Métodos Numéricos

Resolución de Sistemas de Ecuaciones Diferenciales Ordinarias

Roqueta, Matías Daniel

Centro Atómico Bariloche y Instituto Balseiro, Comisión Nacional de Energía Atómica

Resumen

Se realizó un estudio de resolución numérica de ecuaciones diferenciales correspondientes a órbitas del problema de dos cuerpos, comparando soluciones numéricas obtenidas contra la solución analítica conocida del problema.

El estudio se centró en los métodos elementales θ, en particular, los métodos de Euler explícito e implícito y el método de Crank-Nicolson. Los experimentos realizados ponen en evidencia la ventajas de usar el método de Crank-Nicolson para obtener una mayor precisión a un costo de un incremento admisible en complejidad algorítmica.

Habiendo obtenido estos resultados se comparó el algoritmo de Crank-Nicholson desarrollado contra el algoritmo de Adams optimizado provisto por **lsode**, encontrando situaciones para la que el algoritmo de Crank-Nicholson no es el indicado. En particular, se encontró un problema rígido que algoritmo escrito no puede resolver eficientemente.

Introducción

El objeto central del estudio de métodos numéricos para problemas de valor inicial es la ecuación diferencial de primer órden, expresada simplemente como

$$y' = f(y, t) \tag{1}$$

Si se obtienen métodos para resolver la ecuación 1, y se extiende el dominio a $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, se tienen métodos capaces de resolver sistemas de n ecuaciones diferenciales de primer órden. Por la equivalencia 2 esto contempla, además, sistemas que incluyen ecuaciones diferenciales de mayor orden

$$y^{[k]} = f(y^{[k-1]} \dots y, t) \equiv \begin{vmatrix} y_1' = y_2 \\ \vdots \\ y_k' = f(y_{k-1} \dots y_1, t) \end{vmatrix} \tag{2}$$

por lo que un método que resuelva la ecuación 1 en su expresión vectorial es extremadamente versátil.[1]

Los métodos numéricos más elementales surgen de discretizar el intervalo temporal sobre el cual se evalúa la función y en pasos de tamaño h, asumiendo en cada iteración $y_n = y(t_n)$ conocida para aproximar $y*_{n+1} \simeq y(t_{n+1})$ a primer órden.

$$y_{n+1} \simeq y_n + h y_n' \tag{3}$$

Para aproximar y_n' se puede tomar el valor conocida de la misma en el paso n, el valor predictivo en el paso $n+1$, o una combinación de ambos, la generalización de estos métodos es el método θ

$$y_{n+1} \simeq y_n + h\left[\theta f(y_{n+1},\, t_{n+1}) + (1-\theta)\, f(y_n,\, t_n)\right] \tag{4}$$

La elección del parámetro θ del método afecta el órden de error del método, pero también la complejidad algorítmica del mismo.

En este experimento se toma bajo estudio el sistema de dos cuerpos sometidos a fuerzas gravitacionales

$$\begin{aligned} \frac{d^2\vec{r}_1}{dt^2} &= GM_2 \frac{\vec{r}_2(t) - \vec{r}_1(t)}{\left|\vec{r}_1(t) - \vec{r}_2(t)\right|^3} \\ \frac{d^2\vec{r}_2}{dt^2} &= GM_1 \frac{\vec{r}_1(t) - \vec{r}_2(t)}{\left|\vec{r}_1(t) - \vec{r}_2(t)\right|^3} \end{aligned} \tag{5}$$

Este sistema se puede expresar en función de ecuaciones diferenciales de primer orden. Si se define el factor

$$\kappa_x = \frac{GM_x}{\left|\vec{r}_1(t) - \vec{r}_2(t)\right|^3}$$

Entonces el sistema físico se puede expresar con la siguiente ecuación matricial

$$\frac{d}{dt}\begin{bmatrix}\vec{r}_1 \\ \vec{v}_1 \\ \vec{r}_2 \\ \vec{v}_2\end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\kappa_2 & 0 & \kappa_2 & 0 \\ 0 & 0 & 0 & 1 \\ \kappa_1 & 0 & -\kappa_1 & 0 \end{bmatrix}\begin{bmatrix}\vec{r}_1 \\ \vec{v}_1 \\ \vec{r}_2 \\ \vec{v}_2\end{bmatrix} \tag{6}$$

[1] Un ejemplo de la aplicación en el área de telecomunicaciones es el modelo de variables de estado de un sistema lineal invariante en el tiempo. Este modelo y su aplicación para el análisis de filtros se trata en el capítulo 2 de Señales y Sistemas de S. Haykin. [1]

Sobre los elementos de este sistema vamos a definir un método θ, detallado en el apéndice 1. A partir de este método se obtienen directamente las expresiones de los métodos de Euler explícito, Euler implícito, y de Crank-Nicolson.

