

b) $\sin(3x) - 3x = O(x^3)$, $1 - \cos(x) = O(x^2)$, $\cos(x) = O(1)$

c) $\sinh(x) = O(x)$, $\cosh(x) = O(1)$

d) $\sin(x) = o(1)$, $\sin(x^2) = o(x)$, $\sin(7x) = o(1)$

e) $1 - \cos(3x) = o(x)$, $\cos(x) = o(x^{-1/2})$

f) $e^{-1/x} = o(x^n)$, $\forall n$

g) $\frac{1}{3+2x^2} \sim \frac{1}{2x^2}$ ($x \rightarrow \pm\infty$)

h) $\frac{5}{1-x-e^{-x}} \sim \frac{-10}{x^2}$ ($x \rightarrow 0$)

i) $\sin(3x) \sim 3x$ ($x \rightarrow 0$)

$$j) \sin(3x) - 3x \sim -\frac{(3x)^2}{3!} \quad (x \rightarrow 0)$$

2. Demostrar que

$$\frac{\log(1+x)}{x} = 1 - \frac{x}{2} + O(x^2), \quad x \rightarrow 0.$$

3. Para $h \rightarrow 0$ y $n \rightarrow +\infty$ tal que $hn = t$ fijo se cumple

- $(1 - hL)^n \sim e^{-tL}$, es decir, $\lim_{hn=t} (1 - hL)^n = e^{-tL}$,
- $(1 + hL)^n \sim e^{tL}$, es decir, $\lim_{hn=t} (1 + hL)^n = e^{tL}$.
- Si $p > 0$ entonces $(1 + O(h^{p+1}))^n = 1 + O(h^p)$.

donde $\lim_{hn=t}$ representa $h \rightarrow 0$ y $n \rightarrow +\infty$ tal que $hn = t$ fijo.

4. Si para $A, B > 0$ constantes independientes de n los números $\xi_n \geq 0$ cumplen

$$\xi_{n+1} \leq A \xi_n + B, \quad n = 0, 1, 2, \dots, N-1,$$

entonces comprobar que

$$\xi_n \leq A^n \xi_0 + E_n(A) B, \quad n = 1, 2, \dots, N,$$

donde

$$E_n(A) = \frac{A^n - 1}{A - 1}, \quad \text{si } A \neq 1, \quad E_n(A) = n, \quad \text{si } A = 1.$$

Cuando $A = 1 + \delta$ con $\delta > 0$ comprobar que

$$\xi_n \leq A^n \xi_0 + \frac{e^{n\delta} - 1}{\delta} B, \quad n = 1, 2, \dots, N.$$

5. Sabemos que

$$\lim_{n \rightarrow +\infty} \left(1 + \frac{x}{n}\right)^n = e^x.$$

Si $K > 0$ y $|\theta_n| \leq 1$ para $n = 1, 2, \dots$ probar que

$$\lim_{n \rightarrow +\infty} \left(1 + \frac{x}{n} + \theta_n \frac{K}{n^2}\right)^n = e^x.$$

6. Aplicar Euler explícito al problema

$$w'(t) = t, \quad w(0) = 0$$

con solución $w(t) = t^2/2$ tomando como valor de inicio $y_0 = 0$ y con $t_n = nh$. Comprobar que el error global es $O(h)$ pero no puede ser más pequeño, esto es, no puede ser $O(h^2)$.

7. Probar que el método de Euler aplicado a

$$w'(t) = 1 + t^2, \quad w(0) = 0$$

con $y_0 = 0$ y con $t_n = nh$ es de orden 1 pero no 2.

8. Sea $h = 1/N$ con $N \geq 1$ un entero y sea $t_n = nh$ para $n = 0, 1, 2, \dots, N$. Aproximamos $y(t_n)$ por los valores y_n obtenidos con Euler explícito para aproximar la solución del problema

$$y'(t) = -\frac{y(t)}{1+t}, \quad 0 \leq t \leq 1.$$

- a) Suponiendo que $y(0) = y_0 = 1$ deducir expresiones explícitas, libres de sumas o productos, para los valores y_n .
- b) Deducir a partir de estas expresiones, y de su límite, la forma de la solución exacta.
- c) Verificar que $y(nh) - y_n = O(h)$ para cualquier $n \leq M$.

9. Aplicar el método de Euler a la ecuación $y' = ky$ con $y(0) = 1$ con paso constante h y obtener la aproximación clásica a la solución $y(t) = e^{kt}$. Dar una estimación del error y observar el comportamiento del mismo de acuerdo al signo de k .

- 10. a) Hallar una fórmula para y_n cuando se aplica el método de Euler a una ecuación de la forma $y'(t) = \alpha y(t) + \beta$, con y_0 dado.
- b) Calcular la solución verdadera y comprobar que $|y(t) - y_n^h| \rightarrow 0$ cuando $h \rightarrow 0$, $n \rightarrow \infty$ de forma tal que $nh = t$.
- c) Demostrar que si $\alpha \neq 0$ se cumple $|y(t) - y_n^h| = O(h)$ pero no existe $p > 1$ tal que $|y(t) - y_n^h| = O(h^p)$.

- 11. Comprobar que cuando el método de Euler se aplica al problema $y' = 5y$, $y(0) = 1$ en el punto $t = 1$ el valor de la aproximación con paso h es

$$y(1) \approx (1 + 5h)^{1/h}.$$

- 12. Comprobar que cuando el método de Euler se aplica al problema

$$y'(x)\sqrt{1-x} = 1, \quad y(0) = 0$$

tiene orden estrictamente $1/2$ para $x \in [0, 1]$, es decir, $|y(1) - y_N| = O(h^{1/2})$ pero $|y(1) - y_N| \neq O(h^p)$ para $p > 1/2$. Explicar la razón de este comportamiento.

13. Comprobar que cuando el método de Euler se aplica al problema

$$y'(x) = x(1-x), \quad y(0) = 0$$

$x \in [0, 1]$ cumple $|y(1) - y_N| = O(h^2)$ y sin embargo no es de orden 2.

14. Demostrar que el método de Euler explícito falla al aproximar la solución $y(x) = (4x/5)^{5/4}$ del problema $y' = y^{1/5}$ con $y(0) = 0$. Justifica la respuesta. ¿Qué ocurre cuando se aplica el método de Euler implícito?
15. Sea $f \in C^0([a, b] \times \mathbb{R})$ globalmente lipschitziana respecto a su segunda variable. Dado el Problema de Cauchy

$$\begin{cases} y'(t) &= f(t, y) \text{ en } [t_0, t_0 + T], \\ y(t_0) &= \alpha, \end{cases}$$

consideramos su resolución mediante la siguiente variante del método de Euler

$$(M) \quad \begin{cases} y_{n+1} &= y_n + h f(t_n + \theta h, y_n + \theta h f(t_n, y_n)) \\ y_0 &= \alpha \end{cases}$$

donde $t_n = t_0 + nh$, $h = T/N$ y $\theta \in [0, 1]$ es un parámetro a elegir.

Se pide:

- a) Interpretar geométricamente lo que se está haciendo.
 - b) Probar que (M) es estable cualquiera que sea θ .
 - c) Determinar el orden del error local de consistencia en función de θ .
 - d) Obtener estimaciones del error global de discretización para la mejor elección posible de θ . Para esta elección el esquema se denomina **método de Euler Modificado**.
16. Dado el problema de segundo orden

$$\begin{cases} x''(t) &= f(t, x(t), x'(t)), \quad t > 0 \\ x(0) &= x_0, \\ x'(0) &= v_0 \end{cases}$$

poniendo $y_n \approx x(t_n)$ y $z_n \approx x'(t_n)$, consideramos su resolución mediante el siguiente esquema

$$(M) \quad \begin{cases} y_{n+1} &= y_n + h z_n + \frac{h^2}{2} f_n, \\ z_{n+1} &= z_n + h f_n, \\ y_0 &= x_0, \\ z_0 &= v_0. \end{cases}$$

donde $t_n = nh$ y $f_n = f(t_n, y_n, z_n)$.

- a) Interpretar el esquema.
- b) Suponiendo que f es globalmente Lipschitz en la segunda y tercera variable con constante L , esto es

$$|f(t, x_1, y_1) - f(t, x_2, y_2)| \leq L(|x_1 - x_2| + |y_1 - y_2|),$$

demonstrar que el error

$$\theta_n = |z_n - x'(t_n)| + |y_n - x(t_n)|$$

cumple la relación

$$\theta_{n+1} \leq (1 + hL + h^2L/2 + h)\theta_n + h^2(1/2 + h/6)\|x'''\|_\infty.$$

Obtener la estimación del error global de aquí.

17. Consideremos el siguiente problema de Cauchy para un sistema diferencial de dos ecuaciones:

$$\begin{cases} \dot{y}(t) = f(z(t)), & t \in [0, T]; \\ \dot{z}(t) = g(y(t)), & t \in [0, T]; \\ y(0) = a, z(0) = b, \end{cases}$$

donde f y g son funciones reales globalmente lipschitzianas con constantes de Lipschitz L_f y L_g , respectivamente. Dada una partición $0 = t_0 < t_1 < \dots < t_N = T$ (que supondremos uniforme con paso $h = T/N$) de $[0, T]$, ponemos

$$\begin{cases} y_{n+1} = y_n + h f(z_n), \\ z_{n+1} = z_n + h g(y_n) \\ y_0 = a, z_0 = b. \end{cases}$$

Se pide:

- a) Definir el error local de consistencia para este método y determinarlo en términos de h .
- b) Consideremos el esquema perturbado

$$\begin{cases} Y_{n+1} = Y_n + h f(Z_n) + \alpha_n, \\ Z_{n+1} = Z_n + h g(Y_n) + \beta_n, \end{cases}$$

donde los α_n y β_n se suponen conocidos. Obtener la estimación de estabilidad

$$\theta_n \leq e^{Lt_n} \left(\theta_0 + \sum_{k=0}^n \sigma_k \right),$$

con $\theta_n = \max\{|Y_n - y_n|, |Z_n - z_n|\}$, $L = \max\{L_f, L_g\}$, $\sigma_k = \max\{\alpha_k, \beta_k\}$. ¿Cómo se puede utilizar esta estimación para obtener estimaciones de error?

c) Deducir la estimación del error de discretización global.

18. Comprobar que si

$$y'(t) = \sin(e^{y(t)}), \quad t \in [0, 1], \quad y(0) = 0.$$

entonces $|y''(t)| \leq e$.

19. Aplicar la siguiente variante del método de Euler

$$\begin{cases} y_{n+1} &= y_n + h f\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}h f(t_n, y_n)\right) \\ y_0 &= \alpha \end{cases}$$

donde $t_n = t_0 + nh$, $h = T/N$ al problema

$$y' = 4y, \quad y(0) = 1/3$$

y obtener el valor de la aproximación en el punto $t = 1/2$.

20. Sea $f \in C^0([a, b] \times \mathbb{R})$ globalmente lipschitziana respecto a su segunda variable, con constante de Lipschitz L_f y consideramos el siguiente método de un paso

$$(M) \quad \begin{cases} y_{n+1} &= y_n + h f(t_n + \theta h, (1 - \theta)y_n + \theta y_{n+1}), \\ y_0 &= \alpha \end{cases}$$

donde $t_n = a + nh$, $h = (b - a)/N$ y $\theta \in [0, 1]$.

Se pide:

- a) Interpretar geométricamente lo que se está haciendo.
 - b) Probar que si $h < \frac{1}{\theta L}$, entonces (M) está bien definido
 - c) Determinar el orden del error local de (M) en función de θ .
 - d) Obtener la estimación de estabilidad correspondiente.
 - e) Obtener la estimación de error global correspondiente.
21. Sea $f : \mathbb{R} \mapsto \mathbb{R}$ una función globalmente lipschitziana, y consideramos el problema de Cauchy

$$(PC) \quad \begin{cases} y'(t) &= f(y), \\ y(0) &= y_0; \end{cases}$$

que pretendemos resolver mediante el método

$$(M1P) \quad y_{n+1} = y_n + \alpha h f(y_n + \beta h f(y_n)).$$

Se pide:

- a) Interpretar geométricamente el esquema.
- b) Comprobar que el método es estable y determinar α y β para que ($M1P$) sea de orden 2, suponiendo que f tiene la suficiente regularidad, la cual se especificará.
- c) Estimar el error de consistencia en este caso, en términos de f y sus derivadas.

22. Para el problema anterior usamos ahora el esquema implícito

$$(M1P) \quad y_{n+1} = y_n + \alpha h f(\beta y_n + \mu y_{n+1}).$$

Se pide:

- a) Determinar una cota superior para h que permita asegurar que el método está bien definido, suponiendo que $\alpha \neq 0$.
- b) Comprobar que el método es estable y determinar α , β y μ para que ($M1P$) sea de orden 2 al menos, suponiendo que f tiene la suficiente regularidad, la cual se especificará.
- c) Estimar el error de consistencia en este caso, en términos de f y sus derivadas.

23. Consideramos el problema de valor inicial

$$y'(t) = \frac{2}{\pi} \arctan(y(t)), \quad t \in [0, 1], \quad y(0) = 1.$$

- a) Encontrar cotas para $y''(t)$ e $y'''(t)$ en el intervalo $t \in [0, 1]$ sin hallar $y(t)$ explícitamente
- b) Si se resolviese este problema usando el método de Euler explícito con paso h constante sin cometer error en el valor inicial >Qué valor de $h > 0$ habrá que tomar para garantizar que el mayor error cometido sea menor que 10^{-3} ? No usar computador, sólo tener en cuenta las estimaciones de error básicas.

24. Comprobar que $f(y) = \sqrt{|y|}$ no es Lipschitz en ningún intervalo que contenga a $y = 0$. Hallar tres soluciones distintas de $y' = \sqrt{|y|}$ con $y(0) = 0$ en $(-1, 1)$ y estudiar su regularidad. ¿A cuál de ellas converge el método de Euler? Suponiendo la existencia de error inicial y tomando $y_0 = 10^{-10}$ ¿A qué solución converge el método de Euler?

25. Considerar la ecuación $u'(t) = \lambda u(t)$ con $u(0) = u_0$, donde $\lambda \in \mathbb{C}$ y $u : \mathbb{R} \rightarrow \mathbb{C}$.

- Discutir el crecimiento o decaimiento de la solución en términos de $Re(\lambda)$.

- Discutir el papel de $Im(\lambda)$ en el comportamiento de la solución.
- Escribir la ecuación como un sistema real 2×2 y estudiar la equivalencia.

26. Siendo $c(t)$ y $d(t)$ funciones continuas en $[a, b]$ obtener la solución del problema de valor inicial

$$y(0) = t_0, \quad y'(t) = c(t)y(t) + d(t), \quad t > 0.$$

27. Sea $y'(t) = Ay(t)$ con $y \in \mathbb{R}^n$ y A una matriz $n \times n$ real. Probar que en la norma matricial adecuada la constante de Lipschitz correspondiente es $L = \max \sqrt{\lambda_j}$ donde los λ_j son los autovalores de $A^T \cdot A$ (recuérdese que toda matriz simétrica diagonaliza en una base ortonormal).

28. Sea $f \in C^0([a, b] \times \mathbb{R})$ globalmente lipschitziana respecto a su segunda variable, con constante de Lipschitz L . Dado el Problema de Cauchy

$$\begin{cases} y' &= f(x, y) \text{ en } [a, b]; \\ y(a) &= \alpha, \end{cases}$$

consideramos su resolución mediante el siguiente método general de un paso :

$$(M) \quad \begin{cases} y_{n+1} &= y_n + h_n \Phi(x_n, y_n; h_n) \\ y_0 &= \alpha; \end{cases}$$

donde $\Phi : [a, b] \times \mathbb{R} \times [0, h_0] \mapsto \mathbb{R}$ es una función continua y globalmente Lipschitziana con respecto a la segunda variable (constante de Lipschitz L_Φ). Supongamos que el paso $h_n = x_{n+1} - x_n$ es variable.

Consideremos el problema perturbado

$$\begin{cases} z_{n+1} &= z_n + h_n [\Phi(x_n, y_n; h_n) + \delta_n] \\ y_0 &= \tilde{\alpha}; \end{cases}$$

Se pide :

a) Probar la estimación de estabilidad

$$|z_n - y_n| \leq e^{L_\Phi(x_n - a)} |z_0 - y_0| + \frac{e^{L_\Phi(x_n - a)} - 1}{L_\Phi} \max_{0 \leq k \leq n-1} |\delta_k|.$$

b) Suponiendo que el método es de orden p , deducir la estimación de error

$$|y(x_n) - y_n| \leq e^{L_\Phi(x_n - a)} |y(a) - y_0| + C \frac{e^{L_\Phi(x_n - a)} - 1}{L_\Phi} h^p,$$

para cierta constante $C > 0$.

29. Para el sistema lineal $y'(t) = Ay(t)$, ($t > 0$) y donde

$$A = \begin{pmatrix} -5 & -4 \\ 2 & 1 \end{pmatrix}$$

caracterizar el valor de la condición inicial para que al aplicar el método de Euler con $h = 1$ se tenga $\lim_n y_n = 0$.

30. Para el sistema lineal $y'(t) = Ay(t)$, ($t > 0$) y donde

$$A = \begin{pmatrix} 242 & 324 \\ -183 & -245 \end{pmatrix}$$

probar que cualquiera que sea el valor de la condición inicial al aplicar el método de Euler con $h < 1$ se tenga $\lim_n y_n^h = \lim_t y(t) = 0$. ¿Cómo es posible si los coeficientes son muy grandes? Esto indica que no hay relación entre los autovalores y los coeficientes.

31. Considerese el sistema 2×2 dado por $y' = Ay$ en donde la matriz A es real diagonalizable en \mathbb{R} . Demostrar que si los elementos de A son menores que uno en valor absoluto y $\lim_{t \rightarrow +\infty} y(t) = 0$, entonces al aplicar el método de Euler para cualquier $h < 1$ se cumple $\lim_{t \rightarrow +\infty} y_n^h = 0$ (usar el **teorema de los círculos de Gershgorin** que afirma que los autovalores de una matriz $A = (a_{ij})$ pertenecen a la unión de los círculos en el plano complejo dados por los discos B_i de centro a_{ii} y radio r_i en donde $r_i = \sum_{j \neq i} |a_{ij}|$).

32. Hallar el dominio de estabilidad absoluta, o A-estabilidad del método

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

33. Se dice que un método es L-estable si al aplicarlo a $y' = \lambda y$ se cumple $y_{n+1} = R(\lambda h)y_n$ donde R cumple $R(x) \rightarrow 0$ si $x \rightarrow -\infty$. Probar que la regla del trapecio no es L-estable mientras que el método de Euler implícito sí lo es.
34. Hallar la función de amplificación del método de Euler mejorado.
35. Estudiar el comportamiento de Euler explícito e implícito cuando se aplica al sistema $y' = Ay$ con A diagonalizable con autovalores reales.

Capítulo 3

Métodos de Euler implícito y de Crank-Nicolson

Resumen del tema

Primera lectura:

- Estudio de método de Euler implícito
- Estudio de método de Crank-Nicolson
- Sistemas: caso lineal

Sólo hay un parámetro en el método de Euler explícito que nos dice como el algoritmo va a trabajar en un intervalo de longitud T cuando lo apliquemos a un problema concreto. Este parámetro es representado por h o por N . Sabemos que la relación entre ambos es $hN = T$ y que se describen como

1. h la talla del paso a avanzar por el intervalo
2. N el número de puntos para subdividir el intervalo.

Aunque tradicionalmente se usa h , desde el punto de vista de la programación en el computador es mejor usar N para, así, plantear la longitud de los vectores donde se almacenan los valores del cálculo.

Sobre h hemos realizado dos restricciones, una para buscar una precisión deseada y otra, simplemente, porque si no se hace los valores calculados oscilan con amplitud cada vez mayor y divergen dentro del crecimiento exponencial. Esto es lo que hemos denominado como cálculos inestables. Esto es:

1. **Restricción por precisión:** Reducimos h para obtener precisión, esto es, si $E(h) = Ch$ es el error y buscamos $E(h) < \varepsilon$ entonces tenemos que tomar $h < \varepsilon C^{-1}$.

2. **Restricción de estabilidad:** Reducimos h para conseguir que el método no produzca oscilaciones crecientes en los problemas disipativos, por ejemplo $h < 2/\lambda$ en el caso $f(t, y) = -\lambda y$ con $\lambda > 0$.

A veces la segunda restricción es más exigente que la primera, sobre todo en los casos disipativos con $\lambda \gg 1$. En estos casos además sabemos que la constante C depende del tiempo y decrece con el tiempo por lo que ayuda a reducir el error. Eso no es natural ni bueno y es culpa del esquema de cálculo (en este caso Euler explícito) escogido.

Además de lo comentado anteriormente, otras dificultades han quedado al descubierto:

1. Tener un orden de convergencia uno, esto es $O(h)$, en el error global para el método de Euler **no tiene fuerza suficiente para competir** con las constantes que surgen en los problemas expansivos, ya que estas constantes crecen de manera exponencial; Por lo tanto, **mayor precisión sería bueno**.
2. En los problemas disipativos las constantes de la estimación de error podemos mejorarlas pero **necesitamos restringir el valor de h** de una forma que puede ser excesiva. El no hacerlo nos lleva a situaciones donde el cálculo se hace inestable (ver Figuras 3.1, 3.2 y 3.3) y por lo tanto inservible. Luego una **mejor estabilidad**, esto es una menor restricción sobre h , sería bueno.
3. La predicción inicial de la pendiente $f(t_n, y_n)$ es manifiestamente mejorable. En el mismo subintervalo $[t_n, t_{n+1}]$ **la pendiente en el final $f(t_{n+1}, y_{n+1})$ puede cambiar mucho con respecto a $f(t_n, y_n)$** tanto en el caso expansivo como en el contractivo.

Como consecuencia, resulta que el método de Euler explícito es mejorable al menos en dos sentidos

1. **mayor precisión:** mejorando el error local, pasando de $O(h^2)$ a $O(h^{p+1})$ con $p > 1$. Esto llevará a un error global $O(h^p)$ con $p > 1$.
2. **mejor estabilidad:** mejorando la restricción de estabilidad sobre h , o incluso eliminando esta restricción.

Vamos a **elaborar mejor la predicción de la dirección de avance inicial en tiempo**.

3.1. Mejora en la estabilidad: Euler implícito

En el caso de un **sistema disipativo** se pasa de pendientes extremadamente grandes y negativas a pendientes prácticamente nulas.

Ejemplo 79 En el problema tipo $w'(t) = -aw(t)$ con $w(0) = 1$ y $a \gg 1$ la solución es $w(t) = e^{-at}$ y la pendiente en $t = 0$ es $w'(0) = -a$ mientras que en $t = 1$ es $w'(1) = -a/e^a$. Vemos que $|-a| \gg |-a/e^a|$ y gráficamente se observa que usar la pendiente $-a$ para avanzar nos va a llevar muy lejos de la solución si no llevamos cuidado.

El **método de Euler implícito** aplicado en el intervalo $[0, T]$ consiste tomar una partición con $h = T/N$ y avanzar de acuerdo a

$$y_{n+1} = y_n + h f(t_{n+1}, y_{n+1}), \quad n = 0, 1, 2, \dots, N^{(h)} - 1.$$

Se interpreta geométricamente como avanzar en la dirección dada por la pendiente $f(t_{n+1}, y_{n+1})$ que hay en (t_{n+1}, y_{n+1}) en vez de la pendiente $f(t_n, y_n)$ que hay en (t_n, y_n) . Esto supone una mejora si la pendiente decae fuertemente desde t_n a t_{n+1} , es decir, si $|y'(t_n)| \gg |y'(t_{n+1})|$, o lo que es lo mismo, si $|f(t_n, y_n)| \gg |f(t_{n+1}, y(t_{n+1}))|$.

Ejemplo 80 La importancia de usar el método implícito se ve cuando se aplica al problema test

$$\begin{cases} \frac{dy}{dt} = -\lambda y(t), & 0 < t < T, \\ y(0) = 1, \end{cases}$$

para $\lambda \gg 1$. Si usamos el **método de Euler explícito** en una partición uniforme se obtienen los valores

$$y_{n+1} = (1 - \lambda h)^{n+1}, \quad n = 0, 1, 2, \dots, N - 1, \quad (N = T/h).$$

Sabemos que si

$$\frac{1}{\lambda} < h < \frac{2}{\lambda}$$

hay decaimiento pero es oscilatorio, comportamiento que no está presente en la solución verdadera, y si $h > 2\lambda^{-1}$ se tiene crecimiento en la solución aproximada lo que es totalmente contrario a la curva buscada. Por lo tanto, se reproduce el comportamiento cuantitativo y cualitativo de decaimiento uniforme sólo si

$$0 < 1 - \lambda h < 1 \Leftrightarrow \lambda h < 1 \Leftrightarrow h < \frac{1}{\lambda}$$

y esta condición puede ser muy restrictiva si $\lambda \gg 1$ ya que $h = T/N$, luego también influye la longitud del intervalo de cálculo. La restricción clave se puede ver como

$$N > T\lambda$$

o lo que es lo mismo en este caso ($f(t, y) = -\lambda y$)

$$N > T |\partial_y f(t, y)|.$$

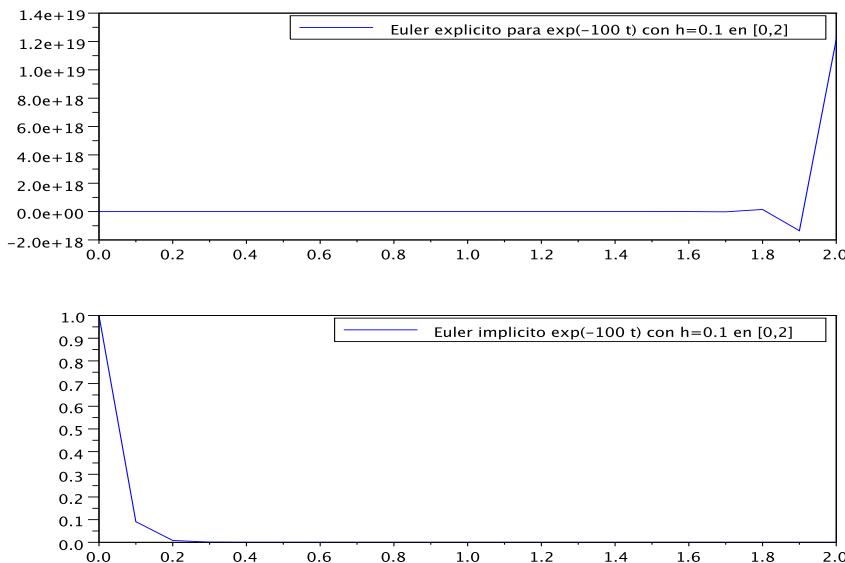


Figura 3.1: Euler explícito con $h = 0.1$ comparado con Euler implícito para $\lambda = 100$. El valor de h óptimo para Euler explícito es $h < 0.01$.

Observación 75 Recordamos otra vez que el producto de la longitud del intervalo de integración por la derivada de la función pendiente con respecto a la segunda variable TL_f es el valor importante que marca la estabilidad intrínseca del problema. Cuanto mayor sea este producto más tenemos que reducir el valor de h , o lo que es lo mismo, más grande hay que tomar N , esto es, más puntos, hay que tomar en el intervalo $[0, T]$.

Por otro lado, si usamos el **método de Euler implícito**, obtenemos la recurrencia

$$(1 + \lambda h)y_{n+1} = y_n, \quad n = 0, 1, 2, \dots, N - 1$$

de donde

$$y_{n+1} = (1 + \lambda h)^{-(n+1)}, \quad n = 0, 1, 2, \dots, N - 1$$

y por lo tanto, el comportamiento cualitativo se puede reproducir para cualquier valor de $h > 0$ (ver Figuras 3.1, 3.2 y 3.3) y la restricción de estabilidad ha desaparecido.

Observación 76 Nuestro problema de partida, $y'(t) = f(t, y(t))$ es no lineal en general, luego una aproximación no lineal funciona mejor que una lineal. Además, ¡no hay mejoras a coste cero!

La solución a la segunda dificultad planteada al principio de este tema parece simple si no fuese porque nos hemos encontrado con la necesidad de resolver una ecuación

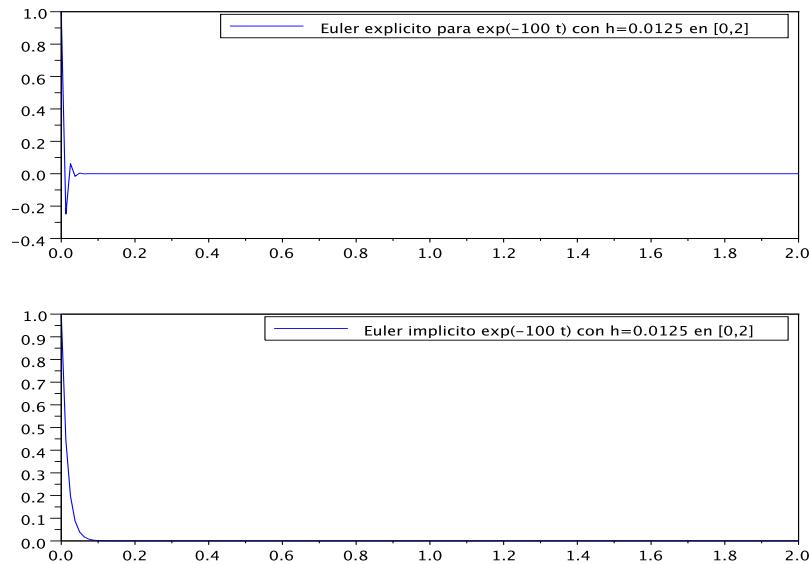


Figura 3.2: Euler explícito con $h = 0.0125$ comparado con Euler implícito para $\lambda = 100$. El valor de h óptimo para Euler explícito es $h < 0.01$.

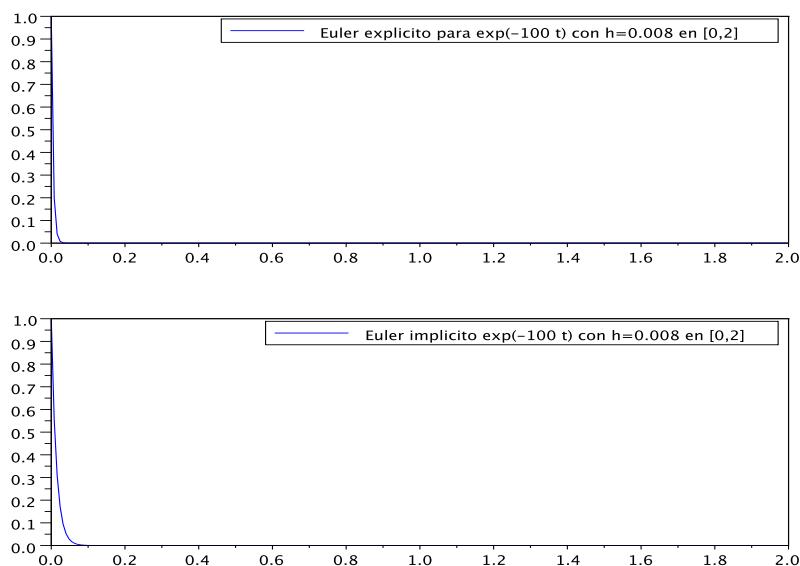


Figura 3.3: Euler explícito con $h = 0.008$ comparado con Euler implícito para $\lambda = 100$. El valor de h óptimo para Euler explícito es $h < 0.01$.

que en general es no lineal. Seguimos la misma idea de avanzar por una linea recta una distancia h , pero en vez de tomar $f(t_n, y_n)$ como pendiente de salida para avanzar de t_n a t_{n+1} tomamos la pendiente en el punto de llegada $f(t_{n+1}, y_{n+1})$. Esto presenta la dificultad de resolver un problema implícito en cada paso del proceso iterativo, a saber, para (t_0, y_0) dado buscamos y_1 que es la solución de la ecuación

$$z = y_0 + h f(t_0 + h, z),$$

y en general, y_{n+1} se obtiene como la solución de

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

Resolvemos para z y entonces $y_{n+1} = z$.

La función de iteración para la ecuación de punto fijo es $z = G(z)$ donde

$$G(z) = y_n + h f(t_{n+1}, z)$$

y por lo tanto

$$G(z) - G(w) = h [f(t_{n+1}, z) - f(t_{n+1}, w)]$$

luego una primera aproximación para garantizar solución es tener en cuenta que

$$|G(z) - G(w)| \leq L_f h |z - w|$$

y tendremos que usar $h < 1/L_f$ para que converja usando punto fijo, lo que obliga a tener h muy pequeño si $L_f \gg 1$. En todo caso, la restricción sobre h no es tan severa en la práctica. Normalmente, se resuelve este tipo de problemas mediante el método de Newton para aproximar la solución de $F(z) = 0$, donde

$$F(z) = z - (y_n + h f(t_{n+1}, z)).$$

Ejemplo 81 *Observar la diferencia entre Euler explícito e implícito cuando se aplican al problema de crecimiento de población limitado*

$$\begin{cases} \frac{dy}{dt} = y(t)(1 - y(t)), & t > 0, \\ y(0) = y_0 > 0, \end{cases} \quad (3.1)$$

Si $f(t, y)$ es lineal en la variable y no hay dificultad puesto que cuando

$$f(t, y) = a(t)y + b(t)$$

entonces se puede despejar y_{n+1} limpiamente:

$$y_{n+1} = y_n + h (a(t_{n+1})y_{n+1} + b(t_{n+1})) \Rightarrow y_{n+1} = \frac{y_n + h b(t_{n+1})}{1 - h a(t_{n+1})}.$$

El esquema de cómputo queda como ($N = T/h$)

$$y_{n+1} = \frac{1}{1 - h a(t_{n+1})} y_n + \frac{h b(t_{n+1})}{1 - h a(t_{n+1})}, \quad n = 0, 1, 2, \dots, N-1 \quad (3.2)$$

y sólo nos debemos de preocupar de buscar $h > 0$ tal que $1 - h a(t_{n+1}) \neq 0$ en cada paso. Esto se tiene garantizado para cualquier valor de $h > 0$ si $a(t) = \partial_y f(t, y) < 0$ o, lo que es lo mismo, si el sistema es disipativo. En particular,

1. si $a(t) \geq 0$ y tomamos $a_{max} = \max_t a(t)$ está claro que con $h < (a_{max})^{-1}$ se tiene que $1 - h a(t_{n+1}) \geq 1 - h a_{max} > 0$ y el esquema está bien definido.
2. si $a(t)$ cambia signos repetimos la idea pero ahora con $a_{max} = \max_t |a(t)|$.
3. si $-B \leq a(t) \leq -b < 0$ para valores positivos $0 < b < B$ entonces

$$1 - h a(t) \geq 1 + hb > 1 > 0, \quad \forall h > 0$$

y no hay que hacer ninguna restricción sobre el paso h que no sea debida a la precisión que se quiera obtener.

Observación 77 *En los problemas lineales se encuentra la gran ventaja de usar este método implícito, ya que no plantea ninguna restricción sobre h en comparación con el uso del método explícito. Además, el coeficiente delante de y_n en (3.2) es menor que uno, lo que respeta la dinámica del problema continuo.*

Observación 78 *Cualquier proceso iterativo para resolver el problema no lineal*

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

se puede empezar en $z^0 = y_n + h f(t_{n+1}, y_n)$ que ya es bastante cercano al punto buscado y_{n+1} . Como consecuencia, las restricciones teóricas usualmente estos métodos iterativos se relajan bastante en la práctica computacional.

Observación 79 *Veremos que Euler explícito y Euler implícito poseen cotas de error similares. Pero son mejores en el caso de Euler implícito sobre todo cuando el problema es contractivo.*

Observación 80 *La restricción sobre h para resolver los problemas no lineales no es importante ya que las restricciones por precisión y por estabilidad son mas fuertes usualmente.*

- **Si el sistema es expansivo** los dos métodos dan similares cotas de error y suele ser más preciso Euler implícito.
- **Si el sistema es disipativo**, esto es, $\partial_u f(t, u) < -b < 0$, los dos métodos dan cotas de error similares pero con Euler implícito la restricción de estabilidad sobre h no existe.
- **Si el problema es lineal**, suele ser mejor usar Euler implícito en cualquier caso.

3.1.1. Error local para Euler implícito

Usualmente se presenta el esquema como

dado u_0 ,

$$u_{n+1} = u_n + h f(t_{n+1}, u_{n+1}), \quad n = 0, 1, \dots, N - 1$$

pero puede ser más interesante y claro hacerlo de la siguiente forma:

Dado u_0 , para $n = 0, 1, \dots, N - 1$

1. obtener z tal que $z = u_n + h f(t_{n+1}, z)$
2. definir $u_{n+1} = z$ y repetir

De momento suponemos que el esquema permite el cálculo de u_{n+1} en cada paso. Para ver el error local en un punto genérico fijo t_\star suponemos, como siempre, que partimos del dato exacto $(t_\star, u(t_\star))$ y aplicamos el esquema de cálculo. Entonces, tenemos \tilde{z} tal que

$$\tilde{z} = u(t_\star) + h f(t_\star + h, \tilde{z})$$

y construimos el error local asociado

$$l(u(t_\star); h) = u(t_\star + h) - \tilde{z} = u(t_\star + h) - u(t_\star) - h f(t_\star + h, \tilde{z})$$

El desarrollo de Taylor con respecto al punto $(t_\star, u(t_\star))$ siendo los incrementos $(h, h f(t_\star + h, \tilde{z}))$ y usando f , f_t , etc... cuando se evalúa en el punto $(t_\star, u(t_\star))$, nos permite escribir:

$$\begin{aligned} f(t_\star + h, \tilde{z}) &= f(t_\star + h, u(t_\star) + h f(t_\star + h, \tilde{z})) \\ &= f + h f_t + h f(t_\star + h, \tilde{z}) f_y + O(h^2) \end{aligned}$$

Usando de forma recursiva esta expresión dentro del segundo miembro obtenemos de forma limpia el término que corresponde a la segunda derivada $y'' = f_t + f f_y$:

$$\begin{aligned} f(t_\star + h, \tilde{z}) &= f + h f_t + h [f + h f_t + h f(t_\star + h, \tilde{z}) f_y + O(h^2)] f_y + O(h^2) \\ &= f + (h f_t + h f f_y) + h^2 f_t f_y + h^2 f(t_\star + h, \tilde{z}) f_y^2 + O(h^3) f_y + O(h^2) \end{aligned}$$

de donde

$$f(t_\star + h, u(t_\star) + h f(t_\star + h, \tilde{z})) = u'(t_\star) + h u''(t_\star) + O(h^2)$$

y el término $O(h^2)$ no se puede eliminar en general. Como consecuencia, se tiene

$$\begin{aligned} l(u(t_\star); h) &= h u'(t_\star) + \frac{h^2}{2} u''(t_\star) + O(h^3) - h [u'(t_\star) + h u''(t_\star) + O(h^2)] \\ &= -\frac{h^2}{2} u''(t_\star) + O(h^3) = O(h^2) \end{aligned}$$

luego, como con Euler explícito, tenemos

$$l(u(t_\star); h) = u(t_\star + h) - \tilde{z} = O(h^2) \Rightarrow \mathcal{T}(u(t_\star); h) = O(h).$$

Observación 81 Se puede intentar ser muy preciso en los desarrollos de Taylor que acabamos de hacer. Pero realmente no importa la forma exacta del coeficiente de la potencia h^2 sino saber que en algún caso no es cero. Esto se puede averiguar simplemente usando el problema modelo $u'(t) = u(t)$ donde el error local se puede obtener de forma explícita y se ve que este coeficiente no es cero. Con que se tenga para un caso ya sirve para poder decir que no se puede mejorar la expresión $\mathcal{T}(u; h) = O(h)$ para toda función u , es decir, por ejemplo, no se puede tener $\mathcal{T}(u; h) = O(h^2)$. Veamoslo:

Al aplicar el esquema para $f(t, u) = u$ y buscar el error local sobre la curva concreta $u' = u$ se puede comprobar que el error local es proporcional a h^2 y no se puede hacer más pequeño: Tenemos que \tilde{z} es tal que

$$\tilde{z} = u(t_*) + h f(t_* + h, \tilde{z}) = u(t_*) + h\tilde{z}$$

luego

$$\tilde{z} = \frac{1}{1-h}u(t_*) = (1 + h + h^2 + h^3 + \dots)u(t_*), \quad h < 1$$

pero como $u'(t) = u(t)$ nos dice que $u(t) = e^t$ y que $u^{(j)}(t) = u(t)$ para cualquier j , entonces

$$\tilde{z} = u(t_*) + hu'(t_*) + h^2u''(t_*) + \dots, \quad h < 1$$

de donde

$$\begin{aligned} l(u(t_*); h) &= u(t_* + h) - u(t_*) - hu'(t_*) - h^2u''(t_*) + \dots \\ &= -\frac{h^2}{2}u(t_*) + O(h^3) = O(h^2), \quad h < 1 \end{aligned}$$

Observación 82 Para el problema modelo no trivial $u'(t) = u(t)$ se cumple que su solución $u(t) = e^t$ tiene todas las derivadas no nulas y que $f(t, u) = u$ tiene nulas todas sus derivadas parciales salvo $f_u = 1$.

3.2. Error global para Euler implícito

Para estudiar el **error global**, vemos la diferencia $u(t_{n+1}) - u_{n+1}$ donde $u_{n+1} = z$ con $z = z_{n+1}$ dado por

$$z_{n+1} = u_n + hf(t_{n+1}, z_{n+1}).$$

Igual que con Euler explícito, tenemos que

$$\begin{aligned} u(t_{n+1}) - u_{n+1} &= u(t_{n+1}) - \tilde{z}_{n+1} + \tilde{z}_{n+1} - z_{n+1} \\ &= l(u(t_n); h) + \tilde{z}_{n+1} - z_{n+1} = O(h^2) + \tilde{z}_{n+1} - z_{n+1} \end{aligned}$$

Para comparar z y \tilde{z} nos vamos a sus definiciones:

$$\begin{aligned}\tilde{z} - z &= u(t_*) + h f(t_* + h, \tilde{z}) - (u_n + h f(t_{n+1}, z)) \\ &= u(t_*) - u_n + h [f(t_{n+1}, \tilde{z}) - f(t_{n+1}, z)] \\ &= u(t_*) - u_n + h \partial_u f(t_{n+1}, s)(\tilde{z} - z)\end{aligned}$$

de donde

$$[1 - h \partial_u f(t_{n+1}, s)](\tilde{z} - z) = u(t_*) - u_n.$$

Tomando valor absoluto y usando la estimación de Lipschitz para la función f , si $1 - hL_f > 0$, es decir, $h < 1/L_f$, entonces

$$|\tilde{z} - z| \leq \frac{1}{1 - hL_f} |u(t_n) - u_n|, \quad h < 1/L_f.$$

Observación 83 El factor de amplificación debido al esquema es $(1 - hL_f)^{-1}$. En el caso $\partial_u f < 0$ si usamos el Teorema del valor medio este factor va a ser menor que uno sin restricciones sobre h . Esto es así porque el factor tendrá la forma $(1 - h\partial_u f)^{-1}$ y es claro que $(1 - h\partial_u f)^{-1} < 1$. Luego la restricción $h < L_f^{-1}$ es algo más teórico que práctico.

Usándolo en la estimación para $u(t_{n+1}) - u_{n+1}$ tenemos

$$|u(t_{n+1}) - u_{n+1}| \leq O(h^2) + \frac{1}{1 - hL_f} |u(t_n) - u_n|$$

es decir, si ponemos $e_n = |u(t_n) - u_n|$, tenemos

$$e_{n+1} \leq \frac{1}{1 - hL_f} e_n + O(h^2)$$

y se ve que hay convergencia de orden 1 con respecto a h siendo además el error global decreciente por un factor de $(1 - hL_f)^{-1}$.

Ejercicio 82 Terminar la prueba de convergencia en el caso general $f = f(t, y)$.

Ejemplo 83 Veamos el caso $f(t, u) = -au$ con $a > 0$. Tenemos

$$u(t_{n+1}) - u_{n+1} = u(t_n) - \tilde{z} + \tilde{z} - u_{n+1}$$

en donde

$$\tilde{z} = \frac{1}{1 + ah} u(t_n), \quad u_{n+1} = \frac{1}{1 + ah} u_n$$

reordenando términos

$$u(t_{n+1}) - u_{n+1} = u(t_n + h) - \frac{1}{1+ah}u(t_n) + \frac{1}{1+ah}(u(t_n) - u_n)$$

y teniendo en cuenta que $u(t) = e^{-at}$ luego $u^{(j)}(t) = (-a)^j u(t)$ tenemos que si $ah < 1$ entonces

$$u(t_{n+1}) - u_{n+1} = \frac{1}{1+ha}(u(t_n) - u_n) - \frac{1}{2}h^2u''(t) + \dots$$

es decir

$$u(t_{n+1}) - u_{n+1} = \frac{1}{1+ha}(u(t_n) - u_n) + O(h^2)$$

y vemos que podemos seguir el mismo análisis que con Euler explícito y con la sola restricción de $ah < 1$ para realizar la convergencia. Como el factor de amplificación es $K(h) = (1+ha)^{-1}$ tenemos una gran ventaja para este tipo de problemas ya que se reduce el error en cada paso, el factor de amplificación $K(h)$ es menor que 1. La desventaja es que sigue siendo un método de primer orden.

Es el momento de recuperar los ejemplos vistos en el tema anterior sobre problemas rígidos en donde comparábamos Euler explícito e implícito. Estos problemas son lineales por lo que para usar Euler implícito hay que trabajar con la fórmula dada por la ecuación (3.2) adaptada a cada caso.

Ejemplo 84 Ecuación de Dahlquist-Bjorck. (1974)

$$\begin{cases} y'(t) = 100(\sin(t) - y(t)), & 0 < t \leq 3, \\ y(0) = 0. \end{cases}$$

posee como solución

$$y(t) = e^{-100t}y_0 + \frac{\sin(t) - 100^{-1}\cos(t) + 100^{-1}e^{-100t}}{1 + 100^{-2}}$$

y el modo transitorio exponencial e^{-100t} decae muy rápido. Este modo debe ser capturado bien por el método explícito y esto nos lleva a un valor de h excesivamente pequeño. Se puede ver en la Figura 3.4 el comportamiento del método de Euler explícito en comparación con el método de Euler implícito. Aquí es $\partial_y f = -100$ y debemos de usar $h < 2/100$ en Euler explícito, luego si trabajamos en $[0, 3]$ debe ser

$$N > 3 * 100/2 = 150$$

para que Euler explícito funcione bien.

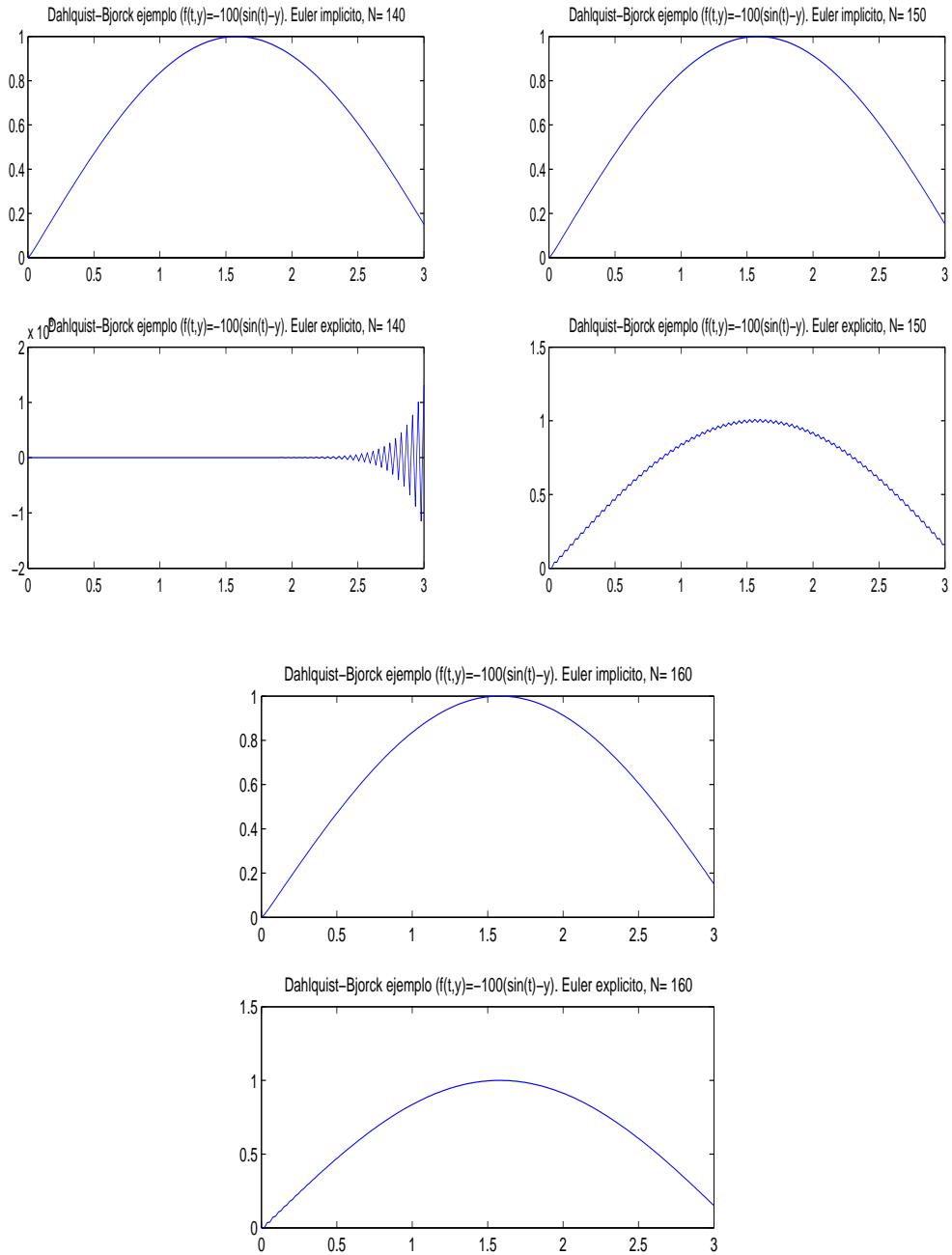


Figura 3.4: Cálculo de la solución del problema de Dahlquist-Bjorck usando Euler explícito y Euler implícito

Ejemplo 85 Ecuación de Prothero-Robinson: parecido al ejemplo anterior,

$$\begin{cases} y'(t) &= L(\varphi(t) - y(t)) + \varphi'(t), & 0 < t, \\ y(0) &= y_0 \end{cases}$$

la solución exacta es

$$y(t) = e^{-L t}(y_0 - \varphi(0)) + \varphi(t)$$

y otra vez el modo rápido $e^{-L t}$ para $L \gg 1$ debe ser capturado correctamente. Aquí $\varphi(t)$ puede ser una función suave sin cambios bruscos y es el modo estacionario.

Incluso en el caso en el que $y_0 = \varphi(0)$ y aparentemente el modo rápido no está presente en la solución, sí que se encuentra en todas las soluciones vecinas y se debe también capturar como si estuviese presente, ver Figura 3.5. Para la Figura 3.6 el dato en $t = 0$ para las curvas que caen sobre la fase transitoria está muy lejos de $y_0 = 1$. De hecho, esta “lluvia de curvas” se ha obtenido usando tiempo inicial $t_0 = 0.5, 1, 1.5$ y valores $y(t_0)$ cercanos al valor de la fase transitoria en estos tiempos.

Observación 84 Podemos ver como la rigidez es también un concepto de eficiencia. Con un simple método de Euler explícito podemos obtener la solución pero necesitamos reducir mucho el parámetro de discretización e incrementar por lo tanto el trabajo computacional a realizar (¡como lo hace el computador parece que no cuesta... pero sí!)

3.3. Mejora en el orden: Método de Crank-Nicolson

Usualmente se presenta el esquema como:

Dado u_0 ,

$$u_{n+1} = u_n + \frac{h}{2} [f(t_n, u_n) + f(t_{n+1}, u_{n+1})], \quad n = 0, 1, \dots, N-1.$$

pero puede ser más interesante y claro hacerlo de la siguiente forma:

Dado u_0 , para $n = 0, 1, \dots, N-1$

1. obtener z tal que $z = u_n + \frac{h}{2} [f(t_n, u_n) + f(t_{n+1}, z)]$,
2. definir $u_{n+1} = z$ y repetir

Se puede interpretar de varias formas:

- antes de avanzar en una dirección dada, muestrear y promediar, esto es, promedio de dos pendientes: la del principio y la del final.

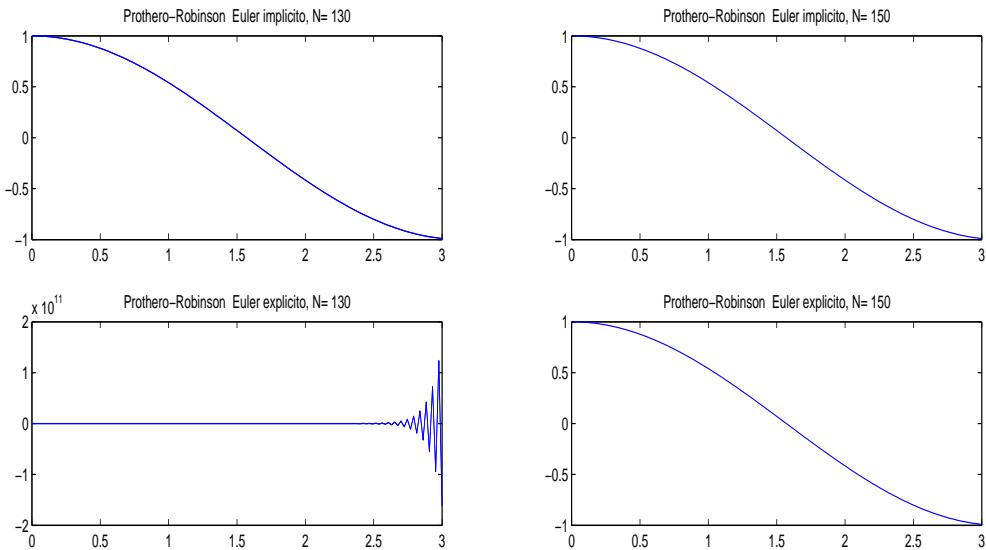


Figura 3.5: Euler Explícito e implícito sobre la ecuación de Prothero-Robinson

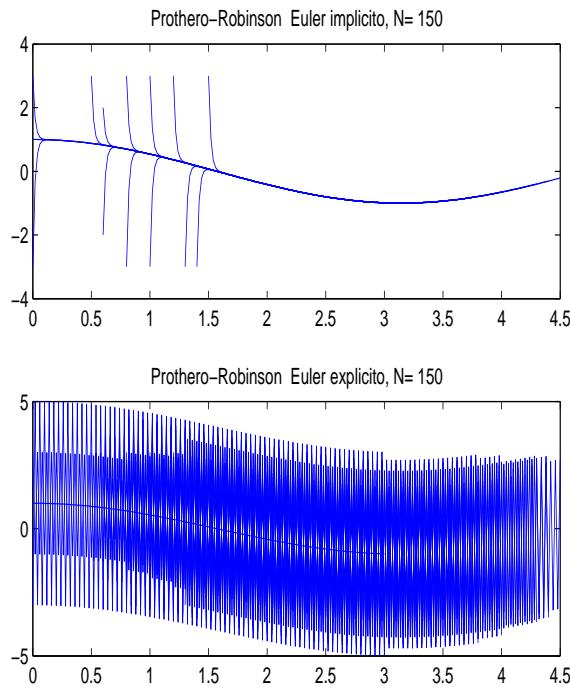


Figura 3.6: Comportamiento de las curvas vecinas a la fase transitoria. Cálculo posible con Euler implícito pero imposible con Euler explícito. El dato en $t = 0$ para las curvas que caen sobre la fase transitoria está muy lejos de $y_0 = 1$.

- Aplicar la fórmula de los trapecios a la expresión integral del problema de Cauchy. Usaremos mejor el punto de vista anterior, pero en muchos textos se usa también esta idea que proviene de escribir la edo como una ecuación integral y usar una fórmula de cuadratura.

Observación 85 *El método de Heun consiste en linealizar esta ecuación y se escribe como:*

Dado u_0 ,

$$\begin{aligned}\tilde{u}_{n+1} &= u_n + hf(t_n, u_n), \\ u_{n+1} &= u_n + \frac{h}{2} [f(t_n, u_n) + f(t_{n+1}, \tilde{u}_{n+1})], \quad n = 0, 1, \dots, N - 1.\end{aligned}$$

Este método tiene el mismo orden que el método de Crank-Nicolson pero una región de estabilidad más pequeña.

Como se hizo antes, podemos garantizar la existencia de solución a cada paso si la ecuación para z

$$z = u_n + \frac{h}{2} [f(t_n, u_n) + f(t_{n+1}, z)]$$

es resoluble. Esto queda garantizado para una iteración de punto fijo si

$$\frac{h}{2} |f_y(t_{n+1}, u)| < 1$$

y de forma más general, si $L_f = \max_{t,u} |f_y(t, u)|$ tenemos garantizada la solución si

$$h < 2/L_f.$$

Observación 86 *La restricción práctica, para un cálculo específico, sobre h es menos severa en general. Usamos Newton en vez de el método de las aproximaciones sucesivas y cuando el problema es rígido la convergencia es bastante rápida.*

3.3.1. Error local para Crank-Nicolson

De momento suponemos que el esquema permite el cálculo de u_{n+1} en cada paso. Para ver el error local en un punto genérico fijo t_\star suponemos, como siempre que partimos del dato exacto $(t_\star, y(t_\star))$ y aplicamos el esquema de cálculo. Entonces, tenemos \tilde{z} tal que

$$\tilde{z} = y(t_\star) + \frac{h}{2} [f(t_\star, y(t_\star)) + f(t_\star + h, \tilde{z})].$$

Observemos que también es

$$\tilde{z} = y(t_\star) + \frac{h}{2} y'(t_\star) + \frac{h}{2} f(t_\star + h, \tilde{z}).$$

Construimos ahora

$$\begin{aligned}
 l(y(t_\star); h) &= y(t_\star + h) - \tilde{z} = y(t_\star + h) - \{y(t_\star) + \frac{h}{2}[f(t_\star, y(t_\star)) + f(t_\star + h, \tilde{z})]\} \\
 &= y(t_\star + h) - y(t_\star) - \frac{h}{2}y'(t_\star) - \frac{h}{2}f(t_\star + h, \tilde{z}) \\
 &= \frac{h}{2}y'(t_\star) + \frac{h}{2}y''(t_\star) + O(h^3) - \frac{h}{2}f(t_\star + h, \tilde{z})
 \end{aligned}$$

en donde hemos hecho el desarrollo de Taylor de $y(t_\star + h)$. Tenemos que

$$f(t_\star + h, \tilde{z}) = f(t_\star + h, y(t_\star) + \frac{h}{2}y'(t_\star) + \frac{h}{2}f(t_\star + h, \tilde{z}))$$

Observación 87 Recordemos que desarrollo de Taylor de una función de dos variables se puede describir de forma simbólica como

$$f(t + \Delta t, y + \Delta y) = \sum_{n \geq 0} \frac{1}{n!} \{\Delta t \partial_t + \Delta y \partial_y\}^n f(t, y).$$

El desarrollo de Taylor con respecto al punto $(t_\star, y(t_\star))$ siendo los incrementos $(h, \frac{h}{2}y'(t_\star) + \frac{h}{2}f(t_\star + h, \tilde{z}))$ y usando f , f_t , etc... cuando se evalua en el punto $(t_\star, y(t_\star))$, nos permite escribir:

$$\begin{aligned}
 f(t_\star + h, \tilde{z}) &= f(t_\star + h, y(t_\star) + \frac{h}{2}y'(t_\star) + \frac{h}{2}f(t_\star + h, \tilde{z})) \\
 &= f + hf_t + \{\frac{h}{2}y'(t_\star) + \frac{h}{2}f(t_\star + h, \tilde{z})\}f_y + O(h^2) \\
 &= f + hf_t + \frac{h}{2}ff_y + \frac{h}{2}f(t_\star + h, \tilde{z})f_y + O(h^2)
 \end{aligned}$$

y si reiteramos el desarrollo en la expresión de $f(t_\star + h, \tilde{z})$ ahora en el segundo miembro llegamos a

$$\begin{aligned}
 f(t_\star + h, \tilde{z}) &= f + hf_t + \frac{h}{2}ff_y + \frac{h}{2}f(t_\star + h, \tilde{z})f_y + O(h^2) \\
 &= f + hf_t + \frac{h}{2}ff_y + \frac{h}{2}[f + hf_t + \frac{h}{2}ff_y + \frac{h}{2}f(t_\star + h, \tilde{z})f_y + O(h^2)]f_y + O(h^2) \\
 &= f + hf_t + \frac{h}{2}ff_y + \frac{h}{2}ff_y + \frac{h^2}{2}f_t f_y + \frac{h^2}{2^2}ff_y^2 + O(h^2) \\
 &= f + hf_t + \frac{h}{2}ff_y + \frac{h}{2}ff_y + O(h^2) = y' + hy'' + O(h^2)
 \end{aligned}$$

luego

$$\frac{h}{2}f(t_\star + h, \tilde{z}) = \frac{h}{2}y'(t_\star) + \frac{h^2}{2}y''(t_\star) + O(h^3)$$

y como consecuencia,

$$\begin{aligned} l(y(t_\star); h) &= \frac{h}{2}y'(t_\star) + \frac{h}{2}y''(t_\star) + O(h^3) - \frac{h}{2}f(t_\star + h, \tilde{z}) \\ &= \frac{h}{2}y'(t_\star) + \frac{h}{2}y''(t_\star) + O(h^3) - \frac{h}{2}y'(t_\star) - \frac{h^2}{2}y''(t_\star) - O(h^3) = O(h^3) \end{aligned}$$

luego hemos ganado una potencia en el orden del error local

$$l(y(t_\star); h) = y(t_\star + h) - \tilde{z} = O(h^3) \Rightarrow \mathcal{T}(y(t_\star); h) = O(h^2) \quad h \rightarrow 0.$$

Esto implica un error global de orden 2 en h y hemos mejorado un orden la precisión tomando más información adecuadamente.

Observación 88 *Al aplicar el esquema $y' = y$ se puede comprobar que el error local es proporcional a h^3 y no se puede hacer más pequeño.*

3.4. Error global para Crank-Nicolson

Se procede como con el método de Euler implícito.

Ejercicio 86 *Realizar la prueba de convergencia en el caso general $f = f(t, y)$.*

Ejemplo 87 *En el caso simple de $y' = -ay$ con $a > 0$ tenemos*

$$e_{n+1} = e_n + \frac{h}{2}(-ae_n + (-ae_{n+1})) + O(h^3),$$

de donde

$$(1 + ha/2)e_{n+1} = (1 - ha/2)e_n + O(h^3),$$

luego tenemos la estimación para el error

$$|y_{n+1} - y(t_{n+1})| \leq K(h) |y_n - y(t_n)| + O(h^3)$$

siendo $K(h) = \frac{1 - ha/2}{1 + ha/2}$, por lo que $-1 < K(h) < 1$. Observar que

$$-1 < K(h) = \frac{1 - ha/2}{1 + ha/2} = 1 - ha + O(h^3 a^3) < 1, \quad h \rightarrow 0$$

Esto es una gran ventaja para este tipo de problemas ya que reduce el error en cada paso, el factor de amplificación $K(h)$ es menor que 1. Una desventaja de este método, es que para valores grandes de a se tiene $|K(h)| \sim 1$, lo que puede producir oscilaciones.

Ejemplo 88 Si tenemos el sistema lineal $y' = Ay$ entonces

1. el método de Euler explícito genera el esquema

$$y_{n+1} = (I_m + hA)y_n$$

2. el método de Euler implícito genera el esquema (recordemos que en general nunca hay que calcular explícitamente la inversa de una matriz)

$$(I_m - hA)y_{n+1} = y_n$$

3. Crank-Nicolson genera el esquema

$$y_{n+1} = y_n + \frac{h}{2}Ay_n + \frac{h}{2}Ay_{n+1}$$

luego

$$(I_m - \frac{h}{2}A)y_{n+1} = (I_m + \frac{h}{2}A)y_n.$$

4. Heun genera el esquema

$$y_{n+1} = y_n + \frac{h}{2}Ay_n + \frac{h}{2}A(y_n + Ay_n)$$

luego

$$y_{n+1} = (I_m + hA + \frac{h^2}{2}A^2)y_n.$$

En el caso A simétrica definida negativa, esto es, diagonalizable con todos sus autovalores negativos, se puede estudiar la restricción de estabilidad en la norma $\|\cdot\|_2$ de estos métodos en términos de los autovalores de una manera muy cómoda.

3.5. Ejercicios

1. Probar que el método del trapecio o de Crank-Nicolson es de segundo orden cuando se aplica a

$$w'(t) = t^2, \quad w(0) = 0.$$

La solución exacta es $w(t) = t^3/3$ y va a ser útil la fórmula de la suma de los cuadrados de números enteros

$$\sum_{i=0}^n i^2 = \frac{n(n+1)(2n+1)}{6}.$$

Esto indica que el orden del método es dos y que no se puede mejorar ya que existe al menos una curva donde es exactamente dos.

2. Hallar una fórmula para y_n cuando se aplica el método del trapecio a una ecuación de la forma $y'(t) = a y(t) + b$, $y(0) = 0$ con $y_0 = 0$. Deducir la expresión de la solución continua a partir del valor de y_n^h .

3. Dada la siguiente variante del método de Euler

$$\begin{cases} y_{n+1} &= y_n + h f\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hf(t_n, y_n)\right) \\ y_0 &= \alpha \end{cases}$$

- a) Comprobar de forma práctica y teórica el orden de convergencia del método en este problema.
- b) Buscar algún contraejemplo que demuestre que el orden del método no puede ser mayor que dos.

4. Interpretar geométricamente el siguiente método

$$y_{n+1} = y_n + h f\left(t_n + \frac{3}{4}h, \frac{1}{4}y_n + \frac{3}{4}y_{n+1}\right).$$

Identificar la función de incremento $\Phi(t, y; h)$ y comprobar que si $f(t, y)$ satisface una condición de Lipschitz con respecto a la variable y también lo hace $\Phi(t, y; h)$. Obtener una constante de Lipschitz para $\Phi(t, y; h)$ uniforme en h .

5. Aplicar el método de Euler implícito a la ecuación $y'(t) = -a y(t)$ con $a > 0$. Obtener el error local y obtener la estimación de error global en la forma más detallada y precisa posible incluyendo el error inicial.
6. Aplicar el método de Crank-Nicolson a la ecuación $y'(t) = -a y(t)$ con $a > 0$. Obtener el error local y obtener la estimación de error global en la forma más detallada y precisa posible incluyendo el error inicial.
7. Considerar la familia de métodos de un paso

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

donde $\theta \in (0, 1)$.

- a) Interpretar geométricamente lo que se está haciendo.
- b) Dar la condición mínima sobre h que garantize que está bien definido.
- c) Determinar el orden del error local en función de θ .
- d) Obtener la estimación de error global correspondiente.

8. Escribir la ecuación lineal

$$y^{(n)} + a_{n-1}y^{(n-1)} + \dots + a_1y' + a_0y = 0$$

en la forma matricial $Y'(t) = AY(t)$ donde $Y = (y, y', \dots, y^{(n-1)})$ y A es una matriz $n \times n$.

9. Sea $y'(t) = Ay(t)$ con $y \in \mathbb{R}^n$ y A una matriz $n \times n$ real. Probar que la constante de Lipschitz correspondiente es $L = \max \sqrt{\sigma_j}$ donde los σ_j son los autovalores de $A^T \cdot A$ (recuérdese que toda matriz simétrica diagonaliza en una base ortonormal).

10. Consideramos el método de segundo orden

$$y_{n+1} = y_n + hg(t_n + \frac{h}{2}, y_n + \frac{h}{2}g(t_n, y_n))$$

y tomamos el problema escalar $y'(t) = g(t, y(t))$ junto con su transformación a problema autónomo usando $u(t) = t$ y $v(t) = y(t)$, luego $u'(t) = 1$; $v'(t) = g(u, v)$ con $u(0) = 1, v(0) = y(0)$. Aplicar el método a ambos problemas y observar si se obtiene el mismo resultado. Motivar la respuesta.

11. Comprobar que la aplicación del método anterior al problema $y' = Ay$ genera

$$y_{n+1} = (1 + Ah + \frac{h^2}{2}A^2)y_n.$$

12. Hallar el dominio de estabilidad absoluta, o A-estabilidad del método

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

13. Para el problema

$$\begin{cases} y'(t) &= (1 - 2t)y(t) \\ y(0) &= 1 \end{cases}$$

y en una partición uniforme obtener el esquema de Taylor correspondiente con residuo (error local) proporcional a h^3 .

14. El sistema

$$\begin{cases} y'_1(t) &= y_1 - 2y_2 + 4\cos(t) - 2\sin(t) & y_1(0) = 1 \\ y'_2(t) &= 3y_1 - 4y_2 + 5\cos(t) - 5\sin(t), & y_2(0) = 2 \end{cases}$$

posee por solución exacta

$$y_1(t) = \cos(t) + \sin(t), \quad y_2(t) = 2\cos(t).$$

a) Escribirlo como un problema de la forma

$$\frac{d}{dt} \vec{y} = A\vec{y} + G(\vec{t})$$

b) Aproximar la solución por el método de Euler explícito en el intervalo $[0, 1]$ con paso uniforme $h = 1/10$ y $h = 1/100$. Comparar con la solución exacta. Hacer estimaciones de error rigurosas y compararlas con los errores reales.

Capítulo 4

Métodos de Runge-Kutta

Resumen del tema

Primera lectura:

- interpretación geométrica
- Conceptos de estabilidad numérica, precisión y orden de convergencia.
- Otros métodos para alcanzar orden alto.

4.1. Introducción

Los métodos de Runge-Kutta se pueden interpretar geométricamente como una extensión del método de Euler progresivo en donde mejoramos el orden de convergencia realizando más evaluaciones de la función pendiente $f(t, y)$. Se pretende ajustar mediante un promedio la mejor pendiente por la que avanzar desde un punto t al punto $t+h$. Tienen la característica de que se generan cálculos no lineales puesto que se anidan evaluaciones de la función pendiente. Estos métodos fueron desarrollados en torno a 1900 por los matemáticos alemanes Carl David Tomé Runge y Martin Wilhelm Kutta.

