Métodos Numéricos II

Ecuaciones diferenciales ordinarias

Método del trapecio

Resumen

El objetivo de este trabajo es el de estudiar uno de los métodos numéricos para la resolución de ecuaciones diferenciales ordinarias: el método del Trapecio. Éste es un método de orden 2 que se suele utilizar para fines didácticos.

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:

$$y'(t)^2 + y(t)^2 + 1 = 0 \tag{1}$$

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 $y(a) = y_0$ [9]. 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 $w_0 = y_0$ [5].

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.

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 $t_i = a + ih\ \forall i = 0 \ldots n$. En el resto del texto se trabajará siempre con puntos equidistantes. El estudio

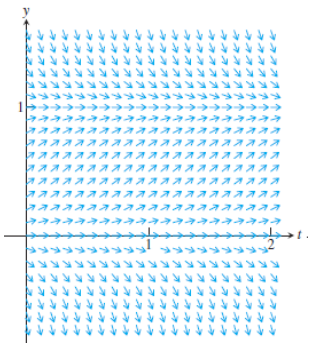

Figura 1: Representación del campo vectorial asociado a la ecuación logística $y'(t) = cy(t)(1 - y(t))$.

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.

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 $i = 0 \dots n$ cuando h tiende a 0.

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

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

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 unos 10000 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.

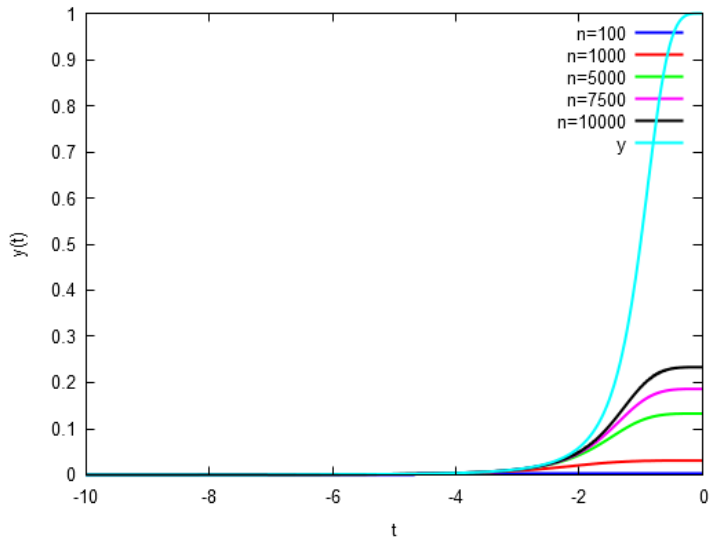

Figura 2: Aproximaciones de $y(t)$ obtenidas con el método de Euler para diferentes valores de n.

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. Además, se muestra cómo los errores de redondeo afectan al comportamiento del método del trapecio explícito. En la Sección 6 se introducen dos ejemplos que muestran el comportamiento de los métodos desarrollados. En la Sección 7 se proponen 4 ejercicios de ámbito teórico - práctico que discuten diferentes facetas del trabajo. En la Sección 8 se resume el artículo de investigación Solving Differential Equations with Constructed Neural Networks [11], 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 $M \geq 0$ tal que que $|f(t, y_1) - f(t, y_2)| \leq M|y_1 - y_2|$ para todo $(t, y_1), (t, y_2) \in \Omega$. En tal caso, se llama constante de Lipschitz de f, y se denota L, al menor M que verifica la definición anterior.

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, denotaremos por y_i a 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, consecuencia del Lema de 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. El error global g_{i+1} 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.

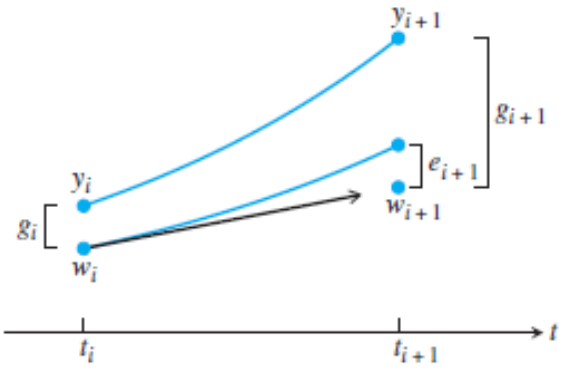

Figura 3: Representación gráfica de los errores locales y globales.

