Métodos Numéricos

Book · November 2016

CITATIONS
0

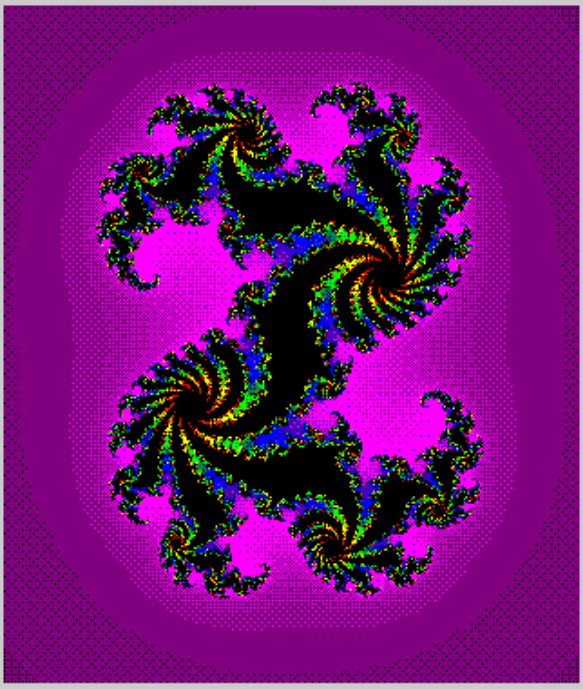
READS
9,997

1 author:

Andrés Granados
Simón Bolívar University
92 PUBLICATIONS **107** CITATIONS

SEE PROFILE

MÉTODOS NUMÉRICOS



Andrés L. Granados M.

Editorial Digitería
2016

http://digiteria.com.ve/wp/?p=342

UNIVERSIDAD SIMÓN BOLÍVAR.

DIVISION DE FÍSICA Y MATEMÁTICAS

Departamento de Mecánica

Valle de Sartenejas.
Caracas, VENEZUELA.

CURSO SOBRE:

MÉTODOS NUMÉRICOS

***

ANDRÉS L. GRANADOS M.

**

Enero, 2025.

♣

MÉTODOS NUMÉRICOS

Un Curso Introductorio Para Ingenieros y Científicos

Andrés L. Granados M.
UNIVERSIDAD SIMON BOLIVAR
Departamento de Mecánica
Sartenejas, Baruta, Edo. Miranda
Apdo.89000, Caracas 1080-A
Caracas, Venezuela.
E-Mail: agrana@usb.ve

Métodos Numéricos, publicado por Editorial Digitería, en Caracas, Venezuela, el 21 de noviembre de 2016

Ha sido realizado el Depósito Legal.

Depósito Legal MI2016000489

digiteria.ca@gmail.com

digiteria.com.ve/wp

Editorial Digitería

Ilustración de la portada: Conjunto de Julia del proceso iterativo complejo $z_{k+1} = z_k^2 + c$ con valor $c = 0.32 + 0.043\,i$.

MÉTODOS NUMÉRICOS

Un Curso Introductorio Para Ingenieros y Científicos

ANDRES L. GRANADOS M.
UNIVERSIDAD SIMON BOLIVAR. Departamento de Mecánica.
Valle de Sartenejas. Caracas.
Estado Miranda. Venezuela.

RESUMEN

En esta monografía se han desarrollado los métodos numéricos fundamentales básicos para un curso introductorio de cálculo numérico. He intentado plasmar toda mi experiencia pedagógica en la Universidad Simón Bolívar de alrededor de 25 años, durante los cursos de pre-grado MC-2416 Métodos Aproximados, MC-2421 Mecánica Computacional I y MC-2422 Mecánica Computacional II. Eventualmente me correspondió dictar también los cursos de post-grado MC-6461 Análisis Avanzado en Ingeniería y MC-7465 Técnicas Aproximadas en Mecánica. Fuera de la Universidad Simón Bolívar ocasionalmente dicté el curso de Doctorado "Técnicas Numéricas" en la Universidad Nacional Experimental Politécnica-UNEXPO, Sede Los Dos Caminos, Av. Sucre. He incluido material para cursos cortos como el de "Flujo en Redes de Tuberías" para PDVSA-ESP OIL (Análisis, Diagnóstico y Simulación de Redes de Fluidos) y PETRO-ECUADOR (Almacenamiento de Petróleo y Transferencia).

También he incluido de forma diluida varios de mis artículos más emblemáticos. Tratando de hacer inserciones continuas en el texto, allanando en lo posible los cambios de nivel, para hacer, de una lectura para un auditorio exclusivo de especializados, una lectura para una público más general e ingenuo y ávido de aprender. He incluido, en lo relativo al tema, mi experiencia durante mi doctorado en España, cuya tesis fué básicamente numérica aplicada a la Mecánica de Fluidos, flujo bifásico sólido-Gas, en régimen turbulento. Particularmente, mi experiencia con soluciones de ecuaciones diferenciales ordinarias, en derivadas parciales para fluidos e interpolaciones polinómicas y derivación numérica están plasmados en el texto.

Las siguientes referencias producidas en las últimas décadas, constituyen la inspiración inicial de esta obra, que se ha extendido lo posible en su contenido y habrá de extenderse todavía más. Principalmente en los temas de Interpolación Libre (Criterios), Regresión No Lineal (Levenberg-Marquardt), Sistemas de Ecuaciones No Lineales (Estabilidad Fractal) y Sistemas de Ecuaciones Diferenciales Ordinarias (Runge-Kutta Implícitos).

REFERENCIAS

[1] Granados M., A. L. **Nuevas Correlaciones para Flujo Multifásico.** INTEVEP S.A. Reporte Técnico No. INT-EPPR/322-91-0001. Los Teques, Febrero de 1991. Trabajo presentado en la Conferencia sobre: *Estado del Arte en Mecánica de Fluidos Computacional.* Auditorium de INTEVEP S.A. Los Teques, del 27 al 28 de Mayo de (1991).

[2] Granados M., A. L. **Second Order Methods for Solving Non-Linear Equations**, INTEVEP, S. A. (Research Institute for Venezuelan Petroleum Industry), Tech. Rep. No.INT-EPPR/322-91-0002, Los Teques, Edo. Miranda, pp.14-36, Jun. 1991.

[3] Granados M., A. L. **Free Order Polynomial Interpolation Algorithm.** INTEVEP S.A. Nota Técnica. Los Teques, Jul. 1991.

[4] Granados M., A.L. **Lobatto Implicit Sixth Order Runge-Kutta Method for Solving Ordinary Differential Equations with Stepsize Control**. INTEVEP S.A. Reporte Técnico No. INT-EPPR/3-NT-92-003. Los Teques, Marzo 1992.

[5] Granados M., A.L. "Fractal Technics to Measure the Numerical Instability of Optimization Methods". **Mecánica Computacional Vol.XV**: Anales del "9° CONGRESO SOBRE METODOS NUMERICOS Y SUS APLICACIONES, ENIEF'95". Hotel Amancay, 6-10 de Noviembre de 1995, San Carlos de Bariloche, Argentina. Compilado por Larreteguy, A. E. y Vénere, M. J. Asociación Argentina de Mecánica Computacional (AMCA), pp.369-374,1995.

[6] Granados M., A. L. "Fractal Techniques to Measure the Numerical Instability of Optimization Methods". **Numerical Methods in Engineering Simulation**: *Proceedings of The Third International Congress on Numerical Methods in Engineering and Applied Sciences, CIMENICS'96.* Cultural Centre Tulio Febres Cordero, March 25-29, 1996. Mérida, Venezuela. Editors: M. Cerrolaza, C. Gajardo, C. A. Brebbia. Computational Mechanics Publications of the Wessex Institute of Technology (UK), pp.239-247, (1996).

[7] Granados M. A. L. "Lobatto Implicit Sixth Order Runge-Kutta Method for Solving Ordinary Differential Equations with Stepsize Control". **Mecánica Computacional Vol.XVI**: *Anales del V Congreso Argentino de Mecánica Computacional, MECOM'96.* Universidad Nacional de Tucumán, Residencia Universitaria Horco Molle, Comuna de Yerba Buena, 10-13 de Septiembre de (1996). San Miguel de Tucumán, Argentina. Compilado por: Etse, G. y Luccioni, B. Asociación Argentina de Mecánica Computacional (AMCA), pp.349-359, (1996).

[8] Granados M., A. L. "Implicit Runge-Kutta Algorithm Using Newton-Raphson Method". **Simulación con Métodos Numéricos: Nuevas Tendencias y Aplicaciones**, Editores: O. Prado, M. Rao y M. Cerrolaza. Memorias del *IV CONGRESO INTERNACIONAL DE METODOS NUMERICOS EN INGENIERIA Y CIENCIAS APLICADAS, CIMENICS'98.* Hotel Intercontinental Guayana, 17-20 de Marzo de 1998, Puerto Ordaz, Ciudad Guayana. Sociedad Venezolana de Métodos Numéricos en Ingeniería (SVMNI), pp.TM9-TM16. Corregido y ampliado Abril, 2016.

[9] Granados M., A. L. "Implicit Runge-Kutta Algorithm Using Newton-Raphson Method". *Fourth World Congress on Computational Mechanics*, realizado en el Hotel Sheraton, Buenos Aires, Argentina, 29/Jun/98 al 2/Jul/98. International Association for Computational Mechanics, **Abstracts**, Vol.**I**, p.37, (1998).

[10] Granados, A. L. "Numerical Taylor's Methods for Solving Multi-Variable Equations", Universidad Simón Bolívar, Mayo, 2015.

[11] Granados, A. L. "Taylor Series for Multi-Variable Functions", Universidad Simón Bolívar, Dic. 2015.

[12] Granados, A. L. "Criterios para Interpolar", Universidad Simón Bolívar, Jul. 2017.

[13] Granados, A. L. "Métodos Numéricos para Redes", Universidad Simón Bolívar, Mar. 2023.

[14] Granados M., A. L. "Redes de Tuberías". Universidad Simón Bolívar, Departamento de Mecánica, Nov., 2020. https://www.academia.edu/20394578/Redes_de_Tuberías

DEDICATORIA

Dedico este trabajo a mi querida esposa Magaly y a mi adoradas hijas Andreína y Andrea, con todo el amor del mundo.

Deseo también dedicar este trabajo a todos aquellos hombres sabios que han hecho posible el desarrollo del ...

ANÁLISIS NUMÉRICO

como una parte importantísima de las Matemáticas Aplicadas.

Andrés L. Granados M.

PREFACIO

Una importante razón motivó la elaboración de este trabajo. En la literatura especializada de habla española no existe un texto que realice una introducción al Cálculo Numérico de una forma sencilla, resumida y completa, simultáneamente con aplicaciones al campo de la ingeniería mecánica. Un texto de Análisis Numérico sería demasiado tedioso para un curso enfocado para estudiantes de ingeniería. Un compendio de algoritmos sin la formalidad del análisis sería demasiado estéril para aquellos estudiantes que quieren un enfoque más general. En esta oportunidad se ha tratado de crear un híbrido de ambos aspectos aparentemente extremos. Esto se ha hecho mediante la estructuración de un recetario de métodos numéricos con inserciones de aspectos analíticos importantes, que en primera instancia pueden ser obviados. Sin embargo, para los estudiantes más curiosos, estos aspectos analíticos pueden ser revisados en una segunda o tercera lectura más detallada.

Esta monografía en primera instancia fue desarrollada para estudiantes de pregrado, sin embargo, puede servir para un curso introductorio a nivel de postgrado, en donde se haga más énfasis a los aspectos analíticos. El curso se ha diseñado para completarse en un trimestre, pero puede fácilmente extenderse a un semestre si los dos últimos capítulos se estudian con mayor profundidad.

Todo el temario de este texto se ha estructurado en cinco (5) capítulos y un (1) apéndice:

- Solución de Ecuaciones No Lineales.
- Solución de Sistemas de Ecuaciones.
- Interpolación, Integración y Aproximación.
- Ecuaciones Diferenciales Ordinarias.
- Ecuaciones en Derivadas Parciales.
  - Series de Taylor.

Todos los temas tratados en este trabajo se han enfocado siguiendo un proceso de desarrollo de los temas de forma inductiva. Se comienzan los temas con problemas o algoritmos particulares y luego se generalizan dentro de un espacio de problemas o de algoritmos. Esto último, como se planteó antes, viene acompañado de su respectivo análisis, lo cual completa y formaliza las ideas vagamente planteadas en un principio.

El Capítulo I presenta un breve resumen de todos los métodos numéricos más importantes para resolver una sóla ecuación algebraica no lineal, en donde intervengan una que otra función trascendental. Esto cubre todo el espectro de posibilidades.

El Capítulo II trata de los sistemas de ecuaciones. Básicamente se distinguen dos grupos de problemas: Sistemas de Ecuaciones Lineales y Sistemas de Ecuaciones No Lineales. Par el primer grupo pueden existir métodos directos y métodos iterativos. Para el segundo grupo, todos los métodos son iterativos.

El Capítulo III contiene métodos para estimar valores de funciones dadas de manera discreta. En los métodos de interpolación la función que estima los valores pasa por todos los puntos discretos. También se presenta un conjunto de métodos para estimar las derivadas e integrales basándose en los métodos de interpolación estudiados en la Sección anterior. En los métodos de aproximación los datos forman una muestra estadística, y por lo tanto es casi imposible que la función que estima los valores pase por todos y cada uno de los puntos datos.

El Capítulo IV desarrolla el tema de integración de ecuaciones diferenciales ordinarias mediante métodos de un sólo paso o métodos de paso múltiples. En ambos casos, se presentan las alternativas de métodos explícitos (predictores) y métodos implícitos (correctores).

El Capítulo V concluye el contenido de esta monografía con el estudio introductorio de métodos para la resolución de ecuaciones diferenciales en derivadas parciales (en desarrollo). Básicamente se distinguen tres categorías de métodos: Diferencia Finita, Volúmenes Finitos y Variacionales. También se presentan algunos

métodos mixtos como lo son el método de la líneas y el métodos de las características. Al final de este caítulo se hace una introducción muy básica a los métodos de los elementos finitos.

En los **Anexos** existe un apéndice que han sido colocado para hacer consultas rápidas acerca de cuestiones de contenido matemático relacionadas con las series de Taylor, que de otra manera recargarían el texto en su parte pricipal. En este apéndice el tratamiento de los temas es formal tratando de ser lo más general posible. Sin embargo, se han omitido demostraciones y fundamentos que son importantes, puesto que no son el objetivo primordial de este texto. Se incluyen dentro de los anexos la bibliografía general compilada de toda esta monografía en un sólo lugar, aunque ya esté redundantemente distribuida por cada capítulo.

Los capítulos han sido numerados con números romanos, como ya se habrá visto, las secciones con números consecutivos y las sub-secciones y subsub-secciones con números de apartados de los números de las secciones y sub-secciones respectivamente. Es decir, por ejemplo, el Capítulo VII tiene una Sección 2., una Sub-sección 2.1. y una Subsub-sección 2.1.3. Cuando se hace dentro del texto una referencia a una sección o sub-sección en particular se menciona de la siguiente manera: ... ver la Sección VII.2. ... o ... ver la Sección VII.2.1.3. En caso de que se esté referenciando una parte del texto perteneciente al mismo capítulo o a la misma sección esta información se omite. Los apéndices han sido ordenados según las letras del alfabeto, por ejemplo, Apéndice A, Apéndice B, etc. La organización interna de cada Apéndice es la misma que para los capítulos. Existe una tabla de contenido general al principio del texto, sin embargo, al principio de cada capítulo se ha colocado una tabla de contenido más detallada para facilitar la búsqueda de los temas de interés para el lector.

Las ecuaciones han sido numeradas de forma consecutiva por sub-secciones. Eventualmente la primera sub-sección puede incluir la sección principal (anterior) dentro de la numeración de ecuaciones. Para referenciar las ecuaciones se hace de la siguiente forma: ... basado en la ecuación VII.2.1.(13) ..., cuyo significado es obvio. Para las ecuaciones también es válida la observación hecha antes con respecto a la información superflua. Así que si estoy dentro del mismo capítulo se diría ... ecuación 2.1.(13) ... , o si se está en la misma sub-sección simplemente se habla de la ecuación (13). En alguna ocasiones un grupo de ecuaciones se numera con un sólo número. En estos casos debe entenderse que las ecuaciones internas están ordenadas con letra de arriba hacia abajo y de izquierda a derecha. Por ejemplo, ... ver ecuación (10.c) ... Aunque el grupo de ecuaciones esté numerado con el número (10) sólamente, se entenderá que la ecuación a la que se hizo referencia es la tercera dentro del grupo.

Los axiomas, definiciones, proposiciones, lemas, teoremas y corolarios han sido numerados de forma consecutiva por sub-secciones, al igual que las ecuaciones, con la particularidad de que el número, en lugar de aparecer entre paréntesis, se presentará en negrillas. Por ejemplo, ... Teorema A.3.2.**1**. Una consideración adicional es que cuando en una sub-sección exista un sólo teorema, axioma, etc., este no se numerará, sin embargo se sobreentenderá que es el teorema, axioma, etc. número 1 de esa sub-sección. En las definiciónes cuando aparezcan por primera vez se colocará la palabra o palabras definidas en *letras inclinadas*.

Para las referencias bibliográficas no se sigue el mismo principio que las ecuaciones para referirlas. Al final de la monografía se dispone de un listado de las bibliografías más importante a las cuales puede o no hacerse referencia. Las bibliografías se han ordenado en un listado de forma alfabética, empleando al mismo tiempo el apellido del autor y año entre corchetes o un úmero entre corchetes, para indicar el lugar que ocupa dentro de dicho ordenamiento. Existen dos formas para hacer mención a una referencia. Una de ellas, la más abreviada, es mediante el número entre corchetes que se mencionó antes, dentro de cada capítulo. La otra forma, es mediante el apellido del primer autor y el año entre corchetes o entre paréntesis. Cuando el año de la publicación está encerrado entre paréntesis significa que la publicación es periódica, y, en caso contrario, significa que es una monografía, por ejemplo, ... ver la referencia [15], o ... ver a [Wilkinson,1972], o ... ver a [Atkinson,(1965)]. Cuando para un mismo autor y un mismo año existen dos publicaciones o más, se anexa al año las diferentes letras minúsculas del alfabeto, por ejemplo, ... [Marquardt,1960a] ... [Marquardt,1960b]. Finalmente, cuando se desea mencionar un nombre o un autor que a su vez es referenciado en otra parte, este debe aparecer fuera de los corchetes, por ejemplo, ... Taylor [Marquardt,1960], o también puede aparecer de la forma ... Taylor [10], aunque Taylor no sea el autor de la referencia [10]. Dentro de los corchetes puede aparecer eventualmente información adicional a la referencia como el capítulo o las páginas como por ejemplo, ... [Marquardt,1960;§.81,p.347]. El símbolo '§' se emplea para indicar los capítulos o secciones, el símbolo '¶' se emplea para indicar los párrafos y el símbolo 'p' para indicar las páginas. Cuando estos símbolos aparecen

dos veces significa que son varios las entidades a la que se hace referencia, las cuales se pueden indicar como un rango de cantidades separadas por el símbolo '-'.

La notación usada en el texto es la convencional para estos temas, sin embargo, al final del texto se ha hecho un anexo con la notación más importante. De manera general, se puede decir que se ha empleado la notación de Gibbs, empleando itálicas para los escalares, negrillas minúsculas para los vectores y negrillas mayúsculas para los tensores de orden dos o más. Esta regla, aunque general tiene algunas excepciones, en cuyo caso el carácter de la cantidad se especifica ampliamente. El producto escalar se especifica con un punto, el producto vectorial se especifica con una cruz y la doble contracción del producto de dos tensores de segundo orden (o producto escalar de dos tensores) se especifica con el doble punto. También se ha definido el producto punto de un tensor y un vector como la transformación de este por aquel, significando al mismo tiempo que existe una contracción en los índices adyacentes en las componentes. Algo similar se ha definido para el producto cruz de un tensor y un vector, donde el producto sólamente afecta los vectores bases adyacentes al símbolo de multiplicación. También se define el producto cuña como el producto exterior y su relación con el producto cruz y con el producto tensorial. El producto tensorial, para los efectos de simplificar la notación en la gran mayoría de los casos, se indica como un producto diádico y no con una cruz encerrada en un círculo, como normalmente se hace en los textos de análisis matemático. Sin embargo, en donde se hace necesario emplear el producto tensorial de forma explícita se emplea el símbolo antes mencionado. La notación matricial se ha prácticamente confinado al Capítulo II.

Cualquier comentario de forma o de fondo acerca de esta obra será bien recibido por el autor, puesto que se está bien seguro que ellos redundarán en mejoras y añadiduras, que de otra forma tardarían mucho tiempo en realizarse.

Deseo dar las gracias a todas aquellas personas que de alguna forma se han interesado en la obra, y espero que sea de mucha utilidad, tanto en los cursos que la emplean, como en su uso en calidad de material de consulta.

Andrés L. Granados M.
UNIVERSIDAD SIMON BOLIVAR
Departamento de Mecánica
Caracas, Venezuela, Agosto de 2016

CONTENIDO

CURSO SOBRE:

MÉTODOS NUMÉRICOS

FUNDAMENTOS

CAPITULO I

SOLUCION DE ECUACIONES NO-LINEALES

CONTENIDO

Es frecuente encontrarse con expresiones matemáticas que involucran funciones trascendentales o polinomios de orden superior, en la cual se desea hallar el valor de alguna de las variables involucradas que satisfaga dicha expresión, de aquí el interés en desarrollar métodos numéricos para solventar dicha necesidad.

Entre los casos mas frecuentes en ingeniería mecánica se encuentran las expresiones correspondientes a las ecuaciones de estado para una sustancia determinada, la cual viene dada por $f(p, v, T)$, donde no todas las variables pueden ser despejadas en forma explícita. También podemos citar las ecuaciones que rigen los

modos de vibración de medios continuos, las cuales involucran funciones exponenciales y trigonométricas, o en la solución de problemas mediante series de Fourier, etc.

Describiremos algunos de los métodos más utilizados para resolver el presente problema, como lo son:

a. Método de la bisección, atribuido a Bolzano.
b. Método de interpolación lineal.
c. Método de la secante.
d. Método iterativo o de punto fijo.
e. Método de Newton-Raphson.
f. Métodos de Segundo Orden cerrados y abiertos.

1. METODOS CERRADOS

1.1. TEOREMA DE BOLZANO

"Sea $f(x)$ una función contínua en el intervalo $[a, b]$ tal que $f(a)f(b) \leq 0$, entonces se puede garantizar que existe un valor r tal que $f(r) = 0$, este valor r se denomina *la raíz* de $f(x)$".

1.2. BISECCION DEL INTERVALO

El método que a continuación se describe está basado en el teorema de Bolzano para funciones contínuas. Por la forma como se implementa el algoritmo se podría decir que este método es de orden cero. Utilizando este teorema, y dada una expresión del tipo $f = f(x)$ se selecciona un intervalo $[a, b]$ donde se verifique la condición de $f(a)\, f(b) \leq 0$ impuesta por el teorema. Si dividimos este intervalo en dos sub-intervalos de igual longitud, se debe verificar la condición del teorema en alguno de los dos sub-intervalos, por lo tanto r está en el intervalo donde la función $f(x)$ cambia de signo. Repitiendo el proceso en forma iterativa se obtendrán intervalos de menor longitud hasta acotar el valor de la raíz r.

El proceso de subdivisión del intervalo se lleva a cabo hasta que se verifique una tolerancia especificada para el problema, entre las cuales se pueden mencionar:

- Tolerancia ϵ_{max} en la variable independiente (error local) en forma absoluta o relativa. El error absoluto representa la distancia (cuando se usa valor absoluto) entre dos estimados consecutivos de la raíz, y el error relativo es el valor absoluto del cociente entre el error absoluto y el último estimado de la raíz.
- Tolerancia d_{max} en el valor absoluto de la función (desviación global). La desviación viene a representar el valor obtenido al evaluar el valor de la función $f(x)$ en el estimado de la raíz.
- Ambas simultáneamente (condicón inclusiva). Cuando es en una variable la tolerancia de la otra variable se escoge de valor grande para que siempre se cumpla la condición.
- O se alcance (condición exclusiva) la tolerancia k_{max} de número de iteraciones máximas permitidas que se puede estimar con $\epsilon_{max} \geq (b - a)/2^{k_{max}}$.

Es responsabilidad de quien utiliza el algoritmo conocer cual de estas formas es más restrictiva en cuanto a la precisión en el valor de la raíz r.

Ya descrito el método podemos organizarlo de la siguiente forma:

1. Sea $f(x)$ una función contínua tal que $f(a)\, f(b) \leq 0$ en el intervalo $[a, b]$. Se escoge un número máximo de iteraciones $k_{max} \geq [\log(b - a) - \log \epsilon_{max}]/\log 2$.

   Denotando $k = 1$, $x_1 = a$ y $x_2 = b$.
2. Se evalúa el estimado de la raíz mediante $c_k = (x_1 + x_2)/2$.
3. Se determina el signo de $f(c_k)f(x_1)$:

   Si el signo es positivo se toma $x_2 = c_k$.

   En caso contrario $x_1 = c_k$.

4. Se evalúa el error y la desviación mediante las expresiones:

   Error local: $\epsilon_k = c_k - c_{k-1}$.

   Desviación global: $d_k = f(c_k)$.

5. Se verifica si el error local ϵ_k y la desviación global d_k son en valor absoluto menores que las tolerancias seleccionadas ϵ_{max} y d_{max}. En caso afirmativo se detiene el proceso de cálculo y el valor deseado de la raíz es igual al valor de c_k, $r = c_k$. En caso contrario, se vuelve al punto 2 y realiza una nueva iteración $(k \to k+1)$.

EJEMPLO:

Hallar el factor de fricción de una tubería comercial($\varepsilon/D = 0.00001$) por la cual circula un fluido con un número de Reynolds igual a $I\!R\!e = 10^6$, utilizando para ello la ecuación de Colebrook:

$$\frac{1}{\sqrt{f}} = -2\,\log\left(\frac{a}{I\!R\!e\,\sqrt{f}} + \frac{\varepsilon/D}{b}\right) \qquad \begin{matrix} a = 2.51 \\ b = 3.71 \end{matrix}$$

Substituyendo los valores donde $x = \mathrm{f}$ es la variable independiente

$$f(x) = \frac{1}{\sqrt{\mathrm{f}}} + 0.86\,\ln\left[2.6954 * 10^{-6} + \frac{2.51 * 10^{-6}}{\sqrt{\mathrm{f}}}\right] = 0$$

Escogiendo $\mathrm{f}_1 = 0.001$ y $\mathrm{f}_2 = 0.05$, basados en el conocimiento del factor fricción (por experiencia), se obtiene la siguiente tabla.

Tabla. Resultados del Ejemplo.

a	$f(a)$	b	$f(b)$	c	$f(c)$
0.0010	23.5319	0.0500	−5.1445	0.0255	−3.1138
		0.0255	−3.1138	0.0133	−0.4593
		0.0133	−0.4593	0.0072	2.8944
0.0072	2.8944			0.0103	0.8218
0.0103	0.8218			0.0118	0.1198
0.0118	0.1198			0.0126	−0.2009
		0.0126	−0.2009	0.0122	−0.0452
		0.0122	−0.0452	0.0120	−0.0363

El valor del factor de fricción es 0.0120 con una tolerancia de $2 * 10^{-4}$ y 8 iteraciones. Una tolerancia en la desviación de $4 * 10^{-2}$ sería suficiente.

1.3. INTERPOLACION LINEAL

1.3.1. Simple

Al igual que el método de bisección del intervalo, el método de interpolación lineal, también denominado *Regula Falsi*, se basa en el teorema de Bolzano, pero con una variante para el cálculo del estimado de la raíz r.

La idea fundamental de este algoritmo es acelerar la convergencia al valor de r, de forma de disminuir el tiempo de cálculo.

De forma de hallar el valor de r que satisface la expresión $f(r) = 0$, se supondrá que la función se comporta como una recta que pasa por los puntos $(a, f(a))$ y $(b, f(b))$, correspondientes a los extremos del intervalo.

La ecuación de la recta que pasa por estos dos puntos viene dada por:

$$\frac{y - f(a)}{f(b) - f(a)} = \frac{x - a}{b - a} \tag{1}$$

Bajo la suposición de un comportamiento la lineal, la aproximación de r sera el punto de corte c de la recta con el eje de las abscisas $y = 0$, lo cual genera dos nuevos sub-intervalos $[a, c]$ y $[c, b]$, en los cuales se deben verificar las condiciones del teorema de Bolzano y escoger el intervalo que las satisfaga. Posteriormente se repite el procedimiento hasta verificar alguna de las tolerancias seleccionadas.

La expresión que permite evaluar el estimado de la raíz queda de la siguiente forma:

$$c = a - f(a)\frac{b - a}{f(b) - f(a)} = b - f(b)\frac{b - a}{f(b) - f(a)} = \frac{a\, f(b) - b\, f(a)}{f(b) - f(a)} \tag{2}$$

Ya descrito el algoritmo, se puede agrupar de la siguiente forma:

1. Sea $f(x)$ una función continua tal que $f(a)f(b) \leq 0$ en el intervalo $[a, b]$.

   Denotando $k = 1$, $x_1 = a$ y $x_2 = b$.
2. Se evalúa el estimado de la raíz c_k mediante la equación (2).
3. Se determina el signo de $f(c_k)f(x_1)$:

   Si el signo es positivo se toma $x_2 = c_k$.

   En caso contrario $x_1 = c_k$.
4. Se evalúa el error y la desviación mediante las expresiones:

   Error local: $\epsilon_k = c_k - c_{k-1}$.

   Desviación global: $d_k = f(c_k)$.
5. Se verifica si el error local ϵ_k o la desviación global d_k son menores en valor absoluto que las tolerancia seleccionada ϵ_{max} y d_{max}. En caso afirmativo se detiene el proceso de calculo y el valor deseado de la raíz r es igual al valor de c_k, $r = c_k$. En caso contrario se vuelve al punto 2 $(k \to k + 1)$.

EJEMPLO:

Hallar la raíz cercana a $x = 1$ de la expresión:

$$f(x) = e^x - 3x^2 = 0$$

Tabla. Resultados del ejemplo.

a	$f(a)$	b	$f(b)$	c	$f(c)$
0.0000	1.0000	1.0000	−0.2817	0.7802	0.3558
0.7802	0.3558			0.9029	0.0212
0.9029	0.0212			0.9097	0.0011
0.9097	0.0011			0.90999	0.000052

La raíz buscada es $r = 0.90999$ con una desviación de $5.2 * 10^{-5}$.

1.3.2. Modificada

El método de interpolación lineal modificada es una variedad de la anterior donde, en el caso de obtener una convergencia marcadamente unilateral como en el ejemplo anterior, el extremo inalterado sufre una modificación en la fórmula algorítmica 1.3.(2), el valor de la función se divide por dos consecutivamente hasta que la convergencia deje de ser unilateral (en el ejemplo $f(b)/2^{[\![k-n]\!]}$, donde n es el número de repeticiones permitida en el extremo x_2). Lo mismo para el otro extremo x_1 en caso de que se repita. Normalmente la convergencia unilateral ocurre siempre por el mismo lado. El símbolo $[\![\cdot]\!]$ indica que es el valor positivo sobre 0, $[\![x]\!] = max[0, x]$, luego de que k se ha reinicializado después de un cambio en la unilateralidad de la convergencia. El método descrito recibe el nombre de algoritmo de Illinois por parte de algunos escolares [Dahlquist & Björck,1974] e incluso han surgido algunas variantes [Ford,1995].

Supóngase que en la k-ésima iteración el intervalo cerrado es $[a_k, b_k]$ y que el valor funcional $f(c_k)$ del nuevo iterado c_k

$$c_k = \frac{a_k\, f(b_k) - b_k\, f(a_k)}{f(b_k) - f(a_k)} \tag{3}$$

tiene el mismo signo que $f(b_k)$. En este caso, el nuevo intervalo cerrado sería $[a_{k+1}, b_{k+1}] = [a_k, c_k]$ y el lado izquierdo ha sido retenido. La convergencia es unilateral por la derecha. Como quiera que el algoritmo de Illinois multiplicaría $f(a_k)$ por $\frac{1}{2}$, el algoritmo de Anderson-Björck multiplicaría por m, donde dicho factor tiene uno de los siguiente dos valores [King,(1983)]

$$m = \begin{cases} m' & \text{si } \; m' > 0 \\ \frac{1}{2} & \text{de otra forma} \end{cases} \qquad\qquad m' = 1 - \frac{f(c_k)}{f(b_k)} \tag{4}$$

Para raćes sencillas, el algoritmo de Anderson-Björck se comporta bien en la práctica. En el algoritmo anteriormente descrito, donde se permiten repeticiones, el factor sería $m^{[\![k-n]\!]}$.

Otro método modificado a considerar es el método ITP (Interpolaton-Truncation-Proyection), que no es más que una mejora del método de la bisección del intervalo. Dados $\kappa_1 \in \mathbb{R}$, $\kappa_2 \in [1, 1+\varphi)$, $n_{1/2} = \log_2[L_0/(2\epsilon_{max})]$, $L_k = (a_k - b_k)$, siendo $\varphi = (1+\sqrt{5})/2$ la proporción áurea, se calcula la estimación x_{ITP} en los siguiente tres pasos:

1. *Interpolation.* Se calcula los puntos de bisección $x_{1/2} = (a_k + b_k)/2$ y regula falsi $x_f = c_k$ con (3);
2. *Truncation* Se perturba la estimación hacia el centro $x_t \equiv x_f + \sigma\,\delta$ (x_t truncamiento de x_f), donde $\sigma \equiv \text{sign}(x_{1/2} - x_f)$ y $\delta \equiv \min\{\kappa_1 |b_k - a_k|^{\kappa_2}, |x_{1/2} - x_f|\}$;
3. *Projection* Se proyecta la estimación hacia el intervalo de *minmax* $x_{ITP} \equiv x_{1/2} - \sigma\,\rho_k$, donde $\rho_k \equiv \min\{r_k, |x_t - x_{1/2}|\}$, siendo $r_k \equiv \epsilon_{max} 2^{n_{1/2}-k} - L_k/2$ el radio del intervalo $I_k = [x_{1/2} - r_k, x_{1/2} + r_k]$.

Finalmente se comprueba que ahora, a cada paso k, el método ITP define $\tilde{x}_k \in (a_k, b_k)$ como la proyección de x_t en el intervalo I_k ($|\tilde{x}_k - x_{1/2}| \leq r_k$) [Oliveira & Takahashi,(2021)].

1.4. METODOS DE SEGUNDO ORDEN

Los métodos de segundo orden se basan en hacer pasar una parábola única por los puntos tres puntos $[a, f(a)]$, $[b, f(b)]$ y $[c, f(c)]$, los extremos y el intermedio. Los puntos extremos en $x = a$ y $x = b$ satisfacen el teorema de Bolzano y el punto intermedio bien sea el obtenido con el método de bolzano $\bar{c}$ o el método de interpolación lineal c' (el "orden" en realidad se refiere al grado del polinomio utilizado).

Estos métodos de basa en los polinomios de Newton en diferencias divididas, como por ejemplo la parábola que pasa por los tres puntos $(a, f(a))$, $(b, f(b))$ y $(c, f(c))$, y que tiene la forma

$$P_2(x) = f[a] + (x-a) f[a, b] + (x-a)(x-b) f[a, b, c] \tag{1}$$

donde el símbolo $f[\ \cdot\ ]$ se denomina *diferencia dividida* y se define de forma recurrente empleando las siguientes expresiones [Carnahan et al.,1969]

$$f[x_0] = f(x_0) \tag{2.a}$$

$$f[x_1, x_0] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} \tag{2.b}$$

$$f[x_2, x_1, x_0] = \frac{f[x_2, x_1] - f[x_1, x_0]}{x_2 - x_0} \tag{2.c}$$

$$f[x_3, x_2, x_1, x_0] = \frac{f[x_3, x_2, x_1] - f[x_2, x_1, x_0]}{x_3 - x_0} \tag{2.d}$$

$$f[x_n, x_{n-1}, \ldots, x_1, x_0] = \frac{f[x_n, x_{n-1}, \ldots, x_2, x_1] - f[x_{n-1}, x_{n-2}, \ldots, x_1, x_0]}{x_n - x_0} \tag{2.e}$$

Siendo $f[x_n, x_{n-1}, \ldots, x_1, x_0] = f[x_0, x_1, \ldots, x_{n-1}, x_n] \ \ \forall n \in I\!N$.

1.4.1. Método de Brent

El método de Brent [Brent,1973] se basa en hacer pasar por los tres puntos antes mencionados una parábola inversa. Es decir, una parábola acostada cuyo polinomio de interpolación (de Lagrange sección III.1.2) de segundo grado inverso es (intercambiando el rol de variables independiente y dependiente)

$$x = \frac{[y - f(b)]\,[y - f(c)]\,a}{[f(a) - f(b)]\,[f(a) - f(c)]} + \frac{[y - f(a)]\,[y - f(c)]\,b}{[f(b) - f(a)]\,[f(b) - f(c)]} + \frac{[y - f(a)]\,[y - f(b)]\,c}{[f(c) - f(a)]\,[f(c) - f(b)]} \tag{3}$$

Colando $y = 0$ obtenemos el nuevo estimado c_k de la raíz, el cual puede ser escrito

$$c_k = b + \frac{P}{Q} \tag{4}$$

donde

$$\begin{aligned} P &= S\,[\,T\,(R - T)\,(c - b) - (1 - R)\,(b - a)\,] \\ Q &= (T - 1)\,(R - 1)\,(S - 1) \end{aligned} \qquad \begin{aligned} R &= f(b)/f(c) \\ S &= f(b)/f(a) \\ T &= f(a)/f(c) \end{aligned} \tag{5}$$

Este método escoge usar la parábola acostada para garantizar que corta al eje x en un único punto c_k.

1.4.2. Método de Interpolación

El método de interpolación al igual que el anterior usa una parábola que para por los tres puntos a, b y c, pero en este caso es una par erguida cuya ecuación es

$$\alpha x^2 + \beta x + \gamma = 0 \tag{6}$$

donde

$\alpha = f[a, b, c]$
$\beta = f[a, b] - (a + b) f[a, b, c]$
$\gamma = f[a] - a\, f[a, b] + ab\, f[a, b, c]$

La solución de esta parábola es el estimado de un nuevo iterado c_k y puede ser obtenida mediante la resolvente

$$c_k = \frac{-\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha} \tag{7}$$

Un ejemplo de este procedimiento se presenta en la figure 1 donde c_k se ha indicado como c' para no confundirlo con c. El punto c en la figura se ha obtenido mediante interpolación lineal (segunda opción del algoritmo), pero también pudo obtenerse mediante la bisección del intervalo (primera opción del algoritmo).

La expresión (7) contiene dos soluciones para el nuevo iterado c_k, sin embargo, una de las soluciones pertenece al interv alo $[a, b]$. Esta restricción resuelve este inconveniente y finalmente (7) puede ser expresada de la siguiente forma

$$c_k = \bar{c} - \delta + \sqrt{\delta^2 + \Delta(\Delta/4 - \delta) - \zeta}\ \text{sign}(\delta) = \bar{c} + \frac{\Delta(\Delta/4 - \delta) - \zeta}{\delta + \sqrt{\delta^2 + \Delta(\Delta/4 - \delta) - \zeta}\ \text{sign}(\delta)} \tag{8}$$

donde

$\Delta = b - a$
$f[a, b] = [f(b) - f(a)]/\Delta$
$\bar{c} = (a + b)/2\ \ , c = \bar{c}$ ó $c = a - f(a)/f[a, b]$
$f[a, b, c] = \{f[a, b] - [f(b) - f(c)]/(b - c)\}/(a - c)$
$\delta = \frac{1}{2} f[a, b]/f[a, b, c]$
$\zeta = f(a)/f[a, b, c]$
$Sign(\delta) = \delta/|\delta|$

El discriminante $\beta^2 - 4\alpha\gamma$ en la solución de la ecuación (7) siempre tiene un valor positivo debido a la condición $f(a).f(b) \leq 0$ que fuerza a la parábola, conteniendo los puntos $(a, f(a))$ y $(b, f(b))$, intersectar la línea $f(x) = 0$ en dos puntos que representan dos soluciones reales distintas de la ecuación (7), una de las cuales pertenece al intervalo cerrado $[a, b]$. El único caso cuando el discriminante mencionado puede ser cero es cuando en alguno de los extremos del intervalo $[a, b]$ exista un mínimo o un máximo de la función parabólica (6), de otra forma el discriminante siempre tiene un valor positivo. Esto garantiza que la raíz cuadrada de la expresiones (7) y por consiguiente (8) está siempre definida en el conjunto de los números reales positivos. La forma final de (8), garantiza que denominador sea el más grande en valor absoluto, evitando la propagación del error de redondeo en la sustracción entre δ y la raíz.

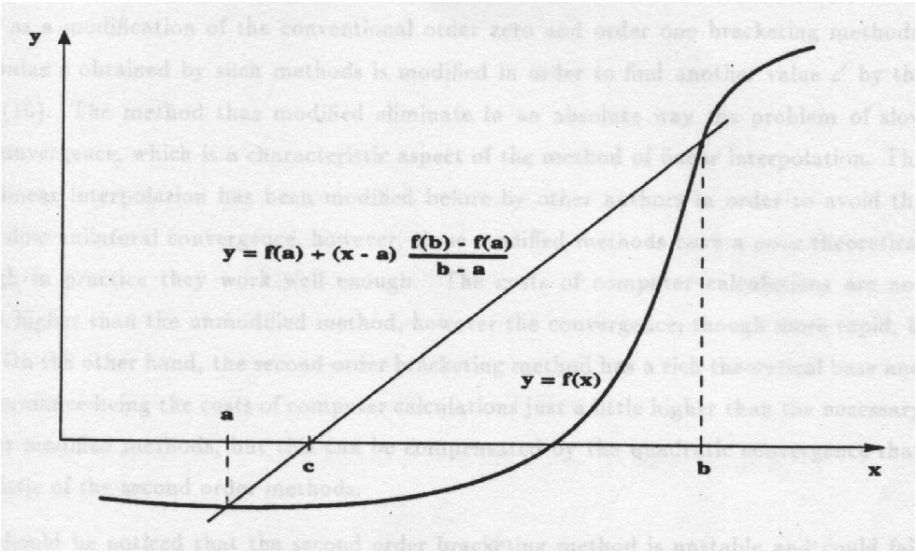


Figura 1.a. Método de Interpolación Lineal mostrando el cálculo del nuevo iterado.

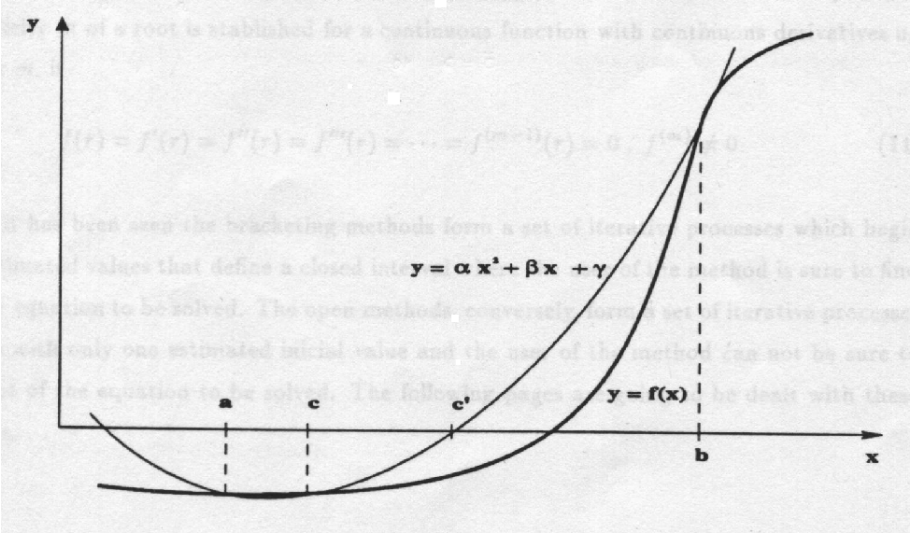


Figura 1.b. Método cerrado de Segundo Orden mostrando la interpolación parabólica.

2. METODOS ABIERTOS

2.1. PUNTO FIJO

Los métodos estudiados anteriormente consisten en la verificación del teorema de Bolzano, el cual nos garantiza la existencia de al menos una raíz. Ahora se presenta un método que no hace uso de tal teorema y que necesita un solo punto inicial para comenzar el proceso iterativo.

La ecuación básica a resolver es en todo momento del tipo

$$f(x) = 0 \tag{1}$$

la cual puede ser manipulada algebraicamente y reescribirla de la forma

$$x = g(x) \tag{2}$$

Ya cumplido este paso, escogemos un "punto arbitrario" x_0 y comenzamos a evaluar dicha expresión en forma iterativa, es decir

$$\begin{aligned} x_1 &= g(x_o) \\ x_2 &= g(x_1) \\ x_3 &= g(x_2) \\ &\vdots \\ x_k &= g(x_{k-1}) \\ x_{k+1} &= g(x_k) \end{aligned}$$

Expresión que, si converge, llegará a verificar la expresión

$$r = g(r) \tag{3}$$

donde r es la raíz buscada de la expresión $f(r) = 0$ y que para este método se denomina el *punto fijo*.

El problema de hallar la raíz como el punto de corte entre la función $f(x)$ y el eje de las abscisas, se ha transformado en hallar el punto de intersección entre la curva $g(x)$ y la recta identidad $y = x$. La convergencia a una raíz r puede ser unilateral si $g'(r) > 0$ o bilateral si $g'(r) < 0$.

EJEMPLO:

Hallar las raíces del polinomio $x^2 - 4x + 3$. Entre las posibles expresiones para $g(x)$ se tienen:

$x_0 = 6$	$x = \frac{x^2+1}{4}$	$x = \sqrt{4x-3}$	$x = \frac{3}{4-x}$
x_1	9.75	4.5825	−1.5
x_2	24.51	3.9154	0.5454
x_3	151.004	3.5583	0.8684
x_4	5701.0111	3.3516	0.9579
x_5		3.2259	0.9862
x_6		3.1470	0.9954
⋮	⋮	⋮	⋮
x_∞	diverge	3.0000	1.0000

Con lo que se muestra que no todos los esquemas convergen a una raíz, como se observa en el primer despeje, el cual diverge. Se puede probar que el tercer despeje diverge si el punto de partida está fuera del intervalo $[-3, 3]$.

Designemos al *error global* como $e_k = x_k - r$, entonces se tiene que

$$e_{k+1} = g'(\zeta)\, e_k \qquad g'(\zeta) = \frac{g(x_k) - g(r)}{x_k - r} \qquad \zeta \in [x_k, r] \tag{4}$$

Entonces el método de punto fijo converge cuando

$$|g'(\zeta)| < 1 \qquad \zeta \in [x_k, r] \tag{5}$$

Es decir, cuando la aplicación g es una *contracción*, o sea, aplicada sobre un subdominio de g alrededor de ζ lo contrae.

2.2. ACELERACION DE AITKEN

La aceleración d Aitken se basa en la suposición de que la pendiente de la curva $y = g(x)$ varía muy poco y, por lo tanto, se puede considerar que

$$g'(\zeta) = \frac{g(x_k) - g(r)}{x_k - r} \approx \frac{g(x_{k+1}) - g(x_k)}{x_{k+1} - x_k} = \frac{x_{k+2} - x_{k+1}}{x_{k+1} - x_k} \tag{1}$$

Substituyendo $r = g(r)$, $x_{k+1} = g(x_k)$ y despejando r de la expresión anterior se obtiene

$$\tilde{r} = x_k - \frac{(x_{k+1} - x_k)^2}{x_{k+2} - 2x_{k+1} - x_k} = x_k - \frac{(\Delta x_k)^2}{\Delta^2 x_k} \tag{2}$$

O sea, calculando las diferencias adelantadas de primer orden $\Delta x_k = x_{k+1} - x_k$ y de segundo orden $\Delta^2 x_k = \Delta x_{k+1} - \Delta x_k = x_{k+2} - 2x_{k+1} - x_k$, en dos iteraciones consecutivas con el método de punto fijo a partir de

x_k, se obtiene un valor $\tilde{r}$ dado por la fórmula (2), donde la convergencia hacia r se habrá acelerado (Una vez comprobada la convergencia $\Delta x_{k+1} < \Delta x_k$. En caso contrario, habrá una aceleración de la divergencia).

2.3. METODO DE LA SECANTE

Al igual que el método de interpolación lineal, el método de la secante supondrá que la función se comporta en forma lineal, pero no hará uso del teorema de Bolzano.

El método interpolará o extrapolará, según sea el caso, utilzando la misma ecuación empleada en el algoritmo de interpolación lineal, pero los puntos involucrados no se seleccionan mediante el cambio de signo de la función $f(x)$ necesariamente, sino en forma secuencial:

Dados x_0 y x_1 se obtiene x_2

con x_1 y x_2 se obtiene x_3

con x_3 y x_4 se obtiene x_5

$\vdots$

con x_{k-2} y x_{k-1} se obtiene x_k

con x_{k-1} y x_k se obtiene x_{k+1}

Por lo que la expresión para determinar el estimado $k+1$ de la raíz es:

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \tag{1}$$

Vemos que el segundo factor del segundo término de (1) es el inverso de la derivada media entre x_{k-1} y x_k

$$f'(\zeta_k) = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \qquad \zeta_k \in [x_k, x_{k-1}] \tag{2}$$

El algoritmo puede ser resumido de la siguiente forma:

1. Se escogen dos puntos cualesquiera x_0 y x_1. Se recomienda verifiquen el teorema de Bolzano, pero no es condición necesaria. Se escoge un número máximo de iteraciones k_{max}.
2. Se determina el valor de x_{k+1} con la expresión (1).
3. Se evalúa el error local y la desviación global mediante las expresiones:

   Error local: $\epsilon_k = x_k - x_{k-1}$.

