Métodos Numéricos II

Ecuaciones diferenciales ordinarias

Método del trapecio

Resumen

En este texto puedes incluir un resumen del documento. Este informa al lector sobre el contenido del texto, indicando el objetivo del mismo y qué se puede aprender de él.

Andrés Herrera Poyatos
Javier Poyatos Amador
Rodrigo Raya Castellano
Universidad de Granada

Índice

1. Motivación: ecuaciones diferenciales ordinarias de primer orden

Definición 1.1. Dada una función $f : \Omega \subseteq \mathbb{R}^2 \to \mathbb{R}$ continua, una un problema de valores iniciales de primer orden consiste en encontrar aquellas funciones $y : [a, b] \to \mathbb{R}$ de clase 1 que verifiquen $G(y) \subset \Omega$, $y'(t) = f(t, y(t))\ \forall t \in [a, b]$ y la condición inicial $y(t_0) = y_0$, donde $t_0 \in [a, b]$.

De forma simplificada, un problema de valores iniciales se representa de la siguiente forma:

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = y_0 \\ t \in [a, b] \end{cases}$$

Resolver de forma exacta un problema de valores iniciales es muy difícil. Existen ecuaciones diferenciales como $y'(t)^2 + y(t)^2 + 1 = 0$ de las cuales no se conoce una solución exacta. Sin embargo, existen múltiples resultados que permiten asegurar la existencia y unicidad de soluciones de la ecuación diferencial incluso cuando no se puedan obtener soluciones explícitamente.

Uno de los objetivos de la teoría del Análisis Numérico en el campo de las ecuaciones diferenciales ordinarias es resolver de forma aproximada problemas de valores iniciales una vez se conoce la existencia y unicidad de soluciones. En este contexto, es estándar en la literatura especializada considerar siempre condiciones iniciales del tipo $y(a) = y_0$ [4]. Este será el tipo de problemas de valores iniciales que se abordarán en este trabajo.

Un conjunto de técnicas muy populares para resolver de forma aproximada problemas de valores iniciales son los métodos de discretización. Estos métodos tratan de obtener valores aproximados de la solución en un conjunto finito de puntos $t_0, t_1, \ldots, t_n \in [a, b]$ donde $a = t_0 < t_1 < \ldots < t_n = b$. A las aproximaciones obtenidas en dichos puntos se las denota $w_0, w_1, \ldots, w_n$. Evidentemente, siempre se toma $w_0 = y_0$.

La primera idea intuitiva para resolver este problema consiste en interpretar la ecuación $y'(t) = f(t, y(t))$ como un campo vectorial aprovechando la definición de derivada como aproximación lineal de la función en un punto. Esto es, f le asigna a cada punto la dirección en la que varía cualquier solución del problema que pase por ese punto.

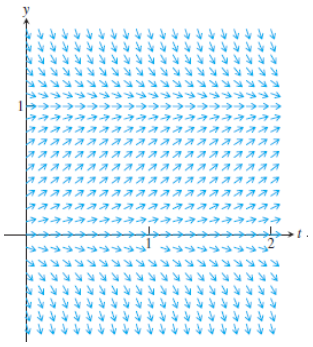

Figura 1: Representación del campo vectorial asociado a la ecuación logística $y'(t) = cy(t)(1 - y(t))$.

Si se conoce la imagen de la solución y en un punto t_i, entonces la dirección de la recta tangente a y en t_i vendrá dada por $f(t_i, y(t_i))$. Por tanto, se puede utilizar la imagen de esta recta tangente en

t_{i+1} para aproximar $y(t_{i+1})$. Esto es, se ha aproximado $y(t_{i+1})$ moviéndose en la dirección que indica el campo vectorial comentado previamente. Repitiendo el proceso para aproximar $y(t_{i+2})$ a partir de w_{i+1}, se obtiene el método de Euler cuya expresión resumida es la siguiente:

$$\begin{cases} w_0 = y_0 \\ h_i = t_{i+1} - t_i \\ w_{i+1} = w_i + h_i f(t_i, w_i) \end{cases}$$

Los mejores resultados se obtienen mediante el uso de puntos equidistantes, esto es, $h = \frac{b-a}{n}$ y $t_i = a + ih \; \forall i = 0 \dots n$. En el resto del texto se trabajará siempre con puntos equidistantes. El estudio del método de Euler concluye que el error global de aproximación cometido es $O(h)$, esto es, existe $M \geq 0$ tal que $|y_i - w_i| \leq Mh$ para todo $i = 0 \dots n$.

A priori, puede parecer que el método de Euler es válido en cualquier aplicación simplemente reduciendo el valor de h, esto es, aproximando un mayor número de puntos. Sin embargo, a continuación se presenta un ejemplo para el cual el método de Euler requiere una excesiva cantidad de puntos para obtener un error de aproximación aceptable.

Ejemplo 1.1: Considérese el siguiente problema de valores iniciales

$$\begin{cases} y'(t) = -4t^3 y^2 \\ y(-10) = 1/10001 \\ t \in [-10, 0] \end{cases}$$

La solución exacta de este problema es $y(t) = \frac{1}{1+t^4}$. La Tabla 1 muestra los resultados de aproximación obtenidos por el método de Euler en $y(0) = 1$ para distintos valores de n. Se observa que la aproximación obtenida deja mucho que desear a pesar de haber llegado a utilizar hasta un 10000 de puntos.

N	h	w_n
100	0.1	0.00390138
1000	0.01	0.03085162
5000	0.002	0.13282140
7500	0.0013	0.18614311
10000	0.001	0.23325153

Tabla 1: Ejemplo de un mal comportamiento del método de Euler.

La Figura 2 muestra las soluciones de la Tabla 1 de forma gráfica. Puede verse que para $n = 100$ la aproximación obtenida es prácticamente nula. Aunque para valores más altos de n las aproximaciones imiten el comportamiento de y, distan mucho del valor real de la función.

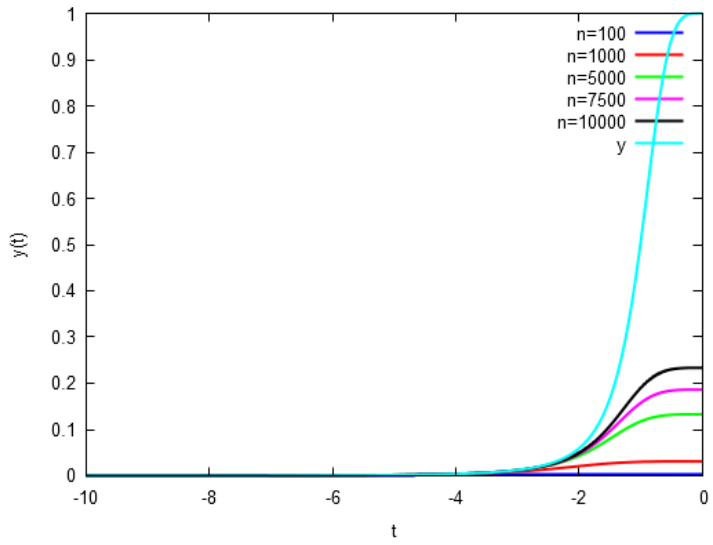

Figura 2: Aproximaciones de $y(t)$ obtenidas con el método de Euler para diferentes valores de n.

El objetivo de este trabajo es introducir un método de discretización para aproximar soluciones de problemas de valores iniciales que presente menor error de aproximación que el método de Euler y consiga resolver el Ejemplo 1.1. El método en cuestión se conoce como método del trapecio y presenta dos variantes denominadas explícita e implícita.