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*

$$g_i \leq \frac{Ch^k}{L}(e^{L(t_i - a)} - 1) \tag{2}$$

Demostración. Fíjese $i = 1 \dots n$. Sea z la solución de $y' = f(t, y)$ que verifica $z(t_{i-1}) = w_{i-1}$ y $z_i = z(t_i)$. Entonces:

$$g_i = |y_i - w_i| \leq |y_i - z_i| + |z_i - w_i| = |y_i - z_i| + e_i$$

La idea es acotar los dos últimos sumandos de la desigualdad. Para el primer sumando, nótese que $|y_i - z_i| = |y(t_i) - z(t_i)|$ es la diferencia de dos soluciones de la ecuación en t_i. Además, f es lipschitziana respecto de la segunda variable y se sabe que $y(t_{i-1}) = y_{t-1}$ y $z(t_{i-1}) = w_{i-1}$ por definición. Por tanto, se puede aplicar el Teorema 2.2, obteniendo

$$|y_i - z_i| \leq e^{L(t_i - t_{i-1})}|y_{i-1} - w_{i-1}| = e^{Lh} g_{i-1}$$

Además, por hipótesis $e_i \leq Ch^{k+1}$. Juntando ambas desigualdades se obtiene

$$g_i \leq e^{Lh} g_{i-1} + Ch^{k+1}$$

Esta igualdad se puede desarrollar aplicándola sobre g_{i-1}, después sobre g_{i-2}, y así sucesivamente. Teniendo en cuenta que $g_0 = 0$, se obtiene

$$g_i \leq \sum_{j=0}^{i-1} Ch^{k+1} \left(e^L H\right)^j = Ch^{k+1} \sum_{j=0}^{i-1} \left(e^L H\right)^j = Ch^{k+1} \frac{\left(e^{Lh}\right)^i - 1}{e^{Lh} - 1}$$

En el último término se aplica $x + 1 \leq e^x$, esto es, $e^{Lh} - 1 \geq Lh$, para obtener la igualdad deseada:

$$g_i \leq Ch^{k+1} \frac{\left(e^{Lh}\right)^i - 1}{e^{Lh} - 1} \leq Ch^{k+1} \frac{e^{iLh} - 1}{Lh} = \frac{e^{L(t_i - a)} - 1}{L} Ch^k$$

□

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 $i = 0 \ldots n$ cuando h tiende a cero.
2. El método es de orden k si existe una constante $C \geq 0$ tal que $g_i \leq Ch^k$ para todo $i = 0 \ldots n$ cuando h tiende a cero.

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 C^2 y lipschitziana respecto de 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 $i = 1 \ldots n$ y sea z la solución del problema de valores iniciales para $z(t_{i-1}) = w_{i-1}$. El teorema de Taylor para orden 2 proporciona la siguiente igualdad para cualquier

$$z_i = w_{i-1} + hf(t_{i-1}, w_{i-1}) + \frac{h^2}{2} z''(\xi_i) = w_i + \frac{h^2}{2} z''(\xi_i) \tag{3}$$

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

$$e_i = |z_i - w_i| = \left| \frac{h^2}{2} z''(\xi_i) \right| \leq \frac{M_i}{2} h^2 \tag{4}$$

donde $M_i = \text{máx}\{z''(t) : t \in [t_{i-1}, t_i]\}$. Tomando $M = \text{máx}_{i=1\ldots n} M_i$, 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.

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 un paso [6]. 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. De aquí en adelante se hablará únicamente de este tipo de métodos, pues son los más extendidos y utilizados.

Una propiedad que debe interesar al estudiar un método de un paso es que este sea convergente.

Definición 2.4. Un método de un paso se dice convergente respecto a la ecuación diferencial que aproxima (con función f lipschitziana con respecto a la segunda variable) si:

$$\lim_{n\to+\infty} \max_{i=0\ldots n} g_i = 0$$

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 está uniformemente acotado por un polinomio en función de h. 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.

Otro concepto importante es la consistencia:

Definición 2.5. Un método de un paso se dice consistente con respecto a la ecuación diferencial que aproxima (con función f lipschitziana con respecto a la segunda variable) si:

$$\lim_{n\to+\infty} \max_{i=0\ldots n} e_i = 0$$

Esto es, los errores locales convergen uniformemente a 0 cuando $n \to +\infty$.

Nótese que si un método es localmente de orden r con $r > 0$, entonces el método es consistente ya que los errores locales están uniformemente acotados por un polinomio que depende de h.

Cuando se considera un problema de valores iniciales:

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

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:

$$\begin{cases} y' = f(t, y) \\ y(a) = y_0 + \epsilon_0 \end{cases}$$

donde a ϵ_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. Si se considera el error cometido en $t_n = b$, este puede incluso diverger cuando $n \to +\infty$. 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, las diferencias 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 2.6. Un método de un paso 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 $K|y_0 - y_0'|$ para cualquier h lo suficientemente pequeño. El siguiente resultado permite asegurar la estabilidad de un método de un paso bajo determinadas condiciones.