Conceptualmente, es una mejora muy notable puesto que:

- tiene en cuenta el campo de soluciones y promedia la pendiente para avanzar entre las distintas curvas vecinas a la buscada, no sólo la solución que se busca.
- Por otro lado, la nolinealidad es una propiedad intrínseca del problema. Por lo que es muy razonable que esta nolinealidad aparezca en el método numérico.

Para un estudio más detallado se puede ver Lambert [20] y varias otras referencias.

Observación 89 Ni el método de Euler implícito ni el método de Crank-Nicolson pertenecen a la familia de métodos de Runge-Kutta explícitos, pertenecen a la de Runge-Kutta implícitos y también a la de métodos multipaso implícitos.

Vamos a ver varios ejemplos de métodos de tipo Runge-Kutta. El primero de ellos ya lo hemos estudiado antes y se incluye porque es una reinterpretacion.

1. **Método de Heun de orden 2:** Tomamos la pendiente inicial

$$k_1 = f(t_n, y_n)$$

y avanzamos en esa dirección hasta llegar al punto final del intervalo donde volvemos a tomar una nueva muestra para la pendiente:

$$k_2 = f(t_n + h, y_n + hk_1).$$

Con los dos valores k_1 y k_2 terminamos haciendo un promedio

$$k = \frac{1}{2}k_1 + \frac{1}{2}k_2$$

siendo esta la dirección de avance definitiva

$$y_{n+1} = y_n + h k.$$

También se puede entender siguiendo la misma idea que en el método de Crank-Nicolson, pero evitando la dificultad de resolver un problema implícito mediante la aproximación de y_{n+1} por la expresión

$$y_\star = y_n + hf(t_n, y_n)$$

para luego avanzar definitivamente con

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

Este método es de orden 2 con respecto a h .

2. **Regla del punto medio:** Tomamos la pendiente inicial

$$k_1 = f(t_n, y_n)$$

y avanzamos en esa dirección hasta llegar al punto medio del intervalo donde volvemos a evaluar la pendiente como muestra.

$$k_2 = f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1\right).$$

Con los dos valores k_1 y k_2 terminamos haciendo un promedio que simplemente consiste en usar el valor k_2

$$k = 0 k_1 + 1 k_2 = k_2$$

siendo esta la dirección de avance definitiva

$$y_{n+1} = y_n + h k.$$

También se puede interpretar como que usamos el punto medio del intervalo para aproximar: tomamos

$$y_* = y_n + \frac{h}{2} f(t_n, y_n)$$

y escribimos

$$y_{n+1} = y_n + h f\left(t_n + \frac{h}{2}, y_*\right).$$

Este método es de orden 2 con respecto a h .

3. El **método de Heun de orden 3**: se avanza desde t_n hasta $t_n + h/3$ con la pendiente k_1 y se genera k_2 . Luego desde t_n hasta $t_n + 2h/3$ con la pendiente k_2 y se genera k_3 . Finalmente se toma el promedio para avanzar definitivamente a $t_n + h$

$$\begin{aligned} k_1 &= f(t_n, y_n) && \text{(primera pendiente)} \\ k_2 &= f\left(t_n + \frac{1}{3}h, y_n + \frac{1}{3}hk_1\right) && \text{(segunda pendiente)} \\ k_3 &= f\left(t_n + \frac{2}{3}h, y_n + \frac{2}{3}hk_2\right) && \text{(tercera pendiente)} \\ y_{n+1} &= y_n + h\left(\frac{1}{4}k_1 + \frac{3}{4}k_3\right) && \text{(uso de pendiente promedio)} \end{aligned}$$

Este método es de orden 3 con respecto a h .

4. **Runge-Kutta clásico de orden 4**: es el ejemplo más popular. Se avanza desde t_n hasta $t_n + h/2$ con la pendiente k_1 y desde t_n también hasta el mismo punto $t_n + h/2$ pero ahora con la pendiente k_2 . Luego avanzamos hasta $t_n + h$ con k_3 y finalmente se toma el promedio para avanzar definitivamente a $t_n + h$

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1\right), \\ k_3 &= f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_2\right), \\ k_4 &= f(t_n + h, y_n + hk_3), \\ y_{n+1} &= y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4). \end{aligned}$$

La popularidad de este último método provienen de la era pre-computacional (antes de los años 50) puesto que los coeficientes son sencillos y además tiene orden 4. Esto fue decisivo para su extensión como el más usado ya que los cálculos se hacían a mano. Este método es de orden 4 con respecto a h .

La forma más general de un **método de Runge-Kutta** consiste en tomar S muestras de pendientes k_1, k_2, \dots, k_S y promediarlas para generar una nueva dirección de avance. En esta nueva dirección se da el paso definitivo.

1. Dado $y_0 \approx y(t_0)$, para $0 \leq n \leq N - 1$
2. Construir las pendientes k_1, k_2, \dots, k_S y obtener la pendiente promedio

$$k = \sum_{i=1}^S b_i k_i$$

siendo $b_1 + b_2 + \dots + b_S = 1$.

3. Dar el paso definitivo en esta dirección promedio generada

$$y_{n+1} = y_n + hk.$$

Al **número de muestras** S obtenidas se le llama **etapas del método** y los valores k_i siguen la misma idea de promediar pendientes. Es importante destacar que el número de muestras S no tiene porqué coincidir con el orden del método p , siempre es $p \leq S$. En los casos donde $S = p$ tenemos la situación óptima, pero esto se consigue sólo si $S \leq 4$.

4.2. Métodos de Runge-Kutta explícitos

Supongamos que partimos del valor $k_1 = f(t_n, y_n)$, entonces usamos la recta

$$r_1(x) = y_n + (x - t_n)k_1,$$

tomamos muestra en $x = t_n + c_2 h$

$$y_{n+c_2} = r_1(t_n + c_2 h) = y_n + c_2 h k_1$$

y calculamos la pendiente en este nuevo punto

$$k_2 = f(t_n + c_2 h, y_{n+c_2}) = f(t_n + c_2 h, y_n + c_2 h k_1).$$

Hemos avanzado desde t_n y por la dirección dada por k_1 una longitud $c_2 h$.

Tenemos ya dos valores distintos k_1 y k_2 y tiene sentido hacer un promedio entre ambos:

$$\alpha_{3,1}k_1 + \alpha_{3,2}k_2 \quad (\alpha_{3,1} + \alpha_{3,2} = 1)$$

por el que volver a avanzar desde t_n hasta $t_n + c_3h$. Tomamos entonces

$$\begin{aligned} k_3 &= f(t_n + c_3h, y_{n+c_3}) = f(t_n + c_3h, y_n + c_3h(\alpha_{3,1}k_1 + \alpha_{3,2}k_2)) \\ &= f(t_n + c_3h, y_n + h(c_3\alpha_{3,1}k_1 + c_3\alpha_{3,2}k_2)) \end{aligned}$$

Usando $a_{2,1} = c_2$ y para $i \geq 2$ la notación $a_{i,j} = \alpha_{i,j}c_i$, $j = 1, \dots, i-1$ lo podemos escribir como

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + c_2h, y_n + ha_{21}k_1), \\ k_3 &= f(t_n + c_3h, y_n + h(a_{31}k_1 + a_{32}k_2)), \end{aligned}$$

y en general, cuando tenemos k_1, k_2, \dots, k_{i-1} construimos

$$y_{n+c_i} = y_n + h \sum_{j=1}^{i-1} a_{i,j}k_j$$

siendo evidente que $c_i = a_{i1} + a_{i2} + \dots + a_{i,i-1}$ ya que

$$a_{i1} + a_{i2} + \dots + a_{i,i-1} = c_i(\alpha_{i,1} + \alpha_{i,2} + \dots + \alpha_{i,i-1}) = c_i.$$

Entonces calculamos la pendiente en este nuevo punto

$$k_i = f(t_n + c_ih, y_{n+c_i}).$$

Resumiendo, tomando $c_1 = 0$ y sabiendo que $c_2 = a_{21}$,

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + c_2h, y_n + ha_{21}k_1), \\ k_i &= f(t_n + c_ih, y_n + h(a_{i1}k_1 + a_{i2}k_2 + \dots + a_{i,i-1}k_{i-1})), \quad i = 3, \dots, S. \end{aligned}$$

Observación 90 *El adjetivo explícito se refiere a la forma de calcular de las pendientes k_j entre sí. La pendiente k_j se calcula en función de los valores k_i para $i < j$ que ya son conocidos.*

4.3. Runge-Kutta implícitos (Si tiempo permite)

La misma idea se puede generalizar haciendo que los valores k_1, k_2, \dots, k_S dependan todos entre sí. La dependencia entre las pendientes ahora no es explícita.

Observación 91 El adjetivo *implícito* se refiere a la forma de calcular de las pendientes k_j . La pendiente k_j se calcula en función todos los valores k_i sean estos conocidos o no.

En cada etapa i para $i = 1, 2, \dots, S$ construimos una pendiente promedio dada por

$$k_i = \sum_{j=1}^S \alpha_{i,j} k_j = \sum_{j=1}^{i-1} \alpha_{i,j} k_j + \sum_{j=i-1}^S \alpha_{i,j} k_j$$

donde tomamos $\alpha_{i,1} + \alpha_{i,2} + \dots + \alpha_{i,S} = 1$ arbitrarios, observar que el segundo sumando es la novedad e incluye los valores k_j para $j \geq i$. Ahora, continuamos igual que en el caso explícito. Avanzamos por la dirección k_i y desde t_n una longitud $c_i h$. Esto es, usamos la recta

$$r_i(x) = y_n + (x - t_n)k_i$$

y el valor tomado nuevo es

$$y_{n+c_i} = r_i(t_n + c_i h) = y_n + c_i h k_i$$

que se puede reescribir como

$$y_{n+c_i} = y_n + h \sum_{j=1}^S a_{i,j} k_j$$

donde $a_{i,j} = \alpha_{i,j} c_i$, $i, j = 1, 2, 3, \dots, S$. Entonces, para ser coherentes con nuestro desarrollo, queremos tener

$$k_i = f(t_n + c_i h, y_{n+c_i}),$$

o lo que es lo mismo,

$$k_i = f(t_n + c_i h, y_n + h(a_{i1} k_1 + a_{i2} k_2 + \dots + a_{iS} k_S)), \quad i = 1, 2, \dots, S.$$

Otra vez, se tiene que

$$a_{i1} + a_{i2} + \dots + a_{iS} = c_i (\alpha_{i,1} + \alpha_{i,2} + \dots + \alpha_{i,S}) = c_i, \quad i = 1, 2, \dots, S.$$

El cálculo de las pendientes k_i se hace mediante la resolución de una ecuación de punto fijo en general no lineal de dimensión S en la forma

$$K = F(K).$$

También puede ocurrir que la dependencia implícita no sea muy fuerte. Por ejemplo, en el caso donde sólo se usa a_{ii} nos encontramos con S ecuaciones no lineales desacopladas

$$k_i = f(t_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{ij} k_j + h a_{ii} k_i), \quad i = 1, 2, \dots, S$$

(separamos la suma en la parte conocida $h \sum_{j=1}^{i-1} a_{ij} k_j$ y la parte desconocida para k_i) es decir, para cada k_i tenemos que resolver una ecuación de punto fijo que sólo involucra a k_i :

$$k_i = g_i(k_i).$$

Aquí el esfuerzo computacional es menor y estos métodos se denominan **métodos semi-implícitos**.

Observación 92 Los métodos Runge-Kutta implícitos tienen propiedades de estabilidad y de orden de convergencia distintas de las que tienen los Runge-Kutta explícitos y son muy usados para los problemas rígidos. A cambio de esta ventaja su dificultad de cómputo es mayor y por ello no vamos a profundizar más.

4.4. Forma general y tablero de Butcher

Resumiendo, la forma más general de un **método de Runge-Kutta** consiste en tomar S muestras de pendientes k_1, k_2, \dots, k_S y promediarlas para avanzar en la nueva dirección obtenida. Esto es,

1. Dado $y_0 \approx y(t_0)$, para $0 \leq n \leq N - 1$
2. Construir las pendientes k_1, k_2, \dots, k_S definidas por

$$k_i = f(t_n + c_i h, y_n + h(a_{i1}k_1 + a_{i2}k_2 + \dots + a_{iS}k_S)), \quad i = 1, 2, \dots, S$$

siendo

$$\sum_{j=1}^S a_{ij} = c_i, \quad i = 1, 2, \dots, S$$

y obtener la pendiente promedio

$$k = \sum_{i=1}^S b_i k_i$$

siendo $b_1 + b_2 + \dots + b_S = 1$.

3. Dar el paso definitivo en esta dirección promedio generada

$$y_{n+1} = y_n + hk.$$

Los coeficientes $\{a_{i,j}\}$, $\{c_i\}$, $\{b_i\}$ determinan completamente el método de Runge-Kutta y normalmente se recolectan en lo que se conoce como **matriz o tablero de Butcher** (introducido por John Butcher [5] en los años 60.)

En el caso implícito el tablero queda relleno completamente en general

c_1	$a_{1,1}$	$a_{1,2}$	\dots	$a_{1,S-1}$	$a_{1,S}$
c_2	$a_{2,1}$	$a_{2,2}$	\dots	$a_{2,S-1}$	$a_{2,S}$
c_3	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	\dots	$a_{3,S}$
\vdots	\vdots	\vdots	\ddots	\ddots	\vdots
c_S	$a_{S,1}$	$a_{S,2}$	\dots	$a_{S,S-1}$	$a_{S,S}$
	b_1	b_2	\dots	b_{S-1}	b_S

mientras que en el caso explícito el tablero queda como

0	0	0	\dots	0	0
c_2	$a_{2,1}$	0	\dots	0	0
c_3	$a_{3,1}$	$a_{3,2}$	0	\dots	0
\vdots	\vdots	\vdots	\ddots	\ddots	0
c_S	$a_{S,1}$	$a_{S,2}$	\dots	$a_{S,S-1}$	0
	b_1	b_2	\dots	b_{S-1}	b_S

En general, se tiene la distribución matricial

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

donde A es una matriz cuadrada $S \times S$ y en el caso explícito contiene ceros en la diagonal y en la parte superior a la diagonal. De acuerdo a esto, los ejemplos del comienzo se pueden describir como sigue:

Ejemplo 89 El **Método de Heun de orden 2** es

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + h, y_n + hk_1) \\ y_{n+1} &= y_n + \frac{h}{2}(k_1 + k_2) \end{aligned}$$

luego su tablero es

0	0	0
1	1	0
	$1/2$	$1/2$

En el caso de la **regla del punto medio**:

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + h/2, y_n + h/2k_1) \\ y_{n+1} &= y_n + hk_2 \end{aligned}$$

y su tablero es

0	0	0
1/2	1/2	0
0	1	

En el **método de Heun de orden 3**:

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + \frac{1}{3}h, y_n + \frac{1}{3}hk_1) \\ k_3 &= f(t_n + \frac{2}{3}h, y_n + \frac{2}{3}hk_2) \\ y_{n+1} &= y_n + h(\frac{1}{4}k_1 + \frac{3}{4}k_3) \end{aligned}$$

su tablero es

0	0	0	0
1/3	1/3	0	0
2/3	0	2/3	0
0	1/4	3/4	

y finalmente para el **método Runge-Kutta clásico de orden 4** dado por

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1), \\ k_3 &= f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_2), \\ k_4 &= f(t_n + h, y_n + hk_3), \\ y_{n+1} &= y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4). \end{aligned}$$

su tablero es

0	0	0	0	0
1/2	1/2	0	0	0
1/2	0	1/2	0	0
1	0	0	1	0
	1/6	2/6	2/6	1/6

mientras que en el caso semi-implícito el tablero queda como

c_1	$a_{1,1}$	0	0	...	0	0
c_2	$a_{2,1}$	$a_{2,2}$	0	...	0	0
c_3	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	0	...	0
\vdots	\vdots	\ddots	\ddots	\ddots	\ddots	0
\vdots	\vdots	\ddots	\ddots	\ddots	0	0
c_S	$a_{S,1}$	$a_{S,2}$	$a_{S,S-1}$	$a_{S,S}$
	b_1	b_2	b_{S-1}	b_S

es decir, en la forma

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

donde A es una matriz cuadrada $S \times S$ con ceros en la parte superior a la diagonal.

Observación 93 *Estos esquemas son los métodos clásicos dentro de la familia general de métodos de Runge-Kutta. Los primeros métodos explícitos aparecieron en 1895 y hasta los años 1960 aproximadamente sólo se consideraron métodos explícitos. Se han desarrollado también métodos implícitos en los valores de k_1, \dots, k_S puesto que, aunque son computacionalmente más costosos, se encuentran mejor adaptados a los problemas rígidos.*

Observación 94 *Se dice que son métodos de un paso ya que se avanza desde t_n a t_{n+1}*

Observación 95 *Siempre son métodos de un paso explícitos con respecto al valor buscado y_{n+1} pero pueden no serlo con respecto a las pendientes k_1, k_2, \dots, k_S .*

4.5. Estudio del error local

Por construcción son todos métodos consistentes. Efectivamente, usando el desarrollo de Taylor en torno al punto $(t, y(t))$

$$k_i = f(t + c_i h, y(t) + h(a_{i1}k_1 + \dots + a_{iS}k_S)) = f(t, y(t)) + O(h) = y'(t) + O(h)$$

luego gracias a la restricción

$$b_1 + b_2 + \dots + b_S = 1,$$

tenemos

$$\begin{aligned} l(y(t); h) &= y(t+h) - y(t) - h \sum_{i=1}^S b_i k_i(t, y(t); h) \\ &= h y'(t) + O(h^2) - h \left(\sum_{i=1}^S b_i y'(t) - O(h) \right) \\ &= h y'(t) - h y'(t) + O(h^2) = O(h^2). \end{aligned}$$

Por lo tanto, el error local o de consistencia es siempre al menos $O(h^2)$. La idea ahora es usar los parámetros y el número de etapas para aumentar el orden local o de consistencia del método, esto lo veremos más adelante.

4.6. Estudio de la estabilidad: 0-estabilidad

Hemos visto la consistencia de los métodos y sabemos que todos son consistentes puesto que cumplen al menos

$$l(h) = O(h^2).$$

Veamos ahora la estabilidad. Como ya hemos visto, un aspecto fundamental es **la propagación de errores** a lo largo de los distintos pasos que tengamos que hacer, esto es la **estabilidad del método**.

Dado y_0 para calcular en $[t_0, t_0 + T]$ con $h = T/N^{(h)}$ consideramos el esquema

$$y_{n+1} = y_n + h \sum_{i=1}^S b_i k_i(t_n, y_n; h), \quad 0 \leq n \leq N^{(h)} - 1$$

junto con su perturbación: dados $z_0 = y_0 + \delta_0^{(h)}$,

$$z_{n+1} = z_n + h [\sum_{i=1}^S b_i k_i(t_n, z_n; h) + \delta_n^{(h)}], \quad 0 \leq n \leq N^{(h)} - 1.$$

Definición 90 Decimos que el **método es 0-estable** (leer cero-estable) cuando fijado el intervalo de cálculo $[t_0, t_0 + T]$ y particiones de talla h con $h = T/N$, si la perturbación cumple $|\delta_n^{(h)}| \leq \epsilon$ para $\epsilon > 0$, entonces existe una constante $C(T, f)$ y $h_0 > 0$ tal que para todo $h < h_0$ se cumple

$$|z_n^{(h)} - y_n^{(h)}| \leq C \{ \epsilon + |z_0 - y_0| \}, \quad 1 \leq n \leq N = T/h.$$

Observación 96 Generalizamos mediante el uso del término extra $\delta_n^{(h)}$ el concepto ya visto de estabilidad, antes usabamos $\delta_n^{(h)} = 0$, pero no afecta la idea básica

Esta es una propiedad fundamental del método numérico que nos permite amortiguar los errores que se introducen en cada paso del cálculo.

Observación 97 Hablamos del comportamiento cuando $h \rightarrow 0^+$.

Vamos a ver que todos los métodos de Runge-Kutta explícitos son cero estables. Para esto sólo tenemos que garantizar que la expresión

$$\Phi_f(t, y; h) = \sum_{i=1}^S b_i k_i(t, y; h)$$

cumple una condición de Lipschitz con respecto a su segundo argumento de manera uniforme para $h < h_0$, si es preciso.

Teorema 91 Existe un $h_0 > 0$ (puede ser $h_0 = +\infty$) tal que para todo $h < h_0$ se tiene M_{Φ_f} independiente de h con

$$|\Phi_f(t, z; h) - \Phi_f(t, y; h)| \leq M_{\Phi_f} |z - y|.$$

Dem: Esto se deduce de la condición de Lipschitz satisfecha por $f(t, y)$ de forma inductiva. \blacksquare

Ejemplo 92 La cero estabilidad para el método de Euler progresivo

$$\Phi(t, y; h) = k_1 = f(t, y)$$

se obtiene de la condición de Lipschitz para f con respecto a la segunda variable, es decir, $M_{\Phi_f} = L_f$.

Ejemplo 93 Para el método de Heun se sigue igual: veamos

$$\Phi(t, y; h) = \frac{1}{2}\{k_1 + k_2\} = \frac{1}{2}\{f(t, y) + f(t + h, y + h f(t, y))\}$$

entonces, si $|f(t, y) - f(t, z)| \leq L_f |y - z|$

$$\begin{aligned} |\Phi(t, y; h) - \Phi(t, z; h)| &\leq \frac{1}{2}|f(t, y) - f(t, z)| \\ &+ \frac{1}{2}|f(t + h, y + h f(t, y)) - f(t + h, z + h f(t, z))| \\ &\leq \frac{L_f}{2}|y - z| + \frac{L_f}{2}|y + h f(t, y) - z - h f(t, z)| \\ &\leq L_f |y - z| + \frac{h L_f^2}{2}|y - z| = (L_f + \frac{h L_f^2}{2})|y - z|. \end{aligned}$$

Entonces, para cualquier $h_0 > 0$ fijamos $M_{\Phi_f} = L_f + h_0 L_f^2 / 2$ y tenemos que

$$L_f + h L_f^2 / 2 \leq M_{\Phi_f}, \quad h \leq h_0$$

por lo que

$$|\Phi(t, y; h) - \Phi(t, z; h)| \leq (L_f + \frac{h L_f^2}{2})|y - z| \leq M_{\Phi_f} |y - z|, \quad h \leq h_0.$$

Teorema 94 Todos los métodos de Runge-Kutta explícitos son cero estables

Dem: Ponemos $w_n^{(h)} = z_n^{(h)} - y_n^{(h)}$ para $n = 0, 1, \dots, N^{(h)}$. Entonces

$$w_{n+1}^{(h)} = w_n^{(h)} + h\{\Phi_f(t_n, z_n^{(h)}; h) - \Phi_f(t_n, y_n^{(h)}; h)\} + h \delta_{n+1}^{(h)}.$$

Supongamos que $|\delta_{n+1}^{(h)}| \leq \epsilon$, entonces aplicando la propiedad de Lipschitz que cumple Φ_f tenemos ($w_n^{(h)} = w_n$ para aliviar notación)

$$|w_{n+1}| \leq |w_n| + h M_{\Phi_f} |w_n| + h \epsilon = (1 + h M_{\Phi_f})|w_n| + h \epsilon, \quad n \geq 0.$$

Una sencilla recursión nos da

$$\begin{aligned} |w_{n+1}| &\leq (1 + h M_{\Phi_f})|w_n| + h \epsilon \leq (1 + h M_{\Phi_f})^2|w_{n-1}| + (1 + h M_{\Phi_f})h \epsilon + h \epsilon \\ &\leq (1 + h M_{\Phi_f})^3|w_{n-2}| + (1 + h M_{\Phi_f})^2h \epsilon + (1 + h M_{\Phi_f})h \epsilon + h \epsilon \\ &\leq \dots \leq (1 + h M_{\Phi_f})^{n+1}|w_0| + h \epsilon \sum_{i=0}^n (1 + h M_{\Phi_f})^i \end{aligned}$$

Usando ahora que $1 + h M_{\Phi_f} \leq e^{h M_{\Phi_f}}$ y la expresión para una suma geométrica de razón $1 + h M_{\Phi_f}$ obtenemos que

$$\begin{aligned} |w_{n+1}| &\leq e^{h(n+1)M_{\Phi_f}}|w_0| + h \epsilon \frac{(1 + h M_{\Phi_f})^{n+1} - 1}{1 + h M_{\Phi_f} - 1} \\ &\leq e^{T M_{\Phi_f}}|w_0| + \epsilon \frac{e^{T M_{\Phi_f}} - 1}{M_{\Phi_f}}, \quad n = 0, 1, 2, \dots N-1. \end{aligned}$$

Si además también se tiene $|w_0| \leq \epsilon$ y ponemos $C = e^{T M_{\Phi_f}} + (e^{T M_{\Phi_f}} - 1) M_{\Phi_f}^{-1}$ entonces para $n = 0, 1, 2, \dots N-1$

$$|w_n| \leq C \epsilon$$

siendo ϵ el parámetro que acota a las perturbaciones y al error inicial. ■

Observación 98 En el caso $\delta_n = 0$ para $n = 1, 2, \dots$ sólo se tiene en cuenta el efecto provocado por el error inicial

$$|w_n^{(h)}| \leq e^{T M_{\Phi_f}} |w_0^{(h)}|, \quad n = 0, 1, 2, \dots N.$$

Observación 99 Las constantes que garantizan la estabilidad dependen de T y de M_{Φ_f} de manera exponencial, es decir, aparece un $e^{T M_{\Phi_f}}$. Por lo que pueden empeorar de manera muy importante. Este resultado cubre un gran espectro de situaciones, por lo tanto es razonable que la estimación sea muy general y tengamos las constantes $e^{T M}$ como ya vimos en el caso de los métodos simples de Euler. Estas estimaciones se podrían mejorar con información sobre las derivadas de Φ_f . Un estudio más fino usando el Teorema del Valor Medio puede dar mejores cotas de error, pero los cálculos se complican en exceso.

4.7. Convergencia de los métodos RKE

La forma general de un **método de Runge-Kutta explícito** de S etapas es

1. Dado $y_0 \approx y(t_0)$, para $0 \leq n \leq N-1$

2. Construir las pendientes k_1, k_2, \dots, k_S definidas por

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + c_2 h, y_n + h a_{21} k_1), \\ k_i &= f(t_n + c_i h, y_n + h(a_{i1} k_1 + a_{i2} k_2 + \dots + a_{ii-1} k_{i-1})), \quad i = 3, \dots, S \end{aligned}$$

siendo $c_2 = a_{21}$ y $\sum_{j=1}^{i-1} a_{ij} = c_j$, $i = 3, \dots, S$. Obtenemos entonces la pendiente promedio

$$k = \sum_{i=1}^S b_i k_i$$

siendo $b_1 + b_2 + \dots + b_S = 1$.

3. Dar el paso definitivo en esta dirección promedio generada

$$y_{n+1} = y_n + h k$$

o bien

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

siendo

$$\Phi_f(t_n, y_n; h) = \sum_{i=1}^S b_i k_i(t_n, y_n; h).$$

Tenemos entonces la misma estructura que con el método de Euler explícito, salvo que hemos reemplazado la dirección de avance $f(t_n, y_n)$ por $\Phi_f(t_n, y_n; h)$. Pero sobre $\Phi_f(t_n, y_n; h)$ tenemos una condición de Lipschitz global para $h < h_0$ y tenemos también un error local de consistencia al menos $O(h^2)$. Por lo tanto, la misma demostración nos da la convergencia con orden al menos $O(h)$. En este caso además podemos tener ordenes mayores dependiendo de la elección del número de etapas S y de los coeficientes.

A parte del argumento analítico se puede otra vez interpretar geométricamente mediante, ver Figura 4.1.

Teorema 95 *Todos los métodos de Runge-Kutta explícitos son convergentes con orden al menos uno. Además, si un método de Runge-Kutta es consistente con orden $O(h^p)$ entonces converge con orden p siempre que el error inicial también sea $O(h^p)$.*

Observación 100 *Es claro que hemos reemplazado $y'(t) = f(t, y(t))$ por la expresión*

$$\frac{y_{n+1} - y_n}{h} = \sum_{i=1}^S b_i k_i, \quad n \geq 0$$

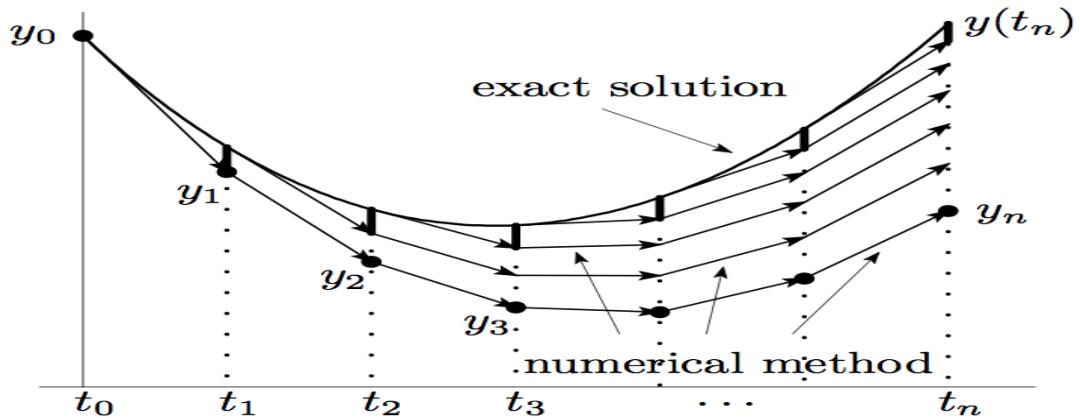


Figura 4.1: Interpretación gráfica del error global para un método de un paso cualquiera. Los trazos rectos se construyen con el método mientras que la curva continua es la solución exacta. En cada vertical t_n se ve el error local en el trazo grueso y el resto depende de aplicar el esquema en puntos originales distintos.

luego queremos

$$\boxed{\frac{y_{n+1} - y_n}{h} \approx y'(t_n)} \quad (4.1)$$

y también

$$\boxed{\sum_{i=1}^S b_i k_i \approx f(t_n, y(t_n))} \quad (4.2)$$

Observación 101 También se pide que el error inicial $|y(t_0^h) - y_0^h|$ tienda a cero con el mismo orden al menos que el método.

Observación 102 Este resultado usa $h \rightarrow 0^+$ por lo que no se preocupa de la posibilidad de que el valor h sea tan pequeño que resulte impracticable. Ya hemos hablado de esto, la reducción del parámetro h puede ser para

- conseguir mayor precisión ó
- simplemente para que el método no se desestabilice.

Una clasificación más cercana a la práctica la ofrece el concepto de **estabilidad absoluta o A-estabilidad** que ya veremos más adelante.

4.8. Familias RKE de acuerdo al orden

Recordemos que un método convergente se dice que tiene orden de convergencia p si su error local tiene orden $p + 1$.

4.8.1. Métodos RKE con $S = 1$

En el caso $S = 1$ reduce la matriz de Butcher a la situación trivial

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

que representa el método de Euler explícito. Sabemos que es de orden 1 y hay ejemplos donde este error es exactamente 1. Se puede ver fácilmente con el campo $f(t, y) = y$ que tiene solución notrivial $y(t) = e^t$.

4.8.2. Métodos RKE con $S = 2$

En este caso tenemos como matriz de Butcher

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \hline c_2 & a_{2,1} & 0 \\ \hline & b_1 & b_2 \end{array}$$

con $c_2 = a_{2,1}$ y $b_1 + b_2 = 1$. Por lo tanto tenemos

$$\begin{aligned} k(t, y; h) &= b_1 k_1 + b_2 k_2, \\ k_1 &= k_1(t, y(t)) = f(t, y(t)) = y'(t), \\ k_2 &= k_1(t, y(t); h) = f(t + c_2 h, y(t) + h c_2 k_1). \end{aligned}$$

El error local resulta ser

$$l(y(t); h) = y(t + h) - y(t) - h \{b_1 k_1(t, y(t); h) + b_2 k_2(t, y(t); h)\}$$

Podemos entonces escribirlo como

$$l(y(t); h) = y(t + h) - y(t) - h b_1 y'(t) - h b_2 k_2$$

Usando el desarrollo de Taylor para $y(t + h)$ respecto a t obtenemos

$$y(t + h) - y(t) = h y'(t) + \frac{h^2}{2} y''(t) + \frac{h^3}{3} y'''(t) + \mathcal{O}(h^4)$$

de donde

$$l(y(t); h) = h y'(t) + \frac{h^2}{2} y''(t) + \frac{h^3}{3!} y'''(t) + \mathcal{O}(h^4) - h b_1 y'(t) - h b_2 k_2.$$

Recordemos que desarrollo de Taylor de una función de dos variables se puede describir de forma simbólica como

$$f(t + \Delta t, y + \Delta y) = \sum_{n \geq 0} \frac{1}{n!} \{\Delta t \partial_t + \Delta y \partial_y\}^n f(t, y).$$

Vamos a desarrollar por Taylor $k_2 = f(t + h c_2, y(t) + h a_{21} f(t, y(t)))$ en el punto $(t, y(t))$. Abreviaremos usando $f = f(t, y(t))$, $\partial_t f = f_t$, $\partial_{tt} f = f_{tt}$, ... y también que $c_2 = a_{2,1}$. Entonces tenemos

$$\begin{aligned} k_2 &= f(t, y(t)) + h c_2 \{f_t + y'(t) f_y\} \\ &+ \frac{1}{2!} \{h^2 c_2^2 f_{tt} + 2h c_2 h a_{21} f f_{ty} + h^2 a_{21}^2 f^2 f_{yy}\} + \mathcal{O}(h^3) \\ &= y'(t) + h c_2 y''(t) + \frac{h^2 c_2^2}{2} \{f_{tt} + 2f f_{ty} + f^2 f_{yy}\} + \mathcal{O}(h^3) \end{aligned}$$

usando que $y'(t) = f(t, y(t))$ y entonces $y''(t) = f_t + f f_y$. Por lo tanto, tenemos

$$\begin{aligned} l(y(t); h) &= h y'(t) + \frac{h^2}{2} y''(t) + \frac{h^3}{3!} y'''(t) + \mathcal{O}(h^4) \\ &- b_1 h y'(t) - h b_2 y'(t) - h^2 b_2 c_2 y''(t) \\ &- \frac{h^3 b_2 c_2^2}{2} \{f_{tt} + 2f f_{ty} + f^2 f_{yy}\} + \mathcal{O}(h^4) \end{aligned}$$

usando que $b_1 + b_2 = 1$ el término $h y'(t)$ se anula y nos queda

$$l(y(t); h) = \left(\frac{1}{2} - b_2 c_2\right) y''(t) h^2 + \left[\frac{1}{6} y'''(t) - \frac{b_2 c_2^2}{2} \{f_{tt} + 2f f_{ty} + f^2 f_{yy}\}\right] h^3 + \mathcal{O}(h^4).$$

Si forzamos $c_2 b_2 = 1/2$ el término $h y''(t)$ también se anula nos queda

$$l(y(t); h) = h^3 \left[\frac{1}{6} y'''(t) - \frac{c_2}{4} \{f_{tt} + 2f f_{ty} + f^2 f_{yy}\} \right] + \mathcal{O}(h^4) \quad h \rightarrow 0.$$

luego tenemos para toda $y(t)$ solución suficientemente regular del problema $u'(t) = f(t, u(t))$ que

$$l(y(t); h) = \mathcal{O}(h^3) \quad h \rightarrow 0.$$

Observación 103 ¿Podemos mejorar esto para cualquier solución de cualquier problema de Cauchy? Es decir, ¿Podemos tener $l(h) = \mathcal{O}(h^4) < \mathcal{O}(h^3)$ para cualquier solución de cualquier problema de Cauchy? La respuesta es que no. Es decir, el coeficiente de la potencia h^3 no se anula en general. Sólo necesitamos encontrar una solución de un problema de Cauchy donde no se pueda conseguir esto. Es decir, sólo hay que encontrar un ejemplo donde

$$\frac{1}{6} y'''(t) - \frac{c_2}{4} \{f_{tt} + 2f f_{ty} + f^2 f_{yy}\} \neq 0$$

y tendremos que no se puede mejorar a $O(h^4)$, es decir, el error local en este caso se mantiene proporcional a $O(h^3)$.

Los casos triviales son normalmente suficientes para aclarar estas situaciones. Por ejemplo, tomamos $f(t, y) = y$ de donde la solución es $y(t) = e^t$ y entonces $y'''(t) = e^t$. Como $f(t, y) = y$ tenemos que $f_y = 1$ y $f_{tt} + 2f f_{ty} + f^2 f_{yy} = 0$ entonces

$$l(y(t); h) = \frac{1}{6}e^t h^3 + \mathcal{O}(h^4) = \mathcal{O}(h^3), \quad h \rightarrow 0.$$

Otra vez se puede observar que si se usa la solución exacta $y(t) = -B/A$ del problema con $f(t, y) = A y + B$ entonces tenemos un error cero, es decir, si empezamos en el valor exacto seguimos en él. Pero esto no contradice el hecho de que $l(y(t); h) \leq Ch^3$ para toda solución del problema de Cauchy planteado con cualquier función $f(t, y)$. Lo que vemos con estos ejemplos son un caso donde la estimación de error se alcanza de forma exacta, es decir, $l(y(t); h) = Ch^3$ y otro donde el error es cero, es decir, $l(y(t); h) = 0 \leq Ch^3$. Pero lo que hemos determinado es que al existir una solución donde el error es $l(y(t); h) = Ch^3$ ya podemos descartar con total seguridad que podamos tener un orden mejor, esto es, $l(y(t); h) \leq O(h^p) < O(h^3)$ con $p > 3$ para toda solución del problema de Cauchy planteado con cualquier función $f(t, y)$ usando este esquema numérico.

Luego el mayor orden posible con $S = 2$ del error local es 3, lo que nos da un error global 2.

Para cualquier solución lo suficientemente regular se garantiza que el error global de este método decréce como $O(h^2)$ cuando $h \rightarrow 0$ y existe al menos una solución donde el decaimiento del error no es más rápido que $O(h^2)$ cuando $h \rightarrow 0$. Por lo tanto el orden de convergencia 2 no es mejorable.

Nos encontramos entonces con una familia de métodos de Runge-Kutta explícitos (RKE) de orden 2 caracterizada por las restricciones sobre los parámetros

$$b_1 + b_2 = 1, \quad c_2 b_2 = 1/2, \quad (c_2 = a_{2,1}).$$

Evidentemente, como tomamos siempre la restricción

$$b_1 + b_2 = 1$$

entonces

$$c_2 b_2 = 1/2 \Rightarrow \text{orden 2}$$

$$c_2 b_2 \neq 1/2 \Rightarrow \text{orden 1}$$

Si usamos $\gamma = c_2$ (sabemos que $0 < \gamma \leq 1$ por ser γh la longitud que avanzamos dentro de $[t_n, t_{n+1}]$) encontramos que básicamente sólo hay un parámetro y el resto está en función de γ . Por ejemplo,

$$b_2 = 1/(2\gamma), \quad b_1 = 1 - 1/(2\gamma), \quad a_{2,1} = \gamma$$

Por lo tanto, aparece **una familia** infinita de métodos de Runge-Kutta explícitos de orden 2 que depende de un sólo parámetro. El tablero de Butcher es

0	0	0
γ	γ	0
	$1 - 1/(2\gamma)$	$1/(2\gamma)$

Ejemplo 96 *El método de Heun de segundo orden coincide con el caso $\gamma = 1$*

dado y_0

$$\begin{aligned}\tilde{y}_{n+1} &= y_n + h f(t_n, y_n), \\ y_{n+1} &= y_n + \frac{h}{2} \left(f(t_n, y_n) + f(t_{n+1}, \tilde{y}_{n+1}) \right), \quad n = 0, 1, \dots, N-1.\end{aligned}$$

También se puede escribir como

dado y_0

$$y_{n+1} = y_n + h \left\{ \frac{1}{2} k_1 + \frac{1}{2} k_2 \right\}, \quad n = 0, 1, \dots, N-1,$$

donde

$$k_1 = f(t_n, y_n), \quad k_2 = f(t_n + h, y_n + h k_1)$$

y el tablero de Butcher es

0	0	0
1	1	0
	$1/2$	$1/2$

Ejemplo 97 *Si $\gamma = 1/2$ tenemos la versión explícita de la regla del punto medio*

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

y sí que tenemos orden 2 ya que $b_2 = 1$ y $b_1 = 0$ y el tablero de Butcher

0	0	0
$1/2$	$1/2$	0
	0	1

Ejemplo 98 *Vemos aquí un método de dos etapas y con orden 1 ya que no cumple las restricciones. Es una aproximación al método de Euler implícito dada por*

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

que se entiende mejor si se escribe como

$$k_1 = f(t_n, y_n), \quad k_2 = y_n + h k_1$$

y después

$$y_{n+1} = y_n + h k_2$$

luego se puede ver como querer resolver el problema no lineal del esquema de Euler implícito mediante una iteración de punto fijo y realizar sólo una iteración. El tablero de Butcher es

0	0	0
1	1	0
	0	1

Ejemplo 99 En general, si $\gamma = 1$ y $\theta \in (0, 1)$ tenemos la familia de esquemas

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

con tablero de Butcher

0	0	0
1	1	0
	θ	1 - θ

que se pueden mirar como una linealización del caso más general e implícito

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

Aquí tenemos $c = 1$ y $b_1 = \theta$ luego y sólo tenemos orden dos cuando $\theta = 1/2$, en el resto de casos es orden 1.

4.8.3. Métodos RKE con $S = 3$

Para el tablero de Butcher

0	0	0	0
c_2	a_{21}	0	0
c_3	a_{31}	a_{32}	0
	b_1	b_2	b_3

usando las mismas ideas las restricciones que se obtienen sobre los parámetros son

$$\begin{aligned} b_1 + b_2 + b_3 &= 1, \\ b_2 c_2 + b_3 c_3 &= 1/2, \\ b_2 c_2^2 + b_3 c_3^2 &= 1/3, \\ b_3 c_2 a_{3,2} &= 1/6. \end{aligned}$$

Tenemos cuatro ecuaciones y seis incógnitas, por lo tanto aparecen **dos familias** infinitas de métodos de RK explícitos de orden 3. Se puede observar también por los desarrollos de Taylor que este es el mayor orden posible con $S = 3$.

Ejemplo 100 Método de Heun de tercer orden

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f\left(t_n + \frac{1}{3}h, y_n + \frac{1}{3}hk_1\right) \\ k_3 &= f\left(t_n + \frac{2}{3}h, y_n + \frac{2}{3}hk_2\right) \\ y_{n+1} &= y_n + h\left(\frac{1}{4}k_1 + \frac{3}{4}k_3\right). \end{aligned}$$

Ejemplo 101 En el siguiente método se observa la posibilidad de usar coeficientes negativos. Se pueden interpretar como usar $-k_1 + 2k_2$ como promedio entre k_1 y k_2 para obtener $k_3 = -k_1 + 2k_2$.

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1\right) \\ k_3 &= f\left(t_n + h, y_n - h k_1 + 2 h k_2\right) \\ y_{n+1} &= y_n + h\left(\frac{1}{6}k_1 + \frac{4}{6}k_2 + \frac{1}{6}k_3\right) \end{aligned}$$

la matriz de Butcher en este caso es

0	0	0	0
1/2	1/2	0	0
1	-1	2	0
<hr/>			
1/6 4/6 1/6			

Ejemplo 102 Un último ejemplo lo constituye el **método de Nystrom**, cuya matriz de Butcher es

0	0	0	0
2/3	2/3	0	0
2/3	0	2/3	0
<hr/>			
1/4 3/8 3/8			

y a partir de aquí el método se escribe como

$$\begin{aligned} y_{n+1} &= y_n + h\left(\frac{1}{4}k_1 + \frac{3}{8}k_2 + \frac{3}{8}k_3\right) \\ k_1 &= f(t_n, y_n) \\ k_2 &= f\left(t_n + \frac{2}{3}h, y_n + \frac{2}{3}hk_1\right) \\ k_3 &= f\left(t_n + \frac{2}{3}h, y_n + \frac{2}{3}h k_2\right) \end{aligned}$$

Ejemplo 103 Para $y'(t) = f(t, y(t))$ usamos la notación

$$f = f(t, y(t)), \quad f_y = f_y(t, y(t)), \quad f_t = f_t(t, y(t)), \quad f_{ty} = f_{ty}(t, y(t)), \quad \dots$$

y definimos

$$F = f_t + f f_y, \quad G = f_{tt} + 2 f f_{ty} + f^2 f_{yy}.$$

Comprobar que:

1. $y'' = f_t + f f_y = F, \quad y''' = F f_y + G$
2. $y(t+h) = y(t) + fh + F \frac{h^2}{2} + [F f_y + G] \frac{h^3}{6} + O(h^4)$

Si se considera ahora la forma general de un método de Runge-Kutta explícito para $S \leq 3$ etapas (si $S = 2$ entonces $c_3 = a_{31} = a_{32} = b_3 = 0$)

0	0	0	0
c_2	a_{21}	0	0
c_3	a_{31}	a_{32}	0
	b_1	b_2	b_3

con $b_1 + b_2 + b_3 = 1$, $c_2 = a_{21}$, $c_3 = a_{31} + a_{32}$. Comprobar que:

1. $k_2 = f + a_{21}hF + \frac{1}{2}c^2h^2G + O(h^3)$
2. $h(a_{31}k_1 + a_{32}k_2) = c_3hf + a_{32}a_{21}h^2F + \frac{1}{2}a_{32}c_2^2h^3G + O(h^4)$
3. $k_3 = f + c_3hF + h^2[a_{32}a_{21}F f_y + \frac{1}{2}c_3^2G] + O(h^3)$
4. Finalmente

$$b_1 k_1 + b_2 k_2 + b_3 k_3 = f + (a_{21}b_2 + c_3b_3)hF + \frac{1}{2}h^2[2a_{32}a_{21}b_3F f_y + (b_3c_3^2 + b_2c_2^2)G] + O(h^3)$$

Entonces

$$y(t) + h \sum_{i=1}^3 b_i k_i(t, y(t)) = y(t) + hf + (a_{21}b_2 + c_3b_3)h^2F + \frac{1}{2}h^3[2a_{32}a_{21}b_3F f_y + (b_3c_3^2 + b_2c_2^2)G] + O(h^4)$$

Comparando con

$$y(t+h) = y(t) + fh + F \frac{h^2}{2} + [F f_y + G] \frac{h^3}{6} + O(h^4)$$

Concluir que para obtener orden 2 es necesario

$$c_2 b_2 + c_3 b_3 = \frac{1}{2}$$

y que para obtener orden 3 necesitamos además tener simultáneamente

$$c_2^2 b_2 + c_3^2 b_3 = \frac{1}{3}, \quad a_{32} a_{21} b_3 = \frac{1}{6}$$

Construir el método de Runge-Kutta de tres etapas que tiene la siguiente matriz de Butcher

0	0	0	0
1	1	0	0
1	1/2	1/2	0
	3/6	1/6	2/6

Comprobar que el método es de orden 2 y no es 3 en general. Comprobar que si que es de orden 3 cuando se aplica al problema $y'(t) = y(t)$ con $y(0) = 1$. Explicar esta aparente contradicción. Dar ejemplos teóricos donde el orden 2 y el orden 3 se alcancen.

4.8.4. Métodos RKE con $S = 4$

Ya vemos que la técnica básica estandard para derivar estos métodos consiste en hacer desarrollos de Taylor y forzar el mayor numero de coeficientes de potencias de h cero en el desarrollo de Taylor de $y(t+h)$ en torno a t . Existen condiciones similares para obtener métodos de cuarto orden pero no las vamos a escribir.

El ejemplo clásico de Runge-Kutta explícito de orden 4 con 4 etapas es

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 &= f(t_n, y_n) \\ k_2 &= f\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1\right) \\ k_3 &= f\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_2\right) \\ k_4 &= f(t_n + h, y_n + hk_3) \end{aligned}$$

y la matriz de Butcher en este caso es

0	0	0	0	0
1/2	1/2	0	0	0
1/2	0	1/2	0	0
1	0	0	1	0
	1/6	2/6	2/6	1/6

Tener coeficientes sencillos y alto orden fue decisivo para su extensión como el más usado en la era pre-computacional, esto es, cuando había que hacer los cómputos a mano.

4.8.5. Resumen

Hemos obtenido

- si $S = 1$ existe un sólo método explícito de orden 1
- si $S = 2$ existe una familia infinita que depende de un parámetro de métodos explícitos de orden 2
- si $S = 3$ existe una familia infinita que depende de dos parámetros de métodos explícitos de orden 3
- si $S = 4$ existe una familia infinita que depende de dos parámetros de métodos explícitos de orden 4

Observación 104 *Varias puntuaciones son interesantes aquí*

- *El resultado para $S = 4$ indica algún tipo de anomalía en la generalización del estudio del orden ya que el número de parámetros sigue siendo dos, como en el caso $S = 3$.*
- *Los términos de error ya no son simples sino que incluyen expresiones complicadas con derivadas de orden alto de la función f .*
- *Se hace uso de los desarrollos de Taylor de forma intensiva. Estos desarrollos cambian cuando se aplica a funciones vectoriales, por ejemplo, el Teorema del Valor Medio ya cambia en el caso vectorial. Esto indica que la extensión al caso vectorial no va a ser directa.*

Estudio de métodos RK explícitos de alto orden

Hemos construido métodos de S etapas con orden S para $S = 1, 2, 3, 4$. Es natural preguntarse si esta pauta se mantiene para $S \geq 5$. La respuesta a esta pregunta se debe a John Butcher en torno a 1960 y es negativa [5]. El siguiente resultado presenta una relación entre el número de etapas de un método y su orden:

Teorema 104 *Un método de Runge-Kutta explícito de S etapas no puede tener un orden mayor que S . No existe un método de Runge-Kutta explícito de S etapas que tenga orden S para $S \geq 5$.*

En particular, para **ordenes** entre 1 y 10 el **mínimo** número de **etapas** requeridas, S_{min} , para obtener un método con ese mismo orden se muestra en la siguiente tabla

orden	1	2	3	4	5	6	7	8
S_{min}	1	2	3	4	6	7	9	11

entonces se puede observar que **4 es el número máximo para el que el orden coincide con las etapas**. Para más de 4 etapas ya el orden es más pequeño que el número de etapas. Por lo tanto, es como decir que el esfuerzo en la construcción del método se ve recompensado si $S \leq 4$ pero ya no para $S \geq 5$. Esta es una respuesta más a la popularidad de los métodos de cuatro etapas y de entre ellos el clásico cuyos coeficientes son más fáciles.

4.9. Extensión a sistemas

Desde el punto de vista práctico la extensión de los esquemas de Runge-Kutta explícitos a sistemas es un aspecto trivial consistente en **repetir las ecuaciones para cada incógnita del sistema** de manera ordenada. Por ejemplo, si tenemos un sistema de dos ecuaciones

$$\begin{aligned} x'(t) &= f(t, x(t), y(t)) \\ y'(t) &= g(t, x(t), y(t)) \end{aligned}$$

y queremos aplicar el método clásico de Runge-Kutta explícito de orden 4 con 4 etapas dado por

$$\begin{aligned} w_{n+1} &= w_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 &= f(t_n, w_n) \\ k_2 &= f(t_n + \frac{1}{2}h, w_n + \frac{1}{2}hk_1) \\ k_3 &= f(t_n + \frac{1}{2}h, w_n + \frac{1}{2}hk_2) \\ k_4 &= f(t_n + h, w_n + hk_3) \end{aligned}$$

entonces debemos de calcular como sigue:

$$\begin{aligned} kx_1 &= f(t_n, x_n, y_n) \\ ky_1 &= g(t_n, x_n, y_n) \\ kx_2 &= f(t_n + \frac{1}{2}h, x_n + \frac{1}{2}hkx_1, y_n + \frac{1}{2}hky_1) \\ ky_2 &= g(t_n + \frac{1}{2}h, x_n + \frac{1}{2}hkx_1, y_n + \frac{1}{2}hky_1) \\ kx_3 &= f(t_n + 0.5h, x_n + 0.5hkx_2, y_n + 0.5hky_2) \\ ky_3 &= g(t_n + 0.5h, x_n + 0.5hkx_2, y_n + 0.5hky_2) \\ kx_4 &= f(t_n + h, x_n + hkx_3, y_n + hky_3) \\ ky_4 &= g(t_n + h, x_n + hkx_3, y_n + hky_3) \end{aligned}$$

y concluir con

$$\begin{aligned}x_{n+1} &= x_n + \frac{h}{6}(kx_1 + 2kx_2 + 2kx_3 + kx_4) \\y_{n+1} &= y_n + \frac{h}{6}(ky_1 + 2ky_2 + 2ky_3 + ky_4).\end{aligned}$$

Observación 105 Obviamente, usar $1/2$ ó 0.5 en la fórmula es simplemente una cuestión meramente estética. Una cosa distinta es lo que pueda ocurrir en un lenguaje de programación donde se distinga el tipo entero del real. En este caso $1/2=0$ y si que tendríamos un error por lo que se debe escribir 0.5 en vez de $1/2$.

La notación compacta vectorial mimetiza exactamente el caso de una ecuación. Ponemos

$$X = (x, y)^*, \quad F(t, X) = (f(t, x, y), g(t, x, y))^* = (f(t, X), g(t, X))^*$$

donde $*$ indica la trasposición. Entonces el sistema se escribe como

$$X'(t) = F(t, X(t))$$

y el esquema numérico como

$$\begin{aligned}X_{n+1} &= X_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4) \\K_1 &= F(t_n, X_n) \\K_2 &= F\left(t_n + \frac{1}{2}h, X_n + \frac{1}{2}hK_1\right) \\K_3 &= F\left(t_n + \frac{1}{2}h, X_n + \frac{1}{2}hK_2\right) \\K_4 &= F(t_n + h, X_n + hK_3)\end{aligned}$$

con la notación obvia para $K_j = (kx_j, ky_j)^*$.

4.9.1. Sobre el orden de convergencia

En la obtención de estos métodos siempre se supuso además que el problema era escalar y que nada importante ocurre cuando pasamos al caso de sistemas. Pero en este caso, los desarrollos de Taylor nos llevan al Jacobiano y a otros entes con derivadas parciales de mayor orden. La dificultad intrínseca en estos objetos matemáticos fue simplificada gracias a la introducción de la Teoría de árboles por parte de John C. Butcher en torno a 1960, ver en su libro [5] por ejemplo. Esta teoría relaciona la forma en la que se van construyendo las sucesivas derivadas con aspectos de la teoría de grafos que simplifican el estudio.

Se esperaba que cuando se obtuviese un método de orden p para un problema escalar, este orden se mantuviese en el caso vectorial. Pero de la Teoría de Butcher se deduce que no es así, siendo este un resultado que sorprendió a la comunidad científica en este campo cuando se dio a conocer.

Un método de RK de orden p en el caso escalar $m = 1$ puede tener orden menor que p en el caso de sistemas $m > 1$.

También hay que mirar con cuidado las ecuaciones autónomas escalares. Vamos a recopilar una serie de resultados sin demostración:

Teorema 105 *Todo método de RK de orden $p = 1, 2, 3, 4$ en el caso $m = 1$ se extiende a un método del mismo orden en el caso de sistemas, $m > 1$.*

Teorema 106 *Consideremos las siguientes hipótesis*

- (A) *el método RK tiene orden p para $y'(t) = f(t, y(t))$ con $m > 1$.*
- (B) *el método RK tiene orden p para $y'(t) = f(t, y(t))$ con $m = 1$.*
- (C) *el método RK tiene orden p para $y'(t) = f(y(t))$ con $m = 1$.*

Tenemos entonces

- *Si $1 \leq p \leq 3$ entonces (A) \iff (B) \iff (C)*
- *Si $p = 4$ entonces (A) \iff (B) \Rightarrow (C) pero (C) $\not\Rightarrow$ B.*
- *Si $p \geq 5$ entonces (A) \Rightarrow (B) \Rightarrow (C) pero (C) $\not\Rightarrow$ (B) y (B) $\not\Rightarrow$ (A).*

Ejemplo 107 *El siguiente método de Runge-Kutta de cuatro etapas*

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + \frac{1}{2}h, y_n + \frac{1}{2}h k_1), \\ k_3 &= f(t_n - h, y_n + \frac{1}{2}h k_1 - \frac{3}{2}h k_2), \\ k_4 &= f(t_n + h, y_n + \frac{4}{3}h k_2 - \frac{1}{3}h k_3), \\ y_{n+1} &= y_n + \frac{h}{6} (k_1 + 4k_2 + k_4), \end{aligned}$$

es de orden 3, sus coeficientes no cumplen las restricciones adecuadas para obtener orden máximo correspondiente a cuatro etapas, pero tiene orden 4 cuando se aplica a problemas autónomos. Es decir, $p = 4$ en el caso (C) pero (C) $\not\Rightarrow$ B. Este efecto se puede ver sobre los problemas:

1. para $t \in [0, 3]$ y con dato inicial $y(0) = 1$ los dos problemas escalares

$$(I) \quad y'(t) = \sqrt{y(t)}, \quad (II) \quad y'(t) = \frac{y(t)}{1 + 0.5t}$$

tienen la misma solución $y(t) = (1 + 0.5t)^2$. Cuando se aplica al problema (II) resulta ser de orden 3, pero si se aplica al problema (I) es de orden 4.

2. También con dato inicial $y(0) = 1$ sobre los dos problemas escalares

$$(I) \quad y'(t) = y(t), \quad (II) \quad y'(t) = t y(t)$$

se presenta orden 4 en (I) y orden 3 en (II).

Ejemplo 108 $y(t) = (t + 1)^2$ es solución de

$$y'(t) = 2(t + 1), \quad y(0) = 1$$

y de

$$y'(t) = \frac{2y(t)}{t + 1}, \quad y(0) = 1.$$

Se puede ver que el método de Heun es exacto cuando se aplica a la primera ecuación pero no lo es cuando se aplica a la segunda.

4.10. Estabilidad absoluta: A-estabilidad

El conjunto de las curvas solución de un campo continuo tiene comportamientos localmente distintos. Para simplificar el estudio nos fijamos en los puntos interesantes a estudiar que son los puntos de equilibrio. Para ello linealizamos el campo usando el desarrollo de Taylor de primer orden. Reemplazamos entonces nuestro campo original por el aproximado y observamos como se comporta el esquema numérico en el problema aproximado. Este análisis se hace con relativa facilidad y es suficiente para empezar a ver las propiedades de estabilidad numérica de un esquema de cálculo, es decir, nos sirve para clasificar la estabilidad del mismo.

Supongamos tenemos el campo autónomo dado por $f(y)$ usando desarrollo de Taylor cerca de un punto de equilibrio $z_* = 0$ (para simplificar)

$$f(y) \sim -\lambda y$$

donde $\lambda = f_y(0)$. Entonces, localmente, podemos trabajar con el **problema modelo lineal**

$$y'(t) = -\lambda y(t), \quad t > 0, \quad y(0) = 1, \quad \lambda \gg 1.$$

El concepto de estabilidad numérica es importante sólo en el caso donde hay decaimiento de la solución o de parte de ella y principalmente cuando este decaimiento es brusco (problemas rígidos), o mejor dicho, cuando la EDO es muy fuertemente estable. Pero esta situación es lo suficientemente importante ya que genera una restricción en el paso de avance h que de no cumplirla el esquema puede ser inestable desde el principio o más adelante.

Observación 106 *Por lo dicho anteriormente, el caso $y'(t) = \lambda y(t)$ con $\lambda \gg 1$ es de por sí inestable y la estabilidad del esquema numérico no es relevante.*

Por lo tanto, nos vamos a fijar en el caso $\lambda > 0$. Aquí sabemos que la solución exacta es $y(t) = e^{-\lambda t}$ y decae exponencialmente para $\lambda > 0$: se dice que **es una edo estable**. En el caso $\lambda < 0$ la solución crece exponencialmente: se dice que **es una edo inestable**.

Cualquier método numérico para aproximar las soluciones de un campo de vectores debe generar aproximaciones cualitativamente similares.

Los esquemas numéricos se clasifican no sólo por su orden de convergencia sino también sus propiedades de estabilidad.

Recordemos que la **0-estabilidad**, al ser un concepto que, cuando se satisface, sólo necesita $h \rightarrow 0$, no se preocupa de lo que ocurre para un valor de h fijo. La **0-estabilidad** nos sirve para poder comprobar la convergencia en general del esquema, pero no para ver su comportamiento para un valor de h fijo.

La **A-estabilidad, o estabilidad absoluta**, indica la restricción necesaria para reproducir un comportamiento cualitativo de la solución. Este concepto nos permite clasificar los métodos convergentes determinando el intervalo de pasos de tiempo h donde el método no se desestabiliza y reproduce cualitativamente una aproximación numérica razonable para el problema modelo $y' = -\lambda y$. Usamos

$$\begin{aligned} y'(t) &= -\lambda y(t), \quad t > 0, \quad \lambda \gg 1 \\ y(0) &= 1 \end{aligned}$$

y se observa cual es la restricción que el esquema numérico impone sobre h para poder reproducir el comportamiento $|y(t)| \rightarrow 0$ de la solución continua. Es decir, queremos asegurarnos de que si la solución continua decae a cero cuando t crece también lo hace la discreta.

Ejemplo 109 *Para Euler explícito tenemos*

$$y_n = (1 - h\lambda)^n, \quad n = 0, 1, 2, \dots, N = T/h$$

luego necesitamos

$$0 < h < \frac{2}{\lambda}$$

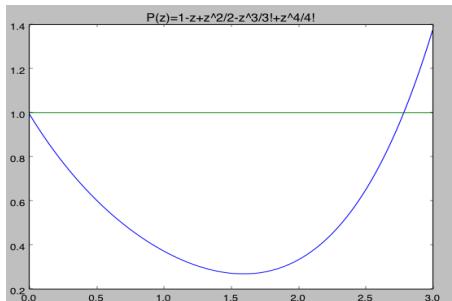


Figura 4.2: Polinomio $p(z) = 1 - z + z^2/2 - z^3/6 + z^4/24$ y su restricción para garantizar $0 < p(z) < 1$.

que garantiza $-1 < 1 - h\lambda < 1$ para tener que los valores de $|y_n|$ decaen y con $0 < h < 1/\lambda$ obtenemos decaimiento sin oscilaciones.

Otra lectura de esta restricción es: dado un intervalo de cálculo $[0, T]$ necesitamos un número de puntos N suficientemente grande para una partición uniforme tal que $N > \frac{T\lambda}{2}$, es decir,

$$0 < \frac{T}{N} < 2/\lambda \Leftrightarrow N > \frac{T\lambda}{2}.$$

Definición 110 Diremos que el **esquema es incondicionalmente A-estable (incondicionalmente absolutamente estable)**, cuando al aplicarlo al problema modelo

$$\begin{aligned} y'(t) &= -\lambda y(t), \quad 0 < t < T, \\ y(0) &= 1, \end{aligned}$$

para cualquier $\lambda > 0$ y en cualquier intervalo de tiempo $T > 0$, la solución discreta generada decae en valor absoluto. En caso contrario, el **esquema es condicionalmente A-estable**, a la restricción sobre h se le llama **restricción de estabilidad absoluta**. Al intervalo donde se cumple **intervalo de estabilidad absoluta**.

Ejemplo 111 La aplicación del **método de Runge-Kutta de orden 4 clásico**

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1) \\ k_3 &= f(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_2) \\ k_4 &= f(t_n + h, y_n + hk_3) \end{aligned}$$

al problema test ($\lambda > 0$)

$$\begin{cases} \frac{d}{dt}y(t) = -\lambda y(t), & 0 < t < T, \\ y(0) = 1, \end{cases}$$

genera, siguiendo la misma idea que con el método de Heun de orden 2,

$$y_{n+1} = \left(1 - h\lambda + \frac{h^2}{2}\lambda^2 - \frac{h^3}{3!}\lambda^3 + \frac{h^4}{4!}\lambda^4\right)^{n+1}.$$

Conseguir que sea

$$\left|1 - h\lambda + \frac{h^2}{2}\lambda^2 - \frac{h^3}{3!}\lambda^3 + \frac{h^4}{4!}\lambda^4\right| < 1$$

nos lleva a observar el polinomio $p(z) = 1 - z + z^2/2 - z^3/6 + z^4/24$ para $z = h\lambda$. Igual que con Heun, vemos, Figura 4.2, que $0 < p(z) < 1$ sólo si $z < z_* = 2.785\dots$ aproximadamente. Luego también aquí, a pesar de haber aumentado el orden, tenemos una restricción de estabilidad que cumplir:

$$h\lambda < 2.785\dots, \Rightarrow h < 2.785\dots/\lambda.$$

Hemos visto que Euler explícito es de primer orden y Euler implícito también lo es. Para ecuaciones no lineales el método de Euler es fácil de aplicar pero no así el método de Euler implícito. ¿Cuál es su ventaja entonces? Pues sabemos ya que **Euler implícito es A-estable** mientras que **Euler explícito es condicionalmente A-estable**.

Ya hemos visto que el caso realmente importante y con origen en la linealización de sistemas no lineales lleva al estudio general de modos de la forma $y(t) = \lambda y(t)$ con $\lambda \in \mathbb{C}$. Repetir los argumentos anteriores pero ahora usando $\lambda \in \mathbb{C}$ nos lleva a las siguientes regiones de A-estabilidad

4.10.1. Efecto de la linealización

Ejemplo 112 Equilibrando esfuerzos, podemos linealizar Crank-Nicolson y obtener una aproximación al valor y_{n+1} usando Euler explícito, esto es:

dado y_0

$$\begin{aligned} \tilde{y}_{n+1} &= y_n + hf(t_n, y_n), \\ y_{n+1} &= y_n + \frac{h}{2} \left(f(t_n, y_n) + f(t_{n+1}, \tilde{y}_{n+1}) \right), \quad n = 0, 1, \dots, N-1, \end{aligned}$$

este método se conoce como **método de Heun** y pertenece a la familia de los métodos de Runge-Kutta explícitos de dos etapas. Se puede ver que su error local es $O(h^3)$ haciendo desarrollos de Taylor. La diferencia con Crank-Nicolson es que

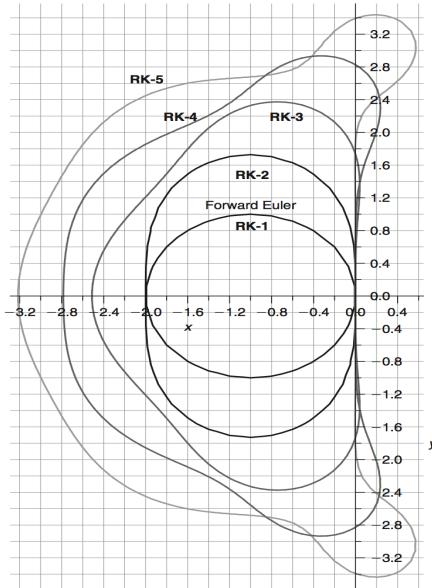


Figura 4.3: Regiones de A-estabilidad para distintos métodos de Runge-Kutta.

ahora sí que necesitamos una restricción para h en el caso $y' = -\lambda y$. Básicamente, esto es lo que hemos perdido al linealizar el problema y hacerlo explícito. El método queda como

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + h, y_n + hk_1) \\ y_{n+1} &= y_n + \frac{h}{2}(k_1 + k_2) \end{aligned}$$

y su aplicación al problema test ($\lambda > 0$)

$$\begin{cases} \frac{d}{dt}y(t) = -\lambda y(t), & 0 < t < T, \\ y(0) = 1, \end{cases}$$

genera

$$\begin{aligned} k_1 &= -\lambda y_n \\ k_2 &= -\lambda(y_n + h(-\lambda y_n)) = -\lambda y_n + h\lambda^2 y_n = y_n(-\lambda + h\lambda^2) \end{aligned}$$

y como

$$y_{n+1} = y_n + \frac{h}{2}(k_1 + k_2)$$

entonces

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

de donde

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

Conseguir que sea

$$\left|1 - h\lambda + \frac{h^2}{2}\lambda^2\right| < 1$$

nos lleva a observar el polinomio $p(z) = 1 - z + z^2/2$ que tiene el mínimo en $z = 1$ con valor $p(1) = 1/2$ y cumple $p(0) = p(2) = 1$. Por lo tanto, no tiene ceros en la recta real y $0 < p(z) < 1$ sólo se obtiene si $0 < z < 2$. Luego la restricción de estabilidad para Heun es

$$h\lambda < 2 \Leftrightarrow h < 2/\lambda.$$

Además, tenemos

$$1/2 < 1 - h\lambda + \frac{h^2}{2}\lambda^2 < 1, \quad \forall h < 2/\lambda.$$

Por lo tanto, la consecuencia de eliminar la propiedad implícita del esquema genera restricciones sobre h para trabajar en problemas disipativos. Luego tiene peores propiedades de estabilidad que Crank-Nicolson, aunque se mantiene el orden.

Ejemplo 113 La misma idea podemos aplicarla al esquema que resulta de linealizar Euler implícito usando aproximación al valor y_{n+1} usando Euler explícito. Esto es

dado y_0

$$\begin{aligned} \tilde{y}_{n+1} &= y_n + hf(t_n, y_n), \\ y_{n+1} &= y_n + hf(t_{n+1}, \tilde{y}_{n+1}), \quad n = 0, 1, \dots, N-1, \end{aligned}$$

este método no tiene adjudicado un nombre por no ser especialmente interesante, pero nos sirve para volver a ilustrar este concepto.

Se puede ver que su error local es $O(h^2)$ haciendo desarrollos de Taylor, por lo que su error global es de primer orden, y que ahora sí que necesita una restricción para h en el caso $y' = -\lambda y$:

En formato Runge-Kuta queda como

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + h, y_n + hk_1) \\ y_{n+1} &= y_n + h k_2 \end{aligned}$$

La aplicación al problema test ($\lambda > 0$)

$$\begin{cases} \frac{dy}{dt}(t) = -\lambda y(t), & 0 < t < T, \\ y(0) = 1, \end{cases}$$

genera

$$\begin{aligned} k_1 &= -\lambda y_n \\ k_2 &= -\lambda(y_n + h(-\lambda y_n)) = -\lambda y_n + h\lambda^2 y_n = y_n(-\lambda + h\lambda^2) \end{aligned}$$

y como

$$y_{n+1} = y_n + h k_2$$

entonces

$$y_{n+1} = y_n + h y_n(-\lambda + h\lambda^2)$$

de donde

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

Conseguir que sea

$$|1 - h\lambda + h^2\lambda^2| < 1$$

nos lleva a observar el polinomio $p(z) = 1 - z + z^2$ que tiene el mínimo en $z = 1/2$ con valor $p(1) = 3/4$ y cumple $p(0) = p(1) = 1$. Por lo tanto, no tiene ceros en la recta real y $0 < p(z) < 1$ sólo se obtiene si $0 < z < 1$. Luego es $h\lambda < 1$ y $h < 1/\lambda$ la restricción de estabilidad absoluta. Además, tenemos

$$3/4 < 1 - h\lambda + h^2\lambda^2 < 1, \quad \forall h < 1/\lambda.$$

Otra vez, la consecuencia de eliminar el cálculo implícito del problema genera restricciones sobre h para trabajar en problemas disipativos y, por lo tanto, tiene peores propiedades de estabilidad absoluta que Euler implícito.