Para estudiar la precisión del método se compara la solución numérica con la solución analítica para parámetros $M_1 = M_2$ y condiciones iniciales

$$\vec{r}_1(t_0) = \begin{bmatrix} 1 & 0 \end{bmatrix} \qquad \vec{r}_2(t_0) = \begin{bmatrix} -1 & 0 \end{bmatrix}$$
$$\vec{v}_1(t_0) = \begin{bmatrix} 0 & 1 \end{bmatrix} \qquad \vec{v}_2(t_0) = \begin{bmatrix} 0 & -1 \end{bmatrix}$$

La solución analítica para el sistema con estas condiciones son órbitas circulares de radio 1 y período 2π, se evalúa el error de cada método luego de una órbita.

Ya que los métodos se están usando para estudiar la evolución temporal de un sistema físico, es de interés estudiar si las soluciones numéricas preservan las leyes de la física como la solución analítica lo hace.

Particularmente, se estudia la conservación de la energía, evaluando la evolución de la energía mecánica del sistema en cada paso, dada por la ecuación 7.

$$E(t) = -\frac{2\,G\,M_1 M_2}{|\vec{r}_1(t) - \vec{r}_2(t)|} + \frac{1}{2}\left(M_1\,|\vec{v}_1(t)|^2 + M_2\,|\vec{v}_2(t)|^2\right) \tag{7}$$

Finalmente, se compara el rendimiento de los métodos θ, de un paso, contra el rendimiento de los métodos multipaso. Para resolver el sistema con métodos multipaso se usa la función **lsode** de Octave para invocar un método Adams.

Durante el experimento se estudia el error global de los métodos. En el experimento, esto se define como la desviación de la posición final de la solución numérica respecto a la posición final de la solución analítica.

$$e^g(h) = \sum_{i\,\in\{1,\,2\}} \left\|\vec{r}_i(t_N) - \vec{r}_{i_{An}(t_N)}\right\| \tag{8}$$

Resultados

Para estudiar el error de cada método se resuelve el problema orbital dividiendo el intervalo $t = [0,\, 2\pi]$ en 100 subdivisiones, por lo que $h \simeq 0,63$. Los tres métodos se implementan con las ecuaciones matriciales 12, 13, y 14 del apéndice 1.

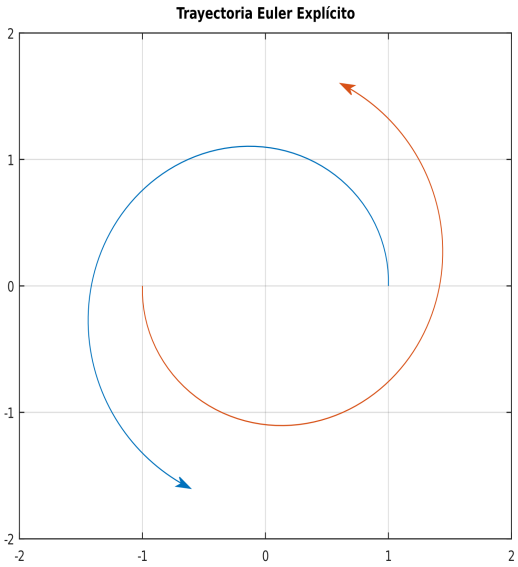

Figura 1: Trayectoria del sistema resuelto numéricamente con el método de Euler explícito y $h \simeq 0,63$

Se observa en la figura 1 que el método de Euler explícito da una solución divergente del sistema, dejando en evidencia un error por exceso en cada paso temporal.

Además, se puede ver que el período orbital de la solución numérica es mayor al período de la solución analítica.

Error global de Euler explícito: $e^g(0,53) = 4,46$

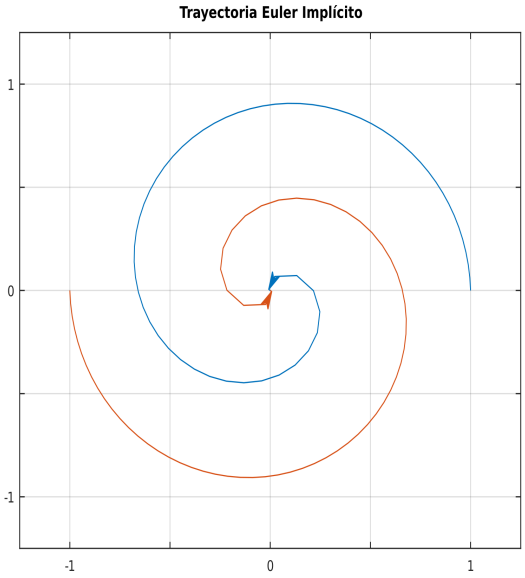

Figura 2: Trayectoria del sistema resuelto numéricamente con el método de Euler implícito y $h \simeq 0,63$.