El trabajo se organiza como sigue. En la Sección 2 se explican algunas definiciones y resultados sobre la existencia y unicidad de las soluciones así como propiedades del error y estabilidad de los métodos. Estas definiciones y resultados serán necesarios posteriormente. En la Sección 3 se muestra la idea a partir de la cual surgen las diferentes versiones del método del trapecio. Posteriormente, en las Secciones 4 y 5 se desarrollan los métodos del trapecio explícito e implícito respectivamente. Ambos métodos se estudian desde una doble perspectiva: cálculo del error y estabilidad. Además, se muestra cómo los errores de redondeo afectan al comportamiento del método. En la Sección 6 se resume el artículo de investigación "Nombre del artículo", que pone de manifiesto que la resolución de problemas de valores iniciales sigue siendo un tema abierto en la actualidad. Por último, en la Sección 9 se destacan las conclusiones obtenidas y las ventajas y desventajas del método del trapecio.

2. Definiciones y resultados previos

En esta sección se proporcionan las definiciones y resultados que se necesitan para el estudio del método del trapecio. En primer lugar, una de las hipótesis con las que se suele trabajar para problemas de valores iniciales es que la función f sea lipschitziana en la segunda variable.

Definición 2.1. Sea $\Omega \subset \mathbb{R}^2$ y sea $f : \Omega \to \mathbb{R}$. Se dice que f es lipschitziana respecto de la segunda variable, y, si existe una constante L, llamada constante de Lipschitz, de forma que $|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|$ para cualquier $(t, y_1), (t, y_2) \in \Omega$.

No tiene sentido aplicar un método numérico para resolver un problema de valores iniciales que no tenga solución. Por tanto, los resultados que garanticen la existencia de soluciones al problema son

fundamentales en este contexto. Además, si el problema admitiese varias soluciones distintas, entonces el método puede no comportarse correctamente pues no se sabe cuál debe calcular. Por tanto, la unicidad de soluciones también es un concepto que se debe estudiar en profundidad. El resultado de este estudio se resume en el teorema de existencia y unicidad de soluciones, que utiliza como hipótesis fundamental el concepto de función lipschitziana en la segunda variable.

Teorema 2.1. *(Existencia y unicidad de soluciones) Sea $f : [a, b] \times I \to \mathbb{R}$, donde I es un intervalo de $\mathbb{R}$, y sea $y_0 \in I$. Entonces:*

1. *Si $I = [\alpha, \beta]$ y f es lipschitziana respecto de la segunda variable en $[a, b] \times [\alpha, \beta]$, entonces existe $c \in [a, b]$ tal que el problema de valores iniciales:*

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(a) = y_0 \\ t \in [a, c] \end{cases}$$

   tiene exactamente una solución.

2. *Si $I =]-\infty, \infty[$ y f es lipschitziana respecto de la segunda variable en $[a, b]\times]-\infty, \infty[$, entonces existe exactamente una solución en $[a, b]$*

Nótese que el resultado es válido para cualquier condición inicial escogida. Esto es, la existencia y unicidad solamente depende de f. De aquí en adelante siempre se supondrá que el problema de valores iniciales a resolver tiene solución y que esta es única. En la práctica este hecho es algo que habrá que comprobar mediante el Teorema 2.1. Bajo hipótesis de existencia y unicidad se pueden definir y_i como los valores que toma la solución en los puntos $t_i = a + ih$ para todo $i = 0 \dots n$, donde $h = \frac{b-a}{n}$.

El estudio de los métodos numéricos para problemas de valores iniciales se centra en la acotación de los errores cometidos y en el análisis de la estabilidad de los métodos. Las demostraciones de resultados asociados a estos conceptos suelen requerir el uso de múltiples desigualdades. El siguiente resultado, basado en la desigualdad de Gronwall, proporciona una de las desigualdades con más aplicaciones en esta área.

Teorema 2.2. *Sean dos soluciones $y(t), z(t)$ de la ecuación diferencial $y'(t) = f(t, y(t))$ para las condiciones iniciales $y(a)$ y $z(a)$ respectivamente. Supóngase que f es lipschitziana respecto de la segunda variable. Entonces $|y(t) - z(t)| \leq e^{L(t-a)}|y(a) - z(a)|$ donde L es la constante de Lipschitz de f.*

Un método será mejor que otro cuanto menor error presenten las aproximaciones obtenidas. Sin embargo, el concepto de error se puede ampliar introduciendo los errores locales y globales.

Definición 2.2. Sean w_i los valores estimados en los puntos t_i por cierto método de discretización. Sea también z_i el valor de la solución exacta en t_i para el problema de valores iniciales

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(t_{i-1}) = w_{i-1} \\ t \in [t_{i-1}, t_i] \end{cases}$$

Se definen los siguientes errores:

- Error global de truncatura o error acumulado en el nodo i-ésimo: $g_i = |y_i - w_i|$
- Error local de truncatura o error en un paso: $e_i = |z_i - w_i|$

Dicho de otro modo, el error local es el error cometido al calcular w_{i+1} suponiendo que w_i es el valor exacto de la solución en t_i. Esto es, el error que introduce el método en cada paso. Por su parte, el error global g_i es el error que presenta la aproximación w_i frente a la solución buscada. El error global depende de los errores locales pero no tiene por qué ser suma de éstos. Sin embargo, el error global g_{i+1} sí puede entenderse por la suma del error local, e_{i+1}, y el error global del paso previo amplificado. Este hecho se puede visualizar en la Figura 3, que ejemplifica los conceptos de errores locales y globales.

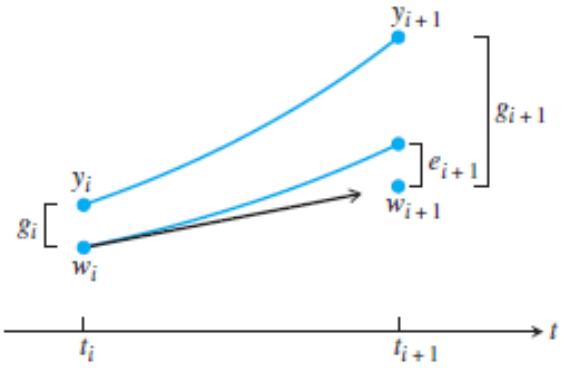

Figura 3: Representación gráfica de los errores locales y globales.

La relación entre errores locales y errores globales viene dada por el siguiente teorema:

Teorema 2.3. *Supóngase que la función f es lipschitziana en la segunda variable con constante de Lipschitz L. Además, supóngase que existen $C \geq 0$ y $k \in \mathbb{N}$ tales que los errores locales verifican $e_i \leq Ch^{k+1}$ para todo $i = 0 \dots n$. Entonces, se verifica la siguiente desigualdad para los errores globales*

$$g_i \leq \frac{Ch^k}{L}(e^{L(t_i - a)} - 1) \tag{1}$$

Demostración. En breve. □

La siguiente definición pone nombre a las acotaciones del Teorema 2.3.

Definición 2.3. Considérese un método de discretización para problemas de valores iniciales. Entonces:

1. El método es localmente de orden k si existe una constante $C \geq 0$ tal que $e_i \leq Ch^k$ para todo $i = 0 \dots n$.
2. El método es de orden k si existe una constante $C \geq 0$ tal que $g_i \leq Ch^k$ para todo $i = 0 \dots n$.

Las constantes de la definición previa dependerán del problema de valores iniciales en cuestión. Nótese que el Teorema 2.3 está diciendo si un método es localmente de orden $k + 1$, entonces es de orden k. Como aplicación directa de este teorema se obtiene fácilmente el orden del método de Euler.