4.10.2. Factor de amplificación de un esquema

Hemos visto que para el problema lineal $u' = -\lambda u$ la solución de varios esquemas se puede expresar como

$$u_{n+1} = G(\lambda h) u_n.$$

A la expresión $G(z) = G(\lambda h)$ se llama el **factor de amplificación** y la solución en el paso final será

$$u_N = G(\lambda h)^N u_0.$$

Por lo tanto, para tener un comportamiento de los valores u_n que decaigan como hace la solución continua $e^{-\lambda t}$ necesitamos

$$|G(\lambda h)| < 1$$

y esta es la condición para tener estabilidad numérica: queremos tener un **factor de amplificación** tal que se reproduzca la dinámica del campo de soluciones del problema modelo, esto es, si las soluciones decaen se debe tener un factor de amplificación menor que uno en módulo.

- Si cumplimos esto para cualquier valor de h el método es **incondicionalmente estable**
- si hace falta restringir h , esto es, $h < h_*$ para algún valor h_* finito, entonces el método es **conditionalmente estable**

Tener $|G(\lambda h)| < 1$ se puede descomponer en dos casos: En el caso

$$-1 < G(\lambda h) < 0$$

se admiten oscilaciones que decaen lo que puede ser admisible en algunas situaciones. Por otro lado, en el caso

$$0 < G(\lambda h) < 1$$

la solución numérica decae sin oscilar, que es lo que debe ocurrir.

Para el problema test en el caso de Euler explícito tenemos

$$G(\lambda h) = 1 - \lambda h, \quad (h_* = 2/a)$$

mientras que para Euler implícito

$$G(\lambda h) = \frac{1}{1 + \lambda h}, \quad (h_* = +\infty).$$

Fácilmente se ve que para Crank-Nicolson es

$$G(\lambda h) = \frac{1 - \lambda h/2}{1 + \lambda h/2}, \quad (h_* = +\infty)$$

mientras que para el Runge-Kutta clásico de cuarto orden es

$$G(\lambda h) = 1 - \lambda h + \frac{(\lambda h)^2}{2!} - \frac{(a\lambda h)^3}{3!} + \frac{(\lambda h)^4}{4!}, \quad (h_* \sim 2.785/\lambda)$$

aquí h_* se calcula con el computador. Por lo tanto, el análisis de estabilidad numérica reside en calcular este valor $G(z)$ y ver las condiciones para tener $|G(z)| < 1$.

Observar que en todos los casos, $G(\lambda h)$ es una aproximación a $e^{-\lambda h} = e^{-\lambda(t_{n+1}-t_n)}$ que es la solución exacta cuando vamos de t_n a t_{n+1} .

Resumiendo, en general, para un campo de curvas solución representado por $u' = f(t, u)$ y un esquema numérico que podemos representar por $u_{n+1} = G(\lambda h)u_n$. El análisis de estabilidad numérica del esquema sigue los pasos:

- Determinar $G(\lambda h)$ para el problema modelo $u'(t) = -\lambda u(t)$
- Determinar condiciones para que $|G(\lambda h)| < 1$.

Observación 107 El factor de amplificación $G(\lambda h)$ es característico del método numérico pero no tiene nada que ver con su error local ni con su orden.

Observación 108 Esto solo se puede hacer para problemas lineales por lo complicado que resulta en los no lineales. En el caso general, supongamos tenemos el campo

$$u'(t) = f(t, u(t))$$

y queremos estudiarlo en torno a un punto (t_0, u_0) . Usando desarrollo de Taylor se tiene que

$$f(t, u) = f(t_0, u_0) + f_t(t_0, u_0)(t - t_0) + f_u(t_0, u_0)(u - u_0) + \dots$$

entonces, olvidándonos de términos de orden mayor que uno tenemos

$$f(t, u) \sim -a u + b(t)$$

donde $a = -f_y(t_0, u_0)$ y $b(t) = f(t_0, u_0) + f_t(t_0, u_0)(t - t_0) - f_y(t_0, u_0)(u - u_0)$ es el resto. Localmente, podemos trabajar con el problema lineal

$$u'(t) = -a u(t) + b(t).$$

Usando el factor integrante e^{at} la solución exacta de este problema continuo es

$$u(t) = e^{-at} u_0 + e^{-at} \int_0^t e^{as} b(s) ds = e^{-at} u_0 + \int_0^t e^{-a(t-s)} b(s) ds.$$

Se puede aproximar aun más si suponemos que b es constante y entonces tenemos que la solución se calcula con facilidad y es

$$u(t) = e^{-at} u_0 + \frac{b}{a} (1 - e^{-at}).$$

entonces nos fijamos que en el caso estable $a > 0$, esto es $f_y(t_0, u_0) < 0$, repetimos la dinámica del caso $u' = -a u$ sólo que trasladada la solución estable a $u(t) \equiv b/a$. Esta es la razón por la que el término importante es el primero ($u' = -a u$) y es donde el esquema numéricico debe mostrar que tiene el comportamiento estable. Es decir, el problema realmente importante en este asunto es

$$u'(t) = -a u(t)$$

puesto que en el caso general

$$u'(t) = -a u(t) + b(t)$$

el término $b(t)$ no influye en la restricción sobre h .

Ejemplo 114 En la ecuación de Dahlquist-Bjorck es

$$\begin{cases} y'(t) = 100(\sin(t) - y(t)), & 0 < t < 3, \\ y(0) = 0 \end{cases}$$

se cumple $f(t, y) = 100(\sin(t) - y)$ luego $\partial_y f = -100$ y el análisis de estabilidad para Euler explícito nos pide $h < 2/100$, y como $T = 3$ y $h = T/N$ entonces debe ser $N > 150$. Valores menores generan oscilaciones en el cálculo discreto.

Ejemplo 115 Igual ocurre con la ecuación de de Prothero-Robinson donde $f(t, y) = L(\varphi(t) - y) + \varphi'(t)$ y $\partial_y f(t_0, y_0) = -L$ o en el ejemplo $f(t, y) = -3y + 3t$ donde $\partial_y f(t_0, y_0) = -3$.

Observación 109 Observar que si fijamos la longitud del intervalo T , entonces también podemos hacer que el número de puntos de la partición intervenga. Como $h = T/N$ y pedimos $h < 2/a$ entonces debemos tener $N > a/(2T)$ para estabilidad numérica en el caso de Euler explícito. En todo caso, es la razón $h = T/N$ la que debe cumplir la restricción de estabilidad numérica.

Observación 110 Sólo interesa hacer este estudio en el problema rígido estandard $y' = -\lambda y$ con $\lambda \gg 1$. Aquí hay decaimiento exponencial en la solución y sirve de modelo para casos más complicados. **El estudio de la estabilidad numérica no es importante si la edo ya es de por sí inestable** y esto es lo que ocurre con $y' = \lambda y$ con $\lambda \gg 1$.

Ejemplo 116 El campo $f(t, y) = f(y) = 1000(1 - y)$ posee una solución de equilibrio $w(t) = 1$ y el campo de vectores es muy contractivo. Al usar Euler explícito con $y_0 = 1$ se reproduce de forma exacta la solución $w(t) \equiv 1$. Si aplicamos el método a $f(t, y) = 1000(1 - y) = -1000y + 1000$ tenemos que el punto crítico es $y_* = 1$ y que $\partial_y f(1) = -1000$ la linealización da

$$z'(t) = -1000 z(t)$$

y necesitamos $h < 2/1000$ para que el esquema sea numéricamente estable. Esto es importante cuando $y_0 \neq 1$. Se puede ver de forma explícita en los cálculos manuales ya que tenemos

$$y_{n+1} = y_n + h 1000(1 - y_n) = (1 - h1000)y_n + 1000h, \quad n \geq 1$$

y la diferencia entre la solución constante $w \equiv 1$ y la calculada es

$$y_n^h - 1 = (1 - h1000)^n (y_0 - 1)$$

con lo que tenemos un factor de amplificación $G(1000h) = 1 - h1000$ de donde $|G(1000h)| < 1$ sólo si $h < 2/1000$.

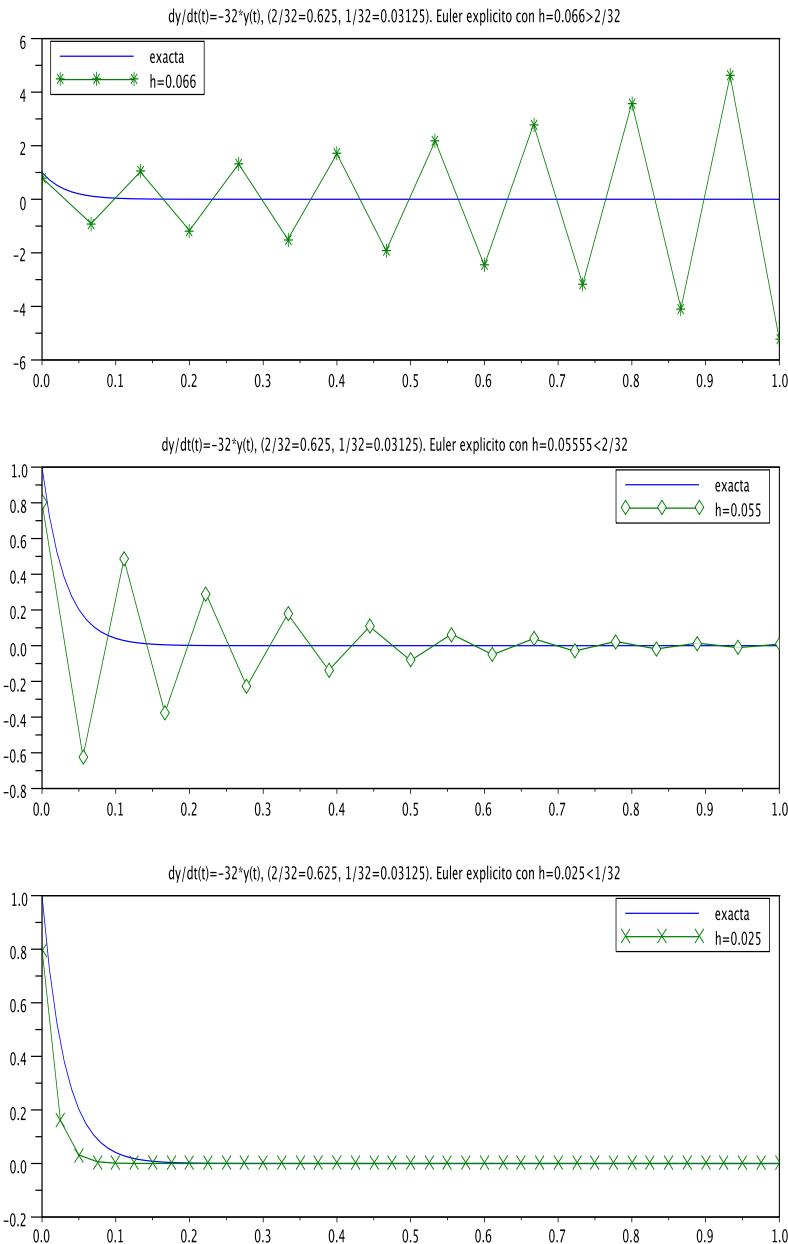


Figura 4.4: **Restricción de estabilidad:** Euler explícito para $y' = -ay$ con $a = 32$. Aproximaciones generadas de acuerdo a $h > 2/a$, $0 < h < 2/a$ ó $0 < h < 1/a$. Por defecto y para visualizar mejor se unen los puntos calculados con un trazo continuo, pero el cálculo sólo genera los puntos.

4.11. Estabilidad vs Precisión

*La restricción por estabilidad lleva a que el método numérico funcione y empiece a ser útil. Pero ahora **falta la precisión**, ya que los resultados no tienen por qué ser precisos. Al usuario le interesa que la restricción predominante sea la producida por la precisión antes que por la estabilidad. Luego lo mejor sería que no tuviésemos que preocuparnos por la estabilidad y sólo por la precisión, pero conseguir esto no es fácil.*

1. *¿De qué depende la restricción de precisión que desea el usuario?*

- *El método numérico usado determina la forma del error local. El valor concreto del error local depende del problema al que se le aplica el método y de la solución que se busca. La regularidad de la solución buscada determina el tamaño de las derivadas que aparece en el error local. Pero estos datos normalmente son desconocidos y sólo se puede tener una idea aproximada de sus acotaciones. Es por ello por lo que nos preocupa principalmente el orden de convergencia.*

2. *¿De qué depende la restricción de estabilidad?*

- *Cada método se puede clasificar según sus propiedades de estabilidad absoluta. Para un problema de valor inicial cualquiera no se conocen sus propiedades de rigidez pero se pueden sospechar. Aplicar un método numérico cualquiera y observar dificultades en la estabilidad nos indica que hay que cambiar a otro método numérico con mejores propiedades de estabilidad.*

3. *¿Qué restricción es más fuerte? ¿La impuesta por la precisión o la impuesta por la estabilidad?*

- *Depende del método usado y del problema de valor inicial al que le aplicamos el método numérico.*

4. *¿Qué hacer en la práctica?*

- *Buscar un equilibrio entre usar un método numérico con buenas propiedades de estabilidad y lo suficientemente preciso. Esto nos defenderá de campos de soluciones con rigideces y comportamientos bruscos. Normalmente los métodos de cuarto orden ofrecen un buen balance, y entre ellos destaca el Runge-Kutta clásico. En otros casos, cuando las evaluaciones de la función $f(t, y)$ se hacen muy pesadas o el problema es fuertemente inestable también son buenos candidatos Crank-Nicolson o incluso Euler implícito.*

4.11.1. Ejemplos

Observemos ahora el error global cometido usando Euler explícito en nuestro problema modelo $u' = -au$ con $u(0) = 1$ y $a \gg 1$. Sabemos que

$$u(t_{n+1}) = u(t_n) - h u'(t_n) + \frac{1}{2} u''(\xi_n) h^2 = u(t_n) - h a u(t_n) + l(u(t_n); h)$$

donde $l(u(t_n); h) = \frac{1}{2} u''(\xi_n) h^2$ y el esquema es

$$u_{n+1} = u_n - h a u_n.$$

Entonces el error global $e_n = u(t_n) - u_n$ es

$$e_{n+1} = (1 - h a) e_n + l(u(t_n); h), \quad n \geq 0.$$

Luego el error global en el paso $n + 1$ es la suma de dos términos:

- la amplificación del error global en el paso n por el factor $G(ah) = 1 - h a$
- el error local $l(u(t_n); h) \sim Cte h^2$ del esquema numérico en el paso n

Por otro lado, como conocemos la solución exacta $u(t) = e^{-at}$, el error local viene dado por

$$l(u(t_n); h) = \frac{1}{2} a^2 e^{-a\xi_n} h^2$$

y el factor $e^{-a\xi_n}$ decrece a cero. Por lo tanto, tener un error local pequeño, menor que ε por ejemplo, equivale a controlar este valor. En el caso $a > 0$ sólo necesitamos tener

$$l(h) = \frac{1}{2} a^2 e^{-a\xi_n} h^2 \leq \varepsilon \Rightarrow h^2 \leq \varepsilon 2 e^{a\xi_n} / a^2 \sim \varepsilon e^{a\xi_n}$$

y esto es fácil porque el factor $e^{a\xi_n}$ es cada vez más grande.

Por otro lado, controlar el factor de amplificación $G(ah) = 1 - ah$ implica tener $h < 2/a$ y si $a \gg 1$ es claro que la restricción de estabilidad $h < h_e = 2/a$ puede ser más exigente que la de precisión $h < h_p = (\varepsilon e^{a\xi_n})^{1/2}$.

Nos hemos encontramos entonces con la necesidad de restringir h por un doble motivo y tendremos que conseguir un error global pequeño que es lo que buscamos:

- **restricción por estabilidad** para conseguir factor de amplificación menor que uno cuando la dinámica sea contractiva, podemos escribir esta restricción como $h \leq h_e$
- **restricción por precisión** para conseguir un residuo $l(t_n; h)$ pequeño. Esto depende del método numérico. El residuo tiene la forma $l(t_n; h) \leq Cte h^{p+1}$ para $p \geq 1$ donde Cte depende de la solución buscada y sus derivadas, del campo de velocidades, del intervalo temporal etc... Por lo tanto, cuanto más grande p mejor será el control sobre $l(t_n; h)$. Podemos describir esta restricción como $h \leq h_p$.

Si queremos que el error global se mantenga pequeño necesitamos entonces ambas cosas y tendremos que imponer

$$h < \min\{h_e, h_p\}.$$

A nosotros nos interesa que el método sea preciso. Entonces si $h_e \ll h_p$ no estamos haciendo un buen negocio ya que predomina la restricción por estabilidad sobre la de precisión. Por ejemplo, si usamos Euler implícito en el campo $y' = -ay$ para $a > 0$ sólo debemos preocuparnos de la precisión y esto es bueno.

Observación 111 Los métodos numéricos son herramientas ajustables mediante algunos parámetros (en nuestro caso, sólo hay un parámetro que es h). Dado un campo de vectores, elegimos un método para determinar una curva buscada dentro de este campo de vectores y ajustamos el paso h para tener el error deseado. La facilidad del proceso dependerá del método elegido.

Ejemplo 117 Supongamos que tenemos el problema

$$y'(t) = L(\varphi(t) - y(t)) + \varphi'(t), \quad y(0) = y_0.$$

La solución exacta es

$$y(t) = e^{-Lt}(y_0 - \varphi(0)) + \varphi(t)$$

- Si $L \ll 0$ dos soluciones continuas cualesquiera se separan muy rápidamente debido al factor

$$e^{-Lt}(y_0 - \varphi(0))$$

y el problema de valor inicial es muy inestable. Aquí Euler explícito tiene un factor $G(h) = 1 - hL = 1 + h|L| \gg 1$ luego siempre amplifica los errores. Tendremos convergencia pero con un coste computacional muy alto, es decir, usando h muy pequeño. Por otro lado, para Euler implícito $G(h) = (1 + hL)^{-1}$ y sólo necesitamos satisfacer la restricción de precisión. Aquí se puede ver con más detalle pues se conoce el error local: la segunda derivada es

$$y''(t) = L^2 e^{|L|t}(y_0 - \varphi(0)) + \varphi''(t) \approx L^2 e^{|L|t}.$$

Por lo tanto, si los valores de $\varphi''(t)$ son moderados lo que manda es

$$y''(t) \approx L^2 e^{|L|t} \gg 1$$

y no se puede controlar con valores moderados de h .

- Si $L \gg 0$ dos soluciones continuas cualesquiera se unen muy rápidamente en la trayectoria $\varphi(t)$ debido al factor

$$e^{-Lt}(y_0 - \varphi(0))$$

y el problema de valor inicial es muy estable. Para Euler explícito necesitamos $h < 2/L$ inviable prácticamente. Mientras que Euler implícito no tiene ninguna restricción. Incluso lo hará mejor cuanto mayor sea L . Por otro lado, la segunda derivada es

$$y''(t) = L^2 e^{-L t} (y_0 - \varphi(0)) + \varphi''(t) \approx L^2 (y_0 - \varphi(0)) + \varphi''(t)$$

y como $L > 0$ entonces

$$y''(t) \approx L^2 (y_0 - \varphi(0)) + \varphi''(t)$$

y el error local se puede controlar con valores moderados de h siempre y cuando $\varphi''(t)$ sea moderada. Así que la restricción por precisión depende sobre todo de la solución buscada, lo que es razonable.

- Si $L \approx 0$ entonces las curvas solución son mas o menos paralelas en intervalos de tiempo moderados y ambas restricciones serán lo exigente que sea la solución buscada, ahora L no es importante.

4.12. Control del paso

En los códigos profesionales se ajusta un paso variable de forma automática para conseguir un error determinado.

Desde el punto de vista de la programación, la situación se complica puesto que ya no sabemos a priori el número de pasos que se va a dar y no podemos prefijar la memoria a usar. El número de pasos va a depender del algoritmo planteado.

Un esquema de un paso con paso variable se describe como:

Dado u_0 y h_0 hacer para $n = 0, 1, 2, \dots$

$$\begin{aligned} u_{n+1} &= u_n + h_n \Phi_f(t_n, u_n; h), \\ t_{n+1} &= t_n + h_n \end{aligned}$$

y el truco está aquí en poder tomar h_n lo más grande posible como para mantener una precisión predeterminada. Aunque no se puede saber el error global en cada paso se puede intentar estimar el error local usando dos evaluaciones distintas y suponiendo que una de ellas es la exacta.

Siguiendo Shampine et al. [24], una idea razonable puede ser que en cada t_n el código seleccione un paso h_n en la forma

$$h_n = \Theta(t_n) H$$

donde $\Theta(t)$ sea una función continua con $0 < \theta < \Theta(t) \leq 1$. La modificación de la demostración resulta ser bastante simple y el resultado es que si $H > 0$ tiende a cero entonces el error cometido es $O(H)$. Veamos como hacerlo:

Sabemos que en un paso de longitud h_n el error cometido es aproximadamente el de consistencia que viene dado por

$$\epsilon_n \approx h_n^2 \frac{|y''(t_n)|}{2},$$

por lo tanto, si $|y''(t_n)|$ es pequeña podremos tomar un h_n más grande. Además, podemos usar que $y'' = f_t + f f_y$ para poner

$$y''(t_n) \approx f_t(t_n, y_n) + y_n f_y(t_n, y_n)$$

y poder así calcular, o acotar, esta derivada en cada punto.

Si fijamos un error máximo admisible $\epsilon_\star > 0$, entonces el valor máximo que podremos tomar para h_n viene dado por

$$h_n \approx \sqrt{\frac{2\epsilon_\star}{|y''(t_n)|}}.$$

En el caso

$$\xi = \min_t \{|y''(t)|\} > 0$$

podemos definir

$$H = \sqrt{\frac{2\epsilon_\star}{\xi}}$$

de donde

$$h_n \approx \Theta(t_n) H, \quad \Theta(t_n) = \sqrt{\frac{\xi}{|y''(t_n)|}}.$$

Tenemos entonces $H = O(\epsilon_\star^{1/2})$ y finalmente $|y(t_n) - y_n| = O(\epsilon_\star^{1/2})$.

Evidentemente, la mayor objeción a esta idea es la necesidad de conocer y'' o bien de usar derivadas de f . Existen alternativas

4.12.1. Uso de paso doble

Vamos a suponer que partimos ya del valor correcto, esto es que $y_n = y(t_n)$. Para aceptar o rechazar un paso h a partir de (t_n, y_n) hacemos lo siguiente

- calculamos y^* el resultado de un paso de Euler con h , esto es

$$y^* = y_n + h f(t_n, y_n)$$

- repetimos el avance de t_n a $t_n + h = t_{n+1}$ pero ahora haciendo dos cálculos con Euler y tomando $h/2$ para avanzar. Primero avanzamos una distancia $h/2$ con Euler

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

y definitivamente, desde $(t_n + h/2, y_{n+1/2})$ recorremos la siguiente $h/2$ distancia también con Euler

$$y^{**} = y_{n+1/2} + \frac{h}{2} f(t_n + h/2, y_{n+1/2}).$$

Determinar si el avance de acuerdo con el paso h es bueno o no lo vamos a ver de acuerdo al error de truncatura que se cometa y esto lo vamos a hacer trabajando con los valores y^* e y^{**} como sigue: Sabemos que por desarrollo de Taylor

$$y(t_{n+1}) - y^* = y(t_n) - y_n + h [f(t_n, y(t_n)) - f(t_n, y_n)] + \frac{1}{2} h^2 y''(t_n) + \frac{1}{6} h^3 y'''(\xi_n).$$

La **hipótesis de que ya en t_n tenemos el valor correcto**, esto es que $y_n = y(t_n)$ nos simplifica la expresión y tenemos

$$y(t_{n+1}) - y^* = \frac{1}{2} h^2 y''(t_n) + \frac{1}{6} h^3 y'''(\xi_n). \quad (4.3)$$

Por otro lado, comparando ahora $y(t_{n+1}) - y^{**}$ mediante desarrollo de Taylor (de $y(t_{n+1})$ y de $f(t_n + h/2, y_{n+1/2})$) que

$$\begin{aligned} y(t_{n+1}) - y^{**} &= y(t_{n+1}) - y_{n+1/2} - \frac{h}{2} f(t_n + h/2, y_{n+1/2}) \\ &= y(t_n) + hy'(t_n) + \frac{h^2}{2} y''(t_n) + \frac{h^3}{6} y'''(t_n) + \dots \\ &\quad - y_n - \frac{h}{2} f(t_n, y_n) - \frac{h}{2} f(t_n + h/2, y_{n+1/2}). \end{aligned}$$

Como suponemos que $y_n = y(t_n)$ y por tanto $y'(t_n) = f(t_n, y_n)$ se simplifica a

$$y(t_{n+1}) - y^{**} = \frac{h}{2} y'(t_n) - \frac{h}{2} f(t_n + h/2, y_{n+1/2}) + \frac{h^2}{2} y''(t_n) + \frac{h^3}{6} y'''(t_n) + \dots$$

y desarrollando

$$f(t_n + h/2, y_{n+1/2}) = f(t_n + h/2, y_n + h/2 f(t_n, y_n)) = f(t_n + h/2, y(t_n) + y'(t_n)h/2)$$

tenemos

$$\begin{aligned} f(t_n + h/2, y_{n+1/2}) &= f(t_n, y_n) + \frac{h}{2} f_t + \frac{h}{2} y'(t_n) f_y + O(h^2) \\ &= y'(t_n) + \frac{h}{2} y''(t_n) + O(h^2) \end{aligned}$$

de donde

$$\begin{aligned} y(t_{n+1}) - y^{**} &= \frac{h}{2}y'(t_n) - \frac{h}{2}y'(t_n) - \frac{h^2}{4}y''(t_n) + O(h^3) + \frac{h^2}{2}y''(t_n) + O(h^3) \\ &= \frac{h^2}{4}y''(t_n) + O(h^3). \end{aligned}$$

Finalmente, hemos llegado a que

$$y(t_{n+1}) - y^* = \frac{1}{2}h^2 y''(t_n) + O(h^3) \quad (4.4)$$

$$y(t_{n+1}) - y^{**} = \frac{1}{4}h^2 y''(t_n) + O(h^3). \quad (4.5)$$

y esto ya nos permite jugar puesto que haciendo la diferencia

$$y^{**} - y^* = \frac{1}{2}h^2 y''(t_n) + O(h^3). \quad (4.6)$$

Por otro lado,

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

luego restando (4.4) obtenemos

$$y(t_{n+1}) = 2y^{**} - y^* + O(h^3). \quad (4.7)$$

Corolario 118 Consecuencias: *Estimamos el error de consistencia usando (4.6) y podemos obtener una mejor aproximación, a tercer orden, de $y(t_{n+1})$ mediante (4.7).*

Ahora podemos plantear un criterio para decidir si nuestro paso h es bueno o no. Tomemos $\epsilon > 0$ una tolerancia fijada. Construimos

$$r = \frac{|y^{**} - y^*|}{h} \approx \frac{1}{2}h |y''(t_n)|.$$

El cociente r aproxima el error de truncatura. Tenemos entonces que

- Si $r < \epsilon$ nos quedamos con este valor de h y tomamos como aproximación a $y(t_{n+1})$ el valor $2y^{**} - y^*$. Así que nuestro nuevo punto para continuar será $(t_n + h, 2y^{**} - y^*)$ que, otra vez, asumimos que nos da el verdadero punto $(t_n + h, y(t_{n+1}))$

- Si $r > \epsilon$ buscamos un nuevo valor de h y repetimos el cálculo

Cómo escogemos el nuevo h llamémosle \tilde{h} ? Buscamos

$$\frac{1}{2}\tilde{h}|y''(t_n)| < \epsilon$$

pero no conocemos y'' , así que la reemplazamos por r

$$\frac{1}{2}\tilde{h}|y''(t_n)| = \frac{\tilde{h}}{h} \frac{1}{2}h|y''(t_n)| \approx \frac{\tilde{h}}{h} \frac{|y^{**} - y^*|}{h} = \frac{\tilde{h}}{h}r < \epsilon$$

entonces tomamos

$$\tilde{h} < h \frac{\epsilon}{r}$$

(4.8)

Por ejemplo, podemos usar

$$\tilde{h} = 0.9h \frac{\epsilon}{r}. \quad (4.9)$$

Por otro lado, si $r < \epsilon$ podemos seguir con este mismo valor de h para continuar o incrementarlo un poco. Usando que ahora es $\epsilon/r > 1$ podemos tomar h un poco más grande, por ejemplo $h = \tilde{h}$ donde \tilde{h} vienen dado por (4.9) o incluso algo mucho más simple como $\tilde{h} = ah$ para algún $a > 1$, en todo caso, aumentar h conlleva la posibilidad de pasar por alto alguna zona sensible en las derivadas.

Como ejemplos, ver Figura 3.2 y Figura 3.3. Este metodo generalizado a ordenes mas grandes puede generar un numero de evaluaciones de la función f demasiado alto.

4.12.2. Runge-Kutta adaptativos: pares encajados

La forma más habitual es mediante el uso de dos métodos de orden distinto que compartan varios coeficientes y etapas. Así se ahorran cálculos: en cada paso se realizan unos cálculos básicos que generan dos aproximaciones, la primera de orden p y la segunda de orden $p+1$. Con estas dos aproximaciones se puede estimar el error de consistencia local de cualquiera de los dos métodos y de acuerdo con este valor se modifica el paso. Sequimos aquí el libro de Burden y Faires [4]:

Supongamos que tenemos el método generado por Φ_f de orden p y el generado por Ψ_f de orden $p+1$ (pondremos Φ y Ψ para abreviar, pero ambas dependen de f). Entonces tenemos valores

$$w_{n+1} = w_n + h_n \Phi(t_n, w_n; h), \quad n \geq 0, \quad (4.10)$$

$$v_{n+1} = v_n + h_n \Psi(t_n, v_n; h), \quad n \geq 0. \quad (4.11)$$

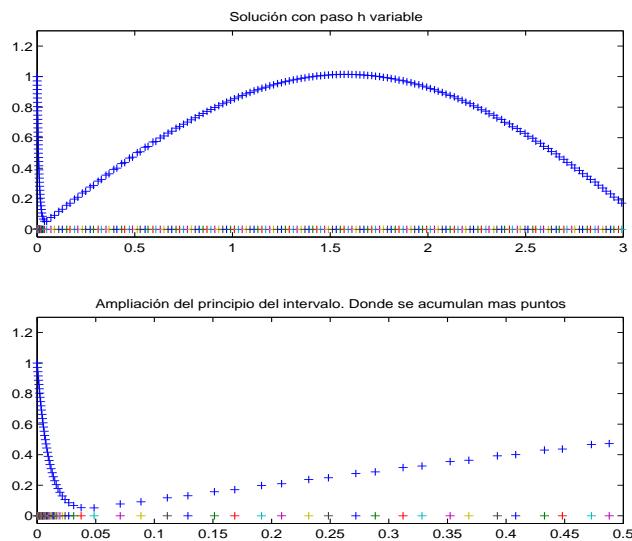


Figura 4.5: Ajuste del paso usando el proceso doble.

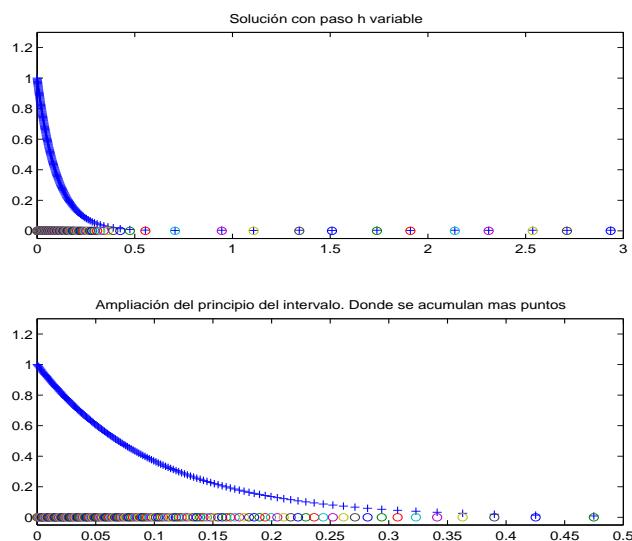


Figura 4.6: Ajuste del paso usando el proceso doble para un decaimiento exponencial.

Cualquiera de los dos esquemas, idealmente debería de satisfacer que para una tolerancia dada $\epsilon > 0$ se cumpliese

$$|y(t_n) - w_n| < \epsilon, \quad n = 0, 1, 2, \dots$$

Tener un número pequeño de puntos donde evaluar y al mismo tiempo controlar el error global parece inconsistente con tener los puntos igualmente distribuidos en el intervalo. Usando métodos de diferente orden podemos predecir el error local de truncatura, y usando esta predicción, elegir un paso que controle este error local y también el error global.

Siendo $y(t)$ la curva exacta que queremos aproximar y partiendo de la **hipótesis de localización**: Damos por buenos los valores en el paso t_n , esto es

$$w_n \approx y(t_n) \approx v_n,$$

los errores de consistencia locales en un punto t_n cuando damos un paso h son

$$\begin{aligned} l_\Phi(y(t_n); h) &= y(t_n + h) - \tilde{w}_{n+1}, \quad (\tilde{w}_{n+1} = y(t_n) + h \Phi(t_n, y(t_n); h)) \\ l_\Psi(y(t_n); h) &= y(t_n + h) - \tilde{v}_{n+1}, \quad (\tilde{v}_{n+1} = y(t_n) + h \Psi(t_n, y(t_n); h)) \end{aligned}$$

y suponemos que

$$l_\Phi(y(t); h) \approx K_1 h^{p+1}, \quad l_\Psi(y(t); h) \approx K_2 h^{p+2}$$

o lo que es lo mismo

$$\tau_\Phi(y(t); h) \approx K_1 h^p, \quad \tau_\Psi(y(t); h) \approx K_2 h^{p+1}.$$

Por lo tanto,

$$\begin{aligned} \tau_\Phi(y(t_n); h) &= \frac{l_\Phi(y(t_n); h)}{h} = \frac{y(t_n + h) - \tilde{w}_{n+1}}{h} = \frac{y(t_n + h) - \tilde{v}_{n+1} + \tilde{v}_{n+1} - \tilde{w}_{n+1}}{h} \\ &= \frac{y(t_n + h) - \tilde{v}_{n+1}}{h} + \frac{\tilde{v}_{n+1} - \tilde{w}_{n+1}}{h} = \tau_\Psi(y(t_n); h) + \frac{\tilde{v}_{n+1} - \tilde{w}_{n+1}}{h} \end{aligned}$$

de donde podemos relacionar los dos errores de truncatura

$$\tau_\Phi(y(t_n); h) = \tau_\Psi(y(t_n); h) + \frac{\tilde{v}_{n+1} - \tilde{w}_{n+1}}{h}.$$

siendo $\tau_\Phi = O(h^p)$ y $\tau_\Psi = O(h^{p+1})$ entonces se tiene

$$O(h^p) = \frac{\tilde{v}_{n+1} - \tilde{w}_{n+1}}{h} + O(h^{p+1})$$

por lo que el término $(\tilde{v}_{n+1} - \tilde{w}_{n+1})h^{-1}$ será el término $O(h^p)$ de la igualdad. Suponiendo que $\tilde{v}_{n+1} = v_{n+1}$ y que $\tilde{w}_{n+1} = w_{n+1}$ tenemos

$$\tau_\Phi(y(t_n); h) = \frac{v_{n+1} - w_{n+1}}{h} + \tau_\Psi(y(t_n); h).$$

Como $\tau_\Phi(y(t); h) \approx K_1 h^p$, $\tau_\Psi(y(t); h) \approx K_2 h^{p+1}$ para h lo suficientemente pequeño, la parte significante de $\tau_\Phi(y(t); h)$ debe de venir de

$$\frac{v_{n+1} - w_{n+1}}{h}.$$

Por lo tanto, podemos aproximar

$$\tau_\Phi(y(t); h) \approx \frac{v_{n+1} - w_{n+1}}{h} = \frac{h\Phi(t_n, v_n; h) - h\Psi(t_n, w_n; h)}{h} = \Phi(t_n, v_n; h) - \Psi(t_n, w_n; h).$$

La expresión

$$\tau_\Phi(y(t); h) \approx \Phi(t_n, v_n; h) - \Psi(t_n, w_n; h)$$

nos indica que **si construimos Φ y Ψ de forma que se parezcan podremos simplificar el coste computacional y obtener τ_Φ de forma rápida.**

Hemos estimado el error de truncatura local, pero también queremos controlar el paso que damos: dada una tolerancia $\epsilon > 0$ si

$$\tau_\Phi(y(t); h) < \epsilon$$

aceptamos el paso h y pasamos al punto $t + h$ usando w_{n+1} . Por otro lado, si

$$\tau_\Phi(y(t); h) > \epsilon$$

rechazamos este h , tomamos un h más pequeño y repetimos el cálculo. Observemos que para el paso sh el error de truncatura local $\tau_\Phi(y(t); sh)$ se puede expresar en términos del error de truncatura local $\tau_\Phi(y(t); h)$ como sigue

$$\tau_\Phi(y(t); sh) \approx s^p K_1 h^p \approx s^p \tau_\Phi(y(t); h)$$

de donde podemos usar

$$|\tau_\Phi(y(t); sh)| = s^p \frac{|v_{n+1} - w_{n+1}|}{h} < \epsilon$$

es decir,

$$s < \left(\frac{h \epsilon}{|v_{n+1} - w_{n+1}|} \right)^{1/p}.$$

Tomando un factor de seguridad, 0.9 por ejemplo, aproximamos s como

$$s = 0.9 \left(\frac{h \epsilon}{|v_{n+1} - w_{n+1}|} \right)^{1/p}$$

o bien

$$s = 0.9 \exp \left(\frac{1}{p} \log \left(\frac{h \epsilon}{|v_{n+1} - w_{n+1}|} \right) \right)$$

y repetimos el proceso avanzando de t a $t + sh$.

Se pretende que los cálculos de Φ_f y de Ψ_f en cada paso sean parecidos para ahorrar así coste computacional. Una forma de abordar este problema es usando el concepto de **pares encajados**:

Dado un método de Runge-Kutta de orden p con tablero

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

si tenemos otro método con las mismas etapas pero con otro orden $p + 1 > p$

$$\begin{array}{c|A} c & \\ \hline & \beta \end{array}$$

podemos calcular sin mucho coste adicional. Por ejemplo, tenemos

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline w_{n+1} & 1 & 0 \\ \hline v_{n+1} & 1/2 & 1/2 \end{array}$$

es decir, avanzamos con Euler explícito y estimamos con Euler mejorado.

Definición 119 *Se dice que un par encajado es de orden $p(q)$ si se avanza con un método de orden p y se estima con uno de orden $q > p$.*

Observación 112 *Se puede avanzar con el método de orden más alto y entonces la idea de justificación anterior se abandona y simplemente consideramos la diferencia entre métodos con el propósito de seleccionar el paso, ver Hairer pag. 168 [14].*

Existen varias familias de pares encajados, se suelen conocer por los nombres de los investigadores que los estudiaron. Siendo la idea general común, los detalles técnicos pueden cambiar de una familia a otra dependiendo del criterio con el que se ha desarrollado la familia de métodos. Por ejemplo están las familias de métodos de Fehlberg, las de Dormand-Prince, las de Cash y Carp, etc... Vemos alguna de ellas como muestra a continuación

4.12.3. Runge-Kutta-Fehlberg 4(5)

Este método es también muy popular. Usa un método de Runge-Kutta con orden de truncatura local 5 (orden global 5) dado por

$$v_{n+1} = v_n + \frac{16}{135}hk_1 + \frac{6656}{12825}hk_3 + \frac{28561}{56430}hk_4 - \frac{9}{50}hk_5 + \frac{2}{55}hk_6$$

para estimar el error local en el Runge-Kutta de orden global 4 dado por

$$w_{n+1} = w_n + \frac{25}{216}hk_1 + \frac{1408}{2565}hk_3 + \frac{2197}{4104}hk_4 - \frac{1}{5}hk_5$$

y donde los coeficientes vienen dados por:

$$\begin{aligned} k_1 &= f(t_n, w_n) \\ k_2 &= f\left(t_n + \frac{1}{4}h, w_n + \frac{1}{4}k_1\right) \\ k_3 &= f\left(t_n + \frac{3}{8}h, w_n + \frac{3}{32}hk_1 + \frac{9}{32}hk_2\right) \\ k_4 &= f\left(t_n + \frac{12}{13}h, w_n + \frac{1932}{2197}hk_1 - \frac{7200}{2197}hk_2 + \frac{7296}{2197}hk_3\right) \\ k_5 &= f\left(t_n + h, w_n + \frac{439}{216}hk_1 - 8hk_2 + \frac{3680}{513}hk_3 - \frac{845}{4104}hk_4\right) \\ k_6 &= f\left(t_n + \frac{1}{2}h, w_n - \frac{8}{27}hk_1 + 2hk_2 - \frac{3544}{2565}hk_3 + \frac{1859}{4104}hk_4 - \frac{11}{40}hk_5\right) \end{aligned}$$

El tablero correspondiente al par de Runge-Kutta-Fehlberg 4(5) es

0	0	0	0	0	0	0
$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	0	0
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$	0	0	0	0
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$	0	0	0
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$	0	0
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	0
$w_{n+1}(\beta_i)$	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0 (avance)
$v_{n+1}(b_i)$	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$ (estimador)

aquí la primera fila determina el método de orden 4 mientras que la segunda el de orden 5 y se usa el **método de orden 4 para avanzar**.

Al igual que antes, se toman

$$\begin{aligned} k_i &= f(t_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{i,j} k_j), \quad i = 1, 2, \dots, 6 \\ v_{n+1} &= v_n + h \sum_{j=1}^6 b_j k_j(t_n, v_n; h), \quad (\text{orden global } 5) \\ w_{n+1} &= w_n + h \sum_{j=1}^6 \beta_j k_j(t_n, w_n; h) \quad (\text{orden global } 4). \end{aligned}$$

En particular, la estrategia en esta familia de métodos consiste dar el paso definitivo con el método de orden inferior w_{n+1} y usar v_{n+1} para estimar el error de truncatura local para el método de cuarto orden (usamos v_{n+1} como valor exacto):

$$\tau_4(h) = \frac{v_{n+1} - w_{n+1}}{h} = \sum_{j=1}^6 (b_j - \beta_j) k_j(t_n, y_n; h)$$

Como suponemos que $\tau_4(h) \sim K h^4$, si queremos dar un paso de talla sh en vez de talla h tenemos la relación

$$\tau_4(sh) \sim K(sh)^4 = K h^4 s^4 = \tau_4(h)s^4$$

y si queremos que sea $\tau_4(sh) < \epsilon$, entonces debe ser

$$\tau_4(sh) \sim \tau_4(h)s^4 < \epsilon$$

de donde

$$s < \left(\frac{\epsilon}{\tau_4(h)} \right)^{1/4} = \left(\frac{\epsilon}{\left| \sum_{j=1}^6 (\beta_j - b_j) k_j \right|} \right)^{1/4}$$

Finalmente, se toma

$$s = 0.9 \left(\frac{\epsilon}{\tau_4(h)} \right)^{1/4} \tag{4.12}$$

y además, para evitar pasos muy largos,

$$0.1 \leq s \leq 10. \tag{4.13}$$

Resumiendo, dado $\epsilon > 0$, si $\tau_4(h) < \epsilon$ entonces aceptamos el valor de h e incluso podemos tomar un paso más grande para realizar el siguiente cálculo, por ejemplo, $h := 10h$, el propio código lo ajustará. En el caso donde $\tau_4(h) > \epsilon$ entonces tomamos un nuevo valor de h como $h := sh$ de acuerdo a (4.14) y a (4.15), con cuidado de no pasarnos de $t_0 + T$. Volvemos a calcular entonces.

4.12.4. Runge-Kutta Dormand-Prince 5(4)

El método Dormand–Prince tiene siete etapas, pero sólo usa seis evaluaciones de función por paso ya que la última etapa de un paso se evalúa en el mismo punto que el primero del paso siguiente. Corresponde a un metodo de 7 etapas de orden 5 que contiene a uno de orden 4.

Dormand y Prince escogieron los coeficientes de su método para minimizar el error de la solución de quinto orden y tiene buenas propiedades en cuanto a predicción de error. Esta es la principal diferencia con el método de Fehlberg, que se construyó de modo que la solución de cuarto orden tenga un error pequeño. Por esa razón, el método de Dormand–Prince **es más adecuado cuando la solución de orden superior se usa para continuar la integración**, una práctica conocida como **interpolación local**. Dormand–Prince es el método principal de resolución de EDOs en MATLAB. Se toman

$$\begin{aligned} k_i &= f(t_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{i,j} k_j), \quad i = 1, 2, \dots, 6, 7 \\ y_{n+1} &= y_n + h \sum_{j=1}^7 \beta_j k_j(t_n, y_n; h), \quad (\text{orden global 5}) \\ z_{n+1} &= z_n + h \sum_{j=1}^7 b_j k_j(t_n, z_n; h) \quad (\text{orden global 4}). \end{aligned}$$

En particular, la estrategia en esta familia de métodos consiste dar el paso definitivo con el método de orden superior y_{n+1} y usar z_{n+1} para construir el estimador de error que permita avanzar:

$$\tau_5(h) = \left| \frac{y_{n+1} - z_{n+1}}{h} \right| = \left| \sum_{j=1}^7 (\beta_j - b_j) k_j(t_n, y_n; h) \right|.$$

Como suponemos que $\tau_5(h) \sim K h^5$, si queremos dar un paso de talla sh en vez de talla h tenemos la relación

$$\tau_5(sh) \sim K (sh)^5 = K h^5 s^5 = \tau_5(h)s^5$$

y si queremos que sea $\tau_5(sh) < \epsilon$, entonces debe ser

$$\tau_5(sh) \sim \tau_5(h)s^5 < \epsilon$$

de donde

$$s < \left(\frac{\epsilon}{\tau_5(h)} \right)^{1/5} = \left(\frac{\epsilon}{\left| \sum_{j=1}^7 (\beta_j - b_j) k_j(t_n, y_n; h) \right|} \right)^{1/5}$$

Finalmente, se toma

$$s = 0.9 \left(\frac{\epsilon}{\tau_5(h)} \right)^{1/5} \quad (4.14)$$

y además, para evitar pasos muy largos,

$$0.1 \leq s \leq 10. \quad (4.15)$$

Resumiendo, dado $\epsilon > 0$, si $\tau_5(h) < \epsilon$ entonces aceptamos el valor de h e incluso podemos tomar un paso más grande para realizar el siguiente cálculo, por ejemplo, $h := 10 h$, el propio código lo ajustará. En el caso donde $\tau_5(h) > \epsilon$ entonces tomamos un nuevo valor de h como $h := s h$ de acuerdo a (4.14) y a (4.15), con cuidado de no pasarnos de $t_0 + T$. Volvemos a calcular entonces.

El tablero de coeficientes de Dormand-Prince 5(4) es el siguiente:

0	0	0	0	0	0	0	0
1	$\frac{1}{5}$	0	0	0	0	0	0
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$	0	0	0	0	0
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$	0	0	0	0
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$	0	0	0
$\frac{1}{1}$	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$	0	0
$\frac{1}{1}$	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
$y_{n+1}(\beta_i)$	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0 (avance)
$z_{n+1}(b_i)$	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$ (estimador)

aquí la primera fila determina el método de orden 5 mientras que la segunda el de orden 4: **avanzamos con el método de orden 5**. Denotaremos este método como RK-DP5(4).

4.13. Otras ideas para mejorar precisión

4.13.1. Métodos de Taylor

La extensión de la idea analítica encontrada en el método de Euler en cuanto al desarrollo de Taylor es trivial y nos lleva a lo que se conoce como métodos de

Taylor. Estos métodos tiene mayor orden de aproximación a la derivada que Euler explícito. La dificultad que encontramos es **la necesidad de calcular derivadas sucesivas de la función pendiente** y esto hace que estos métodos no se usen normalmente en la práctica. En todo caso, es pedagógico su comprensión y tienen algo de uso por el auge del software de cálculo simbólico; por eso los mostramos brevemente.

El método de Euler explícito se puede entender como una aproximación al desarrollo de Taylor

$$u(t_{n+1}) = u(t_n) + hu'(t_n) + \frac{1}{2}h^2u''(\xi_n), \quad \xi_n \in (t_n, t_{n+1})$$

donde se elimina el término con la potencia h^2 y se usa $u'(t_n) = f(t_n, u(t_n))$ para escribir el esquema

$$u_{n+1} = u_n + hf(t_n, u_n).$$

Se interpreta como avanzar desde t_n a $t_n + h$ por la línea recta

$$r(s) = u_n + (s - t_n)f(t_n, u_n)$$

es decir, u_n aproxima a $u(t_n)$ y entonces $u'(t_n)$ se aproxima por $f(t_n, u_n)$.

De la misma forma, podemos también usar un desarrollo con más términos y escribir

$$u(t_{n+1}) = u(t_n) + hu'(t_n) + \frac{1}{2}h^2u''(t_n) + \frac{1}{3!}h^3u'''(\xi_n), \quad \xi_n \in (t_n, t_{n+1}).$$

y entonces usar un polinomio para avanzar desde t_n a $t_n + h$ de orden mayor que uno (dos en este caso):

$$p(s) = y_n + (s - t_n)u'(t_n) + \frac{1}{2}(s - t_n)^2u''(t_n)$$

y ahora tenemos que reemplazar $u'(t_n)$ por $f(t_n, u_n)$ y la segunda derivada $u''(t_n)$ por la derivación de $f(t, y(t))$ con respecto a $y(t)$. Si usamos la derivación en cadena

$$y'(t) = f(t, y(t)) \Rightarrow y''(t) = f_t(t, y(t)) + f(t, y(t))f_y(t, y(t))$$

generamos el esquema

$$u_{n+1} = u_n + hf(t_n, u_n) + \frac{1}{2}h^2\{f_t(t_n, u_n) + f(t_n, u_n)f_y(t_n, u_n)\}$$

eliminando la potencia h^3 , recordemos que $f(t, y)$ siempre es un dato. Aquí tendremos un error local $O(h^3)$.

Ejemplo 120 Tomemos,

$$\begin{cases} y'(t) &= (1 - 2t)y(t) \\ y(0) &= 1 \end{cases}$$

con solución $y(t) = \exp(0.25 - (t - 0.5)^2)$. Entonces

$$f(t, y) = (1 - 2t)y$$

de donde

$$f_t = -2y, \quad f_y = (1 - 2t) \Rightarrow f_t + f f_y = -2y_n + (1 - 2t_n)(1 - 2t_n)y_n$$

y el esquema nos queda como

$$y_{n+1} = y_n + h(1 - 2t_n)y_n + \frac{1}{2}h^2\{-2 + (1 - 2t_n)^2\}y_n$$

con un error local de orden h^3 . Éste es un método de Taylor de orden 2.

Siguiendo esta misma idea podemos desarrollar métodos de orden p cualquiera usando el desarrollo de Taylor de orden p

$$u(t_{n+1}) = u(t_n) + hu'(t_n) + \frac{1}{2}h^2u''(t_n) + \dots + \frac{1}{p!}h^p u^{(p)}(t_n) + \frac{1}{(p+1)!}h^{p+1}y^{(p+1)}(\xi_n)$$

y donde $\xi_n \in (t_n, t_{n+1})$. Entonces el esquema queda, usando la notación obvia, como

$$u_{n+1} = u_n + hu'_n + \frac{1}{2}h^2u''_n + \dots + \frac{1}{p!}h^p u_n^{(p)} := T_p(u_n; h)$$

con un error proporcional a h^{p+1} , sólo tenemos que obtener $u_n^{(j)} \approx u^{(j)}(t_n)$.

Tenemos entonces el método escrito en la forma

$$u_{n+1} = u_n + h\Phi_f(t_n, u_n; h), \quad \Phi_f = T_p$$

y con un error local de orden $p+1$. Sabemos que **si Φ_f es Lipschitz con respecto a la segunda variable y de forma uniforme con respecto a h tendremos la 0-estabilidad** de estos métodos. Para esto sólo necesitamos condiciones semejantes a la condición de Lipschitz para $f(t, y)$ en las distintas derivadas parciales de $f(t, y)$ que surgen y condiciones de acotación sobre la función. Esto se da por supuesto normalmente. Por lo demás, podemos concluir:

Teorema 121 El método de Taylor de orden p converge con orden p .

Observación 113 El método generado es eficaz en tanto en cuanto sea fácil calcular las derivadas de f . Pero esto no suele ser así, y además, cuando lo es, hay que construir el método de forma específica para cada problema. Esta es la razón por lo que los métodos de Taylor no se usan con frecuencia.

En todo caso es útil recordar las derivaciones siguientes:

$$y^{(N)}(t) = (\partial_t + f(t, y(t)) \partial_y)^N f(t, y(t)), \quad N = 1, 2, \dots$$

o brevemente,

$$y^{(N)}(t) = (\partial_t + f \partial_y)^N f, \quad N = 1, 2, \dots$$

4.14. Ejercicios

1. Interpretar geométricamente el siguiente método de Runge-Kutta explícito de 3 etapas

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{9} (2k_1 + 3k_2 + 4k_3) \\ k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1) \\ k_3 &= f(t_n + \frac{3}{4}h, y_n + \frac{3}{4}h k_2). \end{aligned}$$

y aplicarlo al problema $y'(t) = \lambda y(t)$ obteniendo una estimación del error global cometido.

2. Dada la ecuación diferencial $y'(t) = f(t, y(t))$, $y(0) = \alpha$ y la familia de métodos de Runge-Kutta explícitos de tres etapas

$$\begin{aligned} y_{n+1} &= y_n + h \sum_{i=1}^3 b_i k_i, \\ k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + c_2 h, y_n + ha_{21} k_1), \\ k_3 &= f(t_n + c_3 h, y_n + h\{a_{31} k_1 + a_{32} k_2\}), \end{aligned}$$

se tienen las siguientes condiciones para garantizar un orden 3:

$$\begin{aligned} b_1 + b_2 + b_3 &= 1, \\ b_2 c_2 + b_3 c_3 &= 1/2, \\ b_2 c_2^2 + b_3 c_3^2 &= 1/3, \\ b_3 c_2 a_{32} &= 1/6. \end{aligned}$$

El método de Nystrom es el resultante de encontrar una solución a estas restricciones en donde $c_2 = c_3$ y $b_2 = b_3$. Calcular los coeficientes, escribir el método, aplicarlo al problema $y'(t) = \lambda y(t)$, $y(0) = 1$ calculando la solución al esquema y comprobar el error global cometido.

3. La situación de un satélite viene determinada por su posición $\vec{r}(t) \in \mathbb{R}^3$, su velocidad $\vec{v}(t) \in \mathbb{R}^3$ y su aceleración $\vec{a}(t) \in \mathbb{R}^3$, donde

$$\vec{v}(t) = \frac{d}{dt}\vec{r}(t) = \vec{r}'(t) \in \mathbb{R}^3, \quad \vec{a}(t) = \frac{d}{dt}\vec{v}(t) = \frac{d^2}{dt^2}\vec{r}(t) = \vec{r}''(t) \in \mathbb{R}^3.$$

La relación que liga entre sí estas magnitudes viene dada por la segunda Ley de Newton $\vec{F} = m\vec{a}$ y si la fuerza se toma dependiente de la posición y de la velocidad llegamos a una ecuación diferencial ordinaria de segundo orden

$$\frac{d^2}{dt^2}\vec{r}(t) = f(t, \vec{r}(t), \vec{r}'(t))$$

que suele venir complementada con la posición y velocidad inicial, esto es $\vec{r}(0) = \vec{r}_0, \vec{r}'(0) = \vec{v}_0$.

Para simplificar, vamos a suponer que $r(t)$ es un escalar, entonces tenemos la ecuación

$$\begin{cases} r''(t) = f(t, r(t), r'(t)), & t > 0 \\ r(0) = x_0, \\ r'(0) = v_0 \end{cases}$$

Escribir este problema como una ecuación diferencial de primer orden (sistema). Describir las ecuaciones en diferencias resultantes con su notación conveniente si se le aplica un método de Runge-Kutta explícito de 3 etapas. Escoger cualquier método de los expuestos en los apuntes de clase.

4. Construir el método de Runge-Kutta de tres etapas que tiene la siguiente matriz de Butcher

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1/2 & 1/2 & 0 \\ \hline & 3/6 & 1/6 & 2/6 \end{array}$$

Comprobar que el método es de orden 2 y no es tres en general. Comprobar que si es de orden 3 cuando se aplica al problema $y'(t) = \lambda y(t)$ con $y(0) = 1$. Explicar esta aparente contradicción.

5. Para el problema $y'(x) = 0$ con $y(0) = 1$ se considera el método de Runge-Kutta explícito de cuatro etapas

$$\begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ -1 & 1/2 & -3/2 & 0 & 0 \\ 1 & 0 & 4/3 & -1/3 & 0 \\ \hline & 1/6 & 2/3 & 0 & 1/6 \end{array}$$

Estudiar sus propiedades.

6. Comprobar que el método de Runge-Kutta de tres etapas

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 3/4 & 0 & 3/4 & 0 \\ \hline & 2/9 & 1/3 & 4/9 \end{array}$$

tiene orden 3.

7. Construir todos los métodos de Runge-Kutta de tres etapas y que tiene la siguiente matriz de Butcher

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & c_2 & 0 & 0 \\ c_3 & 0 & c_3 & 0 \\ \hline & 0 & 0 & 1 \end{array}$$

Comprobar que el orden de este método no es tres. Aplicarlo al problema $y'(t) = \lambda y(t)$ con $y(0) = 1$ y obtener la estimación de error de global sobre ésta solución concreta y éste campo lo que da un orden global tres. Explicar esta aparente contradicción.

8. Ejercicio 1 de Hairer et al. [14].
 9. Para $y_0 = \alpha$ consideramos el esquema

$$y_{n+1} = y_n + h \left\{ \frac{1}{4} P_n + \frac{3}{4} Q_n \right\}$$

donde

$$P_n = f(t_n + \frac{h}{3}, y_n + \frac{h}{3} f(t_n, y_n)), \quad Q_n = f(t_n + h, y_n + h f(t_n, y_n)).$$

Identificarlo de acuerdo a la notación habitual con k_1, k_2 , etc... Escribir la matriz de Butcher, comprobar que la función de incremento es Lipschitz y obtener el error de consistencia.

10. Consideraremos el siguiente método creado por Heun

$$y_{n+1} = y_n + h \left\{ \frac{1}{4} P + \frac{3}{4} Q \right\}, \quad n \geq 0.$$

donde

$$P = f(t, y_n), \quad Q = f\left(t + \frac{2}{3} h, y_n + \frac{2}{3} h f\left(t + \frac{1}{3} h, y_n + \frac{1}{3} h P\right)\right)$$

Identificarlo de acuerdo a la notación habitual con k_1, k_2 , etc...

- a) Comprobar que su función de incremento satisface una condición de Lipschitz con respecto a la segunda variable.
- b) Escribir su tablero.
- c) Determinar el orden
- d) Aplicar al problema modelo $y'(t) = y(t)$ con $y(0) = 1$ y comprobar que el error cometido para este problema es 3, es decir, $e^{-\lambda t} - y_n^h = O(h^3)$.

11. Para a y b constantes considerar el esquema

Dado y_0 , obtener

$$y_{n+1} = y_n + h \{ b k_1 + (1 - b) k_2 \}, \quad n \geq 0$$

donde

$$k_1 = f(t_n, y_n), \quad k_2 = f(t_n + a h, y_n + a h k_1).$$

Existe una relación entre los valores de a y b para obtener error local de orden 3 con respecto a h . Demostrando todos los pasos que se realizan:

- a) Obtener esta relación imponiendo la regularidad que se vaya necesitando y mostrar el tablero de Butcher asociado.
- b) Para estos valores comprobar la estabilidad del esquema.
- c) Comprobar la convergencia del mismo hacia la solución del problema de Cauchy indicando el orden del método.
- d) Aplicar el caso particular

$$y_{n+1} = y_n + h \left\{ \frac{1}{4} k_1 + \frac{3}{4} k_2 \right\}, \quad n \geq 0.$$

donde

$$k_1 = f(t, y_n), \quad k_2 = f\left(t + \frac{2}{3} h, y_n + \frac{2}{3} h k_1\right)$$

al problema modelo $y'(t) = -\lambda y(t)$ con $\lambda > 0$ con $y(0) = 1$ y comprobar que el error cometido es el que indica el análisis realizado, es decir, $e^{-\lambda t} - y_n^h = O(h^2)$ en el límite estacionario $h n = t$.

- e) Con las mismas condiciones que en el apartado previo comprobar el orden de error para el esquema

$$y_{n+1} = y_n + h \left\{ \frac{3}{4} k_1 + \frac{1}{4} k_2 \right\}, \quad n \geq 0.$$

cuando se aplica al problema modelo $y'(t) = -\lambda y(t)$ con $\lambda > 0$ con $y(0) = 1$. Comprobar ahora que el error cometido es $e^{-\lambda t} - y_n^h = O(h)$ en el límite estacionario $h n = t$. Explicar la razón.

12. Demostrar que cuando el método de Euler mejorado, o de Heun, se aplica al problema

$$y' = 4y, \quad y(0) = 1/3$$

en el punto $t = 1/2$ el valor de la aproximación con paso h es

$$y(1/2) \approx \frac{1}{3}(1 + 4h + 8h^2)^{1/(2/h)}.$$

13. Construir pares encajados 1(2) de dos etapas.
 14. Hallar todos los pares encajados 2(3) en la forma

0	0	0	0
1	1	0	0
1/2	a_{31}	a_{32}	0
w_{n+1}	b_1	b_2	0
v_{n+1}	\hat{b}_1	\hat{b}_2	\hat{b}_3

15. Probar que

0	0	0	0
1/4	1/4	0	0
1/2	0	1/2	0
1	1	-2	2
w_{n+1}	1	-2	2
v_{n+1}	1/6	0	4/6

es un par encajado 2(4).

16. Comprobar que el método de Taylor de orden 3 aplicado a $y'(t) = 2t y(t)$ lleva a la iteración

$$y_{n+1} = y_n + 2t_n y_n h + y_n(1 + 2t_n^2)h^2 + 2t_n y_n(3 + 2t_n^2)h^3/3$$

o, lo que es lo mismo,

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

donde

$$\Phi_f(t_n, y_n; h) = 2t_n y_n h + y_n(1 + 2t_n^2)h^2 + 2t_n y_n(3 + 2t_n^2)h^3/3.$$

Capítulo 5

Métodos Multipaso

Resumen del tema

Primera lectura:

- interpretación geométrica
- ecuaciones en diferencias
- Conceptos de estabilidad numérica, precisión y orden de convergencia.

5.1. Introducción

¿Se pueden usar los datos que se van calculando para mejorar el orden? la respuesta es que sí. Hay toda una familia de métodos bien extensa y popular conocida como los métodos multipaso, en contraposición con los métodos de un paso que son los que hemos visto hasta ahora.

Estos métodos son muy usados en la práctica y tienen muchas virtudes como el hecho de que no hay que realizar evaluaciones anidadas de la función pendiente, que el error local es fácil de calcular y además de forma explícita. Pero también tienen la desventaja de que no todos son 0-estables por construcción como ocurre con los métodos de un paso. Por lo tanto, la consistencia aquí no es garantía de convergencia.

Los métodos multipaso fueron desarrollados por **John Couch Adams** en torno a 1883 para resolver un problema sobre movimiento de fluidos en contra de la gravedad estudiado por **Francis Bashforth**. Adams observó que puede ser útil retener y usar información anterior para aumentar con facilidad el orden de convergencia mediante polinomios de interpolación. Otro investigador importante fue también **Forest Ray Moulton** quien los usó de forma más extensa en torno a 1926 en sus estudios de balística.

En general, este proceso se tiene que hacer con cuidado puesto que va contra la naturaleza del problema continuo que es un problema de valor inicial: para cada tiempo t_* sólo y únicamente con el valor de $y(t_*)$ ya está determinada de manera única la solución para tiempo futuro $t > t_*$. Por lo tanto, introducir valores pasados a t_* produce una perturbación en la solución calculada y es esta perturbación la que debe de estar bajo control.

Existen varias formas de proceder, pero todas se basan en construir polinomios de interpolación para algunos de los datos ya obtenidos: Por ejemplo, se pueden interpolar

- **Métodos de Adams:** algunos de los valores previos $f_j = f(t_j, y_j)$ ya obtenidos (**cálculo explícito**) o también incluyendo el que se busca (**cálculo implícito**).
- **Métodos BDF (Backwards Differentiation Formulae):** algunos de los valores previos y_j ya obtenidos (**cálculo explícito**) o también incluyendo el que se busca (**cálculo implícito**).
- A continuación se deriva o integra el polinomio de interpolación para generar la secuencia de valores.

Planteamos el problema de Cauchy en su versión integral

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(s, y(s)) ds.$$

Una vez conocidos k valores iniciales y_0, y_1, \dots, y_{k-1} en los pasos de tiempo t_0, t_1, \dots, t_{k-1} podemos considerar $P(s)$ un polinomio de interpolación que interpole en los puntos t_j para dar un valor aproximado en t_n a cualquier cantidad, $y(t_n)$ o $f(t_n, y(t_n))$. Por ejemplo, si usamos

$$P(t_n) = f(t_n, y_n), \quad n = 0, 1, 2, \dots, k-1$$

podemos entonces aproximar el valor y_k mediante la expresión

$$y_k = y_{k-1} + \int_{t_{k-1}}^{t_k} P(s) ds.$$

A continuación se traslada el proceso un índice para así obtener y_{k+1} , luego y_{k+2} , etc... Estos son lo que se conoce como **métodos de Adams explícitos** de k pasos. Cuando usamos los valores $y_0, y_1, \dots, y_{k-1}, y_k$, incluyendo el propio valor que buscamos, y pedimos que $P(t)$ interpole en los puntos t_j , esto es

$$P(t_j) = f(t_j, y_j), \quad j = 0, 1, 2, \dots, k-1, k$$

se genera lo que se conoce como **métodos de Adams implícitos** de $k+1$ pasos.

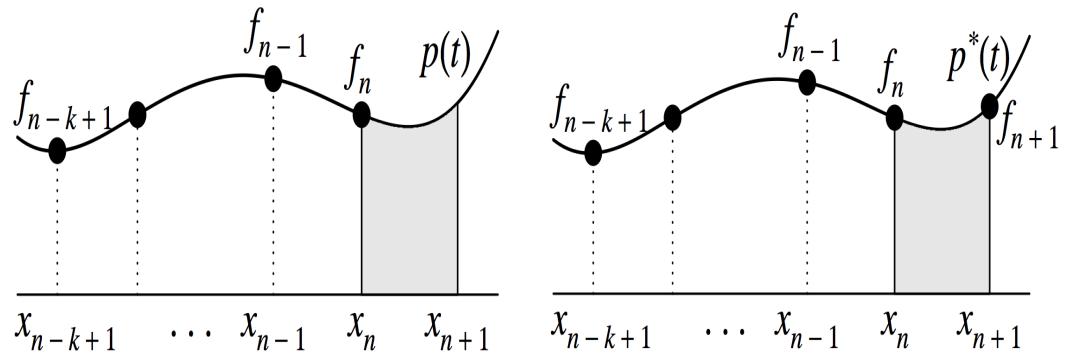


Figura 5.1: Interpretación geométrica en los métodos de Adams explícito e implícito.

Observación 114 Sabemos que la interpolación global, cuando se hace con respecto a muchos puntos, es decir, k grande, puede generar oscilaciones internas del polinomio (recordemos el fenómeno de Runge). Por consiguiente no es habitual tener valores de k muy altos.

Notación: Si k indica el número de puntos que se usan para construir el polinomio de interpolación. Entonces

- Los métodos de tipo explícito se denominan de **Adams-Bashforth** y se pueden denotar por AB k donde k suele coincidir con el número de puntos usados previos a y_n para obtener y_n sin incluir y_n .
- Los métodos implícitos se denominan de **Adams-Moulton**, se pueden denotar por AM k y aquí k suele coincidir con el número de puntos usados previos a y_n para obtener y_n incluyendo el propio y_n .

Observación 115 Los métodos implícitos de **Adams-Moulton** fueron desarrollados también por Adams pero fue **Forest Ray Moulton** quien los usó de forma más extensa en torno a 1926 al observar que se podían combinar con los explícitos. Estos últimos predicen un valor y los implícitos lo corrigen. Las aplicaciones principales de Moulton fueron los estudios de balística.

Se suele decir que los métodos multipaso tienen memoria puesto que usan información de tiempos pasados, en contraposición con aquellos métodos que sólo usan información del paso previo, como por ejemplo los Runge-Kutta. Por esta razón se dice que los métodos de un paso no tienen memoria.

Observación 116 Mientras los métodos de Runge-Kutta aumentan el orden usando la nolinealidad los métodos multipaso usan información previa.