Por el contrario la figura 2 muestra que el método de Euler implícito da una solución que convergería a 0, implicando un error por defecto en cada paso.

Además, se puede ver que el período orbital de la solución numérica es menor al período de la solución

analítica.

Error global de Euler implícito: $e^g(0,63) = 2$

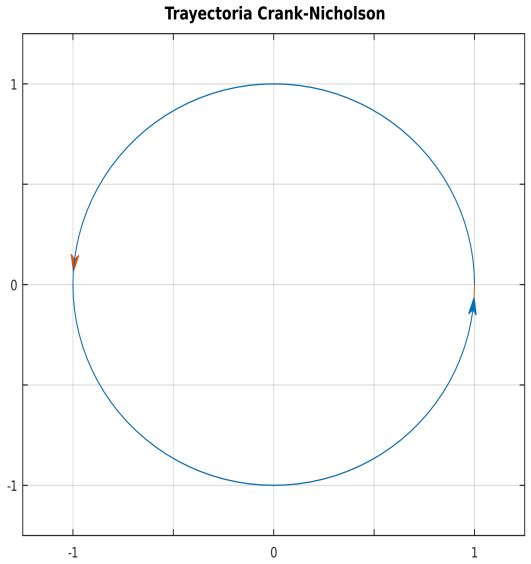

Figura 3: Trayectoria del sistema resuelto numéricamente con el método de Crank-Nicolson y $h \simeq 0,63$

La figura 3 muestra que la trayectoria del método de Crank-Nicolson se mantiene circular, lo que corresponde a la solución analítica del sistema.

Error global de Crank-Nicolson: $e^g(0,63) = 0,002$

Para estudiar como reacciona el sistema para otros valores de h, se repite el mismo experimento incrementando el número de subdivisiones.

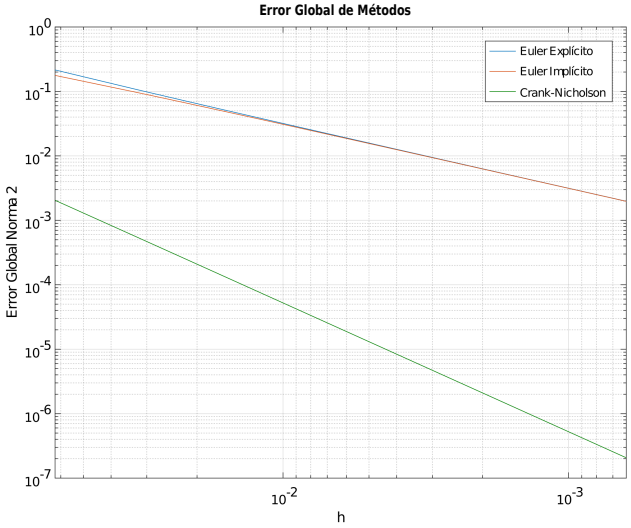

Figura 4: Disminución del error global de cada método por disminución del paso temporal h.

La figura 4 muestra como depende el error de los métodos de h, reflejando su órden de convergencia en la pendiente de la recta en escala loglog.

Específicamente, la convergencia del método de Crank-Nicolson es $O(h^2)$, del Euler implícito $O(h)$, y del Euler explícito, si converge, también $O(h)$.

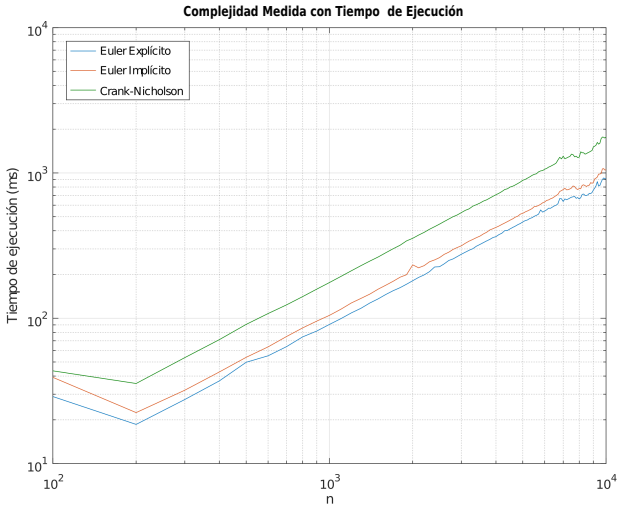

Figura 5: Incremento del tiempo de ejecución de cada método por incremento del número de puntos de evaluación.