Teorema 2.4. *Supóngase que $f : [a, b] \times [\alpha, \beta] \to \mathbb{R}$ es derivable y lipschitziana en la segunda variable. Entonces, el método de Euler es localmente de orden 2. Consecuentemente, el método de Euler es de orden 1.*

Demostración. Sea y la solución del problema de valores iniciales para $y(a) = y_0$. Se fija $i = 1 \dots n$ y sea z la solución del problema de valores iniciales para $z(t_{i-1}) = w_{i-1}$. Por inducción, z es de clase infinito. El teorema de Taylor para orden 2 proporciona la siguiente igualdad para cualquier

$$z_i = w_{i-1} + hf(t_{i-1}, w_{i-1}) + \frac{h^2}{2} z''(\xi_i) = w_i + \frac{h^2}{2} z''(\xi_i) \tag{2}$$

donde $\xi_i \in [t_{i-1}, t_i]$. Por tanto, si se utiliza esta igualdad en la expresión del error local se tiene

$$e_i = |z_i - w_i| = \left| \frac{h^2}{2} z''(\xi_i) \right| \leq \frac{M_i}{2} h^2 \tag{3}$$

donde $M_i = \text{máx}\{z''(t) : t \in [t_{i-1}, t_i]\}$. Tomando $M = \text{máx}_{i=1\dots n} M_i$, se tiene que el método de Euler es localmente de orden 2 como se quería. La prueba la cierra la aplicación del Teorema 2.3. □

El orden de un método permite tener información teórica sobre el error que se va a cometer. Otra propiedad que debe interesar al estudiar un método de discretización es que este sea convergente.

Definición 2.4. Un método de discretización se dice convergente si para cualquier problema de valores iniciales tal que la función f es lipschitziana respecto de la segunda variable se tiene que $\lim_{n \to +\infty} y_n = y(b)$.

Si el orden del método es $O(h^r)$ con $r > 0$, entonces es claro que el método es convergente ya que el error global en cualquier punto tiende a 0 cuando $n \to +\infty$. Sin embargo, en la práctica el método puede no converger por culpa de los errores de redondeo. Se incidirá en ésto en la Sección 4.2.

En la práctica, cuando se considera un problema de valores iniciales $y' = f(t, y)$ con condición inicial $y(a) = y_0$ se puede cometer un error al evaluar la condición inicial, utilizando $y_0 + \epsilon_0$ en lugar de y_0. Al problema de valores iniciales dado por $y' = f(t, y)$ e $y(a) = y_0 + \epsilon_0$ se le denomina perturbación del problema inicial. Cuando se aplica un método al problema perturbado, el error introducido inicialmente puede ir aumentando en cada iteración e incluso diverger cuando h tiende a 0. Por lo tanto, para poder utilizar cualquier valor de h interesa que el método a utilizar sea estable ante perturbaciones. Esto es, si se perturba la condición inicial, la diferencia entre las aproximaciones obtenidas en ambos problemas están acotadas independientemente del número de puntos que se utilicen. Este concepto se formaliza en la siguiente definición.

Definición 2.5. Un método de discretización se dice estable si para cualquier PVI verificando que f es lipschitziana respecto de la segunda variable y para cualquier perturbación de este PVI existen constantes positivas h_0 y K tales que la diferencia entre las aproximaciones obtenidas para ambos PVI están acotadas por $K|y_0 - y_0'|$ para todo $h \in [0, h_0]$. Esto es, si w_i son las aproximaciones obtenidas para el problema sin perturbar y w_i' son las aproximaciones obtenidas para el problema perturbado, utilizando en ambos casos el mismo $h < h_0$, entonces $|w_i - w_i'| \leq K|y_0 - y_0'|$ para todo i.

En definitiva, la estabilidad indica que un error cometido al principio del método no se magnifica al calcular las sucesivas iteraciones del método, siempre permanece acotado por $K|y_0 - y_0'|$.

Muchos de los métodos de discretización pueden ser escritos de la forma $y_{i+1} = y_i + h\phi(t_i, y_i, h)$ donde ϕ es una función de t, y y h que, además, está definida en función de f. A este conjunto de métodos se los denomina métodos de paso. A la función ϕ se la denomina función incremento. Esta generalización permite probar resultados para métodos de paso arbitrarios y aplicarlos después a casos particulares como el método de Euler.

Teorema 2.5. *Si un método de un paso anterior verifica que ϕ es continua en cada una de sus variables y, además, es lipschitziana respecto de la segunda variable en el correspondiente dominio para $h \in [0, h_0]$, entonces:*

1. *el método es estable.*
2. *el método es convergente o, equivalentemente, $\phi(b, y, 0) = f(b, y)$.*

Demostración. i) Puede encontrarse en los ejercicios resueltos de la sección 5.10 del libro de Burden. ii) Puede encontrarse en la sección 4.3 del libro de Gear: "Numerical initial value problems in ordinary differential equations". □

3. Introducción al método del trapecio

El método del trapecio se basa en la siguiente proposición:

Proposición 3.1. *Considérese el problema de valores iniciales dado por la ecuación diferencial $y'(t) = f(t, y(t))$ sobre $[a, b]$ y la condición $y(t_0) = y_0$. Entonces, son equivalentes:*

1. *y es una solución del problema de valores iniciales.*
2. *$y(t) = y_0 + \int_{t_0}^{t} f(s, y(s)))ds \ \forall t \in [a, b]$*

Demostración. Es consecuencia directa del Teorema Fundamental del Cálculo. □

Utilizando la Proposición 3.1, si un PVI con condición inicial $t_0 = a$, $y(t_0) = y_0$ tiene solución única, entonces esta es la única solución de la siguiente ecuación

$$y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))) \ ds \tag{4}$$

En este contexto se pueden aplicar los métodos de integración numérica para aproximar la integral que aparece en la segunda igualdad. Para ello supóngase de aquí en adelante que f es diferenciable. En tal caso una obvia inducción concluye que y es de clase infinito. Por tanto, se puede utilizar la fórmula del trapecio para integración numérica, obteniendo la siguiente igualdad

$$y(t_1) = y_0 + \frac{h}{2}\left[f(t_0, y_0) + f(t_1, y(t_1))\right] - \frac{h^3}{12} y^{3)}(\xi) \tag{5}$$

donde $\xi \in [t_0, t_1]$. Ignorando el último sumando se obtiene la aproximación dada en (6), que tiene error $-\frac{h^3}{12} y^{3)}(\xi)$.

$$y(t_1) \approx w_1 = w_0 + \frac{h}{2}\left[f(t_0, w_0) + f(t_1, y(t_1))\right] \tag{6}$$

El problema reside en que para aproximar el valor de y en t_1 se debe conocer previamente dicho valor. En este contexto se plantean dos soluciones diferentes obteniendo dos métodos, denominados método del trapecio explícito e iterativo respectivamente. En el resto del texto se desarrollan sendos métodos, proporcionando el error teórico cometido y resultados de convergencia y estabilidad.

4. Método del trapecio explícito

Recuérdese en este punto el método de Euler para ecuaciones diferenciales ordinarias que se comentó en la Sección 1. Llámese w'_i a las aproximaciones obtenidas por este método. El valor de la solución y en cada punto se aproxima mediante la siguiente expresión, donde $w'_0 = y_0$:

$$y(t_{i+1}) \approx w'_{i+1} = w'_i + hf(t_i, w'_i))$$

Se comentó previamente que el problema de la aproximación (6) reside en que el valor a aproximar aparece en el segundo miembro de la expresión. Para solventar este hecho se puede utilizar la aproximación dada por el método de Euler en su lugar. De esta forma se obtiene la siguiente aproximación:

$$y(t_{i+1}) \approx w_{i+1} = w_i + \frac{h}{2}\left[f(t_i, w_i) + f(t_i + h, w_i + hf(t_i, w_i))\right] \tag{7}$$

Sean $S_L = hf(t_i, w_i)$ y $S_R = hf(t_{i+1}, w'_{i+1})$. El método de Euler obtiene (t_{i+1}, w'_{i+1}) sumándole S_L a (t_i, w_i) . Por su parte, el método del trapecio explícito obtiene (t_{i+1}, w_{i+1}) como (t_i, w_i) más la media de S_L y S_R. La Figura 4 muestra este hecho de forma visual.