5.2. Ejemplos

Una vez conocidos k valores iniciales y_0, y_1, \dots, y_{k-1} y también f_0, f_1, \dots, f_{k-1} en los pasos de tiempo t_0, t_1, \dots, t_{k-1} , buscamos en t_k el nuevo valor y_k . Podemos hacer lo siguiente:

5.2.1. Interpolar pendientes $f_j = f(t_j, y_j)$ previas

Métodos de Adams explícitos

Consideramos $P(s)$ un polinomio de grado $k-1$ que **interpole las k pendientes calculadas**, esto es,

$$P(t_j) = f(t_j, y_j) = f_j, \quad j = 0, 1, 2, \dots, k-1$$

Tendremos

$$P(s) = \sum_j f_j L_j(s) \in \mathbb{P}^{k-1}$$

para las funciones de base de Lagrange correspondientes. Obtenemos entonces el valor y_k mediante la expresión

$$y_k = y_{k-1} + \int_{t_{k-1}}^{t_k} P(s) ds$$

que da

$$y_k = y_{k-1} + \sum_{j=0}^{k-1} \tilde{c}_j f(t_j, y_j)$$

donde

$$\tilde{c}_j = \int_{t_{k-1}}^{t_k} L_j(s) ds.$$

Se puede ver que los coeficientes \tilde{c}_j no dependen del intervalo $[t_{k-1}, t_k]$, tienen la forma $\tilde{c}_j = h c_j$ (cambio en la variable de integración al $[0, 1]$) y una vez computados sirven para cualquier otro intervalo. A continuación se traslada el proceso un índice para así obtener y_{k+1} , luego y_{k+2} , etc... En este caso se genera lo que se conoce como métodos de **Adams explícitos** de k pasos y se describen por:

Dados y_0, y_1, \dots, y_{k-1} obtener

$$y_{k+n} = y_{k+n-1} + h \sum_{j=0}^{k-1} c_j f(t_{j+n}, y_{j+n}), \quad n \geq 0$$

o simplificando notación

$$y_{n+1} = y_n + h \sum_{j=0}^{k-1} c_j f_{n-j}, \quad n \geq k-1.$$

Métodos de Adams implícitos

Cuando para $k \geq 1$ usamos los valores $y_0, y_1, \dots, y_{k-1}, y_k$, **incluyendo y_k que es el propio valor que buscamos** y usamos un polinomio $P(t) \in \mathbb{P}^k$ que interpole en los puntos t_j los valores f_j para $j = 0, 1, 2, \dots, k$, esto es

$$P(t_j) = f(t_j, y_j), \quad j = 0, 1, 2, \dots, k - 1, k$$

se genera lo que se conoce como métodos de **Adams implícitos** de $k + 1$ pasos.

$$y_k = y_{k-1} + \sum_{j=0}^k \tilde{d}_j f(t_j, y_j)$$

para otros coeficientes distintos

$$\tilde{d}_j = \int_{t_{k-1}}^{t_k} L_j(s) ds.$$

También aquí se puede ver que los coeficientes \tilde{d}_j no dependen del intervalo $[t_{k-1}, t_k]$, tienen la forma $\tilde{d}_j = h d_j$ (cambio en la variable de integración al $[0, 1]$) y una vez computados sirven para cualquier otro intervalo. A continuación se traslada el proceso un índice para así obtener y_{k+1} , luego y_{k+2} , etc...se describen por:

Dados y_0, y_1, \dots, y_{k-1} obtener

$$y_{n+1} = y_n + h \sum_{j=0}^k d_j f_{n+1-j}, \quad n \geq k - 1.$$

Observación 117 En el primer caso integramos $P(t)$ fuera de su intervalo de interpolación; de forma natural se entiende que la aproximación no será muy precisa al realizar una extrapolación. En el segundo caso, este defecto se corrige a cambio de ser implícito, esto es, de introducir el valor de y_k en ambos miembros de la igualdad.

Algunos ejemplos clásicos son (usamos sólo tres puntos como mucho por simplificar)

- **Adams-Bashforth explícito (AB1)**, o Euler explícito:

Dado y_0 calcular para $n \geq 0$

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

aquí $P(s) \equiv f(t_n, y_n) = f_n$.

- **Adams-Bashforth explícito (AB2)**:

Dados y_0, y_1 calcular para $n \geq 1$

$$y_{n+1} = y_n + h \left\{ \frac{3}{2} f(t_n, y_n) - \frac{1}{2} f(t_{n-1}, y_{n-1}) \right\}$$

aquí $P(s)$ interpola a f_n y a f_{n-1} .

■ **Adams-Bashforth explícito (AB3):**

Dados y_0, y_1, y_2 calcular para $n \geq 2$

$$y_{n+1} = y_n + h \left\{ \frac{23}{12} f(t_n, y_n) - \frac{16}{12} f(t_{n-1}, y_{n-1}) + \frac{5}{12} f(t_{n-2}, y_{n-2}) \right\}$$

aquí $P(s)$ interpola a f_n, f_{n-1} y a f_{n-2} .

■ **Adams-Moulton implícito (AM1)**, o método de Euler implícito:

Dados y_0 calcular para $n \geq 0$

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

aquí $P(s) \equiv f(t_{n+1}, y_{n+1}) = f_{n+1}$.

■ **Adams-Moulton implícito (AM2)**, también conocido como Crank-Nicolson

o regla del trapecio:

Dado y_0 calcular para $n \geq 0$

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

aquí $P(s)$ interpola a f_{n+1} y a f_n .

■ **Adams-Moulton implícito (AM3):**

Dados y_0, y_1 calcular para $n \geq 1$

$$y_{n+1} = y_n + h \left\{ \frac{5}{12} f(t_{n+1}, y_{n+1}) + \frac{8}{12} f(t_n, y_n) - \frac{1}{12} f(t_{n-1}, y_{n-1}) \right\}$$

aquí $P(s)$ interpola a f_{n+1}, f_n y f_{n-1} .

5.2.2. Métodos BDF (Backward Differentiation formulas): interpolando las aproximaciones y_j

Podemos tomar $Q(s)$ un polinomio de grado k que interbole los k valores y_j previos calculados junto con el buscado y_k , esto es,

$$Q(t_j) = y_j, \quad j = 0, 1, 2, \dots, k,$$

entonces

$$Q(s) = \sum_{j=0}^k y_j L_j(s) \in \mathbb{P}^k$$

para las funciones de base de Lagrange correspondientes. Entonces aproximamos

$$y'(s) \sim Q'(s)$$

y obtener el valor y_k mediante la expresión

$$Q'(t_{k-1}) = f(t_{k-1}, y_{k-1}) \quad \text{o bien} \quad Q'(t_k) = f(t_k, y_k).$$

Esto da un método explícito si usamos

$$\sum_{j=0}^k y_j L'_j(t_{k-1}) = f(t_{k-1}, y_{k-1})$$

o uno implícito si usamos

$$\sum_{j=0}^k y_j L'_j(t_k) = f(t_k, y_k).$$

Igual que antes, tenemos

$$\sum_{j=0}^k y_j \tilde{r}_j = f(t_k, y_k) \quad \text{o} \quad \sum_{j=0}^k y_j \tilde{r}_j = f(t_{k-1}, y_{k-1})$$

para algunos coeficientes \tilde{r}_j , que se puede comprobar son de la forma $\tilde{r}_j = r_j h^{-1}$ para r_j independiente del intervalo $[t_{k-1}, t_k]$, para llegar a

$$\sum_{j=0}^k y_j r_j = h f(t_k, y_k) \quad \text{o} \quad \sum_{j=0}^k y_j r_j = h f(t_{k-1}, y_{k-1}).$$

Ejemplo 122 Si queremos usar y_{n-1}, y_n, y_{n+1} tenemos

$$\begin{aligned} Q(s) &= y_{n-1} \frac{(s - t_n)(s - t_{n+1})}{(t_{n-1} - t_n)(t_{n-1} - t_{n+1})} + y_n \frac{(s - t_{n-1})(s - t_{n+1})}{(t_n - t_{n-1})(t_n - t_{n+1})} \\ &+ y_{n+1} \frac{(s - t_n)(s - t_{n-1})}{(t_{n+1} - t_n)(t_{n+1} - t_{n-1})} \end{aligned}$$

o bien

$$Q(s) = y_{n-1} \frac{(s - t_n)(s - t_{n+1})}{2h^2} + y_n \frac{(s - t_{n-1})(s - t_{n+1})}{-h^2} + y_{n+1} \frac{(s - t_n)(s - t_{n-1})}{2h^2}$$

de donde, por ejemplo, la derivada es

$$Q'(s) = y_{n-1} \frac{2s - (t_n + t_{n+1})}{2h^2} + y_n \frac{2s - (t_{n-1} + t_{n+1})}{-h^2} + y_{n+1} \frac{2s - (t_n + t_{n-1})}{2h^2}.$$

Con $Q'(t_n) = f(t_n, y_n)$ tenemos el método

$$y_{n+1} - y_{n-1} = 2h f_n.$$

Con $Q'(t_{n+1}) = f(t_{n+1}, y_{n+1})$ (normalizando a uno el coeficiente de y_{n+1})

$$y_{n+1} - \frac{4}{3}y_n + \frac{1}{3}y_{n-1} = \frac{2}{3}h f_{n+1}$$

Ejemplo 123 Supongamos que $Q(s)$ cumple

$$Q(t_{n+1}) = y_{n+1}, \quad Q(t_n) = y_n.$$

La ecuación para $Q(s)$ se deduce usando la interpolación de Lagrange y es

$$Q(s) = y_n \frac{s - t_{n+1}}{t_n - t_{n+1}} + y_{n+1} \frac{s - t_n}{t_{n+1} - t_n}$$

o bien, como $t_{n+1} - t_n = h$ también

$$Q(s) = y_n \frac{s - t_{n+1}}{-h} + y_{n+1} \frac{s - t_n}{h}$$

Derivando, tenemos

$$Q'(s) = y_n \frac{1}{-h} + y_{n+1} \frac{1}{h}$$

Entonces $Q'(t_{n+1}) = f(t_{n+1}, y_{n+1})$ genera

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

que coincide con Euler implícito.

Observación 118 Desde el punto de vista de la programación en el computador resultan ser una iteración a partir de varios valores previos, o iteración de varios pasos. En el caso explícito tenemos algo como

$$z_{n+k} = \Phi_h(z_n, z_{n+1}, z_{n+2}, \dots, z_{n+k-1}), \quad n = 0, 1, 2, \dots$$

mientras que en el implícito hay que resolver una ecuación no lineal (en general)

$$z_{n+k} = z_\star, \quad \text{donde } z_\star = \Phi_h(z_n, z_{n+1}, z_{n+2}, \dots, z_{n+k-1}, z_\star) \quad n = 0, 1, 2, \dots$$

Veamos un ejemplo donde la construcción del método nos lleva a un esquema de cálculo inestable

Ejemplo 124 Inestabilidad numérica: Sabemos que:

$$\frac{y(t+h) - y(t)}{h} = y'(t) + O(h), \quad h \rightarrow 0$$

pero podemos buscar otro desarrollo de Taylor para $y'(t)$ con un menor error local. Por ejemplo, si nos apoyamos en los puntos $t - 2h, t - h$ y $t + h$ podemos escribir

$$\begin{aligned} y(t+h) &= y(t) + hy'(t) + \frac{h^2}{2}y''(t) + \frac{h^3}{3!}y'''(t) + \frac{h^4}{4!}y''''(t) + \dots \\ y(t-h) &= y(t) - hy'(t) + \frac{h^2}{2}y''(t) - \frac{h^3}{3!}y'''(t) + \frac{h^4}{4!}y''''(t) + \dots \\ y(t-2h) &= y(t) - 2hy'(t) + 2^2 \frac{h^2}{2}y''(t) - 2^3 \frac{h^3}{3!}y'''(t) + 2^4 \frac{h^4}{4!}y''''(t) + \dots \end{aligned}$$

una combinación lineal de coeficientes a, b, c nos da

$$\begin{aligned} a y(t+h) + b y(t-h) + c y(t-2h) &= (a+b+c)y(t) + (a-b-2c)hy'(t) \\ &+ (a+b+4c)\frac{h^2}{2}y''(t) + (a-b-8c)\frac{h^3}{3!}y'''(t) \\ &+ (a+b+16c)\frac{h^4}{4!}y^{(4)}(t) + \dots \end{aligned}$$

Eso decir,

$$\begin{aligned} a y(t+h) + b y(t-h) + c y(t-2h) - (a+b+c)y(t) &= (a-b-2c)hy'(t) \\ &+ (a+b+4c)\frac{h^2}{2}y''(t) + (a-b-8c)\frac{h^3}{3!}y'''(t) \\ &+ (a+b+16c)\frac{h^4}{4!}y^{(4)}(t) + \dots \end{aligned}$$

Una elección de parámetros tal que

$$\begin{aligned} a - b - 2c &= \lambda \neq 0 \\ a + b + 4c &= 0 \\ a - b - 8c &= 0 \end{aligned}$$

nos deja con

$$\begin{aligned} a y(t+h) + b y(t-h) + c y(t-2h) - (a+b+c)y(t) &= \lambda hy'(t) \\ &+ (a+b+16c)\frac{h^4}{4!}y^{(4)}(t) + \dots \end{aligned}$$

lo que va a generar un orden local 4 a la hora de aproximar $y'(t)$. Usando λ como referencia, tenemos que

$$a - b - 8c = 0 \Rightarrow \lambda - 6c = 0 \Rightarrow c = \lambda/6$$

usando que

$$a - b - 8c = 0 = a + b + 4c \Rightarrow 2b = -12c \Rightarrow b = -6c = -\lambda$$

y finalmente

$$a = \lambda/3$$

y además

$$a + b + c = \lambda/3 - \lambda + \lambda/6 = -\lambda/2$$

$$a + b + 16c = \lambda/3 - \lambda + 16\lambda/6 = 2\lambda$$

luego tenemos el esquema

$$\lambda y(t+h)/3 - \lambda y(t-h) + \lambda y(t-2h)/6 + \lambda y(t)/2 = \lambda hy'(t) + 2\lambda \frac{h^4}{4!}y^{(4)}(t) + \dots$$

Vemos que λ se puede simplificar y queda en la parte del error para dejarnos el esquema

$$\frac{y(t+h)/3 + y(t)/2 - y(t-h) + y(t-2h)/6}{6h} = -y'(t) + 2\frac{h^4}{4!}y^{(4)}(t) + \dots,$$

Dados y_0, y_1, y_2 este esquema trabaja como

$$\frac{1}{3}y_{n+1} + \frac{1}{2}y_n - y_{n-1} + \frac{1}{6}y_{n-2} = hf(t_n, y_n), \quad n \geq 2$$

y tiene un error local de orden 4. Necesitamos tres valores para empezar, si lo aplicamos al ejemplo

$$y'(t) = 2t(1+y(t)^2), \quad y(0) = 0$$

cuya solución es $y(t) = \tan(t^2)$ y empezamos con valores exactos nos encontramos con la situación de la Figura 5.2 y vemos que reducir el valor de h sólo empeora la situación por lo que a pesar de tener un error local de orden 4 no es un método convergente, nos falla la estabilidad a cero del método.

Esto es debido a que el polinomio $p(z) = \frac{1}{3}z^3 + \frac{1}{2}z^2 - z + \frac{1}{6} = (z-1)(z^2 + \frac{5}{2}z - \frac{1}{2})$ asociado al uso de los valores y_{n+1}, y_n, y_{n-1} e y_{n-2} tiene raíces de módulo mayor que uno. Sabiendo que $r_1 = 1$ es una raíz el resto se obtienen con facilidad y son $r_2 \approx -2.686\dots$ y $r_3 \approx 0.186\dots$. Se reproduce el mismo fenómeno con la simple $y'(t) = 0$ con $y(0) = 1$ si tomamos $y_0 = y_1 = 1$ e $y_2 = 1 + 10^{-13}$ por ejemplo. Luego no es debido a la ecuación diferencial ordinaria que usamos sino al esquema numérico. Básicamente se necesita que las raíces de este polinomio asociado al uso de los datos previos sean de módulo menor que uno.

Si consideramos el esquema implícito

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

y usamos una ecuación lineal, para que no haya que resolver un problema de punto fijo, podemos observar que se tiene el mismo efecto. Luego el problema proviene de la elección de coeficientes para los valores y_{n+1}, y_n, y_{n-1} e y_{n-2} .

Más ejemplos de métodos multipaso

- **Adams-Basforth explícito (AB4)** de cuatro pasos:

Dados y_0, y_1, y_2, y_3 calcular para $n \geq 3$

$$y_{n+1} = y_n + \frac{h}{24} \{ 55f(t_n, y_n) - 59f(t_{n-1}, y_{n-1}) + 37f(t_{n-2}, y_{n-2}) - 9f(t_{n-3}, y_{n-3}) \}$$

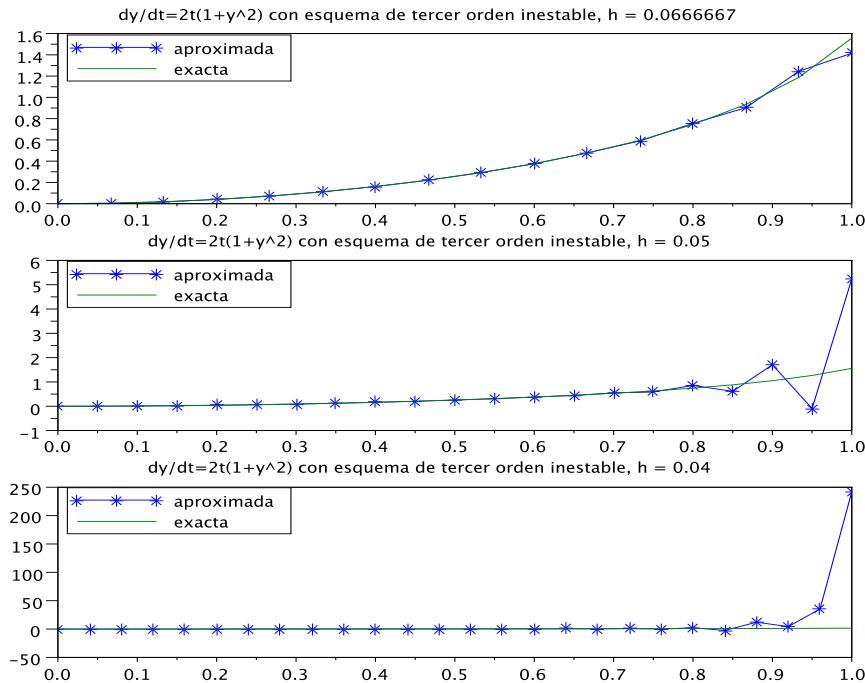


Figura 5.2: Método con error local de orden 3 pero inestable, reducir h incrementa la inestabilidad.

- **Método BDF (Backward Differentiation formulas) de dos pasos (BDF2):**
Dados y_0, y_1 calcular para $n \geq 1$

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

- **Método BDF de tres pasos (BDF3):**
Dados y_0, y_1, y_2 calcular para $n \geq 2$

$$\frac{11}{6}y_{n+1} - 3y_n + \frac{3}{2}y_{n-1} - \frac{1}{3}y_{n-2} = hf(t_{n+1}, y_{n+1})$$

Observar que es distinto del método usado en el ejemplo último.

- **Método de Quade:** dados y_0, y_1, y_2, y_3 obtener

$$y_{n+4} - \frac{8}{19}(y_{n+3} - y_{n+1}) - y_n = \frac{6}{19}h[f_{n+4} + 4f_{n+3} + 4f_{n+1} + f_n], \quad n \geq 0.$$

5.3. Forma general de los métodos multipaso

Denotaremos por y_n a una aproximación a la solución $y(t_n)$

$$y_n \approx y(t_n)$$

y similarmente podemos denotar f_n a una aproximación a $f(t_n, y(t_n))$ en general de la forma: $f_n = f(t_n, y_n)$. Nuestros objetivos son conseguir y_1, y_2, \dots, y_N tales que

- Para cada $t_\star \in [0, T]$ fijo, con $t_\star = t_n = t_0 + nh$, y_n aproxima a $y(t_n)$ cuando $h \rightarrow 0$, $n \rightarrow +\infty$ tal que $t_0 + nh = t_\star$.

La expresión general de un método de k pasos para una ecuación diferencial ordinaria es la siguiente:

Dados y_0, y_1, \dots, y_{k-1} para $n \geq 0$ calcular y_{n+k} donde

$$y_{n+k} + \sum_{j=0}^{k-1} a_j y_{n+j} = h \sum_{j=0}^k b_j f(t_{n+j}, y_{n+j}), \quad n \geq 0.$$

Observación 119 *Hemos normalizado coeficientes para evitar arbitrariedades. Por lo tanto, hemos fijado $a_k = 1$ y suponemos $a_0^2 + b_0^2 \neq 0$ para no estar en la situación de un método de $k - 1$ pasos.*

Observación 120 *El método es explícito si $b_k = 0$ e implícito si $b_k \neq 0$. En efecto, si $b_k = 0$ entonces y_{n+k} se obtiene de forma explícita de los valores y_{n+j} para $j = 0, 1, 2, \dots, k - 1$ y por lo tanto el método lineal de k pasos (5.3) es explícito mientras que si $b_k \neq 0$ entonces es implícito y hay que resolver una ecuación en principio no lineal con la típica restricción*

$$h|b_k L_f| \leq 1.$$

Por ejemplo con AM4 (trasladamos el comienzo a t_n) dados y_0, y_1, y_2 calculamos para $n \geq 0$

$$y_{n+3} = y_{n+2} + h \left\{ \frac{9}{24} f(t_{n+3}, y_{n+3}) + \frac{19}{24} f(t_{n+2}, y_{n+2}) - \frac{5}{24} f(t_{n+1}, y_{n+1}) + \frac{1}{24} f(t_n, y_n) \right\}$$

entonces $a_2 = -1$ y $a_1 = a_0 = 0$ mientras que

$$\begin{aligned} \Phi_f(t_n, y_n, y_{n+1}, y_{n+2}, y_{n+3}; h) &= \frac{1}{24} f(t_n, y_n) - \frac{5}{24} f(t_{n+1}, y_{n+1}) + \frac{19}{24} f(t_{n+2}, y_{n+2}) \\ &\quad + \frac{9}{24} f(t_{n+3}, y_{n+3}). \end{aligned}$$

y recordamos que $t_{n+j} = t_n + j h$ para $j \in \mathbb{Z}$ por lo que la dependencia de Φ_f en los argumentos $(t_n, y_n, y_{n+1}, y_{n+2}, y_{n+3}, h)$ a través de f está bien justificada, es decir, no es necesario explicitar los argumentos t_{n+1} , t_{n+2} y t_{n+3} .

Observación 121 *El método se dice lineal porque la expresión que aparece es lineal en los valores f_j . Esta linealidad del método no tiene nada que ver con f que no tiene porque ser lineal.*

Observación 122 Hemos reemplazado $y'(t) = f(t, y(t))$ por la expresión

$$\frac{y_{n+k} + \sum_{j=0}^{k-1} a_j y_{n+j}}{h} = \sum_{j=0}^k b_j f(t_{n+j}, y_{n+j}), \quad n \geq 0.$$

construida a base de valores puntuales (t_j, y_j) y evaluaciones $f_j = f(t_j, y_j)$, que en el fondo es de lo único que tenemos. Luego queremos

$$\frac{y_{n+k} + \sum_{j=0}^{k-1} a_j y_{n+j}}{h} \approx y'(t_\star) \quad (5.1)$$

y también

$$\sum_{j=0}^k b_j f(t_{n+j}, y_{n+j}) \approx f(t_\star, y(t_\star)) \quad (5.2)$$

donde $t_\star = t_n$ o $t_\star = t_{n+1}$.

5.3.1. Consecuencias sobre los coeficientes a_j

Nuestro esquema general **debe reproducir casos particulares simples**. Por ejemplo,

$$\begin{cases} y'(t) &= 0, \quad t > 0 \\ y(0) &= 1 \end{cases}$$

con solución $y(t) \equiv 1$ para todo $t > 0$. En este caso $f \equiv 0$ luego $\sum_{j=0}^k b_j f_{n+j} \equiv 0$ y del esquema sólo queda: Dados y_0, y_1, \dots, y_{k-1}

$$y_{n+k} = -\sum_{j=0}^{k-1} a_j y_{n+j}, \quad n \geq 0;$$

es decir, una **ecuación en diferencias**: para y_0, y_1, \dots, y_{k-1} con $n \geq 0$ obtener y_{n+k} tal que

$$y_{n+k} + a_{k-1} y_{n+k-1} + a_{k-2} y_{n+k-2} + \dots + a_0 y_0 = 0.$$

Observación 123 En el caso de un sólo paso con $y_0 = 1$ tenemos $a_0 = -1$, esto es decir que los **métodos de un paso** deben escribirse en forma general como

$$y_{n+1} = y_n + h \{b_0 f(t_n, y_n) + b_1 f(t_{n+1}, y_{n+1})\}, \quad n \geq 0$$

y aquí vemos que tenemos el esquema de Crank-Nicolson como un ejemplo práctico.

Para obtener la solución general de esta ecuación en diferencias es útil asociar un polinomio a esta ecuación, llamado **primer polinomio característico**, definido por

$$\rho(z) = z^k + a_{k-1}z^{k-1} + a_{k-2}z^{k-2} + \dots + a_0$$

(en el caso de un paso es $\rho(z) = z - 1$). Si $y_0 = y_1 = \dots = y_{k-1} = 1$ se debe garantizar $y_k = 1$, entonces debe ser

$$y_k = - \sum_{j=0}^{k-1} a_j y_j = - \sum_{j=0}^{k-1} a_j$$

o lo que es lo mismo

$$\rho(1) = 1 + a_{k-1} + a_{k-2} + \dots + a_0 = 0.$$

Po lo tanto, $\rho(z)$ debe de tener a $z = 1$ como raíz. Esta raíz se llama **raíz principal del primer polinomio característico**.

5.3.2. Breve sobre ecuaciones en diferencias

Una sucesión de números $\{y_n\}_n \subset \mathbb{R}$ se puede construir de acuerdo a unas determinadas leyes o ecuaciones, que se denominan ecuaciones en diferencias.

Ejemplo 125 La más conocida es la ecuación de Fibonacci, que es de segundo orden

$$\begin{cases} y_0 = 1, \\ y_1 = 1, \\ y_{n+1} = y_n + y_{n-1}, \quad n \geq 1 \end{cases}$$

Es fácil motivar este tipo de ecuaciones desde el modelado de fenómenos estacionales, donde por estación entendemos cualquier periodo de tiempo discreto, ya sea un segundo, un minuto, un día, año, etc...

Ejemplo 126 **Modelo de poblaciones estacional** Supongamos que tenemos una colonia de bacterias y en cada estación la colonia multiplica por un número $r > 0$ su población. Entonces si B_0 viene dado y B_n es el tamaño de la población en la estación n tenemos

$$B_{n+1} = r B_n, \quad n \geq 0 \Rightarrow B_{n+1} = r^{n+1} B_0$$

que indica decaimiento si $0 < r < 1$ y crecimiento si $r > 1$. En general se puede tener una relación no lineal $B_{n+1} = f(B_n)$ más complicada.

El nombre de ecuación en diferencias proviene de la posibilidad de estudiar el incremento en vez de el nuevo valor y escribir la ecuación en la forma

$$B_{n+1} - B_n = g(B_n).$$

¿Cómo ir de una ecuación en diferencias a una ecuación diferencial? Se ve claro que se puede entender como un cociente incremental en donde la unidad de tiempo es una estación y poniendo $B_n = B(t_n)$ podemos escribir

$$B_{n+1} - B_n = \frac{B_{n+1} - B_n}{1} = \frac{B(t_n + \Delta t) - B(t_n)}{\Delta t} = g(B(t_n)) \quad (\Delta t = 1)$$

Cuando no existe un periodo de tiempo característico y deseamos ver un proceso continuo podemos pasar a la derivada usando el límite para $\Delta t \rightarrow 0$. Suponiendo que $g(B(t_n))$ no dependa de Δt nos encontramos con el modelo continuo

$$B'(t) = g(B(t)).$$

Por ejemplo, en el caso $B_{n+1} = r B_n$ tenemos $B(t_n + \Delta t) - B(t_n) = (r - 1)B(t_n)$ y se genera la ecuación diferencial básica

$$B'(t) = (r - 1)B(t)$$

con solución $B(t) = e^{(r-1)t}B_0$ y que indica decrecimiento si $0 < r < 1$ (exponente negativo) y crecimiento si $r > 1$ (exponente positivo).

Ejemplo 127 Modelo de poblaciones de Leslie (1945) Supongamos que tenemos una población estructurada por m grupos de edades. Entonces si denotamos por y_j^n la componente j del vector y^n podemos usar y_j^n como la cantidad de individuos de grupo de edad j que hay en la población en la estación n . Por ejemplo, y_1^n son los individuos de edad menor que 1, y_2^n son los individuos de edad mayor que 1 y menor que 2, etc... y así hasta llegar a y_m^n que representa el número de individuos de edad entre $m - 1$ y m , pudiendo suponer que ningún individuo vive más de m estaciones. Si suponemos ahora que la población cambia sólo por nacimientos y muertes y ambos fenómenos dependen sólo de la edad de los individuos nos encontramos con unas ecuaciones que pueden tener la forma siguiente:

Muertes: El tanto por ciento de individuos que superan una estación es α_j entonces tendremos las ecuaciones

$$y_{j+1}^{n+1} = \frac{\alpha_j}{100} y_j^n, \quad j = 1, 2, \dots, m - 1$$

Nacimientos: El tanto por ciento de individuos que se reproducen en cada grupo de edad lo llamamos β_j entonces nos encontramos con la ecuación

$$y_1^{n+1} = \frac{\beta_1}{100} y_1^n + \frac{\beta_2}{100} y_2^n + \dots + \frac{\beta_m}{100} y_m^n.$$

Recolectando todas las ecuaciones tenemos una ecuación de la forma

$$y^{n+1} = Ly^n$$

donde L es la matriz de Leslie que está formada por los α_j y los β_j . Además, claramente

$$y^{n+1} = L^{n+1}y^0.$$

Como las propiedades de los autovalores y autovectores de L determinan el comportamiento de L^n podemos deducir su importancia para saber el comportamiento de la población.

Ejercicio 128 Construir la matriz de Leslie.

En general, podemos plantear (a veces se justifican como modelos aplicados o simplemente como ecuaciones matemáticas) la ecuación:

Definición 129 Ecuación en diferencias de orden k . Para y_0, y_1, \dots, y_{k-1} y dados φ_n obtener y_{n+k} tal que

$$y_{n+k} + a_{k-1}y_{n+k-1} + a_{k-2}y_{n+k-2} + \dots + a_0y_n = \varphi_n, \quad \forall n \geq 0.$$

La forma de resolver este tipo de ecuaciones es similar a como se resuelven las ecuaciones diferenciales ordinarias lineales de coeficientes constantes:

1. Buscamos una solución particular y_n^P del problema no homogéneo ($\varphi_n \neq 0$)
2. Buscamos la solución general y_n^H del problema homogéneo ($\varphi_n = 0$)
3. Entonces nuestra solución buscada del problema general es

$$y_n = y_n^H + y_n^P.$$

Observación 124 Obtener una solución particular del problema no homogéneo no es fácil. Sólo en el caso donde $\varphi_n = \varphi$ constante podemos usar $y_n^P = c$ constante dada por

$$c = \frac{\varphi}{1 + a_{k-1} + a_{k-2} + \dots + a_0} = \frac{\varphi}{\rho(1)}$$

para lo que necesitamos $\rho(1) \neq 0$.

Nosotros trabajaremos con el problema homogéneo ($\varphi_n = 0$) ya que normalmente es nuestro caso. Por lo tanto nos centramos en resolver la ecuación:

Para y_0, y_1, \dots, y_{k-1} con $n \geq 0$ obtener y_{n+k} tal que

$$y_{n+k} + a_{k-1}y_{n+k-1} + a_{k-2}y_{n+k-2} + \dots + a_0y_n = 0. \quad (5.3)$$

La forma de buscar una solución del problema homogéneo es suponerla de la forma $y_n = r^n$ para algún número r . Entonces llegamos a la necesidad de encontrar los ceros del polinomio característico asociado

$$\rho(z) = z^k + a_{k-1}z^{k-1} + a_{k-2}z^{k-2} + \dots + a_0.$$

Si tenemos k raíces distintas entonces la solución general es una combinación lineal de potencias de estas raíces

$$y_n = \sum_j d_j r_j^n$$

donde d_j son coeficientes a determinar con los datos iniciales. Si hay raíces con multiplicidad mayor que uno, para cada raíz aparece un polinomio en n asociado de grado la multiplicidad menos 1. Es decir, si tenemos m raíces distintas r_1, r_2, \dots, r_m y cada una de ellas con multiplicidad $\mu_1, \mu_2, \dots, \mu_m$, entonces la forma general de la solución es

$$y_n = \sum_{j=1}^m p_j(n) r_j^n.$$

donde cada $p_j(n)$ es un polinomio de grado $\mu_j - 1$.

Ejemplo 130 Para y_0, y_1 con $n \geq 0$ obtener y_{n+2} tal que

$$y_{n+2} + a_1 y_1 + a_0 y_0 = 0$$

donde a_1 y a_0 son reales. El polinomio característico es

$$\rho(z) = z^2 + a_1 z + a_0$$

y si las raíces son r_1, r_2 tenemos las siguientes situaciones:

1. $r_1 \neq r_2$ reales o complejas. Entonces

$$y_n = Ar_1^n + Br_2^n, \quad n \geq 0$$

donde A y B se determinan usando y_0 e y_1 , esto es

$$y_0 = A + B, \quad y_1 = Ar_1 + Br_2.$$

En el caso complejo, siendo r_1 y r_2 conjugadas, y_n será real.

2. $r_1 = r_2 = r$ real. Entonces

$$y_n = Ar^n + Bnr^n, \quad n \geq 0$$

donde A y B se determinan usando y_0 e y_1 , esto es

$$y_0 = A, \quad y_1 = Ar + Br = (A + B)r.$$

Ya hemos visto que el primer polinomio característico $\rho(z)$ debe de tener $z = 1$ como raíz, esta raíz se llama **raíz principal**. El resto de raíces deben de estar controladas de una forma eficiente (deben tener módulo menor que uno y aquellas de módulo uno deben ser simples) como atestiguan los siguientes ejemplos:

Ejemplo 131 Tomemos $y_0 = 1$ e $y_1 = 1 + \gamma(h)$ donde $\gamma(h)$ es cualquier error que suponemos tiende a cero con h . Si tenemos la recurrencia

$$y_{n+2} - 2y_{n+1} + y_n = 0.$$

El polinomio característico tiene a $r = 1$ por raíz doble. La recurrencia tiene por solución general

$$y_n = Ar^n + Bn r^n, \quad n \geq 0 \quad (r = 1)$$

de donde $y_0^h = A$ y $B = \gamma(h)$ por lo tanto se genera la solución

$$y_n^h = 1 + \gamma(h) n = 1 + \frac{\gamma(h)}{h} hn = 1 + \frac{\gamma(h)}{h} t_n$$

Entonces si $\gamma(h)$ no tiende a cero con suficiente fuerza no se puede controlar el crecimiento de n . Necesitamos $\gamma(h) = h\delta(h)$ con $\delta(h) \rightarrow 0$. Por ejemplo, errores de la forma $\gamma(h) = \sqrt{h}$ o $\gamma(h) = h|\log h|$ no nos sirven. Tampoco nos sirve un error constante, por muy pequeño que sea, por ejemplo, algo de la forma $\gamma(h) = \gamma = 10^{-13}$ genera

$$y_n^h = 1 + 10^{-13} n \rightarrow +\infty, \quad n \rightarrow +\infty.$$

Explicación: Si calculamos en $[t_0, t_0 + T]$ con N puntos y tomando $h = T/N$ los valores los calculamos en los puntos de la partición

$$t_0 = t_0 < t_1^{(h)} < t_2^{(h)} < \dots < t_n^{(h)} < \dots < t_{N-1}^{(h)} < t_N = t_0 + T$$

donde

$$t_n^{(h)} = t_0 + nh.$$

Entonces si h tiende a cero para reducir el error resulta que

$$y_0^{(h)}, y_1^{(h)} \rightarrow y(t_0) = y_0 = 1, \quad h \rightarrow 0^+$$

Por otro lado, el rango de valores posibles para n es creciente desde 0 hasta $N = T/h \rightarrow +\infty$. Esto es, el número de puntos en la partición es cada vez más grande (indexados a través del parámetro $n = 0, 1, 2, \dots, N^{(h)} = T/h$). Luego $h \rightarrow 0$ implica $N^{(h)} \rightarrow +\infty$ y entonces el rango de n será cada vez más grande. Como consecuencia, tarde o temprano hará crecer

$$y_n = y_n^{(h)} = 1 + \gamma n$$

fueras de lo admisible como error. En nuestro ejemplo es $y_0^{(h)} = 1$, que no da problemas, y

$$y_1^{(h)} = 1 + \gamma \rightarrow 1, \quad h \rightarrow 0^+.$$

Luego debe ser $\gamma = \gamma(h) \rightarrow 0$, $h \rightarrow 0^+$ y tal que

$$y_n = y_n^{(h)} = 1 + \gamma(h)n \rightarrow 1, \quad h \rightarrow 0^+.$$

Si tomamos

$$\gamma(h) = h \log(h) \rightarrow 0, \quad h \rightarrow 0^+$$

resulta que

$$\gamma(h)n = t_n \log(h) \rightarrow -\infty, \quad h \rightarrow 0^+$$

Si ahora tenemos la recurrencia

$$y_{n+2} - 3y_{n+1} + 2y_n = 0.$$

Las raíces del polinomio característico son $z = 1$ y $z = 2$ y la solución general es

$$y_n = 1 - \gamma + \gamma 2^n$$

por lo que ahora la amplificación es mucho más rápida.

Definición 132 Polinomio contractivo Es todo polinomio cuyas raíces tienen módulo menor o igual a 1 y aquellas de módulo uno son simples

De la única forma de la que nos libraremos del efecto visto en los ejemplos previos es cuando el polinomio característico

$$\rho(z) = z^k + a_{k-1}z^{k-1} + a_{k-2}z^{k-2} + \dots + a_0.$$

es contractivo.

Observación 125 Se puede interpretar como el precio que hay que pagar por usar la memoria en un proceso que no la necesita. Sólo podemos hacerlo cuando las raíces adicionales que se introducen al método de un paso con polinomio $\rho(z) = z - 1$, tienen un módulo controlado de acuerdo a la condición anterior. Estas raíces se suelen denominar **raíces espúreas** (es decir, falsas, contaminantes, etc...). Lo que se pide es

- $z = 1$ debe ser una raíz simple del polinomio
- no se pueden admitir raíces con módulo mayor que uno
- aquellas de módulo uno deben ser simples.

Esto se resumen en que debemos añadir la siguiente restricción a nuestra forma general:

Definición 133 Condición de Dahlquist: El primer polinomio característico $\rho(z)$ debe ser contractivo y tener $z = 1$ como raíz simple. Esta raíz $z = 1$ se denomina **raíz principal** y el resto de raíces son **raíces contaminantes**.

El siguiente resultado es el clásico y se puede ver en Isaacson-Keller (1966), ya lo hemos comprobado en parte con algunos ejemplos.

Teorema 134 El método general de k pasos es 0-estable sí y sólo sí su primer polinomio característico cumple la condición de Dahlquist.

Estamos exigiendo la condición de Dahlquist sobre $\rho(r)$ y esto equivale a garantizar que el método es 0-estable. La siguiente reescritura de la ecuación en diferencias (5.3) facilita nuestro desarrollo:

Definición 135 Dado el polinomio $\rho(z) = z^k + a_{k-1}z^{k-1} + a_{k-2}z^{k-2} + \dots + a_1z + a_0$ se define su **matriz compañera** como la matriz $A \in \mathbb{R}^{k \times k}$ dada por

$$A = \begin{pmatrix} -a_{k-1} & -a_{k-2} & \dots & \dots & -a_1 & -a_0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix}.$$

Entonces, la ecuación

$$y_{n+k} + \sum_{j=0}^{k-1} a_j y_{n+j} = 0$$

se puede expresar en forma de un paso como sigue:

$$Y_{n+1} = AY_n$$

donde

$$Y_{n+1} = \begin{pmatrix} y_{n+k} \\ y_{n+k-1} \\ \vdots \\ y_{n+1} \end{pmatrix}, \quad Y_n = \begin{pmatrix} y_{n+k-1} \\ y_{n+k-2} \\ \vdots \\ y_n \end{pmatrix}$$

por lo tanto, tenemos que la ecuación en diferencias se resuelve en términos de potencias de A

$$Y_{n+1} = A^n Y_0.$$

La siguiente propiedad, de sencilla comprobación, liga los autovalores de esta matriz con las raíces del polinomio

Lema 3

$$\det(A - \lambda I) = \pm \rho(\lambda)$$

por lo tanto, el espectro de A , es decir, $\sigma(A)$ es el conjunto de las raíces de $\rho(z)$:

$$\sigma(A) = \{r_1, r_2, \dots, r_k\}.$$

donde r_j son las raíces de $\rho(z)$.

Podemos observar entonces que las potencias de A marcan el comportamiento de la solución y es entonces importante tener control sobre los términos $p_j(n)r_j^n$ cuando $n \rightarrow \infty$. Ya que

$$\|A^n\| = \max_{\|V\|=1} \|A^n V\|$$

es trivial obtener bajo la condición de Dahlquist la acotación

$$\|A^n\| \leq M, \quad \forall n \geq 0$$

para cualquier norma matricial inducida.

5.4. Estudio de la consistencia

Como ya sabemos, el **error local** en un punto $t_\star + kh$ es el error cometido en un paso resultante de aplicar el esquema suponiendo que tenemos los valores previos de manera exacta.

Definición 136 Se define el **error local** en un punto t_\star al avanzar a $t_\star + kh$ como ($a_k = 1$, $a_0^2 + b_0^2 \neq 0$)

$$l(y(t_\star); h) = y(t_\star + kh) - \tilde{z}$$

donde

$$\tilde{z} = - \sum_{j=0}^{k-1} a_j y(t_\star + jh) + h \sum_{j=0}^{k-1} b_j f(t_\star + jh, y(t_\star + jh))$$

es decir, usando $f(t_\star + jh, y(t_\star + jh)) = y'(t_\star + jh)$,

$$\tilde{z} = - \sum_{j=0}^{k-1} a_j y(t_\star + jh) + h \sum_{j=0}^{k-1} b_j y'(t_\star + jh)$$

Por lo tanto, de forma simple se puede escribir como

$$l(y(t_\star); h) = y(t_\star + kh) + \sum_{j=0}^{k-1} a_j y(t_\star + jh) - h \sum_{j=0}^{k-1} b_j y'(t_\star + jh)$$

Poniendo $t_{n+j} = t_n + jh$ podemos escribir

Definición 137 El método lineal de k pasos (5.3) tiene orden $p \geq 1$ cuando

$$l(t; h) := \sum_{j=0}^k a_j y(t_{n+j}) - h \sum_{j=0}^k b_j y'(t_{n+j}) = O(h^{p+1}), \quad h \rightarrow 0,$$

para toda función $y(t)$ lo suficientemente regular. Además, existe al menos una función para la que no se puede mejorar este orden.

En el caso de un esquema implícito, $b_k \neq 0$, se modifica de forma habitual. Necesitamos $l(t_\star; h) \rightarrow 0$ cuando $h \rightarrow 0$ con orden mayor que uno puesto que tenemos que controlar la acumulación de errores locales usando la estabilidad. Siguiendo la misma idea que en los métodos de un paso, vamos a considerar cantidades globales

Definición 138 Error de consistencia global: Se define como

$$l(h) = \max_{t_\star \in [t_0, t_0 + T - kh]} \{|l(y(t_\star); h)|\}.$$

para cualquier función $y(t)$ suficientemente regular solución de $u'(t) = f(t, u(t))$.

Definición 139 Error de truncatura local: Para cada valor de h ponemos

$$\tau(t_\star; h) = \frac{1}{h} l(y(t_\star); h)$$

para cualquier función $y(t)$ suficientemente regular solución de $u'(t) = f(t, u(t))$. A la cantidad $\tau(y(t_\star); h)$ se llama el **error de truncatura local** en el punto t_\star .

Definición 140 Error de truncatura global: Se define como

$$\tau(h) = \max_{t_\star \in [t_0, t_0 + T - kh]} \{|\tau(y(t_\star); h)|\}$$

para cualquier función $y(t)$ suficientemente regular solución de $u'(t) = f(t, u(t))$.

Observación 126 El error de truncatura se asocia a la ecuación diferencial y al hecho de que las tangentes también deben de converger, puesto que (por ejemplo en el caso explícito)

$$\tau(y(t_\star); h) = \frac{y(t_\star + kh) + \sum_{j=0}^{k-1} a_j y(t_\star + jh)}{h} - \sum_{j=0}^{k-1} b_j f(t_\star + jh, y(t_\star + jh))$$

mientras que el error local se asocia más a la convergencia hacia la propia función $y(t)$. Estos errores dependen evidentemente de la solución exacta $y(t)$.

Definición 141 El método es consistente si

$$\frac{1}{h} l(h) \rightarrow 0, \quad h \rightarrow 0.$$

y con orden de consistencia $p \geq 1$ si el error de consistencia máximo cumple

$$l(h) = \mathcal{O}(h^{p+1}), \quad h \rightarrow 0.$$

Esta potencia extra de h nos ayuda a controlar la suma de las amplificaciones de los errores locales y conseguir que esta suma tienda a cero si $h \rightarrow 0$. Se puede reescribir la definición de consistencia de una forma más cómoda

Definición 142 El método es consistente con orden de consistencia $p \geq 1$ cuando el error de truncatura global cumple

$$\tau(h) = \mathcal{O}(h^p), \quad h \rightarrow 0.$$

Definición 143 Error global en un punto t_\star es la diferencia

$$e_n^h = y(t_\star) - y_n^h$$

siendo y_n^h el valor calculado en el punto $t_n^h = t_\star$.

Observación 127 Si tomamos \tilde{z} el valor obtenido suponiendo que todos los valores previos son conocidos, entonces

$$e_n^h = y(t_\star) - y_n^h = y(t_\star) - \tilde{z} + \tilde{z} - y_n^h$$

y nos encontramos con que el error global está compuesto de dos partes. La primera es precisamente el error local

$$y(t_\star) - \tilde{z} = l(y(t_\star); h)$$

que debe de ser lo bastante pequeño para que la acumulación en todos los pasos de su método sea controlable. La segunda es consecuencia de aplicar el esquema en puntos distintos

$$\tilde{z} - y_n^h,$$

los exactos (caso de \tilde{z}) y los previos (caso de y_n^h) ya generados por el esquema. Esta parte se controla con la **0-estabilidad** del esquema numérico.

Bajo buenas condiciones, veremos que si el esquema es 0-estable y consistente con orden p entonces el método converge y cumple

$$\max_{0 \leq n \leq N^h} |y(t_n^h) - y_n^h| = O(h^p), \quad h \rightarrow 0^+$$

por lo que diremos entonces que el método converge con orden p .

5.4.1. Cálculo del error local de consistencia

En los métodos multipaso se puede hacer algo más también sobre los coeficientes b_j . Recordemos que dados y_0, y_1, \dots, y_{k-1} calculamos

$$y_{n+k} + \sum_{j=0}^{k-1} a_j y_{n+j} = h \sum_{j=0}^k b_j f(t_{n+j}, y_{n+j}), \quad n \geq 0.$$

Definición 144 Segundo polinomio característico

$$\sigma(z) = \sum_{j=0}^k b_j z^j. \quad (5.4)$$

Si ahora consideramos el problema

$$\begin{cases} y'(t) = 1, & t > 0 \\ y(0) = 0 \end{cases}$$

con solución $y(t) = t$ para todo $t > 0$, tomando los valores exactos $y_j = jh$ la aplicación al esquema general nos da

$$h(n+k) + \sum_{j=0}^{k-1} a_j(n+j)h = h \sum_{j=0}^k b_j, \quad n \geq 0.$$

o lo que es lo mismo

$$hn(1 + \sum_{j=0}^{k-1} a_j) + h(k + \sum_{j=0}^{k-1} j a_j) = h \sum_{j=0}^k b_j, \quad n \geq 0.$$

es decir,

$$hn\rho(1) + h\rho'(1) = h\sigma(1).$$

Como $\rho(1) = 0$ y $\rho'(1) \neq 0$, por la condición de Dahlquist, nos encontramos con

$$\sigma(1) = \rho'(1) \neq 0.$$

Resumiendo, una aplicación del esquema general a problemas básicos nos impone las restricciones siguientes

- **Métodos de un paso:** deben ser de la forma

$$y_{n+1} = y_n + h\Phi_f(t_n, y_n, y_{n+1}; h), \quad n \geq 0.$$

- **Métodos multipaso:** su primer polinomio característico $\rho(r)$ debe satisfacer la condición de Dahlquist y su segundo polinomio característico $\sigma(r)$ cumplir

$$\sigma(1) = \rho'(1) \quad (\rho'(1) \neq 0).$$

Para t_\star fijo, $l(y(t_\star); h)$ es una función de h y tiene un desarrollo de Taylor con respecto a h :

$$l(y(t_\star); h) = l(y(t_\star); 0) + kh\partial_h l(y(t_\star); 0) + \frac{k^2 h^2}{2!} \partial_{hh}^2 l(y(t_\star); 0) + \dots$$

Para tener la consistencia, $l(y(t_\star); h) = O(h^{p+1})$ con $p \geq 1$, tenemos que garantizar simplemente que se tenga al menos

$$l(y(t_\star); 0) = 0, \quad \partial_h l(y(t_\star); 0) = 0.$$

Teniendo en cuenta que en los métodos de un paso

$$l(y(t_\star); h) = y(t_\star + h) - y(t_\star) - h\Phi_f(t_\star, y(t_\star); h)$$

un resultado aparente y vistoso sobre la función de incremento es que el método general de un paso es convergente sí y sólo sí se cumple

$$y'(t) = \Phi_f(t, y(t); 0) \Leftrightarrow f(t, y(t)) = \Phi_f(t, y(t); 0).$$

Efectivamente, en estos casos se tiene la 0-estabilidad por construcción y esta expresión equivale a la consistencia. Además, si suponemos la regularidad $f \in C^p$ que implica $y \in C^{p+1}$ y también $\Phi_f \in C^p$ podemos plantear un desarrollo de Taylor del error de consistencia. Tomemos

$$\Psi(h) = \Phi_f(t_\star, y(t_\star); h)$$

entonces

$$\begin{aligned} l(t_\star; h) &= y(t_\star + h) - y(t_\star) - h\Phi_f(t_\star, y(t_\star); h) = y(t_\star + h) - y(t_\star) - h\Psi(h) \\ &= \{hy'(t_\star) + \frac{h^2}{2!}y''(t_\star) + \frac{h^3}{3!}y'''(t_\star) + \dots + \frac{h^{p+1}}{(p+1)!}y^{(p+1)}(\theta_\star)\} \\ &\quad - h\{\Psi(0) + h\Psi'(0) + \frac{h^2}{2!}\Psi''(0) + \dots \frac{h^p}{p!}\Psi^{(p)}(\xi)\} \end{aligned}$$

donde $\theta_\star \in (t_\star, t_\star + h)$ y $\xi \in (0, h)$. Podemos agrupar términos y tener

$$\begin{aligned} l(t_\star; h) &= h\{y'(t_\star) - \Psi(0)\} + h^2\left\{\frac{1}{2}y''(t_\star) - \Psi'(0)\right\} + \frac{h^3}{2!}\left\{\frac{1}{3}y'''(t_\star) - \Psi''(0)\right\} + \dots \\ &\quad + \frac{h^p}{(p-1)!}\left\{\frac{1}{p}y^{(p)}(t_\star) - \Psi^{(p-1)}(0)\right\} + \frac{h^{p+1}}{p!}\left\{\frac{1}{(p+1)}y^{(p+1)}(\theta_\star) - \Psi^{(p)}(\xi)\right\} \end{aligned}$$

entonces las potencias de h que se anulan en este desarrollo nos dan el orden del método.

Teorema 145 El método es de orden p , esto es $l(t_\star; h) = O(h^{p+1})$ ($h \rightarrow 0$) para todo $t_\star \in [t_0, t_0 + T]$, si todas las potencias menores o iguales a p en el desarrollo previo se cancelan, esto es, para $j = 1, 2, 3, \dots, p$

$$\frac{1}{j} y^{(j)}(t) = \Psi^{(j-1)}(0) = \frac{\partial^j}{\partial h^j} \Phi_f(t, y(t); 0)$$

para cualquier $y(t)$ solución de $y'(t) = f(t, y(t))$.

Volviendo a la forma concreta de los métodos multipaso, recordemos que el método lineal de k pasos (5.3) tiene orden $p \geq 1$ cuando

$$l(t; h) := \sum_{j=0}^k a_j y(t_{n+j}) - h \sum_{j=0}^k b_j y'(t_{n+j}) = O(h^{p+1}), \quad h \rightarrow 0,$$

para toda función $y(t)$ lo suficientemente regular y además, existe al menos una función para la que no se puede mejorar este orden. De momento, se tiene

$$l(t_\star; 0) = y(t_\star) + \sum_{j=0}^{k-1} a_j y(t_\star) = y(t_\star) \rho(1) = 0 \Leftrightarrow \rho(1) = 0$$

luego la condición $\rho(1) = 0$ es necesaria. Una vez la damos por supuesta es suficiente y necesario garantizar

$$\partial_h l(t_\star; 0) = 0.$$

Como $f(t_\star + jh, y(t_\star + jh)) = y'(t_\star + jh)$ entonces

$$l(t_\star; h) = y(t_\star + kh) + \sum_{j=0}^{k-1} a_j y(t_\star + jh) - h \sum_{j=0}^{k-1} b_j y'(t_\star + jh)$$

y teniendo en cuenta que $\partial_h y(t_{n+j}) = \partial_h y(t_n + jh) = jy'(t_{n+j})$ tenemos

$$\begin{aligned} \partial_h l(t_\star; h) &= ky'(t_\star + kh) + \sum_{j=0}^{k-1} ja_j y'(t_\star + jh) - \sum_{j=0}^{k-1} b_j y'(t_\star + jh) \\ &\quad - h \sum_{j=0}^{k-1} jb_j y''(t_\star + jh) \end{aligned}$$

por lo que

$$\partial_h l(t_\star; 0) = ky'(t_\star) + \sum_{j=0}^{k-1} ja_j y'(t_\star) - \sum_{j=0}^{k-1} b_j y'(t_\star) = y'(t_\star)[\rho'(1) - \sigma(1)]$$

Entonces

$$\partial_h l(t_\star; 0) = y'(t_\star)[\rho'(1) - \sigma(1)] = 0 \Leftrightarrow \sigma(1) = \rho'(1)$$

luego si tenemos $\sigma(1) = \rho'(1)$ garantizamos orden al menos 1. Y así sucesivamente, se obtiene

$$l(t; h) = C_0 y(t) + C_1 h y'(t) + C_2 \frac{h^2}{2!} y''(t) + C_3 \frac{h^3}{3!} y'''(t) \dots$$

siendo

$$\begin{aligned} C_0 &= \rho(1) = 0, \\ C_1 &= \rho'(1) - \sigma(1), \\ C_q &= \sum_{j=0}^k [j^q a_j - q j^{q-1} b_j], \quad q = 2, 3, \dots \end{aligned}$$

Se ve fácilmente entonces que

Teorema 146 *El método lineal de k pasos tiene orden $p \geq 1$ sí y sólo sí para cualquier solución del problema de Cauchy lo suficientemente regular*

$$l(t; h) = C_{p+1} \frac{h^{p+1}}{(p+1)!} y^{(p+1)}(t) + \dots, \quad C_{p+1} \neq 0.$$

o bien

$$C_0 = C_1 = \dots = C_p = 0, \quad C_{p+1} \neq 0.$$

siendo

$$\begin{aligned} C_0 &= \rho(1) = 0, \\ C_1 &= \rho'(1) - \sigma(1), \\ C_q &= \sum_{j=0}^k [j^q a_j - q j^{q-1} b_j], \quad q = 2, 3, \dots \end{aligned}$$

Observación 128 *La consistencia en los esquemas multipaso no garantiza convergencia pues puede fallar la propiedad de 0-estabilidad.*

5.4.2. Estudio de la estabilidad: 0-estabilidad

*Como ya hemos visto, un aspecto fundamental es la propagación de errores a lo largo de los distintos pasos que tengamos que hacer y como debemos de poder controlarlos cuando $h \rightarrow 0$. Esto es la **0-estabilidad** del método:*

Dados y_0, y_1, \dots, y_{k-1} para calcular en $[t_0, t_0 + T]$ con $h = T/N^{(h)}$ consideramos el esquema ($a_k = 1$, $a_0^2 + b_0^2 \neq 0$)

$$y_{n+k} + \sum_{j=0}^{k-1} a_j y_{n+j} = h \sum_{j=0}^k b_j f(t_{n+j}, y_{n+j}), \quad 0 \leq n \leq N^{(h)} - k$$

junto con su perturbación: dados z_0, z_1, \dots, z_{k-1} ,

$$z_{n+k} + \sum_{j=0}^{k-1} a_j z_{n+j} = h [\sum_{j=0}^k b_j f(t_{n+j}, z_{n+j}) + \delta_{n+k}], \quad 0 \leq n \leq N^{(h)} - k.$$

Observación 129 Los valores δ_{n+k} pueden representar errores de representación en el computador, de redondeo, o el error local de consistencia, como veremos en la prueba de convergencia.

Definición 147 Decimos que el **método es 0-estable** (leer cero-estable) cuando fijado el intervalo de cálculo $[t_0, t_0 + T]$ y particiones de talla h con $h = T/N$, si la perturbación cumple $\|\delta_n^{(h)}\| \leq \epsilon$ para $\epsilon > 0$, entonces existe una constante $C = C(T, f)$ y $h_0 > 0$ tal que para todo $h < h_0$ se cumple

$$\|z_n^{(h)} - y_n^{(h)}\| \leq C \{ \epsilon + \max_{j=0, \dots, k-1} \{\|z_j - y_j\|\}, \quad k \leq n \leq N = T/h.$$

La forma de comprobar la 0-estabilidad de estos métodos se visualiza mejor usando la matriz compañera del polinomio $\rho(z) = z^k + a_{k-1}z^{k-1} + a_{k-2}z^{k-2} + \dots + a_1z + a_0$ dada por

$$A = \begin{pmatrix} -a_{k-1} & -a_{k-2} & \dots & \dots & -a_1 & -a_0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix}.$$

Teorema 148 Siendo A la matriz compañera del primer polinomio característico $\rho(z)$ se tiene que existe una constante $M > 0$ tal que

$$\|A^n\| \leq M, \quad \forall n \geq 0$$

sí y sólo sí se cumple la propiedad de Dahlquist sobre los ceros de $\rho(z)$.

Dem: Si tenemos m raíces distintas r_1, r_2, \dots, r_m y cada una de ellas con multiplicidad $\mu_1, \mu_2, \dots, \mu_m$, entonces la forma general de la solución es

$$y_n = \sum_{j=1}^m p_j(n) r_j^n.$$

donde cada $p_j(n)$ es un polinomio de grado $\mu_j - 1$. Bajo la condición de Dahlquist se tiene claramente que estos valores están uniformemente acotados

$$|y_n| \leq C, \quad n \geq 0.$$

Esto es así ya que cualquier raíz r_j tal que $|r_j| < 1$ genera un término de la forma $p_j(n)r_j^n$ que tiende a cero cuando n crece y cualquier raíz r_j tal que $|r_j| = 1$ es simple luego genera un término de la forma $p_j r_j^n$ con p_j constante, de donde $|p_j r_j^n| = |p_j|$. Como $Y^n = A^n Y^0$ y

$$\|A^n\| = \max_{\|V\|=1} \|A^n V\|$$

es trivial obtener

$$\|A^n\| \leq M, \quad \forall n \geq 0$$

para cualquier norma matricial inducida. ■

Comprobación de la 0-estabilidad

Usando una norma vectorial cualquiera en \mathbb{R}^k y la norma matricial inducida (recordemos que k está fijo) podemos entonces comparar dos soluciones distintas asumiendo que tenemos la propiedad de Lipschitz de $f(t, y)$ con respecto a la segunda variable:

El esquema de cálculo

$$y_{n+k} + \sum_{j=0}^{k-1} a_j y_{n+j} = h \sum_{j=0}^k b_j f(t_{n+j}, y_{n+j})$$

se puede expresar en forma de un paso en \mathbb{R}^k como sigue:

$$Y^{n+1} = AY^n + h\vec{\Phi}_f(Y^n, Y^{n+1})$$

donde

$$Y^{n+1} = \begin{pmatrix} y_{n+k} \\ y_{n+k-1} \\ \vdots \\ y_{n+1} \end{pmatrix}, \quad Y^n = \begin{pmatrix} y_{n+k-1} \\ y_{n+k-2} \\ \vdots \\ y_n \end{pmatrix}, \quad \vec{\Phi}_f(Y^n, Y^{n+1}) = \begin{pmatrix} \sum_{j=0}^k b_j f(t_{n+j}, y_{n+j}) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Entonces, para comprobar la 0-estabilidad suponemos que tenemos para $n \geq 0$

$$\begin{aligned} y_{n+k} + \sum_{j=0}^{k-1} a_j y_{n+j} &= h \sum_{j=0}^k b_j f(t_{n+j}, y_{n+j}), \\ z_{n+k} + \sum_{j=0}^{k-1} a_j z_{n+j} &= h \left[\sum_{j=0}^k b_j f(t_{n+j}, z_{n+j}) + \delta_{n+k} \right] \end{aligned}$$

donde $\delta_{n+k} \leq \varepsilon$. Entonces, para

$$W^n = Y^n - Z^n = \begin{pmatrix} y_{n+k-1} - z_{n+k-1} \\ y_{n+k-2} - z_{n+k-2} \\ \vdots \\ y_n - z_n \end{pmatrix} \in \mathbb{R}^k$$

se cumple la ecuación

$$W^{n+1} = AW^n + h[\Psi^n + \Delta^n].$$

donde $\Psi^n = \vec{\Phi}_f(Y^n, Y^{n+1}) - \vec{\Phi}_f(Z^n, Z^{n+1})$, es decir,

$$\Psi^n = \begin{pmatrix} \sum_{j=0}^k b_j [f(t_{n+j}, y_{n+j}) - f(t_{n+j}, z_{n+j})] \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \Delta^n = \begin{pmatrix} \delta_{n+k} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Iterando $W^{n+1} = AW^n + h[\Psi^n + \Delta^n]$ llegamos a

$$W^n = A^n W^0 + h \sum_{j=0}^{n-1} A^{n-1-j} [\Psi^j + \Delta^j]$$

de donde, usando que $\|A^n\| \leq M$, $\forall n \geq 0$ y que $\|\delta_{n+k}\| \leq \varepsilon$, $\forall n \geq 0$,

$$\|W^n\| \leq M (\|W^0\| + h \sum_{j=0}^{n-1} \|\Psi^j\| + hn\varepsilon).$$

Pero

$$\Psi^j = \left(\sum_{s=0}^k b_s [f(t_{j+s}, y_{j+s}) - f(t_{j+s}, z_{j+s})], 0, \dots, 0 \right)^T$$

de donde usando $b = \max_{j \leq k-1} |b_j|$ y permitiendo $b_k \neq 0$ tenemos

$$\|\Psi^j\| \leq \sum_{s=0}^k |b_s| L_f |y_{j+s} - z_{j+s}| \leq |b_k| L_f |y_{j+k} - z_{j+k}| + b L_f \sum_{s=0}^{k-1} |y_{j+s} - z_{j+s}|$$

luego teniendo en cuenta que en \mathbb{R}^k cualquier norma es equivalente, podemos escribir

$$\|\Psi^j\| \leq |b_k| L_f \|W^{j+1}\| + b L_f \|W^j\|, \quad j \geq 0$$

de donde

$$\|W^n\| \leq M (\|W^0\| + h L_f \sum_{j=0}^{n-1} \{|b_k| \|W^{j+1}\| + b \|W^j\|\} + hn\varepsilon)$$

distinguiendo la parte de $\|W^n\|$ en la derecha obtenemos

$$\begin{aligned} \|W^n\| &\leq M (\|W^0\| + h L_f |b_k| \|W^n\| + h L_f \sum_{j=0}^{n-2} |b_k| \|W^{j+1}\| + h L_f \sum_{j=0}^{n-1} b \|W^j\| + hn\varepsilon) \\ &\leq M \|W^0\| + h L_f |b_k| M \|W^n\| + h B L_f M \sum_{j=0}^{n-1} \|W^j\| + hnM\varepsilon \end{aligned}$$

donde B depende de b_k y b . Entonces, si, por ejemplo, $h|b_k|ML_f < 1/2$ y ponemos $M = 2M$ en el caso implícito ($b_k \neq 0$), o para cualquier $h > 0$ en el explícito ($b_k = 0$) nos da

$$\|W^n\| \leq M\|W^0\| + hB L_f M \sum_{j=0}^{n-1} \|W^j\| + hnM\varepsilon, \quad \forall n \geq 0.$$

Dado $\xi_0 = \|W^0\|$ si consideramos la recurrencia para ξ_n dada por

$$\xi_n = M\|W^0\| + hB L_f M \sum_{j=0}^{n-1} \xi_j + hnM\varepsilon, \quad \forall n \geq 0.$$

tenemos que $\|W^n\| \leq \xi_n$ y por otro lado,

$$\xi_{n+1} - \xi_n = hB L_f M \xi_n + hM\varepsilon$$

de donde

$$\xi_{n+1} = (1 + hB L_f M)\xi_n + hM\varepsilon$$

Esto nos lleva a la estimación estandarizada en procesos iterativos:

$$\begin{aligned} \xi_n &= (1 + hB L_f M)^n \xi_0 + hM\varepsilon \sum_{j=0}^{n-1} (1 + hB L_f M)^j \\ &= (1 + hB L_f M)^n \xi_0 + hM\varepsilon \frac{(1 + hB L_f M)^n - 1}{(1 + hB L_f M) - 1} \\ &\leq e^{TB L_f M} \xi_0 + M\varepsilon \frac{e^{TB L_f M} - 1}{B L_f M} \end{aligned}$$

de donde

$$\|W^n\| \leq C(T, f)\{\|W^0\| + \varepsilon\}, \quad n \geq 0$$

que es el resultado de estabilidad que buscamos, y el de convergencia si usamos como ε el error de truncatura global. ■

Observación 130 Se puede comparar con el caso $k = 1$ en donde $A = 1$, la ecuación es escalar y la mecánica es la misma pero más sencilla. También sabemos que $\rho(A) = 1$ pero esto no implica $\|A\| = 1$.

Es fundamental tener aquí la acotación

$$\|A^n\| \leq M, \quad \forall n \geq 0$$

y esto se garantiza con la propiedad de Dahlquist.

5.5. Análisis de convergencia

El resultado fundamental es el siguiente

Teorema 149 Teorema de Lax-Richtmyer (caso multipaso)

Un método de k pasos es convergente si y sólo si es consistente y 0-estable. Si el error local es de orden $p + 1$ el orden de convergencia es p siempre que el error en los datos iniciales también sea de orden p .

Este resultado fue demostrado por primera vez por Dahlquist (1956) en su tesis doctoral [8] para métodos lineales multipaso. También una referencia clásica es el texto de Isaacson-Keller (1966) [17].

Observación 131 No debemos de olvidar los k valores para inicializar que se usan en un método de k pasos, esto es, los valores $y_0^{(h)}, y_1^{(h)}, \dots, y_{k-1}^{(h)}$. Como k es fijo, cuando $h \rightarrow 0$, todos estos valores deben de converger al valor inicial puesto que los puntos $t_0^{(h)}, t_1^{(h)}, \dots, t_{k-1}^{(h)}$ asociados se contraen al t_0 y por consiguiente se debe cumplir

$$\lim_{h \rightarrow 0} y_j^{(h)} = y(t_0), \quad j = 0, 1, 2, \dots, k - 1.$$

La convergencia en el caso de los métodos de un paso sólo depende de la consistencia puesto que la propiedad de 0-estabilidad esta garantizada por construcción (la condición de Dahlquist es trivial en métodos de un paso ya que $\rho(z) = z - 1$).

Teorema 150 Teorema de Lax-Richtmyer (caso un paso)

Un método de un paso es convergente si y sólo si es consistente. Si el error local es de orden $p + 1$ el orden de convergencia es p siempre que el error en los datos iniciales también sea de orden p .

Para obtener la convergencia en los métodos multipaso comparamos y_n con $y(t_n)$: dados y_0, y_1, \dots, y_{k-1} e $y(t_0), y(t_1), \dots, y(t_{k-1})$ tenemos

$$\begin{aligned} y_{n+k} + \sum_{j=0}^{k-1} a_j y_{n+j} &= h \sum_{j=0}^k b_j f(t_{n+j}, y_{n+j}), \\ y(t_{n+k}) + \sum_{j=0}^{k-1} a_j y(t_{n+j}) &= h \sum_{j=0}^k b_j f(t_{n+j}, y(t_{n+j})) + l(t_n; h) \end{aligned}$$

luego para $E_n = Y(t_n) - Y^n$, con notación obvia para $Y(t_n)$, se cumple la ecuación

$$E^{n+1} = AE^n + h\Psi_n + L(t_n, h).$$

donde $L(t_n; h) = (l(t_n; h), 0, \dots, 0)^T = O(h^{p+1})$ y de forma natural se aplica el resultado de 0-estabilidad anterior usando $\varepsilon = l(t_n; h)h^{-1}$.

1. si el error inicial cumple $\|E_0\| = O(h^p)$,
2. el método numérico tiene error de consistencia $O(h^{p+1})$
3. y se cumple la 0-estabilidad ($\|A\|^n \leq M$, $n = 1, 2, 3, \dots$)

entonces tenemos convergencia de orden $O(h^p)$ para $h \leq h_0$

$$\|E_n\| \leq C h^p, \quad n = 0, 1, \dots, N = T/h.$$

5.6. Barreras en el orden de convergencia

¿Cuál es el orden más alto alcanzable? Un método de k pasos involucra los $k + 1$ valores $y_n, y_{n+1}, \dots, y_{n+k}$ y los coeficientes ($a_k = 1$) $a_0, a_1, \dots, a_{k-1}, b_0, b_1, \dots, b_{k-1}, b_k$. Por lo tanto, tenemos $2k + 1$ coeficientes a elegir si el método es implícito o bien $2k$ en el caso en el que sea explícito ($b_k = 0$). Puesto que las restricciones sobre el orden se reescriben como un sistema lineal en los coeficientes y tenemos que cumplir con p ecuaciones lineales, en teoría podemos llegar a alcanzar orden $p = 2k$ en el caso implícito o $p = 2k - 1$ en el explícito. Cuando esto ocurre así se dice que el método es maximal. Pero los métodos de orden maximal en general no cumplen la condición de las raíces, esto es, no son 0-estables y por lo tanto no son convergentes.

Teorema 151 Primera barrera de Dahlquist (1956)

Un método lineal de k -pasos y 0-estable puede tener como mucho orden $k + 1$ si k es impar y $k + 2$ si k es par.

Un método de k pasos y cero-estable de orden $k + 2$ se dice que es optimal. Naturalmente, aquí es k par y el método implícito. Se puede mostrar que que todas las raíces espúreas o contaminadas del primer polinomio característico de un método optimal tienen módulo uno. Esto lleva a dificultades en cuanto a la estabilidad absoluta. Por lo tanto, los métodos optimales no son los más eficientes y aquellos con k impar y orden $k + 1$ pueden trabajar mejor que los optimales.

Ejemplo 152 La regla de Simpson dada por

$$y_{n+2} - y_n = \frac{h}{3}[f_{n+2} + 4f_{n+1} + f_n]$$

ocupa un lugar privilegiado puesto que $k = 2$ y tiene orden $p = 4$ siendo entonces maximal y optimal.