Teorema 2.5. *Si un método de un paso verifica que (para cualquier PVI con f lipschitziana respecto de la segunda variable) se tiene que ϕ es continua en cada una de sus variables y, además, es lipschitziana respecto de la segunda variable en el dominio $\Omega \times [0, h_0]$, donde Ω es el dominio de f, entonces:*

1. *el método es estable.*
2. *el método es convergente (o, equivalentemente, $\phi(t, y, 0) = f(t, y)$ para todo $t \in [a, b]$), si , y solo si, el método es consistente.*

Demostración. Se referencian los libros en los que se encuentra la prueba:

1. Puede encontrarse en los ejercicios resueltos de la sección 5.10 del libro de Burden-Faires [2].
2. Puede encontrarse en la sección 4.3 del libro de Gear [6].

□

Como consecuencia, se obtiene el siguiente corolario para el método de Euler:

Corolario 2.6. *El método de Euler es estable, consistente y convergente.*

Demostración. En el caso del método de Euler $\phi(t, y, h) = f(t, y)$. Por tanto, se cumplen las hipótesis del Teorema 2.5. Además, el orden del método de Euler es 1, luego es convergente y, por tanto, consistente. □

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

$$y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))) \; ds \tag{5}$$

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 de clase 3. En tal caso y es de clase 4. Por tanto, se puede utilizar la fórmula del trapecio para integración numérica, obteniendo la siguiente igualdad

$$y(t_1) = y_0 + \frac{h}{2}\left[f(t_0, y_0) + f(t_1, y(t_1))\right] + O(h^3) \tag{6}$$

Ignorando el último sumando se obtiene la aproximación dada en (7), que tiene error $O(h^3)$.

$$y(t_1) \approx w_1 = w_0 + \frac{h}{2}\left[f(t_0, w_0) + f(t_1, y(t_1))\right] \tag{7}$$

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 implícito 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 (7) 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:

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

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.

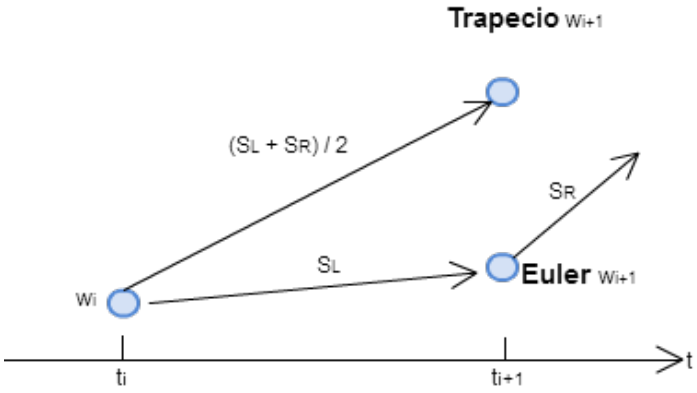

Figura 4: Esquema visual del método del trapecio explícito en contraposición con el método de Euler.

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 4.2 se estudiará el comportamiento del método ante los errores de redondeo y, por último, en la Sección 4.3 se estudiará la estabilidad.

4.1. Error local y global. Convergencia

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 tanto, aplicando la igualdad (6) 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 $j - 1 \equiv 0$ y $j \equiv 1$ 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 $O(h^2)$. □

Como consecuencia, el método del trapecio explícito es convergente.

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. [4]

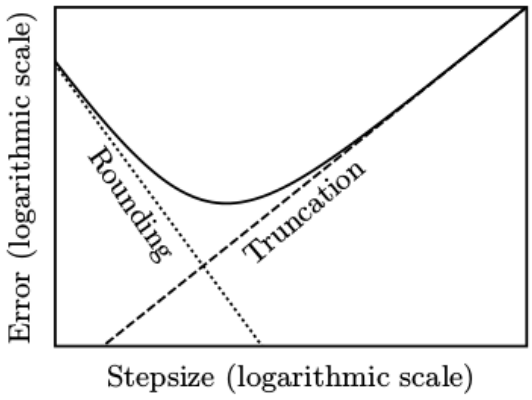

Figura 5: Combinación de los errores de truncatura y redondeo

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

$$n(\epsilon + Ch^3) = \frac{\epsilon}{h} + Ch^2$$

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.[3]

4.3. Estabilidad y convergencia

El estudio de la estabilidad y convergencia del método del trapecio explícito es un corolario del Teorema 2.5. En este caso

$$\phi(t, y, h) = \frac{1}{2}f(t, y) + \frac{1}{2}f(t + h, y + hf(t, y))$$

Asumiendo que f es lipschitziana respecto de la segunda variable 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 se puede aplicar el Teorema 2.5 y se tiene 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.