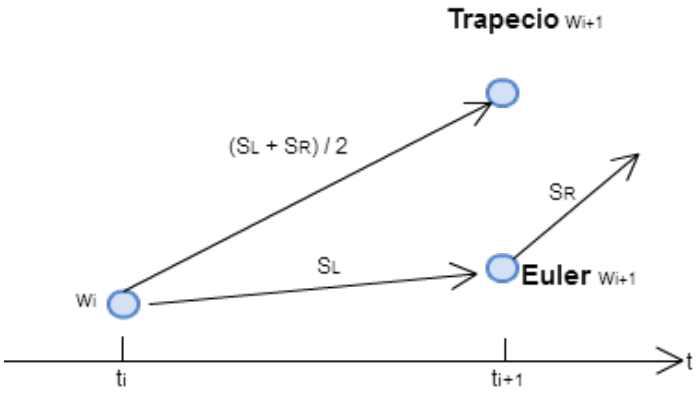

Figura 4: Esquema visual del método del trapecio explícito en contraposición con el método de Euler.

Puesto que la nueva aproximación w_i se consigue "mejorando"la aproximación del método de Euler mediante una media, cabe esperar que el método del trapecio explícito presente un mejor comportamiento. En efecto, este hecho se estudiará en la Sección 4.1. Posteriormente en la Sección 4.3 se estudiará la estabilidad del método del trapecio explícito.

4.1. Error local y global

El estudio del error del método del trapecio explícito se resume en el siguiente teorema.

Teorema 4.1. *El error local del método del trapecio es de orden tres. En consecuencia, el error global del método es de orden dos.*

Demostración. La prueba es similar a la realizada en el Teorema 2.4. Sean $w_0, w_1, \ldots, w_n$ las aproximaciones obtenidas por el método del trapecio explícito, se fija $i = 1, 2, \ldots, m$. Sea z la solución del problema de valores iniciales para la condición $z(t_{i-1}) = w_{i-1}$. Por inducción, z es de clase infinito.

Por tanto, aplicando la igualdad (5) a z, se obtiene la siguiente expresión:

$$z_i = w_{i-1} + \frac{h}{2}\left[f(t_{i-1}, w_{i-1}) + f(t_i, z_i)\right] + 0(h^3)$$

Denótese $w_i' = w_{i-1} + hf(t_{i-1}, w_{i-1})$. Utilizando la definición de w_i y la expresión previa en la definición de error local se obtiene

$$e_i = |z_i - w_i| = \left|\frac{h}{2}\left[f(t_i, z_i) - f(t_i, w_i')\right] + 0(h^3)\right| \leq \frac{h}{2}\left|f(t_i, z_i) - f(t_i, w_i')\right| + 0(h^3)$$

Considérese el desarrollo de Taylor de orden 1 con respecto de la variable y para $f(t_i, w_i')$ en el punto (t_i, z_i):

$$f(t_i, w_i') = f(t_i, z_i) + \frac{\partial f}{\partial y}(t_i, z_i)(w_i' - z_i) = f(t_i, z_i) + O(h^2)$$

donde se ha usado que $z_i - w_i'$ es el error local del método de Euler y, por tanto, es de orden 2. Basta juntar las dos expresiones obtenidas para conseguir

$$e_i \leq \frac{h}{2}\left|f(t_i, z_i) - f(t_i, w_{i-1} + hf(t_{i-1}, w_{i-1})\right| + 0(h^3) = \frac{h}{2}O(h^2) + 0(h^3) = O(h^3)$$

Por último, el Teorema 2.3 implica que el error global es de orden 2.

□

Demostración. Prueba alternativa.

Podemos particularizar la expresión del método para $j - 1 = 0$ y $j = 1$ como $y_1 = y_0 + \frac{h}{2}(f(t_0, y_0) + f(t_0 + h, y_0 + hK_0)$ siendo $K_0 = f(t_0, y_0)$.

Para obtener el error local suponemos que y_0 es exacto, $y_0 = y(t_0)$. Claramente, $K_0 = y'(t_0)$. Considerando el desarrollo de Taylor en varias variables para $K_1 = f(t_0 + h, y_0 + hK_0)$ en el punto (t_0, y_0) se tiene:

$K_1 = f(t_0, y_0) + h\frac{\partial f(t_0, y_0)}{\partial t} + hK_0\frac{\partial f(t_0, y_0)}{\partial y} + O(h^2) = y'(t_0) + hy''(t_0) + O(h^2)$

ya que $\frac{\partial f(t, y(t))}{\partial t} = \frac{\partial f}{\partial t} + y'(t)\frac{\partial f}{\partial y}$

Por lo tanto $y_1 = y_0 + \frac{h}{2}(2y'(t_0) + hy''(t_0) + O(h^2)) = y_0 + hy'(t_0) + \frac{1}{2}h^2y''(t_0) + O(h^3)$

Por otro lado el desarrollo de $y(t_1) = y(t_0 + h)$ en torno a t_0 es $y(t_1) = y(t_0) + hy'(t_0) + \frac{1}{2}h^2y''(t_0) + O(h^3)$.

De donde obtenemos $y(t_1) - y_1 = O(h^3)$ y aplicando el teorema 2.3 sabemos que el error global es $O(h^2)$. □

4.2. Error de redondeo

El error de redondeo debe tenerse en cuenta a la hora de evaluar un método numérico. En el caso de las ecuaciones diferenciales se tiene que la situación es análoga a la encontrada con las fórmulas de derivación: el error de truncamiento disminuye con h, pero el error de redondeo aumenta, existiendo un

valor óptimo para el cual la suma de estos errores es mínima (véase la Figura 5). Este valor óptimo de h suele ser tan pequeño que utilizarlo supone un coste computacional muy grande. Este hecho explica la importancia de utilizar métodos con el mayor orden posible pues son capaces de obtener aproximaciones de calidad sin necesidad de utilizar valores de h muy pequeños. [**Vazquez**]

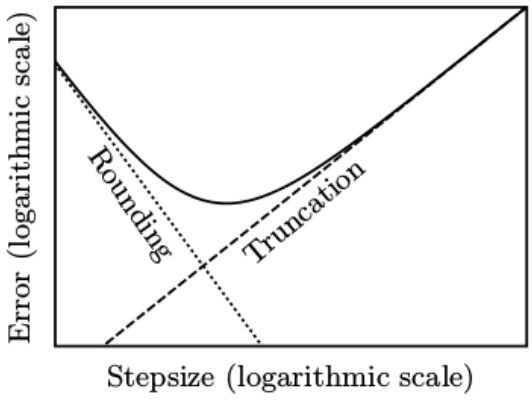

Figura 5: Combinación de los errores de truncatura y redondeo

En el caso del método del trapecio, si se toma $h = \frac{b-a}{n}$, entonces en cada uno de los n pasos se comete un error de redondeo acotado por ϵ además del error local de truncatura Ch^3. Globalmente, por lo tanto, se obtiene el siguiente error

$$n(\epsilon + Ch3) = \frac{\epsilon}{h} + Ch^2$$

En una situación ideal la cota ϵ será del orden de la precisión de la máquina μ y la constante C será del orden del cubo de la constante de Lipschitz L^3 de modo que el valor óptimo para el paso h se obtendría (derivando e igualando a cero) para $h = \frac{\sqrt[3]{\mu}}{L}$.