La precisión del método corresponde directamente a la complejidad del método por la relación $h = \frac{t_n - t_o}{n}$, por lo que es de interés la complejidad algorítmica de los métodos para evaluar el costo de aumentar la precisión.

Aplicando ajuste polinómico a los datos de la figura 5 se puede determinar que la complejidad algorítmica de los tres métodos en función del número de subdivisiones del dominio es $O(n)$.

El estudio de la conservación de la energía se realiza para $n = 100$ y nuevamente para $n = 6000$. Se evalúa la ecuación 7 en cada instante temporal.

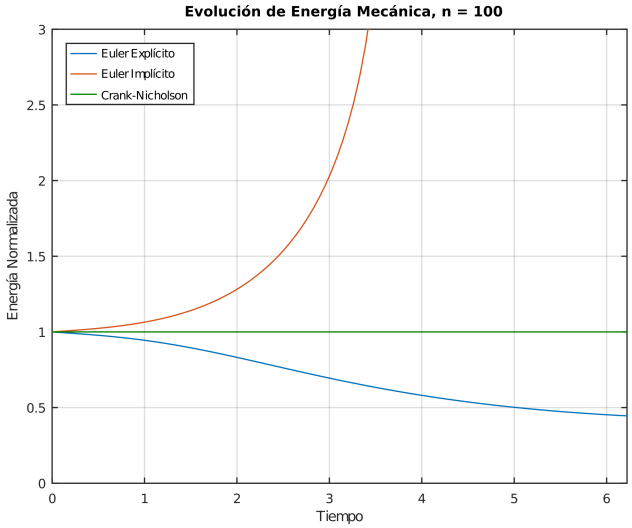

Figura 6: Evolución temporal de la energía mecánica del sistema de 2 cuerpos resuelto numéricamente en 100 pasos.

La figura 6 muestra que solo se conserva la energía si

el método converge a la solución analítica, en particular, el Euler implícito tiende a energía, infinita, que es consistente con el comportamiento de la figura 2.

$$U \xrightarrow[\vec{r}_1 \to 0 \quad \vec{r}_2 \to 0]{} \infty$$

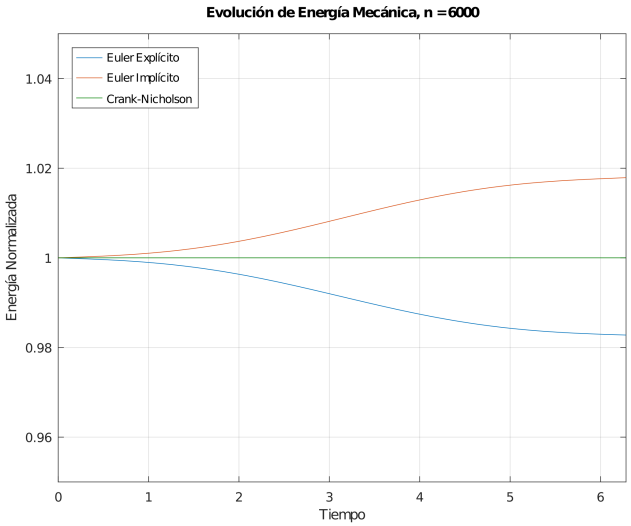

Figura 7: Evolución temporal de la energía mecánica del sistema de 2 cuerpos resuelto numéricamente en 6000 pasos.

Cuando se incrementa el número de pasos por un factor de 60 y el error disminuye según la tendencia marcada por la figura 4, la energía sigue conservándose únicamente para el método C-N, pero la variación de energía de los métodos de Euler no es tan dramática como en la figura 6.

El que la energía para el método de Euler implícito se mantenga acotada en la figura 7 implica que para este valor de n, los radios no tienden a 0 luego de la primera órbita.

Habiendo determinado que el método de Crank-Nicolson supera en convergencia a ambos métodos de Euler, se contrasta contra un método de Adams invocado con **lsode**, usando los parámetros por defecto.

Las figuras 8 y 9 contrastan el tiempo de ejecución y el error global respectivamente.

Ni el error global ni el tiempo de ejecución del método Adams invocado por **lsode** varían apreciablemente para valores de $N > 100$.

Eventualmente, para $N > 700$, la precisión del método de Crank-Nicolson supera la del método de Adams con sus valores por defecto, pero esto viene al costo de un tiempo de ejecución próximo a los 100 ms.