Además, el método del trapecio explícito es convergente y, por tanto, $\phi(t, y, 0) = \frac{1}{2}f(t, y) + \frac{1}{2}f(t, y) = f(t, y)$ y el método es consistente.

5. Método del trapecio implícito

Recuérdese en este punto la aproximación (7), 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 $i = 1 \dots n$ se está resolviendo la siguiente ecuación implícita

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

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 (9). 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 $O(h^3)$. Consecuentemente, el método es localmente de orden 3.*

Demostración. Basta aplicar la igualdad (6), tomando como condición inicial $z(t_{i-1}) = w_{i-1}$, junto con la ecuación implícita (9):

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

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

$$e_i \leq \frac{2}{2 - hL}O(h^3)$$

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 cuadrático.

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

$$w_i^{(j+1)} = g_i(w_i^{(j)}) = w_{i-1} + \frac{h}{2}\left[f(t_{i-1}, w_{i-1}) + f(t_i, w_i^{(j)})\right] \tag{10}$$

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 (9) 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 (10). 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 (10) 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. Ejemplos

6.1. 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 y con el método del Trapecio Explícito e Iterativo, con paso 10^{-3}. se obtienen la siguiente gráfica. El método del Trapecio Iterativo se ha ejecutado con una tolerancia de $h^4 = 10^{-12}$ para el método de iteración funcional y tomando como aproximación inicial el método de Euler.

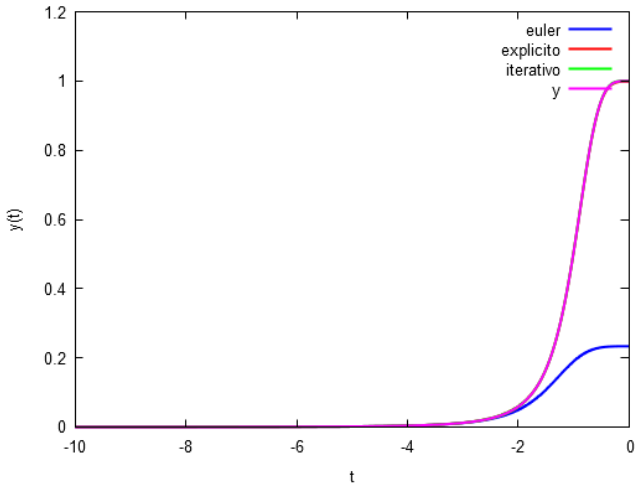

Figura 6: Aproximación a la solución con los distintos métodos.

6.2. 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.

7. Ejercicios teórico-prácticos

7.1. 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 e Iterativo.

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 7.

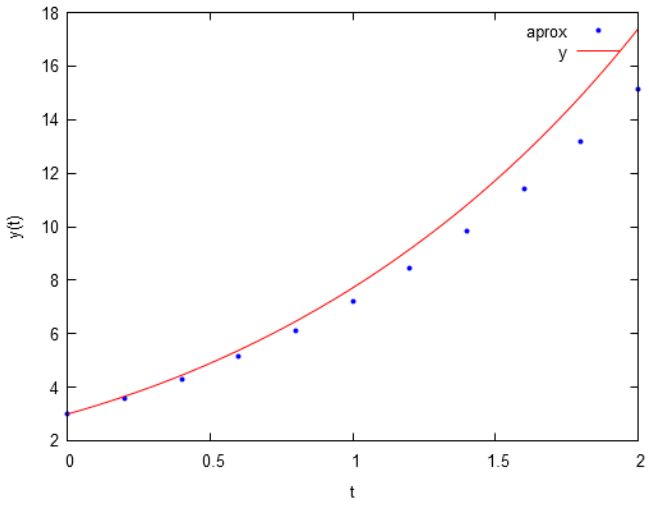

Figura 7: Aproximación a la solución con el método de Euler.

Se realiza ahora el mismo esquema anterior, pero ahora con el método del Trapecio Explícito e Iterativo, este último con la misma configuración que en el Ejemplo 1. Además, se comparan los resultados con el método de Euler:

j	t_j	$y_j Explicito$	$y_j Implicito$	$y_j Euler$
0	0.0	3	3	3
1	0.2	3.656	3.66216	3.6
2	0.4	4.3952	4.453683	4.312
3	0.6	5.361014	5.385540	5.1424
4	0.8	6.433237	6.471135	6.09888
5	1.0	7.671749	7.726853	7.190656
6	1.2	9.095534	9.172720	8.428787
7	1.4	10.727752	10.833211	9.826544
8	1.6	12.596657	12.738239	11.399853
9	1.8	14.736722	14.924363	13.167824
10	2.0	17.190000	17.436269	15.153389

Tabla 5: Tabla comparativa 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 8.