Observación 132 Evidentemente existe una segunda barrera encontrada por Dahlquist pero no la vamos a ver aquí.

5.7. Familias de métodos de multipaso

Al principio de este tema hemos visto varios ejemplos que se pueden clasificar en familias:

Métodos de Adams

Se corresponden con tener orden de consistencia máximo siendo el primer polinomio característico de la forma

$$\rho(r) = r^k - r^{k-1} = r^{k-1}(r - 1).$$

Esta forma proviene del hecho de que interpolamos valores f_{n+j} e integramos. Evidentemente son 0-estables al cumplirse la condición de las raíces.

Se denominan de **Adams-Basforth** cuando son explícitos y de **Adams-Moulton** cuando son implícitos (esto depende de la forma del segundo polinomio característico). Aquí el valor de k depende del número de puntos que tomemos para obtener el polinomio de interpolación de los valores f_{n+j} e integrar.

Métodos de Nystrom

Son explícitos y se corresponden con primer polinomio característico de la forma

$$\rho(r) = r^k - r^{k-2} = r^{k-2}(r^2 - 1).$$

La misma observación sobre k que antes se tiene aquí, esto es, el valor de k corresponde al número de puntos que tomemos para obtener el polinomio de interpolación de los valores f_{n+j} e integrar.

Fórmulas de Diferenciación Regresivas (BDFk)

Entre las muchas familias de métodos multipaso destacan por su buen rendimiento en los problemas rígidos los métodos de diferenciación regresiva, o bien backward differentiation formulae (BDFk) en inglés. Normalizando a uno el coeficiente de y_{n+1} tenemos

$$\begin{aligned} k = 1 : \quad & y_{n+1} - y_n = h f_{n+1} \\ k = 2 : \quad & y_{n+1} - \frac{4}{3}y_n + \frac{1}{3}y_{n-1} = \frac{2}{3}h f_{n+1} \\ k = 3 : \quad & y_{n+1} - \frac{18}{11}y_n + \frac{9}{11}y_{n-1} - \frac{2}{11}y_{n-2} = \frac{6}{11}h f_{n+1} \end{aligned}$$

luego el método BDFk más simple es el método de Euler implícito. En los métodos BDF la condición sobre las raíces falla si $k > 6$, pero el rango $1 \leq k \leq 6$ es más que suficiente para las aplicaciones de estos métodos.

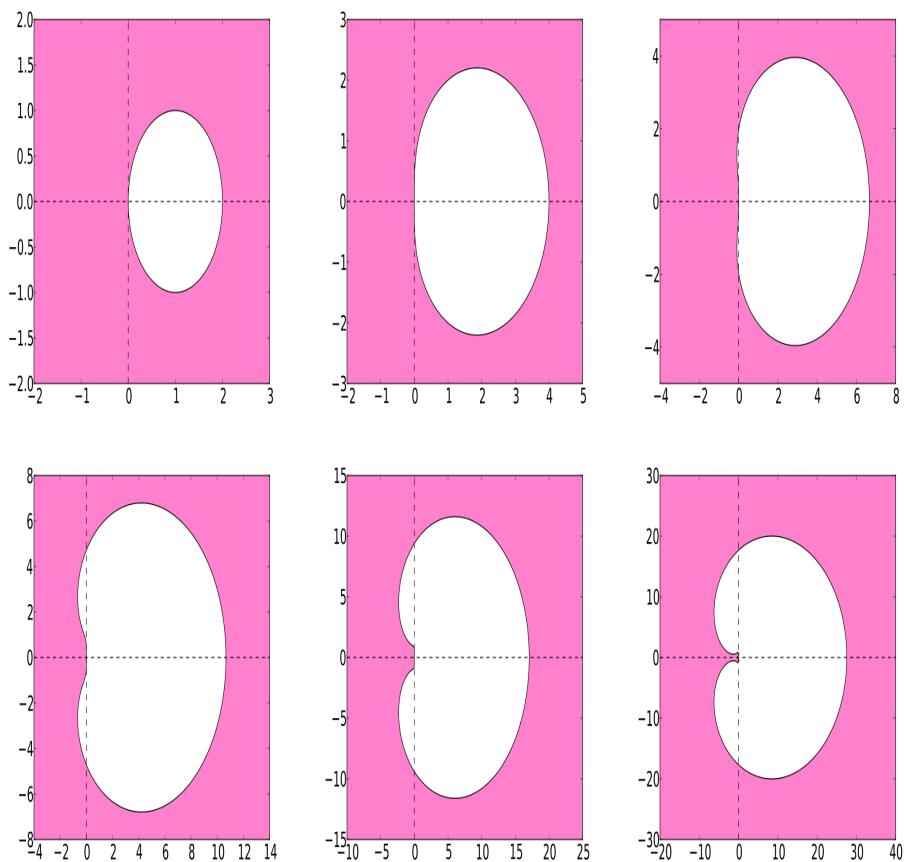


Figura 5.3: Regiones de A-estabilidad para distintos métodos BDF.

Teorema 153 Convergencia de los métodos BDF

El primer polinomio característico asociado al método BDF satisface la condición de las raíces y el método es convergente si y sólo si $1 \leq k \leq 6$.

Ejemplo 154 Obtención de métodos BDF: Tomemos el caso $k = 1$ entonces $P(s)$ interpola

$$P(t_{n+1}) = y_{n+1}, \quad P(t_n) = y_n.$$

La ecuación para $P(s)$ se deduce usando la interpolación de Lagrange y es

$$P(t) = y_n \frac{t - t_{n+1}}{t_n - t_{n+1}} + y_{n+1} \frac{t - t_n}{t_{n+1} - t_n}$$

o bien, como $t_{n+1} - t_n = h$ también

$$P(t) = y_n \frac{t - t_{n+1}}{-h} + y_{n+1} \frac{t - t_n}{h}$$

Derivando, tenemos

$$P'(t) = y_n \frac{1}{-h} + y_{n+1} \frac{1}{h}$$

Entonces $P'(t_{n+1}) = f(t_{n+1}, y_{n+1})$ genera

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

que coincide con Euler implícito. Si ahora queremos usar y_{n-1}, y_n, y_{n+1} tenemos

$$\begin{aligned} P(t) &= y_{n-1} \frac{(t - t_n)(t - t_{n+1})}{(t_{n-1} - t_n)(t_{n-1} - t_{n+1})} + y_n \frac{(t - t_{n-1})(t - t_{n+1})}{(t_n - t_{n-1})(t_n - t_{n+1})} \\ &\quad + y_{n+1} \frac{(t - t_n)(t - t_{n-1})}{(t_{n+1} - t_n)(t_{n+1} - t_{n-1})} \end{aligned}$$

o bien

$$P(t) = y_{n-1} \frac{(t - t_n)(t - t_{n+1})}{2h^2} + y_n \frac{(t - t_{n-1})(t - t_{n+1})}{-h^2} + y_{n+1} \frac{(t - t_n)(t - t_{n-1})}{2h^2}$$

y se procede como antes.

5.8. Ejercicios

1. (Lambert) Aplicar el esquema

$$y_{n+2} - y_{n+1} = \frac{h}{12} [4f(t_{n+2}, y_{n+2}) + 8f(t_{n+1}, y_{n+1}) - f(t_n, y_n)], \quad n \geq 0$$

al problema de valor inicial $y'(t) = t$, $y(0) = 0$ para obtener una ecuación en diferencias de la forma $y_{n+2} - y_{n+1} = \varphi(n, h)$. Intentando una solución particular de la forma $y_n = An^2 + Bn$ encontrar la solución exacta de la ecuación en diferencias que satisface $y_0 = 0$, $y_1 = h^2/2$. Demostrar que la secuencia computada converge en el límite estacionario pero no a la solución buscada. Explicar la razón y dar algún remedio.

2. De acuerdo a los valores de b y c estudiar la solución general de la recurrencia

$$y_0 = \alpha, y_1 = \beta, \quad y_{n+2} + by_{n+1} + cy_n = 0, \quad n \geq 0.$$

3. Obtener aproximaciones de orden 1, 2 y 3 de la derivada $y'(t)$ usando los desarrollos de Taylor de $y(t+h)$, $y(t-h)$ e $y(t-2h)$ en la forma más arbitraria posible y construir un método de tres pasos para resolver el problema de valor inicial explicitando el error local en cada paso. Comprobar que el esquema resultante es el siguiente: Dados y_0, y_1, y_2 obtener y_{n+1} para $n \geq 2$ via

$$\frac{1}{3}y_{n+1} + \frac{1}{2}y_n - y_{n-1} + \frac{1}{6}y_{n-2} = hf(t_n, y_n).$$

Comprobar que, a pesar de tener orden 3, el método no es 0-estable y, por lo tanto, no es convergente.

4. Se considera el método lineal de dos pasos explícito

$$y_{n+2} - (1+a)y_{n+1} + a y_n = h\{\beta_1 f_{n+1} + \beta_0 f_n\}.$$

Para cualquier valor de $a \in \mathbb{R}$ con $|a| < 1$ y si $\beta_0 + \beta_1 = 1 - a$ entonces el método converge con orden uno. Para conseguir orden 2 se necesita la restricción adicional

$$\beta_1 = \frac{3-a}{2}.$$

Usando como ejemplo en el caso de orden uno: $a = 0.5$, $\beta_0 = -0.45$, $\beta_1 = 0.95$ y en el caso de orden dos: $a_0 = 0.5$, $\beta_0 = -0.75$, $\beta_1 = 1.25$, generar una gráfica de pendientes para comprobar el orden del error usando un problema con solución conocida. Resumiendo, para $a, b \in \mathbb{R}$ con $|a| < 1$ la elección

$$\beta_0 = 1 - a - \beta_1, \quad \beta_1 = b + \frac{3-a}{2}$$

genera una familia de métodos de dos pasos de orden 1 para todo $b \neq 0$ y si $b = 0$ se genera un método de orden 2.

5. La función $y(t) \equiv 1$ cumple el problema trivial

$$y(0) = 1, \quad y'(t) = 0, \quad t \in [0, 10].$$

En este caso la parte derecha de los métodos lineales multipaso no influye en el cálculo de la solución dada por el esquema discreto puesto que $f = 0$. Para el problema anterior las raíces del polinomio

$$p(z) = z^2 - (1 + a)z + a$$

asociado a la parte izquierda son a y 1 . Comprueba computacionalmente que si $|a| < 1$ entonces los posibles errores en los valores iniciales $y_0 = 1 \pm \epsilon$ e $y_1 = 1 \pm \delta$ se amortiguan mientras que si $|a| > 1$ entonces el cálculo lleva a resultados inestables e inservibles. Esta situación empeora conforme se aumenta el número de puntos que usamos en el intervalo, esto es, si buscamos eliminar el problema aumentando la precisión lo que hacemos es empeorar. Comprobarlo con los siguientes datos

- a) $T = 10$, $N = 100$, $a = 1.1$, $y_0 = 1 + 10^{-13}$, $y_1 = 1 - 10^{-10}$
- b) $T = 10$, $N = 100$, $a = 0.9$, $y_0 = 1 + 10^{-13}$, $y_1 = 1 - 10^{-10}$

6. Sabiendo que un método lineal de k pasos tiene orden $p \geq 1$ sí y sólo sí

$$C_0 = C_1 = \dots = C_p = 0, \quad C_{p+1} \neq 0.$$

siendo

$$\begin{aligned} C_0 &= \rho(1) = 0, \\ C_1 &= \rho'(1) - \sigma(1), \\ C_q &= \sum_{j=0}^k [j^q a_j - q j^{q-1} b_j], \quad q = 2, 3, \dots \end{aligned}$$

Observemos que si $k = 2$ entonces

$$C_2 = \sum_{j=0}^2 [j^2 a_j - 2 j^{2-1} b_j] = 0^2 a_0 + 1^2 a_2 + 2^2 a_3 - [2 0^1 b_0 + 2 1^1 b_1 + 2 2^1 b_2].$$

Encontrar los valores de α para los que el método

$$y_{n+2} - 2\alpha y_{n+1} + (2\alpha - 1) y_n = h[\alpha f_{n+2} + (2 - 3\alpha) f_{n+1}], \quad n \geq 0$$

es cero estable, consistente y determinar el orden mayor posible.

7. (Griffiths et al.) Comprobar que el orden de consistencia del método

$$y_{n+2} + (\alpha - 1)y_{n+1} - \alpha y_n = \frac{h}{4}[(\alpha + 3)f_{n+2} + (3\alpha + 1)f_n]$$

es 2 si $\alpha \neq -1$ y 3 si $\alpha = -1$. Aplicar el método cuando $\alpha = -1$ al problema $y'(t) = 0$, $t > 0$ con $y(0) = 0$ tomando $y_0 = 0$, $y_1 = h$ como valores iniciales y explicar el comportamiento resultante de la solución numérica.

8. (*Griffiths et al.*) Encontrar los valores de α para los que el método

$$y_{n+2} + 2\alpha y_{n+1} - (2\alpha + 1) y_n = h[(\alpha + 2) f_{n+1} + \alpha f_n], \quad n \geq 0$$

es cero estable, consistente y determinar el orden mayor posible. Deducir que hay una elección del parámetro α que consigue que el método tenga orden 3. Estudiar la validez de ésta elección.

9. Comprobar que un método multipaso tiene orden p sí y sólo sí para las funciones polinomio $y(t) = t^r$ se cumple

$$l_{t^r}(t; h) = 0, \quad r = 0, 1, 2, \dots, p, \quad l_{t^{p+1}}(t; h) \neq 0$$

donde $l_{w(t)}(t; h)$ representa el error local aplicado a la función $w(t)$.

10. (*Lambert*) Comprobar computacionalmente el efecto de la 0-estabilidad usando el método

$$y_{n+2} - (1 + \alpha)y_{n+1} + \alpha y_n = \frac{h}{2}[(3 - \alpha)f(t_{n+1}, y_{n+1}) - (1 + \alpha)f(t_n, y_n)]$$

para calcular la solución de la ecuación $y' = 4ty^{1/2}$ con $y(0) = 1$ y $t \in [0, 2]$ que es $y(t) = (t^2 + 1)^2$. Usar los valores $\alpha = 0$ y $\alpha = -5$ con pasos $h = 0.1, 0.05, 0.025$.

11. Comprobar que el método de Quade es convergente.

Capítulo 6

Ejercicios computacionales

1. Se quiere reproducir la solución de $y'(t) = 10y(t)$ con $y(0) = 1$ en el intervalo $[0, T]$ con $T = 4$ y usando el método de Euler explícito. Tomando $h = T/N$ observar la evolución del cálculo para valores crecientes de N en comparación con la solución exacta. Realizar el mismo estudio para $y'(t) = -10y(t)$ con los mismos datos y confirmar los resultados teóricos.
2. Reproducir el ejercicio anterior tomando como dato inicial $y_0^h = 1 - h^{1/2}$.
3. Aplicar el método de Euler con paso constante y en el intervalo $[0, 0.5]$ a las ecuaciones
 - a) $y' = y^2$ con $y(0) = 1$
 - b) $y' = t^2 + y^2$ con $y(0) = 0$

Hacer estimaciones de error teóricas y compararlas con los errores reales.

4. Resolver el sistema

$$\begin{cases} y'_1(t) &= -100y_1 + y_2, \quad y_1(0) = 1 \\ y'_2(t) &= y_1 - 100y_2, \quad y_2(0) = 0 \end{cases}$$

por el método de Euler explícito, implícito y por el método de Crank-Nicolson en el intervalo $[0, 1]$ con paso uniforme $h = 1/10$ y $h = 1/100$. Comparar con la solución exacta. Hacer estimaciones de error rigurosas y compararlas con los errores reales.

5. El sistema

$$\begin{cases} y'_1(t) &= y_1 - 2y_2 + 4\cos(t) - 2\sin(t) \quad y_1(0) = 1 \\ y'_2(t) &= 3y_1 - 4y_2 + 5\cos(t) - 5\sin(t), \quad y_2(0) = 2 \end{cases}$$

posee por solución exacta

$$y_1(t) = \cos(t) + \sin(t), \quad y_2(t) = 2\cos(t).$$

a) Escribirlo como un problema de la forma

$$\frac{d}{dt} \vec{y} = A\vec{y} + G(\vec{t})$$

- b) Usando la notación matricial en MATLAB, aproximar la solución por el método de Euler explícito, implícito y por el método de Crank-Nicolson en el intervalo $[0, 1]$ con paso uniforme $h = 1/10$ y $h = 1/100$. Hacer estimaciones de error rigurosas y compararlas con los errores reales.
6. Para $y'(t) = a t^{a-1}$, $y(0) = 0$ con $a > 0$ la solución es $y(t) = t^a$. Cuando a no es entero la solución no es infinitamente derivable. En particular hace falta $a \geq 2$ para tener $y(t)$ de clase C^2 . Usar el método de Euler explícito para $a = 2.5, 1.5, 1.1$ con paso $h = 0.2, 0.1, 0.05$. Calcular el error y determinar numéricamente el orden de convergencia.
7. Resolver los siguientes problemas usando el método de Euler explícito con pasos $h = 0.2, 0.1, 0.05$ (deducir N). Calcular el error absoluto usando la solución verdadera $y(t)$. Observar el decaimiento del error conforme se disminuye h .
- a) $y'(t) = [\cos(y(t))]^2$, $0 \leq t \leq 10$, $y(0) = 0$. exacta : $y(t) = \text{atan}(t)$.
 - b) $y'(t) = -y(t)^2$, $1 \leq t \leq 10$, $y(1) = 1$. exacta : $y(t) = 1/t$.
8. Explora la solución del **test del círculo** $z'(t) = iz(t)$ con $z(0) = 1$ donde $z(t) = u(t) + iv(t) \in \mathbb{C}$. Usa el método de Euler explícito, Euler implícito y Crank-Nicolson. Comprueba con lápiz y papel y computacionalmente que las soluciones cumplen
- a) Para Euler explícito $u_{n+1}^2 + v_{n+1}^2 = (1+h^2)(u_n^2 + v_n^2) = (1+h^2)^{n+1}$,
 - b) Para Euler implícito $u_{n+1}^2 + v_{n+1}^2 = (1+h^2)^{-1}(u_n^2 + v_n^2) = (1+h^2)^{-(n+1)}$,
 - c) Para Crank-Nicolson $u_{n+1}^2 + v_{n+1}^2 = u_n^2 + v_n^2 = 1$
- por lo que los dos primeros métodos generan una espiral hacia el exterior o el interior mientras que para Crank-Nicolson se respeta la amplitud.
9. El sistema $y'_1 = y_2$; $y'_2 = y_1$ no se puede escribir en forma compleja. Comprobar que las órbitas son hipérbolas de la forma $y_1^2 - y_2^2 = c$. Computar algunos casos con los datos iniciales $y_1(0) = -1$, $y_2(0) = 1 - \delta$ e $y_1(0) = -1 + \delta$, $y_2(0) = 1$ donde $0 < \delta \ll 1$. Por ejemplo, usar $\delta = 0.005$ y $\delta = 0.02$.
10. Vamos a ver el comportamiento del error para el método de Euler explícito. Consideremos

$$y' + y = 1, \quad y(0) = 0 \tag{6.1}$$

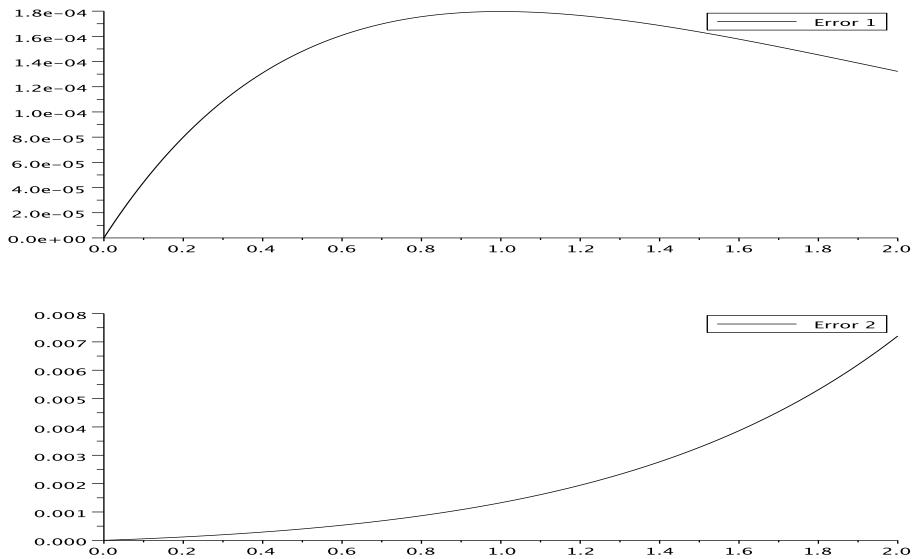


Figura 6.1: Comportamiento del error con el método de Euler en $[0, 2]$ para $h = 2/2048$ para los ejemplos (6.1) y (6.2) respectivamente.

con solución

$$y(t) = 1 - e^{-t}$$

y

$$y' = y, \quad y(0) = 1 \quad (6.2)$$

con solución

$$y(t) = e^t.$$

Calcular el error usando

$$E_n = \max_{k \leq n} |y(t_k) - y_k|$$

Obtener la Figura 6.1 que muestra el comportamiento de los errores en $[0, 2]$ calculados con $h = 2/2048 = 1/1024$ para ambos ejemplos. Podemos ver como para el ejemplo clásico de solución $y(t) = e^t$, el menor error es justo al principio, de hecho el peor error es siempre el previo y y luego crece, el crecimiento parece ser exponencial. Por otro lado, con el ejemplo de solución $y(t) = 1 - e^{-t}$, el peor error ocurre cerca del principio y luego decae. Estos resultados concuerdan con el análisis de convergencia visto en teoría y que tienen que ver con el comportamiento de $\partial_y f(t, y)$. En el primer caso es $f(t, y) = -y + 1$ mientras que en el segundo es $f(t, y) = y$.

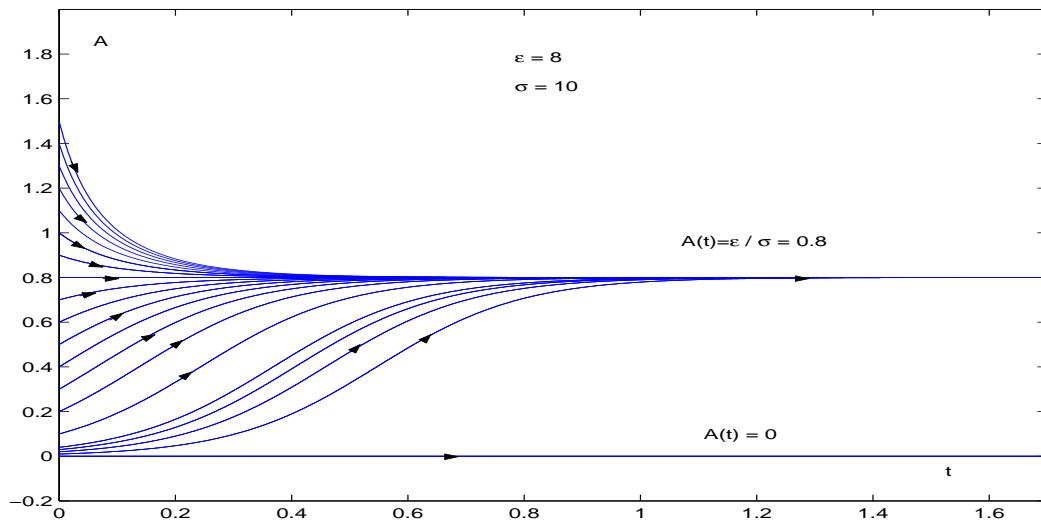


Figura 6.2: Soluciones para distintos valores iniciales de la ecuación logística

11. El problema de valor inicial conocido como **ecuación logística**

$$\begin{cases} \frac{d}{dt}A(t) = \varepsilon A(t) - \sigma A(t)^2, & t > 0, \\ A(0) = A_0 \geq 0, \end{cases}$$

donde $\varepsilon, \sigma > 0$ son números positivos dados, surge en ecología del análisis del desarrollo de una sola especie de tamaño $A(t)$ en el instante t suponiendo que tiene acceso a recursos limitados. A la constante $\varepsilon > 0$ se conoce como la tasa de crecimiento y el término no lineal σA^2 describe la mortalidad como encuentros competitivos de la especie A consigo misma. La razón ε/σ se denomina el **nivel de saturación**. Para $\varepsilon = 8$ y $\sigma = 13$ obtén para distintos valores del dato inicial $A(0)$ las distintas soluciones y realiza una gráfica en la que aparezcan todas ellas junto con la solución estacionaria $A \equiv \varepsilon/\sigma$ con respecto al tiempo (reproducir la Figura 6.2).

12. Dado el sistema lineal

$$\frac{d}{dt}x(t) = -0.1x(t) - 0.2y(t) \quad (6.3)$$

$$\frac{d}{dt}y(t) = 0.8x(t). \quad (6.4)$$

Usa la colección de datos iniciales $x_0 = -1 + i \cdot 0.2$ e $y_0 = -1 + i \cdot 0.3$ para $i = 0, 1, 2, \dots, 10$ y el método de Euler para reproducir los resultados de la Figura 6.3

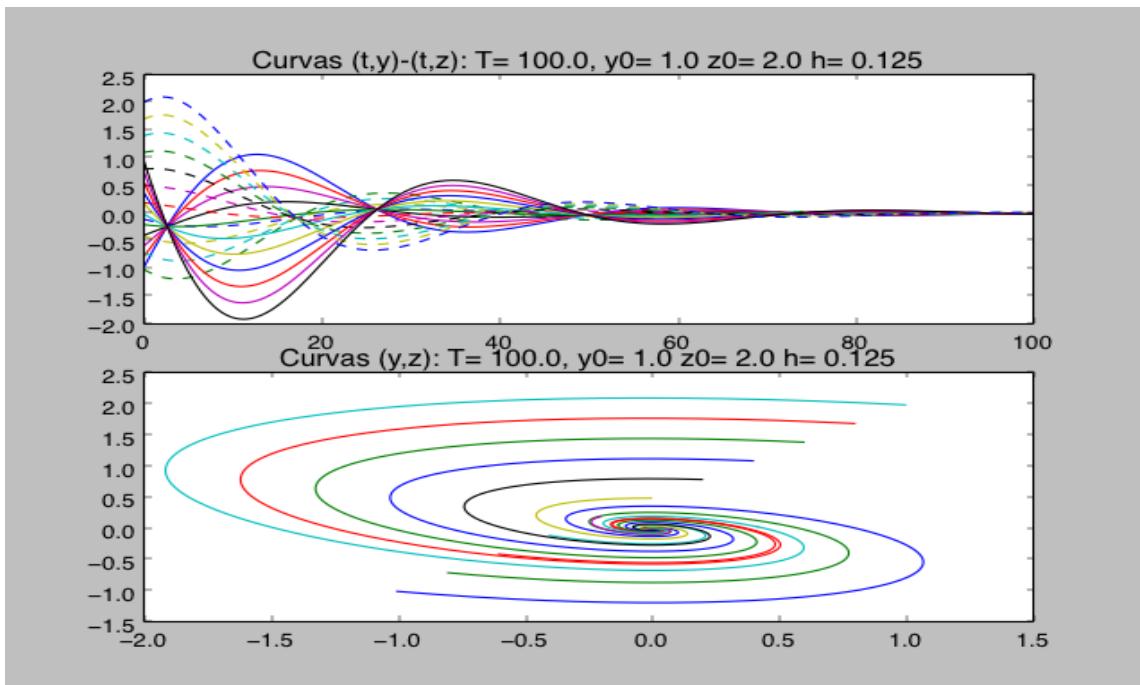


Figura 6.3: Comportamiento de las soluciones de las ecuaciones (6.3) y (6.4).

13. **Modelo depredador-presa:** El modelo de Lotka-Volterra viene expresado por

$$\begin{aligned}\frac{d}{dt}P(t) &= a P(t) - b P(t) D(t) \\ \frac{d}{dt}D(t) &= -c D(t) + d(b(P(t) D(t)))\end{aligned}$$

donde $D(t)$ es la población de depredadores, $P(t)$ es la población de presas en el instante de tiempo t y los valores a, b, c, d son parámetros del modelo. Obtén la gráfica de la solución para los valores $a = 1$, $b = 1/2$, $c = 1$ y $d = 1/2$ y con valores iniciales $P(0) = 3$, $D(0) = 3.87$. Realiza una gráfica con $P(t)$ y $D(t)$ como función del tiempo. Para distintos valores de los datos iniciales $P(0)$, $D(0)$ obtén las distintas soluciones y realiza el plano de fases correspondiente.

- Realizar una gráfica con $P(t)$ y $D(t)$ como función del tiempo.
- Para distintos valores de los datos iniciales $P(0)$, $D(0)$ obtener las distintas soluciones y realizar el plano de fases correspondiente, ver Figura 1.8.
- Usar Euler explícito y comparar de forma gráfica los resultados obtenidos.

14. **Vibraciones forzadas:** la reacción de un oscilador armónico está gobernada por el sistema de segundo orden

$$\begin{cases} \frac{d^2}{dt^2}x(t) + 2b\frac{d}{dt}x(t) + x(t) = \cos(wt), & t > 0, \\ x(0) = 0, \quad \frac{d}{dt}x(0) = 0. \end{cases}$$

En el caso $w = 1$ usando el método de Euler explícito observa la gráfica de $x(t)$ para valores decrecientes de b : $b = 0.2$, $b = 0.15$, $b = 0.1$, $b = 0.05$, $b = 0.01$. El fenómeno que puedes observar en la Figura 6.4 se conoce como **resonancia**: todos los sistemas vibran internamente. Un sistema débilmente estimulado puede exhibir una gran amplificación si la frecuencia de su estimulación coincide con su frecuencia natural.

Reproduce los mismos resultados pero ahora usa RK4. Comparar la diferencia entre los valores de h que se han necesitado con respecto al uso de Euler explícito.

15. Las componentes x_1 , x_3 y x_5 del sistema lineal $x'(t) = Ax(t)$ donde

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & -2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & -1 & 0 \end{pmatrix}.$$

representan ángulos de giro en un sistema físico. Resolver el problema en $[0, 20]$ con dato inicial $x_0 = (-0.05, 0, 0, 0, 0.1, 0)$ usando Euler explícito, Euler implícito y Crank-Nicolson. Usar las facilidades en álgebra lineal en cuanto a notación matricial e inversión de matrices en MATLAB y reproducir la Figura 6.5. Comparar las necesidades de cada método, en cuanto a la restricción sobre h , para reproducir la misma solución cualitativamente hablando.

16. En el intervalo temporal $t \in [1, 10]$ la ecuación diferencial ordinaria

$$t^3 y'''(t) - t^2 y''(t) + 3t y'(t) - 4y(t) = 5t^3 \log(t) + 9t^3$$

con datos iniciales $y(1) = 0$, $y'(1) = 1$, $y''(1) = 3$ tiene por solución

$$y(t) = -t^2 + t \cos(\log(t)) + t \sin(\log(t)) + t^3 \log(t)$$

- a) Transformar esta ecuación en un sistema de ecuaciones diferenciales de primer orden

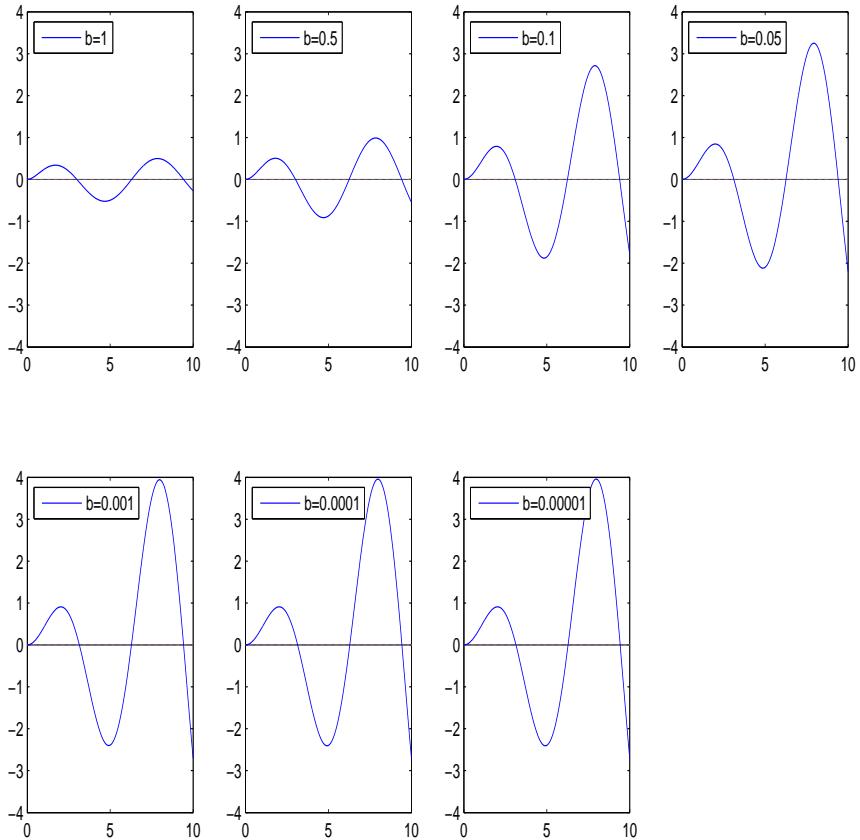


Figura 6.4: Vibraciones forzadas: Soluciones para distintos valores de b .

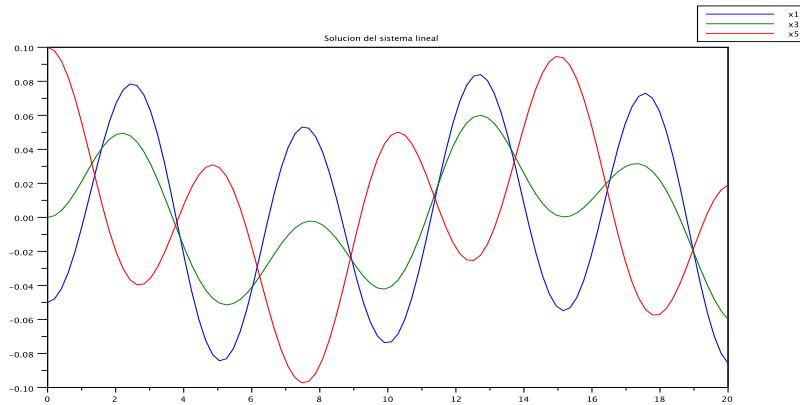


Figura 6.5: Evolución de los ángulos de giro de un sistema físico.

- b) Aproximar la solución en una partición uniforme del intervalo $[1, 10]$ mediante el siguiente método de Runge-Kutta de tres etapas (RK3)

$$\begin{aligned}y_{n+1} &= y_n + \frac{h}{4} (k_1 + 3k_3) \\ k_1 &= f(t_n, y_n), \\ k_2 &= f\left(t_n + \frac{1}{3}h, y_n + \frac{1}{3}h k_1\right), \\ k_3 &= f\left(t_n + \frac{2}{3}h, y_n + \frac{2}{3}h k_2\right).\end{aligned}$$

- c) Realizar un estudio del error y del orden de convergencia del método en el intervalo $[1, 10]$ completando las siguientes tablas e indicar el orden de precisión aproximado del método que se obtiene

N
Error con RK3

N vs. $2N$ vs. vs. vs.
Orden con RK3			

Indicar que norma vectorial se está usando.

- d) Determinar computacionalmente la restricción de estabilidad sobre el parámetro h cuando este método se aplica a este problema.

17. Consideremos el problema

$$y'(t) = -\lambda t y^2 + \lambda/t - 1/t^2, \quad 1 < t < 25, \quad y(1) = 1.$$

Tiene solución exacta $y = 1/t$. Usar los métodos de Euler explícito, implícito y Crank-Nicolson con los valores $\lambda = -1, -10, -50$ y distintos pasos de tiempo h . Aplicar Newton para resolver las ecuaciones de punto fijo que aparecen. Comentar los resultados.

18. **Bifurcación de Hopf:** El sistema

$$\begin{aligned}\frac{dx}{dt} &= a - x(t) - \frac{4x(t)y(t)}{1+x(t)^2} \\ \frac{dy}{dt} &= b x(t) \left(1 - \frac{y(t)}{1+x(t)^2}\right)\end{aligned}$$

aproxima una reacción química. Existe un parámetro $b_c = 3a/5 - 25/a$ tal que para $b > b_c$ las trayectorias de la solución decaen en amplitud y espiral hacia un punto fijo estable en el espacio de fases, mientras que si $b < b_c$ las trayectorias oscilan sin decaer y son atraídas por un ciclo límite estable. Esto es lo que se llama la bifurcación de Hopf.

- a) Usando Euler explícito tomar $a = 10$ y aproximar la solución que empieza con $x(0) = 0$ e $y(0) = 2$ para $t \in [0, 20]$ en los casos $b = 2$ y $b = 4$. En cada situación usar una ventana con dos gráficas para describir las observaciones. La primera gráfica debe contener las dos curvas $(t, x(t))$ y $(t, y(t))$ mientras que la segunda gráfica debe contener la curva $(x(t), y(t))$. Comprobar el comportamiento arriba descrito. Puedes usar **pplane.jar** para contrastar.
- b) Usa $a = 10$ y cualquiera de los métodos vistos con paso $h = 0.01$ para aproximar la solución que empieza con $x(0) = 0$ e $y(0) = 2$ para $t \in [0, 20]$. Hacerlo para $b = 2$ y $b = 4$. Para cada caso hacer gráficas (t, x) y (x, y) para describir las observaciones.
- c) Investigar lo que ocurre cerca del valor crítico $b_c = 3.5$, se puede tener que alargar el intervalo de integración temporal.

19. **Plano de fases:** El sistema lineal

$$\begin{aligned}\frac{d}{dt}x(t) &= ax(t) + by(t) \\ \frac{d}{dt}y(t) &= cx(t) + dy(t)\end{aligned}$$

aproxima (usando desarrollos de Taylor) el comportamiento de un sistema no lineal

$$\begin{aligned}\frac{d}{dt}x(t) &= f(x(t), y(t)) \\ \frac{d}{dt}y(t) &= g(x(t), y(t))\end{aligned}$$

que tenga un punto de equilibrio en $(0, 0)$. El objetivo de este ejercicio es describir el comportamiento de las soluciones del problema lineal de acuerdo a los autovalores de la matriz de coeficientes

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

- a) Determinar los valores de a, b, c, d en términos de las derivadas de f y de g .
- b) Programando Euler explícito crear una ventana con dos gráficas para describir las observaciones. La primera gráfica debe contener las dos curvas $(t, x(t))$ y $(t, y(t))$ mientras que la segunda gráfica debe contener la curva $(x(t), y(t))$. Mostrar todas las curvas de forma progresiva en tiempo para saber el sentido de crecimiento o decrecimiento de las mismas.

Para las siguientes matrices usar datos iniciales $x_0 = y_0 = -1 + j \cdot 0.2$ en la misma ventana

$$A_1 = \begin{pmatrix} 0.5 & 1 \\ -1 & 0.5 \end{pmatrix} \quad A_2 = \begin{pmatrix} -0.2 & 1 \\ -1 & -0.2 \end{pmatrix} \quad A_3 = \begin{pmatrix} -1 & -1 \\ 4 & 1 \end{pmatrix}$$

Describir los resultados obtenidos asociandolos con los autovalores de estas matrices. La situación de la matriz A_1 se denomina **fuente en espiral**, la de la matriz A_2 un **sumidero en espiral**, la de A_3 un **equilibrio central**. Usar **pplane.jar** para contrastar.

20. **Órbita plana:** Las ecuaciones de Newton para el movimiento de una partícula en órbita y con una excentricidad $e \in [0, 1)$ vienen dadas por

$$\begin{aligned} x''(t) &= -\frac{x(t)}{r(t)^3}, \quad x(0) = 1 - e, \quad x'(0) = 0 \\ y''(t) &= -\frac{y(t)}{r(t)^3}, \quad y(0) = 0, \quad y'(0) = \sqrt{\frac{1+e}{1-e}}, \end{aligned}$$

donde $r(t) = \sqrt{x(t)^2 + y(t)^2}$.

- a) Transformándolo a un sistema de primer orden resolverlo con el método de Euler en el intervalo $t \in [0, 20]$.
- b) Para excentricidad nula, $e = 0$, la solución exacta es

$$x(t) = \cos(t), \quad y(t) = \sin(t).$$

Verificar que para $e \in (0, 1)$ la solución a este problema se puede escribir en la forma

$$x(t) = \cos(u(t)) - e, \quad y(t) = \sqrt{1 - e^2} \sin(u(t))$$

donde $u(t)$ es la solución de la ecuación de Kepler.

$$u(t) - e \sin(u(t)) - t = 0.$$

- c) Usar el método de Newton para resolver la ecuación de Kepler y obtener la solución $(x(t), y(t))$ que tomaremos como exacta.
- d) Vamos a aumentar el orden de convergencia de Euler progresivo: el método clásico de Runge-Kutta (RK4) tiene orden cuatro y se obtiene promediando muestras de pendientes. Para un problema escalar se

describe como:

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}h k_1\right) \\ k_3 &= f\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}h k_2\right) \\ k_4 &= f\left(t_n + h, y_n + h k_3\right) \\ y_{n+1} &= y_n + \frac{h}{6} (k_1 + 2k_2 + 2k_3 + k_4) \end{aligned}$$

Implementarlo y comparar resultados.

- e) Calcular la evolución de la constante de error tanto para x como para y en el intervalo propuesto y verificar el orden de convergencia de los dos métodos.
21. Una masa m_1 está unida a un soporte rígido por medio de un resorte de constante elástica k_1 . Una segunda masa m_2 está unida a la primera mediante un segundo resorte de constante elástica k_2 . Las masas se desplazan por una superficie lisa horizontal siendo s_1 y s_2 los desplazamientos de las masas de las posiciones que ocupan cuando ambos resortes están relajados. El movimiento del sistema obedece las ecuaciones

$$\begin{aligned} m_1 s_1''(t) &= -k_1 s_1(t) + k_2(s_2(t) - s_1(t)) \\ m_2 s_2''(t) &= -k_2(s_2(t) - s_1(t)). \end{aligned}$$

Resolver el sistema de ecuaciones en $t \in [0, 20]$ en el caso donde $m_1 = m_2 = k_1 = k_2 = 1$ y con las condiciones iniciales $s_1(0) = 1$, $s_2(0) = 0$, $s_1'(0) = 0.5$ y $s_2'(0) = 1$.

22. La transferencia de calor por radiación se puede modelar por la ecuación

$$T'(t) = -\alpha(T^4(t) - T_a^4), \quad 0 < t$$

con el valor inicial $T(0) = 2500$ y siendo $\alpha = 4 \cdot 10^{-12}$ y $T_a = 250$. Comprobar que para $t = 10$ se tiene $T(10) = 1758.26337470$ como solución exacta. Sabiendo que la solución exacta se puede obtener de forma implícita como

$$\tan^{-1}\left(\frac{T(t)}{T_a}\right) - \tan^{-1}\left(\frac{T_0}{T_a}\right) + 0.5 \log\left[\frac{(T_0 - T_a)(T(t) + T_a)}{(T(t) - T_a)(T_0 + T_a)}\right] = 2\alpha T_a^3 t$$

usa el método de la secante para obtener los valores de $T(t)$ en los puntos t_n y considera estos valores como los valores exactos. A continuación realiza una gráfica donde se comparan los valores exactos con los obtenidos con métodos de orden 1, 2 y 4 en el intervalo $t \in [0, 5]$.

23. Aproximar mediante un método de Runge-Kutta de orden 4 la siguiente ecuación diferencial

$$u''(t) = -2u'(t) - u(t), \quad 0 < t < 4$$

con las condiciones iniciales $u(0) = 0, u'(0) = 1$ y comparar con la solución exacta $u(t) = te^{-t}$.

24. La transferencia de protones en una unión de moléculas de hidrógeno se describe por el sistema lineal

$$\begin{aligned} x'_1 &= -k_1 x_1 + k_2 y \\ x'_2 &= -k_4 x_2 + k_3 y \\ y' &= k_1 x_1 + k_4 x_2 - (k_1 + k_3) y \end{aligned}$$

donde

$$x_1(0) = 0, x_2(0) = 1, y(0) = 0$$

y hay que resolver este problema en el intervalo $0 \leq t \leq 8 \cdot 10^5$ siendo los coeficientes

$$k_1 = 8.4303270 \cdot 10^{-10}, \quad k_2 = 2.9002673 \cdot 10^{11},$$

$$k_3 = 2.4603642 \cdot 10^{10}, \quad k_4 = 8.7600580 \cdot 10^{-6}.$$

Usando Euler explícito observar el comportamiento de las distintas componentes ampliando zonas de cálculo si es preciso. ¿Existe alguna restricción necesaria sobre h para poder realizar el cálculo? Usa Euler implícito, Crank-Nicolson o RK4 usando notación matricial con Matlab.

25. Consideremos las ecuaciones del lanzamiento de un proyectil de masa constante

$$\begin{cases} x'(t) = v(t) \cos(\theta(t)), \\ y'(t) = v(t) \sin(\theta(t)), \\ v'(t) = \frac{1}{m}(T(t) - 0.5 \rho s v(t)^2) - g \sin(\theta(t)), \\ \theta'(t) = \frac{-g}{v(t)} \cos(\theta(t)). \end{cases}$$

Si suponemos un viento horizontal de velocidad $w(t)$ el arrastre aerodinámico es proporcional la cuadrado de la velocidad del proyectil con respecto al viento, esto es

$$T(t) = \frac{c\rho s}{2}((x'(t) - w(t))^2 + y'(t)^2)$$

donde $c = 0.2$ es el coeficiente de arrastre, $\rho = 1.29 \text{ Kg/m}^3$ es la densidad del aire y $s = 0.25 \text{ m}^2$ es el área seccional del proyectil. Resolverlo mediante Euler explícito usando los datos

$$g = 9.81 \text{ m/s}^2, \quad m = 15 \text{ Kg}, \quad v(0) = v_0 = 50 \text{ m/s}$$

con viento constante en contra dado por $w(t) = -10 \text{ m/s}$ y sin viento $w(t) = 0$. Dibujar las trayectorias resultantes para el ángulo de tiro $\theta(0) = \theta_0 = j\pi/18$ para $j = 1, 2, 3, 4, 5$. Para cada trayectoria indicar:

- a) Tiempo de vuelo, esto es, primer valor $t_* > 0$ tal que $y(t_*) \leq 0$.
- b) Distancia alcanzada.
- c) Velocidad de impacto.

26. Consideramos el problema de valor inicial

$$y'(t) = \frac{2}{\pi} \arctan(y(t)), \quad t \in [0, 1], \quad y(0) = 1.$$

- a) Encontrar cotas para $y''(t)$ e $y'''(t)$ en el intervalo $t \in [0, 1]$ sin hallar $y(t)$ explícitamente
- b) Si se resuelve este problema usando el método de Euler explícito con paso h constante sin cometer error en el valor inicial ¿Qué valor de $h > 0$ habrá que tomar para garantizar que el mayor error cometido sea menor que 10^{-3} ?

27. Siguiendo la segunda Ley de Newton, la ecuación que describe una masa que cuelga de un muelle, origen de coordenadas en el punto de sujeción del muelle y por lo tanto $z(t) < 0$, es

$$\begin{aligned} z''(t) &= \frac{k}{m}(l_{rest} - z(t)) - \frac{d}{m}z'(t) + g_z, \\ z(0) &= z_0, \quad z'(0) = v_0. \end{aligned}$$

en donde la masa de la partícula es m , la rigidez del muelle viene dada por la constante k , la longitud de reposo del muelle es l_{rest} y se incluyen dos efectos: la fuerza de la gravedad g_z y el amortiguamiento producido por el rozamiento al moverse, $-\frac{d}{m}z'(t)$, siendo la constante que caracteriza este amortiguamiento d . Normalmente es $k \gg d$

- a) Obtener la solución analítica de esta ecuación
- b) Reescribir la ecuación como un sistema de primer orden usando una formulación matricial.

c) Para los valores $m = 1$, $k = 1000$, $l_{rest} = -1$, $d = 10$, $g_z = 10$, $z_0 = -1$ y $v_0 = -5$ aplicar el método de Euler explícito e implícito y comparar resultados.

28. La ecuación de Van der Pol se puede reescribir como sistema

$$\begin{aligned} y'_1(t) &= y_2(t) \\ y'_2(t) &= (1 - y_1(t)^2)y_2(t) - y_1(t). \end{aligned}$$

Usando datos iniciales

$$\begin{aligned} y_1(0) &= 2.00861986087484313650940188 \\ y_2(0) &= 0 \end{aligned}$$

comprobar que se obtiene una solución periódica de periodo

$$T = 6.6632868593231301896996820305.$$

29. Dado $\alpha \in (-\pi/2, \pi/2)$ la ecuación de un péndulo viene dada por

$$y''(t) + \sin(y(t)) = 0, \quad y(0) = \alpha, \quad y'(0) = 0.$$

Usando el método de Euler explícito aproximar la solución de la ecuación (escribirla primero como una ecuación de primer orden). En Mecánica se conoce que el primer valor conocido donde $y'(t) = 0$ (el semiperíodo) viene dado con gran aproximación por la fórmula

$$\pi \left(1 + \frac{1}{4} \sin^2\left(\frac{\alpha}{2}\right) + \frac{9}{64} \sin^4\left(\frac{\alpha}{2}\right) \right).$$

Usar este resultado para comprobar la precisión del método conforme se cambia la longitud del paso.

30. Para el sistema lineal $x'(t) = Ax(t) + G(t)$, ($t > 0$), $x(0) = (1, 1)'$ y donde

$$A = \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \quad G(t) = \begin{pmatrix} -2e^{-t} + 2 \\ -2e^{-t} - 1 \end{pmatrix}$$

sabemos que la solución verdadera es $x(t) = (e^{-t}, 1)'$. Escribir las ecuaciones explícitamente y usar los métodos de Euler explícito, implícito y Crank-Nicolson en $[0, 10]$ para los valores de h dados por $h = 0.1, 0.05, 0.0025$.

31. Para el sistema lineal $x'(t) = Ax(t)$ donde

$$A = \begin{pmatrix} -82.72 & -58.03 & -49.82 \\ -58.03 & -40.83 & -34.88 \\ -49.82 & -34.88 & -30.44 \end{pmatrix}.$$

queremos aplicar el método de Euler progresivo. Dar la restricción sobre el paso de tiempo para obtener un cálculo estable.

32. Resolver el problema

$$y'(t) = \lambda y(t) + \frac{1}{1+t^2} - \lambda \operatorname{atan}(t), \quad y(0) = 0$$

que tiene como solución $y(t) = \operatorname{atan}(t)$. Usar los métodos de Euler explícito, implícito, Heun y de Crank-Nicolson con los valores $\lambda = -1, -10, -50$ y distintos pasos de tiempo h . Comentar los resultados en términos de la dependencia con respecto a λ de las siguientes cuestiones

- a) Valores óptimos o críticos de h para evitar inestabilidades numéricas (dado N , $h=T/N$), ampliar el rango de valores de h si es necesario.
- b) Orden de convergencia.
- c) Fijado un error global, que esfuerzo (en términos de h o N) necesita cada método para alcanzarlo.

33. El problema

$$y'(t) = \lambda y + (1 - \lambda) \cos(t) - (1 + \lambda) \sin(t), \quad y(0) = y_0.$$

Tiene solución exacta $y(t) = e^{\lambda t} (y_0 - 1) + \sin(t) + \cos(t)$. Resolver este problema en el caso $y_0 = 1$ usando los métodos de Euler explícito e implícito con distintos valores de λ y de h para $0 < t < 5$

- a) $\lambda = -1$; $h = 0.5, 0.25, 0.125$, ($T = 5$, luego equivale a dar $N = 10, 20, 40$)
- b) $\lambda = 1$; $h = 0.5, 0.25, 0.125$,
- c) $\lambda = -5$; $h = 0.5, 0.25, 0.125, 0.0625$,
- d) $\lambda = 5$; $h = 0.5, 0.25, 0.125, 0.0625$

Comentar los resultados en términos de la dependencia con respecto a λ de las siguientes cuestiones

- a) Valores óptimos o críticos de h para evitar inestabilidades numéricas (dado N , $h=T/N$), ampliar el rango de valores de h si es necesario.
- b) Orden de convergencia.
- c) Fijado un error global, que esfuerzo (en términos de h o N) necesita cada método para alcanzarlo.

34. La importancia del orden de convergencia de un método numérico se puede ver en la Figura 6.6 generada approximando la solución de la ecuación diferencial

$$\begin{cases} \frac{dy}{dt} = k \cos(kt) y(t), & t > 0, \\ y(0) = y_0, \end{cases} \quad (6.5)$$

para $k = 7$, $y_0 = 2$ y considerando el intervalo de tiempo $[0, 20]$ (solución exacta $y(t) = y_0 e^{\sin(k t)}$). Reproducir tablas similares a las del ejercicio previo para los métodos RK2 y RK4 y determinar el orden de convergencia de los mismos. Para inicializar estos métodos usar $y_0 = y(0)$.

35. El sistema

$$\begin{cases} y'_1(t) &= y_2, & y_1(0) = 1 \\ y'_2(t) &= -1000y_1 - 1001y_2, & y_2(0) = -1 \end{cases}$$

posee solución exacta $y_1(t) = -y_2(t) = e^{-t}$ mientras que la solución general es $y_1(t) = -y_2(t) = A e^{-t} + B e^{-1000t}$, dando lugar a distintas escalas y a un problema rígido. Comprobar que el método clásico de Runge-Kutta de cuarto orden en $[0, 3]$ debe usar $h < h_\star \sim 0.0027$.

36. **Velocidad terminal** Una partícula esférica de radio a y con uniforme densidad ρ_s cae por acción de la gravedad en un fluido de densidad ρ_f y con viscosidad μ . (King et al [19]) El movimiento de la esfera se puede describir por la ecuación

$$\frac{4}{3}\pi a^3 \rho_s \frac{dv}{dt} + 6\mu\pi av - \frac{4}{3}\pi a^3 (\rho_s - \rho_f)g = 0$$

usando $v = z'(t)$, las condiciones asociadas serían $z(0) = 0$ y $v(0) = z'(0) = 0$ donde $z(t)$ es la distancia de la esfera en sentido de caída. Los términos se describen como sigue:

- El primer término a la izquierda es la aceleración.
- El segundo describe la fuerza de arrastre que es proporcional a la velocidad de la esfera.
- El tercer término es la fuerza debida a la gravedad menos la fuerza de flotación.

Los valores constantes son:

$$a = 10^{-4}m, \rho_s = 1.1g/cm^3, \rho_f = 1g/cm^3, g = 9.81m/s^2, \mu = 0.035.$$

Resolver computacionalmente este problema y obtener la velocidad terminal de la esfera.

37. Para $t > 0$ consideramos el sistema lineal de edos

$$\begin{cases} x'_1(t) &= -298x_1(t) + 99x_2(t), \\ x'_2(t) &= -594x_1(t) + 197x_2(t). \end{cases} \quad (6.6)$$

Para los datos iniciales $x_1(0) = -1/2$ y $x_2(0) = 1/2$ calcular la solución exacta. Aproximarla numéricamente usando Euler explícito e implícito reproduciendo las gráficas de la Figura 2.16. Identificar el modo rápido y el lento

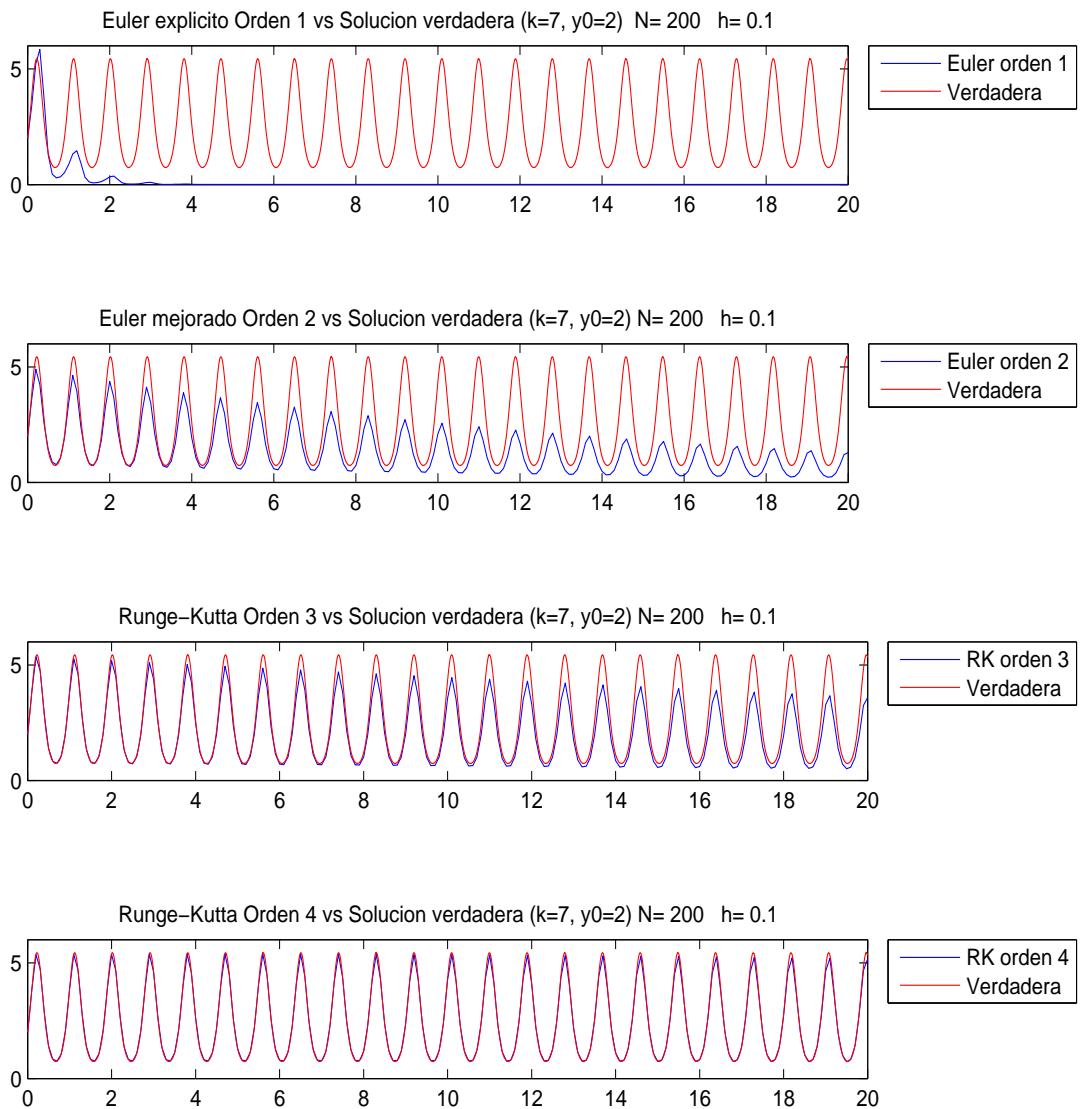


Figura 6.6: Solución verdadera y aproximada para $k = 7$, $y_0 = 2$, $T = 20$ y $h = 0.1$ o bien $N = 200$.

38. El siguiente ejemplo propuesto por Dahlquist y Björck en 1974 ilustra bien este tipo de fenómenos:

$$\begin{cases} y'(t) = 100(\sin(t) - y(t)), & 0 < t \leq 3, \\ y(0) = 0. \end{cases}$$

a) Comprobar que la solución de la edo

$$\begin{cases} y'(t) = a(\sin(t) - y(t)), & 0 < t, \\ y(0) = y_0 \end{cases}$$

es

$$y(t) = e^{-at}y_0 + \frac{\sin(t) - a^{-1}\cos(t) + a^{-1}e^{-at}}{1 + a^{-2}}.$$

Para $a >> 1$ el modo transitorio exponencial e^{-at} decae muy rápido. Este modo debe ser capturado bien por el método explícito y esto nos lleva a un valor de h excesivamente pequeño.

- b) Reproducir las gráficas de las Figuras 2.12, 2.13 y explicar la razón del valor crítico de N en torno a 155 para Euler explícito para $a = 100$. Observa las inestabilidades numéricas que se obtienen con valores de $h = 3/N$ dados por $h = 0.01 \sim N = 300$, $h = 0.015 \sim N = 200$ $h = 0.02 \sim N = 150$, $h = 0.02027.. \sim N = 148$ y $h = 0.02068.. \sim N = 145$ para resolver el problema.
- c) Observa la rigidez de la ecuación cuando se resuelve el mismo problema pero con dato inicial $y(0) = 2$ por ejemplo.
- d) Observar analíticamente y computacionalmente que se tiene el mismo resultado cualitativo con el ejemplo más simple

$$\begin{cases} y'(t) = a(1 - y(t)), & 0 < t, \\ y(0) = y_0 \end{cases}$$

lo que indica que el culpable del comportamiento es el término $-ay(t)$ para valores crecientes de $a > 0$.

39. La siguiente ecuación fue propuesta por Prothero-Robinson:

$$\begin{cases} y'(t) = L(\varphi(t) - y(t)) + \varphi'(t), & 0 < t, \\ y(0) = y_0 \end{cases}$$

la solución exacta es

$$y(t) = e^{-Lt}(y_0 - \varphi(0)) + \varphi(t)$$

y otra vez el modo rápido e^{-Lt} para $L >> 1$ debe ser capturado correctamente. Aquí $\varphi(t)$, es el modo estacionario, y puede ser una función suave sin cambios

bruscos. Incluso en el caso en el que $y_0 = \varphi(0)$ y aparentemente el modo rápido no está en la solución, sí que se encuentra en todas las soluciones vecinas y se debe también capturar como si estuviese presente.

Diseñar un experimento numérico que cumpla $y_0 = \varphi(0)$ y realizar una gráfica donde se vea la solución estacionaria $\varphi(t)$ y cómo todas las soluciones vecinas se aproximan exponencialmente a ella, esto es, reproducir las gráficas de las notas de clase para este ejemplo, ver los ejemplos en las Figuras 2.14 y 2.15.

40. *Al encender una cerilla la bola de fuego crece hasta alcanzar un estado estacionario donde se equilibra al absorción de oxígeno del exterior con el consumo interior. Un modelo simple es el propuesto por Larry Shampine (autor de las librerías de edos para MATLAB, OCTAVE y SCILAB entre otros)*

$$\begin{cases} y'(t) &= y(t)^2 - y(t)^3, & 0 < t < 2/\delta, \\ y(0) &= \delta \end{cases}$$

en donde $y(t)$ es el radio de la bola de fuego, $\delta > 0$ es un radio inicial pequeño y los términos $y(t)^2$ e $y(t)^3$ están relacionados con la superficie y el volumen de la bola de fuego.

El parámetro crítico es el radio inicial $\delta > 0$ que es pequeño y el fenómeno físico de interés ocurre en el tiempo $t_ \approx 1/\delta$. Para $0 < t < 1/\delta$ se observa un crecimiento moderado del radio y un crecimiento repentino en torno a $t_* \approx 1/\delta$ para llegar al valor $y(t) \approx 1$ en donde se estabiliza.*

Reproducir las gráficas de las Figuras 6.7 y 6.8 que se presentan sobre este modelo. La ecuación nolineal que surge en el método de Euler implícito se debe resolver por el Método de Newton y usando como valor inicial la aproximación dada por el método de Euler explícito.

41. *Para $t > 0$ consideramos el sistema lineal de edos*

$$\begin{cases} x'_1(t) &= -500 x_1(t) + 6889 x_2(t), \\ x'_2(t) &= 36 x_1(t) - 500 x_2(t). \end{cases}$$

Para los datos iniciales $x_1(0) = 83$ y $x_2(0) = 6$ posee solución exacta $x_1(t) = 83e^{-2t}$ y $x_2(t) = 6e^{-2t}$ pero es un sistema rígido. Indicar por qué.

Aproximar numéricamente la solución exacta usando Euler explícito, Euler implícito y Crank-Nicolson. Realizar un estudio comparativo de los errores y los órdenes de convergencia de estos tres métodos en el intervalo temporal $[0, 3]$.

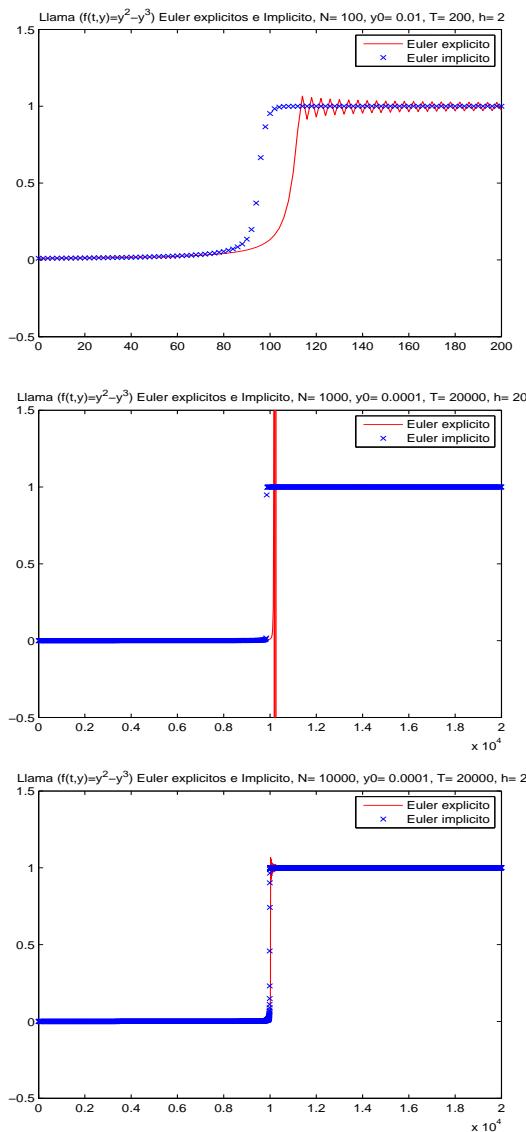


Figura 6.7: Modelo llama de Shampine. Calculos sobre el intervalo temporal $[0, 2\delta^{-1}]$

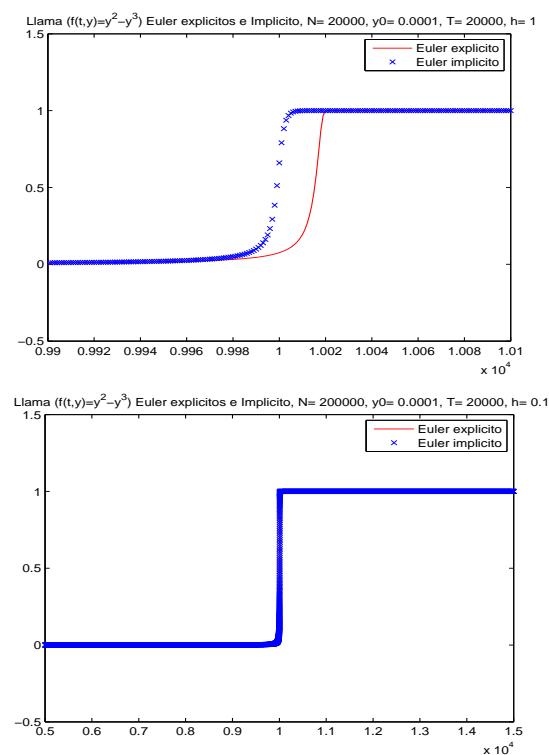


Figura 6.8: Modelo llama de Shampine. Calculos sobre el intervalo temporal $[0, 2\delta^{-1}]$

42. Usando el método de Runge-Kutta clásico de cuarto orden resolver el problema rígido

$$\begin{cases} y'(t) = 1000(\sin(t) - y(t)), & 0 < t \leq 3, \\ y(0) = 0. \end{cases}$$

Realizar una tabla con los errores en $T = 3$ de acuerdo a distintos valores de $N = 3/h$ como, por ejemplo, $N = 100, 200, 300, \dots$. Procurar que la tabla sea representativa de como se comporta el método y sacar conclusiones.

43. El sistema

$$\begin{cases} u' = 9u + 24v + 5\cos(t) - \frac{1}{3}\sin(t), & u(0) = 4/3 \\ v' = -24u - 51v - 9\cos(t) + \frac{1}{3}\sin(t), & v(0) = 2/3 \end{cases}$$

tiene como solución

$$\begin{aligned} u(t) &= 2e^{-3t} - e^{-39t} + \frac{1}{3}\cos(t) \\ v(t) &= -e^{-3t} + 2e^{-39t} - \frac{1}{3}\cos(t) \end{aligned}$$

Usando el método de Runge-Kutta de cuarto orden realizar un estudio comparativo de los errores y los órdenes de convergencia en el intervalo temporal $[0, 1]$ y para $h = 0.1, 0.05, 0.025$.

44. El problema

$$y' = 5e^{5t}(y - t)^2 + 1, \quad y(0) = -1$$

tiene por solución $y(t) = t - e^{-5t}$. Realizar un estudio comparativo de los errores y los órdenes de convergencia en el intervalo temporal $[0, 1]$ usando el método de Runge-Kutta clásico de cuarto orden y el método de Crank-Nicolson para valores de $h = 0.25$ y $h = 0.2$. Para resolver los problemas implícitos usar el método de Newton con valor inicial el que genera el método de Euler explícito.

45. La solución de la ecuación

$$\begin{cases} \frac{d}{dt}y(t) = -200t y^2, & -1 < t < 1, \\ y(-1) = 101^{-1}, \end{cases}$$

viene dada por $y(t) = (1 + 100t^2)^{-1}$. Observar la precisión del método de Runge-Kutta clásico en $t = 0$ donde sabemos que $y(0) = 1$. ¿Qué diferencia se encuentra si tomamos $-3 < t < 3$ con $y(-3) = 901^{-1}$? ¿Qué aproximación a $y(0)$ se encuentra si usamos paso adaptativo con RKF 4(5)?