Sin embargo, el análisis es bastante más complicado que esto y existen al menos dos vías para su estudio. El primero es un modelo más pesimista que se pondría en el peor de los casos. El segundo es un modelo probabilista que se puede encontrar en el libro de Henrici "Discrete Variable Methods in Ordinary Differential Equations". Otros autores como Butcher en su libro "Numerical Methods for Ordinary Differential Equations" en vez de llevar a cabo un análisis detallado de la situación mencionan el uso del llamado algoritmo de Gill-Moller o "suma compensada". Este algoritmo persigue reducir los efectos de los errores de redondeo.[**Butcher**]

4.3. Estabilidad y convergencia

Aplicamos este teorema al método del trapecio. En este caso $\phi(t, y, h) = \frac{1}{2}f(t, y) + \frac{1}{2}f(t+h, y+hf(t, y))$. Asumiendo las condiciones que nos dan existencia y unicidad, si f es Lipschitziana en $\{(t, y) : a \leq t \leq b, y \in \mathbb{R}\}$ con constante de Lipschitz L entonces:

$|\phi(t, y, h) - \phi(t, y', h)| = |\frac{1}{2}f(t, y) + \frac{1}{2}f(t + h, y + hf(t, y)) - \frac{1}{2}f(t, y') - \frac{1}{2}f(t + h, y' + hf(t, y'))| \leq \frac{1}{2}L|y-y'|+\frac{1}{2}L|y+hf(t, y)-y'-hf(t, y')| \leq L|y-y'|+\frac{1}{2}L|hf(t, y)-hf(t, y')| = (L+\frac{1}{2}hL^2)|y-y'|$

Por tanto, ϕ satisface una condición de Lipschitz sobre el conjunto $\{(t, y, h) : a \leq t \leq b, y \in \mathbb{R}, h \in [0, h_0]\}$ con constante de Lipschitz $L' = L + \frac{1}{2}h_0L^2$ para cualquier $h_0 > 0$.

Finalmente, si f es continua en $\{(t,y) : a \leq t \leq b, y \in \mathbb{R}\}$ entonces ϕ es continua en $\{(t,y,h) : a \leq t \leq b, y \in \mathbb{R}, h \in [0,h_0]\}$ directamente por la propia definición de ϕ.

De este modo podemos aplicar el teorema anterior y tenemos demostrado que el método del trapecio es estable.

Considerando ahora $\phi(t,y,0) = \frac{1}{2}f(t,y) + \frac{1}{2}f(t,y) = f(t,y)$ tenemos la condición de consistencia expresada anteriormente lo que nos dice que el método es convergente.

5. Método del trapecio implícito

Recuérdese en este punto la aproximación (6), que se mostraba en la Sección 3, dada por

$$y_1 \approx w_1 = w_0 + \frac{h}{2}\left[f(t_0,w_0) + f(t_1,y_1)\right]$$

En caso de ser una igualdad ($w_1 = y_1$), se tendría la siguiente ecuación implícita

$$w_1 = w_0 + \frac{h}{2}\left[f(t_0,w_0) + f(t_1,w_1)\right]$$

En esta sección se resolverá dicha ecuación implícita mediante métodos numéricos. La solución obtenida será tomada como w_1. Posteriormente, se puede repetir el proceso para obtener w_2. En general, para cada $i = 1 \ldots n$ se está resolviendo la siguiente ecuación implícita

$$w_i = w_{i-1} + \frac{h}{2}\left[f(t_{i-1},w_{i-1}) + f(t_i,w_i)\right] \tag{8}$$

Si se define $g_i(w) = w_{i-1} + \frac{h}{2}\left[f(t_{i-1},w_{i-1}) + f(t_i,w)\right]$, en definitiva se está buscando un punto fijo de g_i. En este contexto se puede aplicar un método de iteración funcional para calcular dicho punto fijo. En caso de obtenerse, el siguiente resultado proporciona el error local cometido.

Proposición 5.1. *Sea w_i una solución de la ecuación implícita (8). Supóngase que f es lipschitziana en la segunda variable con constante de Lipschitz L. En tal caso, si se toma w_i como aproximación y $hL + r < 2$ para cierto $r > 0$, entonces el error local cometido es $O(h^3)$.*

Demostración. Basta aplicar la igualdad (5), tomando como condición inicial $z(t_{i-1}) = w_{i-1}$, junto con la ecuación implícita (8):

$$e_i = |z_i - w_i| = \left|\frac{h}{2}\left[f(t_i,z_i) - f(t_i,w_i)\right] - \frac{h^3}{12}z^{3)}(\xi)\right| \leq \frac{hL}{2}e_i + \left|\frac{h^3}{12}z^{3)}(\xi)\right|$$

Por tanto, juntando los e_i y usando que $2 \neq hL$, se obtiene la siguiente desigualdad:

$$e_i \leq \frac{2}{2-hL}\left|\frac{h^3}{12}z^{3)}(\xi)\right|$$

Se ha tomado h lo suficientemente pequeño de manera que $\frac{2}{2-hL} < \frac{2}{r}$. Por tanto, $e_i = O(h^3)$. □

Normalmente se trabaja con funciones f que sean lipschitzianas respecto de la segunda variable. Por tanto, tomando h lo suficientemente pequeño, siempre se pueden verificar las hipótesis de la proposición previa.

La pregunta que queda por resolver es qué método de iteración funcional se debe utilizar para conseguir aproximar un punto fijo de g_i. Una primera respuesta puede ser el método de Newton en caso de conocer la derivada parcial de f con respecto de la segunda variable. Este método asegura la convergencia en un entorno del punto fijo. Por tanto, si se parte de una aproximación inicial apropiada, como puede ser el método de Euler, es probable que el método de Newton converja y, además, con orden de convergencia 2.

Sin embargo, utilizando que f es lipschitziana en la segunda variable, la función g_i va a ser lipschitziana con constante de Lipschitz $\frac{hL}{2}$. Esto sugiere utilizar el método de iteración funcional dado por g_i partiendo de una aproximación inicial, que se denota $w_i^{(0)}$. La sucesión definida por el método de iteración funcional es la siguiente

$$w_i^{(j+1)} = g_i(w_i^{(j)}) = w_{i-1} + \frac{h}{2}\left[f(t_{i-1}, w_{i-1}) + f(t_i, w_i^{(j)})\right] \tag{9}$$

El objetivo es estudiar cuándo la sucesión $\{w_i^{(j)}\}$ converge. En tal caso, el límite es un punto fijo de g_i y es la aproximación w_i buscada. La aplicación de este método de iteración funcional para resolver la ecuación implícita (8) es lo que se conoce en la literatura especializada como método del trapecio iterativo [1].

La siguiente proposición proporciona una condición suficiente de convergencia que no depende de la aproximación inicial escogida.

Proposición 5.2. *Supóngase que la función f está definida en $[a,b]\times[-\infty,+\infty]$ y es lipschitziana en la segunda variable con constante de Lipschitz L. Si $\frac{LH}{2}<1$, entonces existe w_i tal que $\{w_i^{(j)}\}$ converge a w_i para cualquier aproximación inicial.*

Demostración. La función g_i está definida en $\mathbb{R}$. La constante de Lipschitz de g_i es $\frac{hL}{2}$. Por tanto, si $\frac{LH}{2}<1$, entonces g_i es una contracción sobre $\mathbb{R}$. El resultado se desprende del teorema del punto fijo de Banach. □

La función f puede estar definida en $[a,b]\times[\alpha,\beta]$ con $-\infty<\alpha<\beta<+\infty$ si se toma una buena aproximación inicial. El problema reside en conseguir que $g_i([\alpha,\beta])\subset[\alpha,\beta]$. Esta cuestión ya se estudió durante los métodos de iteración funcional. El resultado del estudio se recoge en la siguiente proposición.

Proposición 5.3. *Sea $w_i^{(0)}$ la aproximación inicial obtenida. Supóngase que la función f es lipschitziana en la segunda variable, con constante de Lipschitz L, en el intervalo $[w_i^{(0)}-r, w_i^{(0)}+r]$ para $r>0$. Si $\frac{LH}{2}<1$ y $\left|w_i^{(1)}-w_i^{(0)}\right|<(1-\frac{Lh}{2})r$, entonces $\{w_i^{(j)}\}$ está bien definida y es convergente.*

Demostración. La función g_i se puede restringir a $[w_i^{(0)}-r, w_i^{(0)}+r]$. La constante de Lipschitz de g_i es $\frac{hL}{2}$. Las otras dos hipótesis proporcionan los siguientes hechos:

1. g_i es una contracción sobre $[w_i^{(0)}-r, w_i^{(0)}+r]$.