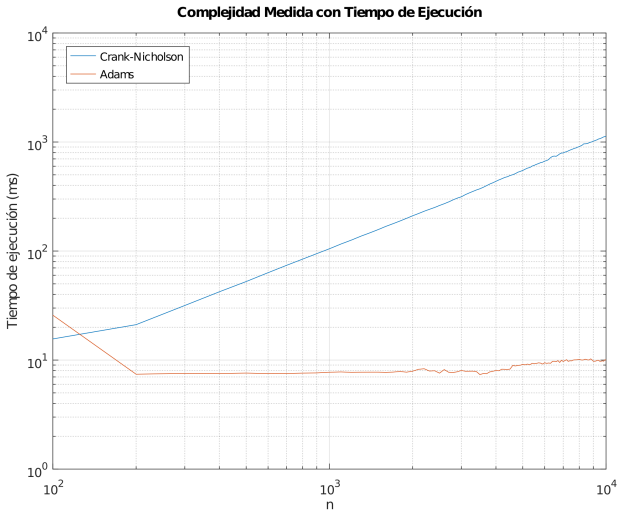

Figura 8: Tiempo de resolución del problema con los métodos en función de número de subdivisiones del intervalo temporal.

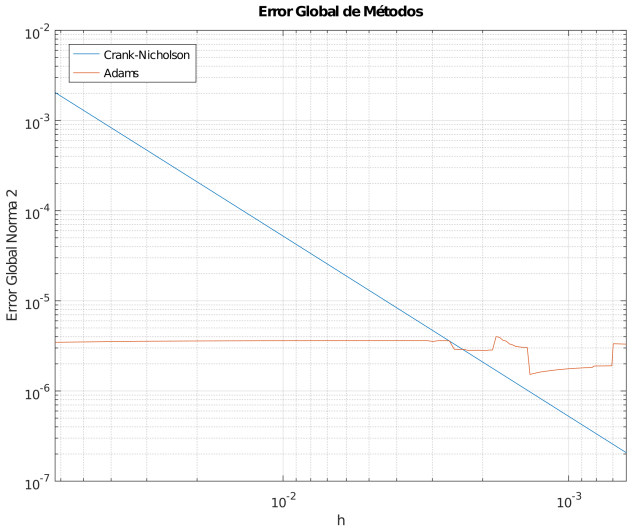

Figura 9: Error global de los métodos en función del paso temporal h.

Finalmente se utiliza el método de Adams y el método de Crank-Nicolson para resolver un problema más rígido [2] que el problema de órbitas circulares.

Se usa el problema definido por la ecuación 2, pero ahora con las condiciones iniciales correspondientes a órbitas elípticas

$$\vec{r}_1(t_0) = \begin{bmatrix} 1 & 0 \end{bmatrix} \qquad \vec{r}_2(t_0) = \begin{bmatrix} -1 & 0 \end{bmatrix}$$
$$\vec{v}_1(t_0) = \begin{bmatrix} 0 & \frac{3}{4} \end{bmatrix} \qquad \vec{v}_2(t_0) = \begin{bmatrix} 0 & -\frac{3}{4} \end{bmatrix}$$

El problema se resuelve usando el mismo intervalo temporal $t = [0, 2\pi]$ discretizado en $N = 100$ puntos, obteniendo con cada método las trayectorias 10 y 11.

[2] Se define como problema rígido, o *stiff*, a aquel que presenta variaciones rápidas en el valor de su derivada.

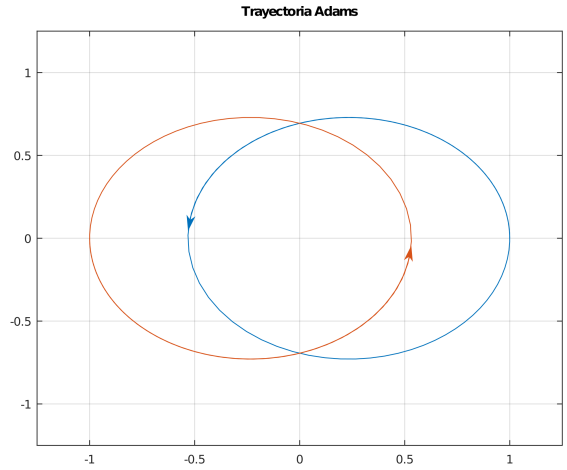

Figura 10: Trayectoria obtenida con el método de Adams para un problema rígido.

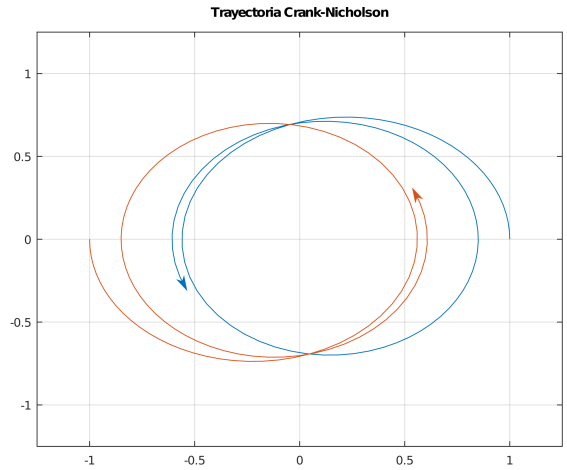

Figura 11: Trayectoria obtenida por el método de Crank-Nicolson para un problema rígido.