46. La solución general de la ecuación

$$\begin{cases} \frac{d}{dt}y(t) = 10y + 11t - 5t^2 - 1, & 0 < t \leq 3, \\ y(0) = y_0, \end{cases}$$

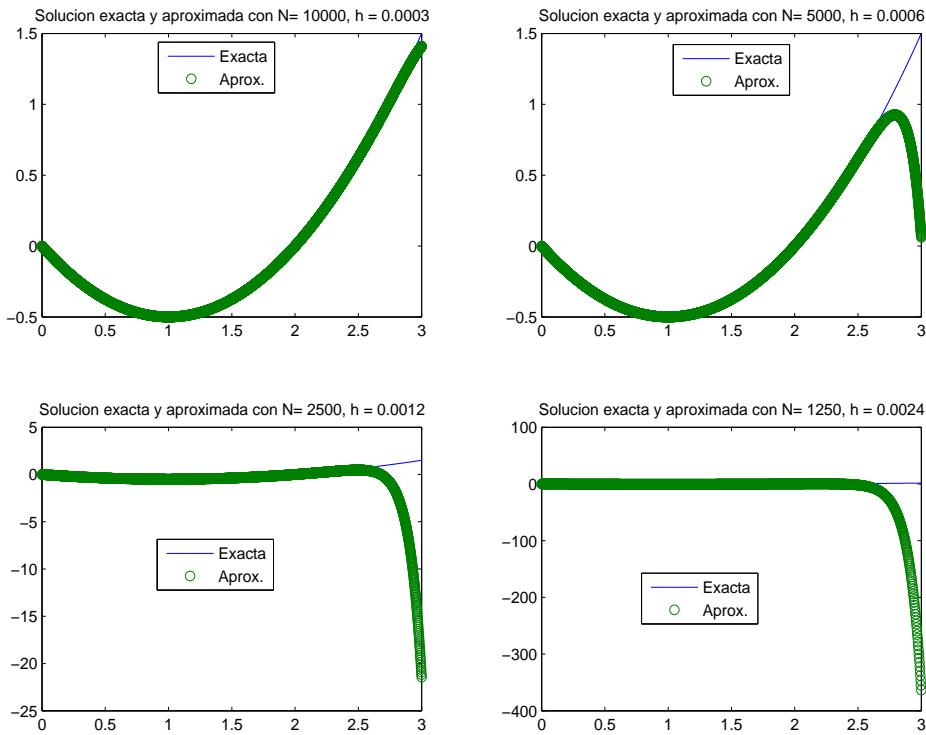


Figura 6.9: Soluciones exacta y aproximada para distintos valores de h

viene dada por $y(t) = y_0 e^{10t} + t^2/2 - t$ por lo que tenemos un campo de trayectorias muy expansivo.

- Si tomamos $y_0 = 0$ tenemos por solución exacta $y(t) = t^2/2 - t$ que es muy inestable. Obtén los resultados gráficos de la Figura 6.9 comparando la solución exacta y la aproximada. Observa como se debe restringir h para poder obtener la solución dependiendo del método que uses.
- Para un valor de h lo suficientemente pequeño como para resolver bien el problema en el caso $y(0) = \epsilon$, por ejemplo $h = 10^{-4}$, observa la discrepancia entre las soluciones numéricas obtenida para $\epsilon > 0$ y la exacta del problema original para $\epsilon = 0$. ¿Qué se puede decir de la sensibilidad del problema respecto al dato inicial?

6.1. Práctica Computacional 1

Nota: Redactar un documento adjuntando gráficas detalladas. Enviar códigos comentados a la dirección de correo eliseo@um.es.

1. La ecuación diferencial

$$\begin{cases} \frac{dy}{dt} = t^2 - ay, & t > 0, \\ y(0) = y_0 \end{cases}$$

tiene por solución

$$y(t) = e^{-at}(y_0 - 2/a^3) + (a^2t^2 - 2at + 2)/a^3.$$

- a) Para $a = 1$ e $y_0 = 1$ aproximar la solución en el intervalo $[0, 1]$ usando el método de Crank-Nicolson (observar que $f(t, y)$ es lineal con respecto a la variable y), el método de Heun, RK3 y el método clásico de Runge-Kutta de cuarto orden.

En cada método y con N dado calculamos el error absoluto entre el vector de la solución exacta y la calculada, a continuación repetimos el cálculo con $2N$. Reproduce y completa las siguientes tablas

N	10	100	500	1000
Error con CN				
Error con Heun	0.0027	2.5587e-005	1.0199e-006	2.5486e-007
Error con RK3				
Error con RK4	1.0506e-006	1.0144e-010	1.6243e-013	1.0214e-014

N vs. $2N$	10 vs. 20	100 vs. 200	500 vs. 1000	1000 vs. 2000
Orden con CN				
Orden con Heun	2.0317	2.0032	2.0006	2.0003
Orden con RK3				
Orden con RK4	4.0279	4.0029	3.9912	1.6656

¿Qué observas y qué explicación puedes dar?

- b) Para $a = 50$ e $y_0 = 3$ aproximar la solución en el intervalo $[0, 5]$ usando el método de Euler explícito y el método clásico de Runge-Kutta de cuarto orden.
- c) Estimar computacionalmente la restricción necesaria sobre h para que cada esquema empiece a generar resultados cualitativos similares a la solución exacta.

- d) Realizar una gráfica donde se comparan de forma simultanea el decaimiento de los errores de ambos métodos para los mismos valores de h .
- e) Realizar una gráfica donde se visualice el orden de convergencia, rectas de pendiente, de ambos métodos de forma simultanea.

6.1.1. Práctica Computacional 1: Soluciones

Respuestas:

1. Para la restricción necesaria sobre h para que cada esquema empiece a generar resultados cualitativos similares a la solución exacta ver Figura 6.10 para Euler explícito y Figura 6.11 para RK4.
2. Para realizar una gráfica donde se comparan de forma simultanea el decaimiento de los errores de ambos métodos para los mismos valores de h ver como hacerlo bien en Figura 6.12 usando escala logarítmica y ver como no se visualiza correctamente en Figura 6.13 al no usar escala logarítmica. Ver también Figura 6.14.
3. Para realizar una gráfica donde se visualice el orden de convergencia, rectas de pendiente, de ambos métodos de forma simultanea ver Figura 6.15.

Observación: En las Figuras 6.16 y 6.17 se observa el comportamiento de los errores cuando el método numérico converge pero a una curva distinta a la buscada.

La razón es que se están calculando los errores con respecto a una curva solución ligeramente distinta a la buscada (por un error se ha tomado $2/a^3 = 0$ en la expresión analítica). Llega un momento donde no hay más decaimiento del error y no hay convergencia. Por poner un simil, el método numérico pasa de largo esta solución ligeramente distinta y va hacia la solución a la que realmente converge.

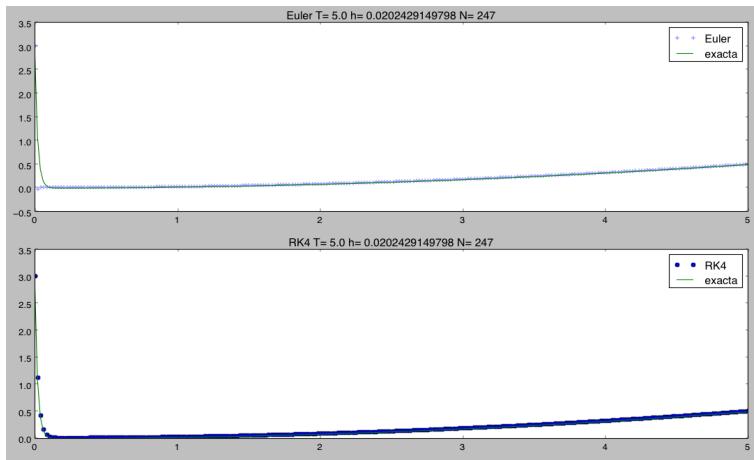


Figura 6.10: Cálculos con Euler y Runge-Kutta orden 4. Para $a = 50$ y Euler se observa el valor crítico $h < 1/a$ correspondiente aproximadamente a $h \approx 1/50 = 0.02$. Es decir, ya que $h = T/N$ y $T = 5$, $N \approx 250 = T/0.02 = 5/0.02$. Por encima de este valor de h Euler explícito es inestable. Por otro lado, RK4 reproduce fielmente la solución.

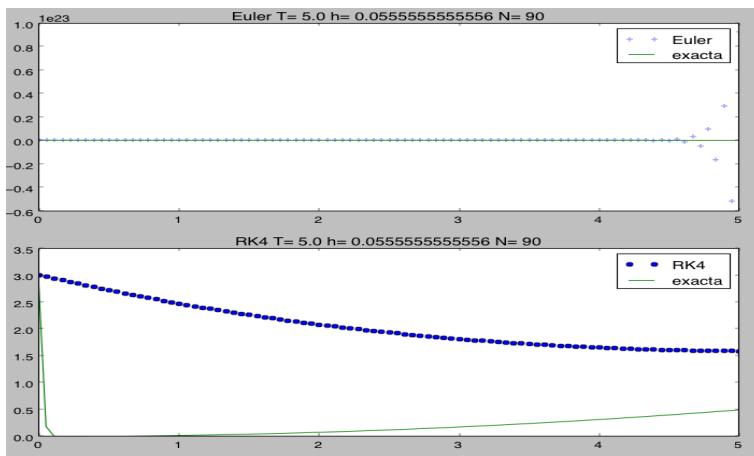


Figura 6.11: Cálculos con Euler y Runge-Kutta orden 4. Para $a = 50$ y Runge-Kutta se observa el valor crítico $h < 2.785/a$ predicho en teoría y correspondiente aproximadamente a $h \approx 2.785/50 = 0.055$. Es decir, ya que $h = T/N$ y $T = 5$, $N \approx 90 = T/0.055 = 5/0.055$. Por encima de este valor de h RK4 es inestable. Por otro lado, Euler explícito es totalmente inestable.

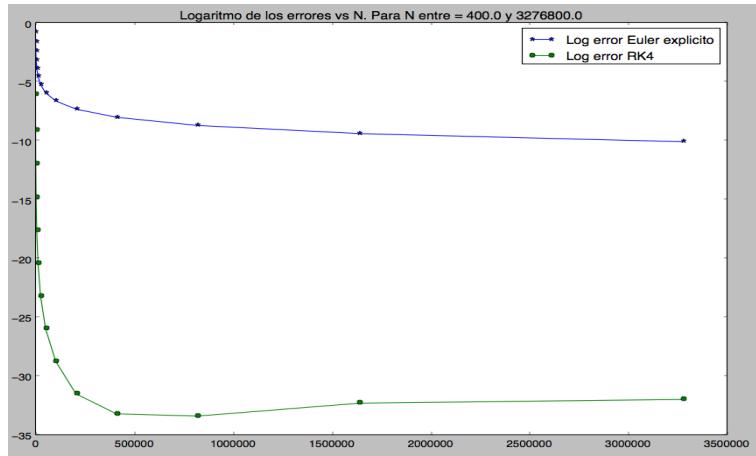


Figura 6.12: Puesto que los errores $E(h)$, o $E(N)$, son valores muy pequeños, una forma de ampliar es usar $\log(E(N))$. Entonces, visualizamos los pares $\{(N, \log(E(N)))\}_N$ para cada método y así podemos comparar el decaimiento. Observar que $\log(1e-13) \approx -29.9$. También se ve como los errores para RK4 dejan de decrecer cuando se alcanza la precisión de la máquina, es decir, $\log(E(N)) \approx -34$

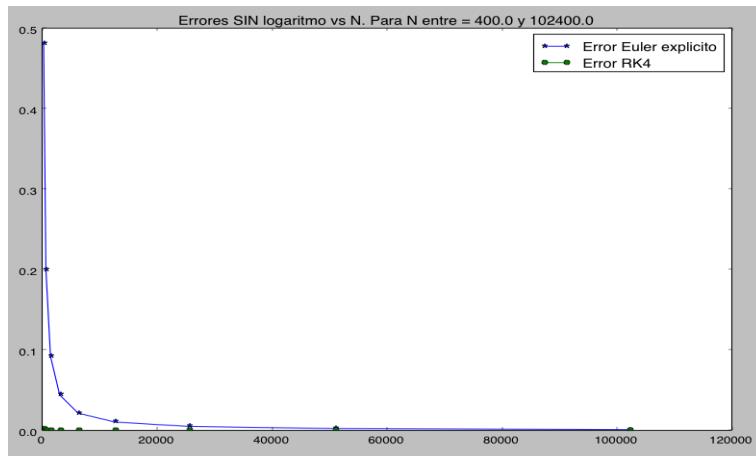


Figura 6.13: Los errores $E(h)$, o $E(N)$ pueden estar en órdenes de magnitud dispares, desde 10^{-1} a 10^{-13} . Es difícil visualizar todas estas escalas al mismo tiempo y por eso es mejor representar $\log(E(N))$ que disminuye el decaimiento de los valores.

```

Para h= 0.0125 error Euler Explicito = 0.480784846374
Para h= 0.0125 error RK4 = 0.00231002503006
Para h= 0.00625 error Euler Explicito = 0.199963985556
Para h= 0.00625 error RK4 = 0.000113680856275
Para h= 0.003125 error Euler Explicito = 0.0923768445159
Para h= 0.003125 error RK4 = 6.23299164948e-06
Para h= 0.0015625 error Euler Explicito = 0.0445520994812
Para h= 0.0015625 error RK4 = 3.65642259581e-07
Para h= 0.00078125 error Euler Explicito = 0.0219123784481
Para h= 0.00078125 error RK4 = 2.21197189454e-08
Para h= 0.000390625 error Euler Explicito = 0.0108664610302
Para h= 0.000390625 error RK4 = 1.36030209319e-09
Para h= 0.0001953125 error Euler Explicito = 0.00541092504604
Para h= 0.0001953125 error RK4 = 8.43303205045e-11
Para h= 9.765625e-05 error Euler Explicito = 0.00269993057929
Para h= 9.765625e-05 error RK4 = 5.24935650503e-12
Para h= 4.8828125e-05 error Euler Explicito = 0.00134859179436
Para h= 4.8828125e-05 error RK4 = 3.27737836869e-13
Para h= 2.44140625e-05 error Euler Explicito = 0.000673952809005
Para h= 2.44140625e-05 error RK4 = 2.07611705605e-14
Para h= 1.220703125e-05 error Euler Explicito = 0.000336890659755
Para h= 1.220703125e-05 error RK4 = 3.77475828373e-15
Para h= 6.103515625e-06 error Euler Explicito = 0.000168423904567
Para h= 6.103515625e-06 error RK4 = 3.10862446895e-15
Para h= 3.0517578125e-06 error Euler Explicito = 8.42065981201e-05
Para h= 3.0517578125e-06 error RK4 = 9.32587340685e-15
Para h= 1.52587890625e-06 error Euler Explicito = 4.21019606245e-05
Para h= 1.52587890625e-06 error RK4 = 1.28785870857e-14

```

Figura 6.14: Resultados de los errores para los cálculos con Euler y Runge-Kutta orden 4 con distintos valores de h . Se observa perfectamente la división por 2 o por 16 de cada error cuando se pasa de h a $h/2$ o bien de N a $2N$. Este patrón **se pierde en RK4 cuando el error es del orden de $1e-15$** . De hecho, se pasa de $3.1...e-15$ a $1.28...e-14...<$ ha crecido el error!

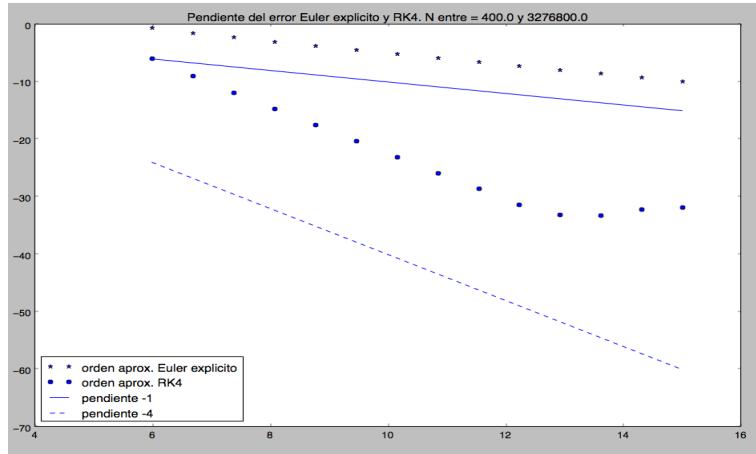


Figura 6.15: Visualizamos los pares $\{(\log(N), \log(E(N)))\}_N$ para estimar las pendientes para los ordenes de los errores. Para $N = 1638400$, es decir, $\log(N) \approx 14.3$ y un $E(N)$ tal que $\log(E(N)) \approx -30$, es decir, $E(N) \approx 1e-15$, ya se pierde cualquier patrón de decaimiento por falta de resolución del computador. Se observa perfectamente en el método RK4 puesto que tiene el decaimiento del error más pronunciado.

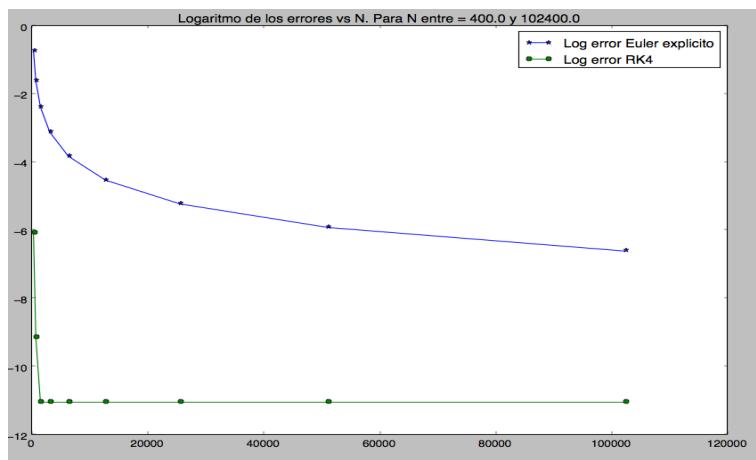


Figura 6.16: Comportamiento de los errores cuando No se converge a la solución exacta buscada pero Sí a una muy cercana: error constante luego no hay convergencia. Se visualiza antes con RK4 por disminuir los errores más rápido.

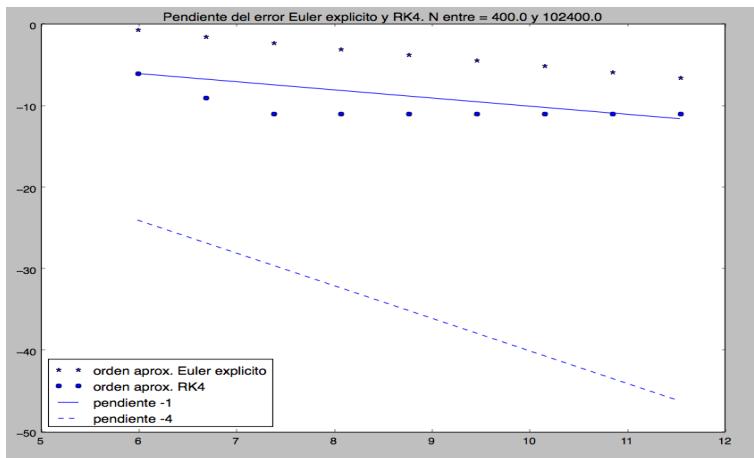


Figura 6.17: Comportamiento de los órdenes cuando No se converge a la solución exacta buscada pero Sí a una muy cercana. Se visualiza antes con RK4 por disminuir los errores más rápido.

6.2. Práctica Computacional 2

Instrucciones: Redactar un documento adjuntando gráficas detalladas. Enviar códigos comentados a la dirección de correo eliseo@um.es.

1. Consideramos una masa que cuelga de un muelle y tomamos el origen de coordenadas en el punto de sujeción del mismo. Si describimos la posición vertical del muelle por $y(t)$ ($y(t) < 0$) y seguimos la segunda Ley de Newton, entonces la ecuación que describe la posición $y(t)$ es

$$\begin{aligned} y''(t) &= \frac{k}{m}(l_{rest} - y) - \frac{d}{m}y'(t) + g_y, \\ y(0) &= y_0, \quad y'(0) = v_0. \end{aligned}$$

en donde la masa de la partícula es m , la rigidez del muelle viene dada por la constante k , la longitud de reposo del muelle es l_{rest} y se incluyen dos efectos: la fuerza de la gravedad g_y y el amortiguamiento producido por el rozamiento al moverse, $-\frac{d}{m}y'(t)$, siendo la constante que caracteriza este amortiguamiento d . Normalmente es $k > d$.

- a) La solución analítica de esta ecuación es

$$y(t) = (l_{rest} + g_y m/k) + e^{-\frac{d}{2m}t}(\alpha \cos(R t) + \beta \sin(R t))$$

siendo

$$\alpha = y_0 - (l_{rest} + g_y m/k), \quad \beta = \frac{v_0 + \alpha d/(2m)}{R}$$

con

$$R = \sqrt{k/m - (d/(2m))^2}.$$

- b) ¿Cuales son los valores límite de $y(t)$ y de $y'(t)$ cuando $t \rightarrow +\infty$ en términos de los parámetros del problema?
- c) Reescribir la ecuación como un sistema de primer orden. Usar también una formulación matricial.
- d) Para $m = 1$, $k = 5000$, $l_{rest} = -1$, $d = 10$, $g = 10$, $y_0 = -1$, $v_0 = -5$ y $t \in [0, 1]$ aplicar los métodos de Euler explícito e implícito.
- e) Estimar computacionalmente la restricción necesaria sobre h para que cada esquema empiece a generar resultados cualitativos similares. Indicar la equivalencia con respecto al número de puntos a usar en la partición del intervalo.
- f) Reproducir las figuras que se presentan en la página siguiente donde se ve la solución exacta y la calculada.

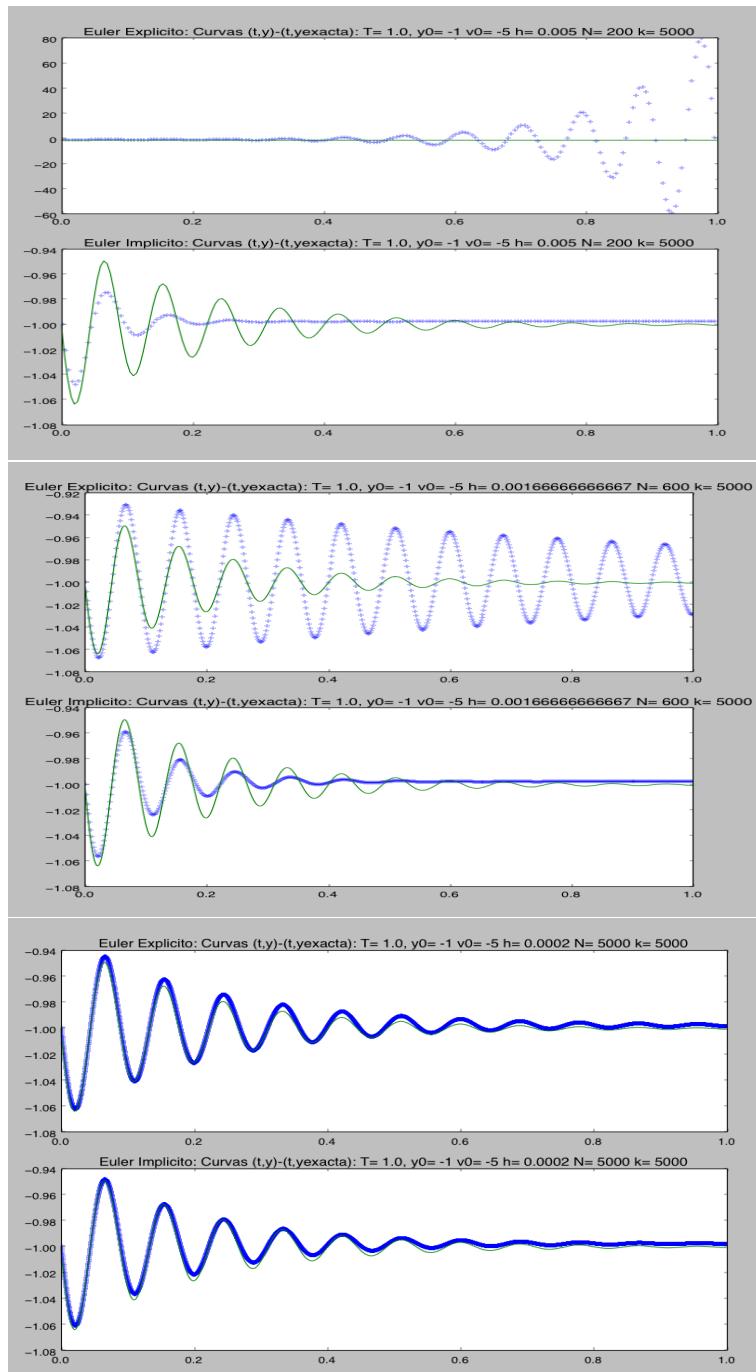


Figura 6.18: Resultados para distintos valores de h donde se ve la solución exacta y la calculada.

6.3. Práctica Computacional 3

Instrucciones: Redactar un documento adjuntando gráficas detalladas. Enviar códigos comentados a la dirección de correo eliseo@um.es.

1. **Órbitas de Arenstorf: Problema de los tres cuerpos.** Un problema básico en astronomía es el problema de los tres cuerpos, la información que sigue está extraída del libro de Hairer-Norsett y Wanner, ver en la bibliografía del curso.

Para fijar ideas, se trata de estudiar el movimiento de un satélite artificial causado por los campos gravitatorios de la tierra y de la luna suponiendo que este movimiento se realiza en un plano. Para empezar, se consideran dos cuerpos de masa μ (la luna) y $\tilde{\mu} = 1 - \mu$ (la tierra) en rotación en un plano y un tercer cuerpo de masa despreciable que se mueve en ese plano. El objetivo es determinar la trayectoria de un objeto ligero (un satélite) en presencia de estos dos cuerpos muy pesados.

Para un movimiento planar, las ecuaciones del cuerpo ligero son

$$\begin{aligned} x''(t) &= x + 2y' - \tilde{\mu} \frac{(x + \mu)}{r^3} - \mu \frac{(x - \tilde{\mu})}{s^3}, \\ y''(t) &= y - 2x' - \tilde{\mu} \frac{y}{r^3} - \mu \frac{y}{s^3} \end{aligned}$$

donde

$$r = \sqrt{(x + \mu)^2 + y^2}, \quad s = \sqrt{(x - \tilde{\mu})^2 + y^2}.$$

El objetivo de la práctica es resolver este problema correspondiente a la Tierra y la Luna como objetos pesados. En este caso es $\mu = 0.0122774714$ y los datos iniciales son

$$\begin{aligned} (x(0), y(0)) &= (0.994, 0), \\ (x'(0), y'(0)) &= (0, -2.0015851063790825224) \end{aligned}$$

Se debe de obtener un movimiento periódico de periodo

$$T = 17.06521656015796.$$

Usar Euler explícito con paso $h = T/24000$ y RK4 clásico con paso $h = T/6000$. La elección de estos dos pasos se realiza de manera que el trabajo numérico sea equivalente puesto que, en cada paso de RK4 se usan cuatro evaluaciones mientras que sólo una evaluación en cada paso de Euler explícito.

Dibujar las trayectorias (x, y) para cada método de forma similar a las figuras que se adjuntan.

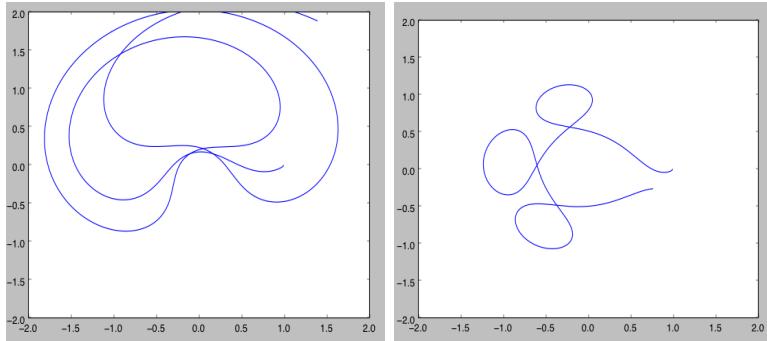


Figura 6.19: Órbitas (x,y) del objeto ligero obtenidas con Euler explicito (izquierda) y RK4 (derecha) usando los datos del ejercicio.

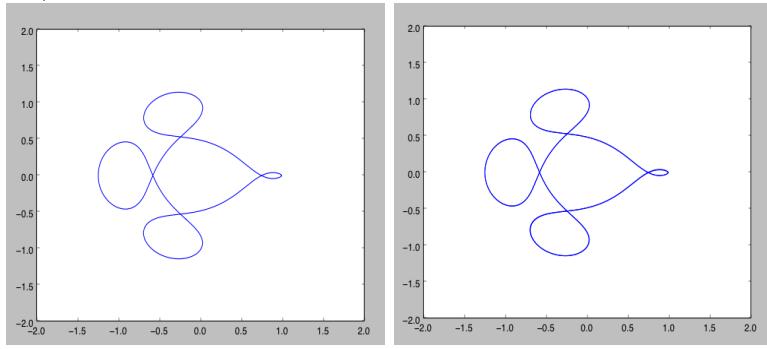


Figura 6.20: Una órbita (x,y) del objeto ligero (izquierda) ya si cerrada obtenida con RK4 usando $N=20000$ puntos para el periodo T anunciado. A la derecha se pueden ver dos órbitas, correspondiente a $2T$.

Se obtiene solución periódica en alguno de los dos métodos reduciendo el paso h ? Entenderemos que es periódica si la curva (x,y) se mantiene cerrada.

La existencia de éstas soluciones periódicas ha fascinado a astrónomos y matemáticos por décadas. Sir George Darwin (1898) realizó cálculos numéricos importantes y la primera vez que se usaron computadores para su determinación fue en 1963 por Richard Arenstorf. Estas órbitas fueron clave en el programa Apollo para ir a la Luna.

Capítulo 7

Prácticas computacionales, curso 2021-22

7.1. Práctica computacional 1

Objetivos:

- *observar la importancia de la precisión.*
- *aprender a estimar y comparar el orden de convergencia de dos métodos distintos*
- *Ejemplos de sistemas*

Observación: *No usar tildes en los comentarios de los códigos.*

Ejemplos:

1. *En el código EulerExplicito.m se muestra como programar un esquema de Euler explícito. El esquema de programación sirve para cualquier método de un paso. Aquí lo aplicamos al problema*

$$\begin{cases} y'(t) = -ay(t), & t \in [0, 3], \\ y(0) = 1 \end{cases}$$

cuya solución es $y(t) = e^{-at}$.

```
clear all;
t0=0;
T=3;
y0=1;
N=20;
```

```

h=T/N;
a=10;
t=t0:h:t0+T;
y=zeros(N+1,1);
y(1)=y0;
true=t0:0.01:t0+T;
ztrue=exp(-a*true);
for j=1:N
    y(j+1)=y(j)+h*mifun(a,t(j),y(j));
end
figure(1);
plot(t,y,'*-',true,ztrue,'r-');

```

El archivo **mifun.m** es:

```

function dydt=mifun(aa,tt,yy)
dydt=-aa*yy;
end

```

2. En el código **RK4vsEulerExplicitoOrdenes.m** se muestra la programación del método de Runge-Kutta clásico y la del método de Euler explícito. Se aplican ambos al problema

$$\begin{cases} y'(t) = \lambda \cos(\lambda t)y(t), & t \in [0, 20], \\ y(0) = y_0 \end{cases}$$

cuya solución es $y(t) = y_0 e^{\sin(\lambda t)}$ con $y_0 = 2$ y $\lambda = 7$. Esta solución es altamente oscilante y necesita precisión para poder ser calculada. Se obtiene además las curvas de pendientes de cada uno de los métodos comprobándose así que el RK4 clásico posee orden 4 mientras que Euler explícito orden 1.

```

%
% Contrastamos Runge-Kutta de orden 4 con Euler explícito
% estimamos los ordenes
% de convergencia usando las rectas de pendiente
%

t0=0; %Tiempo inicial
tf=20; %Tiempo final
T=tf-t0;% Tiempo total
y0=2; %Dato inicial
%Datos para la solución exacta

```

```

kk=7;
% Solucion exacta en una particion fina
ttrue=t0:0.001:t0+T; %Particion fina
ytrue=y0*exp(sin(kk*ttrue));
%
% Numero de calculos a realizar
%
M=18;
nP=zeros(1,M);% guarda el numero de puntos de la particion en cada calculo
errEuler=zeros(1,M); %guarda el error obtenido con Euler
errRK4=zeros(1,M); %guarda el error obtenido con RK4
%
% Numero de puntos iniciales
%
N=10;
for j=1:M
    nP(j)=N;% Se guarda el numero de puntos a usar
    h=T/N; % Talla de la particion
    t=t0:h:t0+T; %Particion
    h=T/N;% talla de la particion
    % Vector para Runge-Kutta orden 4
    yRK4=zeros(1,N+1);% dimensionaliza t e y
    yRK4(1)=y0;
    % Vector para Euler orden 1
    yEuler=zeros(1,N+1);% dimensionaliza t e y
    yEuler(1)=y0;
    % Solucion exacta en la particion
    yt=y0*exp(sin(kk*t));
    for n=1:N
        %Calculo RK4
        k1=mifun(kk,t(n),yRK4(n));
        k2=mifun(kk,t(n)+h/2,yRK4(n)+h/2*k1);
        k3=mifun(kk,t(n)+h/2,yRK4(n)+h/2*k2);
        k4=mifun(kk,t(n)+h,yRK4(n)+h*k3);
        yRK4(n+1)=yRK4(n)+h*(k1+2*k2+2*k3+k4)/6;
        %Calculo Euler
        yEuler(n+1)=yEuler(n)+h*mifun(kk,t(n),yEuler(n));
    end
    figure(1);
    plot(t,yEuler,'+-',t,yRK4,'d-',ttrue,ytrue,'r-');
    legend('Euler','RK4','exacta','Location','Best');
    title([' Modelo con kk= ',num2str(kk),...

```

```

': RK4 vs Euler con N= ',num2str(N));
errEuler(j)=max(abs(yEuler-yt));
errRK4(j)=max(abs(yRK4-yt));
pause(1);
N=2*N; %duplicamos N
disp(['Errores: N= ',num2str(N),' Euler = ',num2str(errEuler(j)),...
      ' RK4 = ',num2str(errRK4(j))])
pause;
end
%
% Visualizamos ahora los datos globales del calculo
%
figure(2)
plot(nP,log(errEuler),'-*',nP,log(errRK4),'-+');
legend('LogErrEuler','LogErrRK4','Location','Best');
title([' Modelo con kk= ',num2str(kk),...
        ': Decaimiento log errores Euler vs RK4 ']);
figure(3)
plot(log(nP),log(errEuler),'*',log(nP),-log(nP),'-',...
      log(nP),log(errRK4),'+',log(nP),-4*log(nP),'-.');
legend('Euler','-1','RK4','-4','Location','Best');
title([' Modelo con kk= ',num2str(kk),...
        ': Ordenens convergencia Euler vs RK4 ']);

```

El archivo **mifun.m** es:

```

function dydt=mifun(kk,tt,yy)
dydt=kk*cos(kk*tt)*yy;
end

```

3. Programamos ahora la resolución de un sistema lineal 2d usando herramientas matriciales de Matlab o de Octave. El problema es

$$\begin{cases} u' = 9u + 24v + 5\cos(t) - \frac{1}{3}\sin(t), & u(0) = 4/3 \\ v' = -24u - 51v - 9\cos(t) + \frac{1}{3}\sin(t), & v(0) = 2/3 \end{cases}$$

que se puede escribir en forma matricial como

$$X'(t) = AX(t) + F(t)$$

donde $X = (u; v)$, $A = [9 \ 24; -24 \ -51]$ y $F(t) = [f1(t); f2(t)]$ donde $f1(t) = 5\cos(t) - \frac{1}{3}\sin(t)$ y $f2(t) = -9\cos(t) + \frac{1}{3}\sin(t)$. El sistema tiene

como solución

$$\begin{aligned} u(t) &= 2e^{-3t} - e^{-39t} + \frac{1}{3} \cos(t) \\ v(t) &= -e^{-3t} + 2e^{-39t} - \frac{1}{3} \cos(t). \end{aligned}$$

Euler explícito usando matrices y la notación obvia queda como

$$X_{n+1} = (I + hA)X_n + hF_n$$

Euler implícito usando matrices y recordando que no se debe de calcular la inversa queda como

$$X_{n+1} = (I - hA) \setminus [X_n + hF_{n+1}]$$

mientras que Crank-Nicolson es

$$X_{n+1} = (I - 0.5hA) \setminus [(I + 0.5hA)X_n + 0.5h(F_n + F_{n+1})]$$

El archivo **sistema2Dgeneral.m** es:

```
clear all;
T=5;
N=100;
x0=1;
y0=2;
h=T/N;
t=[0:h:T];
x0=4/3;
y0=2/3;
%
% Solucion exacta
%
xv=2*exp(-3*t)-exp(-39*t)+cos(t)/3;
yv=-exp(-3*t)+2*exp(-39*t)-cos(t)/3;
x=zeros(N+1,1);
y=zeros(N+1,1);
x(1)=x0;
y(1)=y0;
h=T/N;
ejes=[0 T -10 10];
%
% Matriz del sistema
%
A=[9 24;-24 -51];
%
```

```
% Matriz identidad
%
I=[1 0;0 1];
%
% Matriz de iteracion para Euler explicito
%
Me=I+h*A;
%
% Inicio computo con Euler explicito
%
v=[x(1);y(1)];
for i=1:N
    v=Me*v+h*[f1(t(i));f2(t(i))];
    x(i+1)=v(1);
    y(i+1)=v(2);
end
figure(1);
subplot(2,3,1),plot(t,x,'-o',t,xv,'-');
title([' Euler explicito, N= ',num2str(N)]);
legend('x','xv','location','Best');
%axis(ejes);
subplot(2,3,4),plot(t,y,'-o',t,yv,'-');
title([' Euler explicito, N= ',num2str(N)]);
legend('y','yv','location','Best');%axis(ejes);
%axis(ejes);
%
% Matriz de iteracion para Euler implicito
%
Mi=I-h*A;
%
% Inicio computo con Euler implicito
%
v=[x(1);y(1)];
for i=1:N
    v=Mi\ (v+h*[f1(t(i+1));f2(t(i+1))]);
    x(i+1)=v(1);
    y(i+1)=v(2);
end
subplot(2,3,2),plot(t,x,'*',t,xv,'-');
title([' Euler implicito, N= ',num2str(N)]);
legend('x','xv','location','Best');
axis(ejes);
```

```

subplot(2,3,5),plot(t,y,'*',t,yv,'-');
title([' Euler implicito, N= ',num2str(N)]);
legend('y','yv','location','Best');
axis(ejes);
%
% Matriz de iteracion para Crank-Nicolson
%
MCN=I-0.5*h*A;
MCNaux=I+0.5*h*A;
%
% Inicio computo con Crank-Nicolson
%
v=[x(1);y(1)];
for i=1:N
    v=MCN\ (MCNaux*v+0.5*h*[f1(t(i+1))+f1(t(i));f2(t(i+1))+f2(t(i))]);
    x(i+1)=v(1);
    y(i+1)=v(2);
end
subplot(2,3,3),plot(t,x,'+',t,xv,'-');
title([' Crank-Nicolson, N= ',num2str(N)]);
legend('x','xv','location','Best');
axis(ejes);
subplot(2,3,6),plot(t,y,'+',t,yv,'-');
title([' Crank-Nicolson, N= ',num2str(N)]);
legend('y','yv','location','Best');
axis(ejes);

```

El archivo f1.m es:

```

function valor=f1(tt)
valor=5*cos(tt)-sin(tt)/3;

```

y el archivo f2.m es:

```

function valor=f2(tt)
valor=-9*cos(tt)+sin(tt)/3;

```

4. La programación con el estilo de función para este ejemplo es

```

function Sistema2DNohomogeneo(T,N,x0,y0)

h=T/N;

```

```

t=[0:h:T];
x0=4/3;
y0=2/3;
xv=2*exp(-3*t)-exp(-39*t)+cos(t)/3;
yv=-exp(-3*t)+2*exp(-39*t)-cos(t)/3;
x=zeros(N+1,1);
y=zeros(N+1,1);
x(1)=x0;
y(1)=y0;
h=T/N;
ejes=[0 T -2 2];
%
% Matriz del sistema
%
A=[9 24;-24 -51];
%
% Matriz identidad
%
I=[1 0;0 1];
%
% Matriz de iteracion para Euler explicito
%
Me=I+h*A;
%
% Inicio computo con Euler explicito
%
v=[x(1);y(1)];
for i=1:N
    v=Me*v+h*[f1(t(i));f2(t(i))];
    x(i+1)=v(1);
    y(i+1)=v(2);
end
subplot(2,3,1),plot(t,x,t,xv);
title([' Euler explicito, N= ',num2str(N)]);
legend('x','xv','location','Best');
subplot(2,3,4),plot(t,y,t,yv);
title([' Euler explicito, N= ',num2str(N)]);
legend('y','yv','location','Best');%axis(ejes);
axis(ejes);
%
% Matriz de iteracion para Euler implicito
%

```

```

Mi=I-h*A;
%
% Inicio computo con Euler implicito
%
v=[x(1);y(1)];
for i=1:N
    v=Mi\ (v+h*[f1(t(i+1));f2(t(i+1))]);
    x(i+1)=v(1);
    y(i+1)=v(2);
end
subplot(2,3,2),plot(t,x,t,xv);
title([' Euler implicito, N= ',num2str(N)]);
legend('x','xv','location','Best');
axis(ejes);
subplot(2,3,5),plot(t,y,t,yv);
title([' Euler implicito, N= ',num2str(N)]);
legend('y','yv','location','Best');
axis(ejes);
%
% Matriz de iteracion para Crank-Nicolson
%
MCN=I-0.5*h*A;
MCNaux=I+0.5*h*A;
%
% Inicio computo con Crank-Nicolson
%
v=[x(1);y(1)];
for i=1:N
    v=MCN\ (MCNaux*v+0.5*h*[f1(t(i+1))+f1(t(i));f2(t(i+1))+f2(t(i))]);
    x(i+1)=v(1);
    y(i+1)=v(2);
end
subplot(2,3,3),plot(t,x,t,xv);
title([' Crank-Nicolson, N= ',num2str(N)]);
legend('x','xv','location','Best');
axis(ejes);
subplot(2,3,6),plot(t,y,t,yv);
title([' Crank-Nicolson, N= ',num2str(N)]);
legend('y','yv','location','Best');
axis(ejes);

function valor=f1(tt)

```

```

valor=5*cos(tt)-sin(tt)/3;
function valor=f2(tt)
valor=-9*cos(tt)+sin(tt)/3;

```

5. Finalmente mostramos como se puede programar RK4 para el sistema de Lorenz

$$\begin{aligned}
 x'(t) &= -10(x(t) - y(t)), \\
 y'(t) &= 28x(t) - y(t) - x(t)z(t), \\
 z'(t) &= 2.667(x(t)y(t) - z(t)),
 \end{aligned}$$

```

%
% El sistema de Lorenz se calcula mediante
% Runge-Kutta de cuarto orden en el intervalo
% [0,T] con h=T/N
% con datos iniciales (x0,y0,z0)
%
% Se puede calcular la curva conforme se obtiene simplemente
% dibujando las componentes de los vectores calculadas en cada
% en cada paso de la interacion
hold off;
clear all;

N=1000;
T=10;
x0=1;
y0=2;
z0=3;

h=T/N;
t=0:h:T;
x=zeros(1,N+1);
y=zeros(1,N+1);
z=zeros(1,N+1);
x(1)=x0;
y(1)=y0;
z(1)=z0;
%
% Aunque tengamos funciones de la forma f(x,y,z) y el argumento t
% no sea necesario, es bueno ponerlo y escribir f(t,x,y,z) para
% tener un código base que pueda servir para otros ejemplos.
for i=1:N
    % Calculo de los cuatro pasos de RK4
    k1 = h * f2(t, x(i), y(i), z(i));
    k2 = h * f2(t + h/2, x(i) + k1/2, y(i) + k1/2, z(i));
    k3 = h * f2(t + h/2, x(i) + k2/2, y(i) + k2/2, z(i));
    k4 = h * f2(t + h, x(i) + k3, y(i) + k3, z(i));
    % Actualización de los valores
    x(i+1) = x(i) + (k1 + 2*k2 + 2*k3 + k4)/6;
    y(i+1) = y(i) + (k1 + 2*k2 + 2*k3 + k4)/6;
    z(i+1) = z(i) + (k1 + 2*k2 + 2*k3 + k4)/6;
end

```

```

k1x=fx(t(i),x(i),y(i),z(i));
k1y=fy(t(i),x(i),y(i),z(i));
k1z=fz(t(i),x(i),y(i),z(i));

k2x=fx(t(i)+h/2,x(i)+h*k1x/2,y(i)+h*k1y/2,z(i)+h*k1z/2);
k2y=fy(t(i)+h/2,x(i)+h*k1x/2,y(i)+h*k1y/2,z(i)+h*k1z/2);
k2z=fz(t(i)+h/2,x(i)+h*k1x/2,y(i)+h*k1y/2,z(i)+h*k1z/2);

k3x=fx(t(i)+h/2,x(i)+h*k2x/2,y(i)+h*k2y/2,z(i)+h*k2z/2);
k3y=fy(t(i)+h/2,x(i)+h*k2x/2,y(i)+h*k2y/2,z(i)+h*k2z/2);
k3z=fz(t(i)+h/2,x(i)+h*k2x/2,y(i)+h*k2y/2,z(i)+h*k2z/2);

k4x=fx(t(i)+h,x(i)+h*k3x,y(i)+h*k3y,z(i)+h*k3z);
k4y=fy(t(i)+h,x(i)+h*k3x,y(i)+h*k3y,z(i)+h*k3z);
k4z=fz(t(i)+h,x(i)+h*k3x,y(i)+h*k3y,z(i)+h*k3z);

x(i+1)=x(i)+h*(k1x+2*k2x+2*k3x+k4x)/6;
y(i+1)=y(i)+h*(k1y+2*k2y+2*k3y+k4y)/6;
z(i+1)=z(i)+h*(k1z+2*k2z+2*k3z+k4z)/6;

end
figure(1);
subplot(3,1,1),plot(t,x,t,y,t,z,'--');
title('Modelo de Lorenz')
legend('x','y','z');
hold off;
subplot(3,1,2),plot(x,y)
title('Efecto mariposa: Plano de Fases x-y')
hold off;
subplot(3,1,3),plot(x,z)
title('Efecto mariposa: Plano de Fases x-z')
hold off;

function valor=fx(tt,xx,yy,zz)
valor=-10*(xx-yy);
end

function valor=fy(tt,xx,yy,zz)
valor=-xx*zz+28*xx-yy;
end

function valor=fz(tt,xx,yy,zz)

```

```

valor=2.667*(xx*yy-zz);
end

```

7.2. Práctica computacional 2

Objetivos:

- observar la restricción de estabilidad absoluta en problemas rígidos y usar métodos implícitos para solventarlo.

Ejemplos:

1. Usando **EulerExplicito.m** en el problema

$$\begin{cases} y'(t) = -ay(t), & t \in [0, 13], \\ y(0) = 1 \end{cases}$$

cuya solución es $y(t) = e^{-at}$ con $a = 500$ se observa que hay que restringir el paso h para que el método de resultados que no oscilen y sean inestables.

2. Usando Euler Explícito, Euler Implícito y Crank-Nicolson en el problema test

$$\begin{cases} y'(t) = -ay(t), & t \in [0, 13], \\ y(0) = 1 \end{cases}$$

cuya solución es $y(t) = e^{-at}$ con $a = 500$ se observa como se puede remediar el problema de la inestabilidad absoluta.

El resultado de aplicar los esquemas a este problema se puede expresar usando lápiz y papel como

$$y_{ex,n+1} = (1-ha)^n, \quad y_{im,n+1} = 1/(1+ha)^n, \quad y_{cn,n+1} = [(1-ha/2)/(1+ha/2)]^n.$$

```

%
% Comparacion de Euler explicito, implicito y Crank-Nicolson
% para el problema test y'=-a*y con y(0)=1

```

```

T=1.;
t0=0;
N=100;
h=T/N;
t=t0:h:t0+T;
y0=1;

```

```

figure(1);
a=500;
yex=zeros(1,N+1);
yim=zeros(1,N+1);
ycn=zeros(1,N+1);
yex(1)=1;
yim(1)=1;
ycn(1)=1;
c1=1-h*a;
c2=1/(1+h*a);
c3=(1-h*a/2)/(1+h*a/2);
for n=1:N
    yex(n+1)=c1*yex(n);
    yim(n+1)=c2*yim(n);
    ycn(n+1)=c3*ycn(n);
end;
errex=max(abs(yex-exp(-a*t)));
errim=max(abs(yim-exp(-a*t)));
errcn=max(abs(ycn-exp(-a*t)));
figure(1);
subplot(3,1,1)
plot(t,yex,'*',t,exp(-a*t));
title([' Euler Explicito con N = ',num2str(N),', a = ',num2str(a)]);
subplot(3,1,2)
plot(t,yim,'o',t,exp(-a*t));
title([' Euler Implicito con N = ',num2str(N),', a = ',num2str(a)]);
subplot(3,1,3)
plot(t,ycn,'+',t,exp(-a*t));
title([' Crank-Nicolson con N = ',num2str(N),', a = ',num2str(a)]);

disp(['Para T = ',num2str(T), ' y a = ',num2str(a)]);
disp(['h debe ser menor que h critico (2/|a|) = ',num2str(2/abs(a))]);
disp(['N debe ser mayor que N critico (|a|T/2) = ',num2str(abs(a*T)/2)]);
disp(['Error con h ',num2str(h)]);
disp(['Error con N ',num2str(N)]);
disp(['Error con Euler Explicito ',num2str(errex)]);
disp(['Error con Euler Implicito ',num2str(errim)]);
disp(['Error con Crank-Nicolson ',num2str(errcn)]);

```

3. Ejercicio: Reproducir el resultado para la ecuación de Dahlquist-Bjorck que

es un ejemplo académico del efecto de la rigidez. El modelo es

$$\begin{cases} y'(t) = 100(\sin(t) - y(t)), & 0 < t \leq 3, \\ y(0) = 0 \end{cases}$$

y se puede ver como un caso particular del problema

$$\begin{cases} y'(t) = a(\sin(t) - y(t)), & t > 0, \\ y(0) = y_0 \end{cases}$$

cuya solución es

$$y(t) = e^{-at}y_0 + \frac{\sin(t) - a^{-1}\cos(t) + a^{-1}e^{-at}}{1 + a^{-2}}.$$

Por lo tanto, para $a \gg 1$ el modo transitorio exponencial asociado con e^{-at} decrece muy rápido. Aquí la solución se descompone en

$$\text{modo rápido} \sim e^{-at}y_0 + \frac{a^{-1}e^{-at}}{1 + a^{-2}}.$$

$$\text{modo estacionario o lento} \sim \frac{\sin(t) - a^{-1}\cos(t)}{1 + a^{-2}}.$$

4. En el código **UsodeMatricesEulerExplicitoImplicito.m** se muestra la programación del método del método de Euler explícito e implícito para un sistema lineal que contiene rigidez. Se aplican ambos al problema $x'(t) = Dx(t)$, ($t > 0$), $x(0) = (1, 0, -1)'$ donde

$$D = \begin{pmatrix} -1 & 20 & 0 \\ 0 & -2 & 20 \\ 0 & 0 & -100 \end{pmatrix}$$

Para cada experimento numérico se presenta el cálculo global en $[0, 3]$. El valor de h óptimo (convergencia sin oscilaciones) para Euler explícito indica un autovalor máximo negativo en torno a -100 .

```
%  
% Resolucion del sistema lineal  
%  
% x'(t)=A*x  
%  
% usando Euler implicito y Explicito  
%
```

```
% Aqui A es una matriz 3x3 y NO se calcula la inversa.  
%  
T=3;  
N=200;  
x0=1;  
y0=0;  
z0=-1;  
h=T/N;  
t=0:h:T;  
x=zeros(N+1,1);  
y=zeros(N+1,1);  
z=zeros(N+1,1);  
x(1)=x0;  
y(1)=y0;  
z(1)=z0;  
%  
% Calculo implicito  
%  
A=zeros(3);  
A=[-1 20 0;  
    0 -2 20;  
    0 0 -100];  
  
A  
b=zeros(3,1);  
sol=b;  
%  
% Calculo Euler implicito  
%  
Q=eye(3)-h*A;  
  
for i=1:N  
    b=[x(i),y(i),z(i)];  
    sol=Q\b';  
    x(i+1)=sol(1);  
    y(i+1)=sol(2);  
    z(i+1)=sol(3);  
end  
subplot(2,1,1),plot(t,x,t,y,t,z,t,0);  
title(['Sistema Stiff 3D ',...  
' Euler implicito, N= ',num2str(N)]);
```

```

legend('x1','x2','x3');
%axis(ejes);

%
% Calculo Euler explicito
%
x(1)=x0;
y(1)=y0;
z(1)=z0;
Q=eye(3)+h*A;
for i=1:N
    b=[x(i),y(i),z(i)];
    sol=Q*b';
    x(i+1)=sol(1);
    y(i+1)=sol(2);
    z(i+1)=sol(3);
end
subplot(2,1,2),plot(t,x,t,y,t,z,t,0);
title(['Sistema Stiff 3D ',...
    ' Euler explicito, N= ',num2str(N)]);
legend('x1','x2','x3');

```

5. **Ejercicio:** Para un sistema lineal general $x'(t) = Ax(t) + b(t)$ programar estos métodos y Crank-Nicolson. Comprobar el orden de convergencia en algún ejemplo con solución conocida.

7.3. Práctica computacional 3

Objetivos:

- Implementar y estudiar métodos de paso adaptativos. En los siguientes ejemplos usamos el método de Dormand-Prince 5(4) para resolver los siguientes problemas:

1. **La ecuación de Dahlquist-Bjorck:** El modelo es

$$\begin{cases} y'(t) = 100(\sin(t) - y(t)), & 0 < t \leq 3, \\ y(0) = 1 \end{cases}$$

y se puede ver como un caso particular del problema

$$\begin{cases} y'(t) = a(\sin(t) - y(t)), & t > 0, \\ y(0) = y_0 \end{cases}$$

cuya solución es

$$y(t) = e^{-at}y_0 + \frac{\sin(t) - a^{-1}\cos(t) + a^{-1}e^{-at}}{1 + a^{-2}}.$$

Por lo tanto, para $a >> 1$ el modo transitorio exponencial asociado con e^{-at} decrece muy rápido. Aquí la solución se descompone en

$$\text{modo rápido} \sim e^{-at}y_0 + \frac{a^{-1}e^{-at}}{1 + a^{-2}}.$$

$$\text{modo estacionario o lento} \sim \frac{\sin(t) - a^{-1}\cos(t)}{1 + a^{-2}}.$$

Es conveniente observar las ideas principales del código:

- a) Se introducen los datos del problema y se inicializan variables
- b) Se definen los coeficientes de los métodos Φ y Ψ a utilizar y se establece una tolerancia para empezar a iterar con ellos.
- c) Al entrar en el bucle iterativo se calcula el error local de truncatura
- d) Si este error es menor que la tolerancia se usa este valor para avanzar y se amplia h a $h_{\text{Next}} = 10h$ (la elección $10h$ es arbitraria) teniendo cuidado de no sobrepasar en mucho el valor de tiempo final.
- e) En caso contrario se obtiene un nuevo $h = h_{\text{Next}}$ y se procura que cumpla $0.1 * h < h_{\text{Next}} < 10h$ al mismo tiempo que no sobrepase en mucho el valor de tiempo final.

```
%  
% Código RKDP54ejemplo1DB.m  
%  
% RK Dorman-Prince 5(4)  
% Problema de Dahlquist-Bjork  
%  
t0=0; %Tiempo inicial  
T=3;  
Tfin=t0+T; %Tiempo final  
N=1000; % Número de puntos máximo  
t=zeros(1,N+1);% dimensionaliza t  
t(1)=t0;  
  
% Vector para Runge-Kutta  
y=zeros(1,N+1);% dimensionaliza y
```

```
y0=1; %Dato inicial
y(1)=y0;

%Datos para la solucion exacta
a=100;
% Para no repetir calculos
aa2=a*a;
aux=1/(1+1/aa2);
inva=1.0/a;
% Solucion exacta en una particion fina
ttrue=t0:0.001:t0+T; %Particion fina
ytrue=y0*exp(-a*ttrue)+sin(ttrue)-cos(ttrue)*inva+exp(-a*ttrue)*inva)*aux;

% Coeficientes del tablero

c2 = 0.2;
c3 = 0.3;
c4 = 0.8;
c5 = 8/9.0;
c6 = 1.0;
c7 = 1.0;

beta1 = 35/384;
beta3 = 500/1113;
beta4 = 125/192;
beta5 = -2187/6784;
beta6 = 11/84;

b1 = 5179/57600;
b3 = 7571/16695;
b4 = 393/640;
b5 = -92097/339200;
b6 = 187/2100;
b7 = 1/40;

a21 = 0.2;
a31 = 0.075;
a32 = 0.225;
a41 = 44/45;
a42 = -56/15;
a43 = 32/9;
a51 = 19372/6561;
```

```

a52 = -25360/2187;
a53 = 64448/6561;
a54 = -212/729;
a61 = 9017/3168;
a62 = -355/33;
a63 = 46732/5247;
a64 = 49/176;
a65 = -5103/18656;
a71 = 35/384;
a73 = 500/1113;
a74 = 125/192;
a75 = -2187/6784;
a76 = 11/84;
tol =1e-6; %Tolerancia sobre el error local de truncatura
n=1;% Para el primer tiempo y primer valor
h=0.5; % Paso inicial
Tau=1; % Truncatura inicial para entrar en el bucle

while ((n<=N) && (t(n)<Tfin)&&(Tau>1e-10))
k1 = mifunDB(t(n),y(n),a);
k2 = mifunDB(t(n)+c2*h,y(n)+h*a21*k1,a);
k3 = mifunDB(t(n)+c3*h,y(n)+h*(a31*k1+ a32*k2),a);
k4 = mifunDB(t(n)+c4*h,y(n)+h*(a41*k1+ a42*k2 + a43*k3),a);
k5 = mifunDB(t(n)+c5*h,y(n)+h*(a51*k1+ a52*k2 + a53*k3 + a54*k4),a);
k6 = mifunDB(t(n)+c6*h,y(n)+h*(a61*k1+ a62*k2 + a63*k3 + a64*k4 + a65*k5));
k7 = mifunDB(t(n)+c7*h,y(n)+h*(a71*k1+ a73*k3 + a74*k4 + a75*k5 + a76*k6));
%
% Incremento para la solucion de orden 5
% dada por yRK4(n) + dy
%
dy = h*(beta1*k1 + beta3*k3 + beta4*k4 + beta5*k5 + beta6*k6);
%%
% Vemos los valores calculados
%
tt=t(n)+h;
ytt=y0*exp(-a*tt)+(sin(tt)-cos(tt)*inva+exp(-a*tt)*inva)*aux;
disp(['En tt = ',num2str(tt)]);
disp(['ytrue = ',num2str(ytt),' yaprox (orden 5)= ',num2str(y(n) + dy)]);
%
% Estimamos el error local de truncatura
% haciendo la diferencia
% entre la solucion de orden 5 y la de orden 4

```

```

%
% Error local de truncatura es
%
Tau = abs((beta1-b1)*k1+(beta3-b3)*k3+(beta4-b4)*k4+(beta5-b5)*k5, ...
           +(beta6-b6)*k6-b7*k7);
disp(['Tau = ',num2str(Tau)]);

% Aceptamos el paso dado y avanzamos
if (Tau <= tol)

    disp([' AVANZAMOS CON h= ',num2str(h) ', a tt = ',num2str(t(n)+h)]);
    y(n+1)= y(n) + dy;
    t(n+1)= t(n) + h;

    disp([' Tau<tol: t(n+1)= ',num2str(t(n+1))]);
    if (t(n+1)>=Tfin)% Se sobrepasa Tfin
        disp(' t(n+1) sobrepasa Tfin==> ultima iteracion ');
        Nfin=n;
    else % Si no se sobrepasa
        h=min(10*h,Tfin-t(n+1));
        disp([' Tau<tol: min(10*h,Tfin-t(n+1))= ',num2str(h)]);
    end
    n=n+1;
end % Fin de trabajo cuando se acepta el paso
% No aceptamos el paso dado--> Recalculamos
if (Tau >= tol)
%%
%% hNext =h*s
%% para ser usado si el error local de truncatura no es
%% lo suficientemente pequeno
%%
hNext = 0.9*(tol/Tau)^(1/5.0)*h;
disp([' tol = ',num2str(tol),' Tau = ',num2str(Tau)]);
disp(['h = ',num2str(h),' hNext = ',num2str(hNext)]);
% Controlamos hNext
if (hNext< 0.1*h) % Que no sea menor que 0.1*h
    hNext = 0.1*h;
end
if (hNext > 10*h) % Que no sea mayor que 10*h
    hNext=10*h;
end
if (t(n) + hNext > Tfin)% Que no se sobreponga t0+T

```

```

hNext = Tfin - t(n);
Nfin=n;
end
if(t(n) + hNext < Tfin)
    disp([' RE-CALCULAMOS CON hNext Ajustado = ',num2str(hNext)]);
end
h = hNext; % repetimos el trabajo con h=hNext
end
%pause;
end
figure(1);
%subplot(2,1,1)
plot(t(1:Nfin+1),y(1:Nfin+1),'o-',ttrue,ytrue,'r-');
ytadatp=y0*exp(-a*t(1:Nfin+1))+...
    (sin(t(1:Nfin+1))-cos(t(1:Nfin+1))*inva...
    +exp(-a*t(1:Nfin+1))*inva)*aux;
err=max(abs(y(1:Nfin+1)-ytadatp(1:Nfin+1)));
disp([' Error global = ',num2str(err)]);

```

El archivo **mifunDB.m** es:

```

function dydt=mifunDB(tt,yy,a)
dydt=a*(sin(tt)-yy);
end

```

2. El problema

$$\begin{cases} y'(t) = 51150e^{-50t}y(t)^2, & t \in [0, 3], \\ y(0) = 1/1024 \end{cases}$$

cuya solución $y(t) = (1 + 1023e^{-50t})^{-1}$ (aquí $51150 = 1023 * 50$) presenta un incremento brusco de $y(0)$ a 1.

```

%
% Código RKDP54ejemplo2.m
%
%
% RK Dorman-Prince 5(4)
% Problema con pendiente brusca en la solucion
%
% Solucion: y(t)=(1+Ce^{-Dt})^{-1}
% Derivada: y'(t)=D*C*e^{-Dt}y(t)^2
%
```

```

clear all;
C=1023;
D=50;
t0=0; %Tiempo inicial
T=3;
Tfin=t0+T; %Tiempo final
N=1000; % Numero de puntos maximo
t=zeros(1,N+1);% dimensionaliza t
t(1)=t0;

% Vector para Runge-Kutta orden 4
yRK4=zeros(1,N+1);% dimensionaliza y
y0=1.0/(1+C); %Dato inicial
yRK4(1)=y0;

% Solucion exacta en una particion fina
ttrue=t0:0.001:Tfin; %Particion fina
ytrue=1.0./(1+C*exp(-D*ttrue));

% Coeficientes del tablero

c2 = 0.2;
c3 = 0.3;
c4 = 0.8;
c5 = 8/9.0;
c6 = 1.0;
c7 = 1.0;

beta1 = 35/384;
beta3 = 500/1113;
beta4 = 125/192;
beta5 = -2187/6784;
beta6 = 11/84;

b1 = 5179/57600;
b3 = 7571/16695;
b4 = 393/640;
b5 = -92097/339200;
b6 = 187/2100;
b7 = 1/40;

a21 = 0.2;

```

```

a31 = 0.075;
a32 = 0.225;
a41 = 44/45;
a42 = -56/15;
a43 = 32/9;
a51 = 19372/6561;
a52 = -25360/2187;
a53 = 64448/6561;
a54 = -212/729;
a61 = 9017/3168;
a62 = -355/33;
a63 = 46732/5247;
a64 = 49/176;
a65 = -5103/18656;
a71 = 35/384;
a73 = 500/1113;
a74 = 125/192;
a75 = -2187/6784;
a76 = 11/84;
tol =1e-6; %Tolerancia sobre el error local de truncatura
n=1;% Para el primer tiempo y primer valor
h=0.5; % Paso inicial
Tau=1; % Truncatura inicial para entrar en el bucle

while ((n<=N) && (t(n)<Tfin)&&(Tau>1e-10))
k1 = ejemplo2(t(n),yRK4(n),C,D);
k2 = ejemplo2(t(n)+c2*h,yRK4(n)+h*a21*k1,C,D);
k3 = ejemplo2(t(n)+c3*h,yRK4(n)+h*(a31*k1+ a32*k2),C,D);
k4 = ejemplo2(t(n)+c4*h,yRK4(n)+h*(a41*k1+ a42*k2 + a43*k3),C,D);
k5 = ejemplo2(t(n)+c5*h,yRK4(n)+h*(a51*k1+ a52*k2 + a53*k3 + a54*k4),C,D);
k6 = ejemplo2(t(n)+c6*h,yRK4(n)+h*(a61*k1+ a62*k2 + a63*k3 + a64*k4 + a65*k5),C,D);

k7 = ejemplo2(t(n)+c7*h,yRK4(n)+h*(a71*k1+ a73*k3 + a74*k4 + a75*k5 + a76*k6),C,D);
%
% Incremento para la solucion de orden 5
% dada por yRK4(n) + dy
%
dy = h*(beta1*k1 + beta3*k3 + beta4*k4 + beta5*k5 + beta6*k6);
%
% Vemos los valores calculados
%

```

```

tt=t(n)+h;
ytt=1.0./(1+C*exp(-D*tt));
disp(['En tt = ',num2str(tt)]);
disp(['ytrue = ',num2str(ytt),' yaprox (orden 5)= ',num2str(yRK4(n) + dy)]);
%
% Estimamos el error local de truncatura
% haciendo la diferencia
% entre la solucion de orden 5 y la de orden 4
%
% Error local de truncatura es
%
Tau = abs((beta1-b1)*k1+(beta3-b3)*k3+(beta4-b4)*k4+(beta5-b5)*k5+ . .
(beta6-b6)*k6-b7*k7);
disp(['Tau = ',num2str(Tau)]);

% Aceptamos el paso dado y avanzamos
if (Tau <= tol)
disp([' AVANZAMOS CON h= ',num2str(h) ', a tt = ',num2str(t(n)+h)]);
yRK4(n+1)= yRK4(n) + dy;
t(n+1)= t(n) + h;

disp([' Tau<tol: t(n+1)= ',num2str(t(n+1))]);
if (t(n+1)>=Tfin)% Se sobrepasa Tfin
    disp(' t(n+1) sobrepasa Tfin==> ultima iteracion ');
    Nfin=n;
else % Si no se sobrepasa
    h=min(10*h,Tfin-t(n+1));
    disp([' Tau<tol: min(10*h,Tfin-t(n+1))= ',num2str(h)]);
end
n=n+1;
%pause;
end % Fin de trabajo cuando se acepta el paso
% No aceptamos el paso dado--> Recalculamos
if (Tau >= tol)
%%
%% hNext =h*s
%% para ser usado si el error local de truncatura no es
%% lo suficientemente pequeno
%%
hNext = 0.9*(tol/Tau)^(1/5.0)*h;
disp([' tol = ',num2str(tol),' Tau = ',num2str(Tau)]);
disp(['h = ',num2str(h),' hNext = ',num2str(hNext)]);

```

```
% Controlamos hNext
if (hNext< 0.1*h) % Que no sea menor que 0.1*h
    hNext = 0.1*h;
end
if (hNext > 10*h) % Que no sea mayor que 10*h
    hNext=10*h;
end
if (t(n) + hNext > Tfin)% Que no se sobrepase t0+T
    hNext = Tfin - t(n);
    Nfin=n;
end
if(t(n) + hNext < Tfin)
    disp([' RE-CALCULAMOS CON hNext Ajustado = ',num2str(hNext)]);
    end
h = hNext; % repetimos el trabajo con h=hNext
end
%pause;
end
figure(1);
% subplot(2,1,1)
plot(t(1:Nfin+1),yRK4(1:Nfin+1),'o-',ttrue,ytrue,'r-');
ytadatp=1.0./(1+C*exp(-D*t(1:Nfin+1)));
err=max(abs(yRK4(1:Nfin+1)-ytadatp(1:Nfin+1)));
disp([' Error global = ',num2str(err),' tol = ',num2str(tol)]);
```

El archivo ejemplo2.m es:

```
function dydt=ejemplo2(tt,yy,C,D)
dydt=C*D*exp(-D*tt)*yy*yy;
end
```

3. Extendemos ahora el método a un sistema 2D

$$x'(t) = D x(t) + g(t), \quad 0 < t < 20, \quad x(0) = (4/3, 2/3)'$$

donde

$$D = \begin{pmatrix} 9 & 24 \\ -24 & -51 \end{pmatrix}$$

siendo $g_1(t) = 5\cos(t) - \sin(t)/3$, $g_2(t) = -9\cos(t) + \sin(t)/3$ con solución exacta

$$\begin{aligned} x(t) &= 2e^{-3t} - e^{-39t} + \cos(t)/3, \\ y(t) &= -e^{-3t} + 2e^{-39t} - \cos(t)/3. \end{aligned}$$

Aunque parezca engorroso, en el fondo sólo duplicamos ecuaciones con respecto al caso escalar.

```
%  
% Código RKDP54Ejemplo2D.m  
%  
% RK Dorman-Prince 5(4) en sistema  
%  
clear all;  
t0=0; %Tiempo inicial  
T=20;  
Tfin=t0+T; %Tiempo final  
N=2000; % Número de puntos máximo  
t=zeros(1,N+1);% dimensionaliza t  
t(1)=t0;  
  
% Vector para Runge-Kutta orden 4  
x=zeros(1,N+1);% dimensionaliza y  
y=zeros(1,N+1);% dimensionaliza y  
x0=4/3;  
y0=2/3; %Dato inicial  
x(1)=x0;  
y(1)=y0;  
  
% Solucion exacta en una particion fina  
ttrue=t0:0.001:t0+T; %Particion fina  
xtrue=2*exp(-3*ttrue)-exp(-39*ttrue)+cos(ttrue)/3;  
ytrue=-exp(-3*ttrue)+2*exp(-39*ttrue)-cos(ttrue)/3;  
  
% Coeficientes del tablero  
  
c2 = 0.2;  
c3 = 0.3;  
c4 = 0.8;  
c5 = 8/9.0;  
c6 = 1.0;  
c7 = 1.0;  
  
beta1 = 35/384;  
beta3 = 500/1113;  
beta4 = 125/192;  
beta5 = -2187/6784;
```

```
beta6 = 11/84;

b1 = 5179/57600;
b3 = 7571/16695;
b4 = 393/640;
b5 = -92097/339200;
b6 = 187/2100;
b7 = 1/40;

a21 = 0.2;
a31 = 0.075;
a32 = 0.225;
a41 = 44/45;
a42 = -56/15;
a43 = 32/9;
a51 = 19372/6561;
a52 = -25360/2187;
a53 = 64448/6561;
a54 = -212/729;
a61 = 9017/3168;
a62 = -355/33;
a63 = 46732/5247;
a64 = 49/176;
a65 = -5103/18656;
a71 = 35/384;
a73 = 500/1113;
a74 = 125/192;
a75 = -2187/6784;
a76 = 11/84;
tol = 1e-6; %Tolerancia sobre el error local de truncatura
n=1;
h=0.5;
Tau=1; % Truncatura inicial para entrar en el bucle

while ((n<=N) && (t(n)<Tfin)&&(Tau>1e-10))
k1x = fx(t(n),x(n),y(n));
k1y = fy(t(n),x(n),y(n));

k2x = fx(t(n)+c2*h,x(n)+h*a21*k1x,y(n)+h*a21*k1y);
k2y = fy(t(n)+c2*h,x(n)+h*a21*k1x,y(n)+h*a21*k1y);

k3x = fx(t(n)+c3*h,x(n)+h*(a31*k1x+ a32*k2x),y(n)+h*(a31*k1y+ a32*k2y));
```

```

k3y = fy(t(n)+c3*h,x(n)+h*(a31*k1x+ a32*k2x),y(n)+h*(a31*k1y+ a32*k2y));

k4x = fx(t(n)+c4*h,x(n)+h*(a41*k1x+ a42*k2x + a43*k3x),...
          y(n)+h*(a41*k1y+ a42*k2y + a43*k3y));
k4y = fy(t(n)+c4*h,x(n)+h*(a41*k1x+ a42*k2x + a43*k3x),...
          y(n)+h*(a41*k1y+ a42*k2y + a43*k3y));

k5x = fx(t(n)+c5*h,x(n)+h*(a51*k1x+ a52*k2x + a53*k3x + a54*k4x),...
          y(n)+h*(a51*k1y+ a52*k2y + a53*k3y + a54*k4y));
k5y = fy(t(n)+c5*h,x(n)+h*(a51*k1x+ a52*k2x + a53*k3x + a54*k4x),...
          y(n)+h*(a51*k1y+ a52*k2y + a53*k3y + a54*k4y));

k6x = fx(t(n)+c6*h,x(n)+h*(a61*k1x+ a62*k2x + a63*k3x + a64*k4x + a65*k5x),...
          y(n)+h*(a61*k1y+ a62*k2y + a63*k3y + a64*k4y + a65*k5y));
k6y = fy(t(n)+c6*h,x(n)+h*(a61*k1x+ a62*k2x + a63*k3x + a64*k4x + a65*k5x),...
          y(n)+h*(a61*k1y+ a62*k2y + a63*k3y + a64*k4y + a65*k5y));

k7x = fx(t(n)+c7*h,x(n)+h*(a71*k1x+ a73*k3x + a74*k4x + a75*k5x + a76*k6x),...
          y(n)+h*(a71*k1y+ a73*k3y + a74*k4y + a75*k5y + a76*k6y));
k7y = fy(t(n)+c7*h,x(n)+h*(a71*k1x+ a73*k3x + a74*k4x + a75*k5x + a76*k6x),...
          y(n)+h*(a71*k1y+ a73*k3y + a74*k4y + a75*k5y + a76*k6y));
%
% Incremento para la solucion de orden 5
% dada por yRK4(n) + dy
%
dx = h*(beta1*k1x + beta3*k3x + beta4*k4x + beta5*k5x + beta6*k6x);
dy = h*(beta1*k1y + beta3*k3y + beta4*k4y + beta5*k5y + beta6*k6y);
%
% Estimamos el error local de truncatura
% haciendo la diferencia
% entre la solucion de orden 5 y la de orden 4
%
% Error local de truncatura es
%
Taux = abs((beta1-b1)*k1x+(beta3-b3)*k3x+(beta4-b4)*k4x+(beta5-b5)*k5x, ...
            +(beta6-b6)*k6x-b7*k7x);
disp(['Taux = ',num2str(Taux)]);
Tauy = abs((beta1-b1)*k1y+(beta3-b3)*k3y+(beta4-b4)*k4y+(beta5-b5)*k5y, ...
            +(beta6-b6)*k6y-b7*k7y);
disp(['Tauy = ',num2str(Tauy)]);
Tau=Taux+Tauy;

```

```
% Aceptamos el paso dado y avanzamos
if (Tau <= tol)
    disp([' AVANZAMOS CON h= ',num2str(h) ', a tt = ',num2str(t(n)+h)]);
    x(n+1)= x(n) + dx;
    y(n+1)= y(n) + dy;
    t(n+1)= t(n) + h;
    disp([' Tau<tol: t(n+1)= ',num2str(t(n+1))]);
    if (t(n+1)>=Tfin)% Se sobrepasa Tfin
        disp(' t(n+1) sobrepasa Tfin==> ultima iteracion ');
        Nfin=n;
    else % Si no se sobrepasa
        h=min(10*h,Tfin-t(n+1));
        disp([' Tau<tol: min(10*h,Tfin-t(n+1))= ',num2str(h)]);
    end
    n=n+1;
    %pause;
end % Fin de trabajo cuando se acepta el paso
% No aceptamos el paso dado--> Recalculamos
if (Tau >= tol)
    %%
    %% hNext =h*s
    %% para ser usado si el error local de truncatura no es
    %% lo suficientemente pequeno
    %%
    hNext = 0.9*(tol/Tau)^(1/5.0)*h;
    disp([' tol = ',num2str(tol),' Tau = ',num2str(Tau)]);
    disp(['h = ',num2str(h),' hNext = ',num2str(hNext)]);
    % Controlamos hNext
    if (hNext< 0.1*h) % Que no sea menor que 0.1*h
        hNext = 0.1*h;
    end
    if (hNext > 10*h) % Que no sea mayor que 10*h
        hNext=10*h;
    end
    if (t(n) + hNext > Tfin)% Que no se sobrepase t0+T
        hNext = Tfin - t(n);
        Nfin=n;
    end
    if(t(n) + hNext < Tfin)
        disp([' RE-CALCULAMOS CON hNext Ajustado = ',num2str(hNext)]);
    end
    h = hNext; % repetimos el trabajo con h=hNext
```

```

    end
%pause;
end
figure(1);
subplot(3,1,1)
plot(t(1:Nfin+1),x(1:Nfin+1),'o-',ttrue,xtrue,'+-');
subplot(3,1,2)
plot(t(1:Nfin+1),y(1:Nfin+1),'x-',ttrue,ytrue,'+-');
subplot(3,1,3)
plot(x(1:Nfin+1),y(1:Nfin+1),'d-',xtrue,ytrue,'-');

```

El archivo **fx.m** es:

```

function dydt=fx(tt,xx,yy)
dydt=9*xx+24*yy+5*cos(tt)-sin(tt)/3;
end

```

y el archivo es:

```

function dydt=fy(tt,xx,yy)
dydt=-24*xx-51*yy-9*cos(tt)+sin(tt)/3;
end

```

4. **Ejercicio:** Usar Runge-Kutta Fehlberg 4(5) en cualquiera de estos problemas.
5. **Ejercicio:** Los métodos de paso adaptativos son especialmente útiles en los problemas donde la solución tiene cambios bruscos. Estos problemas se pueden reconocer por la presencia de coeficientes discontinuos o que cambian rápidamente sobre algunas regiones. Este tipo de ecuaciones aparecen naturalmente en problemas que involucran circuitos eléctricos o sistemas mecánicos que contienen fuerzas discontinuas como puede ser en terremotos o mecánica de estructuras.