2. $\left|g_i(w_i^{(0)}) - w_i^{(0)}\right| < (1 - \frac{Lh}{2})r$. Por tanto, si $x \in [w_i^{(0)} - r, w_i^{(0)} + r]$, entonces

$$\left|g_i(x) - w_i^{(0)}\right| \leq \left|g_i(x) - g_i(w_i^{(0)})\right| + \left|g_i(w_i^{(0)}) - w_i^{(0)}\right| \leq \frac{Lh}{2}\left|x - w_i^{(0)}\right| + (1 - \frac{Lh}{2})r \leq r$$

Luego $g_i(x) \in [w_i^{(0)} - r, w_i^{(0)} + r]$.

De nuevo, la tesis se consigue aplicando el teorema del punto fijo de Banach. □

Tomando h lo suficientemente pequeño se pueden conseguir las hipótesis de las dos proposiciones previas. Además, esto también asegura que se verifique la Proposición 5.1. En tal caso, se ha conseguido un método con error local de orden 3 y, por tanto, con error global de orden 2.

En la práctica solo se realiza un número pequeño de iteraciones de la fórmula (9). La ventaja del método del trapecio iterativo reside en que se puede calcular de forma teórica un número de iteraciones, llámese j, de manera que el error $\left|w_i - w_i^{(j)}\right|$ sea tan pequeño como se quiera. Esto es posible gracias a que se parte de que g_i es lipschitziana con constante de lipschitz $\frac{hL}{2}$ menor que 1. Consecuentemente:

$$\left|w_i - w_i^{(j)}\right| = \left|g_i(w_i) - g_i(w_i^{(j-1)})\right| \leq \frac{hL}{2}\left|w_i - w_i^{(j-1)}\right| \leq \ldots \leq \left(\frac{hL}{2}\right)^j \left|w_i - w_i^{(0)}\right|$$

Por tanto, a la j-ésima iteración el error de aproximación cometido por no calcular exactamente w_i es $O(h^j)$. Se necesita que este sea $O(h^3)$ o menor para conseguir mantener un error local de orden 3. En la literatura especializada se recomienda incluso que se reduzca $\left|w_i - w_i^{(j)}\right|$ a $O(h^4)$ para mejorar el comportamiento del método [1].

La aproximación que se toma como $w_i^{(0)}$ suele ser obtenida mediante un método explícito de menor orden, como el método de Euler. Nótese que en tal caso $w_i^{(1)}$ es el resultado de aplicar el método del trapecio explícito partiendo de w_{i-1}. Por tanto, el método del trapecio iterativo puede entenderse como una generalización del método del trapecio explícito. Además, el estudio del error local realizado para el método del trapecio explícito concluye que si se toma $w_i^{(1)}$ como aproximación, entonces el error local es $O(h^3)$. Por tanto, bajo las hipótesis adecuadas siempre se tendrá garantizado un error local de orden 3 tras la primera iteración del método de iteración funcional.

Cabe destacar que el método del trapecio iterativo también puede concebirse como un mecanismo para corregir el error cometido por la aproximación inicial. Al método utilizado para calcular la aproximación inicial se le denomina predictor mientras que a la fórmula (9) se la denomina fórmula correctora. La aplicación de la fórmula correctora al resultado del predictor es lo que se conoce como método predictor-corrector en la literatura especializada.

6. Artículo de investigación

7. Ejemplos

7.1. Ejemplo 1

Considérese el ejemplo de problema de valores iniciales dado en la motivación.

$$\begin{cases} y'(t) = -4t^3y^2 \\ y(-10) = 1/10001 \\ t \in [-10, 0] \end{cases}$$

cuando se resuelve mediante el método de Euler con paso 10^{-3}, 10^{-4} y 10^{-5} se obtienen las siguientes gráficas.

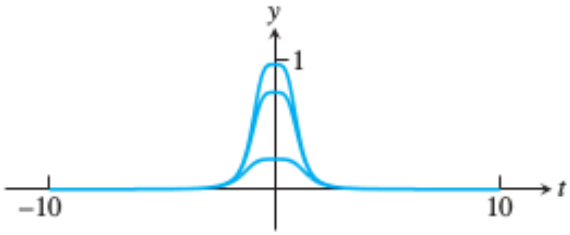

Figura 6: Método de Euler aplicado al problema anterior con paso 10^{-3}, 10^{-4} y 10^{-5}

Aplicando la regla del trapecio explícito con paso 10^{-3} se obtiene la siguiente gráfica:

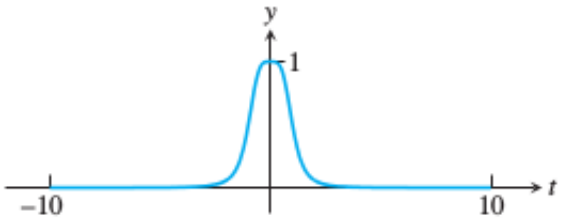

Figura 7: Método del trapecio explícito aplicado al problema anterior con paso 10^{-3}

de modo que queda patente que converge el doble de rápido que el método de Euler.

7.2. Ejemplo 2

Considérese el problema de valores iniciales

$$\begin{cases} y'(t) = -1 + \frac{y}{t} \\ y(1) = 0 \end{cases}$$

calcular el valor de $y(2)$ para $h = 0{,}25$ y $h = 0{,}1$.

j	t_{j-1}	y_{j-1}	t_j	y_j
1	1.00	0.000000	1.25	-0.275000
2	1.05	-0.275000	1.50	-0.600833
3	1.10	-0.600833	1.75	-0.968829
4	1.15	-0.968829	2.00	-1.372859

Tabla 2: Trapecio con $h = 0{,}25$

j	t_{j-1}	y_{j-1}	t_j	y_j
1	0.00	0.000000	0.10	-0.104545
2	0.10	-0.104545	0.20	-0.218216
3	0.20	-0.218216	0.20	-0.340247
4	0.30	-0.340247	0.40	-0.469991
5	0.40	-0.469991	0.50	-0.606896
6	0.50	-0.606896	0.60	-0.750480
7	0.60	-0.750480	0.70	-0.900326
8	0.70	-0.900326	0.80	-1.056065
9	0.80	-1.056065	0.90	-1.217366
10	0.90	-1.217366	1.00	-1.383938

Tabla 3: Trapecio con $h = 0{,}1$

Teniendo en cuenta que $y(2) = -1{,}386294$, los errores relativos son $9{,}6910^{-3}$ para el caso $h = 0{,}25$ y $1{,}7010^{-3}$ para el caso $h = 0{,}10$. Como el método del trapecio es de orden 2 el error relativo es $O(h^2)$ y por tanto el cociente de los errores debería ser $\frac{C(0{,}25)^2}{C(0{,}10)^2} = 6{,}25$ mientras que el valor real es 5.7. La razón de esta diferencia es que el orden es $O(h^2)$ asintóticamente, esto es, cuando $h \to 0$ y los valores de considerados para h no son suficientemente pequeños.

8. Ejercicios teórico-prácticos

8.1. Ejercicio 1

Considérese el problema de valores iniciales

$$\begin{cases} y'(t) = y - t^2 \\ y(0) = 3 \end{cases}$$

calcular una aproximación a la solución del problema de valores iniciales mediante el método de Euler y el método del Trapecio Explícito.

La función $f(t, y) = y - t^2$ es continua, su derivada parcial respecto de y, esto es la función $g(t, y) = 1$ también lo es y esta acotada por $L = 1$ en [0,2], luego se tiene que existe solución y es única.