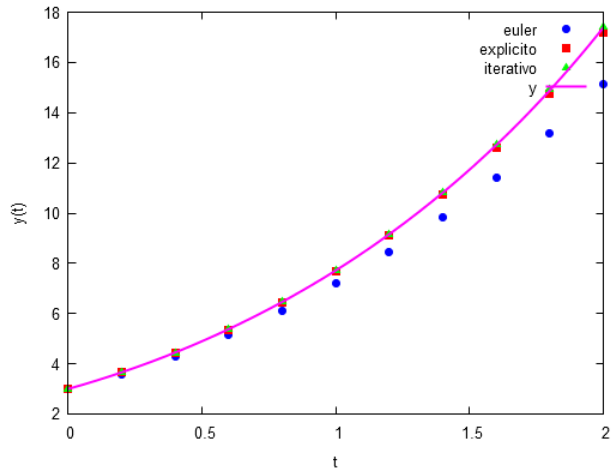

Figura 8: Aproximación a la solución con los métodos.

Como se puede observar en las dos gráficas anteriores, el método del Trapecio Explícito e Iterativo convergen más rápidamente a la solución que el método de Euler, lo cual se debe a que son 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. Análogamente, se calcula el error para el método del Trapecio Iterativo que es 0.047213.

7.2. 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 6: 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 7: 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 $h = 3{,}7\ 10^{-3}$

7.3. 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:

$$\frac{d^2s}{dt^2} = g - \frac{K}{m}(\frac{ds}{dt})^2 \tag{11}$$

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

$$s(0) = s'(0) = 0 \tag{12}$$

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

Finalmente las gráficas generadas son:

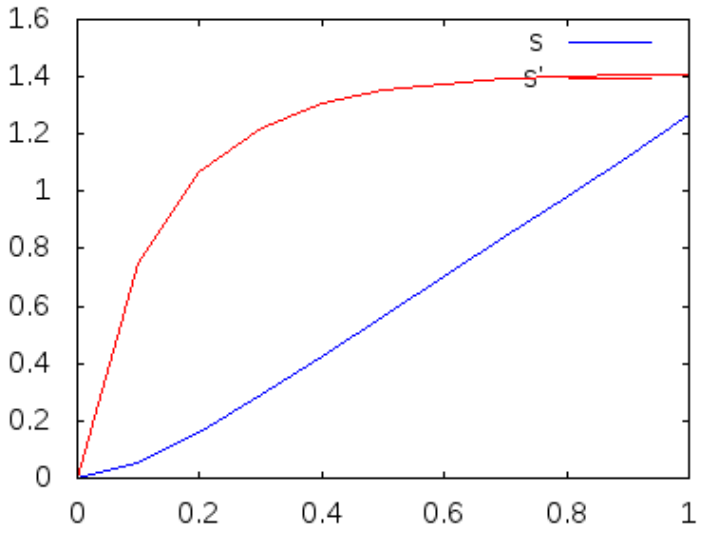

Figura 9: Representación gráfica de s y s'.

7.4. 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 en [12].

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.

8. Artículo de investigación

En esta sección se expone el artículo Solving Differential Equations with Constructed Neural Networks [11]. De esta forma se pone de manifiesto que el desarrollo de métodos numéricos para aproximar soluciones de ecuaciones diferenciales sigue siendo un tema abierto en la actualidad.

En este trabajo se combinan los métodos numéricos con la inteligencia artificial. En concreto, el artículo presenta un método innovador para la resolución de ecuaciones diferenciales ordinarias, sistemas de ecuaciones diferenciales y ecuaciones diferenciales en derivadas parciales. Este método evoluciona una

población de redes neuronales mediante la heurística grammatical evolution con el fin de obtener una red neuronal final que aproxime la solución de la ecuación diferencial. A continuación se explica el funcionamiento del método para ecuaciones diferenciales ordinarias de un orden arbitrario.