Consideremos el modelo

$$\begin{cases} y'(t) = -b(t)y(t) + t, & 0 < t \leq 3, \\ y(0) = 1 \end{cases}$$

donde

$$b(t) = \begin{cases} 1, & 0 \leq t \leq 2, \\ 3, & 2 < t. \end{cases}$$

La solución exacta es

$$y(t) = \begin{cases} t - 1 + 2e^{-t}, & 0 \leq t \leq 2, \\ 1/3t - 1/9 + e^{-3t}(4/9e^6 + 2e^4), & 2 < t. \end{cases}$$

- a) Usar un método adaptativo RK-4(5) para resolver este problema con una tolerancia de 0.0001. Contar el número de pasos usados.
- b) Resolver el mismo problema con el método de RK clásico con un paso escogido tal que el número de iteraciones, o de puntos en $[0, 3]$, sea similar al del apartado anterior.
- c) Comparar ambas aproximaciones con la solución exacta. Observar para ello las Figuras 7.1, 7.2 y 7.3 obtenidas con una tolerancia de $1e - 2$ y de $1e - 5$ en un entorno de radio 0.001 de la discontinuidad en $t = 2$.
6. Estos problemas poseen coeficientes con discontinuidad o con cambios bruscos. Aplicar el ejercicio anterior a alguno de estos ejemplos:

a)

$$\begin{cases} y'(t) &= f(t, y) \\ y(0) &= 2 \end{cases} \quad 0 < t \leq 4, \quad f(t, y) = \begin{cases} ty & t < 2, \\ 1-t & t \geq 2, \end{cases}$$

b)

$$\begin{cases} y'(t) &= f(t, y) \\ y(0) &= 0 \end{cases} \quad 0 < t \leq 4, \quad f(t, y) = \begin{cases} t^2 + y^2 & t < 2, \\ 3y & t \geq 2, \end{cases}$$

c)

$$\begin{cases} y'(t) &= \sin(\frac{1}{2.01-t}) \\ y(0) &= 2 \end{cases} \quad 0 < t \leq 2$$

d)

$$\begin{cases} y'(t) &= f(t, y) \\ y(0) &= -1 \end{cases} \quad 0 < t \leq 4, \quad f(t, y) = \begin{cases} 3y - 2\sin(y) & t < 2, \\ 1 + 2y + \cos(y) & t \geq 2, \end{cases}$$

e)

$$\begin{cases} y'(t) &= f(t, y) \\ y(0) &= 0 \end{cases} \quad 0 < t \leq 4, \quad f(t, y) = \begin{cases} e^{-ty} & t < 1, \\ -4y^{3/2} & t \geq 1, \end{cases}$$

f)

$$\begin{cases} y'(t) &= f(t, y) \\ y(0) &= 2 \end{cases} \quad 0 < t \leq 4, \quad f(t, y) = \begin{cases} 50\sin(50t) & 1.7 < t < 2.5, \\ y - 2t & \text{en otro caso} \end{cases}$$

7. Realizar la misma comparación con la solución del problema

$$\begin{cases} y'(t) &= y(t)^2, \\ y(0) &= 2 \end{cases} \quad 0 < t \leq 1,$$

Explicar porqué la solución no está definida en todo $t \in [0, 1]$.

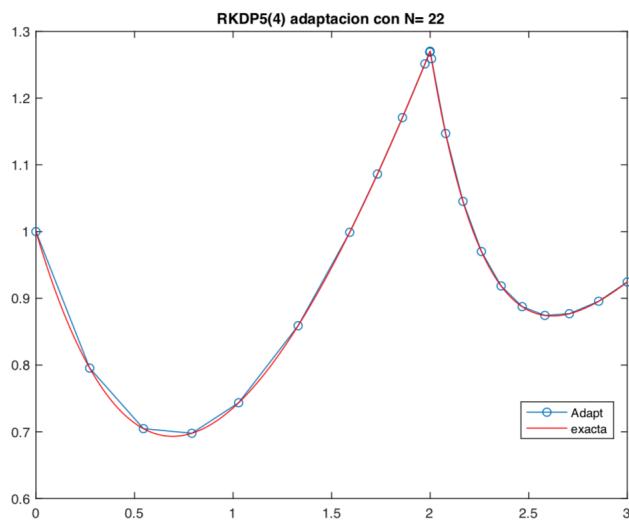


Figura 7.1: Solución con RKDP5(4).

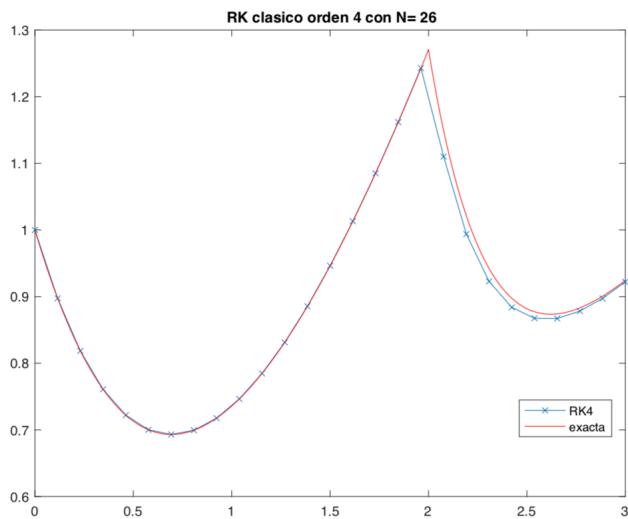


Figura 7.2: Solución con RK4 clásico.

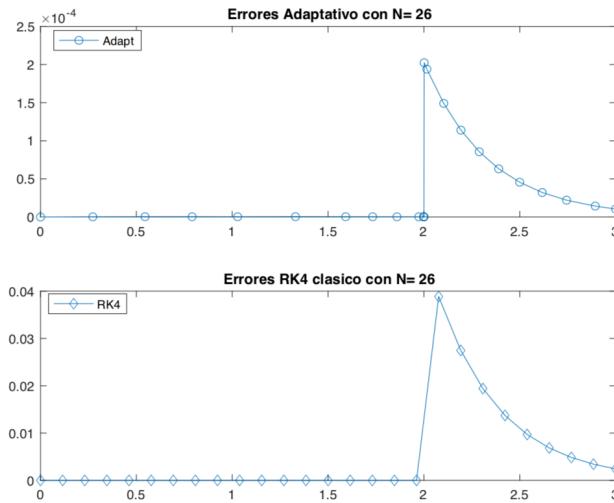


Figura 7.3: Comparación de errores.

8. Comparación en un modelo muy rígido: Consideremos el problema

$$\begin{cases} y'(t) = 101 + 100(t - y), & 0 < t \leq 1, \\ y(0) = 1 \end{cases}$$

La solución exacta $y(t) = 1+t$ viene de la solución general $y(t) = 1+t+ce^{-100t}$ que tiene un término que decrece muy rápidamente.

9. Comparación en un modelo muy inestable: Consideremos el problema

$$\begin{cases} y'(t) = 100y - 101e^{-t}, & 0 < t \leq 3, \\ y(0) = 1 \end{cases}$$

La solución exacta $y(t) = e^{-t}$ viene de la solución general $y(t) = e^{-t} + ce^{100t}$ que tiene un término que crece muy rápidamente.

10. Comparación en un problema no lineal: Al encender una cerilla la bola de fuego crece hasta alcanzar un estado estacionario donde se equilibra al absorción de oxígeno del exterior con el consumo interior. Un modelo simple es el propuesto por Larry Shampine (autor de las librerías de edos para MATLAB, OCTAVE y SCILAB entre otros)

$$\begin{cases} y'(t) = y(t)^2 - y(t)^3, & 0 < t < 2/\delta, \\ y(0) = \delta \end{cases}$$

en donde $y(t)$ es el radio de la bola de fuego, $\delta > 0$ es un radio inicial pequeño y los términos $y(t)^2$ e $y(t)^3$ están relacionados con la superficie y el volumen de la bola de fuego.

El parámetro crítico es el radio inicial $\delta > 0$ que es pequeño y el fenómeno físico de interés ocurre en el tiempo $t_ \approx 1/\delta$. Para $0 < t < 1/\delta$ se observa un crecimiento moderado del radio y un crecimiento repentino en torno a $t_* \approx 1/\delta$ para llegar al valor $y(t) \approx 1$ en donde se estabiliza. Realizar la misma comparación que en los ejercicios anteriores para $\delta = 10^{-3}$.*

7.4. Práctica computacional 4

Objetivos:

- Implementar y estudiar métodos multipaso observando propiedades de 0-estabilidad o de orden de convergencia.

1. Se considera el método lineal de dos pasos explícito

$$y_{n+2} - (1+a)y_{n+1} + a y_n = h\{\beta_1 f_{n+1} + \beta_0 f_n\}.$$

Para cualquier valor de $a \in \mathbb{R}$ con $|a| < 1$ y si $\beta_0 + \beta_1 = 1 - a$ entonces el método converge con orden uno. Para conseguir orden 2 se necesita la restricción adicional

$$\beta_1 = \frac{3-a}{2}.$$

Usando como ejemplo en el caso de orden uno: $a = 0.5$, $\beta_0 = -0.45$, $\beta_1 = 0.95$ y en el caso de orden dos: $a = 0.5$, $\beta_0 = -0.75$, $\beta_1 = 1.25$, generar una gráfica de pendientes para comprobar el orden del error usando un problema con solución conocida. Resumiendo, para $a, b \in \mathbb{R}$ con $|a| < 1$ la elección

$$\beta_0 = 1 - a - \beta_1, \quad \beta_1 = b + \frac{3-a}{2}$$

genera una familia de métodos de dos pasos de orden 1 para todo $b \neq 0$ y si $b = 0$ se genera un método de orden 2.

2. La función $y(t) \equiv 1$ cumple el problema trivial

$$y(0) = 1, \quad y'(t) = 0, \quad t \in [0, 10].$$

En este caso la parte derecha de los métodos lineales multipaso no influye en el cálculo de la solución dada por el esquema discreto puesto que $f = 0$.

Para el problema anterior las raíces del polinomio

$$p(z) = z^2 - (1+a)z + a$$

asociado a la parte izquierda son a y 1 . Comprueba computacionalmente que si $|a| < 1$ entonces los posibles errores en los valores iniciales $y_0 = 1 \pm \epsilon$

e $y_1 = 1 \pm \delta$ se amortiguan mientras que si $|a| > 1$ entonces el cálculo lleva a resultados inestables e inservibles. Esta situación empeora conforme se aumenta el número de puntos que usamos en el intervalo, esto es, si buscamos eliminar el problema aumentando la precisión lo que hacemos es empeorar. Comprobarlo con los siguientes datos

- a) $T = 10$, $N = 100$, $a = 1.1$, $y_0 = 1 + 10^{-13}$, $y_1 = 1 - 10^{-10}$
- b) $T = 10$, $N = 100$, $a = 0.9$, $y_0 = 1 + 10^{-13}$, $y_1 = 1 - 10^{-10}$

3. (Griffiths et al.) Comprobar que el orden de consistencia del método

$$y_{n+2} + (\alpha - 1)y_{n+1} - \alpha y_n = \frac{h}{4}[(\alpha + 3)f_{n+2} + (3\alpha + 1)f_n]$$

es 2 si $\alpha \neq -1$ y 3 si $\alpha = -1$. Aplicar el método cuando $\alpha = -1$ al problema $y'(t) = 0$, $t > 0$ con $y(0) = 0$ tomando $y_0 = 0$, $y_1 = h$ como valores iniciales y explicar el comportamiento resultante de la solución numérica.

4. (Lambert) Comprobar computacionalmente el efecto de la 0-estabilidad usando el método

$$y_{n+2} - (1 + \alpha)y_{n+1} + \alpha y_n = \frac{h}{2}[(3 - \alpha)f(t_{n+1}, y_{n+1}) - (1 + \alpha)f(t_n, y_n)]$$

para calcular la solución de la ecuación $y' = 4ty^{1/2}$ con $y(0) = 1$ y $t \in [0, 2]$ que es $y(t) = (t^2 + 1)^2$. Usar los valores $\alpha = 0$ y $\alpha = -5$ con pasos $h = 0.1, 0.05, 0.025$.

5. Comprobar computacionalmente el efecto de inicializar un método multipaso de orden 3 con valores obtenidos mediante algún método de aproximación de orden menor que tres. Por ejemplo, usar un Adams-Bashforth de orden 3 e inicializarlo con algún Runge-Kutta de orden menor, como Heun o Euler explícito.

Capítulo 8

Introducción a las Ecuaciones en Derivadas Parciales

8.1. Notación

Sea $\Omega \subset \mathbb{R}^d$ un conjunto abierto con frontera $\partial\Omega$ y denotemos por \vec{n} el vector normal exterior a Ω . Para una función $u : \Omega \rightarrow \mathbb{R}$ la derivada parcial en notación multi-índice es

$$D^\alpha u(x) = \frac{\partial^{|\alpha|} u}{\partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \dots \partial_{x_d}^{\alpha_d}}$$

donde $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d) \in \mathbb{N}^d$, es decir, $\alpha_i \geq 0$ entero y denotamos por $|\alpha| = \alpha_1 + \alpha_2 + \alpha_3 + \dots + \alpha_d$. También podemos poner $\partial_{x_i} u = u_{x_i}$. Cuando además hay una dependencia del tiempo

$$u = u(t, x) : [0, +\infty) \times \Omega \rightarrow \mathbb{R}$$

se pueden mezclar derivadas en variables espaciales y el tiempo pero suelen predominar las derivadas primer o segundo orden por tener mayor significación física, como por ejemplo velocidad, aceleración, desplazamientos espaciales, etc...

Operadores diferenciales importantes son

- el vector gradiente

$$\nabla u = Du = (u_{x_1}, u_{x_2}, \dots, u_{x_d}).$$

En el caso escalar es simplemente la derivada $\nabla u = Du = u_x = u'$.

- la derivada direccional

$$\frac{\partial u}{\partial \vec{v}} = \vec{v} \cdot \nabla u = v_1 \partial_{x_1} u + v_2 \partial_{x_2} u + \dots + v_d \partial_{x_d} u,$$

que representa la razón de cambio en la dirección dada por \vec{v} . En el caso escalar $\vec{v} = v$ y tenemos $\frac{\partial u}{\partial \vec{v}} = v u_x$, luego es una amplificación de módulo $|v|$ y signo $\text{sgn}(v)$ de la derivada u_x .

- Cuando tenemos una región acotada Ω donde se puede definir un vector normal en la frontera \vec{n} entonces tenemos la **derivada normal**

$$\frac{\partial u}{\partial \vec{n}} = \vec{n} \cdot \nabla u = n_1 \partial_{x_1} u + n_2 \partial_{x_2} u + \dots + n_d \partial_{x_d} u,$$

que es un caso particular de la derivada direccional. La derivada normal es especialmente útil cuando se usa para representar la razón de cambio en la frontera de una región con vector normal \vec{n} en su frontera.

En el caso escalar, el dominio usualmente es un intervalo y entonces $\vec{n} = \pm 1$. Por lo tanto, tenemos $\frac{\partial u}{\partial \vec{n}} = \pm u'$.

- La **divergencia de un campo vectorial** $\vec{u} = (u_1, u_2, \dots, u_d)$

$$\text{div}(\vec{u}) = \partial_{x_1} u_1 + \partial_{x_2} u_2 + \dots + \partial_{x_d} u_d$$

En el caso escalar \vec{u} es un escalar u y tenemos $\text{div}(\vec{u}) = u'$; luego es simplemente la derivada y representa el movimiento (crecimiento, decrecimiento o estacionario).

- El operador **Laplaciano** para una función escalar

$$\Delta u = u_{x_1,x_1} + u_{x_2,x_2} + \dots + u_{x_d,x_d}$$

que en el caso escalar es simplemente $u_{xx}(x) = u''(x)$ y sabemos que representa la curvatura (concavidad o convexidad) de la función $u(x)$.

Recordemos dos resultados fundamentales

Teorema 155 Teorema de la divergencia: si \vec{u} vectorial y q escalar

$$\int_{\partial\Omega} \vec{u} \cdot \vec{n} q = \int_{\Omega} \text{div}(\vec{u}) q + \int_{\Omega} \vec{u} \cdot \nabla q$$

luego si $q \equiv 1$

$$\int_{\partial\Omega} \vec{u} \cdot \vec{n} = \int_{\Omega} \text{div}(\vec{u}).$$

este resultado nos da una interpretación física del operador divergencia en términos de lo que sale y entra por la frontera de cualquier dominio Ω de acuerdo al campo de vectores dado por \vec{u} . En particular, observando que

$$\vec{u} \cdot \vec{n} \text{ sobre } \partial\Omega$$

indica la proyección del vector \vec{u} en la dirección normal exterior \vec{n} en la frontera de la región Ω dada, entonces

$$\int_{\partial\Omega} \vec{u} \cdot \vec{n}$$

se entiende como la suma de todas las contribuciones en cada elemento de área de la superficie de la región. Por lo tanto, **en promedio** se puede entender que

- Si $\operatorname{div}(\vec{u}) > 0$ es un **campo expansivo**: las trayectorias salen de cualquier región acotada más de lo que entran
- Si $\operatorname{div}(\vec{u}) < 0$ es un **campo contractivo**: las trayectorias entran a cualquier región acotada más de lo que salen.
- Si $\operatorname{div}(\vec{u}) = 0$ es un **campo conservativo**: las trayectorias entran y salen de cualquier región acotada por igual. Se dice que es un **campo incompresible**, por ejemplo, las trayectorias de un fluido como el agua forman un campo incompresible, no se puede comprimir el agua.

Corolario 156 Observar que en el caso de dimensión uno, $d = 1$ y $\Omega = (a, b)$ el Teorema de la divergencia coincide con la **Regla de Barrow** puesto que el vector normal es $\vec{n} = 1$ en $x = b$ y $\vec{n} = -1$ en $x = a$

$$u(b) - u(a) = \int_a^b u'(x)dx.$$

Entonces $u'(x) > 0$ implica que $u(b) > u(a)$, $u'(x) < 0$ que $u(b) < u(a)$ y si $u'(x) = 0$ tendremos $u(b) = u(a)$.

Teorema 157 Teorema de integración por partes de Green: Si u y v son funciones escalares entonces

$$\int_{\partial\Omega} \frac{\partial u}{\partial \vec{n}} v = \int_{\Omega} \Delta u v + \int_{\Omega} \nabla u \cdot \nabla v.$$

También aquí si $v \equiv 1$ tenemos que

$$\int_{\partial\Omega} \frac{\partial u}{\partial \vec{n}} = \int_{\Omega} \Delta u$$

y el operador Laplaciano puede medir la razón de cambio de la magnitud u a través de la frontera de cualquier dominio Ω donde está la materia u . En particular, teniendo la idea de que

$$\frac{\partial u}{\partial \vec{n}} \sim u(\text{fuera}) - u(\text{dentro})$$

y que el sentido natural de expansión es de más a menos

- $\Delta u > 0$ la función es convexa y el sentido de cambio es hacia dentro en promedio ya que hay más fuera que dentro
- $\Delta u < 0$ la función es cóncava y el sentido de cambio es hacia fuera en promedio ya que hay más dentro que fuera
- $\Delta u = 0$ hay un equilibrio en los intercambios.

8.2. Ecuaciones diferenciales

En general, diremos que una ecuación en derivadas parciales es una ecuación que define a una función $u(t, x_1, \dots, x_d)$ usando sus derivadas y puede ser por lo tanto una expresión en la forma

$$F(t, x, u, u_t, u_{t,x_1}, \dots) = 0.$$

Se define el orden de la edp como el orden de derivación más alto que aparece en la ecuación. Una ecuación en derivadas parciales debe de venir acompañada con algunas condiciones extra que se suelen llamar **condiciones de contorno o condiciones iniciales**.

Una regla aproximada es que necesitamos tantas condiciones o datos auxiliares como derivadas aparecen en la ecuación.

Veamos algunos ejemplos muy básicos pero ilustrativos:

Ejemplo 158

$$u' = 1$$

con solución general $u(x) = x + c$ necesita un dato que fije c . Por ejemplo, en $x = 0$ poner $u(0) = c$. Así es un **problema de valor inicial**. Otro ejemplo puede ser

$$u'' = 0, \quad t > 0$$

con datos iniciales $u(0) = u_0$, $u'(0) = v_0$.

Ejemplo 159 Laplaciano y su interpretación física o geométrica: en $d = 1$

$$u'' = 1$$

con solución general $u(x) = x^2 + c_1 x + c_2$ necesita dos datos para fijar c_1 y c_2 . La naturaleza del problema también depende del tipo de datos que se dan. Por ejemplo, si ponemos

$$u(0) = 1; \quad u'(0) = 2$$

volvemos a tener un **problema de valor inicial** pero si usamos

$$u(0) = u(1) = 0$$

tenemos entonces un **problema de valor de contorno**. En este último caso, se genera la función convexa $u(x) = 0.5x(x-1)$ y se puede ver como su forma coincide con las interpretaciones que se han dado del Laplaciano. También puede ser

$$u'' = 1, \quad u(0) = 0, \quad u'(1) = \alpha$$

con solución

$$u(x) = 0.5x^2 + (\alpha - 1)x$$

y coincide con una curva que está en $x = 0$ en la posición 0 mientras que en $x = 1$ tiene tangente α todo bajo la restricción $u'' = 1$ que implica que la curva debe ser convexa. Por lo tanto podemos ver como el mínimo de la parábola es en $x = 1 - \alpha$ y tenemos trasladandose por el eje OX con valor fijo en $x = 0$ a $u(0) = 0$ y cumpliendo que $u'(1) = \alpha$ manteniendo su carácter convexo.

Este problema se generaliza bastante bien al caso $d = 2$ aunque aquí el dominio de cálculo es mucho más importante y las dificultades se abren haciendo imposible una solución analítica en la mayoría de los casos: Tenemos entonces $u = u(x, y)$ definida en Ω y el operador de Laplace es

$$\Delta u = u_{x,x} + u_{y,y}$$

por lo que la ecuación anterior se escribe como

$$\Delta u = 1.$$

Pero ahora $\Omega \subset \mathbb{R}^2$ por lo que su geometría y la frontera juega un papel decisivo. Por ejemplo, si $\Omega = \{x^2 + y^2 < 1\}$ entonces el análogo al primer ejemplo es

$$\begin{aligned} \Delta u &= 1, & \Omega \\ u &= 0, & \partial\Omega. \end{aligned}$$

De manera excepcional, aquí la solución se puede obtener de forma fácil, usando la intuición, y es

$$u(x, y) = \frac{1}{4}(x^2 + y^2 - 1)$$

pero en general esto no es posible. Si además queremos hacer algo similar al segundo ejemplo, debemos entonces partir la frontera en dos trozos y tener algo como

$$\begin{aligned} \Delta u &= 1, & \Omega, \\ u &= 0, & \Gamma_D, \\ \frac{\partial u}{\partial \vec{n}} &= \alpha, & \Gamma_N, \end{aligned}$$

en donde $\partial\Omega = \Gamma_D \cup \Gamma_N$. Observemos que el análogo a la derivada en la frontera del dominio cuando hay varias dimensiones se toma como $\frac{\partial u}{\partial \vec{n}}$.

Esto nos ha introducido la diferencia entre **problema de valor inicial** y **problema de contorno**. Cuando tenemos una variable que juega el papel del tiempo y otras que hacen el de la posición espacial nos encontramos con una ecuación en derivadas parciales dependiente del tiempo. Es la diferencia entre un problema dinámico (dependiendo del tiempo) y otro estático (que no depende del tiempo). En general, hablaremos de:

- **Problemas estacionarios:** todas las variables son espaciales y se describen situaciones de equilibrio frecuentes en mecánica y elasticidad.
- **Problemas evolutivos:** hay una variable temporal y el resto son variables espaciales. Se describen situaciones que cambian en tiempo como movimientos de ondas, difusión de calor o energía, dinámica de poblaciones con cambios globales sin direcciones espaciales o con direcciones espaciales. Los fenómenos evolutivos más habituales son
 - difusión....Ecuaciones elípticas
 - transporte....Ecuaciones hiperbólicas
 - reacción
 - combinaciones de los tres anteriores estacionario
 - combinaciones de los tres anteriores evolutivo...Ecuaciones parabólicas.

Ejemplo 160 Los problemas de contorno son más delicados pues podemos no tener unicidad de soluciones. Por ejemplo

$$u''(x) = u(x)$$

tiene por solución $u(x) = A \cos(x) + B \sin(x)$ entonces imponer $u(0) = A$ y $u(\pi) = -A$ nos lleva a tener infinitas soluciones ya que $\sin(0) = \sin(\pi) = 0$. Por otro lado, imponer $u(0) = A$ y $u(\pi) = C \neq -A$ no tiene solución.

Ejemplo 161 Para c una constante la ecuación

$$u_t(t, x) + cu_x(t, x) = 0, \quad t > 0, \quad x \in \mathbb{R}$$

es una edp conocida como **ecuación de transporte** y para cualquier función $f(z)$ resulta que

$$u(t, x) = f(x - ct)$$

es solución de la misma. Por ejemplo, si $f(z) = \sin(z)$ entonces $u(t, x) = \sin(t - cx)$ es la solución del problema

$$\begin{aligned} u_t(t, x) + cu_x(t, x) &= 0, \quad t > 0, \quad x \in \mathbb{R} \\ u(0, x) &= \sin(x), \quad x \in \mathbb{R} \end{aligned}$$

La visualización de $u(t, x) = \sin(t - cx)$ en el espacio txu nos dice que la solución es la traslación con velocidad constante c del valor inicial $u_0(x)$.

También desde el punto de vista físico se tiene una interpretación muy fácil: tenemos $u_t = -cu_x$ (y si $c > 0$ por ejemplo) entonces si $u_x < 0$ hay crecimiento de u en el sentido positivo de la x luego hay transporte hacia la derecha mientras que si $u_x > 0$ hay decrecimiento de u en el sentido positivo de la x lo que es coherente con el transporte hacia la derecha de la cantidad u . Esto es, si $c > 0$ tenemos una onda que viaja de izquierda a derecha.

Ejemplo 162 Para $k > 0$ una constante la ecuación

$$u_t - k u_{xx} = 0, \quad t > 0, \quad x \in \mathbb{R}$$

es una edp conocida como **ecuación de difusión o del calor**. Aquí la solución exacta ya no es simple. Pero también desde el punto de vista físico se tiene una interpretación:

Tenemos $u_t = k u_{xx}$. Supongamos que $u(t, x)$ representa un perfil de calor que sabemos que se propaga de más a menos. Entonces, si $u_{xx} > 0$ (convexa) hay crecimiento de u ya que hay más calor en el entorno que en el propio punto (t, x) mientras que si $u_{xx} < 0$ (concava) hay decrecimiento de u ya que hay menos calor en el entorno que en el propio punto (t, x) . La rapidez de la difusión del calor depende del medio donde se realiza y este efecto se describe mediante el valor de $k > 0$, a mayor k más rápida es la difusión. Al valor de $k > 0$ se le llama **coeficiente de difusión del medio**.

Observación 133 Mientras que $y'(t) = f(t, y(t))$ representa una familia de curvas, una edp, como por ejemplo, $u_t + au_x = 0$ representa una familia de superficies. Escogiendo los datos se fija la curva o superficie.

8.3. Leyes de conservación $d = 1$

La mayoría de las ecuaciones fundamentales que aparecen en la naturaleza y en las ciencias provienen de una ley de conservación: principio por el cual

la velocidad con la que una cantidad cambia en una región dada debe ser igual a la velocidad con la que entra menos la velocidad a la que sale más la velocidad de creación y/o destrucción de la propia cantidad dentro de la región.

Ejemplos evidentes de leyes de conservación se encuentran en todos los ámbitos: estudios demográficos, de fluidos, reacciones químicas, etc... Matemáticamente las leyes de conservación se traducen a ecuaciones diferenciales que son las que gobiernan el proceso en tiempo. Vamos a formular una ley de conservación en 1D:

Supongamos que tenemos una sustancia descrita por:

- una densidad $u = u(t, x)$ por unidad de volumen que depende de la variable espacial $x \in \mathbb{R}$ y temporal $t > 0$,
- esta cantidad está confinada en un tubo de área sección A y se mueve por la región se difunde, se transporta, etc..., de forma continua.
- en cada sección etiquetada con la abcisa $x = x_*$ la cantidad se mantiene uniforme.

Nos vamos entonces a fijar en el tramo de tubo que va desde $x = a$ a $x = b$ con $a < b$ para ver como cambia la sustancia con el tiempo. El volumen espacial que consideramos es $A(b - a)$. Si en el tiempo t la cantidad fuese constante en cada sección x entonces, $u(t, x) = u(t)$ y la cantidad total sería

$$\text{densidad de cantidad por volumen total} = u(t)\{A(b - a)\}$$

Pero si suponemos que en el tiempo t cambia en cada sección x entonces tenemos por unidad de volumen $A dx$ una cantidad

$$\text{densidad de cantidad por unidad de volumen} = u(t, x)\{A dx\}$$

de donde en total desde $x = a$ a $x = b$

$$\text{cantidad total} = \int_a^b u(t, x) A dx = A \int_a^b u(t, x) dx.$$

Denotamos por $\Phi(t, x)$, que puede ser $\Phi(t, x, u)$, el flujo: velocidad a la que se traspasa una sección de la cantidad u en la sección x , por unidad de área y de tiempo, esta función vendrá dada por la física del proceso. Si \vec{n} es el vector normal exterior entonces tenemos flujo hacia el exterior si

$$\Phi(t, x) \cdot \vec{n} > 0$$

y flujo hacia el interior si

$$\Phi(t, x) \cdot \vec{n} < 0.$$

En el caso escalar $\vec{n} = -1$ a la izquierda y $\vec{n} = 1$ a la derecha. Entonces, por ejemplo, hay salida por la izquierda si $\Phi(x) < 0$ y salida por la derecha si $\Phi(x) > 0$

Suponiéndolo constante por sección transversal tendremos que la cantidad que entra por unidad de área es

$$A\Phi(t, a)$$

y la que sale es

$$A\Phi(t, b)$$

por lo tanto, el flujo total en $[a, b]$ vendrá dado por la expresión

$$\text{flujo total a través de la frontera} = A\Phi(t, a) - A\Phi(t, b)$$

y si la razón de creación o destrucción de u viene dada por una función $f(t, x)$, que puede ser $f(t, x, u)$, entonces la razón de cambio en la sección $[a, b]$ es

$$\int_a^b f(t, x) A dx = A \int_a^b f(t, x) dx.$$

La ley de conservación nos dice entonces que

$$\frac{d}{dt} \int_a^b u(t, x) A dx = A \Phi(t, a) - A \Phi(t, b) + \int_a^b f(t, x) A dx$$

o bien

$$\frac{d}{dt} \int_a^b u(t, x) dx = \Phi(t, a) - \Phi(t, b) + \int_a^b f(t, x) dx.$$

Si las funciones que intervienen son regulares tendremos, usando el Teorema de la divergencia en dimensión uno y permutando derivada con integral,

$$\int_a^b \Phi_x(t, x) dx = \Phi(t, b) - \Phi(t, a), \quad \frac{d}{dt} \int_a^b u(t, x) dx = \int_a^b u_t(t, x) dx$$

de donde

$$\int_a^b u_t(t, x) dx = - \int_a^b \Phi_x(t, x) dx + \int_a^b f(t, x) dx$$

y de aquí

$$\int_a^b \{u_t(t, x) + \Phi_x(t, x) - f(t, x)\} dx = 0.$$

Suponiendo que esto se cumple en cualquier segmento $[a, b]$ y que las funciones son regulares, tenemos la ecuación diferencial en derivadas parciales

$$u_t(t, x) + \Phi_x(t, x) = f(t, x), \quad \forall x \in I, \quad \forall t > 0$$

donde se puede identificar el **término de flujo** $\Phi_x(t, x)$ y el **término fuente o de reacción** $f = f(t, x)$.

8.3.1. Ejemplos

- Si el flujo sigue una **ley de difusión** estandar, esto equivale a decir que la materia se mueve de donde hay más a donde hay menos, que es la ley más común en la mayoría de los fenómenos. Esto depende de la derivada y de su signo. De más a menos implica $u_x < 0$ luego la dirección del flujo es contraria a este signo. Además, podemos medir la resistencia del medio por un coeficiente $k > 0$: a mayor k más rápido es la difusión. Luego el flujo es opuesto a $k u_x(x, t)$ puesto que se propaga cuando $u_x < 0$.

Esto se llama la **Ley de difusión, Ley de Fick o Ley de Fourier** (*dependiendo del campo de la ciencia donde se ha trabajado*) entonces

$$\Phi(t, x, u) = -k u_x(x, t)$$

y tendremos la **ecuación del calor**

$$u_t(t, x) - k u_{xx}(t, x) = f(t, x), \quad \forall x \in I, \quad \forall t > 0$$

propotipo de **ecuación parabólica**.

- Si el flujo es proporcional a la materia existente hay un transporte o convección (convección es sinónimo de transporte)

$$\Phi(t, x) = c u(t, x)$$

entonces tenemos la **ecuación de transporte**

$$u_t(t, x) + c u_x(t, x) = f(t, x), \quad \forall x \in I, \quad \forall t > 0.$$

propotipo de **ecuación hiperbólica**.

- Si al mismo tiempo que tenemos un transporte hay una difusión entonces el flujo es

$$\Phi(t, x) = c u - k u_x$$

y tenemos la **ecuación de convección difusión**

$$u_t + c u_x - k u_{xx} = f(t, x), \quad \forall x \in I, \quad \forall t > 0.$$

- Si también tenemos que la densidad depende de su propia cantidad, por ejemplo un reactante en continua producción o extinción, entonces por ejemplo en el caso más simple de tener una reacción lineal

$$f(t, x) = \lambda u$$

donde $\lambda > 0$ implica extinción de u y $\lambda < 0$ implicará destrucción de u . Tenemos entonces la **ecuación de convección difusión reacción**

$$u_t + c u_x - k u_{xx} = \lambda u, \quad \forall x \in I, \quad \forall t > 0.$$

- La **Ley de Fisher**

$$u_t - k u_{xx} = f(t, x, u), \quad \forall x \in I, \quad \forall t > 0.$$

generaliza el modelo de población al caso donde hay una difusión espacial

$$f(t, x) = \lambda u(1 - \frac{u}{K})$$

donde $\lambda > 0$ implica extinción de u y $\lambda < 0$ implicará destrucción de u . Tenemos entonces la **ecuación de Fisher**

$$u_t + c u_x - k u_{xx} = \lambda u(1 - \frac{u}{K}), \quad \forall x \in I, \quad \forall t > 0.$$

- Cuando la solución alcanza el estado estacionario, $u_t = 0$, podemos tener la ecuación de difusión

$$-u_{xx} = f(x, u), \quad \forall x \in I.$$

Observación 134 Si no hay dependencia espacial, $u = u(t)$, recuperamos los modelos de edos que ya conocemos. Por ejemplo, la **ecuación de convección difusión reacción**

$$u_t + cu_x - k u_{xx} = \lambda u, \quad \forall x \in I, \quad \forall t > 0.$$

se reduce a la reacción

$$u_t = \lambda u, \quad \forall t > 0$$

y la **ecuación de Fisher**

$$u_t + cu_x - k u_{xx} = \lambda u(1 - \frac{u}{K}), \quad \forall x \in I, \quad \forall t > 0.$$

a

$$u_t = \lambda u(1 - \frac{u}{K}), \quad \forall t > 0.$$

Ejemplo 163 Un caso muy importante es cuando el transporte de u es una función no lineal de u y entonces

$$\Phi(t, x, u) = \varphi(u) - k u_x.$$

Esto genera el modelo

$$u_t + \varphi'(u)u_x - k u_{xx} = f(t, x, u), \quad \forall x \in I, \quad \forall t > 0$$

y en el caso $\varphi'(u) = u$ tenemos la versión 1D de las **ecuaciones de Navier-Stokes** que rige el movimiento de un fluido, más conocida como **ecuación de Burgers**. Este es el **modelo fundamental de la dinámica de fluidos** que muestra el acoplamiento entre convección y difusión.

8.4. Leyes de conservación $d > 1$

Pensemos que en una región espacial $\Omega \subset \mathbb{R}^d$ ($d = 1, 2, 3, \dots$) ocupada por una determinada materia, por ejemplo: un fluido, aire, agua, etc... con una determinada propiedad de que queremos estudiar, por ejemplo, una densidad, temperatura, concentración de contaminante, etc... Dada cualquier subdominio $\omega \subset \Omega \subset \mathbb{R}^d$ entonces la **ley de conservación** es

La velocidad de cambio de una materia en w viene dada en términos de la cantidad de esta materia que entra y sale por la frontera de w y la cantidad que se produce o destruye dentro del dominio de interés w .

En términos matemáticos esto lo podemos describir como

$$\frac{d}{dt} \int_{\omega} u \, dx = - \int_{\partial\omega} \vec{\Phi} \cdot \vec{n} \, d\sigma + \int_{\omega} f(x) \, dx \quad (8.1)$$

Aquí, tenemos que $u = u(x, t)$ es la cantidad de materia que estamos observando dentro de ω , $\vec{\Phi}$ es el flujo a través de la frontera $\partial\omega$ y f es la fuente o sumidero que se encuentra presente en ω .

El hecho de que esto se tenga que cumplir para cualquier subdominio ω nos lleva a una relación puntual como una ecuación diferencial en derivadas parciales de la forma

$$\partial_t u + \operatorname{div}(\vec{\Phi}) - f = 0, \quad \forall x \in \Omega. \quad (8.2)$$

La función de flujo $\vec{\Phi}$ y la fuente f tienen que ser determinadas de acuerdo al modelo físico que se pretende estudiar.

- **Ley de difusión:** Por ejemplo, en la difusión del calor o un contaminante, nos encontramos con la **Ley de Fourier-Fick**. Se plantea la hipótesis de que el flujo es proporcional al gradiente de la densidad y en el sentido de mayor a menor densidad:

$$\vec{\Phi} = -k \vec{\nabla} u(x, t) \quad (8.3)$$

en donde $k > 0$ es la conductividad térmica del medio, a mayor valor más fácil la conducción, y puede ser $k = k(x, t)$. Cuando por simplicidad tomamos $k \equiv \text{cte}$ nos encontramos con la ecuación del calor evolutiva

$$\partial_t u - k \Delta u = f(t, x), \quad \forall x \in \Omega \quad (8.4)$$

y si nos interesa el caso estacionario $u(x, t) = u(x)$ entonces tendremos la ecuación

$$-\Delta u = f, \quad \forall x \in \Omega. \quad (8.5)$$

- **Ley de difusión y transporte:** Hemos supuesto que la propagación se hace en un medio que está estacionario, pero podría ocurrir que el medio fuese un fluido que se moviese con velocidad $\vec{V}(t, x)$ (el viento reinante). Entonces el flujo sería una suma de la difusión, como antes, y de la convección o transporte, realizada por el fluido

$$\vec{\Phi} = -k \vec{\nabla} u(x, t) + \vec{V}(t, x) u(x, t), \quad (8.6)$$

usando cálculo diferencial básico,

$$\operatorname{div}(\vec{\Phi}) = -\vec{\nabla} k \cdot \nabla u(x, t) - k \Delta u(x, t) + \operatorname{div}(\vec{V}) u(x, t) + (\vec{V} \cdot \vec{\nabla}) u$$

nos podemos encontrar con la ecuación de convección y difusión

$$\partial_t u(t, x) - k \Delta u(t, x) + (\vec{V} - \vec{\nabla} k) \cdot \vec{\nabla} u + \operatorname{div}(\vec{V}) u(x, t) = f(x, t),$$

Aquí vemos que si simplificamos $k \equiv \text{cte}$ entonces

$$\partial_t u(t, x) - k \Delta u(t, x) + \vec{V} \cdot \vec{\nabla} u = -\operatorname{div}(\vec{V}) u(x, t) + f(x, t),$$

y si el campo es expansivo, $\operatorname{div}(\vec{V}) > 0$ se contribuye a que la materia se pierda, se vaya de la región, mientras que si el campo es contractivo, $\operatorname{div}(\vec{V}) < 0$, se contribuye a la ganancia de materia. Finalmente, si $\operatorname{div}(\vec{V}) = 0$ el viento reinante no influye en la pérdida o ganancia de materia y tenemos la versión simplificada de la ecuación de convección y difusión

$$\partial_t u(t, x) - k \Delta u(t, x) + \vec{V} \cdot \vec{\nabla} u = f(t, x),$$

8.5. Condiciones de contorno y dato inicial

Estas ecuaciones son válidas en todo el dominio de cálculo Ω y debemos de suplementar otras relaciones como son

- **una condición inicial**, que describe el estado inicial de la magnitud. Esta condición se escribe como $u(x, 0) = u_0(x)$ y se puede pensar como el dato que hay que dar debido a la existencia de la derivada $\partial_t u(x, t)$ en la ecuación.
- **una condición de contorno**, que describe que es lo que ocurre en la frontera del dominio de cálculo. Dependerán de la ecuación y se podrán pensar como asociadas a las derivadas espaciales que aparezcan en la ecuación.

Por convención, se toma el instante inicial $t = 0$, y se impone un dato inicial conocido

$$u(x, 0) = u_0(x).$$

El tipo de condición de contorno depende del problema que se está estudiando. Podemos distinguir dos tipos principales:

- **condición de contorno de tipo Dirichlet:** Si la función u tiene un valor predeterminado en todo el contorno tendríamos un dato

$$u(x, t) = d(x, t), \quad \forall x \in \partial\Omega, \quad t > 0.$$

- **condición de tipo Neumann:** Si existe un flujo conocido de la cantidad u a través de la frontera de la región entonces la condición se puede plantear como

$$\frac{\partial u}{\partial \vec{n}} = \vec{n} \cdot \vec{\nabla} u = g(x, t), \quad \forall x \in \partial\Omega, \quad t > 0.$$

- **condición de tipo Fourier o Robin:** Esta es una situación intermedia en la que el flujo a través de la frontera del dominio es proporcional al salto de materia entre el exterior y el interior

$$\frac{\partial u}{\partial \vec{n}} + \alpha u = g(x, t), \quad \forall x \in \partial\Omega, \quad t > 0,$$

donde $\alpha > 0$ es una constante positiva.

Estas condiciones se pueden dar en distintas partes del contorno y hablamos de **condiciones de contorno de tipo mixtas**.

Uno de los pasos claves en la modelización es la elección de las condiciones de contorno. Si escogemos condiciones de tipo Dirichlet y tomamos un coeficiente de difusión constante $k(x) \equiv k$ tendremos entonces el **problema de Cauchy** (debido al dato inicial)

$$\partial_t u(x, t) - k\Delta u(x, t) = f(x, t), \quad \forall x \in \Omega, t \in [0, T], \quad (8.7)$$

$$u(t, x) = 0, \quad \forall x \in \partial\Omega, \quad t > 0, \quad (8.8)$$

$$u(t = 0, x) = u_0(x), \quad \forall x \in \Omega. \quad (8.9)$$

Se puede dibujar el cono temporal $[0, T] \times \Omega$ para tener una visual sencilla. Las ecuaciones (9.1)-(9.1)-(9.1) constituyen un ejemplo de ecuación en derivadas parciales equipada con condiciones de contorno y valor inicial. Estas ecuaciones también son un ejemplo de lo que se conoce como un **modelo de difusión** puesto que modela la difusión por el dominio Ω de la cantidad genérica u . De aquí proviene la tradición de trabajar con el operador $-\Delta$ en vez de Δ .

En general, podemos dar las expresiones generales para el **problema de Cauchy** (debido al dato inicial)

$$\partial_t u + F(t, x, u, u_x, \dots) = 0, \quad \forall x \in \Omega, t \in [0, T], \quad (8.10)$$

$$u(t = 0, x) = u_0(x), \quad \forall x \in \Omega. \quad (8.11)$$

o bien si es de segundo orden en tiempo

$$\partial_{tt} u + F(t, x, u, u_x, \dots) = 0, \quad \forall x \in \Omega, t \in [0, T], \quad (8.12)$$

$$u(t = 0, x) = u_0(x), \quad \forall x \in \Omega, \quad (8.13)$$

$$u_t(t = 0, x) = v_0(x), \quad \forall x \in \Omega. \quad (8.14)$$

Observación 135 Uno de los objetivo más importante a nivel científico y aplicado es resolver las **ecuaciones de Navier-Stokes** de los fluidos

$$\partial_t \vec{u}(t, x) + \vec{u} \cdot \nabla \vec{u} - k\Delta \vec{u} + \nabla p(t, x) = \vec{f}(x, t), \quad \forall x \in \Omega, t > 0, \quad (8.15)$$

$$\operatorname{div}(\vec{u}) = 0, \quad \forall x \in \Omega, t > 0, \quad (8.16)$$

$$u(0, x) = u_0(x), \quad \forall x \in \Omega. \quad (8.17)$$

que incluye todos los efectos antes mencionados, por ejemplo, el propio campo vectorial define el viento reinante $\vec{V} = \vec{u}$, y algunos más.

Capítulo 9

La ecuación de difusión

El método de las diferencias finitas es uno de los más antiguos todavía en uso en muchas aplicaciones como propagación de ondas, sísmicas o electromagnéticas, o en dinámica de fluidos. La base conceptual reside en el uso de los desarrollos de Taylor para aproximar cualquier derivada en un punto.

Otra razón no menos importante es su generalidad y simplicidad conceptual así como poder ser aplicados a cualquier tipo de derivada que aparezca en la ecuación.

*En este tema veremos algunos esquemas numéricos en diferencias finitas. Se definen los conceptos de **estabilidad** y **consistencia** de un esquema. Se demuestra que para un esquema asociado a una edp lineal y de coeficientes constantes, estabilidad más consistencia implica **convergencia**.*

*Uno de los pasos claves en la modelización es la elección de las condiciones de contorno. Si escogemos condiciones de tipo Dirichlet y tomamos un coeficiente de difusión en el medio constante $\kappa(x) \equiv \kappa$. Tendremos entonces el siguiente **problema de contorno y de valor inicial** (debido al dato inicial)*

$$\begin{aligned} u_t(t, x) - \kappa \Delta u(t, x) &= f(t, x), \quad \forall x \in \Omega, t \in [0, T], \\ u(t, x) &= \varphi(x), \quad \forall x \in \partial\Omega, \quad t > 0, \\ u(t = 0, x) &= u_0(x), \quad \forall x \in \Omega. \end{aligned}$$

9.1. Decaimiento debido a la difusión

Vamos a ver como también estamos modelando una propiedad característica de la difusión como es su decaimiento. Fijemos ideas y tomemos el caso $\Omega = (a, b)$ con datos homogéneos. La ecuación queda como

$$u_t(t, x) - \kappa u_{xx}(t, x) = 0, \quad \forall x \in (a, b), t \in [0, T], \quad (9.1)$$

$$u(t, a) = 0, \quad \forall t > 0, \quad (9.2)$$

$$u(t, b) = 0, \quad \forall t > 0, \quad (9.3)$$

$$u(0, x) = u_0(x), \quad \forall x \in [a, b]. \quad (9.4)$$

Se puede interpretar como que no hay fuentes de calor y que en los extremos hay un aislante que anula la temperatura. Vamos a comprobar como entonces la temperatura tiende a extinguirse. El cálculo que vamos a hacer se conoce como **método de la energía** y consiste en estimar la energía total que se representa en cada instante t por:

$$\|u(t)\|_{L^2(a,b)} = \left(\int_a^b u(t,x)^2 dx \right)^{1/2}.$$

Teorema 164 La solución del problema (9.1)-(9.4) verifica

$$\|u(t)\|_{L^2(a,b)} \leq e^{-\kappa(b-a)^{-1}t} \|u_0\|_{L^2(a,b)}, \quad \forall t > 0.$$

Dem: Lo primero que podemos ver es que para cada t la función $u(t,x)$ está controlada por su derivada espacial al ser cero en sus extremos. Esto se hace usando la regla de Barrow que nos permite controlar la función en términos de su derivada:

$$\begin{aligned} u(x,t) - u(a,t) &= \int_a^x u_x(t,x) dx \leq (\int_a^x 1^2 dx)^{1/2} (\int_a^x u_x(t,x)^2 dx)^{1/2} \\ &\leq (b-a)^{1/2} (\int_a^x u_x(t,x)^2 dx)^{1/2}, \end{aligned}$$

de donde, como $u(t,a) = 0$,

$$u(t,x)^2 \leq (b-a) \int_a^b u_x(t,x)^2 dx$$

y entonces integrando en x y observando que el segundo término es una constante

$$\int_a^b u(t,x)^2 dx \leq (b-a)^2 \int_a^b u_x(t,x)^2 dx.$$

Por lo tanto,

$$\|u(t)\|_{L^2(a,b)} \leq (b-a) \|u_x(t)\|_{L^2(a,b)}, \quad \forall t > 0.$$

Por otro lado, usando la ecuación

$$u_t = \kappa u_{xx} \Rightarrow u u_t = \kappa u u_{xx}$$

de donde integrando $x \in [a,b]$ tenemos

$$\begin{aligned} \frac{1}{2} \int_a^b \frac{\partial}{\partial t} u^2(t,x) dx &= \kappa \int_a^b \{(u u_x)_x - u_x^2\} dx \\ &= \kappa(u u_x)_{x=a}^{x=b} - \kappa \int_a^b u_x^2 dx \\ &= -\kappa \int_a^b u_x^2 dx \end{aligned}$$

de donde se obtiene

$$\frac{d}{dt} \int_a^b u^2(t, x) dx = -2 \kappa \int_a^b u_x(t, x)^2 dx$$

es decir,

$$\frac{d}{dt} \|u(t)\|_{L^2(a,b)}^2 = -2 \kappa \|u_x(t)\|_{L^2(a,b)}^2 \leq -2 \kappa (b-a)^{-1} \|u(t)\|_{L^2(a,b)}^2$$

de donde

$$\frac{d}{dt} \|u(t)\|_{L^2(a,b)}^2 + 2 \kappa (b-a)^{-1} \|u(t)\|_{L^2(a,b)}^2 \leq 0$$

y multiplicando por el factor $e^{2\kappa(b-a)^{-1}t}$ tenemos

$$\frac{d}{dt} \{ e^{2\kappa(b-a)^{-1}t} \|u(t)\|_{L^2(a,b)}^2 \} \leq 0$$

de donde

$$\|u(t)\|_{L^2(a,b)} \leq e^{-\kappa(b-a)^{-1}t} \|u_0\|_{L^2(a,b)}$$

y tenemos un decaimiento exponencial de la solución. Más rápido cuanto más grande sea $\kappa > 0$. Luego el calor se difunde hasta desaparecer con una velocidad exponencial. Este resultado también es cierto en dimensión espacial mayor que uno.

9.2. Diferencias finitas

Vamos a fijarnos en el problema modelo para ir introduciendo conceptos que son generales. Nuestro problema de valor inicial y de contorno en el caso $\Omega = (a, b)$ es

$$u_t(t, x) - \kappa u_{xx}(t, x) = f(t, x), \quad \forall x \in (a, b), t \in [0, T], \quad (9.5)$$

$$u(t, a) = r(t), \quad \forall t > 0, \quad (9.6)$$

$$u(t, b) = s(t), \quad \forall t > 0, \quad (9.7)$$

$$u(0, x) = u_0(x), \quad \forall x \in [a, b]. \quad (9.8)$$

Pongamos $\mathcal{C} = [0, T] \times \Omega$ donde $\Omega = (a, b)$, buscamos $u : [0, T] \times \Omega \rightarrow \mathbb{R}$ que sea una solución de este problema. Vamos a generar una parrilla de puntos en \mathcal{C} desde una partición de $[0, T]$ y otra de Ω . Dados enteros $N \geq 1$ y $M \geq 1$ tomamos

$$\Delta t = \frac{T}{N}, \quad \Delta x = \frac{b-a}{M+1}$$

y construimos las particiones

$$0 = t_0 < t_1 < \dots < t_N = T, \quad a = x_0 < x_1 < \dots < x_d < x_{M+1} = b$$

consideramos entonces la parrilla de puntos

$$\mathcal{C}_{\Delta t, \Delta x} = \{(t^n, x_j); \quad t_n = n \Delta t, \quad x_j = j \Delta x\}$$

Al conjunto $\mathcal{C}_{\Delta t, \Delta x}$ lo llamamos **dominio computacional** y nos sirve para aproximar a \mathcal{C} cuando $\Delta t \rightarrow 0$ y $\Delta x \rightarrow 0$ rellenándolo. Queremos buscar valores discretos v_j^n que aproximen a los valores exactos

$$v_j^n \sim u(t^n, x_j).$$

Al instante de tiempo t^n lo llamamos nivel temporal n . Observar que como $u(t, a) = r(t)$ y $u(t, b) = s(t)$ son datos conocidos entonces v_0^n y v_{M+1}^n son datos del problema

$$v_0^n = r(t_n), \quad v_{M+1}^n = s(t_n), \quad \forall n \geq 0.$$

También son datos los valores en el nivel de tiempo $n = 0$ puesto que $u(0, x) = u_0(x)$ luego

$$v_j^0 = u_0(x_j), \quad \forall j = 0, 1, \dots, M + 1.$$

Por lo tanto, nuestras incógnitas son los valores

$$\{v_j^n\}_{j=1,\dots,M}^{n=1,\dots,N}$$

o bien, para cada nivel temporal n , los valores incógnita son

$$\{v_j^n\}_{j=1,\dots,M}.$$

Cuando $\Delta t, \Delta x \rightarrow 0$ entonces $\mathcal{C}_{\Delta t, \Delta x}$ rellena \mathcal{C} y queremos que en el **límite estacionario** cuando $\Delta t, \Delta x \rightarrow 0$ y $n, j \rightarrow \infty$ tal que $n\Delta t = t$ y $j\Delta x = x$ se cumpla

$$\lim_{n\Delta t=t, j\Delta x=x} v_j^n = u(t, x).$$

Veamos como empezar: tenemos que aproximar la solución en un punto interior de $\mathcal{C}_{\Delta t, \Delta x}$, pongamos (t^n, x_j) donde $n \geq 1$ y $1 \leq j \leq M$. Sólo tenemos la ecuación y sabemos que

$$u_t(t^n, x_j) = \kappa u_{xx}(t^n, x_j)$$

pero las derivadas se pueden aproximar por valores puntuales de la función usando Taylor. Para la derivada temporal tenemos

$$\frac{u(t + \Delta t, x) - u(t, x)}{\Delta t} = u_t(t, x) + \frac{\Delta t}{2} u_{tt}(\xi, x) = u_t(t, x) + O(\Delta t)$$

donde $\xi \in (t, t + \Delta t)$ y para la segunda derivada espacial, desarrollando hasta la cuarta derivada,

$$\begin{aligned} \frac{u(t, x + \Delta x) - 2u(t, x) + u(t, x - \Delta x)}{\Delta x^2} &= u_{xx}(t, x) + \frac{\Delta x^2}{12} u_{xxxx}(t, \delta) \\ &= u_{xx}(t, x) + O(\Delta x^2) \end{aligned}$$

donde $\delta \in (x, x + \Delta x)$. Por lo tanto, con un error $O(\Delta t)$

$$\frac{u(t + \Delta t, x) - u(t, x)}{\Delta t} \approx u_t(t, x)$$

y con un error $O(\Delta x^2)$

$$\frac{u(t, x + \Delta x) - 2u(t, x) + u(t, x - \Delta x)}{\Delta x^2} \approx u_{xx}(t, x)$$

Estas son las **diferencias finitas** que nos sirven para poder aproximar las derivadas que aparecen en la ecuación por valores nodales de la función que buscamos.

Tenemos ahora varias opciones a explorar y vamos a ver primero la más sencilla

9.3. Euler explícito en tiempo y diferencias centradas en espacio (Eex-CE)

La primera idea es fijarnos en los datos que tenemos en el nivel t^n para avanzar al nivel t^{n+1} , esto es, seguir la idea de Euler explícito. Veremos que el decaimiento exponencial de las soluciones nos plantea una restricción severa sobre Δt , al igual que ocurre con las ecuaciones diferenciales ordinarias cuyas soluciones poseen un decaimiento exponencial.

Haciendo desarrollo de Taylor en torno a un punto interior en el nivel de tiempo conocido, (t^n, x_j) , y avanzando hacia el siguiente nivel de tiempo, $(t^n + \Delta t, x_j)$, tenemos

$$\begin{aligned} \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} &= u_t(t^n, x_j) + \frac{\Delta t}{2} u_{tt} + \frac{\Delta t^2}{6} u_{ttt} + \dots \\ \frac{u(t^n, x_{j+1}) - 2u(t^n, x_j) + u(t^n, x_{j-1})}{\Delta x^2} &= u_{xx}(t^n, x_j) + \frac{\Delta x^2}{12} u_{xxxx} + \dots \end{aligned}$$

de donde

$$\begin{aligned} f(t^n, x_j) = u_t(t^n, x_j) - \kappa u_{xx}(t^n, x_j) &= \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} \\ &\quad - \kappa \frac{u(t^n, x_{j+1}) - 2u(t^n, x_j) + u(t^n, x_{j-1})}{\Delta x^2} \\ &\quad + O(\Delta t) + O(\Delta x^2) \end{aligned}$$

o abreviadamente, poniendo $u_j^n = u(t^n, x_j)$, $f_j^n = f(t^n, x_j)$

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \kappa \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} - f_j^n + O(\Delta t) + O(\Delta x^2) = 0.$$

Eliminando los términos $O(\Delta t) + O(\Delta x^2)$ tenemos la posibilidad de construir valores v_j^n que cumplan

$$\frac{v_j^{n+1} - v_j^n}{\Delta t} - \kappa \frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{\Delta x^2} = f_j^n, \quad n \geq 0, \quad j = 1, \dots, M$$

con la esperanza de que $v_j^n \sim u_j^n = u(t^n, x_j)$. Además, imponemos los datos conocidos por el dato inicial $v_j^0 = u_j^0$ y por los datos de contorno $v_0^n = r^n$ junto a $v_{M+1}^n = s^n$.

Aquí se puede ver que lo único que no se conoce es el valor v_j^{n+1} para $j = 1, 2, 3, \dots, M$ ya que todo el nivel n se puede dar por conocido usando los datos. Por lo tanto, es un **cálculo explícito**, se puede despejar y escribir como

$$v_j^{n+1} = \mu v_{j+1}^n + (1 - 2\mu)v_j^n + \mu v_{j-1}^n + \Delta t f_j^n, \quad j = 1, \dots, M \quad n \geq 0,$$

en donde

$$\mu = \kappa \frac{\Delta t}{\Delta x^2} \quad \text{se llama el } \mathbf{número de Courant}$$

y para los casos $j = 1$ y $j = M$ usamos los valores conocidos en los extremos. Nuestro esquema de cálculo queda como sigue:

Dados $v_j^0 = u_j^0$ para $j = 0, \dots, M+1$ y los valores $v_0^n = u_0^n = r(t^n) = r^n$, $v_{M+1}^n = u_{M+1}^n = s(t^n) = s^n$ para cualquier $n \geq 0$, se plantea obtener para cada nivel temporal $n \geq 0$ los valores

$$v_j^{n+1} = \mu v_{j+1}^n + (1 - 2\mu)v_j^n + \mu v_{j-1}^n + \Delta t f_j^n, \quad j = 1, 2, \dots, M.$$

Observar que para el cálculo de v_1^{n+1} y de v_M^{n+1} usamos los valores de contorno v_0^n y v_{M+1}^n . Esto es, lo escribimos por una mayor claridad,

$$\begin{aligned} v_1^{n+1} &= \mu v_2^n + (1 - 2\mu)v_1^n + \mu r^n + \Delta t f_1^n, \\ v_j^{n+1} &= \mu v_{j+1}^n + (1 - 2\mu)v_j^n + \mu v_{j-1}^n + \Delta t f_j^n, \quad j = 2, \dots, M-1 \\ v_M^{n+1} &= \mu s^n + (1 - 2\mu)v_M^n + \mu v_{M-1}^n + \Delta t f_M^n. \end{aligned}$$

Al realizar los cálculos indicados por este esquema descubriremos que la calidad de los resultados depende del factor μ . Veremos que la solución es estable sólo para $\mu \leq 1/2$. Por lo tanto, este esquema es **condicionalmente estable** ya que se necesita una relación entre los parámetros dada por $\Delta t \leq \Delta x^2/(2\kappa)$.

Observación 136 Esto se ve fácilmente en el caso $f = 0$ si pedimos que los valores $\{v_j^{n+1}\}$ estén acotados por los valores $\{v_j^n\}$ y ponemos $\mu \leq 1/2$. En efecto,

$$|v_j^{n+1}| \leq \mu |v_{j+1}^n| + (1 - 2\mu)|v_j^n| + \mu |v_{j-1}^n|$$

de donde

$$\begin{aligned}\sum_j |v_j^{n+1}| &\leq \mu \sum_j |v_{j+1}^n| + (1 - 2\mu) \sum_j |v_j^n| + \mu \sum_j |v_{j-1}^n| \\ &= (\mu + 1 - 2\mu + \mu) \sum_j |v_j^n| \\ &\leq \sum_j |v_j^n|.\end{aligned}$$

Por lo tanto, si $\mu \leq 1/2$ los valores se acotan y hay un decaimiento que es el natural por la física del problema.

9.3.1. Uso de matrices

El uso de la notación matricial va a permitirnos simplificar notación y nos abrirá la puerta a una forma de trabajar más simple. Poniendo, para hacerlo más sencillo, los valores del contorno cero, tenemos que

$$\begin{aligned}v_1^{n+1} &= \mu v_2^n + (1 - 2\mu) v_1^n, \\ v_2^{n+1} &= \mu v_3^n + (1 - 2\mu) v_2^n + \mu v_1^n, \\ v_3^{n+1} &= \mu v_4^n + (1 - 2\mu) v_3^n + \mu v_2^n, \\ &\dots = \dots, \\ v_M^{n+1} &= (1 - 2\mu) v_M^n + \mu v_{M-1}^n\end{aligned}$$

Entonces para $A_M = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{M \times M}$, si ponemos

$$P_{Ex} = \text{tridiag}(\mu, 1 - 2\mu, \mu) = I_M - \mu A_M$$

el esquema de cálculo es

$$V^{n+1} = P_{Ex} V^n = P_{Ex}^{n+1} V^0, \quad n \geq 0$$

con el significado obvio de $V^n = (v_1^n, v_2^n, \dots, v_M^n) \in \mathbb{R}^M$.

9.4. Mejora en estabilidad y orden

Cuando hagamos el estudio de Euler explícito veremos que podemos mejorar en dos aspectos, en estabilidad y en orden de convergencia.

Vamos a introducir dos métodos donde los parámetros no necesitan ser restringidos por estabilidad y donde el orden de convergencia es mejor.

Primero usaremos Euler implícito en tiempo y diferencias centradas en espacio. Este es un método con un error $O(\Delta t) + O(\Delta x^2)$ que no necesita relacionar Δt con Δx . Después vamos a ver el método de Crank-Nicolson que tiene un error $O(\Delta t^2) + O(\Delta x^2)$ y tampoco es necesario relacionar o restringir Δt con Δx . Tener el mismo orden para Δt y Δx es un resultado óptimo.