La figura 12 muestra la energía mecánica correspondiente a cada solución numérica.

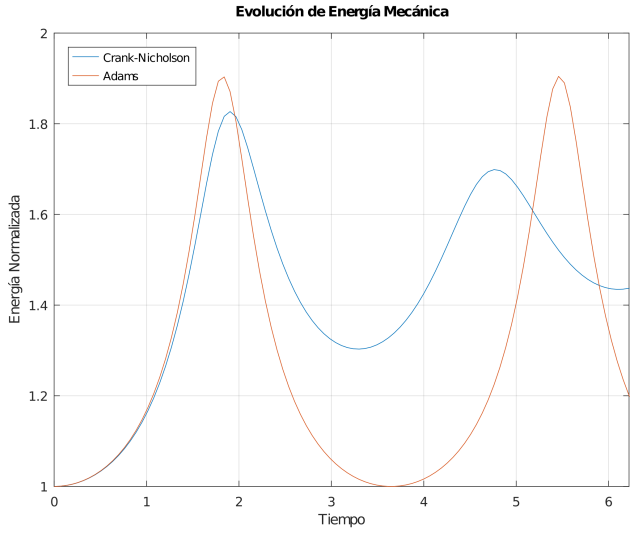

Figura 12: Evolución de la energía mecánica para la solución numérica de órbitas elípticas

La figura 10 muestra que el método de Adams converge a las órbitas elípticas esperadas, mientras que en la figura 11 se ve la tendencia del método de Crank-Nicholson a decaer a órbitas circulares de menor radio para este valor de h.

Si se repite el experimento aumentando la excentricidad de las órbitas y manteniendo la misma discretización del dominio, más evidente es el decaimiento del método de Crank-Nicolson a órbitas circulares.

A pesar de converger a la solución esperada, el método de Adams no preserva la energía mecánica. Sin embargo a diferencia del método de Crank-Nicholson, la energía vuelve a su valor inicial luego de cada periodo orbital, mientras que la de Crank-Nicholson converge a un valor diferente al inicial.

Conclusiones

El método de Crank-Nicolson es muy frecuentemente usado para resolver ecuaciones diferenciales ordinarias, y los resultados de este experimento dan indicios de por que es así.

En este experimento se vio que sus ventajas respecto a los métodos de Euler incluyen

- Convergencia de mayor orden a la solución analítica.
- Estabilidad numérica, sin tener error por defecto como el método de Euler Implícito.
- Preservación de la ley de conservación de la energía en problemas gravitacionales.

Se encontró una complejidad mayor del método en tiempo de ejecución, lo que era de esperarse viendo las ecuaciones del apéndice 1.

En cada paso temporal el método de Euler explícito realiza un producto matriz vector. El de Euler implícito resuelve un sistema lineal. El de Crank-Nicolson realiza ambas operaciones en cada paso.

Sin embargo en el experimento se encontró que el tiempo de ejecución incrementa con el mismo órden para todos los métodos, mientras que el método de Crank-Nicolson tiene un órden de convergencia mayor, permitiendo tener más precisión con menos divisiones.

Era de esperarse que, analizando los métodos como métodos θ, si el método $\theta = 0$ tiene error por exceso y el método $\theta = 1$ error por defecto, entonces el método

$\theta = \frac{1}{2}$ equilibre ambos errores para converger a la solución analítica.

El rendimiento del método de Crank-Nicholson es satisfactorio para problemas simples, tomando en cuenta que es un método muy simple de un paso consistiendo en productos matriciales iterativos, sin ningún tipo de predicción ni optimización. Es cuando se compara con el método Adams provisto por **lsode** que si cuenta con optimizaciones que se ven sus problemas.

En particular, el mayor problema del algoritmo de Crank-Nicholson desarrollado es su incapacidad de resolver de forma eficiente problemas con un mínimo grado de rigidez.

El método desarrollado si puede resolver problemas ligeramente rígidos, pero requiere un paso h muy chico para seguir las zonas de curvatura abrupta, lo cual implica desperdiciar recursos computacionales en pasos innecesarios en las zonas donde la curvatura es suave.

La rutina *lsode* se mantiene eficiente en estas condiciones porque usa un paso dinámico, achica el paso en las regiones donde la segunda derivada es mayor para tener más resolución, y agranda el paso en las regiones donde la segunda derivada es grande evitando operaciones innecesarias.