El objetivo del método es encontrar una red neuronal que aproxime a la solución del problema de valores iniciales. Una red neuronal es una función $N : \mathbb{R}^d \to \mathbb{R}^{d'}$ que se define a partir de los siguientes elementos:

1. Un grafo dirigido y ponderado construido a partir de $c+1$ capas de nodos. Los arcos que salen de los nodos de la i-ésima capa solamente van a los nodos de la capa $i+1$. La primera capa de nodos se denomina input layer o capa de entrada y adquiere el índice 0. El número de nodos en esta capa es d y cada uno de estos nodos representa una variable de entrada de la función. La última capa se denomina output layer o capa de salida. Esta capa consta de d' nodos, que se corresponden con cada una de las variables de salida de la función. El resto de capas se denominan hidden layers o capas ocultas. La Figura 1 muestra un ejemplo de grafo asociado a una red neuronal donde solo hay una capa oculta. Nótese que el número de nodos en cada capa puede ser distinto. El número de nodos de la i-ésima capa se denota n_i. El peso del arco que une el nodo j de la capa i con el nodo k de la capa $i+1$ se denota $w_{jk}^{(i)}$. En caso de que no exista un arco que una ambos nodos, el peso puede considerarse 0.

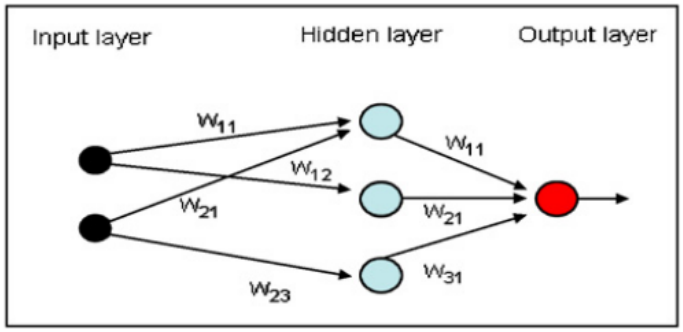

Figura 10: Red neuronal con una capa oculta.

2. Una función $\phi : \mathbb{R} \to \mathbb{R}$ denominada función de activación.
3. Un valor asociado a cada nodo, denominado bias. El bias del k-ésimo nodo de la capa i-ésima se denota $b_k^{(i)}$

Para cada nodo se define una función, denotada $N_k^{(i)}$ para el k-ésimo nodo de la i-ésima capa, de forma recursiva:

$$\begin{cases} N_k^{(0)}(x_1, x_2, \ldots, x_d) = x_k \\ N_k^{(i+1)}(x_1, x_2, \ldots, x_d) = \phi(b_k^{(i)} + \sum_{j=1}^{n_i} w_{jk}^{(i)} N_j^{(i)}(x_1, x_2, \ldots, x_d)) \end{cases} \tag{13}$$

La función N se define como $(N_1^{(c)}, N_2^{(c)}, \ldots, N_{d'}^{(c)})$, esto es, la salida de la última capa. La definición de esta función imita a la estructura del cerebro, de ahí su nombre. La Figura 8 muestra de forma visual la función $N_k^{(i)}$ definida para un nodo que no sea de la capa de entrada.

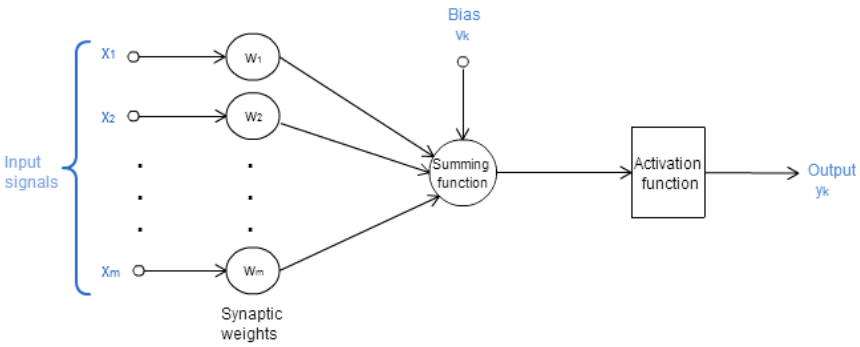

Figura 11: Función definida para un nodo de la red neuronal.

Las redes neuronales se utilizan para aproximar funciones de las cuales no se conoce su expresión. Por ejemplo, una de las aplicaciones de las redes neuronales es la resolución de problemas de regresión o, en este caso, la aproximación de soluciones de ecuaciones diferenciales. Para poder conseguir aproximar una función mediante una red neuronal es necesario entrenarla, esto es, modificar los pesos y bias para que la salida de la red neuronal sea la que se quiere. Entrenar una red neuronal es un problema de optimización. Hay que encontrar los parámetros para los cuales se minimiza el error cometido por la red neuronal. La idea para resolver este problema de optimización es fijar una entrada para la red neuronal y ver el error cometido sobre dicha entrada como una función de sus parámetros (pesos y bias). Si el error es diferenciable, entonces se puede utilizar el método del descenso del gradiente para minimizarlo. Para que el error sea diferenciable hay que tomar una función de activación que también lo sea. En la literatura especializada lo más habitual es utilizar la función sigmoide, $\phi(x) = \frac{1}{1+e^{-x}}$.