A continuación, se va a calcular la aproximación mediante el método de Euler. Para ello calculamos la sucesión de puntos que va converge al valor exacto:

j	t_j	y_j
0	0.0	3
1	0.2	3.6
2	0.4	4.312
3	0.6	5.1424
4	0.8	6.09888
5	1.0	7.190656
6	1.2	8.428787
7	1.4	9.826544
8	1.6	11.399853
9	1.8	13.167824
10	2.0	15.153389

Tabla 4: Trapecio con $h = 0{,}2$

A continuación se dibuja la gráfica de la función para ver la aproximación obtenida junto con la solución de la ecuación diferencial, $y(t) = e^x + t^2 + 2t + 2$. La aproximación se observa en la Figura 8.

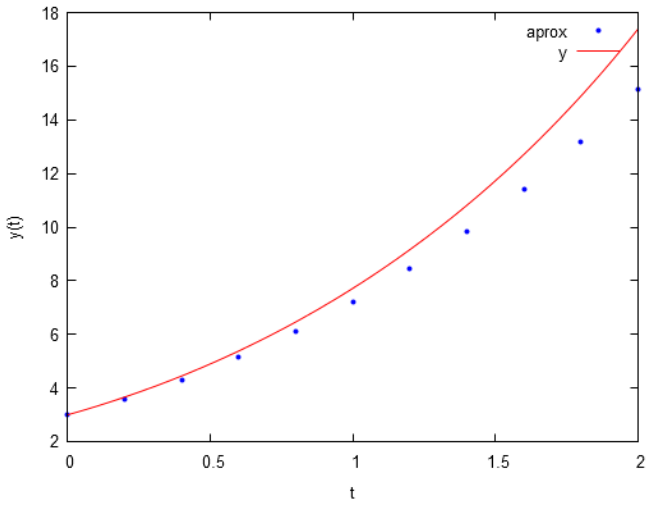

Figura 8: Aproximación a la solución con el método de Euler.

Se realiza ahora el mismo esquema anterior, pero ahora con el método del Trapecio Explícito:

j	t_j	y_j
0	0.0	3
1	0.2	3.656
2	0.4	4.3952
3	0.6	5.361014
4	0.8	6.433237
5	1.0	7.671749
6	1.2	9.095534
7	1.4	10.727752
8	1.6	12.596657
9	1.8	14.736722
10	2.0	17.190000

Tabla 5: Trapecio con $h = 0{,}2$

Se dibuja de nuevo la gráfica de la solución junto con las aproximaciones que se han obtenido.La aproximación se observa en la Figura 9.

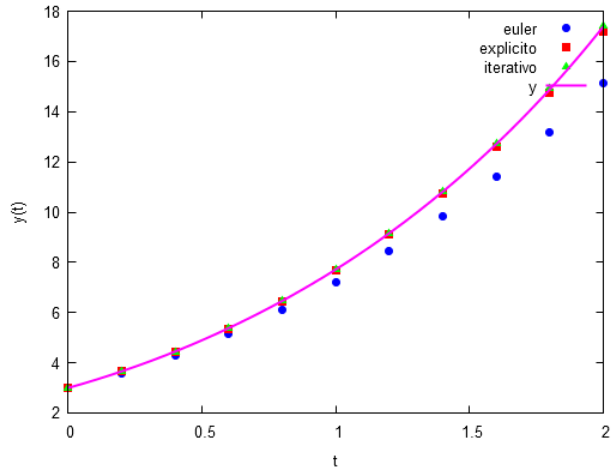

Figura 9: Aproximación a la solución con el método del Trapecio Explícito.

Como se puede observar en las dos gráficas anteriores, el método del Trapecio Explícito converge más rápidamente a la solución que el método de Euler, lo cual se debe a que el primer método es de orden 2 mientras que el de Euler es un método de orden 1.

Además, el error absoluto en el primer método es $2,235666$ obtenido mediante el valor absoluto de la diferencia de la solución calculada menos la solución evaluada en 2, esto es, $|y_{10} - y(2)|$. De la misma forma, se obtiene el error al usar el método del Trapecio Explícito. En este caso, el error cometido es $0,199054$, por lo que se puede ver que este método es mejor que el de Euler.

Por último, se va a calcular la aproximación usando el método del Trapecio Implícito. Para ello calculamos la sucesión de puntos conforme al método y se obtiene el siguiente resultado:

j	t_j	y_j
0	0.0	3
1	0.2	3.66216
2	0.4	4.453683
3	0.6	5.385540
4	0.8	6.471135
5	1.0	7.726853
6	1.2	9.172720
7	1.4	10.833211
8	1.6	12.738239
9	1.8	14.924363
10	2.0	17.436269

Tabla 6: Trapecio con $h = 0{,}2$

La aproximación se observa en la Figura 10. Al igual que se tenía con el método del trapecio explícito, que era de orden 2, se observa en 10 que se ajusta bastante bien a la gráfica de $y(t)$. El error absoluto cometido es $|y_{10} - y(2)|$ al igual que ocurría antes y en este caso es 0.047213.

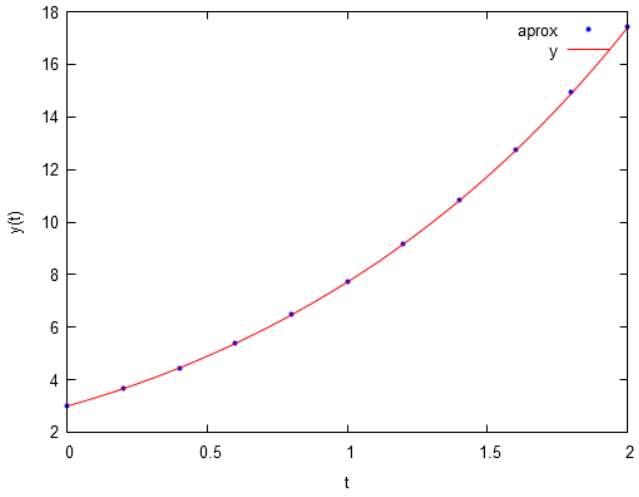

Figura 10: Aproximación a la solución con el método del Trapecio Implícito.

8.2. Ejercicio 2

Dada la ecuación $y' = t + y^2$ con $y(1) = 1$ aproximar mediante el método del trapecio: a) y(1,2) con 2 pasos (h=0,1) y b) y(1,2) con 4 pasos (h=0,05). Si el error global es de la forma Ch^2, estimar el valor de C a partir de los resultados anteriores. Determinar h para que el error sea del orden de 10^{-4}.

Solución:

j	t_{j-1}	y_{j-1}	t_j	y_j
1	1.00	2.000000	1.10	2.617500
2	1.10	2.617500	1.20	3.657368

Tabla 7: Trapecio con $h = 0{,}1$

j	t_{j-1}	y_{j-1}	t_j	y_j
1	1.00	2.000000	1.05	2.277813
2	1.05	2.277813	1.10	2.628941
3	1.10	2.628941	1.15	3.087423
4	1.15	3.087423	1.20	3.712364

Tabla 8: Trapecio con $h = 0{,}05$

Gracias a estos cálculos y como en el enunciado se nos dice que el error global es de la forma Ch^2 (lo que es coherente con el error global del método explícito) podemos escribir:

$y(1{,}2) - 3{,}657368 = C(0{,}1)^2$ $y(1{,}2) - 3{,}712364 = C(0{,}05)^2$ si restamos ambas ecuaciones y despejamos se obtiene: $C = 7{,}33$ Asumiendo entonces que el error global puede representarse mediante $7{,}33h^2$ para que sea de orden 10^{-4} debe ser $h = 3{,}7\ 10^{-3}$

8.3. Ejercicio 3

El movimiento de caída de un cuerpo de masa m en un medio que opone una resistencia proporcional al cuadrado de la velocidad está gobernado por la ecuación diferencial:

$$\frac{d^2s}{dt^2} = g - \frac{K}{m}(\frac{ds}{dt})^2 \tag{10}$$

siendo $g = 10\frac{m}{s^2}$ y $K\frac{kg}{s}$ una constante de proporcionalidad cuyo valor depende del problema concreto. Si el cuerpo se abandona sin velocidad inicial y las condiciones iniciales son

$$s(0) = s'(0) = 0 \tag{11}$$

Calcular una tabla de valores de las funciones $s(t)$ y $s'(t)$ para dibujar sus gráficas en el intervalo $[0, 1]$. Tomar $\frac{K}{m} = 5$.