Referencias

[1] B. V. V. Simon Haykin, *Señales y Sistemas*. Limusa Wiley, 2001.

[2] Octave Forge, https://octave.sourceforge.io/docs.php, *Octave Forge Function Reference*, 2002 - 2008 ed.

Anexo

En una primera instancia se ejecutó una versión del método que precalculaba las matrices A y B asumiendo órbitas circulares. Esto es una equivocación, porque asume convergencia a la solución analítica, sin embargo las gráficas muestran que los métodos de Euler no retornan la solución real aún con esa suposición.

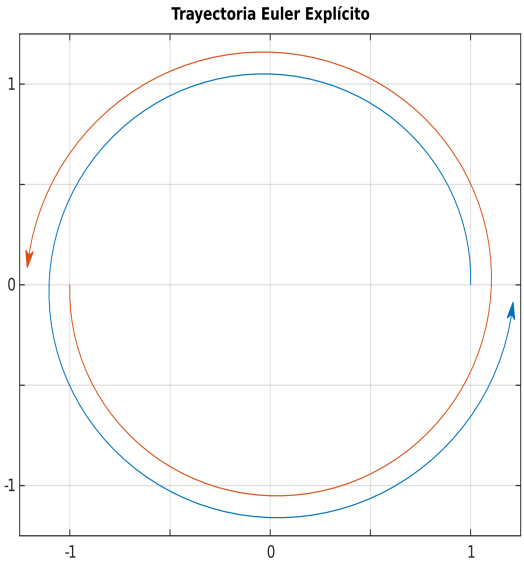

Figura 13: Trayectoria del sistema resuelto numéricamente con el método de Euler explícito y $h \simeq 0,63$, suponiendo convergencia solución real.

Error Global: $e^g(0,63) = 0,430$

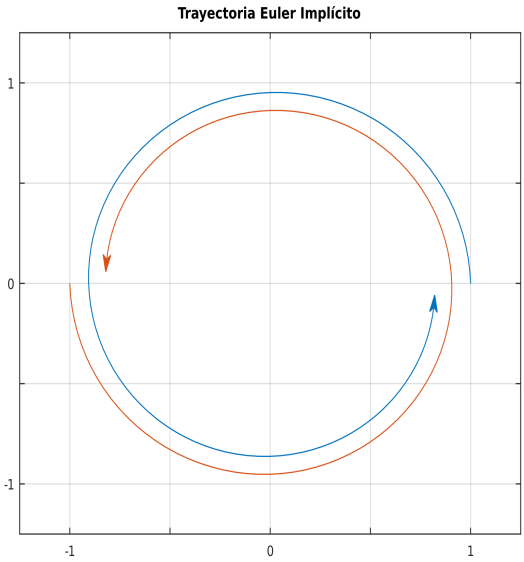

Figura 14: Trayectoria del sistema resuelto numéricamente con el método de Euler implícito y $h \simeq 0,63$, suponiendo convergencia a solución real.

Error Global: $e^g(0,63) = 0,354$

Apéndice 1 - Modelo Método θ

Para desarrollar el modelo del método θ, es necesario para preservar la linealidad precalcular en cada iteración los factores $\kappa_{[n]}$, teniendo estos factores se plantean los valores en la iteración $n+1$.

$$\begin{aligned}
\vec{r}1_{[n+1]} &= \vec{r}1_{[n]} + h\left[\theta\,\vec{v}1_{[n+1]} + (1-\theta)\,\vec{v}1_{[n]}\right] \\
\vec{v}1_{[n+1]} &= \vec{v}1_{[n]} + \kappa_2\,h\left[\theta\left[\vec{r}2_{[n+1]} - \vec{r}1_{[n+1]}\right] + (1-\theta)\left[\vec{r}2_{[n]} - \vec{r}1_{[n]}\right]\right] \\
\vec{r}2_{[n+1]} &= \vec{r}2_{[n]} + h\left[\theta\,\vec{v}2_{[n+1]} + (1-\theta)\,\vec{v}2_{[n]}\right] \\
\vec{v}2_{[n+1]} &= \vec{v}2_{[n]} + \kappa_1\,h\left[\theta\left[\vec{r}1_{[n+1]} - \vec{r}2_{[n+1]}\right] + (1-\theta)\left[\vec{r}1_{[n]} - \vec{r}2_{[n]}\right]\right]
\end{aligned} \tag{9}$$