   Desviación global: $d_k = f(x_k)$.
4. Se verifica si el valor absoluto del error local ϵ_k y la desviación global d_k son menores que la tolerancias seleccionadas ϵ_{max} y d_{max}. En caso afirmativo se detiene el proceso de cálculo y el valor deseado de la raíz r es igual al valor de x_k, $r = x_k$. En caso contrario se vuelve al punto 2.

EJEMPLO:

Hallar la primera raíz mayor que cero de la expresión

$$\cosh(x)\,\cos(x) = 1$$

Tomando $x_0 = 4.7$ y $x_1 = 6.2$ se obtienen los siguientes resultados:

$$x_2 = 4.7102 \qquad x_3 = 4.7170 \qquad x_4 = 4.7303 \qquad x_5 = 4.7300 \qquad x_6 = 4.7300$$

La raíz buscada es r= 4.7300.

La velocidad de convergencia de este método viene determinada por las siguientes relaciones

$$e_{k+1} = C\, e_k e_{k-1} \qquad C \approx \tfrac{1}{2}\, f''(r)/f'(r) \qquad |e_{k+1}| = A\,|e_k|^p \qquad |e_k|^p = |C|\, A^{-(1+1/p)}\, |e_k|^{1+1/p} \quad (3)$$

donde, igualando los últimos exponentes y miembros, $p = \varphi = (1+\sqrt{5})/2$ resulta ser la proporción áurea (solución positiva) y $A = |C|^{\frac{p}{p+1}} = |C|^{p-1}$ [Moheuddin et al.,(2019)] [Ostrowski,1966;Ch.3,12] [Dahlquist & Björck,1974;§.6.4.2,pp.228-229].

2.4. METODO DE NEWTON

2.4.1. Simple

Al igual que el algoritmo de punto fijo, el método de Newton [1736] utiliza un solo punto de partida, pero con una aproximación mucho mas refinada.

La idea fundamental consiste en extrapolar la raíz de la ecuación $f(x) = 0$ como el punto de corte de la recta tangente a $f(x)$ en x_o con el eje de las abscisas, siendo x_o un "punto arbitrario" a escoger. Dado que x_o no se conoce en forma exacta se utiliza la aproximación en forma recurrente hasta obtener convergencia.

Ya descrito el método, intuitivamente, se procede a deducir la fórmula de recurrencia. La ecuación de la recta tangente a $f(x)$ en x_o se puede expresar como

$$f(x) - f(x_o) = f'(x_o)(x - x_o) \quad (1)$$

que particularizada para el punto de corte con el eje de las abscisas queda de la siguiente forma

$$x = x_o - \frac{f(x_o)}{f'(x_o)} = g(x_o) \quad (2)$$

la cual se puede decir que tiene la forma del método de punto fijo con $x_{k+1} = g(x_k)$, si el valor de x substituye a x_o y se vuelve a introducir en la ecuación (2) de forma iterativa.

Se obtiene así el método iterativo de Newton-Raphson ([Raphson,1690] antes que Newton 46 años)

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad (3)$$

donde se dice que este método es del tipo *punto fijo* si denotamos

$$g(x) = x - \frac{f(x)}{f'(x)} \qquad g'(x) = \frac{f(x)\, f''(x)}{[f'(x)]^2} \qquad g''(x) = \frac{f''(x)}{f'(x)} + \frac{f(x)\, f'''(x)}{[f'(x)]^2} - \frac{2\, f(x)\, [f''(x)]^2}{[f'(x)]^3} \quad (4)$$

y se cumple 2.1.(3) que $r = g(r)$ es el punto fijo y 2.1.(5) que $|g'(\zeta)| < 1$ para la convergencia.

Los criterios de parada de este método se basan en el error local ϵ_k

$$\epsilon^k = x^k - x^{k-1} \qquad |\epsilon_k| < \epsilon_{max} \quad (5.a)$$

y la desviación global d_k

$$d^k = f(x^k) \qquad |d^k| < d_{max} \quad (5.b)$$

El error global sería $e_k = x_k - r$ y la deviación local sería $\delta_k = f(x_k) - f(x_{k-1})$.

EJEMPLO:

Hallar la raíz con el método de Newton cercana a $x = 1$ de la expresión:

$$f(x) = \mathrm{e}^x - 3x^2 = 0 \qquad f'(x) = \mathrm{e}^x - 6x$$

Tomando $x_0 = 1.0$ se obtienen los siguientes resultados:

$$x_1 = 0.91415$$
$$x_2 = 0.9100177$$
$$x_3 = 0.91000776$$

La raíz buscada es r= 0.91000.

Hagamos la expansión en series de Taylor de $g(x)$ alrededor de r, hasta el término de segundo orden

$$g(x_k) = g(r) + (x_k - r)\, g'(r) + (x_k - r)^2\, \frac{g''(\zeta)}{2} \qquad \zeta \in [r, x_k] \tag{6}$$

Puesto que por (4.b) $g'(r) = 0$ ya que $f(r) = 0$, el término de primer orden de la serie truncada (6) es nulo, y realizando las substituciones necesarias en el error global $e_k = x_k - r$, se tiene que

$$e_{k+1} = \frac{g''(\zeta)}{2}\, e_k^2 \tag{7}$$

Lo que significa que el método de Newton posee una velocidad de convergencia cuadrática si $f'(r) \neq 0$. Para una raíz r de multiplicidad mayor que 1, $g'(r)$ ya no se anula en (6) y por consiguiente la velocidad de convergencia del método de Newton vuelve a ser lineal.

Hagamos ahora la expasión de la serie de Taylor de la función $f(x)$ alrededor de x y evaluada en r

$$f(r) = f(x) + (r - x)f'(x) + e^2\, \frac{f''(\zeta)}{2} = 0 \qquad \zeta \in [r, x] \tag{8}$$

donde $e = x - r$ y $f(r) = 0$ (el valor de ζ aquí es diferente que en (6)). Evaluada en $x = x_k$, esta expresión da

$$r = x_k - \frac{f(x_k)}{f'(x_k)} - e_k^2\, \frac{f''(\zeta_k)}{2\, f'(x_k)} \qquad \Longrightarrow \qquad e_{k+1} = \frac{f''(\zeta_k)}{2\, f'(x_k)}\, e_k^2 \qquad \zeta_k \in I\!I = [r, x_k] \tag{9}$$

donde el error global $e_k = x_k - r$ y ζ_k permite agrupar el último término. Colocando $e_{k+1} = x_{k+1} - r$, donde $x_{k+1} = x_k - f(x_k)/f'(x_k)$, se obtiene que la fórmula algorítmica del método de Newton (3) tiene una velocidad de convergencia cuadrática (9.b), salvo cuando la raíz tiene multiplicidad mayor que la unidad (entonces la convergencia es lineal $e_{k+1} = g'(\zeta)\, e_k$ y las expansiones en series de Taylor deben hacerse hasta el término de primer orden $e_k = f(x_k)/f'(\zeta_k)$ y $e_{k+1} = e_k - f(x_k)/f'(x_k)$). Las expresiones (7) y (9.b) son equivalentes si se evalúa (4.c) adecuadamente ($f(r) = 0$, $f'(x_k)$ y $f''(\zeta_k)$). Basado en la desigualdad

$$\frac{|e_{k+1}|}{|e_k|} = \frac{1}{2}\left|\frac{f''(\zeta_k)}{f'(x_k)}\right| |e_k| < \frac{M}{2m}\, |e_k| < 1 \qquad \text{donde} \qquad \begin{cases} M = \sup\limits_{x \in I\!I} [|f''(x)|] \\ m = \inf\limits_{x \in I\!I} [|f'(x)|] \end{cases} \tag{9'}$$

obtenida de (9.b), de la última resulta que el método de Newton converge de forma segura si $|e_k| < 2m/M$.

2.4.2. Relajado

La fórmula algorítmica del método de Newton se relaja de la siguiente forma

$$x_k = x_k - \omega\, \frac{f(x_k)}{f'(x_k)} \tag{10}$$

donde se dice que el método está

$\omega < 1$ Subrelajado.

$\omega = 0$ Simple.

$\omega > 1$ Sobrerelajado.

La multiplicidad m de una raíz se establece para una función continua, con cotinuas derivadas hasta del orden m, si

$$f(r) = f'(r) = f''(r) = f'''(r) = \cdots = f^{(m-1)}(r) = 0\ ,\ f^{(m)} \neq 0 \tag{11}$$

Se sabe que para una raíz de multiplicidad m el límite

$$g'(r) = \lim_{x \to r} g'(x) = \lim_{x \to r} \left\{ \frac{f(x)\, f''(x)}{[f'(x)]^2} \right\} = \frac{m-1}{m} \tag{12}$$

donde para obtener este resultado, se ha tenido que aplicar la Regla de L'Hopital para la indeterminación de tipo 0/0.

Para el método de Newton, si se escoge el factor de relajación igual que la multiplicidad, $\omega = m$ la función

$$g(x) = x - m\, \frac{f(x)}{f'(x)} \tag{13}$$

tiene derivada $g'(r)$ siempre nula debido al límite (12), lo que afirma que con esta sobrerelajación siempre la convergencia será de velocidad cuadrática.

Para el método de Newton relajado la función $g(x)$, considerándolo un caso particular de un método de punto fijo, cambia a

$$g(x) = x - \omega\, \frac{f(x)}{f'(x)} \qquad\qquad g'(x) = 1 - \omega + \omega\, \frac{f(x)\, f''(x)}{[f'(x)]^2} \tag{14}$$

y el método converge de forma segura cuando $|g'(\zeta)| < 1$, $\zeta \in [r, x]$. Cuando $\omega = m$ la función $g'(r) = 0$ siempre debido al límite (12), como ya se indicó.

Para evitar la indeterminación antes mencionada en el caso de raíces de multiplicidades mayores que uno, Ralston & Rabinowitz [1978] propusiero un método de Newton modificado que consiste en definir una nueva función $U(x)$

$$U(x) = \frac{f(x)}{f'(x)} \qquad\qquad g(x) = x - \frac{U(x)}{U'(x)} \tag{15}$$

que como parece obvio, posee las misma raíces que la función $f(x)$. Por consiguiente, si se aplica el método de Newton a la función $U(x)$, se obtienen los mismos resultados que para la función $f(x)$. Esto es

$$x_{k+1} = x_k - \frac{U(x_k)}{U'(x_k)} \tag{16}$$

Para expresar esta fórmula recurrente en función de $f(x)$ y sus derivadas, vamos a hallar $U'(x)$

$$U'(x) = \frac{[f'(x)]^2 - f(x)\, f''(x)}{[f'(x)]^2} = 1 - \frac{f(x)\, f''(x)}{[f'(x)]^2} \tag{17}$$

Substituyendo $U(x)$ y $U'(x)$ en la fórmula recurrente (16), se obtiene

$$x_{k+1} = x_k - \frac{f(x_k)\, f'(x_k)}{[f'(x_k)]^2 - f(x_k)\, f''(x_k)} \tag{18}$$

Así que $g'(r)$ es siempre nula para este método sin importar la multiplicidad de la raíz r, lo que se demuestra aplicando la regla de L'Hopital sucesivamente a (15.b). Por lo tanto el método de Newton modificado (16) siempre converge cuadráticamente.

Una generalización de esta idea se consigue haciendo $U(x) = \sqrt[m]{f(x)}$, con lo que se obtiene equivalentemente el método (13) a partir de $(15)-(16)$, al que denominaremos método de Newton relajado para raíces de multiplicidad m y la función $g(x)$ la denotaremos [Gilbert,(2001)]

$$N_m(x) = x - \frac{f(x)^{\frac{1}{m}}}{\frac{1}{m}\,[f(x)]^{\frac{1}{m}-1}\,f'(x)} = x - m\,\frac{f(x)}{f'(x)} \tag{19}$$

Para distinguir un método de otro, denotaremos la función $g(x)$ de $(15)-(18)$ como

$$M(x) = x - \frac{f(x)\,f'(x)}{[f(x)]^2 - f(x)\,f''(x)} \tag{20}$$

De estos resulta el *método de Collatz*, como el promedio de los métodos anteriores

$$C_m(x) = \frac{N_m(x) + M(x)}{2} = x - \frac{f(x)\,\{\,(m+1)\,[f'(x)]^2 - m\,f(x)\,f''(x)\,\}}{2\,f'(x)\,\{\,[f'(x)]^2 - f(x)\,f''(x)\,\}} \tag{21}$$

Si r es la raíz de $f(x)$ de multiplicidad k, entonces

$$C'_m(r) = \frac{N'_m(r) + M'(r)}{2} = \frac{k-m}{2k} \tag{22}$$

de aquí que, si $m > 3k$, la raíz será repelida, si $m < 3k$ pero $m \neq k$, la convergencia será lineal y, si $m = 3k$, la raíz es un punto fijo neutral. Cualquier solución a $(m+1)\,[f'(x)]^2 - m\,f(x)\,f''(x) = 0$, que no son raíces de $f(x)$, son raíces extrañas por este método. Deben ser revisadas para ver si son atractivas o no. El punto en el infinito es siempre un punto fijo repulsivo con autovalor $2d/(d-m)$, donde d es el grado del polinomio $f(x)$.

El *método de Schröder* es una modificación del método de Newton que converge a una raíz de cualquier multiplicidad. Sea x_n una aproximación de una raíz de $f(x)$ en una vecindad donde la derivada $f'(x)$ no es nula. Se puede aproximar $y = f(x)$ mediante un polinomio $x = h(y) = a_0 + a_1 y + a_2 y^2 + \cdots + a_{r-1} y^{r-1}$, tal que las dos curvas se asemejen en x_n hasta la $(r-1)$ derivadas. La curva $x = h(y)$ se encuentra con el eje-x en a_0 y se toma ese punto como la siguiente aproximación x_{n+1} hacia la raíz. Esto define el método iterativo $S_r(x)$ de orden r. El método de Schröder, $S_r(x)$, converge a una raíz simple con orden r. Cuando $r = 2$ la curva se aproxima por una tangente en x_n y así $S_2(x)$ es justamente el método de Newton estándar $N_1(x)$. Cuando $r = 3$, el método se define por

$$S_3(x) = x - \frac{f(x)}{f'(x)} - \frac{f''(x)\,[f(x)]^2}{2\,[f'(x)]^3} = x - \frac{f(x)\,\{\,2\,[f'(x)]^2 + f(x)\,f''(x)\,\}}{2\,[f'(x)]^3} \tag{23}$$

En una raíz de multiplicidad k, el autovalor de $S_3(x)$ es $(k-1)(2k+1)/(2k^2)$ así que $S_3(x)$ converge linealmente a una raíz múltiple. Puntos fijos extraños de este método son las soluciones de $2\,[f'(x)]^2 + f(x)\,f''(x) = 0$, que no son raíces de $f(x)$. El punto en el infinito es siempre un punto fijo repulsivo para $S_3(x)$ con autovalor $2d^3/(2d^3 - 3d^2 + d) > 1$, donde d es el grado del polinomio $f(x)$.

El *método de König* es el método de Newton aplicado a la función $U(x) = f(x)\,h(x)$ donde $h(x)$ es finita y no nula en una raíz simple de $f(x)$, entonces la iteración resultante será siempre al menos de segundo orden en esa raíz simple. En el método de König de orden r, $K_r(x)$, la función $h(x)$ se escoge para que

la iteración converja con orden r hacia una raíz simple de $f(x)$. $K_2(x) = N_1(x)$, mientras que $K_3(x)$ debe satisfacer $K_3''(r) = 0$ en una raíz simple de $f(x)$, y tomamos $h(x) = 1/\sqrt{f'(x)}$. $K_3(x)$ se define por

$$K_3(x) = x - \frac{2\, f(x)\, f'(x)}{2\,[f'(x)]^2 - f(x)\, f''(x)} \tag{24}$$

Este método es también conocido como el *método de Halley*. En una raíz de multiplicidad k, el autovalor de $K_3(x)$ es $(k-1)/(k+1)$ y así converge linealmente hacia una raíz múltiple. Los únicos puntos fijos extraños son las raíces de $f'(x)$ que no son raíces de $f(x)$. Estos puntos fijos extraños son todos repulsivos. El punto en el infinito es siempre un punto fijo repulsivo para $K_r(x)$ con autovalor $(d + r - 2)/(d - 1)$, donde d es el grado del polinomio $f(x)$. Compárese $K_3(x)$ con Richmond 2.5.(4.b) y [Hildebrand,1974] ec.(10.12.20).

El *método de Steffensen* no requiere del cálculo de derivadas y sólo usa el valor inicial (método de tipo punto fijo). El método puede ser derivado de la aceleración de Aitken. Para encontrar las raíces de la función $f(x)$ se itera con

$$S_t(x) = x - \frac{f(x)^2}{f[x + f(x)] - f(x)} = x - \frac{f(x)}{h(x)} \qquad\qquad h(x) = \frac{f[x + f(x)] - f(x)}{f(x)} \tag{25}$$

Cuando el denominador $f[f(x) + x] - f(x) = 0$ se asume $S_t(x) = x$ y se detiene la iteración. El método converge cuadráticamente a una raíz simple y linealmente a una raíz múltiple y pueden haber puntos de no convergencia [Dahlquist & Björck,1974;§.6.4.4,p.230-232]. Requiere de dos evaluaciones funcionales en cada iteración.

2.5. METODOS DE SEGUNDO ORDEN

2.5.1. Método de Richmond

El método de Richmond se basa en la expansión de las series de Taylor de una función escalar, hasta el término de segundo orden [Gundersen,(1982)] [Lapidus,1962,§6.4-6.5,pp.292-297] [Richmond,(1944)]

$$f(r) = f(x) + (r - x)\, f'(x) + \tfrac{1}{2}\,(r - x)^2\, f''(x) + O(|e|^3) \tag{1}$$

donde $e = x - r$ es el error global de x. Reorganizando esta expresión factorizando $(r - x)$, se obtiene

$$[\, f'(x) + \tfrac{1}{2}\,(r - x)\, f''(x)\,]\,(r - x) = -f(x) - O(|e|^3) \tag{2}$$

Sin tener en cuenta el término $O(|e|^3)$ y cambiando luego r por x_{k+1} y x por x_k, se puede aplicar un procedimiento iterativo en la forma (el "orden" en realidad se refiere al grado del polinomio utilizado)

$$x_{k+1} = x_k - [\, f'(x_k) + \tfrac{1}{2} z_k\, f''(x_k)\,]^{-1} f(x_k) \tag{3}$$

donde z_k es el error local ϵ_{k+1} obtenido con el método de Newton 2.4.(3) para la iteración k, así

$$z_k = -\frac{f(x_k)}{f'(x_k)} \qquad\qquad x_{k+1} = x_k - \frac{2\, f(x_k)\, f'(x_k)}{2\,[f'(x_k)]^2 - f(x_k)\, f''(x_k)} \tag{4}$$

Debe notarse que en la expresión (2) la cantidad $(r - x)$, entre los corchetes, se ha substituido por un estimado ofrecido por el método de Newton (4.a), una sóla vez, para pasar de (3) a (4.b), en lugar de volverlo a corregir por un valor más preciso $z_k = (x_{k+1} - x_k)$, siguiendo un esquema iterativo de punto fijo (correcciones sucesivas). Este aspecto hace que el método de Richmond sea un método 'casi' de segundo orden (error del tercer orden). Nótese la similitud de (4.b) con 2.4.(18), salvo el factor de 2, y con 2.4.(24), $K_3(x)$.

Para solventar este último aspecto reformulamos la fórmula algorítmica (3) así

$$x_{k+1} = x_k + z_k \qquad z_{k,t+1} = -[\, f'(x_k) + \tfrac{1}{2} z_{k,t}\, f''(x_k)\,]^{-1} f(x_k) \tag{3'}$$

e implementamos un esquema iterativo secundario en t iteraciones internas de tipo punto fijo en $z_{k,t}$ (para cada k) con la ecuación $(3'.b)$, y el último valor obtenido de z_k de este proceso iterativo secundario (normalmente se usa un número t_{max} de iteraciones internas o correcciones sucesivas, de 3 a 6 son suficientes) es el que se introduce en $(3'.a)$. Se escoge el valor dado por la ecuación (4) como el valor o iterado inicial $z_{k,0}$ del proceso iterativo secundario. Se puede decir que el método así reformulado se vuelve un método predictor en la ecuación (4) y corrector en la ecuación $(3'.b)$, que si es completamente de segundo orden. Se podría decir que el método en $(3')$ para un t_{max} lo suficientemente alto es equivalente al método en la sección 2.5.3 adelante (versión 1), donde z_k se resuelve analíticamente con la resolvente de un polinomio de segundo grado.

Los criterios de parada para este método son los mismos que para el método de Newton ofrecidos por 2.4.(5).

2.5.2. Método de Muller

El método de Muller [(1956)] generaliza el método de la secante (sección 2.3), pero usa una interpolación cuadrática (parabólica) entre tres puntos en lugar de una interpolación lineal entre dos puntos. Resolviendo para los ceros de la parábola permite el método encontrar complejos pares de raíces. Dados tres iterados previos x_{k-2}, x_{k-1} y x_k y sus correspondientes valores de la función $f(x)$ en dichos puntos, la siguiente aproximación x_{k+1} se produce con la siguiente fórmula

$$x_{k+1} = x_k - (x_k - x_{k-1}) \left[\frac{2\, A_k}{B_k \mp \sqrt{D_k}} \right] \tag{5}$$

donde

$$\begin{aligned} A_k &= (1+q)\, f(x_k) \qquad\qquad q = \frac{x_k - x_{k-1}}{x_{k-1} - x_{k-2}} \\ B_k &= (2q+1)\, f(x_k) - (1+q)^2 f(x_{k-1}) + q^2 f(x_{k-2}) \\ C_k &= q\, f(x_k) - q\,(1+q)\, f(x_{k-1}) + q^2 f(x_{k-2}) \\ D_k &= B_k^2 - 4\, A_k C_k \end{aligned} \tag{6}$$

y donde el signo del denominador se escoge para hacer su valor absoluto o módulo (caso complejo) tan grande como sea posible. Se puede comenzar comenzar el proceso iterativo con tres valores cualesquiera, e.g. tres valores igualmente espaciados en la recta real. Nótese que el método debe permitir la posibilidad de un complejo en el denominador, y la subsecuente aritmética compleja.

El método de Muller fué creado en principio para hallar las raíces de un polinomio, pero ha sido usado con éxito con funciones analíticas en el plano complejo, por ejemplo in la rutina IMSL llamada ZANLYT [Press et al.,1986]. La velocidad de convergencia de este método viene determinada por las siguientes relaciones

$$e_{k+1} = C\, e_k\, e_{k-1}\, e_{k-2} \qquad C \approx \tfrac{1}{6}\, f'''(r)/f'(r) \qquad |e_{k+1}| = A\, |e_k|^p \tag{6'}$$

donde $A = |C|^{(p-1)/2}$ y $p \approx 1.84$ es la solución real positiva de la ecuación $p = 1+1/p+1/p^2$ [Hildebrand,1974]. Esta última referencia menciona métodos parecidos como Halley, Bailey, Lambert, Chebyshev, Traub, Frame.

2.5.3. Método de La Parábola Secante

El método de la parábola secante es una generalización del método de Newton. Si en lugar de usar una recta con pendiente igual a $f'(x_k)$ para estimar el nuevo iterado x_{k+1} como en el método de Newton, se usara una parábola (tangente con la curva e igual curvatura) con pendiente $f'(x_k)$ y segunda derivada $f''(x_k)$, iguales que la función $f(x)$ en el punto x_k, como muestra la figura 2, para proyectar el nuevo iterado x_{k+1}, donde dicha parábola corta al eje de las x, se obtendría el método de segundo orden de la parábola secante. Cambiar las derivadas por diferencias divididas reduce el orden de convergencia de $p = 3$ a 1.84.

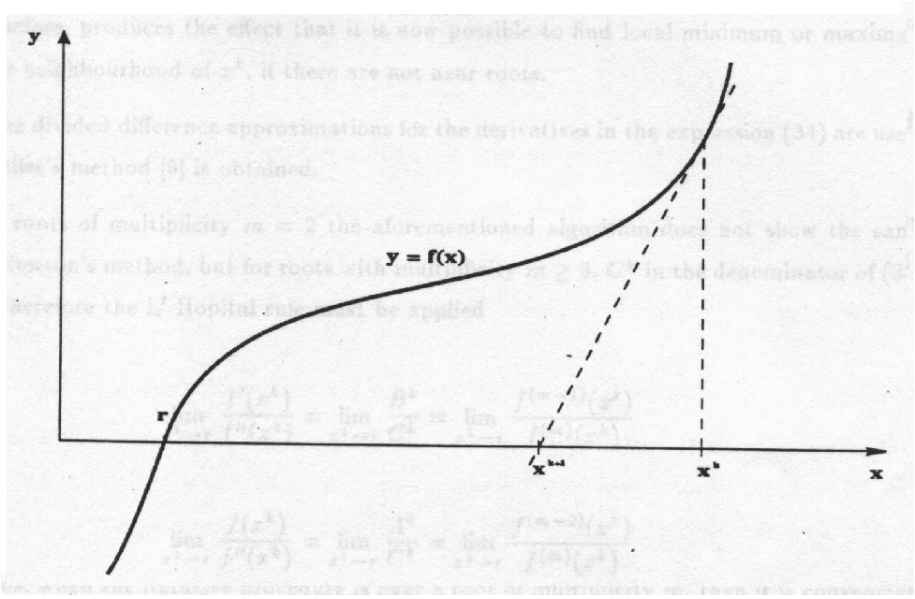


Figura 2.a. Método de Newton-Raphson mostrando el nuevo iterado.

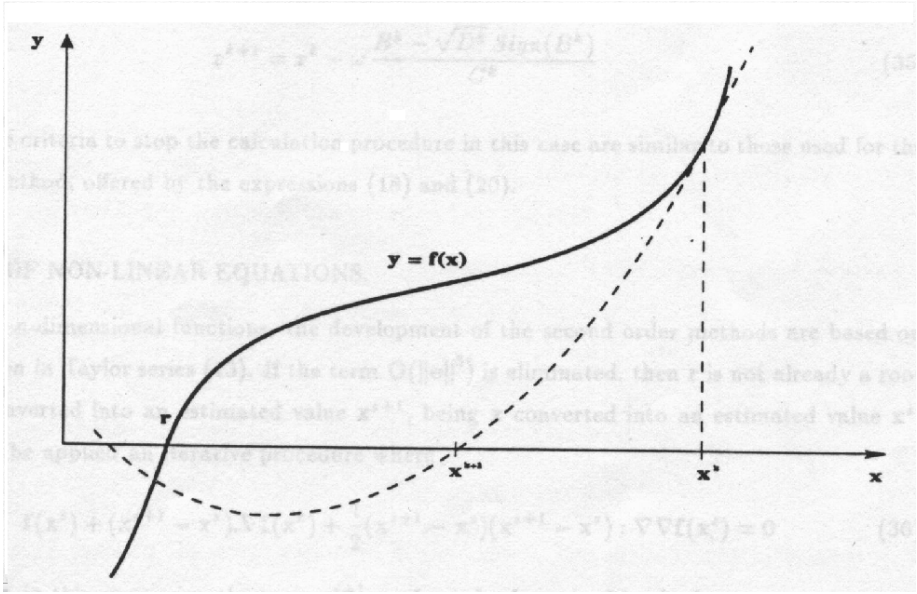


Figura 2.b. Método abierto de Segundo Orden mostrando la extrapolación parabólica.

Su ecuación algorítmica se basa en la expansión en series de taylor, evaluada en r alrededor de x, hasta el término de segundo orden

$$f(r) = f(x) + (r - x)f'(x) + \tfrac{1}{2}(r - x)^2 f''(x) + O(|e|^3) \tag{7}$$

donde $e = x - r$ es el error global del valor x relativo a la raíz r, y por definición de una raíz $f(r) \equiv 0$.

Eliminando el término con $O(|e|^3)$ y haciendo el cambio de r por x_{k+1} y x por x_k (siguiendo el mismo razonamiento que para el método de Newton), se obtiene la siguiente expresión

$$f(x_k) + (x_{k+1} - x_k)\, f'(x_k) + \tfrac{1}{2}\,(x_{k+1} - x_k)^2\, f''(x_k) = 0 \tag{8}$$

el cual representa la ecuación de un polinomio de segundo grado en $(x_{k+1} - x_k)$. Finalmente, resolviendo para x_{k+1} en la ecuación anterior se obtiene (racionalizando el numerador)

$$x_{k+1} = x_k - \frac{f'(x_k) \mp \sqrt{[f'(x_k)]^2 - 2\,f(x_k)f''(x_k)}}{f''(x_k)} = x_k - \frac{2\,f(x_k)}{f'(x_k) \pm \sqrt{[f'(x_k)]^2 - 2\,f(x_k)f''(x_k)}} \tag{9}$$

Se debe observar que existen dos soluciones para la ecuación (8). Un ejemplo gráfico puede observarse en la figura 2, donde este método es comparado con el método de Newton. Con la finalidad de resolver con el doble signo, se selecciona la solución para x_{k+1} más cercana a x_k. Esto se hace modificando la ecuación (9) en la forma

$$x_{k+1} = x_k - \frac{B_k - \sqrt{D_k}\,Sign(B_k)}{C_k} = x_k - \frac{2\,A_k/B_k}{1 + \sqrt{D_k/B_k^2}} \tag{10}$$

donde

$$\begin{aligned} A_k &= f(x_k) \\ B_k &= f'(x_k) \approx f[x_k, x_{k-1}] + (x_k - x_{k-1})\,f[x_k, x_{k-1}, x_{k-2}] \\ C_k &= f''(x_k) \approx 2\,f[x_k, x_{k-1}, x_{k-2}] \\ D_k &= [\![\, (B_k)^2 - 2\,A_k C_k \,]\!] \end{aligned} \tag{11}$$

Aquí existen dos versiones del método. Si se conocen las derivadas $f'(x_k)$ y $f''(x_k)$ (versión 1) o estas se estiman con base a los valores funcionales del punto x_k y de los dos puntos anteriores x_{k-1} y x_{k-2} mediante diferencias divididas (versión 2-Muller). En el el primer caso estaríamos hablado de una parábola con pendiente y segunda derivada igual que la función $f(x)$ en el punto x_k. En el segundo caso estamos hablando de un método equivalente al método de Muller de la sección anterior (aunque la (5) y última de (10) reduce mejor la propagación del error de redondeo). El discriminante D_k si es negativo se asume cero (eso es lo que significa el símbolo $[\![\,\cdot\,]\!]$, excepto cuando se trabaje en los complejos) y el resultado, en lugar de una solución de corte con la recta $y = 0$, da la ubicación del vértice de la parábola. En este último caso, el método también sirve para hallar mínimos o máximos. Finalmente, el método depende de dos parámetros en cada iteración A_k/B_k y A_kC_k/B_k^2. La raíz en el denominador de (10) se aproxima $\sqrt{1-2\alpha} \approx 1 - \alpha$ en los métodos de Richmond, Halley, Bailey, Lambert, König $K_3(x)$, que son realmente el mismo método [Hildebrand,1974;ec.(10.12.20)]. De la aproximación $(1-\alpha)^{-1} \approx 1 + \alpha$ en $K_3(x)$ resulta $S_3(x)$, método de Schröder ec.2.4.(23), también conocido como la fórmula de Chebyshev [idem;ec.(10.12.22)].

2.6. METODO DE BAIRSTOW

El método de *Bairstow* se fundamenta en hallar dos raíces simultáneamente en cada intento, para un polinomio $P_n(x)$ de grado n. En dicho intento se consiguen los coeficientes r y s de una parábola $x^2 - r\,x - s$ que divide de forma exacta a $P_n(x)$. Las dos raíces, $x_{1,2} = (r \pm \sqrt{\Delta})/2$, en el caso $s \geq -r^2/4$ descrito son reales, puesto que el discriminante $\Delta = r^2 + 4s \geq 0$. En el caso complejo, $s < -r^2/4$, el resultado da un discriminante $\Delta < 0$ negativo, por lo que las dos raíces halladas son complejas conjugadas.

Sea un polinomio $P_n(x)$ de grado n

$$\begin{aligned} P_n(x) &= a_n\,x^n + a_{n-1}\,x^{n-1} + \cdots + a_1\,x + a_0 \\ &= (\,x^2 - r\,x - s\,)\,(\,b_{n-2}\,x^{n-2} + b_{n-3}\,x^{n-3} + \cdots + b_1\,x + b_0\,) + b_{-1}\,(x - r) + b_{-2} \end{aligned} \tag{1}$$

se puede factorizar con una parábola $(x^2 - r\,x - s)$ y un polinomio $P_{n-2}(x)$ de forma exacta como se indicó arriba, cuando el residuo $b_{-1}(x - r) + b_{-2}$ correspondiente es nulo. El resultado de dividir por la parábola produce los coeficientes de $P_{n-2}(x)$ y el residuo como

$$\begin{aligned} b_{n-2} &= a_n \\ b_{n-3} &= a_{n-1} + r\,b_{n-2} \\ b_i &= a_{i+2} + r\,b_{i+1} + s\,b_{i+2} \qquad\qquad i = n-4, n-5, \ldots, 0, -1, -2 \end{aligned} \tag{2}$$

Definimos dos funciones objetivos f y g a anular, para así anular el residuo, dependientes de r y s, tales que

$$\begin{array}{lll} f(r,s) = b_{-1} & r^* = r + \Delta r & \begin{bmatrix} \partial f/\partial r & \partial f/\partial s \\ \partial g/\partial r & \partial g/\partial s \end{bmatrix} \begin{Bmatrix} \Delta r \\ \Delta s \end{Bmatrix} = - \begin{Bmatrix} f(r,s) \\ g(r,s) \end{Bmatrix} \\ g(r,s) = b_{-2} & s^* = s + \Delta s & \end{array} \tag{3}$$

El proceso iterativo (3.b) se realiza varias veces hasta que las funciones objetivo (3.a) se anulen. En cada iteración, Δr y Δs se obtienen de resolver el sistema de ecuaciones lineales (3.c), donde la matriz se actualiza en cada iteración. Este método resumido aquí, para resolver dos ecuaciones $f = 0$ y $g = 0$ con dos incógnitas r y s, es el método de Newton-Raphson que veremos más adelante en el próximo capítulo, sección II.2.2. (se puede deducir fácilmente con la expansión en series de Taylor de dos funciones f y g, ver apéndice, alrededor de r y s, hasta el término de primer orden, y anulando las funciones f y g en r^* y s^*).

Bairstow observó que las derivadas parciales requeridas en (3.c) pueden ser obtenidas de las b's, mediante una segunda división sintética por la misma parábola $x^2 - r\,x - s$, en la misma forma que las b's fueron obtenidas a partir de las a's. Definamos un conjunto de coeficientes c's por las relaciones siguientes, obtenidos de la segunda división por la parábola

$$\begin{aligned} c_{n-4} &= b_{n-2} \\ c_{n-5} &= b_{n-3} + r\,c_{n-4} \\ c_i &= b_{i+2} + r\,c_{i+1} + s\,c_{i+2} \qquad\qquad i = n-6, n-5, \ldots, 0, -1, -2, -3 \end{aligned} \tag{4}$$

y compárese con las dos siguientes columnas de derivadas parciales de la primera división por la parábola

$$\begin{array}{ll} \dfrac{\partial b_{n-2}}{\partial r} = 0 & \dfrac{\partial b_{n-2}}{\partial s} = 0 \\ \dfrac{\partial b_{n-3}}{\partial r} = b_{n-2} + r\,\dfrac{\partial b_{n-2}}{\partial r} = c_{n-4} & \dfrac{\partial b_{n-3}}{\partial s} = 0 \\ \dfrac{\partial b_i}{\partial r} = b_{i+1} + r\,\dfrac{\partial b_{i+1}}{\partial r} + s\,\dfrac{\partial b_{i+2}}{\partial r} = c_{i-1} & \dfrac{\partial b_i}{\partial s} = r\,\dfrac{\partial b_{i+1}}{\partial s} + s\,\dfrac{\partial b_{i+2}}{\partial s} + b_{i+2} = c_i \end{array} \tag{5}$$

con i recorriendo los mismo valores que (2).

Finalmente se obtiene que

$$\begin{array}{ll} \dfrac{\partial f}{\partial r} = \dfrac{\partial b_{-1}}{\partial r} = c_{-2} & \dfrac{\partial f}{\partial s} = \dfrac{\partial b_{-1}}{\partial s} = c_{-1} \\ \dfrac{\partial g}{\partial r} = \dfrac{\partial b_{-2}}{\partial r} = c_{-3} & \dfrac{\partial g}{\partial s} = \dfrac{\partial b_{-2}}{\partial s} = c_{-2} \end{array} \tag{6}$$

por lo que el sistema de ecuaciones lineales (3.c), queda como

$$\begin{bmatrix} c_{-2} & c_{-1} \\ c_{-3} & c_{-2} \end{bmatrix} \begin{Bmatrix} \Delta r \\ \Delta s \end{Bmatrix} = - \begin{Bmatrix} b_{-1} \\ b_{-2} \end{Bmatrix} \tag{7}$$

y es más fácil realizar el proceso iterativo y actualizar las b's y las c's en cada iteración [Ralston & Rabinowitz,1978] [Gerald,1978].

BIBLIOGRAFIA

[1] Brent, R. P. **Algorithms for Minimization without Derivatives**. Prentice-Hall (Englewood Cliffs, N. J.), 1973.

[2] Burden R. L.; Faires, J. D. **Numerical Analysis**. 3^{rd} Edition. PWS. Boston, 1985.

[3] Carnahan, B.; Luther, H. A.; Wilkes, J. O. **Applied Numerical Methods**. John Wiley & Sons, 1969.

[4] Dahlquist, G.; Björck, Å. **Numerical Methods**. Dover Publications (New York), 2003. Prentice-Hall, 1974.

[5] Ford, J. A. **Improved Algorithms of Illinois-type for the Numerical Solution of Nonlinear Equations**, Technical Report CSM-257, University of Essex Press, 1995.

[6] Gerald, C. F. **Applied Numerical Analysis**. 2^{nd} Edition. Addison-Wesley, 1978.

[7] Gilbert, W. J. "Generalizations of Newton's Method". **Fractals**, Vol.**9**, No.3, pp.251-262, (2001).

[8] Granados M., A.L. **Second Order Method for Solving Non-Linear Equations**. INTEVEP S.A. Reporte Técnico No. INT-EPPR/322-91-0002. Los Teques, Junio de 1991.

[9] Gundersen, T. "Numerical Aspects of the Implementation of Cubic Equations of State in Flash Calculation Routines". **Computer and Chemical Engineering**, Vol.**6**, No.3, pp.245-255, (1982).

[10] Hildebrand, F. B. **Introduction to Numerical Analysis**, 2^{nd} Edition. Dover Publications (New York), 1974.

[11] Householder, A. S. **The Numerical Treatment of a Single Nonlinear Equation**. McGraw-Hill (New York), 1970.

[12] King, R. F. "Anderson-Bjorck for Linear Sequences". **Mathematics of Computation**, Vol.**41**, No.164, pp.591596, October, (1983).

[13] Lapidus, L. **Digital Computation for Chemical Engineers**. McGraw-Hill (New York), 1962.

[14] Moheuddin, Mir Md.; Jashim Uddin, Md.; Kowsher, Md. "A New Study to Find Out The Best Computational Method for Solving The Nonlinear Equation". **Applied Mathematics and Sciences**, Vol.**6**, No.2/3, pp.15-31, (2019).

[15] Muller, D. E. "A Method of Solving Algebraic Equations Using an Automatic Computer". **Mathematical Tables and Other Aids to Computation (MTAC)**. Vol.**10**, pp.208-215, (1956).

[16] Newton, I. bf De metodis fluxionum et serierum infinitarum, Méthod of Fluxions and Infinite Series. John Colson, 1736.

[17] Oliveira, I. F. D.; Takahashi, R. H. C. "An Enhancement of The Bisection Method Average Performance Preserving Minmax Optimality". **ACM Transactions on Mathematical Software**, Vol.**47**, No.1, pp.5:15:24, (2021).

[18] Ostrowski, A. M. **Solution of Equations and Systems of Equations**, 2^{nd} Edition. Academic Press (New York), 1966.

[19] Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. **Numerical Recipes. The Art of Scientific Computing**. Cambridge University Press, 1986.

[20] Pundir, S. K. **Applied Numerical Analysis**. CBS Publisher & Distributors Pvt. Ltd., 2023.

[21] Ralston, A.; Rabinowitz, P. **A First Course in Numerical Analysis**, 2^{nd} Edition. McGraw-Hill (New York), 1978.

[22] Raphson, J. **Aequationum Universalis**. Royal Society, 1690.

[23] Richmond, H. W. "On Certain Formulae for Numerical Approximation", **J. London Math. Soc.**, Vol.**19**, Issue 73 Part 1, pp.31-38, January (1944).

CAPITULO II

SOLUCION DE SISTEMAS DE ECUACIONES

CONTENIDO

Al momento de resolver problemas de ingeniería, es frecuente encontrar un sistema de ecuaciones algebraicas que representa la solución deseada. Estos sistemas de ecuaciones pueden ser lineales o no-lineales,

segun sea la categoría del problema o la rama a la cual pertenece. En todo momento estos sistemas representan ecuaciones algebraicas.

Como ejemplo de sistemas equaciones algebraicas se pueden citar:

- Resolver una red eléctrica formada por resistencias, la cual origina un sistema de ecuaciones lineales (frecuentemente) para las intensidades que circulan por el circuito.
- Cuando se desea correlacionar un conjunto de datos experimentales o resultados numéricos, es frecuente hacer uso del método de los mínimos cuadrados,el cual origina un sistema de ecuaciones para las constantes de la expresión atrabajar. El sistema de ecuaciones obtenido puede ser líneal o no-líneal dependiendo de la complejidad de la aproximación propuesta.
- Al resolver ecuaciones diferenciales parciales u ordinarias con valor en el contorno se hace uso del método de las diferencias finitas (entre otros) , originando un sistema de ecuaciones líneal o no-líneal dependiendo de las aproximaciones utilizadas o de la misma ecuación diferencial.
- Al momento de obtener los caudales que circulan por una red de tuberías, se presenta un sistema de ecuaciones no-líneal para los caudales y/o las alturas piezométricas en los puntos de unión.

Aquí se tratarán ambos tipos de sistemas de ecuaciones, comenzando con los sistemas lineales, de forma de facilitar la descripción de los algoritmos de solución para los sistemas de ecuaciones no-lineales.

1. SISTEMAS LINEALES

1.1. METODOS DIRECTOS

1.1.1. Eliminación Simple

Este método se basa en las propiedades de una matriz cuadrada, principalmente la que establece que al sumar una fila a otra se mantiene la independencia entre las mismas, es decir, que el determinante no cambia.

El método consiste en que dado un sistema de ecuaciones lineales, que pueda ser representado mediante

$$[\mathbf{A}]\mathbf{x} = \mathbf{b} \tag{1}$$

donde

$[\mathbf{A}]$ es la matriz de coeficientes del sistema,

$\mathbf{x}$ es el vector de incógnitas del problema,

$\mathbf{b}$ es el vector idependiente.

realizar operaciones de adición sustracción entre las filas de forma sistemática, sobre la matriz $[\mathbf{A}]$ y el vector $\mathbf{b}$, hasta obtener una matriz triangular superior $[\mathbf{U}]$ en el lugar de $[\mathbf{A}]$.

Para facilitar las operaciones, se acostumbra a expresar de forma completa la matriz de todo el sistema en un matriz ampliada con una columna adicional, en todo un conjunto, que se denomina *matriz ampliada*, de la forma

$$[\mathbf{A}|\mathbf{b}] \tag{2}$$

de manera que es más fácil exponer las operaciones que sobre ella se realizan para resolver el sistema de ecuaciones. Las operaciones se realizan sobre "toda" la matriz ampliada, para que el sistema de ecuaciones lineales original quede "inalterado".

EJEMPLO:

Hallar la solución del siguente sistema de ecuaciones:

$$2.51\,x_1 + 1.48\,x_2 + 4.53\,x_3 = 0.05$$

$$1.48\,x_1 + 0.93\,x_2 - 1.30\,x_3 = 1.03$$

$$2.68\,x_1 + 3.04\,x_2 - 1.48\,x_3 = -0.53$$

El cual puede ser escrito en forma de matriz ampliada:

$$\left[\begin{array}{ccc|c} 2.51 & 1.48 & 4.53 & 0.05 \\ 1.48 & 0.93 & -1.30 & 1.03 \\ 2.68 & 3.04 & -1.48 & -0.53 \end{array}\right].$$

Se multiplica la primera fila por $M_{21} = -\frac{1.48}{2.51}$ y se le suma a la segunda fila.

Se multiplica la primera fila por $M_{31} = -\frac{2.68}{2.51}$ y se le suma a la tercera fila.

$$\left[\begin{array}{ccc|c} 2.51 & 1.48 & 4.53 & 0.05 \\ 0.00 & 0.059 & -3.96 & 1.00 \\ 0.00 & 1.48 & -6.98 & -0.583 \end{array}\right].$$

Ya culminado el proceso de eliminación para la primera columna se procede con la segunda.

Se multiplica la segunda fila por $M_{32} = -\frac{1.48}{0.059}$ y se le suma a la tercera fila

$$\left[\begin{array}{ccc|c} 2.51 & 1.48 & 4.53 & 0.05 \\ 0.00 & 0.059 & -3.96 & 1.00 \\ 0.00 & 0.00 & 92.7 & -25.5 \end{array}\right].$$

En donde se ha obtenido una matriz triangular superior, que en notación expandida queda

$$\begin{aligned} 2.51\,x_1 + 1.48\,x_2 + 4.53\,x_3 &= 0.05 \\ 0.059\,x_2 - 3.96\,x_3 &= 1.00 \\ 92.7\,x_3 &= -25.5 \end{aligned}$$

de donde despejando de forma ascendente y regresiva se obtiene la siguiente solución

$$\begin{aligned} x_3 &= -0.275 \\ x_2 &= -1.35 \\ x_1 &= 1.30 \end{aligned}$$

Presentado el ejemplo es posible entonces organizar el algoritmo de la siguiente forma:

ALGORITMO:

1.- Dado un sistema de n ecuaciones lineales con n incógnitas, con matriz cuadrada, se construye su matriz ampliada.

2.- Se inicia el proceso de eliminación desde la columna $k = 1$ hasta la columna $k = n - 1$. La última columna $k = n$ no es necesario eliminarla.

3.- Se evaluán los multiplicadores $M_{ik} = -\frac{A_{ik}}{A_{kk}}$ y se realizan las operaciones de multiplicar la fila k por M_{ik} y sumarla a la fila i. Al elemento $A_{k,k}$ se le denomina el *elemento de pivote* k. La fila resultante se almacena en la fila i. La variable i va desde $i = k + 1$ hasta $i = n$. La variable j en la fila i va desde $j = k + 1$ hasta $j = n + 1$, para todos los elementos A_{ij} de la fila i que se han modificado, hasta inclusive la parte ampliada en la columna $j = n + 1$. Los elementos eliminados, A_{ik}, son en teoría nulos, por lo que no es necesario mostrar sus resultados.

4.- Al obtener la matriz triangular superior se inicia un proceso de sustitución regresiva para así obtener la solución del sistema. Si al final el valor de $A_{n,n}$ queda con un valor muy pequeño en valor absoluto, significa que la matriz es singular (con este procedimiento).

1.1.2. Pivotes

Los pivotes pueden ser parciales o totales, según se intercambien sólo filas o filas y columnas, justo antes de la eliminación de la columna correspondiente.

• Pivote Parcial

Se intercambian filas, entre la fila k y las filas $i = k+1$ hasta $i = n$, de manera que al final quede como elemento A_{kk}, en la diagonal principal, el mayor valor "absoluto" de los elementos. Cuando este elemento A_{kk}, que le denominamos elemento de "pivote", esté una vez localizado en su lugar, se procede a hacer la eliminación de la columna k.

El intercambio entre fila debe ocurrir siempre por debajo del elemento de pivote para no alterar los elementos ya eliminados.

• Pivote Total

Se intercambian filas/columnas, entre la fila/columna k y las fila/columna $i = k + 1$ hasta $i = n$, de manera que al final quede como elemento A_{kk}, en la diagonal principal, el mayor valor "absoluto" de los elementos. Cuando este elemento A_{kk}, que le denominamos elemento de "pivote", esté una vez localizado en su lugar, se procede a hacer la eliminación de la columna k.

Al intercambiar columnas se altera el orden de las incógnitas, por lo que es necesario guardar este orden utilizando un puntero en la variable $JJ(j)$ que originalmente tiene el valor j y se puede ir modificando, según se intercambien las columnas.

El intercambio entre filas/columnas debe ocurrir siempre por debajo/derecha del elemento de pivote para no alterar los elementos ya eliminados.

1.1.3. Eliminación de Gauss

Durante el proceso de eliminación es posible que uno de los elementos de la diagonal principal sea nulo, lo cual originaría el problema de una división por cero. De forma de evitar este problema se incluye en el proceso de eliminación el intercambio de filas, comunmente llamado *pivote parcial*, el cual también permite controlar la propagación del error de redondeo que ocurre en sistemas de ecuaciones cuyo determinante es cercanamente singular. El algoritmo propuesto es conocido en la literatura de análisis numérico como el *método de eliminación gaussiana*.

El proceso de intercambio de filas se hace buscando que el valor absoluto de los multiplicadores M_{ik} sea siempre menor o igual a la unidad, $|M_{ik}| \leq 1$. De forma de cumplir con esta restricción, el elemento A_{kk} sobre la columna que se está eliminando deberá ser el mayor en magnitud de todos los elementos que están por debajo de la fila k, $|A_{kk}| \geq |A_{ik}|$ con $i \geq k$.