Observación 137 Uso de norma discreta en la variable espacial: Nos vamos a encontrar matrices simétricas tridiagonales definidas positivas y con autovalores reales y tendremos que controlar sus potencias. Observemos que entonces

$$P = VDV^t, \quad D = \text{diag}(d_1, \dots, d_M)$$

donde $VV^t = I_M$, entonces

$$P^n = VD^nV^t$$

y por lo tanto,

$$\|P^n\|_2 = \|VD^nV^t\|_2 = \|D^n\|_2 = \max_{j=1,\dots,M} |d_j|^n.$$

Luego es más conveniente usar en \mathbb{R}^M la norma vectorial $\|v\|_2 = \sqrt{v_1^2 + \dots + v_M^2}$ ya que induce la norma matricial $\|\cdot\|_2$ que ya vemos que tiene buenas propiedades a la hora de calcular potencias de una matriz.

Por otro lado, vamos también a introducir un factor de reescalamiento $\Delta x^{1/2}$ para tener en cuenta que cuando $\Delta x \rightarrow 0$ entonces $M \rightarrow +\infty$ y la dimensión finita del espacio vectorial crece. Usaremos

$$\|v\|_{\Delta x} = \left(\Delta x \sum_{j=1}^M |v_j|^2 \right)^{1/2}.$$

lo que no afecta para nada a la norma matricial inducida. Es claro que

$$\|g\|_{\Delta x} = \Delta x^{1/2} \|g\|_2.$$

La razón de este factor reside en que si $v_j = g(x_j)$ para una función continua $g(x)$ entonces

$$\lim_{\Delta x \rightarrow 0} \|g\|_{\Delta x} = \lim_{\Delta x \rightarrow 0} \left(\Delta x \sum_{j=1}^M |g(x_j)|^2 \right)^{1/2} = \left(\int_a^b |g(x)|^2 dx \right)^{1/2}$$

usando la suma de Riemann. Por lo tanto, reescalar la norma vectorial por el factor $\Delta x^{1/2}$ nos permite pasar de los valores nodales, en los puntos x_j , a un valor integral para la función, es decir, una norma de la función. También, por ejemplo, si $\mathbf{1} \in \mathbb{R}^M$ representa el vector con todas sus componentes 1 tenemos que

$$\|\mathbf{1}\|_{\Delta x} = \left(\Delta x \sum_{j=1}^M 1^2 \right)^{1/2} = \left(\Delta x M \right)^{1/2} = \left(\frac{M}{M+1} \right)^{1/2} \rightarrow 1.$$

9.4.1. Euler implícito en tiempo diferencias centradas en espacio (Eim-CE)

Siguiendo la idea de Euler implícito podemos buscar la información desde el punto (t^{n+1}, x_j) tomando datos del nivel temporal $n + 1$ y entonces acoplamos las incógnitas. Igual que antes,

$$\begin{aligned}\frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} &= u_t(t^{n+1}, x_j) + \frac{\Delta t}{2} u_{tt} + \dots \\ \frac{u(t^{n+1}, x_{j+1}) - 2u(t^{n+1}, x_j) + u(t^{n+1}, x_{j-1})}{\Delta x^2} &= u_{xx}(t^{n+1}, x_j) + \frac{\Delta x^2}{2} u_{xxxx} + \dots\end{aligned}$$

de donde

$$\begin{aligned}0 = u_t(t^{n+1}, x_j) - \nu u_{xx}(t^{n+1}, x_j) &= \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} \\ &\quad - \kappa \frac{u(t^{n+1}, x_{j+1}) - 2u(t^{n+1}, x_j) + u(t^{n+1}, x_{j-1})}{\Delta x^2} \\ &\quad + O(\Delta t) + O(\Delta x^2)\end{aligned}$$

poniendo $u_j^n = u(t^n, x_j)$,

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \kappa \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{\Delta x^2} = O(\Delta t) + O(\Delta x^2).$$

Eliminando los términos $O(\Delta t) + O(\Delta x^2)$ construimos valores v_j^n que cumplan

$$\frac{v_j^{n+1} - v_j^n}{\Delta t} - \kappa \frac{v_{j+1}^{n+1} - 2v_j^{n+1} + v_{j-1}^{n+1}}{\Delta x^2} = 0, \quad n \geq 0, \quad j = 1, \dots, M$$

con la esperanza de que $v_j^n \sim u_j^n = u(t^n, x_j)$ e imponiendo los datos conocidos por el dato inicial $v_j^0 = u_j^0$ y por los datos de contorno $v_0^n = u_0^n$ junto a $v_{M+1}^n = u_{M+1}^n$.

Aquí se puede ver que lo único que se conoce es el valor v_j^n siendo el resto datos desconocidos. Por lo tanto, es un **cálculo implícito** puesto que cada incógnita v_j^{n+1} involucra las incógnitas vecinas v_{j+1}^{n+1} y v_{j-1}^{n+1} . Cada ecuación se describe entonces como

$$-\mu v_{j+1}^{n+1} + (1 + 2\mu) v_j^{n+1} - \mu v_{j-1}^{n+1} = v_j^n \quad n \geq 0, \quad j = 1, \dots, M$$

en donde aparece otra vez el **número de Courant**

$$\mu = \kappa \frac{\Delta t}{\Delta x^2}.$$

Observación 138 Este número se puede interpretar como la relación de las tallas de las particiones en las variables temporales y espaciales pero teniendo en cuenta el orden de las derivadas que se aproximan, por eso sale Δt y Δx^2 , junto con el coeficiente de difusión que es el que hace que la solución continua se disipe más o menos despacio por el dominio computacional.

Tenemos un sistema de ecuaciones lineales que, teniendo en cuenta que los valores en los extremos son nulos para simplificar, se escribe como

$$\begin{aligned} -\mu v_2^{n+1} + (1 + 2\mu) v_1^{n+1} &= v_1^n, \\ -\mu v_3^{n+1} + (1 + 2\mu) v_2^{n+1} - \mu v_1^{n+1} &= v_2^n, \\ -\mu v_4^{n+1} + (1 + 2\mu) v_3^{n+1} - \mu v_2^{n+1} &= v_3^n, \\ &\dots = \dots, \quad (j = 4, \dots, M-1) \\ (1 + 2\mu) v_M^{n+1} - \mu v_{M-1}^{n+1} &= v_M^n, \end{aligned}$$

Luego en cada paso hay que resolver un sistema lineal donde la matriz es tridiagonal

$$PV^{n+1} = V^n, \quad n \geq 0$$

siendo $P = \text{tridiag}(-\mu, 1 + 2\mu, -\mu)$.

Observación 139 En el caso no homogéneo se resuelve un sistema lineal en la forma

$$PV^{n+1} = V^n + b^{n+1}$$

donde b^{n+1} es el vector $b^{n+1} = \mu(v_0^{n+1}, 0, \dots, 0, v_{M+1}^{n+1})' \in \mathbb{R}^M$

Por lo tanto, nuestro esquema de cálculo queda como:

Dados $v_j^0 = u_j^0$ para $j = 1, \dots, M$ y con $v_0^n = 0$, $v_{M+1}^n = 0$ se resuelve el sistema lineal

$$PV^{n+1} = V^n, \quad n \geq 0.$$

y es un método basado en Euler implícito en tiempo y diferencias centradas en espacio.

9.5. Semidiscretización: Método de líneas

Una forma útil de mirar la discretización consiste en escribir

$$u_t = \mathcal{L}(u)$$

donde $\mathcal{L}(u)$ es cualquier operador diferencial. Por ejemplo, puede ser

$$\mathcal{L}(u) = \kappa u_{xx}(t, x), \quad \mathcal{L}(u) = \kappa u_{xx}(t, x) + f(t, x)$$

en el caso homogéneo ($f = 0$) o en el caso no homogéneo ($f \neq 0$). Entonces podemos discretizar primero en tiempo y luego en espacio:

- *Euler explícito*

$$\frac{u^{n+1} - u^n}{\Delta t} = \mathcal{L}(u^n)$$

- *Euler implícito*

$$\frac{u^{n+1} - u^n}{\Delta t} = \mathcal{L}(u^{n+1})$$

- *Crank-Nicolson*

$$\frac{u^{n+1} - u^n}{\Delta t} = \frac{1}{2}[\mathcal{L}(u^n) + \mathcal{L}(u^{n+1})]$$

Cualquier otra discretización temporal también puede servir y ahora sólo hay que discretizar en espacio los valores $\mathcal{L}(u^n)$.

De forma equivalente, también podemos discretizar en espacio primero y luego en tiempo, lo que se suele denominar el método de líneas ya que se sigue la ecuación a lo largo de la línea dada por $\{x = x_j\}$ para cada j : la ecuación se describe en cada punto interior $\{x = x_j\}$ (para poder usar los vecinos) como

$$\frac{d}{dt}u(t, x_j) = \kappa \frac{u(t, x_{j+1}) - 2u(t, x_j) + u(t, x_{j-1})}{\Delta x^2} + O(\Delta x^2), \quad j = 1, 2, 3, \dots, M.$$

Si eliminamos el término $O(\Delta x^2)$ tenemos entonces el sistema de ecuaciones diferenciales ordinarias

$$\frac{d}{dt}v_j(t) = \kappa \frac{v_{j+1}(t) - 2v_j(t) + v_{j-1}(t)}{\Delta x^2}, \quad j = 1, 2, 3, \dots, M.$$

donde $v_j(t)$ aproxima al verdadero valor $u(t, x_j)$. Entonces, poniendo $V(t) \in \mathbb{R}^M$ dado por $V(t) = (v_1(t), v_2(t), \dots, v_M(t))'$ podemos escribir la semidiscretización como un sistema de ecuaciones diferenciales lineales en la forma

$$\frac{d}{dt}V(t) = -\frac{\kappa}{\Delta x^2}A V(t) + \frac{\kappa}{\Delta x^2}b(t)$$

donde $A = tridiag(-1, 2, -1) \in \mathbb{R}^{M \times M}$ y $b(t) \in \mathbb{R}^M$ proviene de los datos de contorno.

Observación 140 *De una manera intuitiva, se puede decir también que la segunda derivada u_{xx} como función se approxima por el vector $-\frac{1}{\Delta x^2}A \cdot U$ en dimensión finita. Esto es*

$$u_{xx} \approx \frac{1}{\Delta x^2}tridiag(1, -2, 1) \cdot U \in \mathbb{R}^M$$

y que

$$\frac{1}{\Delta x^2}tridiag(1, -2, 1) \cdot U \rightarrow u_{xx}, \quad M \rightarrow +\infty$$

Observación 141 En el caso de tener un término de fuerza y datos de contorno no nulos tendremos la expresión general

$$\frac{d}{dt}V(t) = -\frac{\kappa}{\Delta x^2}AV(t) + \frac{\kappa}{\Delta x^2}b(t) + f(t)$$

para $f(t) = (f_1(t), f_2(t), \dots, f_M(t))'$.

Ejercicio 165 La matriz $A = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{M \times M}$ es simétrica definida positiva y con autovalores positivos dados por

$$\alpha_j = 4 \sin \left(\frac{j\pi}{2(M+1)} \right)^2, \quad j = 1, 2, \dots, M.$$

Como para $j = 1, 2, \dots, M$ los valores $\frac{j}{M+1} \frac{\pi}{2}$ forman una partición de $(0, \pi/2)$ es claro que $\alpha_j \in (0, 4)$. Cuanto más grande es M más cerca de cero se encuentra $\alpha_1 = 4 \sin(\frac{\pi}{2(M+1)})^2$ por lo que la matriz se acerca a ser singular.

Observación 142 Los autovalores λ_j de $A_{\Delta x} = \frac{1}{\Delta x^2} \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{M \times M}$ son $\lambda_j = \alpha_j \Delta x^{-2}$ y cumplen para $M \gg 1$

$$\begin{aligned} \lambda_j &\approx j^2 \pi^2, \quad 1 \leq j \ll M, \\ \lambda_j &\approx 4 \Delta x^{-2} = 4(M+1)^2, \quad 1 \ll j \leq M. \end{aligned}$$

Para ello sólo hay que tener en cuenta que, usando Taylor,

$$\alpha_j = 4 \sin \left(\frac{j\pi}{2(M+1)} \right)^2 \sim 4 \left(\frac{j}{M+1} \frac{\pi}{2} \right)^2 = 4 \left(\Delta x j \frac{\pi}{2} \right)^2 = \Delta x^2 j^2 \pi^2, \quad 1 \leq j \ll M$$

mientras que si $1 \ll j \leq M$ entonces

$$\alpha_j = 4 \sin \left(\frac{j\pi}{2(M+1)} \right)^2 \sim 4 \quad 1 \ll j \leq M$$

ya que $\Delta x = (M+1)^{-1}$. Como $A_{\Delta x}$ es simétrica, usando los autovalores máximo y mínimo tenemos

$$\|A_{\Delta x}\|_2 = \rho(A_{\Delta x}) = \lambda_{\max} \sim M^2, \quad \|A_{\Delta x}^{-1}\|_2 = \rho(A_{\Delta x}^{-1}) = \lambda_{\min} \sim \frac{1}{\pi^2}.$$

Como consecuencia de esto, el condicionamiento de la matriz empeora al disminuir Δx puesto que

$$\text{cond}_2(A_{\Delta x}) = \frac{\lambda_{\max}}{\lambda_{\min}} \sim \frac{M^2}{\pi^2} \sim \frac{1}{\Delta x^2}, \quad \text{si } \Delta x \ll 1$$

Una vez introducida una discretización en espacio, los métodos vistos hasta ahora se pueden describir con facilidad usando una discretización en tiempo. Ponemos ahora $v_j^n \sim v(t^n, x_j)$ y tenemos:

- La aplicación del método de Euler explícito a la variable temporal genera:

$$\frac{V^{n+1} - V^n}{\Delta t} = -\frac{\kappa}{\Delta x^2} A V^n + \frac{\kappa}{\Delta x^2} b^n + f^n$$

o lo que es lo mismo

$$V^{n+1} = (I_M - \mu A_M) V^n + \mu b^n + \Delta t f^n$$

donde $I_M - \mu A_M = \text{tridiag}(\mu, 1 - 2\mu, \mu)$.

- La aplicación del método de Euler implícito a la variable temporal genera:

$$\frac{V^{n+1} - V^n}{\Delta t} = -\frac{\nu}{\Delta x^2} A V^{n+1} + \frac{\nu}{\Delta x^2} b^{n+1} + f^{n+1}$$

o lo que es lo mismo

$$(I_M + \mu A_M) V^{n+1} = V^n + \mu b^{n+1} + \Delta t f^{n+1}$$

donde $I_M + \mu A_M = \text{tridiag}(-\mu, 1 + 2\mu, -\mu)$ y se resuelve el sistema lineal, esto es

$$V^{n+1} = (I_M + \mu A_M)^{-1}(V^n + \mu b^{n+1} + \Delta t f^{n+1})$$

(la inversa no se calcula por ser un proceso numéricamente inestable).

- Dando un paso más preciso podemos aplicar Crank-Nicolson a la variable temporal obteniendo:

$$\begin{aligned} \frac{V^{n+1} - V^n}{\Delta t} &= \frac{1}{2} \left\{ -\frac{\kappa}{\Delta x^2} A V^{n+1} + \frac{\kappa}{\Delta x^2} b^{n+1} + f^{n+1} \right\} \\ &+ \frac{1}{2} \left\{ -\frac{\kappa}{\Delta x^2} A V^n + \frac{\kappa}{\Delta x^2} b^n + f^n \right\} \end{aligned}$$

o lo que es lo mismo

$$(I_M + \frac{1}{2}\mu A_M) V^{n+1} = (I_M - \frac{1}{2}\mu A_M) V^n + \frac{1}{2}\mu \{ b^{n+1} + b^n \} + \frac{1}{2}\Delta t \{ f^{n+1} + f^n \}$$

y se resuelve el sistema lineal

$$V^{n+1} = (I_M + \frac{1}{2}\mu A_M)^{-1} [(I_M - \frac{1}{2}\mu A_M) V^n + \frac{1}{2}\mu \{ b^{n+1} + b^n \} + \frac{1}{2}\Delta t \{ f^{n+1} + f^n \}]$$

En el caso más simple de datos de contorno cero y $f = 0$, entonces entonces nos queda una iteración limpia en cada caso en la forma

$$V^{n+1} = P V^n, \quad n \geq 0$$

o bien

$$V^n = P^n V^0, \quad n \geq 0$$

donde la matriz P es:

- para Euler explícito:

$$P = P_{Ex} = I_M - \mu A_M$$

es decir $I_M - \mu A_M = \text{tridiag}(\mu, 1 - 2\mu, \mu)$.

- para Euler implícito:

$$P = P_{Im} = (I_M + \mu A_M)^{-1}$$

donde $I_M + \mu A_M = \text{tridiag}(-\mu, 1 + 2\mu, -\mu)$.

- para Crank-Nicolson:

$$P = P_{CN} = (I_M + \frac{1}{2}\mu A_M)^{-1}(I_M - \frac{1}{2}\mu A_M)$$

donde, obviamente,

$$I_M + \frac{1}{2}\mu A_M = \text{tridiag}(-\frac{1}{2}\mu, 1 + \mu, -\frac{1}{2}\mu), \quad I_M - \frac{1}{2}\mu A_M = \text{tridiag}(\frac{1}{2}\mu, 1 - \mu, \frac{1}{2}\mu).$$

9.6. Estudio unificado de la estabilidad

Vamos a fijarnos en el caso donde los datos de contorno son cero y no hay fuerza f , es decir, en el caso homogéneo. Usando la norma $\|\cdot\|_{\Delta x}$ vamos a ver la estabilidad, consistencia y convergencia de los métodos introducidos en esta norma. Tomamos P donde P es cualquiera de las matrices P_{Ex} , P_{Im} o P_{CN} .

Dados $V^0 \in \mathbb{R}^M$ y $W^0 = V^0 + \epsilon \in \mathbb{R}^M$ (se sobreentiende $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_M)^T \in \mathbb{R}^M$) tenemos que comparar los cálculos

$$V^{n+1} = P V^n = P^{n+1} V^0$$

con

$$W^{n+1} = P W^n = P^{n+1} W^0$$

de donde por linealidad

$$V^{n+1} - W^{n+1} = P^{n+1} \epsilon, \quad n \geq 0.$$

Entonces

$$\|V^n - W^n\|_{\Delta x} \leq \|P^n\|_{\Delta x} \epsilon$$

donde ahora $\epsilon := \|\epsilon\|_{\Delta x}$ abusando de la notación. Para garantizar estabilidad debemos tener

$$\|P^n\|_{\Delta x} \leq C, \forall n \geq 0$$

puesto que entonces

$$\|V^n\|_{\Delta x} \leq \|P^n\|_{\Delta x} \|V^0\|_{\Delta x} \leq C \|V^0\|_{\Delta x}, \quad n \geq 0,$$

es decir, que todas las potencias de P deben estar acotadas en la norma matricial inducida por $\|\cdot\|_{\Delta x}$ que sabemos coincide con $\|\cdot\|_2$, luego el estudio de los autovalores es fundamental. Pero los autovalores de cada matriz son conocidos:

- Para Euler explícito Eex-CE la matriz de iteración es

$$P_{Ex} = I_M - \mu A_M.$$

donde $I_M - \mu A_M = \text{tridiag}(\mu, 1-2\mu, \mu)$. Los autovalores de $P_{Ex} = I_M - \mu A_M$ son

$$\sigma(I_M - \mu A_M) = \{1 - \mu \alpha_j\}_{j=1,\dots,M}$$

donde $\{\alpha_j\}_{j=1,\dots,M}$ son los autovalores de A_M . Sabemos que los autovalores de A_M cumplen $0 < \alpha_j < 4$, por lo tanto, si $\mu \leq 1/2$ se cumple

$$-1 \leq 1 - \mu \alpha_j \leq 1$$

y por lo tanto, imponiendo $\mu \leq 1/2$, tenemos

$$\|P_{Ex}^n\|_{\Delta x} = \|P_{Ex}^n\|_2 = \max_j |1 - \mu \alpha_j|^n \leq 1, \quad \forall n \geq 0$$

con lo que vemos una **estabilidad con condiciones sobre μ** .

- En el caso de Euler implícito Eim-CE tenemos

$$P_{Im} = (I_M + \mu A_M)^{-1}$$

donde $I_M + \mu A_M = \text{tridiag}(-\mu, 1+2\mu, -\mu)$. Pero los autovalores de $I_M + \mu A_M$ son ahora

$$\sigma(I_M + \mu A_M) = \{1 + \mu \alpha_j\}_{j=1,\dots,M},$$

que son todos positivos y mayores que 1, y los de $(I_M + \mu A_M)^{-1}$ son

$$\sigma((I_M + \mu A_M)^{-1}) = \left\{ \frac{1}{1 + \mu \alpha_j} \right\}_{j=1,\dots,M}$$

y por lo tanto, sin ningún tipo de restricción sobre μ ,

$$\|P_{Im}^n\|_{\Delta x} = \|P_{Im}^n\|_2 = \max_j \frac{1}{|1 + \mu \alpha_j|^n} < 1, \quad \forall n \geq 0$$

con lo que vemos una **estabilidad sin condiciones sobre μ** .

- En el caso de Crank-Nicolson CN-CE:

$$P_{CN} = (I_M + \frac{1}{2}\mu A_M)^{-1}(I_M - \frac{1}{2}\mu A_M)$$

donde

$$I_M + \frac{1}{2}\mu A_M = tridiag(-\frac{1}{2}\mu, 1+\mu, -\frac{1}{2}\mu), \quad I_M - \frac{1}{2}\mu A_M = tridiag(\frac{1}{2}\mu, 1-\mu, \frac{1}{2}\mu).$$

Entonces los autovalores son

$$\lambda \in \sigma(P_{CN}) \Leftrightarrow \lambda = \frac{1 - 0.5\mu\alpha_j}{1 + 0.5\mu\alpha_j}$$

y evidentemente se cumple para cualquier valor de μ

$$-1 < \frac{1 - 0.5\mu\alpha_j}{1 + 0.5\mu\alpha_j} < 1.$$

luego también aquí tenemos estabilidad al ser

$$\|P_{CN}^n\|_{\Delta x} = \|P_{CN}^n\|_2 = \max_j \left| \frac{1 - 0.5\mu\alpha_j}{1 + 0.5\mu\alpha_j} \right|^n < 1, \quad \forall n \geq 0$$

y es una estabilidad sin condiciones sobre μ .

Observación 143 Uso de la serie de Fourier Si tomamos $x \in (0, 1)$ (por fijar ideas) y usamos desarrollo de Fourier en la variable espacial, entonces

$$u(t, x) = \sum_m c_m(t) e^{2\pi m x i}$$

donde $c_m(t)$ denota la amplitud de la onda de frecuencia m dada por $e^{2\pi m x i}$ (recordar <https://www.falstad.com/fourier/> o cualquier otra applet ilustrativa de la representación de una función mediante desarrollos de Fourier). Sabemos que son necesarias infinitas frecuencias, es decir, la serie infinita entera, para representar cualquier falta de regularidad.

Permutando la derivación con la suma infinita podemos interpretar el comportamiento de la solución para operadores diferenciales lineales:

- **Ecuación de difusión:** Si tenemos $u_t - ku_{xx} = 0$ entonces $u_t = ku_{xx}$ y llegamos a

$$c'_m(t) = -k(2\pi m)^2 c_m(t)$$

para la amplitud de cada modo m . Luego vemos que las amplitudes de cada frecuencia decaen de forma exponencial y más rápido cuanto mayor es m :

$$c_m(t) = e^{-k(2\pi m)^2 t} c_m(0), \quad t > 0$$

con lo que hay un efecto de difusión más fuerte cuanto más alta es la frecuencia. Esto implica que la regularidad mejora con el tiempo, esto es el **efecto regularizante del operador de difusión** $\partial_t - k\partial_{xx}$.

- **Ecuación de transporte:** Si tenemos $u_t - au_x = 0$ entonces $u_t = au_x$ y llegamos a

$$c'_m(t) = a(2\pi m i)c_m(t)$$

para la amplitud de cada modo m . Luego vemos que las amplitudes de cada frecuencia mantienen módulo constante para cada frecuencia:

$$c_m(t) = e^{a(2\pi m i)t} c_m(0) \Rightarrow |c_m(t)| = |c_m(0)|, \quad t > 0.$$

Con lo que vemos que se mantienen en tiempo las irregularidades que existiesen originalmente. Esta es la esencia del **operador de transporte** $\partial_t + a\partial_x$: la señal inicial se transporta manteniendo su regularidad.

Observación 144 Como consecuencia de lo anterior, los métodos implícitos para edos son los más naturales para discretizar en tiempo los problemas de difusión ya que la amplitud de cada modo decae exponencialmente.

Pero el uso del esquema numérico de Crank-Nicolson para datos irregulares, o discontinuos, no está aconsejado puesto que amortigua con dificultad las altas frecuencias y no regulariza bien el dato inicial, cosa que debe ser natural si el esquema numérico funciona bien.

El esquema numérico Crank-Nicolson genera los valores numéricos

$$c_m^n = \lambda^n c_m(0), \quad (c_m^n \approx c_m(t_n))$$

donde

$$\lambda = \frac{1 + 0.5k(2\pi m)^2}{1 - 0.5k(2\pi m)^2}$$

La existencia de discontinuidades implica la existencia de valores $m \gg 1$ ya que deben de existir modos altos o vibraciones altas $c_m(t)e^{2\pi m x_i}$ para capturar las discontinuidades. Pero esto implica que $\lambda \approx 1$ por lo que el método numérico no disminuye el dato inicial con rapidez. Pero sabemos que la solución exacta sí que lo hace, luego este esquema numérico no es aconsejable.

Por otro lado, Euler implícito sí que funciona bien, aunque sea de primer orden en tiempo, ya que genera los valores numéricos

$$c_m^n = \lambda^n c_m(0), \quad (c_m^n \approx c_m(t_n))$$

donde

$$\lambda = \frac{1}{1 + k (2\pi m)^2}$$

y hay decaimiento garantizado de los modos altos.

9.7. Estudio de la consistencia

El error de truncatura local es el error que se comete al reemplazar la ecuación diferencial por el cociente incremental. Vamos a recordar como se define en el caso de los métodos numéricos aplicados a las soluciones del problema $y'(t) = f(t, y(t))$.

1. En el caso del método de Euler explícito es

$$\tau_{\Delta t}(t) = \frac{y(t + \Delta t) - y(t)}{\Delta t} - f(t, y(t))$$

de donde Taylor en $(t, y(t))$ lleva a

$$\tau_{\Delta t}(t) = O(\Delta t).$$

2. En el caso del método de Euler implícito es

$$\tau_{\Delta t}(t) = \frac{y(t + \Delta t) - y(t)}{\Delta t} - f(t + \Delta t, y(t + \Delta t))$$

de donde Taylor en $(t, y(t))$ lleva a

$$\tau_{\Delta t}(t) = O(\Delta t).$$

3. En el caso del método de Crank-Nicolson es

$$\tau_{\Delta t}(t) = \frac{y(t + \Delta t) - y(t)}{\Delta t} - \frac{1}{2}[f(t, y(t)) + f(t + \Delta t, y(t + \Delta t))]$$

de donde Taylor en $(t, y(t))$ lleva a

$$\tau_{\Delta t}(t) = O(\Delta t^2).$$

Extendemos entonces la idea a nuestro caso donde ahora discretizamos derivadas en variables t y x .

9.7.1. Consistencia para Euler explícito

Definición 166 Si $u(t, x)$ es la solución de la ecuación diferencial y se cumple

$$u_t(t, x) - \kappa u_{xx}(t, x) = f(t, x)$$

entonces se define el **error de truncatura local** como

$$\tau_{\Delta t, \Delta x}(x, t) = \frac{u(t + \Delta t, x) - u(t, x)}{\Delta t} - \kappa \frac{u(t, x + \Delta x) - 2u(t, x) + u(t, x - \Delta x)}{\Delta x^2} - f(t, x)$$

Usando que $u_t - \kappa u_{xx} = f$ y los desarrollos de Taylor sabemos que

$$\tau_{\Delta t, \Delta x}(x, t) = O(\Delta t) + O(\Delta x^2).$$

Por lo tanto,

$$\tau_{\Delta t, \Delta x} = \max_{(t, x)} |\tau_{\Delta t, \Delta x}(x, t)| = O(\Delta t) + O(\Delta x^2).$$

donde las constantes involucradas contienen los máximos en las derivas u_t y u_{xxxx} en algunos puntos intermedios.

Definición 167 Consistencia de un esquema: Un método se dice consistente si su error de truncatura máximo $\tau_{\Delta t, \Delta x}$ tiende a cero cuando los parámetros de discretización así lo hacen.

Por lo tanto, el método de Euler explícito es consistente con orden 1 con respecto a Δt y orden 2 con respecto a Δx .

9.7.2. Consistencia para Euler implícito

Definición 168 Si $u(t, x)$ es la solución de la ecuación diferencial y se cumple

$$u_t(t, x) - \kappa u_{xx}(t, x) = f(t, x)$$

entonces se define el **error de truncatura local** para Euler implícito en tiempo y centrado en espacio Eim-CE como

$$\begin{aligned} \tau_{\Delta t, \Delta x}(t, x) &= \frac{u(t + \Delta t, x) - u(t, x)}{\Delta t} \\ &- \kappa \frac{u(t + \Delta t, x + \Delta x) - 2u(t + \Delta t, x) + u(t + \Delta t, x - \Delta x)}{\Delta x^2} \\ &- f(t + \Delta t, x). \end{aligned}$$

Usando que $u_t(t, x) - \kappa u_{xx}(t, x) = f(t, x)$ tenemos que

$$\tau_{\Delta t, \Delta x} = \max_{(t, x)} |\tau_{\Delta t, \Delta x}(t, x)| = O(\Delta t) + O(\Delta x^2)$$

Por lo tanto, el método es de orden 1 con respecto a Δt y de orden 2 con respecto a Δx .

9.7.3. Consistencia para Crank-Nicolson

Veamos como obtenemos ganancia de un orden en tiempo usando Crank-Nicolson

Definición 169 Si $u(t, x)$ es la solución de la ecuación diferencial y se cumple

$$u_t(t, x) - \kappa u_{xx}(t, x) = f(t, x)$$

entonces el **error de truncatura local** para Crank-Nicolson es

$$\begin{aligned} \tau_{\Delta t, \Delta x}(t, x) &= \frac{u(t + \Delta t, x) - u(t, x)}{\Delta t} \\ &- \frac{1}{2} \kappa \frac{u(t, x + \Delta x) - 2u(t, x) + u(t, x - \Delta x)}{\Delta x^2} \\ &- \frac{1}{2} \kappa \frac{u(t + \Delta t, x + \Delta x) - 2u(t + \Delta t, x) + u(t + \Delta t, x - \Delta x)}{\Delta x^2} \\ &- \frac{1}{2} [f(t, x) + f(t + \Delta t, x)]. \end{aligned}$$

Teorema 170 El error de truncatura del método de Crank-Nicolson es de orden 2 en espacio y 2 en tiempo

$$\tau_{\Delta t, \Delta x}(t, x) = O(\Delta t^2) + O(\Delta x^2)$$

Dem: Usando desarrollos de Taylor en (t, x) tenemos

$$\frac{u(t + \Delta t, x) - u(t, x)}{\Delta t} = u_t(t, x) + \frac{\Delta t}{2} u_{tt}(t, x) + \frac{\Delta t^2}{6} u_{ttt}(t, x) + \dots$$

por otro lado

$$\begin{aligned} &\frac{u(t + \Delta t, x + \Delta x) - 2u(t + \Delta t, x) + u(t + \Delta t, x - \Delta x)}{\Delta x^2} \\ &= u_{xx}(t + \Delta t, x) + \frac{\Delta x^2}{12} u_{xxxx}(t + \Delta t, x) + \dots \end{aligned}$$

pero además, desarrollando con respecto a (t, x) los valores en $(t + \Delta t, x)$, tenemos que

$$u_{xx}(t + \Delta t, x) = u_{xx}(t, x) + u_{xxt}(t, x)\Delta t + \frac{\Delta t^2}{2} u_{xxtt}(t, x) + \dots$$

luego

$$\begin{aligned} &\frac{u(t + \Delta t, x + \Delta x) - 2u(t + \Delta t, x) + u(t + \Delta t, x - \Delta x)}{\Delta x^2} = \\ &= u_{xx}(t, x) + u_{xxt}(t, x)\Delta t + \frac{\Delta t^2}{2} u_{xxtt}(t, x) + \dots + \frac{\Delta x^2}{12} u_{xxxx}(t + \Delta t, x) + \dots \end{aligned}$$

y también

$$\frac{u(t, x + \Delta x) - 2u(t, x) + u(t, x - \Delta x)}{\Delta x^2} = u_{xx}(t, x) + \frac{\Delta x^2}{12}u_{xxxx}(t, x) + \dots$$

de donde sumando ambos desarrollos

$$\begin{aligned} & \frac{u(t, x + \Delta x) - 2u(t, x) + u(t, x - \Delta x)}{\Delta x^2} \\ & + \frac{u(t + \Delta t, x + \Delta x) - 2u(t + \Delta t, x) + u(t + \Delta t, x - \Delta x)}{\Delta x^2} \\ & = u_{xx}(t, x) + u_{xxt}(t, x)\Delta t + \frac{\Delta t^2}{2}u_{xxtt}(t, x) + \dots \\ & + \frac{\Delta x^2}{12}u_{xxxx}(t + \Delta t, x) + \dots + u_{xx}(t, x) + \frac{\Delta x^2}{12}u_{xxxx}(t, x) + \dots \\ & = 2u_{xx}(t, x) + u_{xxt}(t, x)\Delta t + \frac{\Delta t^2}{2}u_{xxtt}(t, x) + \dots \\ & + \frac{\Delta x^2}{12}u_{xxxx}(t + \Delta t, x) + \frac{\Delta x^2}{12}u_{xxxx}(t, x) + \dots \\ & = 2u_{xx}(t, x) + u_{xxt}(t, x)\Delta t + O(\Delta t^2) + O(\Delta x^2). \end{aligned}$$

por lo tanto,

$$\begin{aligned} \tau_{\Delta t, \Delta x}(x, t) &= u_t(t, x) + \frac{\Delta t}{2}u_{tt}(t, x) + O(\Delta t^2) \\ &- \kappa u_{xx}(t, x) - \frac{\Delta t}{2}\kappa u_{xxt}(t, x) + O(\Delta t^2) + O(\Delta x^2) \\ &- \frac{1}{2}[f(t, x) + f(t + \Delta t, x)]. \end{aligned}$$

Pero como

$$f(t + \Delta t, x) = f(t, x) + \frac{\Delta t}{1!}\partial_t f(t, x) + \frac{\Delta t^2}{2!}\partial_{tt} f(t, x) + \dots$$

entonces

$$\frac{1}{2}[f(t, x) + f(t + \Delta t, x)] = f(t, x) + \frac{1}{2}\frac{\Delta t}{1!}\partial_t f(t, x) + \frac{1}{2}\frac{\Delta t^2}{2!}\partial_{tt} f(t, x) + \dots$$

Usando que

$$u_t(t, x) - \kappa u_{xx}(t, x) = f(t, x), \quad u_{tt}(t, x) - \kappa u_{xxt}(t, x) = \partial_t f(t, x)$$

y nos queda

$$\begin{aligned}\tau_{\Delta t, \Delta x}(x, t) &= [u_t(t, x) - \kappa u_{xx}(t, x) - f(t, x)] \\ &+ \frac{\Delta t}{2} [u_{tt}(t, x) - \kappa u_{xxt}(t, x) - \partial_t f(t, x)] + O(\Delta t^2) + O(\Delta x^2) \\ &= O(\Delta t^2) + O(\Delta x^2).\end{aligned}$$

■

Aprovechando también la idea dada por el método de líneas podemos escribir el error de truncatura como vector $\tau_{\Delta t, \Delta x}(t^n) = (\tau_{\Delta t, \Delta x}(t^n, x_1), \dots, \tau_{\Delta t, \Delta x}(t^n, x_M))'$ y entonces, usando ahora U^n los valores exactos, escribir los esquemas en forma vectorial

- Para Euler explícito (Eex-CE):

$$U^{n+1} = (I_M - \mu A_M) U^n + \Delta t \tau_{\Delta t, \Delta x}(t^n)$$

- Para Euler implícito (Eim-CE):

$$U^{n+1} = (I_M + \mu A_M)^{-1} U^n + \Delta t (I_M + \mu A_M)^{-1} \tau_{\Delta t, \Delta x}(t^n)$$

o bien

$$U^{n+1} = (I_M + \mu A_M)^{-1} U^n + \Delta t \tilde{\tau}_{\Delta t, \Delta x}(t^n)$$

donde $\tilde{\tau}_{\Delta t, \Delta x}(t^n)$ posee el mismo orden que $\tau_{\Delta t, \Delta x}(t^n)$ con respecto a Δt y a Δx al estar $\|(I_M + \mu A_M)^{-1}\|_{\Delta x}$ acotada.

- Para Crank-Nicolson (CN-CE):

$$U^{n+1} = (I_M + 0.5\mu A_M)^{-1} (I_M - 0.5\mu A_M) U^n + \Delta t \tilde{\tau}_{\Delta t, \Delta x}$$

donde $\tilde{\tau}_{\Delta t, \Delta x}$ posee el mismo orden que $\tau_{\Delta t, \Delta x}(t^n)$ con respecto a Δt y a Δx al estar $\|(I_M + 0.5\mu A_M)^{-1}\|_{\Delta x}$ acotada.

En todos estos casos entendemos $\tilde{\tau}_{\Delta t, \Delta x}(t^n)$ como un vector de M componentes $\tilde{\tau}_{\Delta t, \Delta x}(t^n, x_j)$ cada una de ellas del mismo orden que el error local en un punto (t_n, x_j) . Entonces, si por ejemplo tenemos una truncatura local

$$\tilde{\tau}_{\Delta t, \Delta x}(t^n, x_j) = O(\Delta t) + O(\Delta x^2)$$

esta se mantiene al tomar la norma $\|\cdot\|_{\Delta x}$, ya que las constantes que aparecen son valores de las derivadas de la solución exacta y podemos suponer que están acotadas uniformemente. Entonces

$$\begin{aligned}\|\tilde{\tau}_{\Delta t, \Delta x}(t^n)\|_{\Delta x} &= (\Delta x \sum_{j=1}^M \tilde{\tau}_{\Delta t, \Delta x}(t^n, x_j)^2)^{1/2} \\ &= (\Delta x M [O(\Delta t) + O(\Delta x^2)]^2)^{1/2} = O(\Delta t) + O(\Delta x^2)\end{aligned}$$

9.8. Estudio unificado de la convergencia

Los esquemas numéricos se pueden escribir en el caso homogéneo como

$$V^{n+1} = PV^n \in \mathbb{R}^M, \quad n \geq 0$$

mientras que para la solución exacta $U^n = (u(t^n, x_1), \dots, u(t^n, x_M))^T$, suponiendo que tenemos toda la regularidad necesaria para las derivadas que aparecen y usando el error de truncatura podemos escribir

$$U^{n+1} = PU^n + \Delta t \tilde{\tau}_{\Delta t, \Delta x}, \quad n \geq 0$$

donde sabemos que para Euler explícito

$$\tilde{\tau}_{\Delta t, \Delta x} = O(\Delta t) + O(\Delta x^2) \in \mathbb{R}^M,$$

para Euler implícito

$$\tilde{\tau}_{\Delta t, \Delta x} = O(\Delta t) + O(\Delta x^2) \in \mathbb{R}^M$$

y para Crank-Nicolson

$$\tilde{\tau}_{\Delta t, \Delta x} = O(\Delta t^2) + O(\Delta x^2) \in \mathbb{R}^M.$$

Por la linealidad del esquema, si $E^n = V^n - U^n$ entonces

$$E^{n+1} = PE^n + \Delta t \tilde{\tau}_{\Delta t, \Delta x}, \quad n \geq 0.$$

Iterando (I_M es la matriz identidad de talla M)

$$\begin{aligned} E^1 &= PE^0 + \Delta t \tilde{\tau}_{\Delta t, \Delta x}, \\ E^2 &= PE^1 + \Delta t \tilde{\tau}_{\Delta t, \Delta x} = P^2 E^0 + \{P + I_M\} \Delta t \tilde{\tau}_{\Delta t, \Delta x} \\ E^3 &= PE^2 + \Delta t \tilde{\tau}_{\Delta t, \Delta x} = P^3 E^0 + \{P^2 + P + I_M\} \Delta t \tilde{\tau}_{\Delta t, \Delta x} \end{aligned}$$

de donde se observa fácilmente la recurrencia

$$E^n = P^n E^0 + \{P^{n-1} + \dots + P^2 + P + I_M\} \Delta t \tilde{\tau}_{\Delta t, \Delta x}.$$

Como ya hemos visto, si $\mu \leq 1/2$ para Euler explícito o sin ninguna restricción sobre μ para Euler implícito o Crank-Nicolson, se cumple $\|P^n\|_{\Delta x} \leq 1$ para todo $n \geq 0$ y las matrices que multiplican a los vectores de truncatura $\tilde{\tau}_{\Delta t, \Delta x}$ también están acotadas en norma $\|\cdot\|_2 \leq 1$. Por lo tanto,

$$\|E^n\|_{\Delta x} \leq \|E^0\|_{\Delta x} + n \Delta t \|\tilde{\tau}_{\Delta t, \Delta x}\|_{\Delta x}.$$

de donde usando que $n \Delta t \leq T$ y metiendo el tiempo máximo T dentro del error de truncatura global, lo que no cambia el orden, obtenemos

$$\|E^n\|_{\Delta x} \leq \|E^0\|_{\Delta x} + O(\Delta t^q) + O(\Delta x^2), \quad \forall n \geq 0$$

donde $q = 1$ o $q = 2$ dependiendo de si usamos Euler o Crank-Nicolson.

Esto quiere decir que

$$(\Delta x \sum_{j=1}^M |U^{n+1}(x_j) - V_j^{n+1}|^2)^{1/2} \leq (\Delta x \sum_{j=1}^M |U^0(x_j) - V_j^0|^2)^{1/2} + O(\Delta t^q) + O(\Delta x^2).$$

Esto nos da la velocidad de decaimiento del error $O(\Delta t^q) + O(\Delta x^2)$, siempre y cuando el error inicial la respete, es decir, el error inicial decaiga al menos como $O(\Delta x^2)$. Por otro lado, tenemos la convergencia a cero del error en la norma L^2 en espacio y L^∞ en tiempo

$$\max_t \int_0^1 |u(t, x) - V^n|^2 \rightarrow 0$$

donde V^n representa una extensión a todo el intervalo de los valores V_j^n .

Recopilando:

Teorema 171 Para $\mu \leq 1/2$ el método de Euler explícito en tiempo y diferencias centradas en espacio (Eex-CE) es convergente con orden 1 con respecto al parámetro temporal Δt y orden 2 con respecto al parámetro espacial Δx (la restricción $\mu \leq 1/2$ obliga a tener $\Delta t \leq \frac{1}{2\kappa} \Delta x^2$) (el recíproco también es cierto, aunque no lo vemos, es decir, si hay convergencia entonces debe de ser $\mu \leq 1/2$).

Teorema 172 Sin restricción sobre μ , el método de Euler implícito en tiempo y diferencias centradas en espacio (Eim-CE) es convergente con orden 1 con respecto al parámetro temporal Δt y orden 2 con respecto al parámetro espacial Δx .

Tener ordenes distintos no es la situación más conveniente y Crank-Nicolson aporta la solución.

Teorema 173 Sin restricción sobre μ , el método de Crank-Nicolson en tiempo y diferencias centradas en espacio (CN-CE) es convergente con orden 2 con respecto al parámetro temporal Δt y orden 2 con respecto al parámetro espacial Δx .

9.9. Problemas de contorno estacionario

En general, para problemas del tipo

$$-\nu u_{xx}(x) + b u_x(x) + \lambda u(x) = f(x), \quad \forall x \in (a, b), \quad (9.9)$$

$$u(a) = u_a, \quad (9.10)$$

$$u(b) = u_b. \quad (9.11)$$

que contienen los efectos típicos de difusión, transporte y reacción se usa la discretización que permite mantener orden 2 en Δx para u_x (Taylor hasta la tercera derivada) y se invierte el sistema lineal resultante. De acuerdo al juego de parámetros ν, b, λ y a las condiciones de contorno impuestas podemos tener o no solución al sistema lineal. Necesitamos por lo tanto algunos resultados sobre matrices que se salen del contenido y tiempo de este curso.

Uno de los problemas más interesantes involucra el transporte y la difusión de una sustancia y se puede describir como:

$$-\varepsilon u''(x) + \lambda u'(x) = f(x) \quad x \in (0, 1), \quad (9.12)$$

$$u(0) = 0, \quad (9.13)$$

$$u(1) = 0. \quad (9.14)$$

La función f y los números reales $\varepsilon > 0$ y λ son tales que existe solución única al problema. Esta ecuación modela el transporte y difusión de un producto en un medio con velocidad λ (hacia la derecha si $\lambda > 0$ y hacia la izquierda si $\lambda < 0$). El parámetro $\varepsilon > 0$ mide la difusión del producto en el medio. La producción y destrucción de la sustancia se modela por el término f , que en general depende de la propia u .

Si suponemos f constante y ε y λ constantes entonces la solución exacta se puede calcular como

$$u(x) = \frac{f}{\lambda} \left(x - \frac{e^{Pe x} - 1}{e^{Pe} - 1} \right) \quad (9.15)$$

donde $Pe = \lambda/\varepsilon$. La razón $Pe = \lambda/\varepsilon$ se llama el **número de Péclét** y mide la importancia del transporte comparado con la difusión. Cuando $|Pe| \gg 1$ el esquema numérico planteado de forma más natural experimenta dificultades y la matriz generada se aproxima a ser singular y esto es porque $\lambda \gg \varepsilon$, es decir, la velocidad de transporte mucho mayor que la difusión y el sistema pierde elipticidad.

Al realizar una aproximación por diferencias finitas a la solución de este problema la matriz tridiagonal resultante es

$$A = \text{tridiag}\left(-1 - Pe \frac{\Delta x}{2}, 2, -1 + Pe \frac{\Delta x}{2}\right).$$

El valor $Pe_{\Delta x} = Pe \frac{\Delta x}{2}$ que se llama el **número de Peclet de la partición** y relaciona el número de Peclet Pe con la talla de la partición. La condición para que se pueda resolver el sistema de forma estable, ver Figura 9.5 es

$$Pe_{\Delta x} \leq 1$$

lo que equivale a

$$\Delta x \leq 2/Pe = 2 \frac{\varepsilon}{\lambda}.$$

9.9.1. Algunos ejemplos

Vamos a usar Euler explícito en tiempo y diferencias centradas en espacio con el propósito de observar la restricción de estabilidad $\mu \leq 0.5$:

Ejemplo 174 La función $u(t, x) = e^{-\pi^2 t} \sin(\pi x)$, $x \in [0, 1]$ cumple

$$\begin{aligned} u_t(t, x) - u_{xx}(t, x) &= 0, \quad \forall x \in (0, 1), t > 0, \\ u(t, 0) &= 0, \quad \forall t > 0, \\ u(t, 1) &= 0, \quad \forall t > 0, \\ u(0, x) &= \sin(\pi x), \quad \forall x \in [0, 1]. \end{aligned}$$

Podemos realizar una tabla con los errores obtenidos para la solución exacta en un tiempo positivo fijo $T > 0$ y contrastar la reducción esperada de acuerdo al orden predicho por la teoría. Para fijar ideas, la norma en la que medimos el error en el nivel temporal n va a ser

$$\|e^n\|_{\Delta x} = \left(\Delta x \sum_{j=0}^{M+1} |u_j^n - u(t^n, x_j)|^2 \right)^{1/2}.$$

Ponemos $\mu = 0.4$ y entonces $\Delta t = \mu \Delta x^2$ con $\Delta x = (1 - 0)/(M + 1)$. Podemos ver un segundo orden de convergencia claramente. También es de destacar que no podemos usar un T muy grande por ser Δt cada vez más pequeño lo que nos obliga a hacer muchas iteraciones.

M	error en $T = 0.02$
20	0.0005...
40	0.00013...
80	0.000034...
160	0.0000086...

Ejemplo 175 Consideremos el problema

$$\begin{aligned} u_t(x, t) - u_{xx}(x, t) &= 0, \quad \forall x \in (0, 1), t > 0, \\ u(t, 0) &= \varphi_0(t), \quad \forall t > 0, \\ u(t, 1) &= \varphi_1(t), \quad \forall t > 0, \\ u(0, x) &= u_0(x), \quad \forall x \in [0, 1]. \end{aligned}$$

con datos

$$\varphi_0(t) = 0, \quad \varphi_1(t) = e^{-\pi^2 t/4}, \quad t > 0$$

$$u_0(x) = \sin(0.5\pi x) + 0.5 \sin(2\pi x), \quad x \in [0, 1].$$

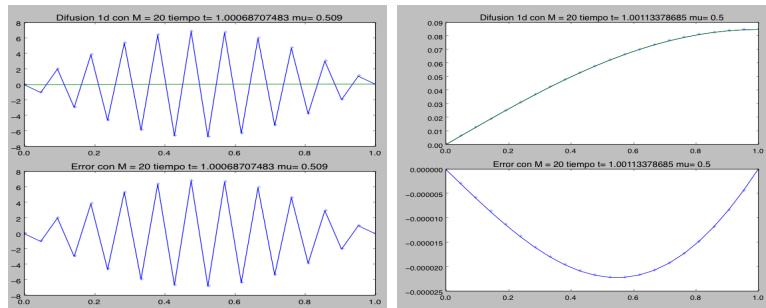


Figura 9.1: Solución y error para $T = 1$ con $M = 20$ puntos espaciales interiores y usando $\mu = 0.5$ y $\mu = 0.509$. Se observa el deterioro de la solución para $\mu > 0.5$.

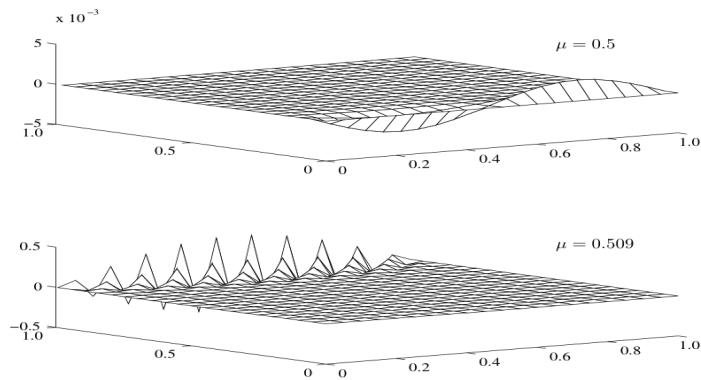


Figura 9.2: Figura extraída del libro de Iserles [18] pag 354, donde se aprecia la evolución temporal del error en $[0, 1]$ con $M = 20$ puntos espaciales y tomando $\mu = 0.5$ y $\mu = 0.509$. Se observa también el deterioro para $\mu > 0.5$.

Entonces la solución exacta es

$$u(t, x) = e^{-\pi^2 t/4} \sin(0.5\pi x) + 0.5 e^{-4\pi^2 t} \sin(2\pi x), \quad x \in [0, 1].$$

Las Figuras 9.1, 9.2 y 9.3 son del libro de Iserles [18]. En la Figura 9.1 se puede ver la solución y el error para $T = 1$ y $M = 20$ puntos espaciales y en la Figura 9.2 se puede ver la evolución temporal del error en $[0, 1]$; se usa $\mu = 0.5$ y $\mu = 0.509$. Se observa el deterioro de la solución para $\mu > 0.5$ debido a la falta de estabilidad. En la Figura 9.3 se puede ver la evolución temporal del error en $[0, 1]$ usando $M = 20, 40, 80, 160$ puntos espaciales interiores y $\mu = 0.4$.

Ejemplo 176 En este ejemplo no disponemos de la solución analítica. En la siguiente Figura 9.4 se observa como una fuente de calor irregular se difumina de forma suave cuando se aplica la ecuación de difusión sobre ella.

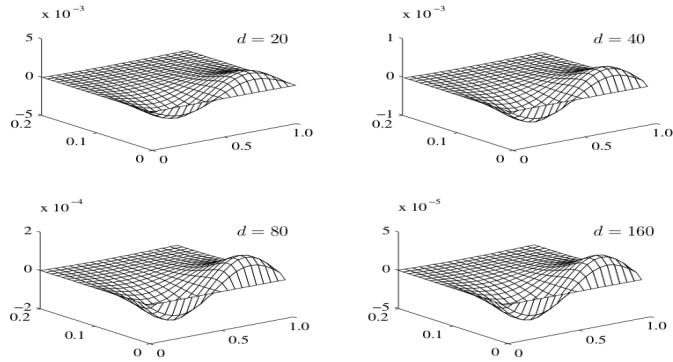


Figura 9.3: Figura extraída del libro de Iserles [18] pag 355, donde se aprecia la evolución temporal del error en $[0, 1]$ usando $M = 20, 40, 80, 160$ puntos espaciales y $\mu = 0.4$.

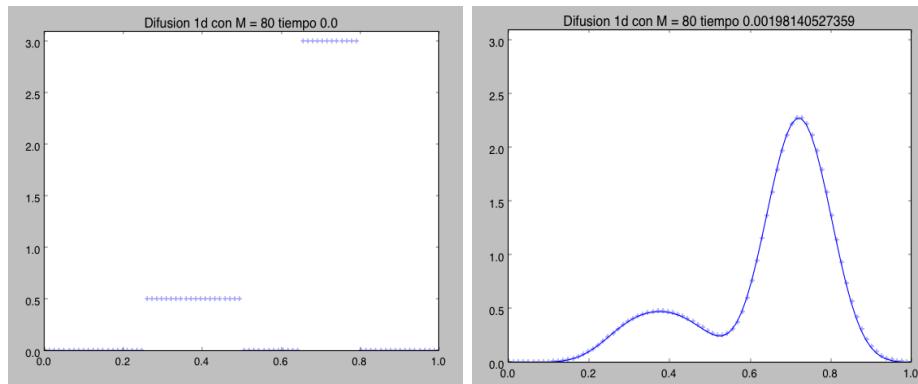


Figura 9.4: Evolución temporal de una fuente de calor inicial discontinua y constante a trozos.

9.10. Ejercicios

1. Consideramos los siguientes problemas de contorno.

- a) $-u''(x) + (1+x)u(x) = f(x)$, $x \in (0, 1)$, $u(0) = \alpha$, $u(1) = \beta$.
 b) $-u''(x) + u'(x) = f(x)$, $x \in (0, 1)$, $u(0) = \alpha$, $u(1) = \beta$.

Para cada uno de ellos desarrollar de forma razonada un esquema en diferencias finitas de orden 2 que permita resolver el problema mediante la resolución de un sistema lineal. Describir explícitamente todos los términos y conceptos involucrados en el planteamiento.

2. La ecuación de contorno

$$\begin{aligned} -u''(x) &= 100e^{-10x}, \quad x \in (0, 1) \\ u(0) &= 0, \\ u(1) &= 0 \end{aligned}$$

tiene la solución exacta $u(x) = 1 - (1 - e^{-10})x - e^{-10x}$. Aproximarla con el computador y realizar un estudio de convergencia usando la norma del máximo.

3. La ecuación de difusión evolutiva

$$\begin{aligned} u_t(t, x) &= u_{xx}(t, x) \quad x \in (0, 1), \quad t > 0 \\ u(t, 0) = u(t, 1) &= 0, \quad t > 0 \\ u(0, x) &= u_0(x), \quad x \in [0, 1]. \end{aligned}$$

con $u_0(x) = 3\sin(\pi x) + 5\sin(4\pi x)$ posee como solución $u(t, x) = 3e^{-\pi^2 t} \sin(\pi x) + 5e^{-16\pi^2 t} \sin(4\pi x)$. Comprobarlo derivando y usar esta solución para confirmar el orden de convergencia de los distintos métodos considerados en el tema.

4. Comprobar para el problema no lineal

$$\begin{aligned} u_t(t, x) &= u_{xx}(t, x) - u^3(t, x) \quad x \in (0, 1), \quad t > 0 \\ u(t, 0) = u(t, 1) &= 0, \quad t > 0 \\ u(0, x) &= u_0(x), \quad x \in [0, 1]. \end{aligned}$$

que se verifica la desigualdad

$$\int_0^1 u(t, x)^2 dx \leq \int_0^1 u_0(x)^2 dx.$$

5. (Tveito [25]) Comprobar para el problema lineal

$$\begin{aligned} u_t(t, x) &= u_{xx}(t, x) + u(t, x) \quad x \in (0, 1), \quad t > 0 \\ u(t, 0) = u(t, 1) &= 0, \quad t > 0 \\ u(0, x) &= u_0(x), \quad x \in [0, 1]. \end{aligned}$$

que se verifica la desigualdad

$$\frac{d}{dt} [e^{-2t} \int_0^1 u(t, x)^2 dx] \leq 0.$$

Concluir que

$$\int_0^1 u(t, x)^2 dx \leq e^{2t} \int_0^1 u_0(x)^2 dx.$$

6. (Tveito [25]) Considerar el problema lineal

$$\begin{aligned} u_t(t, x) &= 4u_{xx}(t, x) - 10u(t, x) + q(t, x) \quad x \in (l_1, l_2), \quad t > 0 \\ u(t, l_1) &= a(t), \quad t > 0 \\ u(t, l_2) &= b(t), \quad t > 0 \\ u(0, x) &= u_0(x), \quad x \in [l_1, l_2]. \end{aligned}$$

- Obtener un esquema explícito.
- Obtener un esquema implícito.
- Suponer

$$l_1 = -2, \quad l_2 = 3, \quad a(t) = e^t - 2, \quad b(t) = e^t + 3, \quad u_0(x) = 1 + x$$

y poner

$$q(t, x) = 11e^t + 10x.$$

Comprobar que

$$u(t, x) = e^t + x$$

es la solución exacta del problema.

- Implementar los esquemas obtenidos en los dos primeros apartados y comparar con la solución analítica obtenida.

7. Siendo $\kappa > 0$ y $a > 0$ consideramos la ecuación de convección difusión

$$\begin{aligned} u_t(t, x) &= \kappa u_{xx}(t, x) + a u_x(t, x) \quad x \in (0, 1), \quad t > 0 \\ u(t, 0) &= 0, \quad t > 0 \\ u(t, 1) &= 0, \quad t > 0 \\ u(0, x) &= u_0(x), \quad x \in [0, 1]. \end{aligned}$$

y el siguiente esquema en diferencias finitas basado en Euler explícito

$$u_l^{n+1} = u_l^n + \mu(u_{l-1}^n - 2u_l^n + u_{l+1}^n) + \frac{1}{2}\theta\mu\Delta x(u_{l+1}^n - u_{l-1}^n)$$

donde tenemos $\mu = \kappa \frac{\Delta t}{\Delta x^2}$ y ponemos $\theta = \frac{a}{\kappa}$. La razón θ mide la importancia del transporte comparado con la difusión. Cuando $|\theta| >> 1$, se transporta más rápido de lo que se difunde, este esquema numérico planteado de forma muy natural experimenta dificultades.

- a) Deducir razonadamente el esquema introduciendo toda la notación conveniente y explicar lo que se está haciendo.
 - b) Obtener el error de consistencia de este esquema con respecto a los parámetros Δt y Δx .
 - c) Comprobar la estabilidad del esquema.
 - d) Demostrar la convergencia e indicar el orden del método con respecto a los parámetros Δt y Δx .
8. Deducir el esquema que surge cuando se aplica Euler implícito y comprobar que si $\theta\Delta x \leq 2$ entonces el sistema lineal resultante es invertible.
9. Para la ecuación de transporte-difusión 1d del ejercicio anterior aplicar computacionalmente el método de Crank-Nicolson para el dato inicial $u(0, x) = 4x(1-x)$, para el valor del coeficiente de difusión $\nu = 0.1$ y para valores de $a = -1$ y $a = +1$. Comentar los resultados computacionales.
10. Consideremos el problema de transporte y difusión con un término fuente general

$$\begin{aligned} w_t(t, x) - \kappa w_{xx}(t, x) + a w_x(t, x) &= g(t, x) \quad x \in (0, 1), \quad t > 0, \\ w(t, 0) = 0, \quad w(t, 1) &= 0, \\ w(0, x) &= w_0(x). \end{aligned}$$

Usando la solución $u(x)$ dada en (9.15) para $f = 1$, construir $g(t, x)$ tal que

$$w(t, x) = \exp\left(\frac{1}{1+t}\right)u(x)$$

sea la solución exacta y comprobar con ella el orden de convergencia del esquema de Crank-Nicolson tanto en espacio como en tiempo.

11. Práctica: Estudio de un problema de transporte y difusión

El objetivo de esta práctica es resolver la ecuación de transporte difusión siguiente:

$$-\varepsilon u''(x) + \lambda u'(x) = f(x) \quad x \in (0, 1), \quad (9.16)$$

$$u(0) = 0, \quad (9.17)$$

$$u(1) = 0. \quad (9.18)$$

La función f y los números reales $\varepsilon > 0$ y λ son tales que existe solución única al problema. Vamos a realizar una aproximación por diferencias finitas a la solución de este problema y estudiar los resultados.

Esta ecuación modela fenómenos como, por ejemplo, lo son el transporte y difusión de un producto químico en un fluido con velocidad λ (si $\lambda > 0$ hay movimiento hacia la derecha y si $\lambda < 0$ lo hay hacia la izquierda). El parámetro $\varepsilon > 0$ mide la difusión del producto en el medio. La producción y destrucción de la especie química se modela por el término f que en general depende de la propia u . Aquí vamos a suponer que es constante o como mucho depende de la posición. También tendremos ε y λ constantes. La razón $\theta = \lambda/\varepsilon$ mide la importancia del transporte comparado con la difusión. Cuando $|\theta| >> 1$, es decir, θ es grande, el esquema numérico planteado de forma más natural experimenta dificultades y la matriz generada se aproxima a ser singular.

Cuestiones teóricas y prácticas a responder:

- a) Obtener una fórmula explícita para la solución del problema en el caso f constante no cero.
- b) Fijemos $\lambda > 0$. Probar la existencia de un valor $x_\theta \in (0, 1)$ que depende sólo de θ tal que la función $(\lambda/f)u(x)$ es estrictamente creciente sobre $(0, x_\theta)$ y estrictamente decreciente sobre $(x_\theta, 1)$. Calcular $\lim_{|\theta| \rightarrow +\infty} x_\theta$.
- c) Para $\lambda > 0$ fijo nos interesamos en el comportamiento de la solución u para $\varepsilon \rightarrow 0^+$, es decir $\theta \rightarrow +\infty$. Calcular $u(x_\theta)$ y $\lim_{\varepsilon \rightarrow 0^+} u(x_\theta)$. Demostrar que

$$\lim_{\varepsilon \rightarrow 0^+} \lim_{x \rightarrow 1} u(x) \neq \lim_{x \rightarrow 1} \lim_{\varepsilon \rightarrow 0^+} u(x).$$

Intentar explicar el significado de la sentencia: para valores pequeños de ε la solución a este problema contiene una estrecha capa límite en el entorno de $x = 1$.

- d) Realizar un código que calcule una aproximación a la solución de este problema en los puntos de una partición. Dibujar la solución aproximada y la exacta para los datos $f = 1$, $\lambda = -1, 1$ y para los valores de ε dados por $\varepsilon = 1, 0.5, 0.1, 0.01$.

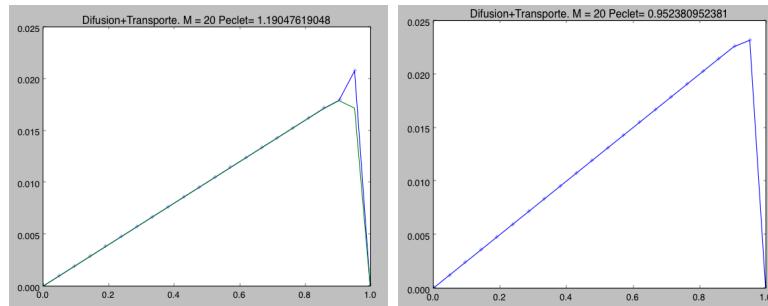


Figura 9.5: Resultados con $M + 1 = 21$ con producidas con $\varepsilon = 1$, $\lambda = 40$ o $\lambda = 50$ y $f = 1$. Distintos valores de $Pe_{\Delta x}$.

- e) **Análisis de error:** Fijado $\varepsilon = 0.1$, $\lambda = 1$ y $f = 1$ y para $M = 10, 20, 30, \dots, 100$ dibujar la curva $\log(M) \rightarrow \log \|e_M\|_{\Delta x}$ donde

$$\|e_M\|_{\Delta x} = \left(\Delta x \sum_{j=0}^{j=M+1} \{|u(x_j) - u_j|^2\} \right)^{1/2}.$$

Deducir un decrecimiento del error en la forma

$$\|e_M\|_{\Delta x} \sim cte M^{-s} = cte (\Delta x)^s, \quad M \rightarrow +\infty$$

y determinar el valor de $s > 0$.

- f) **Análisis de estabilidad:** Fijado $\varepsilon = 0.01$, $\lambda = 1$ y $f = 1$ y para $M = 10$ se obtiene un resultado oscilatorio especialmente cerca de la capa límite que no es satisfactorio, ver la Figura 9.5. Fijado $\varepsilon > 0$, ¿qué valor de M necesitamos para resolver bien el problema? Para responder a esta pregunta tomar $\lambda = 1$ y $f = 1$. Para valores de $\varepsilon \in [0.005, 0.02]$ determinar el entero $M = M(\varepsilon)$ para el que la solución numérica es una razonable aproximación a la exacta, principalmente en la capa límite, esto es: no hay oscilaciones y es una buena aproximación a la capa límite.
Indicación: Fijado ε lanzar el código para $M = 10, 20, 30, \dots$ y para cada valor de M dibujar ambas soluciones, la exacta y la aproximada. Gracias a estos gráficos decidir si la aproximación es buena. Calcular el valor $Pe_{\Delta x} = |\theta| \Delta x / 2$ que se llama el **número de Peclet de la partición**. ¿Qué conclusión se puede sacar del estudio con respecto a este número?

12. Para $\varepsilon > 0$ consideramos la ecuación de difusión evolutiva

$$\begin{aligned} u_t(t, x) - \varepsilon u_{xx}(t, x) &= f(t, x) \quad x \in (0, 1), \quad t > 0 \\ u(t, 0) &= \varphi_0(t), \quad t > 0 \\ u(t, 1) &= \varphi_1(t), \quad t > 0 \\ u(0, x) &= u_0(x), \quad x \in [0, 1]. \end{aligned}$$

Usar la función $u(t, x) = \sin(x) \cos(t)$ como solución exacta calculando los valores adecuados para $f(t, x)$, $\varphi_0(t)$ y $\varphi_1(t)$. Estimar entonces el error en el instante temporal $T = 1$ y para $\varepsilon = 1$ usando la medida de error

$$\|e_M\|_{\Delta x} = \left(\Delta x \sum_{j=0}^{j=M+1} \{|u(x_j) - u_j|^2\} \right)^{1/2}.$$

Comprobar el orden de convergencia de los tres métodos introducidos en clase: Euler explícito, Euler implícito y Crank-Nicolson. Realizar una comparativa sobre el número de iteraciones temporales necesarias en cada método.

Bibliografía

- [1] G. Allaire and S.M. Kaber, *Numerical Linear Algebra; TAM 55*, Springer, New York, 2008.
- [2] V.I. Arnold, *Ordinary Differential Equations*; Springer-Verlag, Berlin, 1992.
- [3] K.E. Atkinson, W. Han, D.E. Stewart, *Numerical solutions of ordinary differential equations*; Wiley, 2009.
- [4] Richard L. Burden, J. Douglas Faires, *Numerical analysis*, Brooks/Cole Cengage Learning, 2010.
- [5] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*, John Wiley and Sons, Ltd, 2008
- [6] W. Cheney, D. Kincaid, *Numerical Mathematics and Computing.*; Brooks/Cole Publishing Company, 1999.
- [7] P.G. Ciarlet, *Introduction a l'analyse numérique matricielle et a l'optimisation*, Ed. Masson 1982
- [8] G. Dahlquist, *Stability and error bounds in the numerical integration of ordinary differential equations*, Thesis Diss. 1958; reprinted in *Trans. Royal Inst. of Technology*, No 130, Stockholm, Sweden, 1959.
- [9] G. Dahlquist G and A. Björck, *Numerical Methods*; Prentice-Hall, Englewood Cliffs, N.J., 1974.
- [10] J.F. Epperson, *An introduction to numerical methods and analysis* Wiley, 2007.
- [11] D. F. Griffiths, D. J. Higham, *Numerical methods for ordinary differential equations*; Springer-Verlag, 2010.
- [12] G.H. Golub, J.M. Ortega, *Scientific Computing and Differential Equations*. Academic Press, 1992.
- [13] G.H. Golub, C.F. Van Loan, *Matrix Computations*, 3rd edition. The Johns Hopkins University Press, 1996.

- [14] E. Hairer, S.P. Norsett, G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff problems* Springer-Verlag, 2008.
- [15] E. Hairer, G. Wanner, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems.* Springer-Verlag, 2010.
- [16] M. W. Hirsch, S. Smale, *Ecuaciones diferenciales, sistemas dinámicos y álgebra lineal* Alianza Universidad, 1974.
- [17] E. Isaacson, H.B. Keller, *Analysis of Numerical Methods*; John Wiley and Sons, New York 1966.
- [18] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, 2nd edition, Cambridge Texts in Applied Mathematics, 2009.
- [19] M. R. King, N.A. Mody , *Numerical and statistical Methods for Bioengineering, Cambridge Texts in Biomedical Engineering*, 2010.
- [20] J.D. Lambert, *Numerical methods for ordinary differential equations*; Wiley, 1991.
- [21] R.J. Leveque, *Finite Difference Methods for Ordinary and Partial Differential Equations Steady State and Time Dependent Problems*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, July, 2007
- [22] J. H. Mathews, K. D. Finn, *Métodos Numéricos con MATLAB*, Prentice Hall, 2000.
- [23] A. Quarteroni, R. Sacco y F. Saleri, *Numerical Mathematics*. Springer-Verlag, 2000.
- [24] L. F. Shampine, R.C. Allen, S. Pruess, *Fundamental of Numerical Computing* John Wiley and Sons, 1997.
- [25] A. Tveito, R. Winther, *Introduction to Partial Differential Equations, a computational approach* Springer, 1998.
- [26] A. Tveito, H.P. Langtangen et al., *Elements of Scientific computing* Springer, 2010.
- [27] John C. Strikwerda *Finite Difference Schemes and Partial Differential Equations*; Wadsforth Brooks/Cole Advanced Books and Software, 1989.