Uno de los algoritmos que utilizan el descenso del gradiente para optimizar los parámetros de una red neuronal es backpropagation. Este suele ser el método estándar para entrenar redes neuronales.

En el caso particular de las ecuaciones diferenciales ordinarias, la función a aproximar es una función real de una variable real. Por tanto, se busca una red con un único nodo en la capa de entrada y en la capa de salida. La ecuación diferencial ordinaria a resolver se puede escribir como sigue:

$$f(t, y, y^{1)}, y^{2)}, \ldots, y^{r)}) = 0, t \in [a, b] \tag{14}$$

Puesto que el orden de la ecuación diferencial es r, se requieren r condiciones iniciales. Estas se pueden escribir de la siguiente forma:

$$\Psi_i(t, y, y^{1)}, y^{2)}, \ldots, y^{r-1)}) = 0 \ \forall \ i = 1, 2, \ldots, r \tag{15}$$

La red neuronal se entrena minimizando el error cometido en las condiciones iniciales y el error cometido en la ecuación (14). La fórmula del error utilizada para las condiciones iniciales es la siguiente:

$$\sum_{i=1}^{r} \Psi_i(t_0, y(t_0), y^{1)}(t_0), \ldots, y^{r-1)}(t_0))^2 \tag{16}$$

Para medir el error cometido en la ecuación (14), se consideran $n+1$ puntos equidistantes, $t_i = a+ih \ \forall i = 0, 1, \ldots, n$ donde $h = \frac{b-a}{n}$. El error cometido por la red neuronal en (14) se define únicamente sobre

estos $n+1$ puntos:

$$\sum_{i=0}^{n} f(t_i, N(t_i), N^{1)}(t_i), \ldots, N^{r)}(t_i))^2 \tag{17}$$

El error total de aproximación que se pretende minimizar se mide como una media ponderada de los errores (16) y (17).

Habitualmente este tipo de problemas se resuelve definiendo de antemano el grafo dirigido asociado a la red neuronal y optimizando posteriormente el resto de los parámetros. Sin embargo, este procedimiento conlleva una serie de pruebas hasta dar con el grafo dirigido apropiado. Por tanto, los autores del artículo deciden utilizar un algoritmo evolutivo denominado grammatical evolution que optimiza la estructura del grafo dirigido utilizado en la red neuronal [10].

Grammatical evolution es el nombre dado a la aplicación de algoritmos genéticos [7] a problemas de optimización en los que las soluciones del problema son palabras generadas por una gramática libre de contexto [8]. Una palabra generada por la gramática libre de contexto se representa por un vector de enteros. Este vector indica qué reglas hay que utilizar para derivar la palabra a partir del símbolo inicial. La primera componente del vector indica cuál de las reglas que substituye el símbolo inicial se debe aplicar. Tras su uso, se busca la variable más a la izquierda de la palabra obtenida y se aplica la regla indicada por la segunda componente del vector. Este proceso se repite hasta obtener la palabra.

Los autores consideran solamente redes neuronales con una única capa oculta con el fin de que la red neuronal se pueda representar mediante una palabra de una gramática libre de contexto. De esta forma, la población del algoritmo genético se puede inicializar con representaciones de redes neuronales aleatorias obtenidas por la gramática libre de contexto escogida. El valor objetivo de un elemento de la población es el error que comete la red neuronal asociada como aproximación de la solución de la ecuación diferencial. La población del algoritmo genético evoluciona mediante el uso de los operadores de selección, cruce y mutación como es habitual en la literatura especializada. Cada cierto número de iteraciones del algoritmo genético, se escogen varias redes neuronales de la población y se mejoran mediante backpropagation u otro algoritmo de entrenamiento. El método finaliza cuando se verifica determinado criterio elegido por el usuario. Este criterio puede ser número de iteraciones del algoritmo genético o cuando se consiga una red neuronal cuyo error sea menor que un determinado umbral. La mejor red neuronal obtenida durante el proceso es la función que devuelve el método.

La Figura 8 muestra la imagen de la red neuronal calculada para el PVI dado por $y' = \frac{2x-y}{x}$ y $y(1) = 3$. Se puede observar que la red neuronal aproxima casi perfectamente a la solución del problema.