EJEMPLO:

De forma de observar la propagación del error de redondeo, se repetirá la solución del ejemplo, de la sub-subsección 1.1.1, pero utilizando como algoritmo de solución el método de eliminación de Gauss.

Partiendo del sistema de ecuaciones, ya en su forma de matriz ampliada

$$\left[\begin{array}{ccc|c} 2.51 & 1.48 & 4.53 & 0.05 \\ 1.48 & 0.93 & -1.30 & 1.03 \\ 2.68 & 3.04 & -1.48 & -0.53 \end{array}\right].$$

en el cual se puede observar que el elemento de mayor magnitud en la columna 1 es el elemento A_{31}, intercambiando la fila 1 con la fila 3 la matriz queda de la forma:

$$\left[\begin{array}{ccc|c} 2.68 & 3.04 & -1.48 & -0.53 \\ 1.48 & 0.93 & -1.30 & 1.03 \\ 2.51 & 1.48 & 4.53 & 0.05 \end{array}\right].$$

Se multiplica la primera fila por $M_{21} = -\frac{1.48}{2.68}$ y se le suma a la segunda fila.

Se multiplica la primera fila por $M_{31} = -\frac{2.51}{2.68}$ y se le suma a la tercera fila.

Se puede observar que ambos multiplicadores tienen magnitud menor o igual a la unidad. Después de hacer las operaciones indicadas, la matriz ampliada es

$$\left[\begin{array}{ccc|c} 2.68 & 3.04 & -1.48 & -0.53 \\ 0.00 & -0.74 & -0.484 & 1.32 \\ 0.00 & -1.368 & 5.91 & 0.546 \end{array}\right].$$

Nuevamente se puede observar que el elemento A_{32} es de mayor magnitud que el elemento A_{22}, por lo cual se hace necesario un nuevo cambio de fila entre la fila 2 y la fila 3.

Si se multiplica la segunda fila por $M_{32} = -\frac{0.74}{1.36}$ y se le suma a la tercera fila, la nueva matriz ampliada es

$$\left[\begin{array}{ccc|c} 2.68 & 3.04 & -1.48 & -0.53 \\ 0.00 & -1.36 & 5.91 & 0.546 \\ 0.00 & 0.00 & -3.69 & 1.02 \end{array}\right].$$

Ya obtenida la matriz ampliada triangular superior, se realiza el proceso de sustitución regresiva para obtener la solución del sistema propuesto

$$x_3 = -0.276$$

$$x_2 = -1.59$$

$$x_1 = 1.45$$

Como puede observarse el resultado no es el mismo que el obtenido con el método de eliminación simple debido a la propagación del error de redondeo. La solución exacta del sistema es

$$x_3 = -0.2749$$

$$x_2 = -1.5892$$

$$x_1 = 1.4531$$

El algoritmo del método de eliminación de Gauss quedaría así

ALGORITMO:

1.- Dado un sistema de n ecuaciones lineales con n incógnitas, con matriz cuadrada, se construye su matriz ampliada.

2.- Se inicia el proceso de eliminación desde la columna $k = 1$ hasta la columna $k = n - 1$.

3.- Se verifica que el elemento A_{kk} es el de mayor magnitud en valor absoluto de todos los elementos por debajo de la fila k, $|A_{kk}| \geq |A_{ik}|$ con $i \geq k$. En caso negativo, hacer el cambio de fila que garantice tal condición.

4.- Se evalúan los multiplicadores $M_{ik} = -\frac{a_{ik}}{a_{kk}}$ y se realizan las operaciones de multiplicar la fila k por M_{ik} y sumarla a la fila i, la fila resultante se almacena en la fila i. La variable i va desde $i = k + 1$ hasta $i = n$. La variable j en la fila i va desde $j = k + 1$ hasta $j = n + 1$, para todos los elementos A_{ij} de la fila i que se han modificado, hasta inclusive la parte ampliada en la columna $j = n + 1$. Los elementos eliminados, A_{ik}, son en teoría nulos, por lo que no es necesario mostrar sus resultados.

5.- Al obtener la matriz triangular superior se inicia un proceso de sustitución regresiva para así obtener la solución del sistema.

1.1.4. Eliminación de Gauss-Jordan

En este método se realiza la eliminación de los elementos por columnas ($i = 1$ hasta $i = n$, $i \neq k$) con la excepción del elemento de pivote k. Finalmente se ejecuta un despeje simple $x_k = b_k/A_{kk}$ en la matriz diagonal obtenida mediante, eliminación simple, eliminación de Gauss o con pivote total.

1.1.5 Normalización

Este procediemiento busca obtener una matriz equivalente (con la misma solución) cuyos elementos sean en valor absoluto menor o igual a la unidad.

• **Por Filas** En este procedimiento se busca el elemento de mayor valor absoluto por filas y luego se divide la correspondiente fila entre dicho elemento.

• **Global** En este procedimiento se busca el elemento de mayor valor absoluto por filas/columnas en toda la matiz y luego se divide toda la matriz entre dicho elemento.

Puede aplicarse de forma inicial o de forma intermedia después de el proceso de eliminación de cada columna. Puede incluir o no la parte ampliada de la matriz.

1.1.6 Descomposición L-U

La decomposición L-U busca obtener la descomposición de la matriz de un sistema $[\mathbf{A}] = [\mathbf{L}][\mathbf{U}]$, donde $[\mathbf{L}]$ es una matriz triangular inferior y $[\mathbf{U}]$ es una matriz triangular superior, todas cuadradas. Una vez teniendo en cuenta los elementos nulos en cada matriz, el producto se desarrolla como

$$A_{ij} = \sum_{k=1}^{i-1} L_{ik}U_{kj} + L_{ii}U_{ij} \qquad (i \leq j) \tag{3.a}$$

$$A_{ij} = \sum_{k=1}^{j-1} L_{ik}U_{kj} + L_{ij}U_{jj} \qquad (i \geq j) \tag{3.b}$$

De donde se obtienen las siguientes ecuaciones

$$U_{ij} = \frac{A_{ij} - \sum_{k=1}^{i-1} L_{ik}U_{kj}}{L_{ii}} \qquad (i \leq j) \tag{4.a}$$

$$L_{ij} = \frac{A_{ij} - \sum_{k=1}^{j-1} L_{ik}U_{kj}}{U_{jj}} \qquad (i \geq j) \tag{4.b}$$

Para resolver el problema de la igualdad, se estipula o impone el valor de L_{ii}, U_{jj} o ambos.

Luego el problema de resolver un sistema de ecuaciones, una vez obtenidas $[\mathbf{L}]$ y $[\mathbf{U}]$, se replantea como dos sistemas

$$[\mathbf{A}]\mathbf{x} = [\mathbf{L}][\mathbf{U}]\mathbf{x} = \mathbf{b} \qquad [\mathbf{U}]\mathbf{x} = \mathbf{z} \qquad [\mathbf{L}]\mathbf{z} = \mathbf{b} \tag{5}$$

con la ayuda de una variable auxiliar $\mathbf{z}$. Por la forma de las matrices esto se reduce a hacer una sustitución progresiva

$$z_i = \frac{b_i - \sum_{k=1}^{i-1} L_{ik}z_k}{L_{ii}} \tag{6}$$

y una sustitución regresiva

$$x_i = \frac{z_i - \sum_{k=i+1}^{n} U_{ik}x_k}{U_{ii}} \tag{7}$$

que permite obtener la solución del sistema original.

Lo interesante de este método es que sólamente hace falta hacer la descomposición de la matriz $[\mathbf{A}]$ una vez y, en el caso de resolver varios sistemas con distintos vectores independientes $\mathbf{b}$, sólo hace falta hacer las dos sustituciones progresiva y regresiva para cada $\mathbf{b}$.

El procedimiento propuesto para elmacenar la información del método conocido como el *método de descomposición LU*, es tal que

$$[\mathbf{A}]\mathbf{x} = [\mathbf{L}][\mathbf{U}]\mathbf{x} = [\mathbf{L}]\mathbf{z} = \mathbf{b} \qquad [\mathbf{A}|\mathbf{b}] \rightarrow [\mathbf{L}\backslash\mathbf{U}|\mathbf{z}] \rightarrow [\mathbf{L}\backslash\mathbf{U}|\mathbf{x}] \tag{8}$$

sin haber interferencia de memoria. La matrices [**A**], [**L**] y [**U**] y los vectores **b**, **z** y **x** pueden ocupar el mismo espacio de memoria de una matriz ampliada sin problema como se verá más adelante.

• Método de Doolittle

En el método de Doolittle se escoge $L_{ii} = 1$, por lo que las ecuaciones anteriores se reducen a encontrar una fila $p = 1, 2, 3, \ldots, n$ de [**U**] y una columna $p = 1, 2, 3, \ldots, n-1$ de [**L**] de forma alternante con las ecuaciones

$$U_{pj} = A_{pj} - \sum_{k=1}^{p-1} L_{pk}U_{kj} \qquad j = p, p+1, \ldots, n+1 \tag{9.a}$$

$$L_{ip} = \frac{A_{ip} - \sum_{k=1}^{p-1} L_{ik}U_{kp}}{U_{pp}} \qquad i = p+1, p+2, \ldots, n \tag{9.b}$$

El caso de [**U**] con $j = n+1$ coincide con la substitución progresiva cuando se trabaja con la matriz ampliada [**A**|**b**]. Para $p = 1$ las sumatorias de (9) son nulas.

La sustitución regresiva viene dada por

$$x_p = \frac{U_{p,n+1} - \sum_{k=p+1}^{n} U_{pk}x_k}{U_{pp}} \qquad p = n, n-1, \ldots, 1 \tag{10}$$

Cuando $p = n$ la sumatoria es nula, al igual que en todos los casos de sumatorias donde el límite inferior supera al límite superior.

El método está basado en descomponer la matriz de coeficientes [**A**] en el producto de dos matrices [**L**] y [**U**], las cuales son matrices triangulares inferior (Lower) y superior (Upper) respectivamente. La matriz [**L**] tiene la particularidad de que todos los elementos de la diagonal principal son iguales a la unidad por lo que no es necesario guardar su valor en memoria.

En forma matricial, la descomposición quedaría de la siguiente forma

$$\begin{bmatrix} 1 & 0 & \ldots & 0 \\ L_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \ldots & 1 \end{bmatrix} \quad \begin{bmatrix} U_{11} & U_{12} & \ldots & U_{1n} \\ 0 & U_{22} & \ldots & U_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & U_{nn} \end{bmatrix} = \quad \begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1n} \\ A_{21} & A_{22} & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \ldots & A_{nn} \end{bmatrix} \tag{11}$$

Para determinar los coeficientes L_{ij} y U_{ij} se utilizan las expresiones (9).

EJEMPLO:

Construir las matrices [**L**] y [**U**] para el sistema de ecuaciones del ejemplo de la subsección 1.1.1.

El primer paso es construir la primera fila de [**U**] y la primera columna de [**L**]:

$$U_{1j} = A_{1j} \qquad j = 1, 2, 3$$

$$L_{i1} = \frac{A_{1j}}{U_{11}} \qquad i = 2, 3$$

El segundo paso es construir la segunda fila de [**U**] y la segunda columna de [**L**], las cuales quedan:

$$U_{2j} = A_{2j} - L_{21}U_{1j} \qquad j = 2, 3$$
$$L_{32} = \frac{A_{32} - L_{31}U_{12}}{U_{22}}$$

El último paso (en este ejemplo) es hallar la última fila de [**U**]

$$U_{33} = A_{33} - L_{31}U_{13} - L_{32}U_{23}$$

Quedando las matrices [**L**] y [**U**] de la siguiente forma:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0.589 & 1 & 0 \\ 1.06 & 25.0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 2.51 & 1.48 & 4.53 \\ 0 & 0.059 & -3.96 \\ 0 & 0 & 92.7 \end{bmatrix}$$

En las cuales se observa que en la matriz [**L**] se almacena la información correspondiente a los multiplicadores, y en la matriz [**U**] se tiene la matriz triangular superior final del proceso de eliminación, donde podemos incluir adicionalmente la parte ampliada del vector independiente modificado con las operaciones.

Conocidas las matrices [**L**] y [**U**] debemos resolver el sistema de ecuaciones planteado. Recordando que el sistema de ecuaciones viene dado por la expresión $(5.a)$ y expresando el producto $[\mathbf{U}]x$ como el vector $\mathbf{z}$, el sistema se desdobla en los sistemas $(5.c)$ y $(5.b)$, que pueden ser resueltos por simples sustituciones progresiva y regresiva, respectivamente.

• Método de Crout-Cholesky

En el método de Crout se escoge $U_{jj} = 1$, por lo que el procedimiento se reduce a encontrar una columna $p = 1, 2, 3, \ldots, n$ de [**L**] y una fila $p = 1, 2, 3, \ldots, n-1$ de [**U**] de forma alternante con la ecuaciones

$$L_{ip} = A_{ip} - \sum_{k=1}^{p-1} L_{ik}U_{kp} \qquad i = p, p+1, \ldots, n \tag{12.a}$$

$$U_{pj} = \frac{A_{pj} - \sum_{k=1}^{p-1} L_{pk}U_{kj}}{L_{pp}} \qquad j = p+1, p+2, \ldots, n+1 \tag{12.b}$$

El indice $j = n + 1$ en $(12.b)$ es la parte ampliada de la matriz [**A**].

Luego la sustitución regresiva se calcula con

$$x_p = U_{p,n+1} - \sum_{k=p+1}^{n} U_{pk}x_k \qquad p = n, n-1, \ldots, 1 \tag{13}$$

El método de Choleski $U_{pp} = L_{pp}$, por lo que los elementos de la diagonal principal, tanto de [**L**] como de [**U**] se calculan como

$$L_{pp} = U_{pp} = \sqrt{A_{pp} - \sum_{k=1}^{p-1} L_{pk}U_{kp}} \qquad p = 1, 2, \ldots, n \tag{14}$$

y se comienza indistintamente con una fila p de [**U**] y una columna p de [**L**] con $p = 1, 2, 3, \ldots, n-1$ usando las ecuaciones (4).

1.1.7 Sistemas Tridiagonales

Son los sistemas de ecuaciones lineales con tres diagonales principales, con elementos a_i en la diagonal inferior, b_i en la diagonal central y principal y c_i en la diagonal superior. Los elementos a_1 y c_n no existen en este sistema. Los elementos del vector independiente son d_i.

Para resolverlo hemos escogido el método de eliminación simple o descomposición L-U (Doolittle), que son equivalentes

• Eliminación

como normalmente son sistemas diagonalmente dominante no hace falta pivotear. Los elementos a_i se eliminan formando los elementos β_i también en la diagonal principal

$$\beta_i = b_i - \frac{a_i}{\beta_{i-1}}\, c_{i-1} \tag{15}$$

para $i = 1, 2, \ldots, n$. Los elementos del vector independiente d_i se tranforman en δ_i

$$\delta_i = d_i - \frac{a_i}{\beta_{i-1}}\, \delta_{i-1} \tag{16}$$

también para $i = 1, 2, \ldots, n$. Los elementos c_i quedan inalterados.

Luego la sustitución regresiva es

$$x_i = \frac{\delta_i - c_i\, x_{i+1}}{\beta_i} \tag{17}$$

para $i = n, n-1, \ldots, 1$. Este algoritmo usa tres ecuaciones β_i, δ_i y x_i.

• Algoritmo de Thomas

Se hace el siguiente cambio de variables $\gamma_i = z_i/U_{ii} = \delta_i/\beta_i$, $\lambda_i = c_i/\beta_i$ siendo $U_{ii} = b_i - L_{i,i-1}U_{i-1,i} \equiv \beta_i = b_i - a_i\, c_{i-1}/\beta_{i-1} = b_i - a_i\, \lambda_{i-1}$, con $U_{i,i+1} = c_i$ y $L_{i,i-1} = a_i/\beta_{i-1}$, por lo que la ecuación de z_i queda

$$z_i = d_i - L_{i,i-1}\, z_{i-1} \qquad\qquad \gamma_i = \frac{d_i - \gamma_{i-1}\, a_i}{\beta_i} \qquad\qquad (\lambda_i = c_i/\beta_i) \tag{18}$$

Finalmente la sustitución regresiva da la solución

$$x_i = \frac{z_i - U_{i,i+1}\, x_{i+1}}{U_{ii}} = \gamma_i - \frac{c_i}{\beta_i}\, x_{i+1} = \gamma_i - \lambda_i\, x_{i+1} \tag{19}$$

donde se ha empleado parcialmente la notación de la descomposición L-U (Doolittle). A este algoritmo (por Llewellyn Thomas,(1949)) también se le denomina TDMA (Three Diagonal Matrix Algorithm). Este algoritmo utiliza también tres ecuaciones β_i ó λ_i, γ_i y x_i, pero de dice que es mucho más eficiente numéricamente que el algoritmo de eliminación. Los espacios de memoria utilizados pueden ser los mismos que los de las variable originales sin interferencia [Conte & de Boor, 1972].

1.1.8. Determinante

El determinante es de fácil cálculo, pues

$$\det([\mathbf{A}]) = (-1)^p\, \det([\mathbf{U}]) \tag{20}$$

donde [**U**] es la matriz triangular superior que queda después del proceso de eliminación y p es el número de pivotes que cambia el signo al determinante.

1.1.9. Matriz Inversa

Si con el método de Gauss-Jordan se trabaja con la matrix [**A**] ampliada con la matrix identidad [**I**] en la forma [**A**|**I**], y se logra convertir con el procedimiento planteado a la matriz [**A**] en la matriz [**I**], entonces la matriz [**I**] en la parte ampliada se convierte en la matriz [**B**], tal que [**A**] [**B**] = [**I**], por lo que [**B**] es realmente $[\mathbf{A}]^{-1}$.

1.1.10. Autovalores y Autovectores

Las definiciones de los autovalores λ y autovectores $\mathbf{e}$ se resume en las siguientes expresiones (ver por ejemplo [Hoffman & Kunze,1971])

$$\mathbf{A}.\mathbf{e} = \lambda\,\mathbf{e} \qquad P(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = 0 \qquad \rho(\mathbf{A}) = \max_{1\le i\le k} |\lambda_i| \qquad \begin{array}{c} \lambda = \alpha + i\beta \\ |\lambda| = \sqrt{\alpha^2+\beta^2} \end{array} \tag{21}$$

$P(\lambda)$ es el polinomio característico cuya raíces son los autovalores. ρ es el radio espectral. La multiplicidad d_k de los autovalores λ_k originan varios sub-espacios $\mathbb{W}_k$, cuya unión es el espacio completo $\mathbb{V} = \mathbb{R}^n$

$$P(\lambda) = \prod_{i=1}^{k} (\lambda - \lambda_i)^{d_i} \qquad \sum_{i=1}^{k} \dim \mathbb{W}_i = \dim \mathbb{V} = n \qquad \dim \mathbb{W}_i = d_i \tag{22}$$

Defínase la matriz $\mathbf{S}$ con columnas siendo los autovectores $\mathbf{e}_{i,j}$ ($j \le d_i$, puede haber más de un autovector para cada autovalor) que generan el subespacio $\mathbb{W}_i$ para cada autovalor λ_i ($1 \le i \le k \le n$)

$$(\mathbf{A} - \lambda_i\,\mathbf{I})\,.\,\mathbf{x} = \mathbf{0} \qquad \mathbf{S} = [\,\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_k\,] = [\,\mathbb{B}_1, \mathbb{B}_2, \ldots, \mathbb{B}_k\,] \tag{23}$$

El sistema lineal (23.a) sirve para obtener las componentes de cada autovector $\mathbf{e}_{i,j}$ en la construcción de $\mathbf{S}$. Cada base $\mathbb{B}_i$ contiene el número d_i de autovectores que generan el subespacio $\mathbb{W}_i$. Entonces, la matriz $\mathbf{S}$ diagonaliza $\mathbf{A}$ de la siguiente manera

$$\mathbf{A}\,.\,\mathbf{S} = \mathbf{\Lambda}.\mathbf{S} = \mathbf{S}.\mathbf{\Lambda} \qquad \mathbf{S}^{-1}.\,\mathbf{A}\,.\,\mathbf{S} = \mathbf{\Lambda} = \mathrm{diag}\{\,\lambda_1, \lambda_2, \ldots, \lambda_k\,\} \tag{24}$$

El valor λ_i puede estar repetido en la matriz $\mathbf{\Lambda}$ dependiendo de d_i. Las matrices $\mathbf{A}$ y $\mathbf{\Lambda}$ (over $\mathbb{F} = \mathbb{R}$) son *semejantes* y las matrices $\mathbf{S}$ y $\mathbf{\Lambda}$ permutan. Cuando la matriz $\mathbf{A}$ es simétrica (o hermítica en el caso complejo) los autovalores son todos reales, la transformación $\mathbf{S}$ es también ortogonal (cuando los autovectores se normalizan), y $\mathbf{A}$ y $\mathbf{\Lambda}$ (over $\mathbb{F} = \mathbb{R}$) son también *congruentes*.

Dada la definición del polinomio característico [Pennington,1970]

$$P(\lambda) = |\,[\mathbf{A}] - \lambda\,[\mathbf{I}]\,| = 0 \tag{21}$$

(el símbolo $|[\mathbf{A}]|$ significa $\det([\mathbf{A}])$), se ha implementado las siguientes fórmulas recurrentes

$$\begin{array}{ll} [\mathbf{A}_1] = [\mathbf{A}] & P_1 = \mathrm{tr}[\mathbf{A}_1] \\ [\mathbf{A}_2] = [\mathbf{A}]\,(\,[\mathbf{A}_1] - P_1\,[\mathbf{I}]\,) & P_2 = \dfrac{1}{2}\,\mathrm{tr}[\mathbf{A}_2] \\ [\mathbf{A}_3] = [\mathbf{A}]\,(\,[\mathbf{A}_1] - P_2\,[\mathbf{I}]\,) & P_3 = \dfrac{1}{3}\,\mathrm{tr}[\mathbf{A}_3] \\ \quad\vdots & \quad\vdots \\ [\mathbf{A}_n] = [\mathbf{A}]\,(\,[\mathbf{A}_{n-1}] - P_{n-1}\,[\mathbf{I}]\,) & P_n = \dfrac{1}{n}\mathrm{tr}[\mathbf{A}_n] \end{array} \tag{22}$$

que permite finalmente obtener

$$|\,[\mathbf{A}_n - P_n\,[\mathbf{I}]\,| = 0 \tag{23}$$

y con las cuales se puede calcular la inversa de la matriz $[\mathbf{A}]$

$$[\mathbf{A}]^{-1} = \frac{1}{P_n}\,(\,[\mathbf{A}_{n-1}] - P_{n-1}\,[\mathbf{I}]\,) \tag{24}$$

y el polynomio caractrístico

$$\lambda^n - P_1\,\lambda^{n-1} - P_2\,\lambda^{n-2} - \cdots - P_n = 0 \tag{25}$$

que al resolver sus raíces nos permite obtener los autovalores.

1.1.11. Normas

Las normas permiten obtener una medida positiva de vectores y matrices y se define como la función $\|\cdot\|$, que se puede usar para estimar longitudes, distancias y órdenes de magnitud.

• Norma de Vectores

Las normas de los vectores en $\mathbb{R}^n$ tienen las siguientes propiedades:

i) $\|\mathbf{x}\| \geq 0$ para todo $\mathbf{x} \in \mathbb{R}^n$ (no negatividad).
ii) $\|\mathbf{x}\| = 0$ si y sólo si $\mathbf{x} = \mathbf{0}$.
iii) $\|\alpha\,\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$ para toda $\alpha \in \mathbb{R}$ y $\mathbf{x} \in \mathbb{R}^n$ (homogeneidad).
iv) $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ para todo $\mathbf{x}$ y $\mathbf{y} \in \mathbb{R}^n$ (desigualdad triangular).
V $|\mathbf{x}.\mathbf{y}| \leq \|\mathbf{x}\|\,\|\mathbf{y}\|$ para todo $\mathbf{x}$ y $\mathbf{y} \in \mathbb{R}^n$ (Desiguladad de Cauchy-Schwarz).

Existe muchos tipos de normas para vectores, pero las más importantes son:

◦ Norma 1

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| \tag{26}$$

◦ Norma ∞

$$\|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} |x_i| \tag{27}$$

◦ Norma 2 (euclidiana)

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}.\mathbf{x}} \tag{28}$$

◦ Norma p

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \tag{29}$$

La norma en (29) con $p = 2$ es equivalente a la norma 2 en (28) (triangular←Minkowski, Cauchy←Hölder).

• Norma de Matrices

Una norma matricial en el conjunto de todas las matrices de $\mathbb{R}^n \times \mathbb{R}^n$ es una función de valores reales positivos, definida en este conjunto que satisface las siguientes propiedades, para todas las matrices $\mathbf{A}$ y $\mathbf{B}$ de $\mathbb{R}^n \times \mathbb{R}^n$ y todo escalar $\alpha \in \mathbb{R}$:

i) $\|\mathbf{A}\| \geq 0$ (no negatividad).
ii) $\|\mathbf{A}\| = 0$ si y sólo si $\mathbf{A} = \mathbf{O}$.
iii) $\|\alpha\,\mathbf{A}\| = |\alpha|\,\|\mathbf{A}\|$ (homogeneidad).
iv) $\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ (desigualdad triangular).
v) $\|\mathbf{A}.\mathbf{B}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|$ (Desiguladad de Cauchy-Schwarz).

Existe muchos tipos de normas para Matrices, pero las más importantes son:

◦ Norma 1

$$\|\mathbf{A}\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |A_{ij}| \tag{30}$$

◦ Norma ∞

$$\|\mathbf{A}\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |A_{ij}| \tag{31}$$

- Norma 2 (euclidiana)

$$\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^{\mathrm{h}}.\mathbf{A})} \tag{32}$$

- Norma p

$$\|\mathbf{A}\|_p = \left(\sum_{j=1}^{n}\sum_{i=1}^{n}|A_{ij}|^p\right)^{1/p} \tag{33}$$

El *radio espectral* se define como (cociente de Rayleigh $r_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{x}^*.\mathbf{A}.\mathbf{x}}{\mathbf{x}^*.\mathbf{x}}$, $\mathbf{A}$ hermítica, $\mathbf{x} \neq \mathbf{0}$)

$$\rho(\mathbf{A}) = \max_{\|\mathbf{x}\|\neq 0} r_{\mathbf{A}}(\mathbf{x}) = \max_{1\leq i\leq k}|\lambda_i| \qquad \begin{array}{c}\lambda = \alpha + i\beta \\ |\lambda| = \sqrt{\alpha^2+\beta^2}\end{array} \tag{34}$$

el mayor de todos los módulos de los autovalores y satisface que [Schatzman,2002,Chp.10][Hämmerlin & Hoffmann,1991,pp.68-72] (fórmula de Gelfand (1041), $\forall k \in \mathbb{N}$)

$$\rho(\mathbf{A}) = \lim_{r\to\infty}\|\mathbf{A}^r\|^{1/r} \leq \|\mathbf{A}^k\|^{1/k} \tag{35}$$

es menor que cualquier norma natural $\|\cdot\|$. Cuando $p = 2$ en la fórmula (33) la norma se denomina de Frobenius o de Hilbert-Schmidt (no subordinada)

$$\|\mathbf{A}\|_F = \left(\sum_{j=1}^{n}\sum_{i=1}^{n}|A_{ij}|^2\right)^{1/2} = \sqrt{\mathrm{tr}(\mathbf{A}^{\mathrm{h}}.\mathbf{A})} \qquad \|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{n}\,\|\mathbf{A}\|_2 \tag{36}$$

Las normas de matrices se dice que son normas subordinadas o inducida por las normas de los vectores, o norma natural, en el sentido que

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1}\|\mathbf{A}.\mathbf{x}\| \qquad \|\mathbf{A}\| = \max_{\|\mathbf{x}\|\neq 0}\frac{\|\mathbf{A}.\mathbf{x}\|}{\|\mathbf{x}\|} \qquad \|\mathbf{A}\| \leq \sqrt{n}\max_{1\leq i\leq n}\|\mathbf{A}.\mathbf{e}_i\|_2 \tag{37}$$

que son equivalentes ($\mathbf{A}.\mathbf{e}_i$ es la fila i de la matriz $[\mathbf{A}]$).

1.1.12 Condicionamiento

El *número de condición* $K(\mathbf{A})$ de la matriz no singular $\mathbf{A}$, relativo a la norma $\|\cdot\|$, se define como

$$K(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\| \tag{38}$$

Se sabe que

$$\mathbf{A}.\mathbf{x} = \mathbf{b} \qquad \mathbf{A}.\mathbf{x}^k = \mathbf{b}^k \qquad \mathbf{A}.\mathbf{r} = \mathbf{b} \qquad \mathbf{A}.\mathbf{e}^k = \mathbf{d}^k \tag{39}$$

donde $\mathbf{r}$ es la solución exacta del sistema. De la última ecuación (39.d) y de la definición (38) se obtiene

$$\|\mathbf{e}^k\| \leq K(\mathbf{A})\,\frac{\|\mathbf{d}^k\|}{\|\mathbf{A}\|} \tag{40}$$

o introduciendo (39.c) $\|\mathbf{A}\|\,\|\mathbf{r}\| \geq \|\mathbf{b}\|$

$$\frac{\|\mathbf{e}^k\|}{\|\mathbf{r}\|} \leq K(\mathbf{A})\,\frac{\|\mathbf{d}^k\|}{\|\mathbf{b}\|} \tag{41}$$

Ya que para cualquier matriz no-singular $\mathbf{A}$

$$1 = \|\mathbf{I}\| = \|\mathbf{A}.\mathbf{A}^{-1}\| \leq \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\| = K(\mathbf{A}) \tag{42}$$

se espera que la matriz $\mathbf{A}$ tenga un buen comportamiento (llamada formalmente una matriz bien condicionada) si $K(\mathbf{A})$ está cerca de uno y un comportamiento defectuoso (llamada mal condicionada) cuando $K(\mathbf{A})$ sea significativa mayor que uno. El comportamiento en esta situación se refiere a la relativa seguridad de que un vector residual $\mathbf{d}^k$ pequeño implique correspondientemente una solución aproximada precisa. La expresión (41) da una interrelación entre el vector error relativo $\|\mathbf{e}^k\|/\|\mathbf{r}\|$ y desviación relativa $\|\mathbf{d}^k\|/\|\mathbf{b}\|$ con el número de condición $K(\mathbf{A})$.

1.2. METODOS ITERATIVOS

1.2.1. Método de Jacobi

Dentro de los métodos de solución de ecuaciones diferenciales parciales figuran las diferencias finitas, los elementos finitos, los elementos de frontera, entre otros, pero todos tienen el denominador común de generar sistemas de ecuaciones algebraicas de gran tamaño. Una de las características de estos sistemas de ecuaciones es la presencia de elementos nulos dentro de la matriz en una forma bien determinada, representando el mayor porcentaje de elementos dentro de la matriz, normalmente del 80% al 95%.

Debido a la existencia de una gran cantidad de elementos nulos no es conveniente trabajar con métodos directos, en los cuales se debe almacenar la matriz de coeficientes, sino utilizar métodos iterativos entre los cuales figura el *método de Jacobi*. Dicho algoritmo consiste en suponer un vector solución inicial $\mathbf{x}_o$ y determinar la solución mediante un procedimiento iterativo o de punto fijo, el cual tiene la siguiente forma

$$x_i^{k+1} = \frac{1}{A_{ii}}\left(b_i - \sum_{j=1}^{i-1} A_{ij}\, x_j^k - \sum_{j=i+1}^{n} A_{ij}\, x_j^k\right) \qquad i = 1, 2, 3, \ldots, n \tag{1}$$

Para obtener convergencia, es necesario que la matriz $[\mathbf{A}]$ sea una matriz diagonalmente dominante, es decir, la magnitud del elemento de la diagonal debe ser mayor que la suma en valor absoluto de todos los elementos restantes en la fila

$$|A_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |A_{ij}| = |A_{i1}| + |A_{i2}| + \ldots + |A_{i,i-1}| + |A_{i,i+1}| + \ldots + |A_{in}| \tag{2}$$

Teorema. Si la matriz $[\mathbf{A}]$ es diagonalmente dominante por filas en forma estricta (expresión (2)), entonces los métodos de Jacobi y de Gauss-Seidel convergen.

1.2.2. Método de Gauss-Seidel

El *método de Gauss-Seidel* es una variante del método de Jacobi, donde las variables se van actualizando en la medida que se van calculando en la forma

$$x_i^{k+1} = \frac{1}{A_{ii}}\left(b_i - \sum_{j=1}^{i-1} A_{ij}\, x_j^{k+1} - \sum_{j=i+1}^{n} A_{ij}\, x_j^k\right) \qquad i = 1, 2, 3, \ldots, n \tag{3}$$

Teorema (Stein-Rosemberg). Si la matriz $[\mathbf{A}]$ es diagonalmente dominante por filas en forma estricta (expresión (2)), y adicionalmente los signos de los elementos de la diagonal principal son de signos opuestos a los elementos fuera de esta, el método de Gauss-Seidel converge más rápido.

1.2.3 Relajación Sucesiva

El método de relajación sucesivas (SOR-Succesive Over Relaxation-Southwell) es una variante del método de Gauss-Seidel, donde se sobre-relaja el método. Si definimos el vector independiente aproximado

$$\mathbf{b}^k = [\mathbf{A}]\mathbf{x}^k \qquad b_i^k = \sum_{j=1}^{i-1} A_{ij}\, x_j^{k+1} + A_{ii}\, x_i^k + \sum_{j=i+1}^{n} A_{ij}\, x_j^k \qquad i = 1, 2, 3, \ldots, n \tag{4}$$

y la desviación global o residuo (a veces definido como -residuo)

$$\mathbf{d}^k = \mathbf{b}^k - \mathbf{b} \qquad d_i^k = \sum_{j=1}^{i-1} A_{ij}\, x_j^{k+1} + A_{ii}\, x_i^k + \sum_{j=i+1}^{n} A_{ij}\, x_j^k - b_i \qquad i = 1, 2, 3, \ldots, n \tag{5}$$

entonce el método de relajación sucesiva se expresa algorítmicamente como

$$x_i^{k+1} = x_i^k - \omega\, \frac{d_i^k}{A_{ii}} \tag{6}$$

donde el factor de relajación ω es:

$\omega < 1$ Subrelajado

$\omega = 1$ Gauss-Seidel

$\omega > 1$ Sobrerelajado

La desviación local $\boldsymbol{\delta}^k$ y los errores locales $\boldsymbol{\epsilon}^k$ y globales $\mathbf{e}^k$ son definidos como

$$\boldsymbol{\delta}^k = \mathbf{b}^k - \mathbf{b}^{k-1} \qquad \boldsymbol{\epsilon}^k = \mathbf{x}^k - \mathbf{x}^{k-1} \qquad \mathbf{e}^k = \mathbf{x}^k - \mathbf{r} \tag{7}$$

extendiendo los conceptos antes definidos.

Teorema (Ostrowski-Reich). Si $\mathbf{A}$ es una matriz positiva definida y $0 < \omega < 2$, entonces el método SOR converge para cualquier elección del iterado inicial $\mathbf{x}^o$ o aproximación inicial del vector de solución.

1.2.4. Estabilidad

Sea una cierta perturbación en $\mathbf{b}$ igual a $\delta\mathbf{b}$

$$[\mathbf{A}]\mathbf{x} = \mathbf{b} \qquad [\mathbf{A}](\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$$

$$[\mathbf{A}]\mathbf{x} + [\mathbf{A}]\delta\mathbf{x} = \mathbf{b} + \delta\mathbf{b} \qquad [\mathbf{A}]\delta\mathbf{x} = \delta\mathbf{b}$$

$$\delta\mathbf{x} = [\mathbf{A}]^{-1}\delta\mathbf{b} \tag{8}$$

Si $[\mathbf{A}]$ es casi singular, cualquier modificación en $\mathbf{b}$ producirá grandes cambios en la solución de $\mathbf{x}$.

Sea una cierta perturbación en $[\mathbf{A}]$

$$[\mathbf{A}]\mathbf{x} = \mathbf{b} \qquad ([\mathbf{A}] + \delta[\mathbf{A}])\,(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}$$

$$[\mathbf{A}]\mathbf{x} + [\mathbf{A}]\delta\mathbf{x} + \delta[\mathbf{A}]\mathbf{x} + \delta[\mathbf{A}]\delta\mathbf{x} = \mathbf{b} \qquad [\mathbf{A}]\delta\mathbf{x} + \delta[\mathbf{A}]\mathbf{x} = \mathbf{0}$$

despreciando los términos de segundo orden $\delta[\mathbf{A}]\delta\mathbf{x}$, queda

$$\delta\mathbf{x} = -[\mathbf{A}]^{-1}\delta[\mathbf{A}]\mathbf{x} \tag{9}$$

Si $[\mathbf{A}]$ es casi singular, $\delta\mathbf{x}$ puede ser grande, por consiguiente, $[\mathbf{A}]$ y $\mathbf{b}$ se deben trabajar con la máxima capacidad, de lo contrario, al no ser valores exactos, no hay manera de obtener una solución aceptable.

1.3. OTROS METODOS

1.3.1. Método de la Potencia

Sea $\mathbf{A}$ la matriz cuyo autovalor se desea hallar. Aplicamos dicha matriz aun vector inicial $\mathbf{e}_0$, el vector resultante se normaliza y se le vuelve a aplicar la matriz $\mathbf{A}$ otra vez, y así sucesivamente. Esto es,

$$\mathbf{A}\mathbf{e}_k = \mathbf{v}_{k+1} \qquad \mathbf{e}_{k+1} = \frac{\mathbf{v}_{k+1}}{\|\mathbf{v}_{k+1}\|} \tag{1}$$

El proceso iterativo denominado *método de la potencia* es de tipo punto fijo [Gerald,1978]. Sea un vector $\mathbf{v}^{(0)}$ cualquiera y $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_n$ los autovectores para los autovalore λ_1, λ_2, ..., λ_n. Entonces

$$\mathbf{v}^{(0)} = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \cdots + c_n\mathbf{e}_n \tag{2}$$

Cualquier vector es una combinación lineal de los autovectores. Si aplicamos a $\mathbf{v}^{(0)}$ la matriz $\mathbf{A}$ un número de m veces, se obtiene

$$\mathbf{v}^{(m)} = \mathbf{A}^m\mathbf{v}^{(0)} = c_1\lambda_1^m\mathbf{e}_1 + c_2\lambda_2^m\mathbf{e}_2 + \cdots + c_n\lambda_n^m\mathbf{e}_m \tag{3}$$

Si un autovalor, sea λ_1, es el más grande que todos en valor absoluto, los valores de λ_i^m, $i \neq 1$, será despreciable en comparación a λ_1^m, cuando m sea muy grande y

$$\mathbf{A}^m\mathbf{v}^{(0)} \longrightarrow c_1\lambda_1^m\mathbf{e}_1 \qquad\qquad \|\mathbf{v}_m\| \longrightarrow |\lambda_1| \tag{4}$$

Este es el principio detrás de el método de la potencia.

Sea $\mathbf{Ae} = \lambda\mathbf{e}$, multiplicando por $\mathbf{A}^{-1}$ se obtiene

$$\mathbf{A}^{-1}\mathbf{A}\,\mathbf{e} = \mathbf{A}^{-1}\lambda\,\mathbf{e} = \lambda\,\mathbf{A}^{-1}\mathbf{e} \qquad\qquad \mathbf{A}^{-1}\mathbf{e} = \frac{1}{\lambda}\,\mathbf{e} \tag{5}$$

Lo que es lo mismo que decir que la matriz inversa $\mathbf{A}$ tiene los autovalores inversos que la original. Esto hace que el método de la potencia descrito anteriormente permite hallar el menor autovalor de la matriz $\mathbf{A}^{-1}$.

Dado $\mathbf{Ae} = \lambda\mathbf{e}$, substrayendo $s\mathbf{Ie} = s\mathbf{e}$ en ambos miembros, se obtiene

$$(\mathbf{A} - s\mathbf{I})\mathbf{e} = (\lambda - s)\mathbf{e} \tag{6}$$

La expresión anterior puede ser aplicada de dos formas diferentes. Si deseamos obtener un autovalor cercano a s, basta con extraer s de la diagonal principal de $\mathbf{A}$, y al invertir la matriz modificada y aplicar el método de la potencia se obtendría el valor inverso $1/(\lambda - s)$ (el cual es grande). Luego volviendo al problema original se puede afinar el resultado buscado. La otra forma es escoger s un valor ya obtenido, con lo que aplicar el método de la potencia a la matriz modificada en la diagonal principal brindaría otro valor diferente del autovalor antes hallado, donde $(\lambda - s)$ sea el más grande en valor absoluto.

1.3.2. Ortogonalización de Gram-Schmidt

Este algoritmo recibe su nombre de los matemáticos Jørgen Pedersen Gram y Erhard Schmidt. Sea un conjunto de vectores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ linealmente independiente en el espacio vectorial $\mathbb{V}$. Entonces [Hoffman & Kunze,1971]

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{v}_1 \\
\mathbf{u}_2 &= \mathbf{v}_2 - \frac{\langle \mathbf{v}_2, \mathbf{u}_1\rangle}{\langle \mathbf{u}_1, \mathbf{u}_1\rangle}\,\mathbf{u}_1 \\
\mathbf{u}_3 &= \mathbf{v}_3 - \frac{\langle \mathbf{v}_3, \mathbf{u}_1\rangle}{\langle \mathbf{u}_1, \mathbf{u}_1\rangle}\,\mathbf{u}_1 - \frac{\langle \mathbf{v}_3, \mathbf{u}_2\rangle}{\langle \mathbf{u}_2, \mathbf{u}_2\rangle}\,\mathbf{u}_2 \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots
\end{aligned} \tag{7}$$

De forma general se puede expresar como

$$\mathbf{u}_k = \mathbf{v}_k - \sum_{j=1}^{k-1} \frac{\langle \mathbf{v}_k, \mathbf{u}_j\rangle}{\langle \mathbf{u}_j, \mathbf{u}_j\rangle}\,\mathbf{u}_j \tag{8}$$

donde $k = 1, 2, \ldots, n$. Cuando en la sumatoria el límite superior es menor que el límite inferior la sumatoria no se realiza o su resultado es nulo.

Siendo $\mathbf{u}_k$ ortogonales entre sí y se pueden normalizar como

$$\mathbf{e}_k = \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|} \qquad\qquad \|\mathbf{u}_k\| = \sqrt{\langle \mathbf{u}_k, \mathbf{u}_k \rangle} \tag{9}$$

donde $\langle \mathbf{a}, \mathbf{b} \rangle$ en el cuerpo $\mathbb{F}$, es el producto interior de $\mathbf{a}$ y $\mathbf{b}$, cualquier par de vectores del espacio vectorial $\mathbb{V}$ correspondiente de dimensión finita n. Entonces el conjunto de vectores $\mathbf{e}_k$ forma una base ortonormal.

Recurriendo al método de ortogonalización de Gram-Schmidt, con las columnas de $\mathbf{A}$ como los vectores a procesar $[\mathbf{A}] = [\mathbf{a}_1|\mathbf{a}_2|\cdots|\mathbf{a}_n]$. Entonces

$$\mathbf{u}_k = \mathbf{a}_k - \sum_{j=1}^{k-1} \langle \mathbf{a}_k, \mathbf{e}_j \rangle \, \mathbf{e}_j \tag{10}$$

Despejando $\mathbf{a}_k$, queda

$$\mathbf{a}_k = \sum_{j=1}^{k-1} \langle \mathbf{e}_j, \mathbf{a}_k \rangle \, \mathbf{e}_j + \mathbf{e}_k \, \|\mathbf{u}_k\| \tag{11}$$

En vista de que $[\mathbf{Q}] = [\mathbf{e}_1|\mathbf{e}_2|\cdots|\mathbf{e}_n]$, entonces

$$[\mathbf{A}] = [\mathbf{Q}][\mathbf{R}] = [\mathbf{e}_1|\mathbf{e}_2|\cdots|\mathbf{e}_n] \begin{bmatrix} \|\mathbf{u}_1\| & \langle \mathbf{e}_1, \mathbf{a}_2 \rangle & \langle \mathbf{e}_1, \mathbf{a}_3 \rangle & \cdots \\ 0 & \|\mathbf{u}_2\| & \langle \mathbf{e}_2, \mathbf{a}_3 \rangle & \cdots \\ 0 & 0 & \|\mathbf{u}_3\| & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{12}$$

siendo $[\mathbf{R}]$ es una matriz triangular superior. Por lo que

$$[\mathbf{R}] = [\mathbf{Q}]^{\mathrm{t}}[\mathbf{A}] = \begin{bmatrix} \langle \mathbf{e}_1, \mathbf{a}_1 \rangle & \langle \mathbf{e}_1, \mathbf{a}_2 \rangle & \langle \mathbf{e}_1, \mathbf{a}_3 \rangle & \cdots \\ 0 & \langle \mathbf{e}_2, \mathbf{a}_2 \rangle & \langle \mathbf{e}_2, \mathbf{a}_3 \rangle & \cdots \\ 0 & 0 & \langle \mathbf{e}_3, \mathbf{a}_3 \rangle & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{13}$$

Nótese que $\langle \mathbf{e}_j, \mathbf{a}_j \rangle = \|\mathbf{u}_j\|$, $\langle \mathbf{e}_j, \mathbf{a}_j \rangle = 0$ para $j > k$, y $[\mathbf{Q}][\mathbf{Q}]^{\mathrm{t}} = [\mathbf{I}]$, entonces $[\mathbf{Q}]^{\mathrm{t}} = [\mathbf{Q}]^{-1}$ es ortogonal.

1.3.3. Reflexiones de Householder

Una *reflexión de Householder* es una transformación que refleja el espacio con respecto a un plano determinado. Esta propiedad se puede utilizar para realizar la transformación QR de una matriz si tenemos en cuenta que es posible elegir la matriz de Householder de manera que un vector elegido quede con una única componente no nula tras ser transformado (es decir, premultiplicando por la matriz de Householder). Gráficamente, esto significa que es posible reflejar el vector elegido respecto de un plano de forma que el reflejo quede sobre uno de los ejes de la base cartesiana [Householder,1975] [Burden & Faires,1985].

La manera de elegir el plano de reflexión y formar la matriz de Householder asociada es el siguiente:

Sea $\mathbf{a}_k$ un vector columna arbitrario m-dimensional tal que $\|\mathbf{a}_k\| = |\alpha_k|$, donde α_k es un escalar (si el algoritmo se implementa utilizando aritmética de coma flotante, entonces α debe adoptar el signo contrario que $\mathbf{a}_k$ para evitar pérdida de precisión).

Entonces, siendo $\mathbf{e}_k$ el vector $\{1, 0, ..., 0\}^{\mathrm{t}}$, y $\|\cdot\|$ la norma euclídea, se define

$$\mathbf{u} = \mathbf{a}_k - \alpha_k \mathbf{e}_k \qquad\qquad \mathbf{v} = \frac{\mathbf{u}}{\|\mathbf{u}\|} \qquad\qquad \mathbf{Q} = \mathbf{I} - 2\,\mathbf{v}\mathbf{v}^{\mathrm{t}} \tag{14}$$

El vector $\mathbf{v}$ unitario perpendicular al plano de reflexión elegido. $\mathbf{Q}$ es una matriz de Householder asociada a dicho plano, tal que

$$\mathbf{Q}\mathbf{a}_k = \{\alpha_k, 0, \cdots, 0\}^{\mathrm{t}} \tag{15}$$

Esta propiedad se puede usar para transformar gradualmente los vectores columna de una matriz $\mathbf{A}$ de dimensiones m por n en una matriz triangular superior. En primer lugar, se multiplica $\mathbf{A}$ con la matriz de Householder $\mathbf{Q}$ que obtenemos al elegir como vector $\mathbf{a}_k$ la primera columna de la matriz ($k = 1$). Esto proporciona una matriz $\mathbf{Q}_1\mathbf{A}$ con ceros en la primera columna (excepto el elemento de la primera fila en la diagonal principal, donde aparece una α_k). Esto es,

$$\mathbf{Q}_1\mathbf{A} = \begin{bmatrix} \alpha_1 & \star & \cdots & \star \\ 0 & \ddots & & \\ \vdots & & [\mathbf{A}'] & \\ 0 & & & \ddots \end{bmatrix} \tag{16}$$

y el procedimiento se puede repetir para $\mathbf{A}'$ (que se obtiene de $\mathbf{A}$ eliminando la columna 1), obteniendo así una matriz de Householder $\mathbf{Q}'_2$. Hay que tener en cuenta que $\mathbf{Q}'_2$ es menor que $\mathbf{Q}_1$. Para conseguir que esta matriz opere con $\mathbf{Q}_1\mathbf{A}$ en lugar de $\mathbf{A}'$ es necesario expandirla hacia arriba a la izquierda, completando con unos en $\mathbf{Q}'_k$ y con ceros en $\mathbf{a}_k$ y $\mathbf{e}_k$, sobre la diagonal principal, o en general

$$\mathbf{Q}_k = \begin{bmatrix} \mathbf{I}_{k-1} & 0 \\ 0 & \mathbf{Q}'_k \end{bmatrix} \tag{17}$$

donde $\mathbf{I}_{k-1}$ es la matriz identidad de dimensión $k-1$. Tras repetir el proceso t veces, donde $t = \min(m-1, n)$,

$$\mathbf{R} = \mathbf{Q}_t \cdots \mathbf{Q}_2\mathbf{Q}_1\mathbf{A} \tag{18.a}$$

es una matriz triangular superior. De forma que, tomando

$$\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2 \cdots \mathbf{Q}_t \tag{18.b}$$

donde $\mathbf{A} = \mathbf{Q}^{\mathrm{t}}\mathbf{R}$ es una descomposición QR de la matriz $\mathbf{A}$.

Este método tiene una estabilidad numérica mayor que la del método de Gram-Schmidt descrito arriba.