Si se agrupan ahora los términos $[n]$ y los términos $[n+1]$

$$\begin{aligned}
\vec{r}1_{[n+1]} - h\theta\,\vec{v}1_{[n+1]} &= \vec{r}1_{[n]} + h\,(1-\theta)\,\vec{v}1_{[n]} \\
\vec{v}1_{[n+1]} - \kappa_2 h\theta\left[\vec{r}2_{[n+1]} - \vec{r}1_{[n+1]}\right] &= \vec{v}1_{[n]} + \kappa_2\,h\,(1-\theta)\left[\vec{r}2_{[n]} - \vec{r}1_{[n]}\right] \\
\vec{r}2_{[n+1]} - h\theta\,\vec{v}2_{[n+1]} &= \vec{r}2_{[n]} + h\,(1-\theta)\,\vec{v}2_{[n]} \\
\vec{v}2_{[n+1]} - \kappa_1 h\theta\left[\vec{r}1_{[n+1]} - \vec{r}2_{[n+1]}\right] &= \vec{v}2_{[n]} + \kappa_1\,h\,(1-\theta)\left[\vec{r}1_{[n]} - \vec{r}2_{[n]}\right]
\end{aligned} \tag{10}$$

La ecuación 10 se puede sintetizar a una única expresión matricial.

$$\begin{bmatrix} 1 & -h\theta & 0 & 0 \\ \kappa_2 h\theta & 1 & -\kappa_2 h\theta & 0 \\ 0 & 0 & 1 & -h\theta \\ -\kappa_1 h\theta & 0 & \kappa_1 h\theta & 1 \end{bmatrix} \begin{bmatrix} \vec{r}1 \\ \vec{v}1 \\ \vec{r}2 \\ \vec{v}2 \end{bmatrix}_{[n+1]} = \begin{bmatrix} 1 & h(1-\theta) & 0 & 0 \\ -\kappa_2 h\,(1-\theta) & 1 & \kappa_2 h(1-\theta) & 0 \\ 0 & 0 & 1 & h(1-\theta) \\ \kappa_1 h(1-\theta) & 0 & -\kappa_1 h(1-\theta) & 1 \end{bmatrix} \begin{bmatrix} \vec{r}1 \\ \vec{v}1 \\ \vec{r}2 \\ \vec{v}2 \end{bmatrix}_{[n]} \tag{11}$$

Si se especializa la ecuación 11 con $\theta = 0$ se obtiene el método de Euler explícito. Con $\theta = 1$, el método de Euler implícito, y con $\theta = \frac{1}{2}$, el método de trapecios de Crank-Nicolson.

- Euler Explícito

$$\begin{bmatrix} \vec{r}1 \\ \vec{v}1 \\ \vec{r}2 \\ \vec{v}2 \end{bmatrix}_{[n+1]} = \begin{bmatrix} 1 & h & 0 & 0 \\ -\kappa_2 h & 1 & \kappa_2 h & 0 \\ 0 & 0 & 1 & h \\ \kappa_1 h & 0 & -\kappa_1 h & 1 \end{bmatrix} \begin{bmatrix} \vec{r}1 \\ \vec{v}1 \\ \vec{r}2 \\ \vec{v}2 \end{bmatrix}_{[n]} \tag{12}$$

- Euler Implícito

$$\begin{bmatrix} 1 & -h & 0 & 0 \\ \kappa_2 h & 1 & -\kappa_2 h & 0 \\ 0 & 0 & 1 & -h \\ -\kappa_1 h & 0 & \kappa_1 h & 1 \end{bmatrix} \begin{bmatrix} \vec{r}1 \\ \vec{v}1 \\ \vec{r}2 \\ \vec{v}2 \end{bmatrix}_{[n+1]} = \begin{bmatrix} \vec{r}1 \\ \vec{v}1 \\ \vec{r}2 \\ \vec{v}2 \end{bmatrix}_{[n]} \tag{13}$$

- Crank-Nicolson

$$\begin{bmatrix} 1 & -\frac{1}{2}h & 0 & 0 \\ \frac{1}{2}\kappa_2 h & 1 & -\frac{1}{2}\kappa_2 h & 0 \\ 0 & 0 & 1 & -\frac{1}{2}h \\ -\frac{1}{2}\kappa_1 h & 0 & \frac{1}{2}\kappa_1 h & 1 \end{bmatrix} \begin{bmatrix} \vec{r}1 \\ \vec{v}1 \\ \vec{r}2 \\ \vec{v}2 \end{bmatrix}_{[n+1]} = \begin{bmatrix} 1 & \frac{1}{2}h & 0 & 0 \\ -\frac{1}{2}\kappa_2 h & 1 & \frac{1}{2}\kappa_2 h & 0 \\ 0 & 0 & 1 & \frac{1}{2}h \\ \frac{1}{2}\kappa_1 h & 0 & -\frac{1}{2}\kappa_1 h & 1 \end{bmatrix} \begin{bmatrix} \vec{r}1 \\ \vec{v}1 \\ \vec{r}2 \\ \vec{v}2 \end{bmatrix}_{[n]} \tag{14}$$