Una de las ventajas de este método es que el resultado es una función, al contrario de lo que sucede con los métodos de discretización. Además, se pueden realizar tantas iteraciones del algoritmo genético como se quiera, mejorando todavía más la aproximación obtenida. Sin embargo, el problema subyacente en este tipo de métodos es la dificultad para obtener resultados teóricos que proporcionen cotas del error o aseguren la convergencia y la estabilidad del método. La mayoría de las heurísticas estudiadas en inteligencia artificial son intuitivas y proporcionan buenos resultados en la práctica pero carecen de fundamento teórico. Este es el caso de los algoritmos genéticos y, por tanto, del método que presenta el artículo.

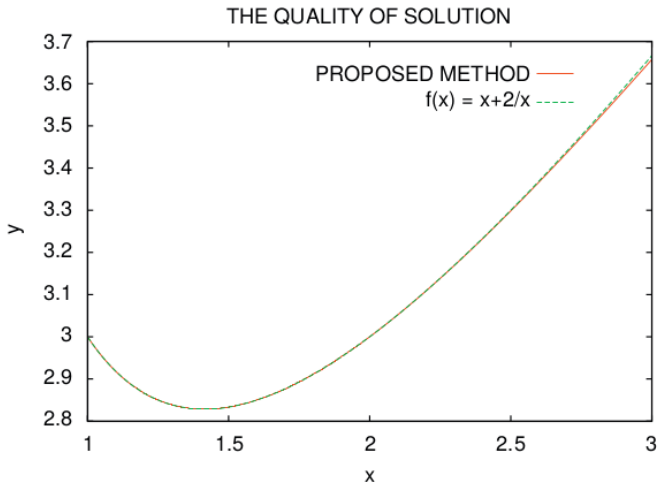

Figura 12: Solución del PVI $y' = \frac{2x-y}{x}$ con condición inicial $y(1) = 3$ y red neuronal obtenida por el método.

9. Conclusión. Ventajas y desventajas del método del trapecio

En este trabajo se han desarrollado las versiones explícitas e implícitas del método del trapecio. Estos métodos han permitido resolver de forma aproximada problemas de valores iniciales para los cuales el método de Euler no obtenía resultados satisfactorios (véase la Figura 6). Tras los estudios realizados, se pueden destacar las siguientes ventajas:

- El método del trapecio es secillo y puede deducirse como una aplicación de las fórmulas de integración numérica. Además, el método del trapecio explícito puede entenderse como una mejora del método de Euler como se puntualiza en la Figura 4.
- El número de evaluaciones de la función f que se requiere para ejecutar el método del trapecio explícito es $2n$. Por tanto, un método eficiente si la evaluación de la función f es muy costosa. En el caso del método del trapecio implícito, este número de evaluaciones dependerá del método numérico utilizado para obtener puntos fijos. Sin embargo, se pueden controlar según se desee.
- El orden del método es 2 y, por tanto, se pueden resolver problemas de valores iniciales con un error aceptable.
- El método del trapecio es un ejemplo de método de discretización para el cuál surgen dos variantes, la explícita y la implícita. Por tanto, tiene una gran utilidad didáctica. Permite poner en práctica casi todo el temario de la asignatura en un único método.

Sin embargo, el método del trapecio no suele ser utilizado en la práctica [3] ya que existen métodos más complejos que presentan un mayor orden. Este es el caso del método de Runge-Kutta de cuarto orden. Hay que decir en este momento que el método del trapecio explícito puede verse como un caso particular de los métodos de Runge-Kutta por lo que también es útil para introducir éstos.

En definitiva, el método del trapecio es un método sencillo y permite introducir los conceptos presentes

en la resolución de PVIs mediante métodos de discretización sin necesidad de recurrir a métodos más complejos.

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] J.C. Butcher. *Numerical Methods for Ordinary Differential Equations.* John Wiley & Sons, 2008.

[4] Carlos Vázquez Espí. *Análisis Numérico.* García Maroto Editores S.L., 2013.

[5] M. Gasca. *Cálculo Numérico.* UNED, 1991.

[6] C. W. Gear. *Numerical initial value problems in ordinary differential equations.* Prentice Hall PTR, 1971.

[7] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, 1989.

[8] J.E. Hopcroft y Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley series in computer science, 1979.

[9] Timothy Sauer. *Numerical Analysis.* Pearson Education, Inc, 2006.

[10] I. Tsoulos, D. Gavrilis y E. Glavas. «Neural network construction and training using grammatical evolution». En: *Neurocomputing* (2008), págs. 269-277.

[11] Ioannis Tsoulos, Dimitris Gavrilis y Euripidis Glavas. «Solving differential equations with constructed neural networks.» En: *Neurocomputing* (2009), págs. 2385-2391.

[12] E. Vitasek. «The numerical stability in solution of differential equations». En: *Conference on Numerical Solution of Differential Equations, Springer* (1969).