Una pequeña variación de este método se utiliza para obtener matrices semejantes con la forma de Hessenberg, muy útiles en el cálculo de autovalores por acelerar la convergencia del algoritmo QR reduciendo así enormemente su coste computacional. Existen otros métodos de factorización de tipo QR como el método de rotaciones de Givens, etc.

En cualquiera de los casos antes dscrito el determinante de $\mathbf{A}$ es fácilmente obtenible. Es posible utilizar la descomposición QR para encontrar el valor absoluto del determinante de una matriz. Suponiendo que una matriz se descompone según $\mathbf{A} = \mathbf{QR}$. Entonces se tiene

$$\det(\mathbf{A}) = \det(\mathbf{Q}) \cdot \det(\mathbf{R}) \tag{19}$$

Puesto que $\mathbf{Q}$ es unitaria, $|\det(\mathbf{Q})| = 1$. Por tanto,

$$|\det(\mathbf{A})| = |\det(\mathbf{R})| = \Big|\prod_{i=1}^{n} r_{ii}\Big| \tag{20}$$

donde r_{ii} son los valores de la diagonal de $\mathbf{R}$.

La factorización QR también puede usarse para obtener la solución de un sistema de ecuaciones lineales en la forma

$$[\mathbf{A}].\mathbf{x} = [\mathbf{Q}].[\mathbf{R}].\mathbf{x} = \mathbf{b} \qquad\qquad [\mathbf{R}].\mathbf{x} = [\mathbf{Q}]^{\mathrm{t}}.\mathbf{b} \tag{21}$$

donde, en la expresión (21.b), la solución $\mathbf{x}$ se obtiene de hacer una substitución regresiva, debido a que $[\mathbf{R}]$ es una matriz triangular superior. Al pasar $[\mathbf{Q}]$ al otro miembro, simplemente se traspone debido a que $[\mathbf{Q}]$ es ortogonal $[\mathbf{Q}]^{-1} = [\mathbf{Q}]^{\mathrm{t}}$.

1.3.4. Algoritmo de QR

Sea $\mathbf{A} = \mathbf{A}_0 \in \mathbb{C}^{n\times n}$. El algoritmo QR produce una secuencia $\mathbf{A}_0$, $\mathbf{A}_1$, $\mathbf{A}_2$, ... de matrices similares, como sigue [Stewart,1973]. Dada una matriz $\mathbf{A}_k$, un escalar λ_k llamado *deplazamiento original* se determina de los elementos de $\mathbf{A}_k$ (a medida que las iteraciones convergen, λ_k se acerca a uno de los autovalores de $\mathbf{A}$). La matriz $\mathbf{A}_k - \lambda_k\mathbf{I}$ se puede factorizar de la siguiente forma

$$\mathbf{A}_k - \lambda_k\mathbf{I} = \mathbf{Q}_k\mathbf{R}_k \tag{22}$$

donde $\mathbf{Q}_k$ es unitaria (ortogonal en el caso real) y $\mathbf{R}_k$ es triangular superior. Se sabe que dicha factorización existe bajo ciertas condiciones ($\mathbf{A}_k$ tiene que ser no singular invertible) y es esencialmente única, provisto que $\mathbf{A} - \lambda_k\mathbf{I}$ no es singular. Finalmente, $\mathbf{A}_{k+1}$ se calcula como

$$\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{Q}_k + \lambda_k\mathbf{I} \tag{23}$$

Nótese que de (22), se tiene que $\mathbf{R}_k = \mathbf{Q}_k^{\mathrm{h}}(\mathbf{A}_k - \lambda_k\mathbf{I})$ (la hermitiana $\mathbf{Q}^{\mathrm{h}} = \bar{\mathbf{Q}}^{\mathrm{t}}$ es el transpuesto conjugado), y de aquí a partir de (23) se obtiene

$$\mathbf{A}_{k+1} = \mathbf{Q}_k^{\mathrm{h}}(\mathbf{A}_k - \lambda_k\mathbf{I})\mathbf{Q}_k + \lambda_k\mathbf{I} = \mathbf{Q}_k^{\mathrm{h}}\mathbf{A}_k\mathbf{Q}_k \tag{24}$$

asi que $\mathbf{A}_{k+1}$ es en verdad *unitariamente similar* a $\mathbf{A}_k$.

De aquí en adelante la variantes de los métodos son infinitas. Hay métodos para matrices tridiagonal, matrices de Hessenberg, matrices hermíticas, etc. Dejamos al lector profundice más de acuerdo a sus intereses.

1.3.5. Algoritmo Simplex

Este popular método fue creado en el año de 1947 por el estadounidense George Bernard Dantzig y el ruso Leonid Vitalievich Kantorovich, con el ánimo de crear un algoritmo capaz de solucionar problemas lineales de optimización de m restricciones y n variables positivas ($n < m$) [Dantzig & Thapa,1997/2003].

Consideremos el modelo de programación lineal

$$\begin{cases} \text{Maximizar} \\ \quad z = c_1x_1 + c_2x_2 + \cdots + c_nx_n \\ \text{Sujeto a:} \\ \quad a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \le b_1 \\ \quad a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \le b_2 \\ \qquad \vdots \qquad\quad \vdots \qquad\quad \vdots \\ \quad a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \le b_m \\ \text{con} \\ \qquad x_1, x_2, \ldots, x_n \ge 0 \end{cases} \tag{25}$$

puede ser representado mediante matrices de la siguiente forma

$$\begin{cases} \text{Maximizar} \\ \quad z = \mathbf{c}.\mathbf{x} \\ \text{Sujeto a:} \\ \quad \mathbf{A}.\mathbf{x} \le \mathbf{b} \\ \text{con} \\ \quad \mathbf{x} \ge \mathbf{0} \end{cases} \tag{26}$$

donde

$$\mathbf{c} = \{c_1, c_2, \cdots, c_n\} \qquad \mathbf{x} = \begin{Bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Bmatrix} \qquad \mathbf{b} = \begin{Bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{Bmatrix} \tag{27.a,b,c}$$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \qquad \mathbf{0} = \begin{Bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{Bmatrix}_{m\times 1} \quad \text{ó} \quad \{0, 0, \cdots, 0\}_{1\times m} \tag{27.d,e}$$

Esto es

$$\mathbf{c} \in \mathbb{R}^{1\times n},\ \mathbf{x} \in \mathbb{R}^{n\times 1},\ \mathbf{b} \in \mathbb{R}^{m\times 1},\ \mathbf{A} \in \mathbb{R}^{m\times n} \text{ y } \mathbf{0} \in \mathbb{R}^{n\times 1} \text{ ó } \mathbb{R}^{1\times n} \tag{28}$$

En su totalidad, realmente son $n+m$ restricciones, m restricciones del tipo $\mathbf{A}.\mathbf{x} \leq \mathbf{b}$ y n restricciones del tipo $\mathbf{x} \geq \mathbf{0}$. Consideradas como igualdades, las restricciones son hiperplanos cada una de ellas, que rodean en parte la frontera de una región convexa de $\mathbb{R}^n$, siendo n la dimensión del espacio donde está sumergido el problema. Dicha región poliédrica convexa de factibilidad sería las intersecciones de todas las restricciones, consideradas como desigualdades. Cada dos restricciones se intersectan en una hiperarista (para $n = 2$ la hiperarista es un punto-vértice) y cada n restricciones se intersectan en un único vértice, considerado uno o varios de ellos la solución básica factible al problema (se puede demostrar por reducción al absurdo que la solución al problema debe estar en un vértice y no en una hiperarista o hiperplano [Hillier & Lieberman,2010]). El número total d vértices que pueden ofrecer una solución tiene como cota superior a la combinación $\binom{n+m}{n} = \frac{(n+m)!}{m!n!}$ (que es el mismo número de ecuaciones posibles de formar del total), de las cuales sólamente la mitad o un tercio son factibles.

Para obtener la forma aumentada del problema de programación lineal, introducimos el vector columna de las variables de holgura $\mathbf{x}_s$, esto es

$$\mathbf{x}_s = \begin{Bmatrix} x_{n+1} \\ x_{n+2} \\ \vdots \\ x_{n+m} \end{Bmatrix}_{m\times 1} \tag{29}$$

de tal manera que las restricciones se convierten en una igualdad (sistema de ecuaciones lineales, puesto que cada variable de holgura compensa cada una de las desigualdades)

$$[\mathbf{A}|\,\mathbf{I}\,] \,.\, \begin{Bmatrix} \mathbf{x} \\ \mathbf{x}_s \end{Bmatrix} = \mathbf{b} \qquad \begin{Bmatrix} \mathbf{x} \\ \mathbf{x}_s \end{Bmatrix} \geq \mathbf{0} \tag{30}$$

donde

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{m\times m} \tag{31}$$

es la matriz identidad de orden $m \times m$ (en (30), $\{\mathbf{0}\}_{(n+m)\times 1}$).

Debemos identificar las variables básicas y no básicas de

$$[\mathbf{A}|\,\mathbf{I}\,] \,.\, \begin{Bmatrix} \mathbf{x} \\ \mathbf{x}_s \end{Bmatrix} = \mathbf{b} \tag{32}$$

dado que se tienen que eliminar las variables no básicas al igualarlas a cero entonces queda un conjunto de m ecuaciones con m incógnitas (las variables básicas). Este sistema de ecuaciones lo denotamos por $\mathbf{B}.\mathbf{x}_b = \mathbf{b}$, donde el vector de variables básicas

$$\mathbf{x}_b = \begin{Bmatrix} x_{b1} \\ x_{b2} \\ \vdots \\ x_{bm} \end{Bmatrix} \tag{33}$$

se obtiene al eliminar las variables no básicas del total de variables ($n+m$ variables)

$$\begin{Bmatrix} \mathbf{x} \\ \mathbf{x}_s \end{Bmatrix} \tag{34}$$

escogiendo por defecto las básicas (m variables) y la matriz base $\mathbf{B}$ a partir del sistema más grande (32) (existen en total $\binom{n+m}{m} = \frac{(n+m)!}{m!n!}$ opciones, algunas válidas, invertibles, esas son el número de iteraciones, y otras no, descartadas). Esta matriz denotada por $\mathbf{B} \in \mathbb{R}^{m \times m}$ es

$$\mathbf{B} = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mm} \end{bmatrix} \tag{35}$$

y se obtiene al eliminar las columnas de

$$[\mathbf{A}|\,\mathbf{I}\,] \tag{36}$$

correspondientes a las variables no básicas. Como la matriz base $\mathbf{B}$ es invertible entonces la solución deseada para las variables básicas es $\mathbf{x}_b = \mathbf{B}^{-1}.\,\mathbf{b}$. La escogencia de las variables básicas es fundamental en el método, puesto que cada escogencia debe ser un sistema invertible y da un valor de z diferente, y la correcta es la solución al problema con el mayor valor de z.

Sea $\mathbf{c}_b$ el vector renglón cuyos elementos son los coeficientes de la función objetivo (incluye los ceros para las variables de holgura) que corresponden a los elementos de $\mathbf{x}_b$. Así, el vector de la función objetivo de la solución básica $\mathbf{x}_b = \mathbf{B}^{-1}.\,\mathbf{b}$ es $z = \mathbf{c}_b.\mathbf{x}_b = \mathbf{c}_b.\mathbf{B}^{-1}.\,\mathbf{b}$.

En el caso del conjunto de ecuaciones originales del modelo inicial aumentado, incluyendo la ecuación de la función objetivo z, se puede representar como

$$\begin{bmatrix} 1 & -\mathbf{c} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \mathbf{I} \end{bmatrix} \cdot \begin{Bmatrix} z \\ \mathbf{x} \\ \mathbf{x}_s \end{Bmatrix} = \begin{Bmatrix} 0 \\ \mathbf{b} \end{Bmatrix} \tag{37}$$

dado que $\mathbf{x}_b = \mathbf{B}^{-1}.\,\mathbf{b}$ y $z = \mathbf{c}_b.\mathbf{B}^{-1}.\,\mathbf{b}$ entonces

$$\begin{Bmatrix} z \\ \mathbf{x}_b \end{Bmatrix} = \begin{Bmatrix} \mathbf{c}_b.\mathbf{B}^{-1}.\,\mathbf{b} \\ \mathbf{B}^{-1}.\,\mathbf{b} \end{Bmatrix} = \begin{bmatrix} 1 & \mathbf{c}_b.\mathbf{B}^{-1} \\ \mathbf{0} & \mathbf{B}^{-1} \end{bmatrix} \cdot \begin{Bmatrix} 0 \\ \mathbf{b} \end{Bmatrix} \tag{38}$$

Premultiplicando la última matriz con la expresión hallada anteriormente obtenemos

$$\begin{bmatrix} 1 & \mathbf{c}_b.\mathbf{B}^{-1} \\ \mathbf{0} & \mathbf{B}^{-1} \end{bmatrix} \cdot \begin{bmatrix} 1 & -\mathbf{c} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \mathbf{I} \end{bmatrix} \cdot \begin{Bmatrix} z \\ \mathbf{x} \\ \mathbf{x}_s \end{Bmatrix} = \begin{bmatrix} 1 & \mathbf{c}_b.\mathbf{B}^{-1} \\ \mathbf{0} & \mathbf{B}^{-1} \end{bmatrix} \cdot \begin{Bmatrix} 0 \\ \mathbf{b} \end{Bmatrix} \tag{39}$$

$$\begin{bmatrix} 1 & \mathbf{c}_b.\mathbf{B}^{-1}.\,\mathbf{A} - \mathbf{c} & \mathbf{c}_b.\mathbf{B}^{-1} \\ \mathbf{0} & \mathbf{B}^{-1}.\,\mathbf{A} & \mathbf{B}^{-1} \end{bmatrix} \cdot \begin{Bmatrix} z \\ \mathbf{x} \\ \mathbf{x}_s \end{Bmatrix} = \begin{Bmatrix} \mathbf{c}_b.\mathbf{B}^{-1}.\,\mathbf{b} \\ \mathbf{B}^{-1}.\,\mathbf{b} \end{Bmatrix} \tag{40}$$

Está forma matricial proporciona el conjunto de ecuaciones de cualquier iteración. Se utilizan las expresiones matriciales $\mathbf{c}_b.\mathbf{B}^{-1}.\mathbf{A}-\mathbf{c}$ y $\mathbf{c}_b.\mathbf{B}^{-1}$ para calcular los coeficientes de las variables no básicas de la función objetivo, a partir de la primera de las dos ecuaciones matriciales de arriba, (40).

Consideremos ahora el problema de la dieta de minimización

$$\begin{cases} \text{Minimizar} \\ \quad z=\mathbf{c}.\mathbf{x} \\ \text{Sujeto a:} \\ \quad \mathbf{A}.\mathbf{x}\geq\mathbf{b} \\ \text{con} \\ \quad \mathbf{x}\geq\mathbf{0} \end{cases} \tag{41}$$

La forma ampliada en este caso se logra mediante

$$[\mathbf{A}|-\mathbf{I}]\,.\left\{\begin{array}{c}\mathbf{x}\\ \mathbf{x}_s\end{array}\right\}=\mathbf{b} \qquad\qquad \left\{\begin{array}{c}\mathbf{x}\\ \mathbf{x}_s\end{array}\right\}\geq\mathbf{0} \tag{42}$$

en lugar de (30), donde se ha invertido el procedimiento restando la variables de holgura, ahora de exceso, una para cada desigualdad para compensar la diferencia. El procedimiento iterativo se vuelve similar al anterior, sólo que en éste, al contrario, se busca minimizar z. Las ecuaciones (37) y (39) ahora cambian sustituyendo $-\mathbf{I}$ en lugar de $\mathbf{I}$, -1 en lugar de 1 y $\mathbf{c}$ en lugar de $-\mathbf{c}$, como se indica a continuación

$$\begin{bmatrix}-1 & \mathbf{c} & \mathbf{0}\\ \mathbf{0} & \mathbf{A} & -\mathbf{I}\end{bmatrix}\cdot\left\{\begin{array}{c}z\\ \mathbf{x}\\ \mathbf{x}_s\end{array}\right\}=\left\{\begin{array}{c}0\\ \mathbf{b}\end{array}\right\} \tag{43}$$

lógicamente también cambia de signo la última columna de la matrix de (40)

Una forma de evitarse el trabajo de escoger las variables básicas en el sistema (37), en cada iteración, de

$$\tilde{\mathbf{A}}.\hat{\mathbf{x}}=\tilde{\mathbf{b}} \qquad \tilde{\mathbf{A}}=\begin{bmatrix}1 & -\mathbf{c} & \mathbf{0}\\ \mathbf{0} & \mathbf{A} & \mathbf{I}\end{bmatrix} \qquad \tilde{\mathbf{b}}=\left\{\begin{array}{c}0\\ \mathbf{b}\end{array}\right\} \qquad \hat{\mathbf{x}}=\left\{\begin{array}{c}z\\ \mathbf{x}\\ \mathbf{x}_s\end{array}\right\} \tag{44}$$

y resolver el problema en un solo paso, realizamos la pre-multiplicación del sistema de ecuaciones lineales anterior por la matriz $\tilde{\mathbf{A}}^{\mathrm{t}}$, con lo cual queda un sistema completo en toda las variables $\hat{\mathbf{x}}$, con igual número $(n+m+1)$ de variables y ecuaciones

$$\hat{\mathbf{A}}.\hat{\mathbf{x}}=\hat{\mathbf{b}} \qquad \hat{\mathbf{A}}=\tilde{\mathbf{A}}^{\mathrm{t}}.\tilde{\mathbf{A}}=\begin{bmatrix}1 & -\mathbf{c} & \mathbf{0}\\ -\mathbf{c}^{\mathrm{t}} & \mathbf{c}^{\mathrm{t}}\mathbf{c}+\mathbf{A}^{\mathrm{t}}.\mathbf{A} & \mathbf{A}^{\mathrm{t}}\\ \mathbf{0} & \mathbf{A} & \mathbf{I}\end{bmatrix} \qquad \hat{\mathbf{b}}=\tilde{\mathbf{A}}^{\mathrm{t}}.\tilde{\mathbf{b}}=\left\{\begin{array}{c}0\\ \mathbf{A}^{\mathrm{t}}.\mathbf{b}\\ \mathbf{b}\end{array}\right\} \tag{45}$$

el cual es compatible determinado (la matriz $\tilde{\mathbf{A}}$ es simétrica y diagonalmente dominante). El término $\mathbf{c}^{\mathrm{t}}\mathbf{c}$ equivalente a la diádica $\mathbf{cc}=\mathbf{c}\otimes\mathbf{c}$, es una matriz (componentes de un tensor 2^{do} orden diádico $\mathbf{cc}$) cuadrada simétrica de $n\times n$, al igual que $\mathbf{A}^{\mathrm{t}}.\mathbf{A}$. Para la minimización lo único que cambia en $\hat{\mathbf{A}}$ es $\mathbf{A}$, $\mathbf{A}^{\mathrm{t}}$ y $\mathbf{b}$, que en su lugar son los opuestos a estos, todo lo demás queda igual.

Para los problemas no lineales de maximizar o minimizar el planteamiento es el siguiente

$$\begin{cases} \text{Maximizar} \\ \quad z=h(\mathbf{x}) \\ \text{Sujeto a:} \\ \quad \mathbf{g}(\mathbf{x})\leq\mathbf{b} \\ \text{con} \\ \quad \mathbf{x}\geq\mathbf{0} \end{cases} \qquad\qquad \begin{cases} \text{Minimizar} \\ \quad z=h(\mathbf{x}) \\ \text{Sujeto a:} \\ \quad \mathbf{g}(\mathbf{x})\geq\mathbf{b} \\ \text{con} \\ \quad \mathbf{x}\geq\mathbf{0} \end{cases} \tag{46}$$

donde las funciones objetivo $h(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ y de restricciones $\mathbf{g}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$ ahora son no lineales, lo que hace que en parte el poliedro de la región de factibilidad tenga como fronteras hipersuperficies curvas en lugar de hiperplanos. Se replantea el problema para generar un sistema de ecuaciones no lineales homogénea, de la forma

$$\left\{\begin{array}{l} \text{Maximizar } z \\ \quad z - h(\mathbf{x}) = 0 \\ \text{Sujeto a:} \\ \quad \mathbf{g}(\mathbf{x}) + \mathbf{I}.\mathbf{x}_s - \mathbf{b} = \mathbf{0} \\ \text{con} \\ \quad \mathbf{x} \geq \mathbf{0} \end{array}\right. \qquad \left\{\begin{array}{l} \text{Minimizar } z \\ \quad h(\mathbf{x}) - z = 0 \\ \text{Sujeto a:} \\ \quad \mathbf{g}(\mathbf{x}) - \mathbf{I}.\mathbf{x}_s - \mathbf{b} = \mathbf{0} \\ \text{con} \\ \quad \mathbf{x} \geq \mathbf{0} \end{array}\right. \tag{46'}$$

Un procedimiento para resolver este sistema de ecuaciones no lineales con un algoritmo iterativo es usando el método de Newton-Raphson (sección 2.2.(5), aunque el número de ecuaciones $m+1$ no sea igual al número de incógnitas $m+n+1$), trabajando con las variables totales $\hat{\mathbf{x}} = \{z, \mathbf{x}, \mathbf{x}_s\}^{\mathrm{t}}$. De aquí resulta que la solución del siguiente sistema de ecuaciones lineales evaluado en la iteración k

$$[\mathbf{A}(\hat{\mathbf{x}}^k)].\Delta\hat{\mathbf{x}}^k = -\mathbf{f}(\hat{\mathbf{x}}^k) \qquad [\mathbf{A}(\hat{\mathbf{x}})] = \begin{bmatrix} \pm 1 & \mp\boldsymbol{\nabla} h(\mathbf{x}) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{\mathbf{g}}(\mathbf{x}) & \pm\mathbf{I} \end{bmatrix} \qquad \mathbf{f}(\hat{\mathbf{x}}) = \left\{\begin{array}{c} \pm z \mp h(\mathbf{x}) \\ \mathbf{g}(\mathbf{x}) \pm \mathbf{I}.\mathbf{x}_s - \mathbf{b} \end{array}\right\} \tag{47}$$

(Signo de arriba para maximizar y el de abajo para minimizar) nos permite encontrar el nuevo iterado en la iteración $k+1$

$$\hat{\mathbf{x}}^{k+1} = \hat{\mathbf{x}}^k + \omega\, \Delta\hat{\mathbf{x}}^k \qquad \mathbf{J}_{\mathbf{g}}(\mathbf{x}) = [\boldsymbol{\nabla}\mathbf{g}(\mathbf{x})]^{\mathrm{t}} \tag{48}$$

donde $\boldsymbol{\nabla} h(\mathbf{x})$ es el gradiente de la función escalar $h(\mathbf{x})$ evaluada en $\mathbf{x}$ y $\mathbf{J}_{\mathbf{g}}(\mathbf{x})$ es el jacobiano de la función vectorial $\mathbf{g}(\mathbf{x})$ evaluada en $\mathbf{x}$ y ω es el factor de relajación ($\boldsymbol{\nabla} = \mathbf{e}_i \partial_i \in \mathbb{R}^n$).

Cuando pre-multiplicamos el sistema de ecuaciones lineales con desiguales incógnitas y ecuación por $[\mathbf{A}(\hat{\mathbf{x}}^k)]^{\mathrm{t}}$ igualamos el número de ecuaciones y variables a $(n+m+1)$, similar a como se hizo en (45). Con lo cual nos queda el sistema lineal para la iteración k de la forma

$$[\hat{\mathbf{A}}(\hat{\mathbf{x}}^k)].\Delta\hat{\mathbf{x}}^k = \hat{\mathbf{b}}(\hat{\mathbf{x}}^k) \qquad [\hat{\mathbf{A}}(\hat{\mathbf{x}}^k)] = [\mathbf{A}(\hat{\mathbf{x}}^k)]^{\mathrm{t}}.\,[\mathbf{A}(\hat{\mathbf{x}}^k)] \qquad \hat{\mathbf{b}}(\hat{\mathbf{x}}^k) = -[\mathbf{A}(\hat{\mathbf{x}}^k)]^{\mathrm{t}}.\,\mathbf{f}(\hat{\mathbf{x}}^k) \tag{49}$$

donde, después de operar con las matrices queda la matriz del sistema

$$[\hat{\mathbf{A}}(\hat{\mathbf{x}})] = \begin{bmatrix} 1 & -\boldsymbol{\nabla} h(\mathbf{x}) & \mathbf{0} \\ -\boldsymbol{\nabla} h(\mathbf{x}) & \boldsymbol{\nabla} h(\mathbf{x})^{\mathrm{t}}\boldsymbol{\nabla} h(\mathbf{x}) + [\mathbf{J}_{\mathbf{g}}(\mathbf{x})]^{\mathrm{t}}.[\mathbf{J}_{\mathbf{g}}(\mathbf{x})] & \pm[\mathbf{J}_{\mathbf{g}}(\mathbf{x})] \\ \mathbf{0} & \pm\mathbf{J}_{\mathbf{g}}(\mathbf{x})] & \mathbf{I} \end{bmatrix} \tag{50.a}$$

y el término independiente

$$\hat{\mathbf{b}}(\hat{\mathbf{x}}) = -\left\{\begin{array}{c} z - h(\mathbf{x}) \\ \boldsymbol{\nabla} h^{\mathrm{t}}[\, h(\mathbf{x}) - z\,] + [\mathbf{J}_{\mathbf{g}}(\mathbf{x})]^{\mathrm{t}}.\,[\,\mathbf{g}(\mathbf{x}) \pm \mathbf{I}.\mathbf{x}_s - \mathbf{b}\,] \\ \pm[\,\mathbf{g}(\mathbf{x}) \pm \mathbf{I}.\mathbf{x}_s - \mathbf{b}\,] \end{array}\right\} \tag{50.b}$$

los cuales cambian ambos en cada iteración. La observaciones hechas antes para el termino central de (45.b) son también válidas para el término central de (50.a)

2. SISTEMAS NO-LINEALES

A diferencia de los sistemas de ecuaciones lineales, los sistemas de ecuaciones no-lineales no pueden ser agrupados en forma matricial por lo tanto ninguno de los algoritmos discutidos en la sección anterior podrían ser aplicados sobre ellos. Un sistema de ecuaciones no-lineales no es más que una extensión del problema de hallar la raíz de una ecuación no-líneal a hallar n raíces para las n incógnitas que se tienen, por ello uno

de los métodos más utilizados para resolverlos es el *método de Newton-Raphson* extendido para sistemas de ecuaciones, sin olvidar que existen otros algoritmos muy eficientes para casos particulares.

El objetivo es resolver un sistema de ecuaciones algebraicas de la forma:

$$\begin{aligned} f_1(\mathbf{x}) =& f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(\mathbf{x}) =& f_2(x_1, x_2, \ldots, x_n) = 0 \\ & \vdots \\ f_n(\mathbf{x}) =& f_n(x_1, x_2, \ldots, x_n) = 0 \end{aligned} \tag{1}$$

o en forma más compacta usando la notación simbólica

$$\mathbf{f}(\mathbf{x}) = 0 \tag{2}$$

A esta ecuación le llamaremos la *ecuación homogénea* y a su solución $\mathbf{r}$, raíz de la ecuación $\mathbf{f}(\mathbf{r}) \equiv \mathbf{0}$.

2.1. METODO DEL PUNTO FIJO

Cualquier manipulación algebraica del sistema (1), espejando una componente diferente en cada ecuación nos da un sistema, que de forma resumida se expresa como

$$\mathbf{x} = \mathbf{g}(\mathbf{x}) \tag{3}$$

Se puede entonces implementar un esquema iterativo de la forma

$$\mathbf{x}^{k+1} = \mathbf{g}(\mathbf{x}^k) \tag{4}$$

Encontraremos la solución $\mathbf{r}$ de dicho sistema (también solución del sistema (2)), cuando comenzando con un iterado inicial $\mathbf{x}^o$ se llega a un punto donde

$$\mathbf{r} \equiv \mathbf{g}(\mathbf{r}) \tag{5}$$

A este punto le denominamos el *punto fijo* $\mathbf{r}$.

Una expansión en series de Taylor de primer orden de la función $\mathbf{g}(\mathbf{x})$, centrado en $\mathbf{r}$ y evaluado en $\mathbf{x}^k$ nos da que

$$\mathbf{g}(\mathbf{x}^k) = \mathbf{g}(\mathbf{r}) + (\mathbf{x}^k - \mathbf{r}).\boldsymbol{\nabla}\mathbf{g}(\mathbf{r}) + O(\|\mathbf{e}^k\|^2) \tag{6}$$

Eliminando el término con $O(\|\mathbf{e}^k\|^2)$, introduciendo (4) y (5), y evaluando en gradiente de $\mathbf{g}$ en un punto intermedio $\boldsymbol{\zeta}$

$$\mathbf{x}^{k+1} - \mathbf{r} = [\mathbf{J}_\mathbf{g}(\boldsymbol{\zeta})] \,.\, (\mathbf{x}^k - \mathbf{r}) \qquad \mathbf{e}^{k+1} = [\mathbf{J}_\mathbf{g}(\boldsymbol{\zeta})] \,.\, \mathbf{e}^k \qquad \boldsymbol{\zeta} \in \mathbb{B}(\mathbf{r}, \|\mathbf{e}^k\|) \tag{7}$$

donde $\mathbb{B}(\mathbf{r}, \|\mathbf{e}^k\|)$ es la $\mathbb{R}^n$ bola cerrada con centro en $\mathbf{r}$ y radio $\|\mathbf{e}^k\|$. El tensor $\mathbf{J}_\mathbf{g}(\mathbf{x})$ es el jacobiano de la función $\mathbf{g}$ definido como

$$\mathbf{J}_\mathbf{g}(\mathbf{x}) = [\boldsymbol{\nabla}\mathbf{g}(\mathbf{x})]^{\mathrm{t}} \qquad [\mathbf{J}_\mathbf{g}(\mathbf{x})]_{ij} = \frac{\partial g_i}{\partial x_j} \tag{8}$$

Del lado derecho se muestra como se calculan las componente matricial del tensor jacobiano.

Obteniendo la norma de la expresión (7.*b*), resulta

$$\|\mathbf{e}^{k+1}\| \leq \|\mathbf{J}_\mathbf{g}(\boldsymbol{\zeta})\| \, \|\mathbf{e}^k\| \tag{9}$$

Lo que nos dice esta expresión es que el proceso iterativo (4) es convergente, o sea los errores $\mathbf{e}^k$ son cada vez menores, si se satisface que

$$\|\mathbf{J_g}(\boldsymbol{\zeta})\| < 1 \tag{10}$$

En términos geométricos significa que la función $\mathbf{g}$ debe ser una contacción alrededor de $\mathbf{r}$. La función $\mathbf{g}$ deforma el espacio alrededor de $\mathbf{r}$, tal que contrae sus dimensiones o volumen.

EJEMPLO:

Hallar la solución del siguiente sistema de ecuaciones no-lineales

$$x^2 + y^2 = 4$$
$$\mathrm{e}^x - y = 1$$

Utilizando un método iterativo de punto fijo se pueden obtener dos tipos de despejes

Caso I $\quad \begin{cases} x = -\sqrt{4 - y^2} \\ y = 1 - \mathrm{e}^x \end{cases}$ $\qquad$ Caso II $\quad \begin{cases} x = \ln(1 - y) \\ y = -\sqrt{4 - x^2} \end{cases}$

Resultados

Caso I

x		-1.83	-1.815	-1.8163	-1.8162
y	0.8	0.84	0.8372	0.8374	0.8374

Caso I

x		1.05	0.743	1.669	Imaginario
y	-1.7	-1.857	-1.102	-4.307	

Caso II

x		0.993	1.006	1.0038	1.0042	1.0042
y	-1.7	-1.736	-1.7286	-1.7299	-1.7296	-1.7296

2.2. METODOS DE NEWTON-RAPHSON

los *métodos de Newton-Raphson* se deducen a partir de la expansión en series de Taylor de la función $\mathbf{f}$ alrededor de $\mathbf{x}$ y evaluado en $\mathbf{r}$. Esta ecuación pueden ser truncadas después del término de primer orden. Igualándola a cero, como indica 2.(2) para la función, se obtiene

$$\mathbf{f}(\mathbf{r}) = \mathbf{f}(\mathbf{x}) + (\mathbf{r} - \mathbf{x}) \,.\boldsymbol{\nabla}\mathbf{f}(\mathbf{x}) + \mathbf{O}(\|\mathbf{e}\|^2) \equiv \mathbf{0} \tag{1}$$

donde $\mathbf{e} = \mathbf{x} - \mathbf{r}$ es el error global.

Al despejar $\mathbf{r}$ de esta ecuación queda

$$\mathbf{r} = \mathbf{x} - \{\, [\boldsymbol{\nabla}\mathbf{f}(\mathbf{x})]^{\mathrm{t}} \}^{-1} . [\, \mathbf{f}(\mathbf{x}) + \mathbf{O}(\|\mathbf{e}\|^2) \,] \tag{2}$$

donde se ha invertido el transpuesto de $\boldsymbol{\nabla}\mathbf{f}$. Veremos más adelante que cambiando la notación y estandarizando el procedimiento resulta más sencillo.

2.2.1. Simple

Si no se toma en consideración el término $\mathbf{O}(\|\mathbf{e}\|^2)$, la expresión anterior no se iguala a $\mathbf{r}$ exactamente, pero si a un valor cercano $\mathbf{r}$. De acuerdo a este razonamiento, se puede substituir $\mathbf{r}$ por $\mathbf{x}^{k+1}$ y $\mathbf{x}$ por $\mathbf{x}^k$ y aplicar un procedimiento iterativo de la forma

$$\mathbf{x}^{k+1} = \mathbf{x}^k - [\mathbf{J_f}(\mathbf{x}^k)]^{-1}.\mathbf{f}(\mathbf{x}^k) \tag{3}$$

donde $[\mathbf{J_f}(\mathbf{x}^k)] = [\boldsymbol{\nabla}\mathbf{f}(\mathbf{x}^k)]^{\mathrm{t}}$ es la matriz del jacobiano de la función $\mathbf{f}$ evaluada en el punto $\mathbf{x}^k$. Así, es más práctico escribir la expresión (3) como

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{z}^k \tag{4}$$

donde $\mathbf{z}^k$ es la solución del sistema de ecuaciones lineales

$$[\mathbf{J_f}(\mathbf{x}^k)]\,.\,\mathbf{z}^k = -\mathbf{f}(\mathbf{x}^k) \tag{5}$$

El proceso iterativo establecido se aplica, comenzando con un estimado del valor inicial de la raíz que se llamará $\mathbf{x}^o$, en forma sucesiva, hasta que se satisfagan las tolerancias ϵ_{max} y d_{max} (impuestas al principio al igual que k_{max})

$$\|\boldsymbol{\epsilon}^k\| < \epsilon_{max} \qquad \text{y} \qquad \|\mathbf{d}^k\| < d_{max} \tag{6}$$

donde $\boldsymbol{\epsilon}^k$ y $\mathbf{d}^k$ son el error local y la desviacion global, respectivamente, y se definen de la siguiente manera

$$\boldsymbol{\epsilon}^k = \mathbf{x}^k - \mathbf{x}^{k-1} \qquad\qquad \mathbf{d}^k = \mathbf{f}(\mathbf{x}^k) \tag{7}$$

La norma $\|\cdot\|$ usada en este análisis es la norma euclidiana que se define como

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}.\mathbf{x}} \qquad\qquad \|\mathbf{A}\| = \sqrt{\rho(\mathbf{A}^{\mathrm{t}}.\,\mathbf{A})} \tag{8}$$

2.2.2. Relajado

El método de Newton-Raphson puede ser relajado en la forma

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \omega\,\mathbf{z}^k \tag{9}$$

donde ω es el factor de relajación, y podrá tomar los siguientes valores

$\omega > 1$ sobrerelajado

$\omega < 1$ subrelajado

El valor de $\mathbf{z}_k$ igualmente que antes, se obtiene del sistema de ecuaciones lineales (5).

2.3. METODOS CUASI-NEWTON

Los *métodos cuasi-Newton* se basa en los dos siguientes lemas preparativos para formular sus fórmulas algorítmicas [Dennis & Moré,(1977)]

Lema 1. Sea $\mathbf{A}$ una transformación lineal $\mathbf{A}.\mathbf{x} = \tilde{\mathbf{x}}$, tal que a todo vector perpendicular a $\mathbf{x}$, es decir $\mathbf{x}.\mathbf{y} = 0$, lo envía a $\mathbf{C}.\mathbf{y}$, con $\mathbf{C}$ siendo otra transformación lineal conocida. Con estas condiciones la transformación $\mathbf{A}$ que definida univocamente como

$$\mathbf{A} = \mathbf{C} + \frac{(\tilde{\mathbf{x}} - \mathbf{C}.\mathbf{x})\mathbf{x}}{\|\mathbf{x}\|^2} \tag{1}$$

La prueba de este lema se hace con $\mathbf{A}.\mathbf{y} = \mathbf{C}.\mathbf{y} \Longrightarrow (\mathbf{A}-\mathbf{C}).\mathbf{y} = \mathbf{0} \Longrightarrow \mathbf{A}-\mathbf{C} = \mathbf{a}\mathbf{x} \Longrightarrow (\mathbf{A}-\mathbf{C}).\mathbf{x} = \mathbf{a}\mathbf{x}.\mathbf{x}$ $\Longrightarrow \mathbf{a} = (\mathbf{A}-\mathbf{C}).\mathbf{x}/\|\mathbf{x}\|^2$.

Lema 2. Sea $\mathbf{A}$ una transformación lineal cuya matriz es no singular, es decir $\det \mathbf{A} \neq 0$. Ai $\mathbf{a}$ y $\mathbf{b}$ son dos vectores, tales que $\mathbf{b}.\mathbf{A}^{-1}\!.\mathbf{a} \neq -1$, entonces $\mathbf{A} + \mathbf{ab}$ es no singular y se tiene que

$$(\mathbf{A} + \mathbf{ab})^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\!.\mathbf{ab}\,.\,\mathbf{A}^{-1}}{1 + \mathbf{b}\,.\,\mathbf{A}^{-1}\!.\mathbf{a}} \tag{2}$$

La prueba de este lema se hace multiplicando $(\mathbf{A} + \mathbf{ab}).(\mathbf{A} + \mathbf{ab})^{-1} = \mathbf{I}$.

2.3.1. Método de Broyden

La fórmula algorítmica del *método de Broyden* es [Broyden,(1965)]

$$\mathbf{x}^{k+1} = \mathbf{x}^k - [\mathbf{A}^k]^{-1}\!.\mathbf{f}(\mathbf{x}^k) \tag{3}$$

tal que

$$[\mathbf{A}^k].\boldsymbol{\epsilon}^k = \boldsymbol{\delta}^k \qquad [\mathbf{A}^k].\mathbf{y} = [\mathbf{A}^{k-1}].\mathbf{y} \qquad \boldsymbol{\epsilon}^k.\mathbf{y} = 0 \tag{4}$$

donde $\boldsymbol{\epsilon}^k = \mathbf{x}^k - \mathbf{x}^{k-1}$ es el error local y $\boldsymbol{\delta}^k = \mathbf{f}(\mathbf{x}^k) - \mathbf{f}(\mathbf{x}^{k-1})$ es la desviación local.

La transformación $[\mathbf{A}^k]$ es una aproximación de la transformación jacobiana $[\mathbf{J_f}(\mathbf{x}^k)]$, que deja el espacio ortogonal a $\boldsymbol{\epsilon}^k$ tal como lo deja la transformación anterior $(4.b)$. Esto define por el lema 1 de forma única

$$\mathbf{A}^k = \mathbf{A}^{k-1} + \frac{(\,\boldsymbol{\delta}^k - [\mathbf{A}^{k-1}]\,.\,\boldsymbol{\epsilon}^k\,)\,\boldsymbol{\epsilon}^k}{\|\boldsymbol{\epsilon}^k\|^2} \tag{5}$$

Por el lema 2 se tiene una forma alterna de $[\mathbf{A}^k]^{-1} = [\mathbf{B}^k]$

$$\mathbf{B}^k = \left[\mathbf{I} + \frac{(\,\boldsymbol{\epsilon}^k - [\mathbf{B}^{k-1}]\,.\,\boldsymbol{\delta}^k\,)\,\boldsymbol{\epsilon}^k}{\boldsymbol{\epsilon}^k\,.\,[\mathbf{B}^{k-1}]\,.\,\boldsymbol{\delta}^k}\right].\mathbf{B}^{k-1} \tag{6}$$

Entonces la fórmula algorítmica queda

$$\mathbf{x}^{k+1} = \mathbf{x}^k - [\mathbf{B}^k]\,.\,\mathbf{f}(\mathbf{x}^k) \tag{7}$$

donde $[\mathbf{B}^k]$ es una aproximación del jacobiano inverso $[\mathbf{J_f}(\mathbf{x}^k)]^{-1}$ calculado con (6) de forma recurrente.

2.4. METODOS DE MINIMOS CUADRADOS

Este método reformula el objetivo del problema a resolver, que de encontrar la solución de la ecuación homogénea 2.(2), se pasa ahora a encontrar el mínimo de la función S definida por

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \qquad S(\mathbf{x}) = \mathbf{f}(\mathbf{x})\,.\,\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{n}[f_i(\mathbf{x})]^2 \tag{1}$$

El gradiente de esta función es

$$\boldsymbol{\nabla} S(\mathbf{x}) = 2\,[\boldsymbol{\nabla}\mathbf{f}(\mathbf{x})]\,.\,\mathbf{f}(\mathbf{x}) = 2\,\mathbf{f}(\mathbf{x})\,.\,[\mathbf{J_f}(\mathbf{x})] \tag{2}$$

Indica la dirección en la cual S aumenta. Por consiguiente, para encontrar el mínimo la iteraciones debes dirigirse en sentido contrario $-\boldsymbol{\nabla} S$ y este gradiente se afecta con un factor de relajación $\omega/2$ (el $1/2$ es simplemente para eliminar el 2 en la ecuación (2)) que se ajusta de manera tal que el descenso de S sea el máximo posible. De esta forma se implementa el esquema iterativo

$$\begin{aligned} \mathbf{x}^{k+1} &= \mathbf{x}^k - \frac{\omega}{2}\,\boldsymbol{\nabla} S(\mathbf{x}^k) \\ &= \mathbf{x}^k - \omega\,\mathbf{f}(\mathbf{x}^k)\,.\,[\mathbf{J_f}(\mathbf{x}^k)] \end{aligned} \tag{3}$$

Si se tiene que $S(\mathbf{x}^{k+1}) < S(\mathbf{x}^k)$, entonces $\omega' = \tau\omega$ $(\tau > 1)$ para la próxima iteración. En caso contrario $\omega' = \rho\omega$ $(\rho < 1)$ y se prueba de nuevo calculando otra vez $\mathbf{x}^{k+1}$ y $S(\mathbf{x}^{k+1})$. Normalmente se escoge el crecimiento de ω menor que su disminución $((\tau - 1) < (1 - \rho))$.

En realidad hay que hacer como mínimo tres intentos ω_1, ω_2 y ω_3, obteniendo $S_1(\mathbf{x}^{k+1})$, $S_2(\mathbf{x}^{k+1})$ y $S_3(\mathbf{x}^{k+1})$, para luego de hacer una interpolación o extrapolación cuadrática, y obtener un valor de ω óptimo (usar por ejemplo el método de Muller sección I.2.5.2 o el método de la parábola secante sección I.2.5.3). Este valor óptimo estará cerca del valor ofrecido por

$$\left.\frac{\partial S}{\partial \omega}\right|^{k+1} = -\mathbf{f}(\mathbf{x}^k) \,.\, [\mathbf{J_f}(\mathbf{x}^k)] \,.\, \boldsymbol{\nabla} S(\mathbf{x}^{k+1}) = 0 \tag{4}$$

que da el óptimo analítico de ω para minimizar $S(\mathbf{x}^{k+1})$.

2.5. METODOS DE SEGUNDO ORDEN

Los *método se segundo orden* se originan mediante la expansión en series de Taylor de segundo orden de la función $\mathbf{f}(\mathbf{x})$ de la ecuación homogénea 2.(2). Se hace la expansión alrededor de $\mathbf{x}$ y se evalúa en la raíz $\mathbf{r}$ en la forma

$$\mathbf{f}(\mathbf{r}) = \mathbf{f}(\mathbf{x}) + (\mathbf{r} - \mathbf{x}) \,.\, \nabla\mathbf{f}(\mathbf{x}) + \frac{1}{2}\,(\mathbf{r} - \mathbf{x})(\mathbf{r} - \mathbf{x}) \;:\; \nabla\nabla\mathbf{f}(\mathbf{x}) + \mathbf{O}(\|\mathbf{e}\|^3) \tag{1}$$

donde $\mathbf{e} = \mathbf{x} - \mathbf{r}$ es el error global del valor de $\mathbf{x}$ respecto a la raíz $\mathbf{r}$, y por definición de una raíz, $\mathbf{f}(\mathbf{r}) \equiv 0$.

La operación doble producto ":" es una doble contracción de los índices contiguos de las componentes de los factores (identificable como el producto escalar de dos tensores de segundo orden), mientras que un solo punto es una contracción simple (identificable como el producto escalar de dos vectores). Esto hace que los vectores y tensores descritos pertenezcan a espacios de Hilbert. Los dos vectores contiguos (sin ninguna operación en el medio) es lo que se denomina una diádica equivalente a un tensor de segundo orden $(\mathbf{ab} \equiv \mathbf{a} \otimes \mathbf{b})$.

Eliminando el término con $\mathbf{O}(\|\mathbf{e}\|^3)$ y cambiando la notación en (1), se puede expresar como

$$\mathbf{f}(\mathbf{r}) = \mathbf{f}(\mathbf{x}) + [\mathbf{J_f}(\mathbf{x})] \,.\, (\mathbf{r} - \mathbf{x}) + \tfrac{1}{2}\,[\mathbf{H_f}(\mathbf{x})] \;:\; (\mathbf{r} - \mathbf{x})(\mathbf{r} - \mathbf{x}) = \mathbf{0} \tag{2}$$

El tensor de segundo orden $\mathbf{J}_f$ en la expansión en serie anterior se denomina el tensor *jabobiano*, se define como $[\mathbf{J_f}(\mathbf{x})] \equiv [\boldsymbol{\nabla}\mathbf{f}(\mathbf{x})]^{\mathrm{t}}$, y agrupados de forma matricial en un arreglo de dos índices tiene componentes

$$[\mathbf{J_f}(\mathbf{x})]_{ij} \equiv J_{ij} = \frac{\partial f_i}{\partial x_j} \tag{3.a}$$

El tensor de tercer orden $\mathbf{H}_f$ en la expansión en serie anterior se denomina el tensor *hessiano*, se define como $[\mathbf{H_f}(\mathbf{x})] \equiv [\boldsymbol{\nabla}[\boldsymbol{\nabla}\mathbf{f}(\mathbf{x})]^{\mathrm{t}}]^{\mathrm{t}}$, y agrupados de forma indicial en un arreglo de tres índices tiene componentes

$$[\mathbf{H_f}(\mathbf{x})]_{ijk} \equiv H_{ijk} = \frac{\partial^2 f_i}{\partial x_j \partial x_k} \tag{3.b}$$

Los índices $i, j, k = 1, \ldots, n$.

Substituyendo $\mathbf{x}^{k+1}$ por $\mathbf{r}$ y $\mathbf{x}^k$ por $\mathbf{x}$, queda la expresión

$$\mathbf{f}(\mathbf{x}^k) + \{\, [\mathbf{J_f}(\mathbf{x}^k)] + \tfrac{1}{2}\,[\mathbf{H_f}(\mathbf{x}^k)] \,.\, \mathbf{z}^k \,\} \,.\, \mathbf{z}^k = \mathbf{0} \tag{4}$$

donde $\mathbf{z}^k = \mathbf{x}^{k+1} - \mathbf{x}^k = \boldsymbol{\epsilon}^{k+1}$ es el error local, con lo que el Método de Segundo Orden queda implementado como

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{z}^k \tag{5}$$

y el incoveniente se traslada a la forma como obtener $\mathbf{z}^k$ en los métodos que siguen.

2.5.1. Método de Richmond

El *método de Richmond* [Lapidus,1962] pretende resolver la ecuación (4) introduciendo en la parte interna de la ecuación, dentro de las llaves, el estimado ofrecido por el método de Newton-Raphson 2.2.(3)

$$\mathbf{z}^{k,0} = -[\mathbf{J_f}(\mathbf{x}^k)]^{-1}.\mathbf{f}(\mathbf{x}^k) \tag{6}$$

y luego resolver el sistema lineal resultante en la parte externa (ver sección I.2.5.1 para una sola ecuación).

Aquí se propone, como un método más completo, un esquema iterativo secundario interno para cada k de tipo punto fijo con la ecuación (4) de la forma

$$\{\,[\mathbf{J_f}(\mathbf{x}^k)] + \tfrac{1}{2}\,[\mathbf{H_f}(\mathbf{x}^k)]\,.\,\mathbf{z}^{k,s}\,\}\,.\,\mathbf{z}^{k,s+1} = -\mathbf{f}(\mathbf{x}^k) \tag{7}$$

en iteraciones secundarias s. Se escoge como iterado inicial ($s = 0$) de este proceso iterativo secundario interno (para cada k) el valor dado por (6). Luego se resuelve (7) de forma iterativa, tantas veces como sea necesaria hasta que $\Delta\mathbf{z}^k = \mathbf{z}^{k,s+1} - \mathbf{z}^{k,s}$ sea menor en valor absoluto que una tolerancia Δ_{max}, mucho menor que ϵ_{max} por supuesto. Normalmente entre unas 5 a 10 iteraciones secundarias, s_{max}, son suficientes. El método de Richmond es con sólo una iteración secundaria interna (hasta $s = 1$). El método aquí propuesto corrige más veces el valor de $\mathbf{z}$. Viéndolo como un método, predictor con (6) y corrector con (7), cuantas veces se quiera. El equivalente de I.2.5.(4.*b*) para sistemas es (una iteración secundaria, $s_{max} = 1$)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \left\{\, 2\,[\mathbf{J_f}(\mathbf{x}^k)].[\mathbf{J_f}(\mathbf{x}^k)] - [\mathbf{J_f}(\mathbf{x}^k)].[\mathbf{H_f}(\mathbf{x}^k)].[\mathbf{J_f}(\mathbf{x}^k)]^{-1}.\mathbf{f}(\mathbf{x}^k) \,\right\}^{-1}.\, 2\,[\mathbf{J_f}(\mathbf{x}^k)]\,.\,\mathbf{f}(\mathbf{x}^k) \tag{8}$$

Como se observa, el jacobiano (derivada) no se cancela, como si lo hace con una ecuación.