Solución:

El problema que se nos propone resolver es

$$\begin{cases} s'' + 5(s')^2 - 10 = 0 \\ s(0) = s'(0) = 0 \end{cases}$$

Una formulación equivalente se obtiene haciendo $s(t) \equiv u(t)$ y $s'(t) \equiv v(t)$ de modo que se tiene el sistema:

$$\begin{cases} u' = v \\ v' = -5v^2 + 10 \\ u(0) = 0, v(0) = 0 \end{cases}$$

como el objetivo es dibujar la gráfica de las funciones no es necesaria mucha exactitud y por su sencillez en este caso es ideal el uso del método del trapecio. Tomaremos $h = 0{,}1$.

La forma que toma el método del trapecio para sistemas de dos ecuaciones es:

$$\begin{cases} u_j = u_{j-1} + \frac{1}{2}(\Delta u_0 + \Delta u_1) \\ \Delta u_0 = hf(t_{j-1}, u_{j-1}, v_{j-1}) \\ \Delta u_1 = hf(t_{j-1}, u_{j-1} + \Delta u_0, v_{j-1} + \Delta v_0) \\ v_j = v_{j-1} + \frac{1}{2}(\Delta v_0 + \Delta v_1) \\ \Delta v_0 = hg(t_{j-1}, u_{j-1}, v_{j-1}) \\ \Delta v_1 = hg(t_{j-1}, u_{j-1} + \Delta u_0, v_{j-1} + \Delta v_0) \end{cases}$$

cuya expresión será deducida en el correspondiente trabajo. La tabla de valores obtenida es la siguiente:

j	t_{j-1}	(u_{j-1}, v_{j-1})	t_j	(u_j, v_j)
1	0.00	0.000000,0.000000	0.10	0.050000,0.750000
2	0.10	0.050000,0.750000	0.20	0.160938,1.070068
3	0.20	0.160938,1.070068	0.20	0.289318,1.223146
4	0.30	0.289318,1.223146	0.40	0.424231,1.305143
5	0.40	0.424231,1.305143	0.50	0.562160,1.351168
6	0.50	0.562160,1.351168	0.60	0.701635,1.377549
7	0.60	0.701635,1.377549	0.70	0.841949,1.392822
8	0.70	0.841949,1.392822	0.80	0.982733,1.401712
9	0.80	0.982733,1.401712	0.90	1.123784,1.406900
10	0.90	1.123784,1.406900	1.00	1.264990,1.409933

Tabla 9: Trapecio para sistemas con $h = 0{,}1$

Finalmente las gráficas generadas son:

8.4. Ejercicio 4

En este caso ilustraremos con un ejemplo el algoritmo de la suma compensada de Kahan cuyo objetivo primordial es reducir el error de redondeo cuando se suma una sucesión de números en coma flotante con suma finita. Este enfoque ha sido aplicado en distintos escenarios, en particular, en el ámbito de las ecuaciones diferenciales puede consultarse "The numerical statibility in solution of differential equations"de Vitasek.

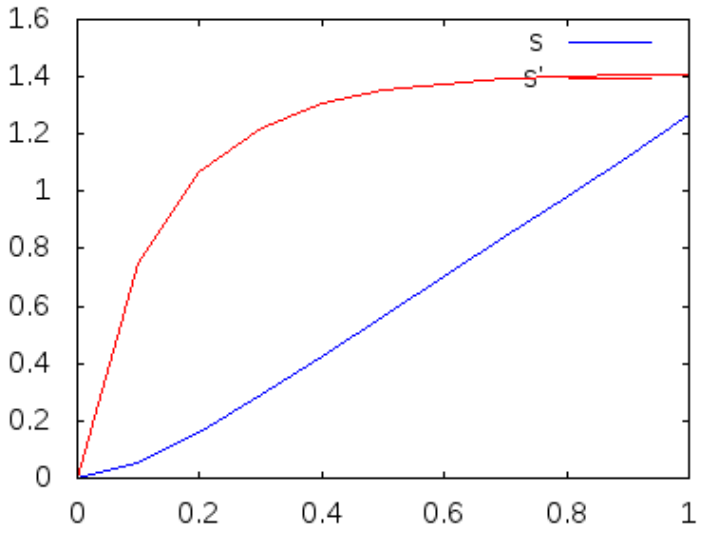

Figura 11: Representación gráfica de s y s'.

Entre las ventajas más importantes del método es que al sumar una sucesión numérica de n números se tiene un error en el caso peor que crece de manera proporcional a n con una cota que es independiente de n y sólo depende de la precisión en coma flotante.

Nuestro objetivo es presentar el algoritmo y mostrar un ejemplo de su uso.

Algoritmo 1 Algoritmo de Kahan

```
function SUMA-COMPENSADA(vector-entrada)
    suma=0.0
    c=0.0
    for i=1 to longitud(vector-entrada)  do
        (1) y = vector-entrada[i]-c
        (2) t = suma + y
        (3) c = (t - suma) - y
        (4) suma = t
    end for
     return suma
end function
```

Explicación del algoritmo:

c es una compensación para los bits de orden pequeño perdidos por redondeo
(2) Al acumular en suma, el valor de suma es grande y el de y es pequeño por lo que los dígitos de orden pequeño de y se pierden.
(3) (t-suma) cancela los dígitos de orden grande de y y restar y recupera de forma negativa los dígitos de orden pequeño de y.
(4) Teóricamente, c debería ser siempre cero pero el hecho de que el algoritmo funcione se basa precisamente en esto ya que al volver a iterar los dígitos de orden pequeños perdidos se añaden de nuevo a y.

Ejemplo:

Supóngase que se utiliza aritmética decimal de seis dígitos, que el valor de suma es 10000.0 y que los siguientes dos valores en el vector de entrada son 2.14159 y 2.71828. Resuélvase este ejemplo mediante la suma usual con redondeo y utilizando el algoritmo de Kahan.

Solución:

Notemos que el resultado exacto sería 10005.85987 que se redondea a 10005.9.

Método usual

Tras la primera suma con redondeo: 10003.1
Tras la segunda suma con redondeo: 10005.8
Este no era el resultado deseado.

Método de Kahan

Tenemos los siguientes cálculos: c = 0.0
y = 3.14159
t = 10000.0 + 3.14159 = 10003.14159 = 10003.1
c = (10003.1 - 10000.0) - 3.14159 = 3.10000 - 3.14159 = -.0415900
suma = 10003.1

Esto es, en la primera iteración coincide con el método usual pero lo importante es que la suma es tan grande que sólo los dígitos de orden elevado están siendo acumulados pero la gran diferencia es que ahora c tiene almacenada la compensación.

y = 2.71828 - -.0.415900 = 2.75987
t = 10003.1 + 2.75987 = 10005.85987 = 10005.9
c = (10005.9 - 10003.1) - 2.75987 = 2.80000 - 2.75987 = .040130
suma = 10005.9

que era el resultado deseado.

9. Conclusión

Referencias

[1] Kendall E. Atkinson. *An Introduction To Numerical Analysis.* John Wiley &Sons, Inc., 1988.
[2] R. Burden y J. Douglas Faires. *Numerical Analysis.* Thomson Learning, 2005.
[3] C. W. Gear. *Numerical initial value problems in ordinary differential equations.* Prentice Hall PTR, 1971.
[4] Timothy Sauer. *Numerical Analysis.* Pearson Education, Inc, 2006.