2.5.2. Método del Paraboloide Secante

Un método más completo que el anterior consiste en resolver el problema (4) que es un *paraboloide*

$$\mathbf{F}(\mathbf{z}) = \mathbf{f}(\mathbf{x}^k) + \{\,[\mathbf{J_f}(\mathbf{x}^k)] + \tfrac{1}{2}\,[\mathbf{H_f}(\mathbf{x}^k)]\,.\,\mathbf{z}\,\}\,.\,\mathbf{z} = \mathbf{0} \tag{9}$$

en la incógnita $\mathbf{z}$, con el método de Newton-Raphson con el jacobiano

$$[\mathbf{J_F}(\mathbf{z})] = [\mathbf{J_f}(\mathbf{x})] + [\mathbf{H_f}(\mathbf{x})]\,.\,\mathbf{z} \tag{10}$$

Todos los valores dependientes de las iteraciones k permanecen constante en este proceso iterativo secundario interno en s, resumido de la siguiente manera

$$\mathbf{z}^{k,s+1} = \mathbf{z}^{k,s} + \Delta\mathbf{z}^{k,s} \qquad\qquad [\mathbf{J_F}(\mathbf{z}^{k,s})]\,.\,\Delta\mathbf{z}^{k,s} = -\mathbf{F}(\mathbf{z}^{k,s}) \tag{11}$$

Luego de finalizado este proceso iterativo secundario interno en s (para cada k), bien por satisfacer la tolerancia $|\Delta\mathbf{z}^{k,s}| < \Delta_{max}$ o por número de iteraciones $s_{max} = 3 \sim 6$, el último valor de $\mathbf{z}$ se substituye en (5)

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{z}^k \tag{12}$$

y se continúa con las iteraciones en k.

El método del *plano secante* es una modificación del método de Newton-Raphson (3), donde los elementos de la matriz jacobiana se aproximan con los dos últimos iterados $\mathbf{x}^k$ y $\mathbf{x}^{k-1}$ por

$$[\mathbf{J_f}(\mathbf{x}^k)] = \left[\frac{\partial f_i}{\partial x_j}\right]^k \approx \frac{f_i(x_1^k, x_2^k, x_3^k, \ldots, x_j^k, \ldots, x_n^k) - f_i(x_1^k, x_2^k, x_3^k, \ldots, x_j^{k-1}, \ldots, x_n^k)}{x_j^k - x_j^{k-1}} \tag{13}$$

El mismo argumento se puede seguir para calcular las componentes del tensor hessiano de forma aproximada con los tres últimos iterado $\mathbf{x}^k$, $\mathbf{x}^{k-1}$ y $\mathbf{x}^{k-2}$ por

$$[\mathbf{H_f}(\mathbf{x}^k)] = \left[\frac{\partial^2 f_i(\mathbf{x}^k)}{\partial x_j \partial x_k}\right] \approx \frac{[\frac{\partial f_i}{\partial x_k}]^k - [\frac{\partial f_i}{\partial x_k}]^k_j}{x_j^k - x_j^{k-1}} \qquad (j \neq k) \tag{14.a}$$

$$\approx 2\,\frac{[\frac{\partial f_i}{\partial x_j}]^k - [\frac{\partial f_i}{\partial x_j}]^k_j}{x_j^k - x_j^{k-2}} \qquad (j = k) \tag{14.b}$$

donde

$$\left[\frac{\partial f_i}{\partial x_k}\right]^k_j = \frac{f_i(x_1^k, x_2^k, .., x_j^{k-1}, .., x_k^k, .., x_n^k) - f_i(x_1^k, x_2^k, .., x_j^{k-1}, .., x_k^{k-1}, .., x_n^k)}{x_k^k - x_k^{k-1}} \qquad (j \neq k) \tag{14'.a}$$

$$\left[\frac{\partial f_i}{\partial x_j}\right]^k_j = \frac{f_i(x_1^k, x_2^k, x_3^k, \ldots, x_j^{k-1}, \ldots, x_n^k) - f_i(x_1^k, x_2^k, x_3^k, \ldots, x_j^{k-2}, \ldots, x_n^k)}{x_j^{k-1} - x_j^{k-2}} \qquad (j = k) \tag{14'.b}$$

En estos casos, el método recibe adicionalmente el apelativo de *secante*.

Cuando se usa el procedimiento de la perturbación, entonces $x_j^{k-1} = x_j^k - \Delta x$ y $x_j^{k-2} = x_j^k - 2\,\Delta x$ (esquema atrasado) o $x_j^{k-2} = x_j^k + \Delta x$ (esquema central), donde Δx es una cantidad pequeña (fracción de la tolerancia para el error local ϵ_{max}). No confundir k como super-índice "iteración" y k como sub-índice "componente".

Estos métodos se pueden relajar utilizando tres factores de relajación ω, ω_h y ω_z de la siguiente forma

$$\mathbf{F}(\mathbf{z}) = \mathbf{f}(\mathbf{x}^k) + \{\,[\mathbf{J_f}(\mathbf{x}^k)] + \frac{\omega_h}{2}\,[\mathbf{H_f}(\mathbf{x}^k)]\,.\,\mathbf{z}\,\}\,.\,\mathbf{z} = \mathbf{0} \tag{15.a}$$

en la incógnita $\mathbf{z}$, con el método de Newton-Raphson con el jacobiano

$$[\mathbf{J_F}(\mathbf{z})] = [\mathbf{J_f}(\mathbf{x})] + \omega_h\,[\mathbf{H_f}(\mathbf{x})]\,.\,\mathbf{z} \tag{15.b}$$

relajando con ω_h cunta influencia se quiere del término de segundo orden. Todos los valores dependientes de las iteraciones k permanecen constante en este proceso iterativo secundario interno en s, resumido de la siguiente manera relajando con ω_z

$$\mathbf{z}^{k,s+1} = \mathbf{z}^{k,s} + \omega_z\,\Delta\mathbf{z}^{k,s} \qquad\qquad [\mathbf{J_F}(\mathbf{z}^{k,s})]\,.\,\Delta\mathbf{z}^{k,s} = -\mathbf{F}(\mathbf{z}^{k,s}) \tag{15.c}$$

Luego de finalizado este proceso iterativo secundario interno en s (para cada k), bien por satisfacer la tolerancia $|\Delta\mathbf{z}^{k,s}| < \Delta_{max}$ o por número de iteraciones s_{max}, el último valor de $\mathbf{z}$ se substituye en (5), relajando también con ω

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \omega\,\mathbf{z}^k \tag{15.d}$$

Una vez escogido los factores de relajación mencionados iniciales, el procedimiento puede modular el valor de dichos factores en cada iteración o grupos de iteraciones k (principio-medio-final).

2.5.3. Método de Taylor

Los métodos de Newton-Raphson, Richmond, Paraboloide caen dentro de esta categoría para $n = 2$. Así como se pueden anidar los diferentes términos de un polinomio de grado n en la forma (algoritmo de Horner) [Horner,(1819)]

$$P_1(x) = a_0 + a_1 x$$

$$P_2(x) = a_0 + (a_1 + a_2 x)x$$

$$
\begin{aligned}
P_3(x) &= a_0 + (a_1 + (a_2 + a_3 x)x)x \\
P_4(x) &= a_0 + (a_1 + (a_2 + (a_3 + a_4 x)x)x)x \\
&\cdots \\
P_n(x) &= a_0 + (a_1 + (a_2 + (a_3 + (a_4 + \cdots + (a_{n-1} + a_n x)\cdots \overbrace{x)x)x)x)x}^{n-1}
\end{aligned}
\tag{16}
$$

así también de igual manera se pueden anidar los términos de la expansión en series de Taylor

$$
\begin{aligned}
\mathbf{f}(\mathbf{x}) &= \mathbf{P}_n(\mathbf{z}) + \mathbf{R}_n(\mathbf{z}) \\
&= \mathbf{f}_o + [\,\mathbf{J}_{\mathbf{f}_o}/1! + [\,\mathbf{J}^2_{\mathbf{f}_o}/2! + [\,\mathbf{J}^3_{\mathbf{f}_o}/3! + \cdots + [\,\mathbf{J}^{n-1}_{\mathbf{f}_o}/(n-1)! + \mathbf{J}^n_{\mathbf{f}_o}/n!\,.\,\mathbf{z}\,]\cdots \overbrace{.\,\mathbf{z}\,]\,.\,\mathbf{z}\,]\,.\,\mathbf{z}}^{n-1} + \mathbf{R}_n(\mathbf{z})
\end{aligned}
\tag{17}
$$

donde el desplazamiento es $\mathbf{z} = \mathbf{x} - \mathbf{x}^o$, el jacobiano generalizado es $\mathbf{J}^n_{\mathbf{f}_o} = \mathbf{J}^n_{\mathbf{f}}(\mathbf{x}^o)$ y el término del residual es $\mathbf{R}_n(\mathbf{z}) = O(\|\mathbf{z}^{(n+1)^\otimes}\|)$ (ver apéndice del libro).

Teniendo esto en cuenta, los métodos de Taylor se implementa de la siguiente manera, una vez escogido el grado n del polinomio con el que se desee trabajar ($\mathbf{x}^o = \mathbf{x}^k$ y $\mathbf{x} = \mathbf{x}^{k+1} \approx \mathbf{r}$)

$$
\begin{aligned}
\mathbf{F}(\mathbf{z}) &= \mathbf{P}_n(\mathbf{z}) = \mathbf{0} \\
&= \mathbf{f}(\mathbf{x}^k) + [\,\mathbf{J}_{\mathbf{f}_k}/1! + [\,\mathbf{J}^2_{\mathbf{f}_k}/2! + [\,\mathbf{J}^3_{\mathbf{f}_k}/3! + \cdots + [\,\mathbf{J}^{n-1}_{\mathbf{f}_k}/(n-1)! + \mathbf{J}^n_{\mathbf{f}_k}/n!\,.\,\mathbf{z}\,]\cdots \overbrace{.\,\mathbf{z}\,]\,.\,\mathbf{z}\,]\,.\,\mathbf{z}}^{n-1} = \mathbf{0}
\end{aligned}
\tag{18}
$$

y luego resolver este problema al estilo de Richmond con el iterado inicial estimado o predicho con Newton-Raphson (8) interiormente en la ecuación, y la incógnita más exterior resolverla con un esquema de punto fijo e irla corrigiendo y substituyendo de nuevo interiormente. También se puede utilizar un método de Newton-Raphson para resolver el problema (18) con el jacobiano de $\mathbf{F}(\mathbf{z})$

$$
\mathbf{J}_{\mathbf{F}}(\mathbf{z}) = \mathbf{J}_{\mathbf{f}_k} + [\,\mathbf{J}^2_{\mathbf{f}_k}/1! + [\,\mathbf{J}^3_{\mathbf{f}_k}/2! + \cdots + [\,\mathbf{J}^{n-1}_{\mathbf{f}_k}/(n-2)! + \mathbf{J}^n_{\mathbf{f}_k}/(n-1)!\,.\,\mathbf{z}\,]\cdots \overbrace{.\,\mathbf{z}\,]\,.\,\mathbf{z}\,]\,.\,\mathbf{z}}^{n-2}
\tag{19}
$$

siguiendo un procedimiento similar al método del paraboloide en las ecuaciones (11) y (12). El jacobiano (19) se ha calculado manteniendo constante los coeficientes tensoriales $\mathbf{J}^n_{\mathbf{f}_k} = \mathbf{J}^n_{\mathbf{f}}(\mathbf{x}^k)$.

Por ejemplo, el caso $n = 1$ (Newton-Raphson) ya es conocido y el caso $n = 2$ ya fue explicado antes en (7) (Richmond) y en (9)(11) (paraboloide). Para los casos $n = 3$ y $n = 4$ las expresiones (18) – (19) se simplifican como una extensión de (7) de la forma (al estilo Richmond)

$$
\{\,\mathbf{J}_{\mathbf{f}}(\mathbf{x}^k) + [\,\frac{1}{2}\,\mathbf{H}_{\mathbf{f}}(\mathbf{x}^k) + \frac{1}{6}\,\mathbf{K}_{\mathbf{f}}(\mathbf{x}^k).\mathbf{z}^{k,s}\,].\mathbf{z}^{k,s}\,\}.\mathbf{z}^{k,s+1} = -\mathbf{f}(\mathbf{x}^k)
\tag{20}
$$

$$
\big\{\,\mathbf{J}_{\mathbf{f}}(\mathbf{x}^k) + \{\,\frac{1}{2}\,\mathbf{H}_{\mathbf{f}}(\mathbf{x}^k) + [\,\frac{1}{6}\,\mathbf{K}_{\mathbf{f}}(\mathbf{x}^k) + \frac{1}{24}\,\mathbf{Q}_{\mathbf{f}}(\mathbf{x}^k).\mathbf{z}^{k,s}\,].\mathbf{z}^{k,s}\,\}.\mathbf{z}^{k,s}\,\big\}.\mathbf{z}^{k,s+1} = -\mathbf{f}(\mathbf{x}^k)
\tag{21}
$$

donde el jacobiano de tercer orden es $\mathbf{K}_{\mathbf{f}}(\mathbf{x}^k) = \mathbf{J}^3_{\mathbf{f}}(\mathbf{x}^k) = \{\boldsymbol{\nabla}\{\boldsymbol{\nabla}[\boldsymbol{\nabla}\mathbf{f}(\mathbf{x}^k)]^{\mathrm{t}}\}^{\mathrm{t}}\}^{\mathrm{t}}$ con elementos $K^{\alpha}_{\cdot\beta\gamma\delta} = \frac{\partial^3 f^\alpha}{\partial x^\beta \partial x^\gamma \partial x^\delta}$ y el jacobiano de cuarto orden es $\mathbf{Q}_{\mathbf{f}}(\mathbf{x}^k) = \mathbf{J}^4_{\mathbf{f}}(\mathbf{x}^k) = \{\boldsymbol{\nabla}\{\boldsymbol{\nabla}\{\boldsymbol{\nabla}[\boldsymbol{\nabla}\mathbf{f}(\mathbf{x}^k)]^{\mathrm{t}}\}^{\mathrm{t}}\}^{\mathrm{t}}\}^{\mathrm{t}}$ with elements $Q^{\alpha}_{\cdot\beta\gamma\delta\epsilon} = \frac{\partial^4 f^\alpha}{\partial x^\beta \partial x^\gamma \partial x^\delta \partial x^\epsilon}$. Aplicando Newton-Raphson (15.c), estas ecuaciones se resuelven como las ecuaciones homogéneas en $\mathbf{z}$

$$
\mathbf{F}(\mathbf{z}) = \mathbf{f}(\mathbf{x}^k) + \{\,\mathbf{J}_{\mathbf{f}}(\mathbf{x}^k) + [\,\frac{1}{2}\,\mathbf{H}_{\mathbf{f}}(\mathbf{x}^k) + \frac{1}{6}\,\mathbf{K}_{\mathbf{f}}(\mathbf{x}^k).\mathbf{z}\,].\mathbf{z}\,\}.\mathbf{z} = \mathbf{0}
\tag{22}
$$

$$
\mathbf{F}(\mathbf{z}) = \mathbf{f}(\mathbf{x}^k) + \big\{\,\mathbf{J}_{\mathbf{f}}(\mathbf{x}^k) + \{\,\frac{1}{2}\,\mathbf{H}_{\mathbf{f}}(\mathbf{x}^k) + [\,\frac{1}{6}\,\mathbf{K}_{\mathbf{f}}(\mathbf{x}^k) + \frac{1}{24}\,\mathbf{Q}_{\mathbf{f}}(\mathbf{x}^k).\mathbf{z}\,].\mathbf{z}\,\}.\mathbf{z}\,\big\}.\mathbf{z} = \mathbf{0}
\tag{23}
$$

y sus correspondientes jacobianos $\mathbf{J_F}(\mathbf{z})$

$$\mathbf{J_F}(\mathbf{z}) = \mathbf{J_f}(\mathbf{x}^k) + [\,\mathbf{H_f}(\mathbf{x}^k) + \frac{1}{2}\,\mathbf{K_f}(\mathbf{x}^k).\mathbf{z}\,].\mathbf{z} \tag{24}$$

$$\mathbf{J_F}(\mathbf{z}) = \mathbf{J_f}(\mathbf{x}^k) + \{\,\mathbf{H_f}(\mathbf{x}^k) + [\,\frac{1}{2}\,\mathbf{K_f}(\mathbf{x}^k) + \frac{1}{6}\,\mathbf{Q_f}(\mathbf{x}^k).\mathbf{z}\,].\mathbf{z}\,\}.\mathbf{z} \tag{25}$$

Los demás detalles son similares a los mencionados. Una forma alternativa, es resolver el problema de manera escalonada para cada iteración k

$$\mathbf{P}_s(\mathbf{z}^{k,s-1}) = \mathbf{0} \qquad\qquad \forall s = 1, 2, 3, \ldots, n \tag{26}$$

utilizando una grado diferente s para cada solución $s-1$ obtenida con el grado anterior como iterado inicial en la iteración interna en s, aumentando el grado del polinomio multivariable en cada escalón, comenzando con la solución $\mathbf{z}^{k,o}$ de Newton-Raphson (6) en $\mathbf{P}_1(\mathbf{z}) = \mathbf{0}$. Luego se sigue con la solución de $\mathbf{P}_2(\mathbf{z}) = \mathbf{0}$ del paraboloide (9)-(11), y así sucesivamente (una o varias iteraciones internas en cada escalón o grado).

2.6. CONVERGENCIA

2.6.1. Criterios

El método de Newton-Raphson al ser un método de tipo punto fijo la función $\mathbf{g}(\mathbf{x})$ que lo define es

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} - \omega\,[\mathbf{J_f}(\mathbf{x})]^{-1}.\,\mathbf{f}(\mathbf{x}) \qquad\qquad [\mathbf{J_f}(\mathbf{x})].(\mathbf{x} - \mathbf{g}(\mathbf{x})) = \omega\,\mathbf{f}(\mathbf{x}) \tag{1}$$

Extrayendo el gradiente de la segunda expresión, y usando la primera, da (ver [Granados,2022;A.2.2.(16.g)])

$$[\mathbf{J_f}(\mathbf{x})]\,.\,\big(\,[\mathbf{I}] - [\mathbf{J_g}(\mathbf{x})]\,\big) + \omega\,\big\{\,[\mathbf{H_f}(\mathbf{x})]\,.\,[\mathbf{J_f}(\mathbf{x})]^{-1}.\,\mathbf{f}(\mathbf{x})\,\big\}^{\mathrm{t}} = \omega\,[\mathbf{J_f}(\mathbf{x})]^{\mathrm{t}} \tag{2}$$

$$[\mathbf{J_f}(\mathbf{x})].[\mathbf{J_g}(\mathbf{x})] = [\mathbf{J_f}(\mathbf{x})] - \omega\,[\mathbf{J_f}(\mathbf{x})]^{\mathrm{t}} + \omega\,\big\{\,[\mathbf{H_f}(\mathbf{x})]\,.\,[\mathbf{J_f}(\mathbf{x})]^{-1}.\,\mathbf{f}(\mathbf{x})\,\big\}^{\mathrm{t}} \tag{3}$$

Finalmente se obtiene el gradiente de $\mathbf{g}$ (la última transposición puede substituirse por $^{\mathrm{t}^{-}}$ sobre $[\mathbf{H_f}(\mathbf{x})]$)

$$[\mathbf{J_g}(\mathbf{x})] = [\,\mathbf{I}\,] - \omega\,[\mathbf{J_f}(\mathbf{x})]^{-1}.\,[\mathbf{J_f}(\mathbf{x})]^{\mathrm{t}} + \omega\,[\mathbf{J_f}(\mathbf{x})]^{-1}.\,\big\{\,[\mathbf{H_f}(\mathbf{x})]\,.\,[\mathbf{J_f}(\mathbf{x})]^{-1}.\,\mathbf{f}(\mathbf{x})\,\big\}^{\mathrm{t}} \tag{4}$$

Este gradiente para un punto intermedio $\boldsymbol{\zeta}$ debe satisfacer

$$\|\mathbf{J_g}(\boldsymbol{\zeta})\| < 1 \qquad\qquad \boldsymbol{\zeta} \in \mathbb{B}(\mathbf{r}, \|\mathbf{e}^k\|) \tag{5}$$

para que el método de Newton-Raphson converga. Las diferencias con una sola ecuación pueden verse comparando con I.2.4.(14). Para una sola ecuación el primer y el segundo términos se calcelan ($\omega = 1$) y el tercero da el resultado esperado en I.2.4.(4.b).

2.6.2. Tipos

Haciendo la expansión en series de Taylor de la función $\mathbf{g}(\mathbf{x})$ alrededor de $\mathbf{r}$ y evaluada en $\mathbf{x}^k$, se obtiene

$$\mathbf{g}(\mathbf{x}^k) = \mathbf{g}(\mathbf{r}) + [\mathbf{J_g}(\mathbf{r})]\,.\,(\mathbf{x}^k - \mathbf{r}) + \tfrac{1}{2}\,[\mathbf{H_g}(\boldsymbol{\zeta})]\;:\;(\mathbf{x}^k - \mathbf{r})(\mathbf{x}^k - \mathbf{r}) \qquad\qquad \boldsymbol{\zeta} \in \mathbb{B}(\mathbf{r}, \|\mathbf{e}^k\|) \tag{6}$$

Según esto, el método de Newton-Raphson debería ser siempre de convergencia lineal porque $[\mathbf{J_g}(\mathbf{r})] = [\,\mathbf{I}\,] - [\mathbf{J_f}(\mathbf{r})]^{-1}.\,[\mathbf{J_f}(\mathbf{r})]^{\mathrm{t}} \neq \mathbf{0}$ siempre ($\mathbf{r}$ de multiplicidad $\omega = 1$), usando (3), y donde el último término se anula por ser $\mathbf{f}(\mathbf{r}) = \mathbf{0}$. Pero en lugar de seguir el desarrollo deductivo del tipo I.2.4.(6) – (7), seguiremos el del tipo I.2.4.(8) – (9), más apropiado para el método de Taylor, que los engloba a casi todos (Newton-Raphson, Richmond, segundo orden, tercer orden, etc.).

El métodos de Taylor de grado $p-1$, es aproximadamente de convergencia computacional de orden p, dependiendo si las derivadas son analíticas o numéricas (perturbación o últimos iterados) y si la solución de $\mathbf{z}$ es convergente exacta o no (iteraciones secundarias internas $s_{max} \to \infty$).

El orden de convergencia computacional p y el índice de eficiencia η del método se definen como

$$p = \lim_{k \to k_{max}} \frac{\log(\|\boldsymbol{\epsilon}^{k+1}\|/\|\boldsymbol{\epsilon}^{k}\|)}{\log(\|\boldsymbol{\epsilon}^{k}\|/\|\boldsymbol{\epsilon}^{k-1}\|)} \qquad \eta = p^{1/m} \qquad m = \alpha n + \beta n^2 + \gamma n^3 + \cdots \tag{7}$$

$$\text{Error Local } \boldsymbol{\epsilon}^k = \mathbf{x}^k - \mathbf{x}^{k-1} \qquad \begin{matrix} \text{Orden de convergencia } p \\ \text{Eficiencia } \eta \text{ para } n \times n \text{ Sistema} \end{matrix} \qquad m \text{ Evaluaciones: } \begin{matrix} \alpha - \text{funcionales}, \\ \beta - \text{Jacobianos}, \\ \gamma - \text{Hessianos, etc.} \end{matrix}$$

Como ejemplo haremos el desarrollo para el Método de Segundo Orden 2.5.(1)-(5). Haciendo la expansión en series de Taylor alrededor de $\mathbf{x}^k$ y evaluada en $\mathbf{r}$, hasta del tercer orden (incluyendo el término residual evaluado en un punto intermedio $\boldsymbol{\xi}^k$, porque la serie se ha truncado. Ver Apéndice)

$$\mathbf{0} = \mathbf{f}(\mathbf{r}) = \mathbf{f}(\mathbf{x}^k) + \{\, [\mathbf{J_f}(\mathbf{x}^k)] + \tfrac{1}{2}\, [\mathbf{H_f}(\mathbf{x}^k)] \,.\, \mathbf{z}^k \,\} \,.\, \mathbf{z}^k - \tfrac{1}{6} \mathbf{K_f}(\boldsymbol{\xi}^k) \vdots \mathbf{e}^{k^{3\otimes}} \tag{8}$$

($\mathbf{z}^k = \mathbf{r} - \mathbf{x}^k = -\mathbf{e}^k$, lo que justifica el signo negativo del término residual impar) donde $\boldsymbol{\xi}^k \in \mathbb{B}(\mathbf{x}^k, \|\mathbf{e}^k\|)$ es la n-bola de centro $\mathbf{x}^k$ y de radio $\|\mathbf{e}^k\|$. Multiplicando toda la ecuación anterior por $\{[\mathbf{J_f}(\mathbf{x}^k)] + \frac{1}{2}\, [\mathbf{H_f}(\mathbf{x}^k)] \,.\, \mathbf{z}^k \,\}^{-1}$, se obtiene el resultado

$$\mathbf{z}^k = -\mathbf{e}^k - \tfrac{1}{6} \{\, [\mathbf{J_f}(\mathbf{x}^k)] + \tfrac{1}{2}\, [\mathbf{H_f}(\mathbf{x}^k)] \,.\, \mathbf{z}^k \,\}^{-1} .\, \mathbf{K_f}(\boldsymbol{\xi}^k) \vdots \mathbf{e}^{k^{3\otimes}} \tag{9}$$

donde del lado izquierdo queda la fórmula algorítmica del método de segundo orden

$$\mathbf{x}^{k+1} - \mathbf{x}^k = -\{\, [\mathbf{J_f}(\mathbf{x}^k)] + \tfrac{1}{2}\, [\mathbf{H_f}(\mathbf{x}^k)] \,.\, \mathbf{z}^k \,\}^{-1} .\, \mathbf{f}(\mathbf{x}^k) = \mathbf{z}^k \tag{10}$$

donde $\mathbf{z}^k = \mathbf{x}^{k+1} - \mathbf{x}^k$ (ahora diferente, porque se ha eliminado el término residual en (10) y $\mathbf{r}$ se convierte en $\mathbf{x}^{k+1}$, la raíz numéricamente aproximada en la iteración siguiente) es la solución exacta de la ecuación anterior (la última igualdad), resuelta, por ejemplo, con el método de punto fijo 2.5.(7) o el método de Newton-Raphson 2.5.(11).

Teniendo en cuenta que $\mathbf{e}^{k+1} = \mathbf{z}^k + \mathbf{e}^k$, entonces de (9) se obtiene

$$\mathbf{e}^{k+1} \leq -\tfrac{1}{6}\, \mathbf{A} . \mathbf{K_f}(\boldsymbol{\xi}^k) \vdots \mathbf{e}^{k^{3\otimes}} \tag{11}$$

y (10) resulta ser un método de convergencia cúbica ($p = 3$), siendo

$$\mathbf{A} = \max_{\mathbf{x} \in \mathbb{B}} \{\, [\mathbf{J_f}(\mathbf{x})] + \tfrac{1}{2}\, [\mathbf{H_f}(\mathbf{x})] \,.\, \mathbf{z} \,\}^{-1} = \{\, [\mathbf{J}_{min}] + \tfrac{1}{2}\, [\mathbf{H}_{min}] \,.\, \mathbf{e}_{max} \,\}^{-1} \tag{12}$$

($\mathbb{B} \equiv \mathbb{B}(\mathbf{x}^k, \|\mathbf{e}^k\|)$) donde $\mathbf{J}_{min} = \inf_{\mathbf{x} \in \mathbb{B}} \mathbf{J_f}(\mathbf{x})$ y $\mathbf{H}_{min} = \inf_{\mathbf{x} \in \mathbb{B}} \mathbf{H_f}(\mathbf{x})$ son los ínfimos que tienen la misma estructura tensorial de sus predecesores (esto se logra calculando los ínfimos para cada componente en valor absoluto) y $\mathbf{e}_{max}$ es la tolerancia en el error global ($\|\mathbf{e}_{max}\| = \sqrt{n\, e_{max}^2}$, sistema $n \times n$ de n ecuaciones y n incógnitas, $e_{max} \gtrsim \epsilon_{max}$).

2.7. ESTABILIDAD

Los métodos iterativos, todos generan un mapa fractal del procedimiento, si cada punto del espacio visto como iterado inicial, se colorea con un color distinto dependiendo de la raíz a la cual converge. De acuerdo a Mandelbrot [1983], quien acuño el término "*Fractal*" un subconjunto del espacio A es llamado fractal, dependiendo si su dimensión de Hausdorff-Besicovitch $D_H(A)$ es un número fraccionado y no un entero. Intuitivamente D_H mide el crecimiento del número de esferas de diámetro ε necesarias para cubrir el dominio analizado A, cuando $\varepsilon \to 0$. Más precisamente, si el dominio A es un subconjunto de $\mathbb{R}^n$, sea $N(\varepsilon)$ el mínimo de bolas n-dimensionales de diámetro ε necesario para cubrir el dominio.

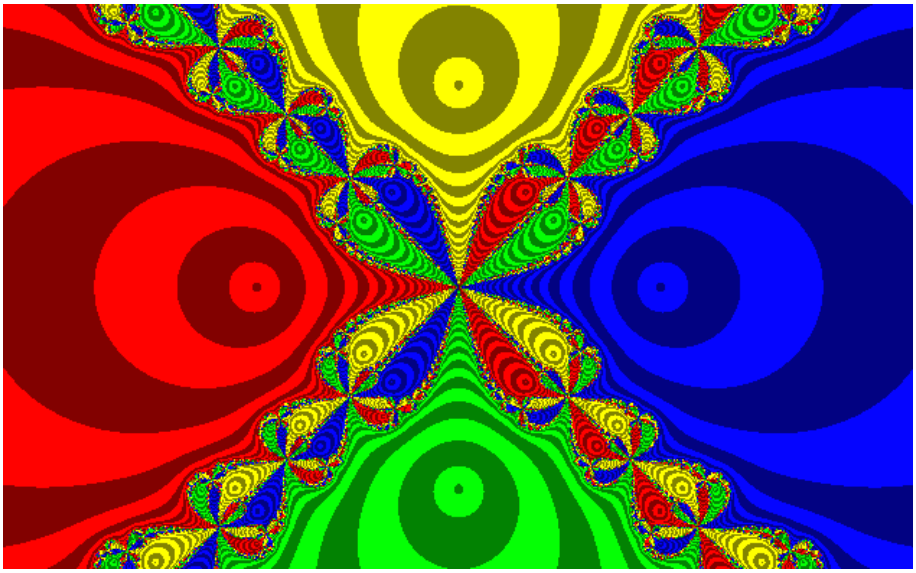


Fig. 1. Método de Newton-Raphson ($K_{min} = 3$, $K_{med} = 9.6$, $K_{max} = 87$, $D_H = 1.976620$).

Luego, si $N(\varepsilon)$ crece exponencialmente como ε^{-D_H} cuando $\varepsilon \to 0$, se dice el dominio A tiene una dimensión de Hausdorff D_H [Peitgen & Richter,1986]. No es dicícil mostrar que la dimensión de Hausdorff puede ser obtenida por

$$D_H = \lim_{\varepsilon \to 0} \frac{\log[N(\varepsilon)]}{\log(k/\varepsilon)} \tag{1}$$

donde $K = k^{D_H}$ es la constante de proporcionalidad cuando

$$N \to K\,\varepsilon^{-D_H} \qquad\qquad \varepsilon \to 0 \tag{2}$$

Por ejemplo, un rectángulo monocolor con lados a y b puede ser cubierto con círculos de diámetro $\varepsilon = a/n = b/m$, ordenados en n filas por m columnas, y la dimensión de Hausdorff se obtiene como

$$D_H = \lim_{n,m \to \infty} \frac{\log(nm)}{\log[k'\sqrt{ab}/\sqrt{(a/n)(b/m)}]} = 2 \tag{1'}$$

donde $k' = 1$ es la constante de proporcionalidad ($k = k'\sqrt{ab}$). Para un cuadrado el resultado es trivial y es exactamente el mismo, sin importar el número de círculos. Para una figura fractal el resultado no es entero como se mencionó, pero el procedimiento usado puede ser el mismo.

De acuerdo a esto, la dimensión de Hausdorff representa una medida de cuan fractal es una figura inmersa en $\mathbb{R}^n$. Consecuentemente, mientras más fractal sea la figura menos estable o caótico es el sistema

dinámico, y por lo tanto menos estable es el proceso iterativo representado por su mapa fractal [Devaney,1987]. Mas cercano al entero siguiente (dimensión topológica, que en el plano es 2). Usaremos los nombre de *región fractal* para las zonas donde los colores están más dispersos con una forma intrincada y alternada, y *cuenca de convergencia* donde los colores son más uniformes alrededor de un punto de convergencia final. Las distintas zonas tiene un sombreado en forma de cebra que indican las iteraciones. Al pasar de un sombreado (impar) a no-sombreado (par) indica una sola iteración.

Fig. 2. Método de La Secante ($K_{min} = 3$, $K_{med} = 12$, $K_{max} = 889$, $D_H = 1.943245$) .

El problema que vamos a usar como ejemplo es el de $f(z) = z^4 - 1 = 0$ en el plano complejo $z = z + iy$. Consiste en hallar las cuatro raíces del problema que son $\mathbf{r} = +1,\ +i,\ -1,\ -i$ coloreados los puntos azul, amarillo, rojo y verde. El problema en el plano $\mathbb{R}^2$ se representa como el sistemas de 2 ecuaciones no-lineales con 2 incógnitas ($\{\mathbf{f}(\mathbf{x})\} = \{f_x(\mathbf{x}), f_y(\mathbf{x})\} = \{\mathbf{0}\}$, $\{\mathbf{x}\} = \{x, y\}$)

$$f(z) = z^4 - 1 = 0 \qquad \begin{aligned} f_x(x,y) &= (x^2 - y^2)^2 - 4\,x^2y^2 - 1 = 0 \\ f_y(x,y) &= 4\,xy\,(x^2 - y^2) = 0 \end{aligned} \tag{3}$$

con la matriz jacobiana $[J_{ij}] = [\partial f_i / \partial x_j]$

$$[\mathbf{J_f}(\mathbf{x})] = \begin{bmatrix} 4x(x^2 - 3y^2) & -4y(3x^2 - y^2) \\ 4y(3x^2 - y^2) & 4x(x^2 - 3y^2) \end{bmatrix} \tag{4.a}$$

y el tensor Hessiano $[H_{ijk}] = [\partial^2 f_i / \partial x_j \partial x_k]$

$$[\mathbf{H_f}(\mathbf{x})] = \begin{bmatrix} 12(x^2 - y^2) & -24xy \\ 24xy & 12(x^2 - y^2) \end{bmatrix} \begin{bmatrix} -24xy & -12(x^2 - y^2) \\ 12(x^2 - y^2) & -24xy \end{bmatrix} \tag{4.b}$$

en el último caso las matrices contiguas contienen las derivadas de los jacobianos $[\partial \mathbf{J_f} / \partial x]$ y $[\partial \mathbf{J_f} / \partial y]$.

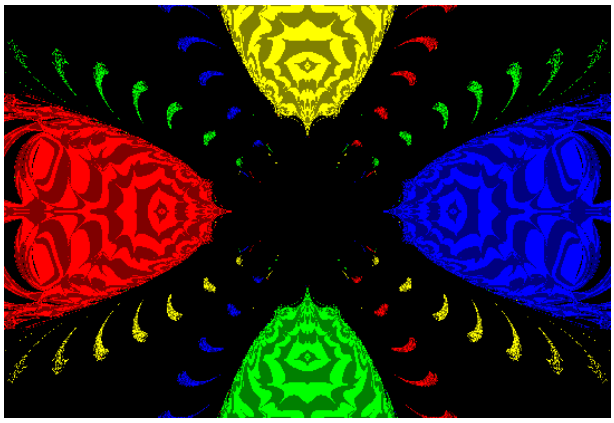


Fig. 3.a. Método de Broyden, con 1 iteration Newton-Raphson al inicio ($K_{min} = 1$, $K_{med} = 11.8$, $K_{max} = 698626$, $D_H = 1.811758$).

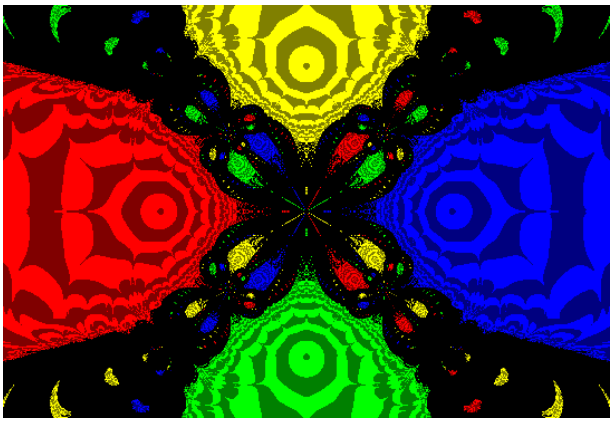


Fig. 3.b. Método de Broyden, con 3 iteraciones previas de Newton-Raphson ($K_{min} = 1$, $K_{med} = 8.2$, $K_{max} = 100000$, $D_H = 1.909757$).

La dimensión de Hausdorff aquí calculada es sólo una aproximación debido al número limitado de píxeles utilizados, y se obtiene numéricamente con

$$D_H \approx \frac{\log[\, N(COLOR)\,]}{\log[\, \sqrt{N_T(COLOR)}\,]} \qquad (1'')$$

donde $N(COLOR)$ es el número de puntos de color "$COLOR$" que está totalmente rodeado por (8) puntos con el mismo color, y $N_T(COLOR)$ es el total de puntos de color "$COLOR$". Nótese que la dimensión fractal de Hausdorff promedio par todos los colores, llamada simplemente *dimensión fractal*, es aproximadamente 2, pero un poco menor. Mientras la región fractal ocupe mayor área de toda la figura, entonces la dimensión fractal se aparta del valor 2 (por debajo). En este caso, el sistema determinado por el método es menos estable.

La figura 1 muestra el método de Newton-Raphson de la sección 2.2.1 con jacobiano calculado analíticamente (fórmulas (4.a)). Muestra las cuencas de convergencia bien definidas y las zonas fractales denotan un proceso iterativo caótico.

La figura 2 muestra el método de la secante, método de la sección 2.2.1 con jacobiano calculado de forma aproximada con las dos últimas iteraciones (fórmula 2.5.(13). Las regiones fractales se difuminan (polvo fractal) y las cuencas de convergencia presenta iteraciones (zonas sombreadas) en forma de cúspides. Levemente peor que el método de Newton-Raphson.

La figura 3 muestra el método de Broyden del segundo tipo de la sección 2.3.1 (fórmulas 2.3.(6)-(7)) con 1 y 3 iteraciones previas iniciales con el método de Newton-Raphson. Presenta zonas inestables de color negro donde no se llega a ninguna raíz. El método de Newton-Raphson previo estabiliza un poco el método. Las cuencas de convergencia son reducidas. Es el peor de todos los métodos.

La figura 4 muestra el método de segundo orden de la sección 2.5.2, paraboloide con jacobiano y hessiano calculados numéricamente con perturbación (fórmulas 2.5.(13) – (14)). Las cuencas de convergencia aumentan de tamaño y se reducen las regiones fractales en la medida que se incrementa el número de las iteraciones internas. Hasta ahora es el mejor de todos los métodos.

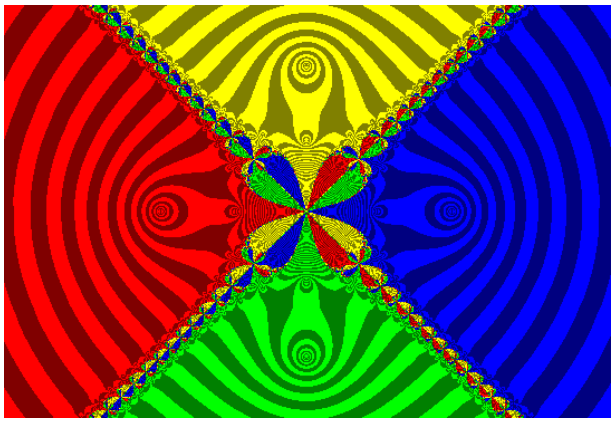


Fig. 4.a. Método de Segundo orden con 2 iteración interna ($K_{min} = 25$, $K_{med} = 48.6$, $K_{max} = 300$, $D_H = 1.988387$).

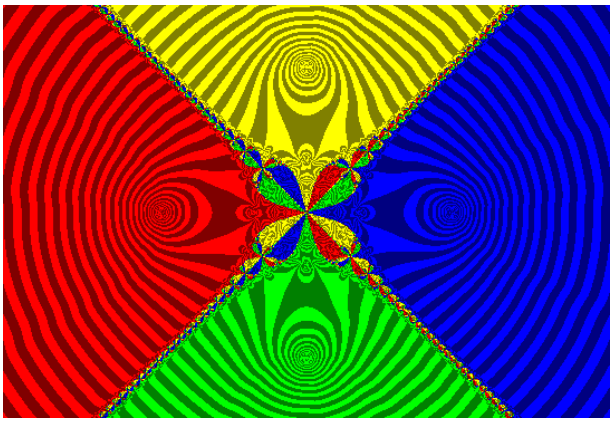


Fig. 4.b. Método de Segundo Orden con 3 iteracciones internas ($K_{min} = 47$, $K_{med} = 94.3$, $K_{max} = 469$, $D_H = 1.990834$).

La figura 5 muestra el método del paraboloide secante (2 y 3 iteraciones internas) de la sección 2.5.2, con jacobiano y hessiano calculados de forma aproximada con las tres últimas iteraciones (fórmulas 2.5.(13)-(14)). Igual que el caso anterior, pero las regiones fractales están difuminadas y las cuencas de convergencia presentan la características de cúspides de los métodos secantes. Es un método intermedio entre el método de Newton-Raphson y levemente peor que el método de segundo orden analítico.

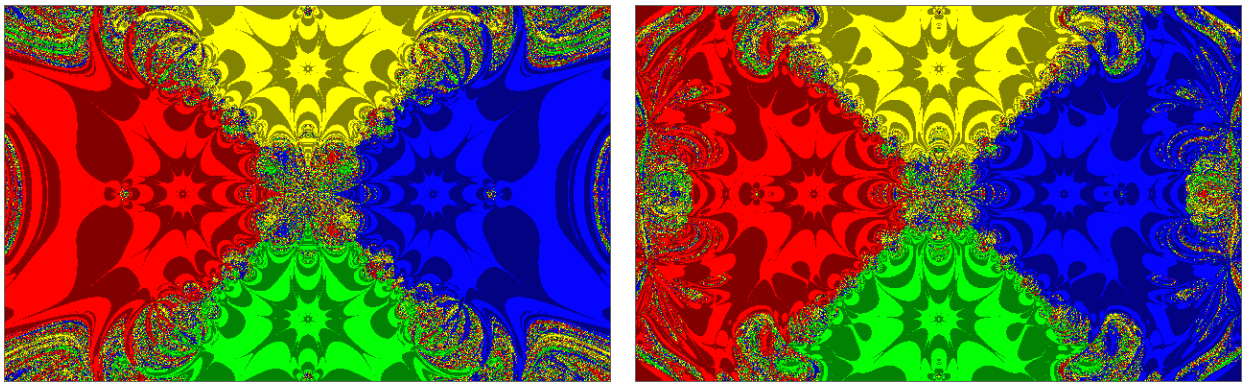


Fig. 5. Método del Paraboloide Secante. 2 y 3 iteraciones internas ($K_{med} = 11.7$, 10.3, $D_H = 1.944819$, 1.946966).

Para el mismo problema (3), el jacobiano de tercer orden es $\mathbf{K_f}(\mathbf{x}^k) = \mathbf{J}^3_{\mathbf{f}}(\mathbf{x}^k) = \{\boldsymbol{\nabla}\{\boldsymbol{\nabla}[\boldsymbol{\nabla}\mathbf{f}(\mathbf{x}^k)]^{\mathrm{t}}\}^{\mathrm{t}}\}^{\mathrm{t}}$ con elementos $K^{\alpha}_{\cdot\beta\gamma\delta} = \frac{\partial^3 f^{\alpha}}{\partial x^{\beta}\partial x^{\gamma}\partial x^{\delta}}$ es

$$[\mathbf{K_f}(\mathbf{x})] = \begin{bmatrix} \begin{bmatrix} 24x & -24y \\ 24y & 24x \end{bmatrix} & \begin{bmatrix} -24y & -24x \\ 24x & -24y \end{bmatrix} \\ \begin{bmatrix} -24y & -24x \\ 24x & -24y \end{bmatrix} & \begin{bmatrix} -24x & 24y \\ -24y & -24x \end{bmatrix} \end{bmatrix} \tag{4.c}$$

En este caso, las matrices por líneas contienen las derivadas del Hessiano (4.*b*), $[\partial \mathbf{H_f}/\partial x]$ and $[\partial \mathbf{H_f}/\partial y]$. Justo para este jacobiano la transposición se evitó o fue innecesaria para las matrices bloques (teorema de Clairaut).

El jacobiano de cuarto orden $\mathbf{Q_f}(\mathbf{x}^k) = \mathbf{J}^4_{\mathbf{f}}(\mathbf{x}^k) = \{\boldsymbol{\nabla}\{\boldsymbol{\nabla}\{\boldsymbol{\nabla}[\boldsymbol{\nabla}\mathbf{f}(\mathbf{x}^k)]^{\mathrm{t}}\}^{\mathrm{t}}\}^{\mathrm{t}}\}^{\mathrm{t}}$ con elementos $Q^{\alpha}_{\cdot\beta\gamma\delta\epsilon} = \frac{\partial^4 f^{\alpha}}{\partial x^{\beta}\partial x^{\gamma}\partial x^{\delta}\partial x^{\epsilon}}$ es

$$[\mathbf{Q_f}(\mathbf{x})] = \begin{bmatrix} \begin{bmatrix} 24 & 0 \\ 0 & 24 \end{bmatrix} & \begin{bmatrix} 0 & -24 \\ 24 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & -24 \\ 24 & 0 \end{bmatrix} & \begin{bmatrix} -24 & 0 \\ 0 & -24 \end{bmatrix} \end{bmatrix} \quad \begin{bmatrix} \begin{bmatrix} 0 & -24 \\ 24 & 0 \end{bmatrix} & \begin{bmatrix} -24 & 0 \\ 0 & -24 \end{bmatrix} \\ \begin{bmatrix} -24 & 0 \\ 0 & -24 \end{bmatrix} & \begin{bmatrix} 0 & 24 \\ -24 & 0 \end{bmatrix} \end{bmatrix} \tag{4.d}$$

En este caso, cada matriz grande tiene las derivadas $[\partial \mathbf{K_f}/\partial x]$ y $[\partial \mathbf{K_f}/\partial y]$, respectivamente. Los jacobianos de orden superiores son nulos.

Aunque, el esfuerzo computacional son casi lo mismo como en 2.5.(20)−(21), las fórmulas 2.5.(22)−(25), con el algoritmo $(13.a, b)$, son más rápidas a una relación de 1 versus 5 iteraciones internas par el mismo avance.

La figura 6.a,b muestra los mapas fractales del algoritmo 2.5.(15) donde las derivadas son calculadas analíticamente con las fórmulas $(4.a, b)$, mientras que la figura 4.a,b las derivadas son calculadas numéricamente con 2.5.(13) − (14). Las diferencias gráficas son drásticas porque el cálculo numérico de las segundas derivadas en la fig. 4 estuvo erróneamente desplazado, lo cual fue corregido aquí en las fig. 6.a,b que correspondientemente substituyen la fig. 4.a,b (pequeños cambios en las derivadas producen grandes cambios en el mapa fractal, e.g. método secante). El mapa tiene una región de cuenca con convergencia estable hacia las raíces con un número pequeño de iteraciones (cambio en la intensidad del color da una idea de las iteraciones: par-claro impar-sombreado). Una región negra ($k > k_{max} = 1000$ iteraciones principales) de no convergencia está situada alrededor y en el centro (fig. 6.b).

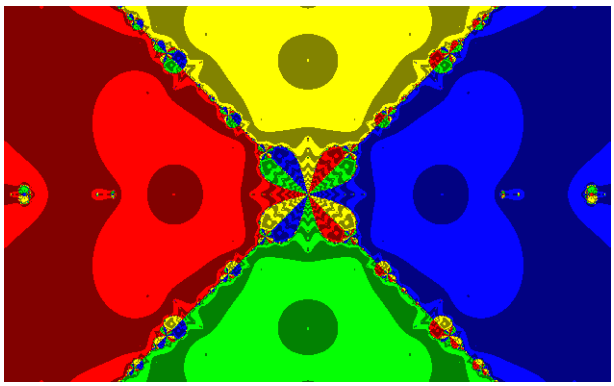


Fig. 6.a. Método de Segundo orden (derivadas analíticas) con 2 iteración interna ($K_{min} = 2$, $K_{med} = 4.8$, $K_{max} = 35$, $D_H = 1.993255$).

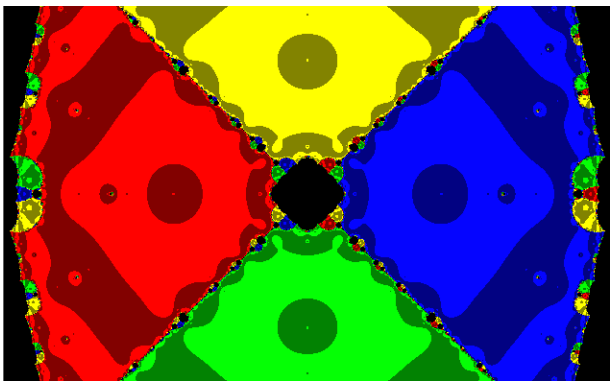


Fig. 6.b. Método de Segundo Orden (derivadas analíticas) con 3 iteraciones internas ($K_{min} = 2$, $K_{med} = 93.3$, $K_{max} = 1000$, $D_H = 1.998690$).

La región fractal con un comportamiento caótico se reduce a las diagonales principales y la frontera, y algunas islas pequeñas dentro de las cuencas de convergencia. El incremento del número de las iteraciones internas elimina las zonas negras (excepto la vecindad cercana del punto central) y envía la frontera fractal lejos y reduce la región fractal en las digonales como puede ser visto en la figura 7, con un máximo de estabilidad y de la dimensión fractal de Hausdorff. Esta correlación puede ser vista en [Granados,1995] donde un análisis (dimensión fractal vs. estabilidad) para los métodos de primer (Newton-Raphson) y segundo orden (Richmond [1944]-Taylor [Granados,2015]) son estudiados.

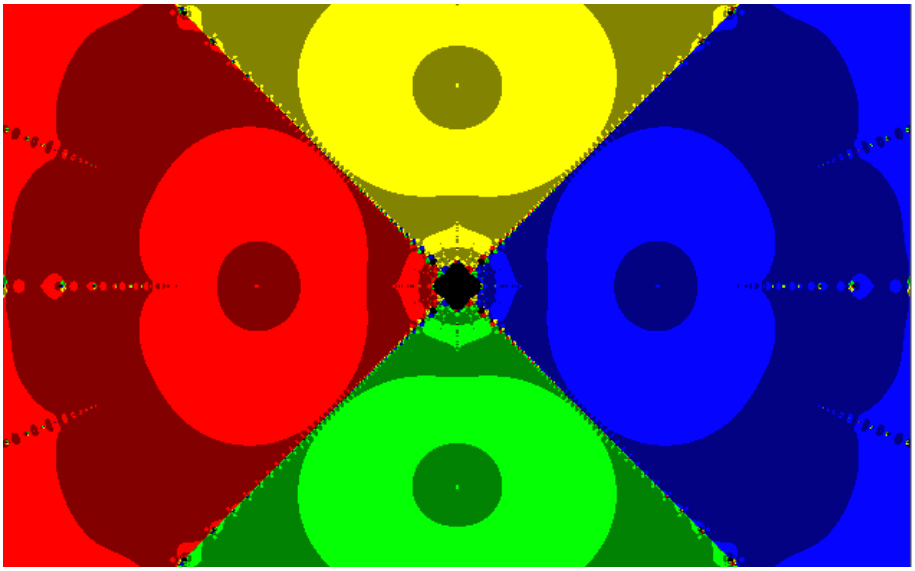


Fig. 7. Método de Segundo Orden (derivadas analíticas) con 6 iteraciones internas ($K_{min} = 2$, $K_{med} = 40.5$, $K_{max} = 1000$, $D_H = 1.996344$).

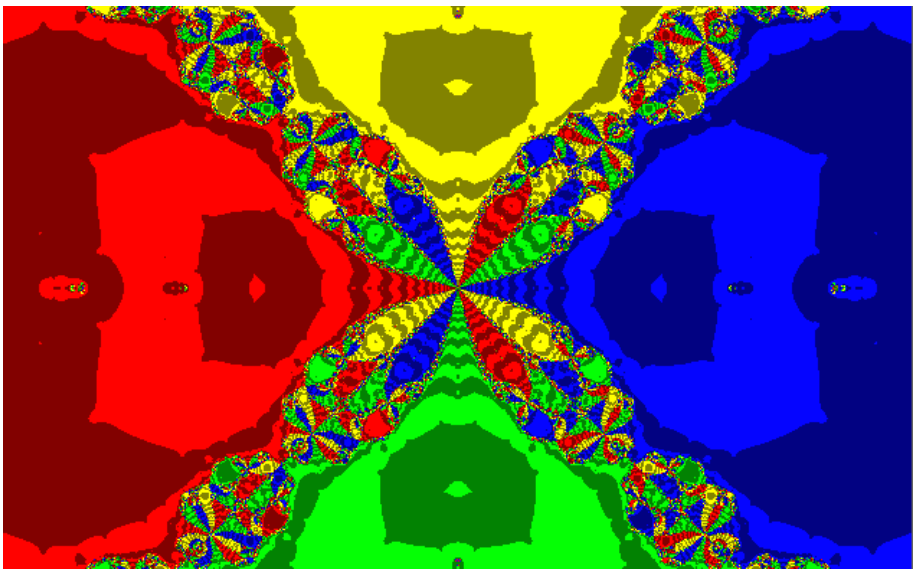


Fig. 8. Método de Tercer Orden (derivadas analíticas) con 3 iteraciones internas ($K_{min} = 2$, $K_{med} = 5.8$, $K_{max} = 57$, $D_H = 1.975170$).

Las figuras 8 y 9 ofrecen los mapas fractales de los métodos de tercer orden 2.5.(22) y 2.5.(24) en el cual se percibe que el incremento del orden del método más del segundo no es buena idea: la región de no convergencia (negra) no desaparece del todo y la región fractal no reduce su tamaño (comparar con la fig.1), incluyendo con 10 o 15 iteraciones internas (fig. 9).

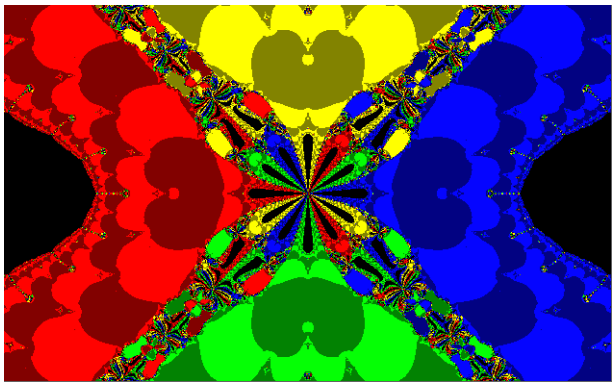


Fig. 9.a. Método de Tercer Orden (derivadas analíticas) con 10 iteraciones internas ($K_{min} = 2$, $K_{med} = 144.9$, $K_{max} = 1000$, $D_H = 1.948307$).

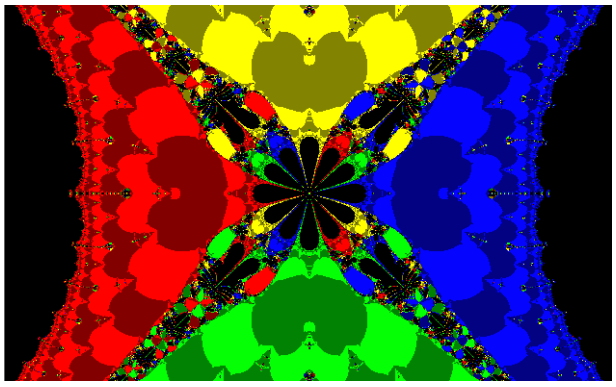


Fig. 9.b Método de Tercer Orden (derivadas analíticas) con 15 iteraciones internas ($K_{min} = 2$, $K_{med} = 287$, $K_{max} = 1000$, $D_H = 1.919773$).

Además, mientras más iteraciones internas, más grandes son las regiones negras, pero la región fractal continúa siendo del mismo tamaño. La únice ventaja usando alto grado es el bajo número de iteraciones principales, porque el trabajo duro ya está hecho por las iteraciones internas. Sin embargo, la estabilidad es peor para mayores iteraciones internas como se podrá observar comparando las dimensiones fractales de las figuras 8 y 9. El mapa fractal en la figura 10 confirma lo que previamente se discutió para los métodos de cuarto orden 2.5.(23) y 2.5.(25) (con 2.5.(15.c, d)). En conclusión, la figura 7 muestra el paradigma del mapa fractal del mejor método, aunque la región fractal es la más pequeña de todas, apenas perceptible. Las figuras 8 y 10 son casi las mismas, lo que significa que un incremento del orden de 3 a 4 no produce mejores resultados, al menos con el problema (3) con el que estamos tratando.

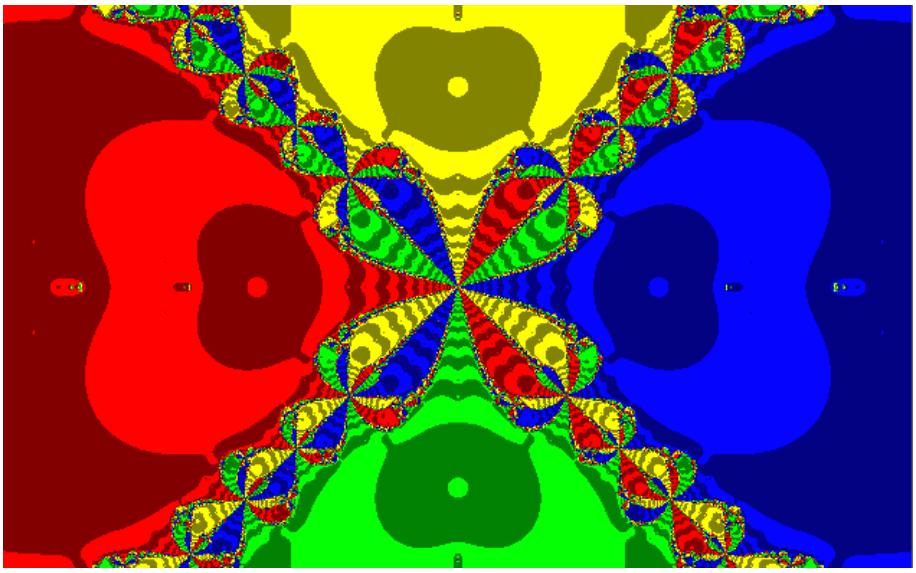


Fig. 10. Método de Cuarto Orden (derivadas analíticas) con 3 iteraciones internas ($K_{min} = 2$, $K_{med} = 5.3$, $K_{max} = 54$, $D_H = 1.979413$).

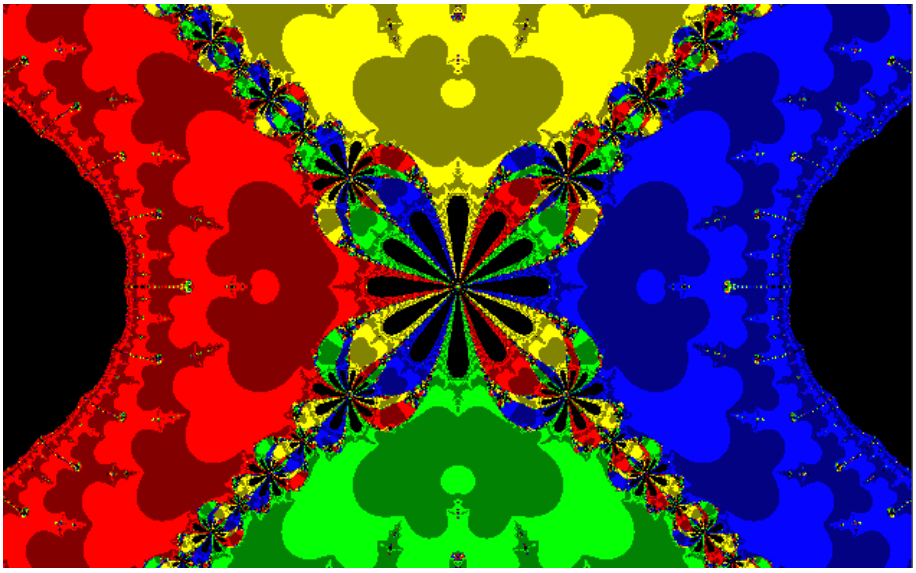


Fig. 11. Método de Cuarto Orden (derivadas analíticas) con 15 iteraciones internas ($K_{min} = 2$, $K_{med} = 183.7$, $K_{max} = 1000$, $D_H = 1.942521$).

Tomando en consideración la antes mencionada conclusión, si se introduce una fórmula dentro de otra para el método de segundo orden (estilo Newton, 2.5.(9) – (12)) y se considera sólo 1 iteración interna (sin factores de relajación) se obtiene la siguiente fórmula compacta

$$\begin{aligned} \mathbf{x}^{k+1} &= \mathbf{x}^k - \left[\, \mathbf{J_f}(\mathbf{x}^k) + \mathbf{H_f}(\mathbf{x}^k).\mathbf{z} \,\right]^{-1}. \left\{ \mathbf{f}(\mathbf{x}^k) - \tfrac{1}{2} \left[\mathbf{H_f}(\mathbf{x}^k).\mathbf{z} \right].\mathbf{z} \right\} \\ \mathbf{z} &= -[\mathbf{J_f}(\mathbf{x}^k)]^{-1}.\mathbf{f}(\mathbf{x}^k) \end{aligned} \tag{5}$$

donde $\mathbf{z}$ es la iteración con Newton-Raphson. El mapa fractal de este método es la fig. 12.b

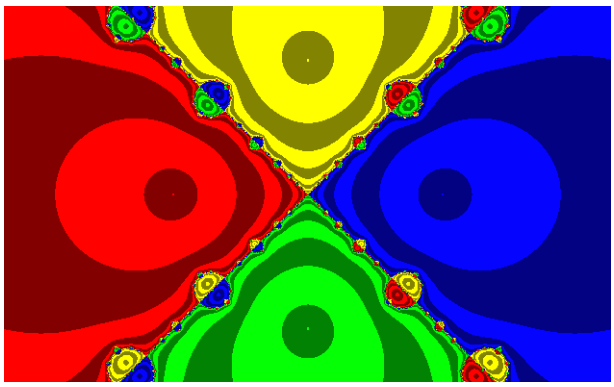


Fig. 12.a. Método (Richmond) de Segundo orden con 1 iteración interna de punto fijo ($K_{min} = 2$, $K_{med} = 5.2$, $K_{max} = 23$, $D_H = 1.994446$).

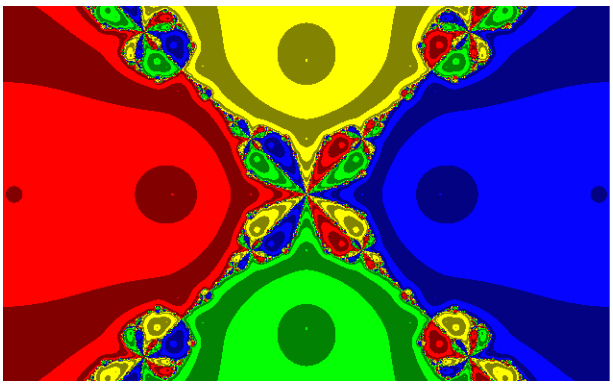


Fig. 12.b. Método de Segundo orden con 1 iteración interna ($K_{min} = 2$, $K_{med} = 5.0$, $K_{max} = 24$, $D_H = 1.987760$).

No importa donde se comience dentro de una cuenca de convergencia, se toma sólamente 5 iteraciones (máximo de 24 en la región fractal) para encontrar la raíz más cercana con esta fórmula algorítmica (para un

problema de cuarto orden). Comparar (5) con la fórmula de Richmond I.2.5.(4), cuya extensión a multivariables se obtendría con 2.5.(7) (el resultado sería (5) sin el $\mathbf{H_f}$ en el "numerador" y multiplicando por $\frac{1}{2}$ en el $\mathbf{H_f}$ del "denominador"), pero como se dijo antes, esta iteración con "punto fijo" es muy lenta y, comparada con (5), peor ($K_{med} = 5.2$, fig. 12.a). Para sólo una ecuación (5) se reduce a

$$x_{k+1} = x_k - \frac{f(x_k)\,\{\,2\,[f'(x_k)]^2 - f(x_k)\,f''(x_k)\,\}}{2\,f'(x_k)\,\{\,[f'(x_k)]^2 - f(x_k)\,f''(x_k)\,\}} \tag{6}$$

la cual es más elaborada que la fórmula de Richmond mostrada en I.2.5.(4), aunque el esfuerzo computacional es un poquito más alto (sin embargo, la fórmula de Richmond es más estable: dimensión fractal un poco más alta). Si se escoge $f''(x_k) = 0$ en (6), se obtiene el método de Newton-Raphson otra vez. Por cierto, da el mismo resultado resolviendo una ecuación $(3.a)$ en los números complejos $x, f(x) \in \mathbb{C}$ con (6), que resolviendo un sistema de ecuaciones reales 2×2 $(3.b, c)$ con (5) en su descomposición real / imaginario $f(z) = f_x(x,y) + i\,f_y(x,y)$ donde $z = x + i\,y$. La fórmula (6) coincide con el método de Collatz $C_m(x)$ $(m = 1)$ I.2.4.(21).

Una inspección de las figuras 6.a, 8, 10 y 12.b inclina a pensar que pocas iteraciones internas es una mejor elección, inclusive elimina las zonas negras, por lo tanto una iteración interna se presenta a continuacón para el método de cuarto orden

$$\begin{aligned}\mathbf{x}^{k+1} = \mathbf{x}^k - \Big[\,\mathbf{J_f}(\mathbf{x}^k) + \{\,\mathbf{H_f}(\mathbf{x}^k) + [\,\tfrac{1}{2}\mathbf{K_f}(\mathbf{x}^k) + \tfrac{1}{6}\mathbf{Q_f}(\mathbf{x}^k).\mathbf{z}\,].\mathbf{z}\,\}.\mathbf{z}\,\Big]^{-1}.\\ \Big\{\,\mathbf{f}(\mathbf{x}^k) + \big[\,\{\,-\tfrac{1}{2}\,\mathbf{H_f}(\mathbf{x}^k) + [\,-\tfrac{1}{3}\mathbf{K_f}(\mathbf{x}^k) - \tfrac{1}{8}\mathbf{Q_f}(\mathbf{x}^k).\mathbf{z}\,].\mathbf{z}\,\}.\mathbf{z}\,\big]\,.\mathbf{z}\,\Big\}\end{aligned} \tag{7}$$

$$x_{k+1} = x_k - \frac{f(x_k)\,\{\,24\,[f'(x_k)]^4 - 12\,f(x_k)\,[f'(x_k)]^2 f''(x_k) + 8\,[f(x_k)]^2 f'(x_k)\,f'''(x_k) - 3\,[f(x_k)]^3 f^{IV}(x_k)\,\}}{f'(x_k)\,\{\,24\,[f'(x_k)]^4 - 24\,f(x_k)\,[f'(x_k)]^2 f''(x_k) + 12\,[f(x_k)]^2 f'(x_k)\,f'''(x_k) - 4\,[f(x_k)]^3 f^{IV}(x_k)\,\}} \tag{8}$$

en ambas formulaciones para sistemas de ecuaciones o para una ecuación. El vector $\mathbf{z}$ en (7) significa lo mismo que en la ecuación (5), el iterado inicial con el método de Newton-Raphson. Es evidente el uso de las expresiones anteriores para métodos de órdenes menores con la eliminación de las derivadas de más alto orden y la reducción de los coeficientes y las potencias de las derivadas. Otra vez, el método de Newton-Raphson se obtiene cuando se eliminan las derivadas de segundo orden o mayores.

La información de las figuras 12.a y 12.b acerca de la dimensión de Hausdorff D_H (estabilidad) y las iteraciones promedio K_{med} (rapidez) indican que un método híbrido puede ser implementado con un factor de relajación ω para un sistema de ecuaciones o para una ecuación

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \big[\,\mathbf{J_f}(\mathbf{x}^k) + \tfrac{\omega+1}{2}\,\mathbf{H_f}(\mathbf{x}^k).\mathbf{z}\,\big]^{-1}.\,\big\{\,\mathbf{f}(\mathbf{x}^k) - \tfrac{\omega}{2}\,[\,\mathbf{H_f}(\mathbf{x}^k).\mathbf{z}\,].\mathbf{z}\,\big\} \tag{9}$$

$$x_{k+1} = x_k - \frac{f(x_k)\,\{\,[f'(x_k)]^2 - \frac{\omega}{2}\,f(x_k)\,f''(x_k)\,\}}{f'(x_k)\,\{\,[f'(x_k)]^2 - \frac{\omega+1}{2}\,f(x_k)\,f''(x_k)\,\}} \tag{10}$$

donde $\mathbf{z}$ es la convencional iteración con Newton-Raphson $(5.b)$. Un factor $\omega \in [0,1]$ selectivo puede modular desde el método de Richmond $\omega = 0$ hacia el método de segundo orden $\omega = 1$. Estabilizando $\omega \to 0$ o acelerando $\omega \to 1$ el método de acuerdo al comportamiento del proceso iterativo. Si es oscilante (caótico inestable en una región fractal), $\omega \to 0$ conduce el método hacia el método de Richmond para estabilizarlo. Si es focalizado hacia una raíz (estable en una cuenca de convergencia), $\omega \to 1$ conduce el método hacia el método de segundo orden para acelerarlo. Un algoritmo para producir este resultado de forma automática es el siguiente: $\omega_{\text{inicial}} = 10^{-3}$, entonces si $\|\mathbf{f}(\mathbf{x}^{k+1})\| < \|\mathbf{f}(\mathbf{x}^k)\|$ se modifica el factor de relajación $\omega' = \min(\omega/\tau, 1)$, $\tau < 1$ y se propone $\tau = 10^{-1}$. Si no se cumple entonces $\omega' = \tau\omega$. Este algoritmo es similar al del método de Levenberg-Marquardt sección III.3.2.3.

2.8. METODOS NUMERICOS PARA REDES

Llamemos redes a los sistemas conformados por "elementos" que están unidos por "nodos". Existen dos tipos de variables "puntual" y "distribuída" tanto en elementos como nodos.

En cada elemento la variable puntual puede ser: Caudal, Intensidad de Corriente, Fuerza-Momento, etc. La variable distribuída puede ser: Presión, Voltaje, Deformación axial-angular, etc. Estas variables se relacionan mediante una ecuación homogénea, por ejemplo, $\Delta P + f(Q) = 0$. El diferencial de la variable distribuída es función de una potencia α de la variable puntual Q en cada elemento. Estos elementos pueden ser: Tuberías, Líneas Eléctricas, Vigas, etc. Puede ser que $f(Q) = C\,|Q|^{\alpha-1}Q$, lo que determina unívocamente el sentido de Q. Las derivadas de $f(Q)$ particularmente son, $f'(Q) = \alpha\,C\,|Q|^{\alpha-1}$, necesaria para el cálculo del la matriz jacobiana y, $f''(Q) = \alpha\,(\alpha-1)\,C\,|Q|^{\alpha-2}\,\mathrm{sign}(Q)$, necesaria para el cálculo del tensor hessiano. El coeficiente C también puede depender de Q, normalmente de forma no-lineal, aunque cuando se linealiza el sistema, este coeficiente se considera constante, al menos para una iteración (linealización de Wood), o para el cálculo de la matriz jacobiana o el tensor hessiano. Luego de cada iteración se actualiza el valor de C (o el valor de $f'(Q)$) al valor de Q actual, cuando corresponda.

El exponente α, en los casos corrientes como redes de tuberías, suele ser $\alpha = 2$, por lo que $f''(Q) = \alpha\,(\alpha-1)\,C\,\mathrm{sign}(Q)$ en estos casos. Para estructuras $\alpha = 1$ y C es la inversa de la matriz de rigidez (3×3 bidimensional, 6×6 tridimensional) y el producto $C\,Q$ en $f(Q)$ es matricial (Matriz×vector). Para redes eléctricas $\alpha = 1$ y C es la resistencia eléctrica simple (constante). Cuando, adicionalmente, la resistencia eléctrica depende de la temperatura, o sea depende (de forma aproximada) linealmente de la potencia $C = C(W) \approx \beta\,W$, donde la potencia $W = C\,Q^2$, entonces el modelo definitivo es equivalente a hacer $\alpha \approx 3$ ($2 < \alpha \lesssim 3$), con C ($= \beta$) una constante global.

En cada nodo, las variables se invierten, las puntuales se convierten en distribuídas y viceversa. La sumatoria de todas las variables distribuídas es nula, por ejemplo, $\sum Q + Q_o = 0$, y existe una única variable puntual P en el nodo. La ecuación por supuesto también es homogénea (aunque la variable P no interviene explícitamente en su ecuación). La convención de suma para estas ecuaciones es simple: lo que entra en el nodo se suma, lo que sale se resta. La constante Q_o es una fuente ($Q_o > 0$) o sumidero ($Q_o < 0$) del sistema (fuerza-momento exteriores para estructuras). Para redes eléctricas o circuitos electrónicos las ecuaciones planteadas en los tres últimos párrafos no son más que las Leyes de Kirchhoff [Wikipedia].

Existen tantas ecuaciones homogéneas como nodos y elementos (una variable y una ecuación por cada uno), y el sistema de ecuaciones planteado para las incógnitas, variables puntuales y distribuídas, se pueden resolver con estos métodos. Para que el sistema sea compatible determinado al menos una P en un nodo debe ser conocida. Los elementos se pueden agrupar en circuitos no redundantes o dependientes (teoría de grafos) para eliminar variables P. De otra forma el sistema se convierte en compatible indeterminado. Otros elementos de la red, no incluidos aquí, también se pueden modelar en función de las variables Q y ΔP, como se puede observar por ejemplo en [Granados,2020].

2.8.1 Introducción

El problema que se desea resolver es un sistema de ecuaciones algebraicas de los siguientes dos tipos

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \qquad\qquad \mathbf{x} = \mathbf{g}(\mathbf{x}) \tag{1}$$

La primera ecuación de la izquierda se denomina *ecuación homogénea*, la segunda en la derecha es cualquier despeje de la primera (o cualquier formulación algorítmica). La solución de las ecuaciones anterior, designada con la letra $\mathbf{r}$, satisface las siguientes definiciones

$$\mathbf{f}(\mathbf{r}) \equiv \mathbf{0} \qquad\qquad \mathbf{r} \equiv \mathbf{g}(\mathbf{r}) \tag{2}$$

En el primer caso $\mathbf{r}$ se denomina la *raíz* de la función $\mathbf{f}$, y en el segundo caso $\mathbf{r}$ se denomina el *punto fijo* de la función $\mathbf{g}$.

La función $\mathbf{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ es una función vectorial de la variable $\mathbf{x}$ también vectorial. Es decir, existe una dependencia de las componentes de la forma $f_i(x_1, x_2, \ldots, x_j, \ldots, x_n)$ con $i = 1, 2, \ldots, n$. Lo mismo es válido para la función $\mathbf{g}$.

2.8.2 Expansión en Series de Taylor

• Serie de Taylor

La expansión en series de Taylor de la función $\mathbf{f}$ involucrada en la ecuación homogénea hasta el término de segundo orden es

$$\mathbf{f}(\mathbf{r}) = \mathbf{f}(\mathbf{x}) + [\mathbf{J}_f(\mathbf{x})].(\mathbf{r} - \mathbf{x}) + \tfrac{1}{2}\,[\mathbf{H}_f(\mathbf{x})] : (\mathbf{r} - \mathbf{x})(\mathbf{r} - \mathbf{x}) + \cdots \tag{3}$$

La serie se desarrolla alrededor del punto $\mathbf{x}$ y esta evaluada en $\mathbf{r}$. La operación producto ":" es una doble contracción de los índices contiguos de las componentes de los factores (identificable como el producto escalar de dos tensores de segundo orden), mientras que un solo punto es una contracción simple (identificable como el producto escalar de dos vectores). Esto hace que los vectores y tensores descritos pertenezcan a espacios de Hilbert. Una generalización de las expansiones en Series de Taylor para funciones multi-variables puede verse en el Apéndice.

• Matriz Jacobiana

El tensor de segundo orden $\mathbf{J}_f$ en la expansión en serie anterior se denomina el *tensor jabobiano* y tiene componentes

$$[\mathbf{J}_f(\mathbf{x})]_{ij} \equiv J_{ij} = \frac{\partial f_i}{\partial x_j} \tag{4}$$

agrupados de forma matricial en un arreglo de dos índices.

• Tensor Hessiano

El tensor de tercer orden $\mathbf{H}_f$ en la expansión en serie anterior se denomina el *tensor hessiano* y tiene componentes

$$[\mathbf{H}_f(\mathbf{x})]_{ijk} \equiv H_{ijk} = \frac{\partial^2 f_i}{\partial x_j \partial x_k} \tag{5}$$

agrupados en un arreglo de tres índices.

2.8.3. Métodos Algebraicos

• Punto Fijo

Utilizando el despeje de la ecuación homogénea en la forma

$$\mathbf{x}^{s+1} = \mathbf{g}(\mathbf{x}^s) \tag{6}$$

se puede implementar un esquema iterativo convergente tal que $\|\boldsymbol{\epsilon}^{s+1}\| < \|\boldsymbol{\epsilon}^s\|$, donde $\boldsymbol{\epsilon}^s = \mathbf{x}^s - \mathbf{x}^{s-1}$ es el *error local*. Dicho esquema se detendría si se satisface la *condici'on de parada* en el error local $\epsilon^s = \|\boldsymbol{\epsilon}^s\| \leq \epsilon_{max}$ y simultáneamente en la *desviación global* $\delta^s = \|\mathbf{f}(\mathbf{x}^s)\| \leq \delta_{max}$, donde los valores ϵ_{max} y δ_{max} son las *tolerancias* permitidas. También se impone una condición de parada en el número de iteraciones $s > s_{max}$ para evitar procesos iterativos no convergentes en un número máximo razonable de iteraciones s_{max}.

Este método se puede relajar en la forma

$$\mathbf{x}^{s+1} = \mathbf{x}^s + \omega\,[\,\mathbf{g}(\mathbf{x}^s) - \mathbf{x}^s\,] \tag{7}$$

siendo ω el factor de relajación.

• Linealización de Wood

Una forma particular del método de punto fijo es mediante la linealización del tipo [Wood & Charles,(1972)]

$$[\mathbf{A}(\mathbf{x}^s)].\,\mathbf{x}^{s+1} = \mathbf{b} \tag{8}$$

donde lo que no depende del valor actual $\mathbf{x}^{s+1}$ se aglomera en una matriz $[\mathbf{A}]$ dependiente de la iteración anterior $\mathbf{x}^s$. Por ejemplo, se considera "momentáneamente constante" $C\,|Q|^{\alpha-1}$, en la ecuación de $f(Q)$ para la iteración inmediata anterior. Resolviendo el sistema de ecuaciones lineales anterior para cada iteración s se obtiene el esquema iterativo deseado. Los criterios de convergencia y las condiciones de parada seguirán siendo los mismos para todos los esquemas iterativos propuestos antes.

Este método se relajaría de forma parecida como

$$\mathbf{x}^{s+1} = \mathbf{x}^s + \omega\,\{\,[\mathbf{A}(\mathbf{x}^s)]^{-1}.\mathbf{b} - \mathbf{x}^s\} \tag{9}$$

utilizando la matriz inversa $[\mathbf{A}]^{-1}$ (realmente es un método de punto fijo). Esto es equivalente a hallar $\mathbf{z}$ de $[\mathbf{A}(\mathbf{x}^s)].\mathbf{z} = \mathbf{b} - [\mathbf{A}(\mathbf{x}^s)].\mathbf{x}^s$ y subtituir en $(10.a)$ (lo mismo que hacer $\mathbf{f}(\mathbf{x}) = [\mathbf{A}(\mathbf{x}^s)].\mathbf{x} - \mathbf{b} = \mathbf{0}$ en el método de Newton-Raphson descrito abajo).

2.8.4 Métodos Analíticos

Dos métodos iterativos, que también son métodos del tipo punto fijo, se deducen de forma analítica a partir de la expansión en series de Taylor hasta el término de primer orden. Luego reasignando $\mathbf{r} = \mathbf{x}^{s+1}$ y $\mathbf{x} = \mathbf{x}^s$ se obtienen los dos siguientes métodos.

• Newton-Raphson

Directamente con la reasignación antes descrita y despejando $\mathbf{r}$ se obtiene el método de Newton-Raphson, cuya fórmula algorítmica se traduce en las siguientes expresiones [Granados,1991]

$$\mathbf{x}^{s+1} = \mathbf{x}^s + \omega\,\mathbf{z}^s \qquad\qquad [\mathbf{J}_f(\mathbf{x}^s)].\mathbf{z}^s = -\mathbf{f}(\mathbf{x}^s) \tag{10}$$

donde la expresión de la derecha es un sistema de ecuaciones lineales. El vector $\mathbf{z}^s = \boldsymbol{\epsilon}^{s+1}$ es el error local en la iteración $s+1$, en el caso de que no haya relajamiento del método ($\omega = 1$). El método de Newton-Raphson es un caso especial de método de punto fijo si se define la siguiente aplicación

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} - \omega\,[\mathbf{J}_f(\mathbf{x})]^{-1}.\mathbf{f}(\mathbf{x}) \tag{11}$$

• Hardy-Cross

Asumiendo un valor de $\mathbf{z}_i^k$ de avance del método de Newton-Raphson anterior, que es igual para todas las componentes de la variable $\mathbf{x}^k$, pero únicamente las envueltas en la ecuación i. Se establece entonces el siguiente método de *Hardy-Cross* [Cross,(1936)]

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \omega\,\mathbf{z}_i^k \qquad\qquad z_i^k = -f_i(\mathbf{x}^k)/\sum_{j=1}^{n} J_{ij} \qquad\qquad \mathbf{z}_i^k = z_i^k\,\mathbf{1} \tag{12}$$

donde la sumatoria aparece de sumar todos los elementos de la fila i de la matriz jacobiana J_{ij}. El rango m del índice i ($i = 1, 2, \ldots, m$) puede ser menor que el rango n del índice j ($j = 1, 2, \ldots, n$), a diferencia de los métodos anteriores, donde el sistema de ecuaciones debe ser "compatible determinado" (igual número de ecuaciones que de incógnitas). El vector $\mathbf{1}$ tiene unos en todas sus componentes.

El recorrido de la red se hace en m circuitos cerrados o pseudo-circuitos abiertos, tales que la variable P_j tenga el mismo valor en el nodo de partida y de llegada (o un diferencial ΔP conocido, al menos que se eliga dicho diferencial como incógnita), y así se elimine dicha variable de la ecuación i correspondiente (o sólo quede ΔP como incógnita). Las variables P intermedias también quedan eliminadas en todo su recorrido

del circuito dentro de este proceso. Los circuitos no deben ser redundantes, es decir, dos circuitos no deben tener las mismas incógnitas.

• Otros

Un método que ha dado muy buenos resultados es el método de segundo orden. Puesto que el sistema de ecuaciones no lineales es aproximadamente de segundo grado ($\alpha = 2$) cuando $I\!R\!e_{\rm D} \longrightarrow \infty$ (en redes de tuberías), entonces los métodos de segundo orden son los idóneos para este tipo de problemas. A esto se le agrega el hecho de que muestran una mayor estabilidad y convergencia segura en los casos de alta no-linealidad [Granados,1991] [idem,1995]. Cuando $\alpha = 2$ no se justifica usar métodos de Taylor de tercer orden en adelante. Cuando $\alpha = 1$ no se justifica usar métodos de Taylor de segundo orden en adelante.

2.8.5. Análisis

De todos los métodos anteriores, el mejor de ellos es el método de Newton-Raphson, por su simpleza, pero mejor aún es el método de segundo orden. Le sigue de cerca el método de Hardy-Cross. De último, el peor de ellos, es el método de linealización de Wood. Por esta razón, en este último método es casi imprescindible sub-relajarlo ($\omega < 1$) para obtener convergencia segura, en caso de escoger un iterado inicial $\mathbf{x}^0$ lejano a la solución.

Todo método de punto fijo (todos lo son) converge a la solución $\mathbf{r}$, si se satisface que la aplicación $\mathbf{g}$ del algoritmo es una contracción del espacio en su cercanía. Esto es,

$$\|\mathbf{J}_g(\boldsymbol{\zeta})\| < 1 \tag{13}$$

donde $\boldsymbol{\zeta}$ pertenece a un entorno $\boldsymbol{\mathcal{V}}^s_{\mathbf{r}} = \mathbb{B}(\mathbf{r}, \|\mathbf{r} - \mathbf{x}^s\|)$, o sea, la n-bola abierta con centro en $\mathbf{r}$ y radio $\|\mathbf{r} - \mathbf{x}^s\|$. La norma $\| \cdot \|$ del tensor debe estar subordinada a la norma sobre vectores.

Aunque el método no se muestra aquí, y por la forma de la ecuación de los elementos de la red, el mejor método de todos, con convergencia casi segura, es el método de *Segundo Orden* mostrado en la sección 2.5 (comprobado por experiencia propia). [Granados,1991] [idem,1995/1996] [idem,2015]

BIBLIOGRAFIA

[1] Cross, H. "Analysis of Flow in Networks of Conduits or Conductors". **Univ. Illinois Bull.**, Engineering Experiment Station, No.**286**, (1936).

[2] Bonnans, J. F.; Gilbert J. Ch.; Lemaréchal, C.; Sagastizábal, C. A. **Numerical Optimization - Theoretical and Practical Aspects**, Second Edition, Springer-Verlag (Berlin), 2006.

[3] Broyden, C. G. "A Class of Methods for Solving Non-Linear Simultaneous Equations", **Mathematics of Computation**, Vol.**19**, pp.577-593, (1965).

[4] Burden R. L.; Faires, J. D. **Numerical Analysis**. 3^{rd} Edition. PWS (Boston), 1985.

[5] Conte, S.D.; deBoor, C. **Elementary Numerical Analysis**. McGraw-Hill (New York), 1972.

[6] Cottle, R. W.; Thapa, M. N. **Linear and Nonlinear Optimization**. Springer Science+Business Media (New York), 2017.

[7] Dantzig, G. B.; Thapa, M. N. **Linear Programming**. Vol.**1**: "Introduction". Vol.**2**: "Theory and Extension". Springer (New York), 1997/2003.

[8] Dennis, J. E. Jr.; Moré, J. J. "Cuasi-Newton Methods, Motivation and Theory", **SIAM Review**, Vol.**19**, No.1, pp.46-89, (1977).

[9] Devaney, R. L. **An Introduction to Chaotic Dynamical Systems**. Addison-Wesley, 1987.

[10] Gerald, C. F. **Applied Numerical Analysis**. 2^{nd} Edition. Addison-Wesley, 1978.

[11] Granados M., A. L. **Second Order Methods for Solving Non-Linear Equations**, INTEVEP, S. A. (Research Institute for Venezuelan Petroleum Industry), Tech. Rep. No.INT-EPPR/322-91-0002, Los Teques, Edo. Miranda, Jun, 1991, págs. 14-36.

[12] Granados M., A.L. "Fractal Technics to Measure the Numerical Instability of Optimization Methods". **Mecánica Computacional Vol.XV**: Anales del "9° CONGRESO SOBRE METODOS NUMERICOS Y SUS APLICACIONES, ENIEF'95". Hotel Amancay, 6-10 de Noviembre de 1995, San Carlos de Bariloche, Argentina. Compilado por Larreteguy, A. E. y Vénere, M. J. Asociación Argentina de Mecánica Computacional (AMCA), pp.369-374,1995.

[13] Granados M., A. L. "Fractal Techniques to Measure the Numerical Instability of Optimization Methods". **Numerical Methods in Engineering Simulation**: *Proceedings of The Third International Congress on Numerical Methods in Engineering and Applied Sciences, CIMENICS'96.* Cultural Centre Tulio Febres Cordero, March 25-29, 1996. Mérida, Venezuela. Editors: M. Cerrolaza, C. Gajardo, C. A. Brebbia. Computational Mechanics Publications of the Wessex Institute of Technology (UK), pp.239-247, (1996).

[14] Granados M., A. L. "Numerical Taylor's Methods for Solving Multi-Variable Equations", Universidad Simón Bolívar, Mayo, 2015. https://www.academia.edu/12520473/Numerical_Taylors_Methods_for_Solving_Multi-Variable_Equations

[15] Granados M., A. L. "Taylor Series for Multi-Variable Functions", Universidad Simón Bolívar, Dic. 2015. https://www.academia.edu/12345807/Taylor_Series_for_Multi-Variables_Functions

[16] Granados M., A. L. **Mecánica y Termodinámica de Sistemas Materiales Continuos**. Universidad Simón Bolívar, 2022. ISBN 980-07-2428-1.

[17] Granados M., A. L. "Métodos Numéricos para Redes", Universidad Simón Bolívar, Mar. 2023.

[18] Granados M., A. L. "Redes de Tuberías". Universidad Simón Bolívar, Departamento de Mecánica, Nov., 2020. https://www.academia.edu/20394578/Redes_de_Tuberías

[19] Granados M., A. L. "Algoritmo Simplex Lineal y No Lineal". Universidad Simón Bolívar, Departamento de Mecánica, Sep., 2024. https://www.academia.edu/123412738/Algoritmo_Simplex_Lineal_y__No_Lineal

[20] Hämmerlin, G.; Hoffmann, K.-H. **Numerical Mathematics**. Springer-Verlag (New York), 1991.

[21] Hillier, F. S.; Lieberman, G. J. **Introducción a La Investigación de Operaciones**, Novena Edición. McGraw-Hill (México), 2010.

[22] Hoffman, K.; Kunze, R. **Linear Algebra**, 2^{nd} Edition. Prentice-Hall (Englewood Cliff-New Jersey), 1971.

[23] Horner, W. G. "A new method of solving numerical equations of all orders, by continuous approximation". **Philosophical Transactions of the Royal Society of London**, pp.308-335, july (1819).

[24] Householder, A. S. **The Theory of Matrices in Numerical Analysis**. Blaisdell Publishing Company (New York), 1964. Dover Publications (new York), 1975.

[25] Lapidus, L. **Digital Computation for Chemical Engineers**. McGraw-Hill (New York), 1962.

[26] Layton, W.; Sussman, M. **Numerical Linear Algebra**. World Scientific Publishing (New Jersey), 2020.

[27] Luenberger, D. G. **Optimization by Vector Space Methods**. John Wiley & Sons, 1969.

[28] Luenberger, D. G.; Ye, Y. **Linear and Nonlinear Programming**, 5^{th} Edition. Springer Nature (Switzerland), 2021.

[29] Mandelbrot, B. B. **The Fractal Geometry of Nature**, Updated and Augmented Edition. W. H. Freeman and Company (New York), 1983.

[30] Méndez, M. V. **Tuberías a Presión.** En Los Sistemas de Abastecimiento de Agua. Fundación Polar & Universidad Católica Andrés Bello, 1995.

[31] Ortega, J. M. **Numerical Analysis**, A Second Course. SIAM, 1990.

[32] Ortega, J. M.; Rheinboldt, W. C. **Iterative Solution of Nonlinear Equations in Several Variables**. Academic Press, 1970.

[33] Peitgen, H.-O.; Richter, P. H. **The Beauty of Fractals. Images of Complex Dynamical Systems**. Springer-Verlag, 1986.

[34] Pennington, R. H. **Introductory Computer Methods and Numerical Analysis**, 2^{nd} Edition. Collier Macmillan Ltd., 1970.

[35] Pundir, S. K. **Applied Numerical Analysis**. CBS Publisher & Distributors Pvt. Ltd., 2023.

[36] Schatzman, M. **Numerical Analysis**, A Mathematical Introduction. Oxford University Press, 2002.

[37] Stewart, G. W. **Introduction to Matrix Computations.** Academic Press (New York), 1973.

[38] Wood, D. J.; Charles, C. O. A. "Hydraulic Network Analysis Using Linear Theory". **Journal of The Hydraulics Division**, ASCE, Vol.**98**, No.HY7, July (1972).

[39] https://es.wikipedia.org/wiki/Leyes_de_Kirchhoff

CAPITULO III

INTERPOLACION, INTEGRACION Y APROXIMACION

CONTENIDO

En la *interpolación* las funciones usadas, normalmente polinomios, pasan por los puntos dados como datos. No puede haber puntos con abscisas repetidas. En la *aproximación* las funciones pasan aproximadamente por los puntos datos, que conforman una nube de puntos, minimizando los errores en promedio cuadrática. Pueden haber abscisas repetidas, inclusive puntos repetidos.

1. INTERPOLACION

Sean $(x_0, f(x_0))$, $(x_1, f(x_1))$, $(x_2, f(x_2))$ hasta $(x_n, f(x_n))$ los $n+1$ puntos discretos que representan a una función $y = f(x)$. Como se sabe, existe un único polinomio $y = P_n(x)$ de grado n que pasa por los $n+1$ puntos mencionados. Estos polinomios son adecuados para realizar estimaciones de la función $y = f(x)$ para un valor x cualquiera perteneciente al intervalo $[x_0, x_1, x_2, x_3, \ldots, x_n]$ que contiene a todos los puntos, estando los valores x_i no necesariamente ordenados, ni habiendo valores repetidos. A este proceso se le denomina "Interpolación". Si el valor x está fuera del intervalo de los puntos entonces el proceso se denomina "Extrapolación".

En esta sección se ha desarrollado un algoritmo de interpolación usando los polinomios de Newton en diferencias divididas. Se han usado dos criterios para hacer la interpolación lo más consistente posible con los puntos discretos dados: Simetría y monotonía.

El criterio de la simetría consiste en escoger la distribución de puntos lo más simétricamente posible, alrededor de donde se desee interpolar. Esto se puede hacer de dos maneras: mediante el número de puntos o mediante la distancia de influencia a uno y otro lado del punto donde se vaya a interpolar. En el caso de intervalos regulares una de las formas implica a la otra, pero no cuando los datos son irregulares o no están ordenados.

El criterio de la monotonía se basa en la definición de monotonía de una función: Una función se dice que es monótona hasta el orden m, en un determinado intervalo, si todas sus derivadas de hasta dicho orden conservan siempre su signo en dicho intervalo. Las diferencias divididas son proporcionales a las derivadas en

su entorno, por ello el criterio de monotonía implica escoger hasta el mayor orden en las diferencias divididas que tengan igual signo. La última diferencia dividida deberá tener signo opuesto a una o ambas de las diferencias divididas vecinas. La falta de monotonía implica que pueden producirse oscilaciones indeseables de la función alrededor o entre los puntos dados.

Los criterios de simetría y monotonía se complementan para indicar que puntos y cuantos de ellos se deben usar en la interpolación. En cualquier caso, el grado del polinomio será siempre una unidad menor que el número de puntos. El algoritmo se resume de la siguiente manera: se escogen los puntos más cercanos al punto donde se desee interpolar, en un número (o distancia) simétrica, hasta que dicho número de puntos reflejen, en las diferencias divididas, que la función conserva la monotonía deseada.

El algoritmo antes explicado puede usarse para hacer interpolaciones en una o en varias dimensiones. También permite la interpolación sin necesidad de pre-ordenar los puntos usados. En varias dimensiones, lo único que se exige es que los valores de las funciones sean siempre para los mismos y todos los valores discretos en cada dimensión. El algoritmo tampoco necesita escoger un grado del polinomio anticipadamente, durante el proceso de la interpolación el algoritmo decide el grado del polinomio óptimo que garantice satisfacer los criterios de simetría y monotonía.

Los algoritmos explicados adelante se han utilizado, por ejemplo, para encontrar el campo de velocidades y sus derivadas en todo el dominio del flujo, basado en los valores de dicho campo en puntos discretos en el espacio. Se ha escogido interpolaciones polinómicas de hasta cuarto grado (cinco puntos en cada dirección espacial) para hacer las interpolaciones, siguiendo el criterio de que el error de las interpolaciones debe ser menor que el de los valores discretos usados (segundo orden). Además, el número de puntos se justifica al usar el criterio de la simetría. Luego la monotonía elimina el orden innecesario. Durante el proceso de convergencia, apenas se ha usado interpolaciones parabólicas (tres puntos en cada dirección) para agilizar los tiempos de ejecución.

1.1. DATOS IRREGULARES

Los datos discretos no necesariamente están ordenados, y en el caso de que así lo sean, las distancias entre dos puntos consecutivos es constante. A esto es lo que denominamos *datos irregulares*.

1.1.1. Diferencias Divididas

Las diferencias divididas [Carnahan et al.,1969] simbolizadas por $f[\ \cdot\ ]$ se definen de manera recurrente de la siguiente forma

$$f[x_0] = f(x_0) \tag{1.a}$$

$$f[x_1, x_0] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} \tag{1.b}$$

$$f[x_2, x_1, x_0] = \frac{f[x_2, x_1] - f[x_1, x_0]}{x_2 - x_0} \tag{1.c}$$

$$f[x_3, x_2, x_1, x_0] = \frac{f[x_3, x_2, x_1] - f[x_2, x_1, x_0]}{x_3 - x_0} \tag{1.d}$$

$$\vdots$$

$$f[x_n, x_{n-1}, \ldots, x_1, x_0] = \frac{f[x_n, x_{n-1}, \ldots, x_2, x_1] - f[x_{n-1}, x_{n-2}, \ldots, x_1, x_0]}{x_n - x_0} \tag{1.e}$$

Las diferencias divididas cumplen con la propiedad

$$f[x_n, x_{n-1}, \ldots, x_1, x_0] = f[x_0, x_1, \ldots, x_{n-1}, x_n] \qquad \forall n \in \mathbb{N} \tag{2}$$

Esta propiedad, expresada para cualquier n, lo que significa es que, sin importar el orden en que están los valores x_i dentro de una diferencia dividida, el resultado es siempre el mismo. Dicho de otra forma concisa, la

diferencia dividida es invariante a cualquier permutación de sus argumentos. Esta propiedad la hace adecuada para los cálculos, como veremos en adelante.

Una forma de expresar todas las diferencias divididas posibles de generar mediante, por ejemplo, un conjunto de cuatro puntos $(x_0, f(x_0))$, $(x_1, f(x_1)$, $(x_2, f(x_2))$, $(x_3, f(x_3))$ y $(x_4, f(x_4))$, no necesariamente ordenados, es lo que se denomina el *Diagrama Romboidal* de diferencias divididas. Para el ejemplo propuesto se tiene que el diagrama romboidal se representa como

$$\begin{array}{c|c|c|c|c|c} x_0 & f[x_0] & & & & \\ & & f[x_0,x_1] & & & \\ x_1 & f[x_1] & & f[x_0,x_1,x_2] & & \\ & & f[x_1,x_2] & & f[x_0,x_1,x_2,x_3] & \\ x_2 & f[x_2] & & f[x_1,x_2,x_3] & & f[x_0,x_1,x_2,x_3,x_4] \\ & & f[x_2,x_3] & & f[x_1,x_2,x_3,x_4] & \\ x_3 & f[x_3] & & f[x_2,x_3,x_4] & & \\ & & f[x_3,x_4] & & & \\ x_4 & f[x_4] & & & & \end{array} \tag{3}$$

Se puede observar que para obtener cualquier diferencia dividida en un vértice de un triángulo imaginario, basta con restar las diferencias divididas contiguas y dividirla entre la resta de los valores extremos de x de la base de dicho triángulo.

Manipulación algebraíca de la diferencias de órdenes crecientes conlleva, mediante inducción, a una forma simétrica similar para la k-ésima diferencia dividida, en término de los argumentos x_i y de los valores funcionales $f(x_i)$. Esta forma simétrica puede ser escrita de manera compacta como

$$f[x_0, x_1, x_2, \ldots, x_{k-1}, x_k] = \sum_{i=0}^{k} \frac{f(x_i)}{\prod\limits_{\substack{j=0 \\ j\neq i}}^{k} (x_i - x_j)} \tag{4}$$

Substituir esta expresión (4) en los polinomios de Newton en diferencias divididas (6.b) luego, no conlleva directamente a los polinomios de Lagrange 1.2.(1)-(2), como veremos más adelante.

1.1.2. Polinomios en Diferencias Divididas

Estos polinomios se les conoce como *polinomios de Newton en diferencias divididas*. Los polinomios de Newton $P_n(x)$ de grado n en diferencias divididas [Carnahan et al.,1969], como se dijo antes, permiten hacer estimaciones de la función $y = f(x)$ de puntos intermedios (o estrapolaciones en puntos extramedios) en la forma

$$f(x) = P_n(x) + R_n(x) \tag{5}$$

donde $P_n(x)$ es el polinomio de grado n

$$\begin{aligned} P_n(x) =& f[x_0] + (x - x_0)\ f[x_0, x_1] + (x - x_0)(x - x_1)\ f[x_0, x_1, x_2] + \cdots \\ &+ (x - x_0)(x - x_1)(x - x_2)\cdots(x - x_{n-1})\ f[x_0, x_1, x_2, \ldots, x_{n-1}, x_n] \\ =& \sum_{k=0}^{n} \prod_{j=0}^{k-1} (x - x_j)\ f[x_0, x_1, x_2, \ldots, x_{k-1}, x_k] \end{aligned} \tag{6}$$

y la función $R_n(x)$ es el error cometido en la interpolación

$$R_n(x) = \prod_{j=0}^{n} (x - x_j)\ f[x_0, x_1, x_2, \ldots, x_{n-1}, x_n, x] \tag{7.a}$$

$$= \prod_{j=0}^{n} (x - x_j)\ \frac{f^{(n+1)}(\xi)}{(n+1)!} \qquad \xi \in [x_0, x_1, \ldots, x_{n-1}, x_n] \tag{7.b}$$

siendo ξ el valor comprendido entre el menor y mayor de los valores $\{x_0, x_1, \ldots, x_{n-1}, x_n\}$. Naturalmente $R_n(x_i) = 0$ para $i = 1, 2, 3, \ldots, n$, ya que el polinomio pasa por cada uno de los puntos $(x_i, f(x_i))$. Cuando el límite superior de una productoria es menor que el límite inferior, como ocurre con el primer término de (6), el resultado de dicha productoria es la unidad.

La expresión (6)-(7.a) se obtiene de tomar inicialmente $f[x] = f[x_0] + (x - x_0)\, f[x_0, x]$ y luego mediante inducción $f[x_0, x_1, \ldots, x_{n-1}, x] = f[x_0, x_1, \ldots, x_{n-1}, x_n] + (x - x_n)\, f[x_0, x_1, \ldots, x_{n-1}, x_n, x]$.

Un ejemplo sencillo es la parábola que pasa por los tres puntos $(a, f(a))$, $(b, f(b))$ y $(c, f(c))$

$$P_2(x) = f[a] + (x - a)f[a, b] + (x - a)(x - b)f[a, b, c] \tag{8}$$

donde $f[a, b]$ es la pendiente de recta entre los puntos a y b y $f[a, b, c]$ es la curvatura de la parábola, que si es positiva es abierta hacia arriba y si es negativa es abierta hacia abajo. Esta parábola ya se ha usado antes en los métodos de segundo orden cerrados sección I.1.4.2 y abiertos secciones I.2.5.2 y I.2.5.3.

En general, los datos utilizados en interpolación no estarán ordenados, ni serán regulares. Al final se usan los polinomios en diferencias divididas por la razón adicional, justificada adelante, de que se pueden aplicar fácilmente los criterios de interpolación sin tener que pre-ordenar los datos, y al agregar datos nuevos cercanos a la interpolación que no estaban antes. Esto, sin tener que armar todo el polinomio de interpolación otra vez, como en el caso del polinomio de Lagrange.

1.1.3. Residual

A continuación se hará la deducción de la expresión (7.b) para $R_n(x)$ [Carnahan et al.,1969]. Consideremos las fórmulas (5), (6) y (7.a) fundamentales de Newton

$$f(x) = P_n(x) + R_n(x) = P_n(x) + \left[\prod_{i=0}^{n}(x - x_i)\right] G(x) \tag{9}$$

con

$$G(x) = f[x_0, x_1, x_2, \ldots, x_{n-1}, x_n, x] \tag{10}$$

en la cual $P_n(x)$ es el polinomio de interpolación de orden n dado por (6) y $R_n(x)$ es el término residual o residuo (7.a) y $G(x)$ es el cociente incremental que incluye x de orden $n + 1$ y que es deconocido.

Para los puntos que forman la base de datos $x_0, x_1, \ldots, x_{n-1}, x_n$, $R_n(x_i) = 0$, pero para cualquier otro punto, en general $R_n(x) \neq 0$. Consideremos por otro lado una nueva función $Q(t)$, tal que

$$Q(t) = f(t) - P_n(t) - \left[\prod_{i=0}^{n}(t - x_i)\right] G(x) \tag{11}$$

Cuando $t = x_i$, $i = 0, 1, 2, \ldots, n$, $Q(t) = 0$; y cuando $t = x$ también $Q(t) = 0$, ya que el término de la derecha de (11) desaparece (véase (6)). Es decir, que la función $Q(t)$ se anula $n + 2$ veces, o sea que tiene $n + 2$ raíces en el intevalo más pequeño que contenga x y los $n + 1$ puntos base $x_0, x_1, \ldots, x_{n-1}, x_n$. si $f(t)$ es continua y convenientemente diferenciable, se le puede aplicar el siguiente teorema:

Teorema 1. (Teorema de Rolle). Sea $f(x)$ una función continua en el intervalo $a \leq x \leq b$ y diferenciable en $a < x < b$; si $f(a) = f(b)$, entonces existe por lo menos un punto ξ, siendo $a < \xi < b$, para el cual $f'(\xi) = 0$.

El teorema exige que la función $Q'(t)$ se anule por lo menos $n + 1$ veces en intervalo de los puntos base. Aplicando el teorema repetidamente a las derivadas de orden superior, se observa que $Q''(t)$ debe tener n raíces, $Q'''(t)$, $n - 1$ raíces, etc... y que $Q^{(n+1)}(t)$ debe anularse por lo menos una vez en el intervalo que contenga los puntos bases. Sea dicho punto $t = \xi$. Derivando la expresión (11) $n + 1$ veces, se obtiene

$$Q^{(n+1)}(t) = f^{(n+1)}(t) - P_n^{(n+1)}(t) - (n + 1)!\, G(x) \tag{12}$$

Pero $P_n(t)$ es un polinomio de grado n, de modo que $P_n^{(n+1)}(t) = 0$, y por tanto, para $t = \xi$ se satisface

$$G(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \qquad \xi \in [x_0, x_1, \ldots, x_{n-1}, x_n, x] \tag{13}$$

o sea que se justifica (7.b), cuando x está en el intervalo base (interpolación) y

$$R_n(x) = \prod_{j=0}^{n}(x - x_j)\,\frac{f^{(n+1)}(\xi)}{(n+1)!} \qquad \xi \in [x_0, x_1, \ldots, x_{n-1}, x_n, x] \tag{14}$$

El valor de ξ es desconocido, salvo que se conoce que está contenido en el intervalo formado por x y los valores $x_0, x_1, \ldots, x_{n-1}, x_n$. Si la función $f(x)$ se describe sólamente de forma tabular, la expresión (14) es de poca utilidad, ya que $f^{(n+1)}(\xi)$ no se puede determinar. No obstante, agregando uno o más puntos adicionales al cálculo, se puede usar la diferencia dividida del mismo orden que la derivada para tener un valor estimativo del error. Por el contrario, si $f(x)$ se conoce de forma analítica, entonces (14) es útil para establecer una cota superior al error.

1.2. POLINOMIOS DE LAGRANGE

Los polinomios de *Lagrange* [Hildebrand,1956] son otra forma de expresar los mismos polinomios $P_n(x)$ de la ecuación (7), pero a través de (4). De manera que se tiene [Carnahan et al.,1969]

$$P_n(x) = \sum_{i=0}^{n} L_i(x)\; f(x_i) \tag{1}$$

donde

$$L_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)} \tag{2}$$

El error $R_n(x)$ continúa siendo el mismo que la expresión 1.1.(7). Cada valor funcional $f(x_i)$ incluido en la expresión (1) es multiplicado por $L_i(x)$, que son todos polinomios de grado n. Por ello reciben el nombre de *multiplicadores de Lagrange*. Un incoveniente que tienen los polinomios de Lagrange es que, para aumentar el grado del polinomio en una unidad, implica el proceso engorroso de agregar un factor adicional a cada multiplicador (productos), y hay que calcular todo de nuevo. Este inconveniente no lo presenta los polinomios de Newton, donde para aumentar un grado al polinomio, sólo hay que agregar un punto y calcular un término adicional (sumas), y todos los cálculos anteriores siguen sirviendo.

EJEMPLO:

Sea la funcion $f(x) = \ln x$. Dada la tabla de valores

Datos

x_i	0.40	0.50	0.70	0.80
$f(x_i)$	-0.916291	-0.693147	-0.356675	-0.223144

estimar el valor de $\ln 0.60$. Evaluando los coeficientes de Lagrange para $i = 1, 2, 3, 4$: $L_0(0.60) = -\frac{1}{6}$, $L_1(0.60) = \frac{2}{3}$, $L_2(0.60) = \frac{2}{3}$, $L_3(0.60) = -\frac{1}{6}$. Por lo que

$$P_3(0.60) = L_0(0.60)\, f(x_0) + L_1(0.60)\, f(x_1) + L_2(0.60)\, f(x_2) + L_3(0.60)\, f(x_3)$$

Sustituyendo, se obtiene que la interpolación del $\ln 0.60$ es $P_3(0.60) = -0.5099075$, el cual comparado con el valor exacto de $\ln(0.60) = -0.5108256$ muestra un desviación global igual a 0.000918.

1.3. DATOS REGULARES

Cuando se tienen datos regulares, estos deberán estár ordenados en la variable independiente x. Por lo que dos puntos consecutivos se distancian en x en un valor constante que designaremos con la letra h. Es decir, $h = x_i - x_{i-1} = x_{i+1} - x_i$ constante para todo i, sin importar si es positivo (datos crecientes) o negativo (datos decrecientes).

1.3.1. Diferencias Adelantadas

Las diferencias adelantadas se obtienen con el mismo procedimiento que las diferencias divididas, sólo que no se dividen.

Para intervalos regulales en x, donde los x_i están ordenados, se define la *diferencia adelantada* $\Delta^k f_i$, tal que

$$\Delta^k f_i = k!\, h^k\, f[x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k-1}, x_{i+k}] \tag{1}$$

donde h es el tamaño del intervalo en x consecutivos ($h = x_{i+1} - x_i$). Se ordenan de forma romboidal al igual que antes, sólo que como son no divididas, por ello aparece el factor $k!\, h^k$.

1.3.2. Polinomios de Newton-Gregory

Para intervalos regulares, el polinomio de Newton en diferencias divididas 1.1.(6) se convierte en

$$P_n(x) = \sum_{k=0}^{n} \binom{s}{k} \Delta^k f_0 \qquad\qquad P_n(x) = \sum_{k=0}^{n} \binom{s-c}{k} \Delta^k f_c \tag{2}$$

denominado polinomio de *Newton-Gregory* progresivo (a la derecha centrado en $c \neq 0$) y donde el número combinatorio significa

$$\binom{s}{k} = \frac{\Gamma(s+1)}{\Gamma(s-k+1)\,\Gamma(k+1)} = \frac{1}{(s+1)}\,\frac{1}{\mathrm{B}(s-k+1, k+1)} \qquad\qquad s = \frac{x - x_0}{h} \tag{3}$$

Particularmente, $\Gamma(k+1) = k!$ por ser k un entero positivo y, aunque la función $\Gamma(s)$ no es un factorial siempre, se satisface $\Gamma(s+1)/\Gamma(s-k+1) = s(s-1)(s-2)\ldots(s-k+1)$ [Gerald,1970]. La función $\mathrm{B}(x, y)$ es la *función Beta de Euler*. La expresión $(2.a)$ se ha obtenido de substituir

$$\prod_{j=0}^{k-1}(x - x_j) = h^k\,\frac{\Gamma(s+1)}{\Gamma(s-k+1)} = k!\, h^k \binom{s}{k} \tag{4}$$

y la ecuación (1), despejada en $f[\,\cdot\,]$ para $i = 0$, en la definición 1.1.(6).

Para intervalos regulares, el error 1.1.(7) se convierte en

$$R_n(x) = \binom{s}{n+1} h^{n+1}\, f^{(n+1)}(\xi) \qquad\qquad \xi \in [x_0, x_1, \ldots, x_{n-1}, x_n, x] \tag{5}$$

teniendo ξ el mismo significado que antes. Los polinomios para intervalos regulares se muestran aquí sólo como caso particular. Aunque es muy difícil o poco práctico conseguir los datos ordenados regularmente siempre.

1.3.3. Diagrama Romboidal

Al igual que en 1.1.(3), las diferencias adelantadas se pueden ordenar de forma tabular, siguiendo un ordenamiento romboidal, insertando en las caras de los rombos los números combinatorios correspondientes, de la forma indicada en el diagrama de abajo.

Si se hace un recorrido del diagrama de izquierda a derecha: Al bajar la diferencia se multiplica por el número combinatorio de arriba. Al subir la diferencia se multiplica por el número combinatorio de abajo. Al hacer un recorrido horizontal la diferencia se multiplica por la semisuma de los números combinatorios o el número combinatorio se multiplica por la semisuma de las diferencias (combinaciones lineales de (2)).

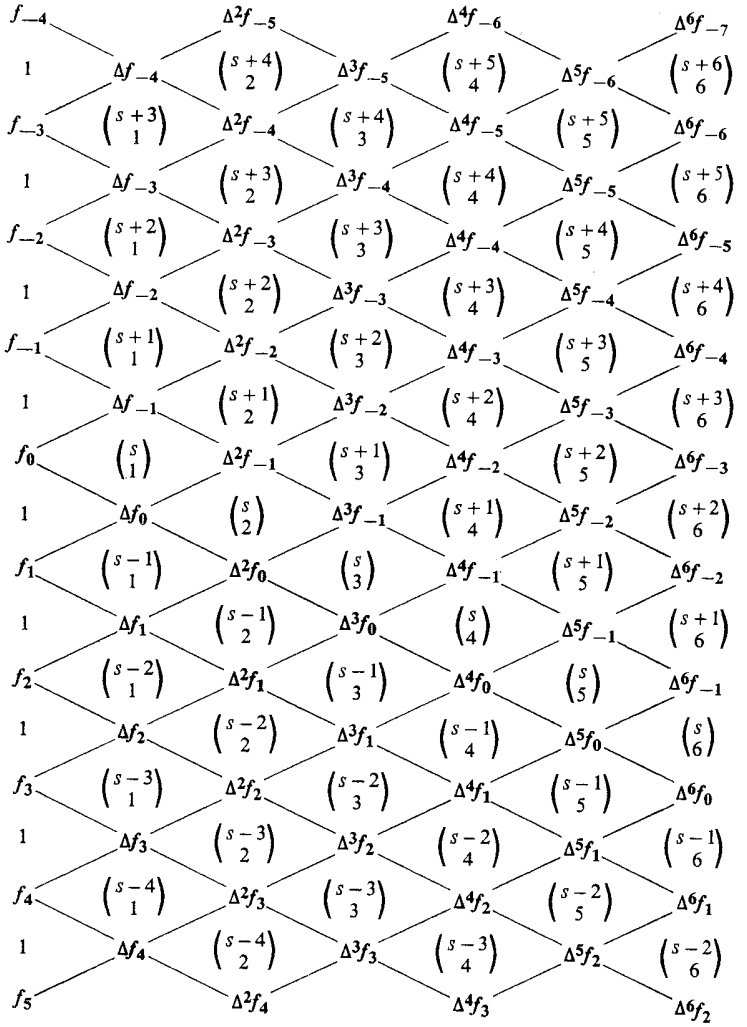


Figura. Diagrama Romboidal para la interpolación en datos regulares.
$\searrow$ Progresivo $\nearrow$ Regresivo $Zig-Zag$ Gauss $\rightarrow$ Stirling $y_0 \rightarrow$ Bessel $y_0 y_1$.

Los números combinatorios en las caras de los rombos cumplen con la regla SE+PC=NE (SE=sureste, PC=central, NE=noreste). Las diferencias en los vértices de los rombos cumplen con la regla SW+VC=NW (SW=suroeste, VC=central, NW=noroeste). Para resumir algorítmicamente $\binom{s}{k+1} + \binom{s}{k} = \binom{s+1}{k+1}$, y $\Delta^{k-1} f_{c+1} + \Delta^k f_c = \Delta^{k-1} f_c$.

Sea la primera columna a la izquierda de los valores funcionales del diagrama romboidal la base de un triángulo isosceles, cuyos lados iguales son la diagonal descendente y diagonal ascendente de diferencias, que se intersectan en el vértice a la derecha de mayor orden. Siempre que se comience en la base y se haga cualquier recorrido del diagrama romboidal, sin salir del mencionado triángulo isósceles, llegando al vértice de mayor orden, el polinomio de interpolación será siempre el mismo.

• Polinomios Regresivos

Se hace una recorrido del diagrama romboidal siguiendo una diagonal ascendente se obtiene el *polinomio de Newton-Gregory regresivo*. Son progresivos si se sigue un recorrido descendente como en la sección 1.3.2.

• Polinomios de Gauss

Si se sigue un recorrido del diagrama romboidal en zig-zag se denomina *polinomios de Gauss*. Progresivo si comienza subiendo o regresivo si comienza bajando.

• Polinomios de Stirling

Si se hace un recorrido horizontal del diagrama romboidal comenzando en y_0 se denomina *polinomio de Stirling*.

• Polinomios de Bessel

Si se hace un recorrido horizontal del diagrama romboidal comenzando entre y_0 y y_1 se denomina *polinomios de Bessel*.

1.4. CRITERIOS DE INTERPOLACION

Se ha desarrollado un algoritmo de interpolación usando los polinomios de Newton en diferencias divididas. Para hacer eficientemente la interpolación se han usado dos criterios que hacen de ésta la más consistente posible con los puntos discretos dados. Estos criterios son el de *Simetría* y el de *Monotonía*. Estos criterios, aplicados en conjunto, permiten determinar el grado del polinomio óptimo a utilizar durante una interpolación.

A continuación se describen los dos criterios utilizados en el algoritmo: simetría y monotonía. Luego se formula como se acoplan en el algorithm.

1.4.1. Simetría

El criterio de la *simetría* consiste en escoger la distribución de puntos lo más simétricamente posible, alrededor de donde se desee interpolar. Esto se puede hacer de dos maneras: mediante el número de puntos o mediante la distancia de influencia, a uno y otro lado del punto donde se vaya a interpolar. En el caso de intervalos irregulares, la segunda opción se convierte en un criterio de *Proximidad*. En el caso de intervalos regulares una de las formas implica a la otra. En cualquier caso, el número de puntos próximos lo determina el criterio de monotonía descrito abajo. En los extremos del intervalo que contiene a los puntos, a veces es imposible seguir el criterio de simetría de forma estricta, y entonces se hace necesario en su lugar seguir el criterio de proximidad, si se desea alcanzar un mayor orden de monotonía como se explica abajo.

El criterio de simetría tiene otra ventaja. Por ejemplo, en los esquemas de diferencias finitas centradas las formulaciones presentan un menor error, que cuando no lo son, usando inclusive el mismo número de puntos (En el método de las diferencias finitas se prefiere más al criterio de monotonía que al criterio de simetría, sobre todo en los bordes del dominio y en los términos no lineales).

1.4.2. Monotonía

El criterio de la *monotonía* se basa en la definición de monotonía de una función: Una función se dice que es monótona hasta el orden m, en un determinado intervalo, si todas sus derivadas de hasta dicho orden conservan siempre su signo en dicho intervalo. En otras palabras, una función continua $f(x)$ es monótona de orden m en un intervalo $[a, b]$, si

$$\begin{aligned} f'(x) \neq 0 \qquad f''(x) \neq 0 \qquad f'''(x) \neq 0 \quad \cdots \quad f^{(m)}(x) \neq 0 \qquad &\text{para todo} \quad x \in [a, b] \\ f^{(m+1)}(x) = 0 \qquad &\text{para algún} \quad x \in [a, b] \end{aligned} \tag{16}$$

En el ejemplo mostrado en (9), la parábola tiene monotonía de orden 2 en los intervalos $(-\infty, v)$ y (v, ∞), separadamente, donde $v = \frac{1}{2}\{\, a + b - f[a, b]/f[a, b, c] \,\}$ es la localización del vértice de dicha parábola.

Las diferencias divididas son proporcionales a las derivadas en su entorno, tal como lo indica la siguiente relación reflejada en 1.1.(7.*b*)

$$f[x_0, x_1, x_2, \ldots, x_{n-1}, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!} \qquad \xi \in [x_0, x_1, \ldots, x_{n-1}, x_n, x] \tag{17}$$

Por ello, el criterio de monotonía implica escoger hasta el mayor orden en las diferencias divididas que tengan igual signo por columna en el diagrama romboidal. La última diferencia dividida a evita a partir de aquí, junto con el último punto que la originó, deberá tener signo opuesto a todas las demás diferencias divididas vecinas del mismo orden (misma columna). Esto significa que el criterio de la monotonía acepta polinomios de interpolación hasta el grado m. La falta de monotonía en las interpolaciones implica que pueden producirse oscilaciones indeseables de la función alrededor o entre los puntos dados. Como el último punto agregado es el más lejano, por el criterio de la simetría, no existe inconveniente en dejarlo (no recomendado), ya que los cálculos están hechos y son del último orden en el error. Entre más parecidas sean la monotonía de la función discreta y el grado del polinomio de interpolación, en esa misma medida la interpolación será más consistente (Regla seguida de forma estricta en el método de diferencias finitas).

1.4.3. Algorítmo

Los criterios de simetría y monotonía se complementan para indicar cuales puntos y el número de ellos se deben usar en la interpolación. En cualquier caso, el grado del polinomio será siempre una unidad menor que el número de puntos usados. El algoritmo se resume de la siguiente manera: se escogen los puntos más cercanos al punto donde se desee interpolar, en un número (distancia) simétrico (próxima), uno a uno (calculando cada vez las diferencias divididas de la diagonal incluido el vértice), hasta que dicho número de puntos, reflejado en las diferencias divididas, conserve el máximo orden posible de monotonía del polinomio de interpolación igual que el de la función discreta.

El algoritmo antes explicado puede usarse para hacer interpolaciones en una o en varias dimensiones. También permite la interpolación sin necesidad de pre-ordenar los puntos usados o pre escoger su número. En varias dimensiones lo único que se exige es que los valores de las funciones sean siempre para los mismos y todos los puntos discretos en cada dimensión. El algoritmo tampoco necesita escoger un grado del polinomio anticipadamente, durante el proceso de la interpolación. El algoritmo decide el grado del polinomio óptimo que garantice satisfacer los criterios de simetría y monotonía.

1.5. INTERPOLACION ESPACIAL

Las interpolaciones con funciones dependientes de más de una variable se hacen mediante el mismo algoritmo de interpolación en una variable, repetido varias veces en curvas (dos dimensiones) o superficies paralelas (tres dimensiones), y a su vez, las interpolaciones en las superficies paralelas se realizan como en funciones de dos variables.

1.5.1. Dos Dimensiones

El algoritmo para la interpolación en dos dimensiones para la función discreta $z_{ij} = z(x_i, y_j)$, con $i = 0, \ldots, n_x - 1$ en x y $j = 0, \ldots, n_y - 1$ en y, se describe de forma estructurada a continuación:

- Para $i = 0, \ldots, n_x - 1$
- Para $j = 0, \ldots, n_y - 1$
- Se asigna $\eta_i(y_j) = z_{ij} = z(x_i, y_j)$
- Siguiente j
- Para cada curva i se interpola en el punto y_* con los valores de $\eta_i(y_j)$, lo que dan los valores interpolados $\zeta_i = \zeta(x_i) = z(x_i, y_*)$ de la función $z(x, y)$ en la curva que pasa por y_* y está parametrizada con los valores x_i.
- Siguiente i.

- Finalmente se interpola en el punto x_* con los valores $\zeta_i = \zeta(x_i)$, lo que da como resultado el valor deseado $z_* = z(x_*, y_*)$.

1.5.2. Tres Dimensiones

El algoritmo para la interpolación en tres dimensiones para la función discreta $t_{ijk} = t(x_i, y_j, z_k)$, con $i = 0, \ldots, n_x - 1$ en x, $j = 0, \ldots, n_y - 1$ en y y $k = 0, \ldots, n_z - 1$ en z, se describe de forma estructurada a continuación:

- Para $k = 0, \ldots, n_z - 1$.
- Para $j = 0, \ldots, n_y - 1$.
- Para $i = 0, \ldots, n_x - 1$.
- Se asigna $\eta_k(x_i, y_j) = t_{ijk} = t(x_i, y_j, z_k)$.

  Siguiente i.
- Siguiente j.
- Para cada superficie k se interpola en dos dimensiones en el punto (x_*, y_*) con los valores de $\eta_k(x_i, y_j)$, lo que dan los valores interpolados $\zeta_k = \zeta(z_k) = t(x_*, y_*, z_k)$ de la función $t(x, y, z)$ en la curva que pasa por (x_*, y_*) y está parametrizada con los valores z_k.
- Siguiente k.
- Finalmente se interpola en el punto z_* con los valores $\zeta_k = \zeta(z_k)$, lo que da como resultado el valor deseado $t_* = t(x_*, y_*, z_*)$.

Para mayores dimensiones se sigue la misma práctica de apoyar el algoritmo en algoritmos para dimensiones menores.

1.6. TRAZADORES

Dado un comjunto de n puntos (x_i, y_i), $i = 1, 2, 3, \ldots, n$, se denominan *trazadores* (splines), al conjunto de $n - 1$ polinomios de orden m

$$y = y_i + \sum_{j=1}^{m} a_{ij} \, (x - x_i)^j \qquad\qquad x \in [x_i, x_{i+1}) \tag{1}$$

con coeficientes a_{ij}, tales que garanticen la continuidad de la función y y de sus derivadas y', y'', y''',..., $y^(m-1)$ en todo el dominio $[x_1, x_n]$.

1.6.1. Trazadores Rectilíneos ($m = 1$)

Sea

$$y = y_i + b_i \, (x - x_i) \tag{2}$$

un polinomio de primer orden que pasa por los puntos (x_i, y_i) y (x_{i+1}, y_{i+1}). Si tenemos en cuenta que el tamaño del intervalo $[x_i, x_{i+1})$ es $h_i = x_{i+1} - x_i$, entonces

$$y_{i+1} = y_i + b_i h_i \tag{3}$$

De esta expresión se obtiene que

$$b_i = \frac{y_{i+1} - y_i}{h_i} \tag{4}$$

Es obvio que el comjunto de trazadores rectilíneos hallados de esta forma garantizan la continuidad de la función y en todo el dominio $[x_1, x_n]$, lo cual está de acuerdo con la definición de los trazadores polinómicos de orden $m = 1$. La función y' de primera derivada representa yna función escalonada que por supuesto no es continua.

1.6.2. Trazadores Parabólicos ($m = 2$)

Sea

$$y = y_i + b_i\,(x - x_i) + c_i\,(x - x_i)^2 \tag{5}$$

un polinomio de segundo grado, que pasa por los puntos (x_i, y_i) y (x_{i+1}, y_{i+1}).

Sean

$$y' = b_i + 2\,c_i\,(x - x_i) \qquad\qquad y'' = 2\,c_i \tag{6}$$

la primera y segunda derivadas del polinomio respectivo. Si tenemos en cuenta que el tamaño del intervalo $[x_i, x_{i+1})$ es $h_i = x_{i+1} - x_i$ y llamamos a p_i a la primera derivada y' evaluada en x_i, esto es

$$h_i = x_{i+1} - x_i \qquad\qquad p_i = y'_i \tag{7}$$

entonces, para que el polinomio parabólico pase por los puntos (x_i, y_i) y (x_{i+1}, y_{i+1}), los coeficientes b_i y c_i de ben cumplir con las siguientes condiciones

$$y_{i+1} = y_i + b_i h_i + c_i h_i^2 \qquad\qquad p_i = b_i \qquad\qquad p_{i+1} = b_i + 2\,c_i h_i \tag{8}$$

De estas relaciones se obtiene que

$$b_i = p_i \qquad\qquad c_i = \frac{p_{i+1} - p_i}{2\,h_i} \tag{9}$$

Si ahora substituimos b_i y c_i en función de los P_i en la expresión de y_{i+1}, queda

$$y_{i+1} = y_i + p_i h_i + \left(\frac{p_{i+1} - p_i}{2\,h_i}\right) h_i^2 \tag{10}$$

Rorganizando esta ecuación, finalmente se obtiene

$$p_i + p_{i+1} = 2\left(\frac{y_{i+1} - y_i}{h_i}\right) \tag{11}$$

Esta ecuación se puede aplicar sólo para los puntos x_2, x_3, …, x_{n-1}. Para los puntos extremos x_1 y x_n se puede asumir cualquiera de las siguientes condiciones:

a:) Los polinomios en los intervalos 1 y n son rectas

$$p_1 - p_2 = 0 \qquad\qquad p_{n-1} - p_n = 0 \tag{12}$$

b:) Los polinomios en los intervalos 1 y n son parábolas

$$\frac{p_2 - p_1}{h_1} = \frac{p_3 - p_2}{h_2} \qquad\qquad \frac{p_n - p_{n-1}}{h_{n-1}} = \frac{p_{n-1} - p_{n-2}}{h_{n-2}}$$

$$-h_2 p_1 + (h_1 + h_2)\,p_2 - h_1 p_3 = 0 \qquad\qquad -h_{n-1} p_{n-2} + (h_{n-2} + h_{n-1})\,p_{n-1} - h_{n-2} p_n = 0 \tag{13}$$

Con todas estas ecuaciones se obtiene el siguiente sitemas de n ecuaciones lineales con n incognitas p_i $i = 1, 2, 3, \ldots, n$

$$\begin{bmatrix} U_1 & V_1 & W_1 & & & & \\ & 1 & 1 & & & & \\ & & 1 & 1 & & & \\ & & & \ddots & \ddots & & \\ & & & & \ddots & \ddots & \\ & & & & & 1 & 1 \\ & & & & U_n & V_n & W_n \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ \vdots \\ p_{n-1} \\ p_n \end{bmatrix} = 2 \begin{bmatrix} T_1 \\ \frac{y_3 - y_2}{h_2} \\ \frac{y_4 - y_3}{h_3} \\ \vdots \\ \vdots \\ \frac{y_n - y_{n-1}}{h_{n-1}} \\ T_n \end{bmatrix} \tag{14}$$

donde

a:) $U_1 = 1 \quad V_1 = -1 \quad W_1 = 0 \quad T_1 = 0$

$U_n = 0 \quad V_n = -1 \quad W_n = 1 \quad T_n = 0$

b:) $U_1 = -h_2 \quad V_1 = h_1 + h_2 \quad W_1 = -h_1 \quad T_1 = 0$

$U_n = -h_{n-1} \quad V_n = h_{n-2} + h_{n-1} \quad W_n = -h_{n-2} \quad T_n = 0$

Aplicando un proceso de eliminación, se puede lograr eliminar algunos términos y así convertir el sistema de ecuaciones en bidiagonal. Como se dabe, un sistema de ecuaciones así puede ser resuelto por substitución progresiva o regresiva. Esto significa que sólamente puede aplicarse una de las condiciones nombradas anteriormente para un extremo y el otro extremo debe quedar libre, sin condición. Una vez halladas las incógnitas p_i, se pueden calcular los coeficientes de los trazadores usando la expresión (9). De acuerdo a esto los primeros y últimos coeficientes de la matriz cambian a

a:) $U_1 = 1 \quad V_1 = 0 \quad W_1 = 0 \quad T_1 = \frac{y_2 - y_1}{2\,h_1}$

$U_n = 0 \quad V_n = 0 \quad W_n = 1 \quad T_n = \frac{y_n - y_{n-1}}{2\,h_{n-1}}$

b:) $U_1 = 1 \quad V_1 = 0 \quad W_1 = 0 \quad T_1 = \frac{(2h_1+h_2)\frac{y_2-y_1}{h_1} - h_1\,\frac{y_3-y_2}{h_2}}{2\,(h_1+h_2)}$

$U_n = 0 \quad V_n = 0 \quad W_n = 1 \quad T_n = \frac{(2h_{n-1}+h_{n-2})\frac{y_n-y_{n-1}}{h_{n-1}} - h_{n-1}\,\frac{y_{n-1}-y_{n-2}}{h_{n-2}}}{2\,(h_{n-1}+h_{n-2})}$ Es obvio que el conjunto de trazadores parabólicos hallados de esta forma garantizan la continuidad de la función y y su primera derivada y' en todo el dominio $[x_1, x_n]$, lo cual está de acuerdo con la definición de los trazadores polinómicos de orden $m = 2$. La función y'' de segunda derivada representa una función escalonada que por supuesto no es continua.

1.6.3. Trazadores Cúbicos ($m = 3$)

La interpolación numérica no debe ser solo vista como una herramienta para cálculo sino también como una herramienta para el dibujante moderno, ya que es muy útil al momento de desarrollar algoritmos para el dibujo asistido por computador.

Dado un conjunto de puntos (x_i, f_i) se desea construir la curva de la función $f(x)$ en el intervalo (x_1, x_n) por lo cual se hace necesario obtener puntos adicionales para una mejor representación de la función $f(x)$. Una de las metodologías existentes es la de utilizar polinomios a trozos en cada sub-intervalo garantizando continuidad de la función y sus derivadas en los extremos de los sub-intervalos, estas expresiones son denominadas *curva especiales*.

Dado un sub-intervalo $[x_i, x_{i+1})$ se propone un polinomio cúbico de la forma

$$f(x) = y_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{15}$$

donde las contantes b_i, c_i, d_i son validas unicamente en el sub-intervalo $[x_i, x_{i+1})$. El polinomio de tercer grado pasa por los puntos (x_i, y_i) y (x_{i+1}, y_{i+1}).

Sean

$$\begin{aligned} y' &= b_i + 2\,c_i\,(x - x_i) + 3\,d_i\,(x - x_i)^2 \\ y'' &= 2\,c_i + 6\,d_i\,(x - x_i) \\ y''' &= 6\,d_i \end{aligned} \tag{16}$$

la primera, segunda y tercera derivadas del polinomio respectivo.

Si tenemos en cuenta que el tama no del intervalo $[x_i, x_{i+1})$ es $h_i = x_{i+1} - x_i$ y llamamos p_i y s_i a la primera derivada y' y a la segunda derivada y'', respectivamente evaluadas en x_i, esto es

$$h_i = x_{i+1} - x_i \qquad p_i = y'_i \qquad s_i = y''_i \tag{17}$$

entonces, para que el polinomio cúbico pase por los puntos (x_i, y_i) y (x_{i+1}, y_{i+1}), dado que el polinomio cúbico admite continuidad en el valor de la función, en su primera y segunda derivadas, los coeficientes b_i, c_i y d_i deben cumplir con las siguientes condiciones

$$\begin{aligned} y_{i+1} &= y_i + b_i h_i + c_i h_i^2 + d_i h_i^3 \\ p_i &= b_i \qquad\qquad p_{i+1} = b_i + 2\,c_i h_i + 3\,d_i h_i^2 \\ s_i &= 2\,c_i \qquad\qquad s_{i+1} = 2\,c_i + 6\,d_i h_i \end{aligned} \tag{18}$$

De estas relaciones se obtienen que

$$b_i = \frac{y_{i+1} - y_i}{h_i} - \frac{h_i\,(2\,s_i + s_{i+1})}{6} \qquad c_i = \frac{s_i}{2} \qquad d_i = \frac{s_{i+1} - s_i}{6\,h_i} \tag{19}$$

La derivadas en el punto x_i usando un polinomio válido en el intervalo $[x_i, x_{i+1})$ y usando un polinomio válido para el intervalo $[x_{i-1}, x_i)$, deben ser las mismas

$$p_i = b_i = b_{i-1} + 2\,c_{i-1} h_{i-1} + 3\,d_{i-1} h_{i-1}^2 \tag{20}$$

Si ahora substituimos b_i, b_{i-1}, c_{i-1} y d_{i-1} en función de los y_i y s_i en la expresión de p_i anterior, al imponer la condición de continuidad de la primera derivada se, obtiene la siguiente expresión

$$\frac{y_{i+1} - y_i}{h_i} - \frac{h_i(2s_i + s_{i+1})}{6} = \frac{y_i - y_{i-1}}{h_{i-1}} - \frac{h_{i-1}(2s_{i-1} + s_i)}{6} + 2\left(\frac{s_{i-1}}{2}\right) h_{i-1} + 3\left(\frac{s_i - s_{i-1}}{6\,h_{i-1}}\right) h_{i-1}^2 \tag{21}$$

Reorganizando esta ecuación, finalmente se obtiene

$$h_{i-1} s_{i-1} + 2\,(h_{i-1} + h_i)\,s_i + h_i s_{i+1} = 6\left(\frac{y_{i+1} + y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}\right) \tag{22}$$

Esta ecuación se puede aplicar sólo para los puntos x_2, x_3, ..., x_{n-1}. Para los puntos x_1 y x_n se pueden asumir cualquiera de las siguientes condiciones:

a:) Los polinomios en los intervalos 1 y n se empalman con rectas, es decir, los extremos son puntos de inflexión

$$s_1 = 0 \qquad\qquad s_n = 0$$

b:) Los polinomios en los intervalos 1 y n son parábolas ($d_1 = d_{n-1} = 0$)

$$s_1 - s_2 = 0 \qquad\qquad s_{n-1} - s_n = 0$$

c:) Los polinomios en los intervalos 1 y n son cúbicas "condición natural" ($d_1 = d_2$, $d_{n-2} = d_{n-1}$)

$$\frac{s_2 - s_1}{h_1} = \frac{s_3 - s_2}{h_2} \qquad\qquad \frac{s_n - s_{n-1}}{h_{n-1}} = \frac{s_{n-1} - s_{n-2}}{h_{n-2}}$$

$$-h_2\,s_1 + (h_1 + h_2)\,s_2 - h_1\,s_3 = 0 \qquad\qquad -h_{n-1}\,s_{n-2} + (h_{n-2} + h_{n-1})\,s_{n-1} - h_{n-2}\,s_n = 0$$

Estas expresiones representan un sistema de ecuaciones lineales tridiagonal para las incognitas s_i, $i = 1, 2, \ldots, n$, y el mismo puede ser resuelto utilizando el algoritmo de Thomas.

El sistema de ecuaciones lineales planteado se muestra a continuación

$$\begin{bmatrix} U_1 & V_1 & W_1 & & & & & \\ h_1 & 2(h_1+h_2) & h_2 & & & & & \\ & h_2 & 2(h_2+h_3) & h_3 & & & & \\ & & & \ddots & \ddots & & & \\ & & & & \ddots & \ddots & & \\ & & & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ & & & & U_n & V_n & W_n \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \\ \vdots \\ s_{n-1} \\ s_n \end{bmatrix} = 6 \begin{bmatrix} T_1 \\ \frac{y_3-y_2}{h_2} - \frac{y_2-y_1}{h_1} \\ \frac{y_4-y_3}{h_3} - \frac{y_3-y_2}{h_2} \\ \vdots \\ \vdots \\ \frac{y_n-y_{n-1}}{h_{n-1}} - \frac{y_{n-1}-y_{n-2}}{h_{n-2}} \\ T_n \end{bmatrix} \tag{23}$$

donde

a:) $U_1 = 1 \quad V_1 = 0 \quad W_1 = 0 \quad T_1 = 0$
$U_n = 0 \quad V_n = 0 \quad W_n = 1 \quad T_n = 0$

b:) $U_1 = 1 \quad V_1 = -1 \quad W_1 = 0 \quad T_1 = 0$
$U_n = 0 \quad V_n = -1 \quad W_n = 1 \quad T_n = 0$

c:) $U_1 = -h_2 \quad V_1 = h_1 + h_2 \quad W_1 = -h_1 \quad T_1 = 0$
$U_n = -h_{n-1} \quad V_n = h_{n-2} + h_{n-1} \quad W_n = -h_{n-2} \quad T_n = 0$

Aplicando un proceso de eliminación, see puede lograr eliminar algunos términos y así convertir el sistema de ecuaciones en tridiagonal. Como se sabe un sistema de ecuaciones así puede ser resuelto utilizando el algoritmo de Thomas (sección II.1.1.7). Una vez halladas las incógnitas s_i, se pueden calcular los coeficientes de los trazadores cúbicos con las expresiones (19).

De acuerdo a esto entonces los coeficientes cambian a

c:) $U_1 = h_1 - h_2 \quad V_1 = 2h_1 + h_2 \quad W_1 = 0 \quad T_1 = \frac{h_1}{h_1+h_2}\left(\frac{y_3-y_2}{h_2} - \frac{y_2-y_1}{h_1}\right)$

$U_n = 0 \quad V_n = h_{n-2} + 2h_{n-1} \quad W_n = h_{n-1} - h_{n-2}$

$$T_n = \frac{h_{n-1}}{h_{n-2}+h_{n-1}}\left(\frac{y_n-y_{n-1}}{h_{n-1}} - \frac{y_{n-1}-y_{n-2}}{h_{n-2}}\right)$$

Es obvio que el conjunto de trazadores cúbicos hallados de esta forma garantizan la continuidad de la función y, sus primeras derivadas y' y sus segundas derivadas y'' en todo el dominio $[x_1, x_n]$, lo cual está de acuerdo con la definición de los trazadores polinómicos de grado $m = 3$. La función y''' tercera derivada representa una función escalonada que por supuesto no es continua.

Si ocurre que $h_1 = h_2$, entonces $U_1 = 0$ y ya no se puede aplicar el algoritmo de Thomas. En este caso, la ecuación a aplicar es la obtenida haciendo eliminaciones de términos con las primera tres ecuaciones y evaluando las dos primeras segundas derivadas. También se puede aplicar lo mismo para los últimos puntos. Basándonos en esto se obtiene

c:) $U_1 = -h_1 \quad V_1 = h_1 \quad W_1 = 0$

$$T_1 = \frac{h_1^2}{h_1+h_2+h_3}\left(\frac{\frac{y_4-y_3}{h_3} - \frac{y_3-y_2}{h_2}}{h_2+h_3} - \frac{\frac{y_3-y_2}{h_2} - \frac{y_2-y_1}{h_1}}{h_1+h_2}\right)$$

$U_n = 0 \quad V_n = h_{n-1} \quad W_n = -h_{n-1}$

$$T_n = \frac{h_{n-1}^2}{h_{n-3}+h_{n-2}+h_{n-1}}\left(\frac{\frac{y_n-y_{n-1}}{h_{n-1}} - \frac{y_{n-1}-y_{n-2}}{h_{n-2}}}{h_{n-2}+h_{n-1}} - \frac{\frac{y_{n-1}-y_{n-2}}{h_{n-2}} - \frac{y_{n-2}-y_{n-3}}{h_{n-3}}}{h_{n-3}+h_{n-2}}\right)$$

EJEMPLO:

Determine el "spline cubico natural" que interpola a la función $f(x)$ en el intervalo $[0.25, 0.53]$, a partir de la siguiente tabla de datos

Datos

x_i	0.25	0.30	0.39	0.45	0.53
$f(x_i)$	0.5000	0.5477	0.6245	0.6708	0.7280

De los datos de la tabla se pueden determinar los valores de los h_i

$$h_1 = 0.05 \qquad h_2 = 0.09 \qquad h_3 = 0.06 \qquad h_4 = 0.08$$

Construyendo el sistema de ecuaciones para los s_i con i=2,3,4, recordando la condición natural en los extremos, se obtiene

$$\begin{aligned} 0.28s_2 + 0.09s_3 &= -0.604 \\ 0.09s_2 + 0.30s_3 + 0.06s_4 &= -0.490 \\ 0.06s_3 + 0.28s_4 &= -0.340 \end{aligned}$$

cuya solución es:

$$s_2 = -1.8806 \qquad s_3 = -0.8226 \qquad s_4 = -1.0261$$

1.7. DERIVACION

Las derivadas de cualquier orden se calculan numéricamente utilizando los polinomios de interpolación y luego derivándolo según requerimiento. Escogiento los valores funcionales diferentes y diferentes órdenes en los polinomio de interpolación, se han generado la siguientes fórmulas para las derivadas, siguiendo las reglas dictadas al final de la sección 1.3.3.

Fórmulas para la Primera Derivada

$$\begin{aligned} f'(x_0) &= \frac{1}{h}(f_1 - f_0) + O(h) \\ f'(x_0) &= \frac{1}{2h}(f_1 - f_{-1}) + O(h^2) && \text{(Diferencia Central)} \\ f'(x_0) &= \frac{1}{2h}(-f_2 + 4f_1 - 3f_0) + O(h^2) \\ f'(x_0) &= \frac{1}{12h}(-f_2 + 8f_1 - 8f_{-1} + f_{-2}) + O(h^4) && \text{(Diferencia Central)} \end{aligned} \tag{1}$$

Fórmulas para la Segunda Derivada

$$\begin{aligned} f''(x_0) &= \frac{1}{h^2}(f_2 - 2f_1 + f_0) + O(h) \\ f''(x_0) &= \frac{1}{h^2}(f_1 - 2f_0 + f_{-1}) + O(h^2) && \text{(Diferencia Central)} \\ f''(x_0) &= \frac{1}{h^2}(-f_3 + 4f_2 - 5f_1 + 2f_0) + O(h^2) \\ f''(x_0) &= \frac{1}{12h^2}(-f_2 + 16f_1 - 30f_0 + 16f_{-1} - f_{-2}) + O(h^4) && \text{(Diferencia Central)} \end{aligned} \tag{2}$$

Fórmulas para la Tercera Derivada

$$\begin{aligned} f'''(x_0) &= \frac{1}{h^3}(f_3 - 3f_3 + 3f_1 - f_0) + O(h) \\ f'''(x_0) &= \frac{1}{2h^3}(f_2 - 2f_1 + 2f_{-1} - f_{-2}) + O(h^2) && \text{(Diferencia Promedio)} \end{aligned} \tag{3}$$

Fórmulas para la Cuarta Derivada

$$
\begin{aligned}
f^{iv}(x_0) &= \frac{1}{h^4}\,(f_4 - 4\,f_3 + 6\,f_2 - 4\,f_1 + f_0) + O(h) \\
f^{iv}(x_0) &= \frac{1}{h^4}\,(f_2 - 4\,f_1 + 6\,f_0 - 4\,f_{-1} + f_{-2}) + O(h^2) \qquad \text{(Diferencia Central)}
\end{aligned}
\tag{4}
$$

Para intervalos irregulares con puntos no ordenados se utiliza la expresión 1.1.(6), derivada y evaluada en $x = x_0$, para la primera derivada, lo cual da

$$
P_n'(x_0) = \sum_{k=1}^{n} \prod_{j=1}^{k-1} (x - x_j)\; f[x_0, x_1, x_2, \ldots, x_{k-1}, x_k] \tag{5}
$$

Todos los términos que contienen $x - x_0$ al derivar, cuando se evalúa en $x = x_0$, se anulan. El resultado es la misma expresión 1.1.(6), sin el primer término y sin el primer factor en la productoria. Haciendo el uso de esta ecuación (5), cambiando cada vez el punto designado como x_0, se tiene una tabla de valores de las primeras derivadas en distintos puntos, tabla con la cual se puede interpolar donde se desee. Si se aplica este mismo procedimiento a los valores de primeras derivadas, se obtienen los valores de las segundas derivadas, y así sucesivamente. Cuando el límite superior de una productoria es menor que el límite inferior, como ocurre con el primer término de (5), el resultado de dicha productoria es la unidad.

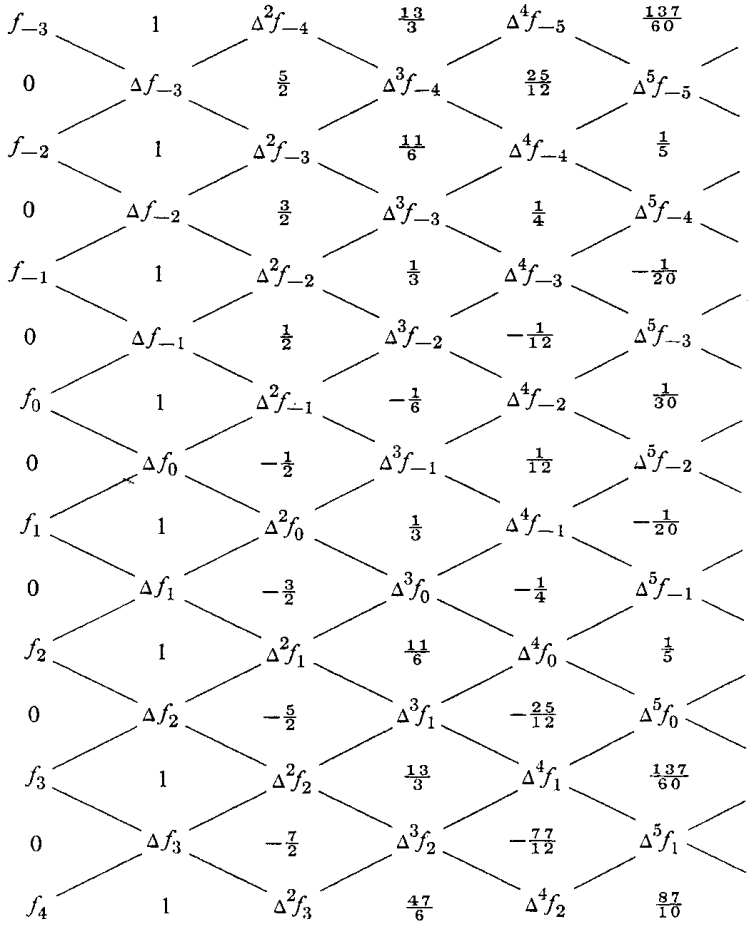


Figura 1. Diagrama Romboidal de la primera derivada. Multiplicar por $1/h$.

La figura 1 anterior se obtuvo de derivar respecto a s las caras del diagrama romboidal (números combianatorios) de la sección 1.3.3, y luego evaluarla en $s = 0$. Por eso hay que multiplicar por $1/h$ para obtener $f'(x_0)$ ($dx = h\,ds$), cualesquiera de los resultados en su aplicación (reglas al final de la sección 1.3.3).

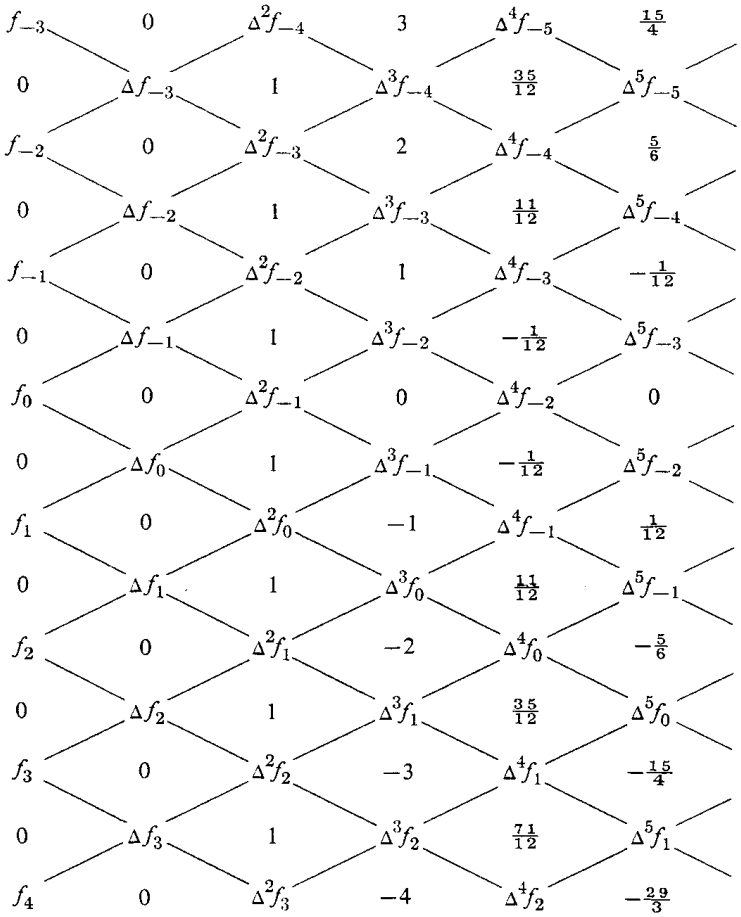


Figura 2. Diagrama Romboidal de la segunda derivada. Multiplicar por $1/h^2$.

La figura 2 anterior se obtuvo de derivar doblemente respecto a s las caras del diagrama romboidal (números combianatorios) de la sección 1.3.3, y luego evaluarla en $s = 0$. Por eso hay que multiplicar por $1/h^2$ para obtener $f''(x_0)$ ($dx = h\,ds$), cualesquiera de los resultados en su aplicación (reglas al final de la sección 1.3.3).

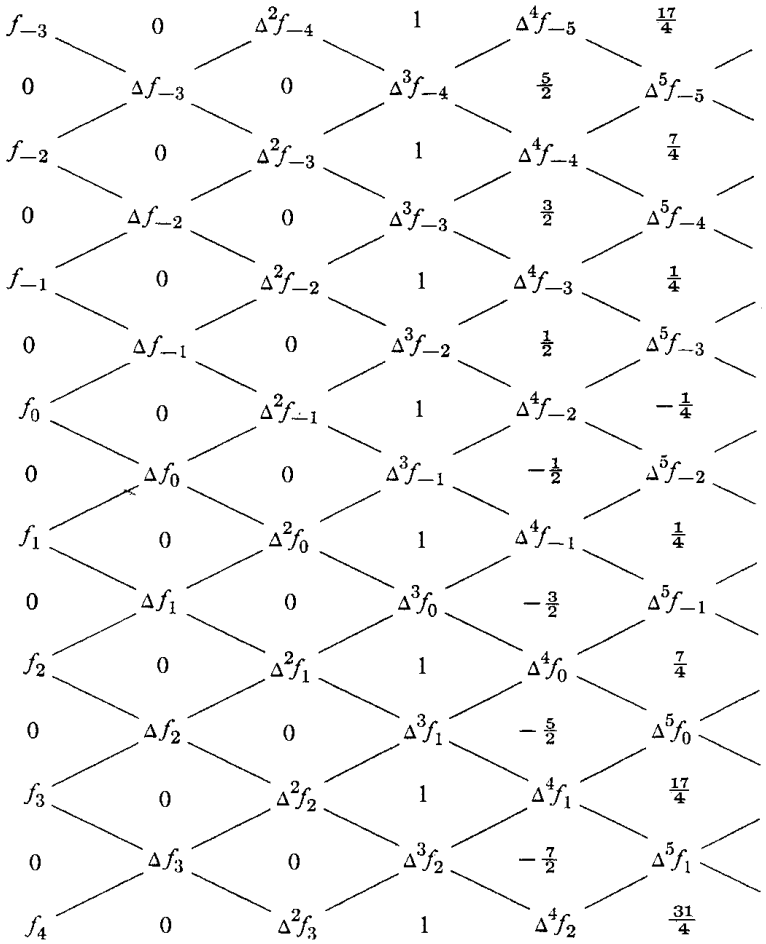


Figura 3. Diagrama Romboidal de la tercera derivada. Multiplicar por $1/h^3$.

La figura 3 anterior se obtuvo de derivar tres veces respecto a s las caras del diagrama romboidal (números combianatorios) de la sección 1.3.3, y luego evaluarla en $s = 0$. Por eso hay que multiplicar por $1/h^3$ para obtener $f'''(x_0)$ ($dx = h\,ds$), cualesquiera de los resultados en su aplicación (reglas al final de la sección 1.3.3).

2. INTEGRACION

La integración de funciones es una operación matemática de mucha importancia, y al estudiante de cálculo le toma tiempo en aprender a dominar las distintas tecnicas analíticas para resolverlas.

Con mucha frecuencia es necesario integrar a función que es conocida en forma tabular, por ejemplo un conjunto de datos experimentales. Los métodos numéricos nos permiten llevar a cabo esta operación. Adicionalmente, la integración analítica de funciones no es un proceso de fácil manejo para el computador, por lo cual el uso de técnicas numéricas para su evaluación son necesarias.

Entre las técnicas numéricas más conocidas están las fórmulas de Newton-Cotes, la fórmulas de la Cuadratura de Gauss, y distintas variantes de estas.

2.1. DATOS REGULARES

Cuando los datos son regulares, estos están ordenados y la distancia entre dos puntos en x es denotada $h = x_{i+1} - x_i$, $i = 1, 2, \ldots, N$

2.1.1. Fórmulas de Newton-Cotes

Al momento de evaluar una integral sobre un conjunto de datos discretos, dados en forma tabular, se hace necesario desarrollar métodos particulares. Al conjunto de $N+1$ puntos se le subdivide en grupos de $n+1$ puntos ($n < N$) cada uno, siendo el extremo final de cada grupo el extremo comienzo del siguiente. Entre los métodos para evaluar este tipo de integrales estan las Fórmulas de Newton-Cotes, entre las cuales se agrupan a las conocidas fórmulas del trapezoide y de Simpson.

Para cada grupo de $n+1$ puntos se utiliza un polinomio de interpolaciń $P_n(x)$ de grado n, con el cual se calcula un estimado de la integral. El polinomio que en este caso es el más apropiado, es el polinomio de Lagrange (sección 1.2). Con esto se generan las fórmulas de Newton-Cotes. Cuando aplicamos algunas de estas "Fórmulas" a un conjunto grandes de $N+1$ puntos le denominamos "La Regla". Pueden usarse combinaciones de fórmulas cuando el número de puntos así lo amerite.

Usando los polinomios de Lagrange

$$f(x) = P_n(x) + R(x) \qquad P_n(x) = \sum_{i=0}^{n} L_i(x)\, f(x_i) \tag{1}$$

las fórmulas de Newton-Cotes tienen la forma

$$\int_{x_0}^{x_n} f(x)\, dx = \int_{x_0}^{x_n} P_n(x)\, dx + \int_{x_0}^{x_n} R(x)\, dx = I^n + E^n \tag{2}$$

donde

$$I^n = \int_{x_0}^{x_n} P_n(x)\, dx = \int_{x_0}^{x_n} \left[\sum_{i=0}^{n} L_i(x)\, f(x_i)\right] dx = h \sum_{i=0}^{n} C_i^n\, f(x_i) \qquad C_i^n = \frac{1}{h}\int_{x_0}^{x_n} L_i(x)\, dx \tag{3}$$

$$E^n = \int_{x_0}^{x_n} R(x)\, dx = h^{n+2} f^{(n+1)}(\xi) \int_0^n \binom{s}{n+1}\, ds = -n\, K_n\, h^{n+2} f^{(n+1)}(\xi) \qquad \xi \in [x_0, x_n] \tag{4}$$

Se ha usado el residual para intervalos regulares 1.3.(5) ($dx = h\, ds$)

$$R_n(x) = \binom{s}{n+1} h^{n+1}\, f^{(n+1)}(\xi) \qquad \xi \in [x_0, x_n] \tag{5}$$

porque es el más adecuado para este caso.

Ocurre para los casos n par, que la integral (4) es nula, por lo que se le agrega un grado más al polinomio de interpolación (cuya integración da nula) y el residual se incrementa en un grado, por lo que el resultado de su integración da un grado mayor en el exponente de h y el orden de la derivación (identificado con m adelante, a veces $m = n + 2$ (n par), a veces $m = n + 1$ (n impar)).

La tabla siguiente resume estos resultados para varios valores de n, dándole nombre en cada caso para las fórmulas de Newton-Cotes

Tabla. Coeficientes de Las Fórmulas de Newton-Cotes.

n m	Factor	C_0^n	C_1^n	C_2^n	C_3^n	C_4^n	C_5^n	C_6^n	$n \times K_n$	K_n
1 2	$\frac{1}{2}\times$	1	1						$\frac{1}{12}$	$\frac{1}{12}$
2 4	$\frac{1}{3}\times$	1	4	1					$\frac{1}{90}$	$\frac{1}{180}$
3 4	$\frac{3}{8}\times$	1	3	3	1				$\frac{3}{80}$	$\frac{1}{80}$
4 6	$\frac{2}{45}\times$	7	32	12	32	7			$\frac{8}{945}$	$\frac{2}{945}$
5 6	$\frac{5}{288}\times$	19	75	50	50	75	19		$\frac{275}{12096}$	$\frac{55}{12096}$
6 8	$\frac{1}{140}\times$	41	216	27	272	27	216	41	$\frac{9}{1400}$	$\frac{3}{2800}$

Nota: N debe ser múltiplo de n de la fórmula.
1-Trapecio, **2**-Simpson1/3, **3**-Simpson3/8-Newton, **4**-Villarceau-Boole, **5**-Villarceau, **6**-Hardy.

Aplicando la fórmula para cada grupo de datos $x \in [x_i, x_{i+n}]$ ($f_i = f(x_i)$)

$$\begin{aligned}\int_{x_i}^{x_{i+n}} f(x)\,dx &= h\sum_{j=0}^{n} C_j^n f_{i+j} - n\,K_n\,h^{m+1} f^{(m)}(\zeta_i) \qquad \zeta_i \in [x_i, x_{i+n}] \qquad \sum f^{(m)}(\zeta_i) = \frac{N}{n} f^{(m)}(\zeta) \\ &= I^n + E^n \qquad\qquad m = \frac{2\,n + 3 + (-1)^n}{2} \qquad\qquad E^n = -n\,K_n\,h^{m+1} f^{(m)}(\zeta_i)\end{aligned} \tag{6}$$

Aplicando la regla para todo el conjunto de puntos de los datos $x \in [x_0, x_{_N}]$

$$\begin{aligned}\int_{x_0}^{x_N} f(x)\,dx &= h\sum_{\substack{i=0\\ n}}^{N-n}\sum_{j=0}^{n} C_j^n f_{i+j} - K_n\,(b-a)\,h^m f^{(m)}(\zeta) \qquad \zeta \in [x_0, x_N] \qquad a = x_0 \quad b = x_N \\ &= I_N^n + E_N^n \qquad\qquad N = \frac{(b-a)}{h} \qquad\qquad E_N^n = -K_n(b-a)h^m f^{(m)}(\zeta)\end{aligned} \tag{7}$$

EJEMPLO:

Hallar la integral de la función $F(x)$, dada en foma tabular, en el intervalo [0.0, 0.6]. Usar la formula de Newton-Cotes basada en un polinomio de tercer grado ($n = 3$), también conocida como la formula de Simpson 3/8.

Datos

x_i	0.0	0.1	0.2	0.3	0.4	0.5	0.6
$f(x_i)$	0.0000	0.0998	0.1987	0.2955	0.3894	0.4794	0.5646

La expresión para la Regla de Simpson 3/8 correspondiente sería

$$I = \frac{3}{8}\, h\, (f_0 + 3\, f_1 + 3\, f_2 + 2\, f_3 + 3\, f_4 + 3\, f_5 + f_6)$$

donde $h = 0.1$. Sustituyendo los valores de la tabla se obtiene que el valor de la integral I es 0.1747.

2.1.2. Extrapolación de Richardson

Si denotamos con I^* el valor exacto de la integral de la función en un intervalo $[a, b]$ con datos regulares y luego hacemos la integración numérica del mismo orden n, pero con distintos números totales de puntos N_1 y N_2 en dos oportunidades

$$I^* = I^n_{N_1} + E^n_{N_1} = I^n_{N_2} + E^n_{N_2} \tag{8}$$

Asumiendo que $f^{(m)}(\zeta_1) \approx f^{(m)}(\zeta_2)$, queda que

$$\frac{E^n_{N_1}}{E^n_{N_2}} \approx \left(\frac{N_2}{N_1}\right)^m = \left(\frac{h_1}{h_2}\right)^m \tag{9}$$

Substituyendo esta expresión queda

$$I^* = I^n_{N_1} + E^n_{N_1} = I^n_{N_2} + E^n_{N_1} \left(\frac{N_1}{N_2}\right)^m \qquad E^n_{N_1} = \frac{I^n_{N_2} - I^n_{N_1}}{1 - (N_1/N_2)^m} \tag{10}$$

y resulta la fórmula de *extrapolación de Richardson*

$$I^* = I^n_{N_1} - \frac{I^n_{N_1} - I^n_{N_2}}{1 - (N_1/N_2)^m} = \frac{(N_2/N_1)^m\, I^n_{N_2} - I^n_{N_1}}{(N_2/N_1)^m - 1} \tag{11}$$

2.1.3. Algoritmo de Romberg

Si tomamos $N_2 = 2\, N_1$, y aumimos que $I^* = I^{n+2}_{N_2}$, se obtiene la *fórmula de Romberg*

$$I^{n+2}_{N_2} = \frac{2^m\, I^n_{N_2} - I^n_{N_1}}{2^m - 1} \tag{12}$$

Tambien se acostumbra a colocarla como (biparticiones sucesivas)

$$I_{i+1,j+1} = \frac{4^j\, I_{i+1,j} - I_{i,j}}{4^j - 1} \tag{13}$$

comenzando con la regla del trapecio ($j = 1$), y siguiendo $j = 1, 2, 3, \ldots$, $i = j, j+1, \ldots$, $h = (b-a)/N$ $N = 2^i$, $m = n + 1 = 2\, j \Rightarrow n = 2\, j - 1$. Las diferentes integraciones se ordenan de forma triangular, cada fila i es la bipartición de la anterior y las columnas j indican el orden del error.

2.1.4. Fórmula de Euler-Maclaurin

En matemáticas, la fórmula de Euler-Maclaurin relaciona a integrales con series. Esta fórmula puede ser usada para aproximar integrales por sumas finitas o, de forma inversa, para evaluar series (finitas o infinitas) resolviendo integrales. La fórmula fue descubierta independientemente por Leonhard Euler y Colin Maclaurin en 1735. Euler usó esta fórmula para calcular valores de series infinitas con convergencia lenta y Maclaurin la utilizó para calcular integrales.

Sea $f(x)$ es una función suave (suficientemente derivable) definida $\forall x \in [0, n]$, entonces, la integral

$$I = \int_0^n f(x)\, dx \tag{14}$$

puede ser aproximada por la siguiente suma

$$S = \frac{f(0) + f(n)}{2} + \sum_{k=1}^{n-1} f(k) \tag{15}$$

(ver regla del trapecio). La fórmula de Euler-Maclaurin nos da una expresión para la diferencia entre la suma y la integral en función de derivadas de $f(x)$ en los extremos del intervalo de integración (0 y n). Para cualquier entero positivo p, tenemos que se cumple

$$S - I = \sum_{k=1}^{p} \frac{B_{k+1}}{(k+1)!} \left[f^{(k)}(n) - f^{(k)}(0) \right] + R \tag{16}$$

donde B_n son los números de Bernoulli y R es una estimación del error normalmente pequeña.

Realizando un cambio de variable en la integral, se puede modificar esta fórmula para funciones $f(x)$ definidas en otros intervalos de la recta real.

El término de error R es

$$R = (-1)^p \int_0^n f^{(p+1)}(x) \, \frac{B_{p+1}(x - \lfloor x \rfloor)}{(p+1)!} \, dx \tag{17}$$

donde $B_i(x - \lfloor x \rfloor)$ son los polinomios de Bernoulli periódicos (el operador $\lfloor \cdot \rfloor$ es el entero $\leq$ argumento). El término de error se puede acotar por

$$|R| \leq \frac{2}{(2\pi)^{2p}} \int_0^n \left| f^{(p+1)}(x) \right| \, dx \tag{18}$$

Cuando se quiere calcular la expansión asintótica de series, la forma más cómoda de la fórmula de Euler-Maclaurin es

$$\sum_{n=a}^{b} f(n) \approx \int_a^b f(x) \, dx + \frac{f(a) + f(b)}{2} + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} \left[f^{(2k-1)}(b) - f^{(2k-1)}(a) \right] \tag{19.a}$$

donde a y b son enteros. Puede ocurrir que esta fórmula siga siendo válida incluso tomando el límite $a \to -\infty$ o $b \to +\infty$, o ambos. Se satisface que

$$\zeta(2k) = \frac{(-1)^{k-1} (2\pi)^{2k} B_{2k}}{2\,(2\,k)!} \qquad\qquad \zeta(-k) = -\frac{B_{k+1}}{k+1} \tag{19.b, c}$$

Los polinomios de Bernoulli $B_n(x)$ con $n = 0, 1, 2, \ldots n$ se pueden definir recursivamente como sigue

$$B_0(x) = 1 \qquad B_n'(x) = n\, B_{n-1}(x) \qquad \int_0^1 B_n(x) \, dx = 0 \qquad B_k^{(r)}(x) = r! \binom{k}{r} B_{k-r}(x) \tag{20}$$

Los primeros 6 son

$$B_1(x) = x - \frac{1}{2} \qquad B_2(x) = x^2 - x + \frac{1}{6} \qquad B_3(x) = x^3 - \frac{3}{2}x^2 + \frac{1}{2}x \tag{21.a, b, c}$$

$$B_4(x) = x^4 - 2x^3 + x^2 - \frac{1}{30} \qquad B_5(x) = x^5 - \frac{1}{2}x^4 + \frac{1}{3}x^3 - \frac{1}{30}x \tag{21.d, e}$$

$$B_6 = x^6 - \frac{3}{5}x^5 + \frac{1}{2}x^4 - \frac{1}{10}x^2 + \frac{1}{42} \qquad (21.f)$$

Los valores $B_n(0) = B_n$ son los números de Bernoulli y se calculan para que la integral $(20.c)$ sea nula. Para $n \geq 2$ se cumple $B_n(0) = B_n(1)$. Los primeros 8 son

$$B_1 = -\frac{1}{2} \qquad B_2 = \frac{1}{6} \qquad B_3 = 0 \qquad B_4 = -\frac{1}{30} \qquad (22.a,b,c,d)$$

$$B_5 = 0 \qquad B_6 = \frac{1}{42} \qquad B_7 = 0 \qquad B_8 = -\frac{1}{30} \qquad (22.e,f,g,h)$$

Las funciones periódicas de Bernoulli $P_n(x)$ se definen como

$$P_n(x) = B_n(x - \lfloor x \rfloor) \qquad (23)$$

Es decir, son iguales a los polinomios de Bernoulli en el intervalo [0,1), pero son funciones periódicas de periodo 1 en el resto del eje real.

2.2. DATOS IRREGULARES

Estos métodos se obtienen al hacer pasar un polinomio $P_n(x)$ en diferencias divididas de grado n en los $n+1$ puntos de cada grupo. Cada grupo termina en x_{i+n} donde el siguiente comienza, $i = 0$ hasta N de n en n.

2.2.1. Polinómica

La regla del trapecio

$$\int_{x_0}^{x_N} y(x)\, dx = \frac{1}{2} \sum_{i=1}^{N} (y_i + y_{i-1})\,(x_i - x_{i-1}) \qquad (1)$$

obtenida al hacer pasar un polinomio $P_1(x)$ por los puntos x_i y x_{i-1}

La regla de Simpson (N par)

$$\begin{aligned} \int_{x_0}^{x_N} y(x)\, dx = \sum_{i=2}^{N} \Bigg\{ & (x_i - x_{i-2}) \left[y_{i-2} + \frac{(x_i - x_{i-2})}{(x_{i-1} - x_{i-2})} \frac{(y_{i-1} - y_{i-2})}{2} \right] \\ & + \frac{1}{6} \left(2\, x_i^2 - x_i x_{i-2} - x_{i-2}^2 + 3\, x_{i-1} x_{i-2} - 3\, x_i x_{i-1}\right) \left[\frac{(y_i - y_{i-1})}{(x_i - x_{i-1})} \frac{(y_{i-1} - y_{i-2})}{(x_{i-1} - x_{i-2})} \right] \Bigg\} \end{aligned} \qquad (2)$$

obtenida al hacer pasar un polinomio $P_2(x)$ por los puntos x_i, x_{i-1} y x_{i-2}.

2.2.2. Cuadratura de Gauss-Legendre

La cuadratura de Gauss-Legendre utiliza los polinomios de Legendre como auxiliares para realizar el cómputo de las integrales, adicionalmente utiliza las raíces de dichos polinomios en el intervalo [-1,1] como puntos de colocación.

Los polinomios de Legendre son ($P_k(1) = 1$, $P_k(-x) = (-1)^k P_k(x)$, $P_k'(1) = k(k+1)/2$)

$$P_0(x) = 1 \qquad P_1(x) = x \qquad P_2(x) = \frac{1}{2}(3x^2 - 1) \qquad P_3(x) = \frac{1}{2}(5x^3 - 3x) \qquad P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$$

$$P_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x) \qquad P_6(x) = \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5) \qquad (3)$$

los demás se pueden hallar con las siguientes expresiones

$$P_n(x) = \frac{2n-1}{n}\, x\, P_{n-1}(x) - \frac{n-1}{n}\, P_{n-2}(x) \qquad P_n(x) = \frac{1}{2^n n!}\,\frac{d^n}{dx^n}[(x^2-1)^n] = \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}^2 (x-1)^k (x+1)^{n-k} \tag{4}$$

La primera se conoce como la relación de recurrencia, la segunda es la fórmula de Rodriges. Estos polinomios satisfacen la ortogonalidad dentro del operador integral en el intervalo [-1,1]

$$\langle P_n, P_m\rangle = \int_{-1}^{1} P_n(x)\, P_m(x)\, dx = \begin{cases} 0 & \text{si } n \neq m \\ c(n) = \dfrac{2}{2n+1} \neq 0 & \text{si } n = m \end{cases} \tag{5}$$

En general los métodos de cuadratura se formulizan como

$$\int_a^b f(x)\, dx \approx \sum_{i=0}^{n} w_i\, f(x_i) \tag{6}$$

Esta expresión es exacta si:

a:) Los x_i son prefijados y regulares, $i = 0, 1, 2, \ldots, n$, y los $n+1$ parámetros w_i pueden ser definidos suponiendo que $f(x)$ es un polinomio de grado n (cuadratura de Newton).

b:) Los x_i como los w_i, $i = 0, 1, 2, \ldots, n$, no están prefijados y estos $2n+2$ parámetros pueden ser definidos suponiendo que $f(x)$ es un polinomio de grado $2n+1$ (cuadratura de Gauss).

Haciendo el siguiente cambio de variables

$$z = \frac{2x - (a+b)}{(b-a)} \qquad x = \frac{(b-a)\, z + (a+b)}{2} \qquad dx = \frac{(b-a)}{2}\, dz \tag{7}$$

el problema de integrar en el intervalo $[a, b]$ en x, se lleva a el problema de integrar en el intervalo $[-1, 1]$ en z. Si utilizamos los polinomios de Lagrange en este último intervalo entonces

$$f(z) = P_n(z) + R_n(z) \qquad P_n(z) = \sum_{i=0}^{n} L_i(z)\, f(z_i) \qquad R(z) = \prod_{j=0}^{n}(z - z_j)\,\frac{f^{(n+1)}(\zeta)}{(n+1)!} = S_{n+1}(z)\, Q_n(z) \tag{8}$$

donde $\zeta \in [-1, 1]$, $S_{n+1}(z)$ es un polinomio de grado $n+1$, y $Q_n(z)$ es un polinomio de grado n.

Para hallar los z_i, llamados puntos de colocación, tales que anulen la integral de $R_n(z)$, vamos a expandir los polinomios S_{n+1} y Q_n en términos de los polinomios de legendre

$$S_{n+1}(z) = \prod_{j=0}^{n}(z - z_j) = \sum_{j=0}^{n+1} a_j P_j(z) \qquad Q_n(z) = \frac{f^{(n+1)}(\zeta)}{(n+1)!} = \sum_{j=0}^{n} b_j P_j(z) \tag{9}$$

Basándonos en la propiedad de ortogonalidad

$$\int_{-1}^{1} R_n(z)\, dz = \sum_{i=0}^{n} a_i b_i \int_{-1}^{1} [P_i(z)]^2\, dz \tag{10}$$

Una forma de anular esta expresión es especificando que $b_i = 0$ con $i = 0, 1, 2, \ldots, n$, o sea que

$$S_{n+1}(z) = \prod_{j=0}^{n}(z - z_j) = a_{n+1} P_{n+1}(z) \tag{11}$$

Esta última ecuacion nos indica que a_{n+1} es inverso del coeficiente que acompaña a z^{n+1} en el polinomio de Legendre $P_{n+1}(z)$ y que las raíces z_i, $i = 0, 1, 2, \ldots, n$ de $S_{n+1}(z) = 0$ son las mismas que las del polinomio de Legendre $P_{n+1}(z) = 0$, con lo cual se obtienen los puntos de colocación z_i.

Entonces la integral de la función de calcula como

$$\int_a^b f(x)\,dx = \frac{(b-a)}{2}\int_{-1}^{1} f(z)\,dz = \sum_{i=0}^{n} w_i\, f(z_i) + E_n \qquad\qquad w_i = \int_{-1}^{1} L_i(z)\,dz \tag{12}$$

donde el error se estima con

$$E_n = \frac{2^{2n+3}[(n+1)!]^4}{[(2n+2)!]^3(2n+3)} f^{(2n+2)}(\eta) \qquad\qquad \eta \in [-1, 1] \tag{13}$$

y otra forma de calcular w_i es

$$w_i = \frac{-2}{(n+2)P'_{n+1}(z_i)\,P_{n+2}(z_i)} \qquad\qquad i = 0, 1, 2, \ldots, n \tag{14}$$

EJEMPLO:

Evaluar $I = \int_0^{\pi/2} \operatorname{sen} x\; dx$ usando el método de cuadratura de Gauss-Legendre con dos puntos de colocación ($n = 1$). El cambio de variables es

$$a = 0 \qquad b = \frac{\pi}{2} \qquad x = \frac{\pi}{4}(z+1) \qquad dx = \frac{\pi}{4}dz$$

$$z_0 = 0.57735 \qquad z_1 = -0.57735 \qquad w_0 = w_1 = 1$$

$$I \approx \frac{\pi}{4}\left[sen(0.10566\,\pi) + \operatorname{sen}(0.39434\,\pi)\right] = 0.99847$$

con un error de $E_n = 1.53 \times 10^{-3}$ (el valor exacto es 1). Un error equivalente a haber usado simpson 3/8 (polinomio de grado $2n + 1 = 3$).

2.3. INTEGRACION MULTIPLE

Sea la siguiente función $z = f(x, y)$ definida en el plano $x - y$ de forma discreta donde $f_ij = f(x_i, y_j)$, $h_x = x_{i+1} - x_i$ y $h_y = y_{j+1} - y_j$ (intervalos regulares).

Hallar la integral

$$I\!I = \int_{y_0}^{y_4}\int_{x_0}^{x_5} f(x, y)\,dx\,dy \tag{1}$$

usando las reglas del trapecio en x y la regla de Simpson en y.

Llamemos

$$I(y) = \int_{x_o}^{x_5} f(x, y)\,dx \tag{2}$$

y su aproximación $I_j = I(y_j)$, donde

$$I_j = \frac{h_x}{2}\left(f_{0j} + 2\,f_{1j} + 2\,f_{2j} + 2\,f_{3j} + 2\,f_{4j} + f_{5j}\right) \qquad\qquad j = 0, 1, 2, 3, 4 \tag{3}$$

Así se obtiene que

$$I\!I = \int_{y_0}^{y_4} I(y)\,dy \approx \frac{h_y}{3}\left(I_0 + 4\,I_1 + 2\,I_2 + 4\,I_3 + I_4\right) \tag{4}$$

3. APROXIMACION

Sea un conjunto de p valores $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, ..., $\mathbf{x}_p$, donde cada $\mathbf{x}_i$ representa una $(m+1)$-upla de la forma

$$\mathbf{x}_i = (x^1, x^2, x^3, \ldots, x^m, x^{m+1})_i \tag{1}$$

con todas sus componentes independientes entre sí.

Sea $y = f(\mathbf{x})$ una función escalar que expresa una de las componentes de la $(m+1)$-upla en función de las restantes. Es decir, por ejemplo, que

$$x^{m+1} = f(x^1, x^2, x^3, \ldots, x^m) \tag{2}$$

Esto se puede hacer sin pérdida de generalidad, puesto que siempre se puede hacer una transformación del tipo

$$\tilde{\mathbf{x}} = \mathbf{H}(\mathbf{x}) \tag{3}$$

tal que los $\tilde{\mathbf{x}}_i$ tengan todos sus componentes independientes entre sí, al igual que los $\mathbf{x}_i$ en (1).

Dados los valores $\mathbf{x}_i$ y definida la función $y = f(\mathbf{x})$, se puede ahora tratar de encontrar una función de aproximación $Y = F(\mathbf{x}, \mathbf{c})$ dependiente, no sólo de los $\mathbf{x}$, sino también de n parámetros c_j, los cuales son expresados en la forma de una n-upla como

$$\mathbf{c} = (c_1, c_2, c_3, \ldots, c_n) \tag{4}$$

Estos parámetros $\mathbf{c}$ se escogen tales que la función $Y_i = F(\mathbf{x}_i, \mathbf{c})$ se aproxime lo mejor posible a la función $y_i = f(\mathbf{x}_i)$, para todos los valores $\mathbf{x}_i$, con $i = 1, 2, 3, \ldots, p$.

El método de los mínimos cuadrados en particular lo que trata es de encontrar los valores de los parámetros $\mathbf{c}$ de la función $Y = F(\mathbf{x}, \mathbf{c})$, tales que el valor

$$S = \sum_{i=1}^{p} (Y_i - y_i)^2 = \sum_{i=1}^{p} (\delta_i)^2 \tag{5}$$

sea el mínimo posible. El valor S representa la sumatoria de todas las desviaciones, entre la función definida para los puntos y la función de aproximación encontrada, al cuadrado.

El valor S se puede interpretar de dos formas posibles. Se puede interpretar como un funcional de la función F, es decir, $S = S[F(\mathbf{x}, \mathbf{c})]$, donde a su vez la función F depende de unos parámetros $\mathbf{c}$ que forman parte de la misma. En este caso el método de mínimos cuadrados se convierte en un problema variacional. El valor S también se puede interpretar como una función de los parámetros, es decir, $S = S(\mathbf{c})$, asumiendo una función de aproximación ya encontrada. Esta última forma es la que vamos a interpretar aquí.

Con la aclaratoria precedente, entonces la definición (5) se puede expresar como

$$S(\mathbf{c}) = \sum_{i=1}^{p} [\, F(\mathbf{x}_i, \mathbf{c}) - f(\mathbf{x}_i) \,]^2 = \sum_{i=1}^{p} [\delta_i(\mathbf{c})]^2 = [\boldsymbol{\delta}(\mathbf{c})]^2 \qquad \boldsymbol{\delta}(\mathbf{c}) = \mathbf{F}(\mathbf{c}) - \mathbf{f} \tag{6}$$

En algunos casos se alteran la desviaciones con una función de peso $W(\mathbf{x})$ para hacer que el ajuste de los parámetros tienda a hacer la aproximación mejor para unos valores de $\mathbf{x}_i$ que para otros. Esto es

$$S(\mathbf{c}) = \sum_{i=1}^{p} W(\mathbf{x}) \, [\, F(\mathbf{x}_i, \mathbf{c}) - f(\mathbf{x}_i) \,]^2 \tag{7}$$

Sin embargo, todas las deducciones se harán para el método de los mínimos cuadrados expresado como está en (6). Extender estos resultados a como está expresado el método en (7) es muy sencillo.

3.1. LINEAL

El método de mínimos cuadrados es en sí un procedimiento para encontrar el valor mínimo de la función (6) de $S = S(\mathbf{c})$. Para ello deben encontrarse los valores de c_j, tales que hagan las derivadas de $S(\mathbf{c})$ todas nulas. En otras palabras,

$$\frac{\partial S}{\partial c_j} = 2\sum_{i=1}^{p} [\, F(\mathbf{x}_i, \mathbf{c}) - f(\mathbf{x}_i)\,] \left(\frac{\partial F}{\partial c_j}\right)_{\mathbf{x}_i} = 2\,[\mathbf{J_F}(\mathbf{c})]^{\mathrm{t}}.\,\boldsymbol{\delta}(\mathbf{c}) = 0 \qquad\qquad [\mathbf{J_F}(\mathbf{c})] = [\boldsymbol{\nabla}_{\mathbf{c}}\mathbf{F}(\mathbf{c})]^{\mathrm{t}} \tag{8}$$

Las expresiones (8) se denominan "Ecuaciones Normales" y deben cumplirse simultáneamente para $j = 1, 2, 3, \ldots, n$. Su nombre se debe a que la derivadas son calculadas para una hipersuperficie donde las direcciones c_j son ortogonales entre sí y están evaluadas en un punto $\mathbf{c}$ donde todas son nulas. Las direcciones son ortogonales debido a que los parámetros c_j son todos independientes entre sí.

Es bueno hacer notar que lo que se halla mediante este procedimiento es un mínimo de la función escalar $S(\mathbf{c})$ y no un máximo, puesto que la función $Y_i = F(\mathbf{x}_i, \mathbf{c})$ puede estar tan alejada de los valores $\mathbf{x}_i$ como se quiera, variando los valores de los parámetros c_j.

EJEMPLO:

En el análisis de la aproximación de múltiples variables, el método de los mínimos cuadrados es utilizada con bastante frecuencia, para una función de aproximación del tipo $F(x, y, \mathbf{a}) = a_1 + a_2 x + a_3 y$, encuentre el sistema de ecuaciones lineales a resolver para determinar las constantes de la aproximación.

Con la siguiente tabla de datos, determine las constantes de la aproximación suponiendo que la funcion se comporta linealmente en las dos variables independientes.

Datos

x_i	0	1.2	2.1	3.4	4.0	4.2	5.6	5.8	6.9
y_i	0	0.5	6.0	0.5	5.1	3.2	1.3	7.4	10.2
$f(x_i, y_i)$	1.2	3.4	−4.6	9.9	2.4	7.2	14.3	3.5	1.3

3.1.1. Series de Funciones Bases

El ajuste lineal es el más empleado y el más reportado en la literatura. Su nombre se debe a que la función de aproximación posee una expresión lineal de la forma

$$F(\mathbf{x}, \mathbf{c}) = \sum_{k=1}^{n} c_k\, g_k(\mathbf{x}) = \mathbf{c}.\mathbf{G}(\mathbf{x}) \tag{9}$$

lo que no es más que una serie de funciones $g_j(\mathbf{x})$ todas diferentes entre sí, por lo que son consideradas que forman parte de una base de un espacio de funciones.

Para el caso particular de la función de aproximación definida por (8) se tiene que

$$\frac{\partial F}{\partial c_j} = g_j(\mathbf{x}) \qquad\qquad \boldsymbol{\nabla}_{\mathbf{c}}\mathbf{F}(\mathbf{x}) = \mathbf{G}(\mathbf{x}) \tag{10}$$

Substituyendo este resultado en la expresión (8) de la sub-sección 3.1, junto con la definición (1) e intercambiando las sumatorias de k con la sumatoria de i, se obtiene

$$\frac{\partial S}{\partial c_j} = 2\sum_{i=1}^{p} \left[\sum_{k=1}^{n} c_k\, g_k(\mathbf{x}_i) - f(\mathbf{x}_i) \right] g_j(\mathbf{x}_i) = 0 \tag{11.a}$$

$$\sum_{i=1}^{p}\sum_{k=1}^{n} c_k\, g_k(\mathbf{x}_i)\, g_j(\mathbf{x}_i) = \sum_{i=1}^{p} f(\mathbf{x}_i)\, g_j(\mathbf{x}_i) \tag{11.b}$$

$$\sum_{k=1}^{n}\left[\sum_{i=1}^{p} g_j(\mathbf{x}_i)\, g_k(\mathbf{x}_i)\right] c_k = \sum_{i=1}^{p} g_j(\mathbf{x}_i)\, f(\mathbf{x}_i) \tag{11.c}$$

Al final queda un sistema de ecuaciones lineales de la forma

$$\sum_{k=1}^{n} A_{jk}\, c_k = b_j \qquad\qquad [\mathbf{A}].\mathbf{c} = \mathbf{b} \tag{12}$$

donde los elementos de la matriz del sistema y el vector independiente se expresan como

$$A_{jk} = \sum_{i=1}^{p} g_j(\mathbf{x}_i)\, g_k(\mathbf{x}_i) \qquad\qquad [\mathbf{A}] = [\mathbf{G}\mathbf{G}^{\mathrm{t}}] \tag{13.a}$$

$$b_j = \sum_{i=1}^{p} g_j(\mathbf{x}_i)\, f(\mathbf{x}_i) \qquad\qquad \mathbf{b} = \mathbf{G}.\mathbf{f} \tag{13.b}$$

EJEMPLO:

Hallar la aproximación cuadrática para la siguiente tabla de datos

Datos

x_i	0.05	0.11	0.15	0.31	0.46	0.52	0.70	0.74	0.82	0.98	1.17
$f(x_i)$	0.956	0.890	0.832	0.717	0.571	0.539	0.378	0.370	0.306	0.242	0.104

Las funciones base son $g_1(x) = 1$, $g_2(x) = x$ y $g_3(x) = x^2$. El sistema de ecuaciones toma la forma

$$11\, a_1 + 6.01\, a_2 + 4.65\, a_3 = 5.905$$
$$6.01\, a_1 + 4.65\, a_2 + 4.12\, a_3 = 2.1839$$
$$4.65\, a_1 + 4.12\, a_2 + 3.92\, a_3 = 1.3357$$

el cual tiene como solución $a_1 = 0.998$, $a_2 = -1.018$ y $a_3 = 0.225$.

3.1.2. Series de Polinomios

Como ejemplos de funciones de aproximación más utilizadas se tienen las series de funciones polinómicas

$$F(x, \mathbf{c}) = \sum_{k=1}^{n} c_k\, x^{k-1} \tag{14}$$

y la serie de funciones trigonométricas

$$F(x, \mathbf{c}) = \sum_{k=1}^{n} c_k\, \cos[(k-1)x] \tag{15}$$

También existen series de funciones racionales, hiperbólicas, polinomios de Chebyshev, polinomios de Legendre, etc. También se pueden tener combinaciones de estas funciones.

3.2. NO LINEAL

En el ajuste no lineal de los parámetros c_j la función de aproximación $F(\mathbf{x}, \mathbf{c})$ tiene una expresión distinta a la expresión (1) de la sub-sección 3.2, por consiguiente, lo que se obtiene es un sistema de ecuaciones no lineales en las variable c_j que puede ser resuelto con cualquier método para tales tipo de sistemas, como, por ejemplo, el método de Newton-Raphson. Sin embargo esto trae como consecuencia que el procedimiento de mínimos cuadrados se vuelva más complicado, ya que hay que calcular la matriz jacobiana del sistema de funciones no lineales.

Para evitar el incoveniente mencionado se han desarrollado varios métodos, dentro los cuales están:

- Método del máximo descenso.
- Método de Gauss-Newton.
- Método de Levenberg-Marquardt.

Todos estos métodos se derivan del siguiente análisis.

Sea la expansión en series de Taylor hasta el término de primer orden de la función de aproximación $F(\mathbf{x}_i, \mathbf{c})$ alrededor de un valor estimado $\mathbf{c}^*$ de los parámetros. Esto es,

$$F(\mathbf{x}_i, \mathbf{c}) = F(\mathbf{x}_i, \mathbf{c}^*) + \sum_{k=1}^{n} \Big(\frac{\partial F}{\partial c_k}\Big)^*_{\mathbf{x}_i} \Delta c_k^* + O(\|\Delta \mathbf{c}^*\|^2) \tag{1.a}$$

donde

$$\Delta c_k^* = c_k - c_k^* \tag{1.b}$$

Substituyendo este resultado en la expresión (8) de la sub-sección 3.1, e intercambiando las sumatorias de k con la sumatoria de i, se obtiene un sistema de ecuaciones de la forma

$$\sum_{k=1}^{n} \Big[\sum_{i=1}^{p} \Big(\frac{\partial F}{\partial c_j}\Big)_{\mathbf{x}_i} \Big(\frac{\partial F}{\partial c_k}\Big)^*_{\mathbf{x}_i} \Big] \Delta c_k^* = - \sum_{i=1}^{p} [\, F(\mathbf{x}_i, \mathbf{c}^*) - f(\mathbf{x}_i) \,] \Big(\frac{\partial F}{\partial c_j}\Big)_{\mathbf{x}_i} \tag{2}$$

Si se supone que las derivadas $\partial F / \partial c_j$ no sufren gran variación del punto valor c_k^* al valor c_k, entonces la expresión (2) se podría reescribir aproximadamente como

$$\sum_{k=1}^{n} A_{jk}^* \, \Delta c_k^* = b_j^* \tag{3.a}$$

donde

$$A_{jk}^* = \sum_{i=1}^{p} \Big(\frac{\partial F}{\partial c_j}\Big)^*_{\mathbf{x}_i} \Big(\frac{\partial F}{\partial c_k}\Big)^*_{\mathbf{x}_i} \tag{3.b}$$

$$b_j^* = - \sum_{i=1}^{p} [\, F(\mathbf{x}_i, \mathbf{c}^*) - f(\mathbf{x}_i) \,] \Big(\frac{\partial F}{\partial c_j}\Big)^*_{\mathbf{x}_i} \tag{3.c}$$

Con base en este análisis, entonces se pueden aplicar los diferentes métodos que se explican a continuación.

3.2.1. Método del Máximo Descenso

El método del máximo descenso está basado en el hecho de que

$$\frac{\partial S}{\partial c_j} \Big|^s = -b_j^s \qquad\qquad \boldsymbol{\nabla}_{\mathbf{c}} S(\mathbf{c}^s) = -\mathbf{b}^s \tag{4}$$

Es decir, que $S(\mathbf{c}^s)$ se incrementa en la dirección indicada por el gradiente (4). Si se escoge una dirección $\Delta\mathbf{c}^s$ opuesta a este gradiente tal que

$$\Delta c_j^s = \omega\, b_j^s \tag{5}$$

se obtendrá el máximo descenso de la función $S(\mathbf{c})$.

La expresión (5) se puede reescribir como

$$\sum_{k=1}^{n} D_{jk}^s\, \Delta c_k^s = b_j^s \qquad\qquad [\mathbf{D}^s].\Delta\mathbf{c}^s = \mathbf{b}^s \tag{6}$$

donde

$$D_{jk}^s = \delta_{jk} \qquad\qquad [\mathbf{D}^s] = [\,\mathbf{I}\,] \tag{7}$$

y se tiene que

$$\mathbf{c}^{s+1} = \mathbf{c}^s + \omega\, \Delta\mathbf{c}^{s+1} \tag{8}$$

Sin embargo, el método puede ser modificado de manera tal que la matriz D_{jk}^s tenga dimensiones acorde con la función $S(\mathbf{c})$, y, por consiguiente, se puede hacer

$$D_{jk}^s = \|\mathbf{A}^s\|\, \delta_{jk} \qquad\qquad [\mathbf{D}^s] = \|\mathbf{A}^s\|\, [\,\mathbf{I}\,] \tag{9}$$

El valor de ω se modifica de igual forma que el método de Gauss-Newton para asegurar la convergencia, pero por el contrario el método del máximo descenso converge muy lentamente y por lo tanto no es recomendable su uso.

3.2.2. Método de Gauss-Newton

El método de Gauss-Newton consiste en un procedimiento iterativo que se origina a partir de las expresiones (1) junto con la definición (3.*b*). De esta forma resulta el siguiente algoritmo iterativo con s como indicador del número de la iteración

$$\sum_{k=1}^{n} A_{jk}^s\, \Delta c_k^s = b_j^s \qquad\qquad [\mathbf{A}^s].\Delta\mathbf{c}^s = \mathbf{b}^s \tag{10}$$

donde

$$A_{jk}^s = \sum_{i=1}^{p} \Big(\frac{\partial F}{\partial c_j}\Big)_{\mathbf{x}_i}^s \Big(\frac{\partial F}{\partial c_k}\Big)_{\mathbf{x}_i}^s \qquad\qquad [\mathbf{A}^s] = [\mathbf{J_F}(\mathbf{c}^s)]^{\mathrm{t}}.\,[\mathbf{J_F}(\mathbf{c}^s)] \tag{11.a}$$

$$b_j^s = -\sum_{i=1}^{p} [\,F(\mathbf{x}_i, \mathbf{c}^s) - f(\mathbf{x}_i)\,] \Big(\frac{\partial F}{\partial c_j}\Big)_{\mathbf{x}_i}^s \qquad\qquad \mathbf{b}^s = -[\mathbf{J_F}(\mathbf{c}^s)]^{\mathrm{t}}.\,\boldsymbol{\delta}^s \tag{11.b}$$

y luego se obtiene

$$\mathbf{c}^{s+1} = \mathbf{c}^s + \Delta\mathbf{c}^s \tag{12}$$

La expresión (10) representa un sistema de ecuaciones lineales que se resuelve en cada iteración s, conociendo los parámetros $\mathbf{c}^s$. Después se substituye este resultado en la expresión (12) para obtener los valores $\mathbf{c}^{s+1}$ de los parámetros en la siguiente iteración. El procedimiento se continúa hasta obtener convergencia hacia la solución $\mathbf{c}$ de las ecuaciones normales (8) de la sub-sección, aplicando un criterio de parada de la forma

$$\|\Delta\mathbf{c}^s\| < \varepsilon_{max} \tag{13}$$

en el error local de las variables $\mathbf{c}^s$ y donde ε_{max} es la tolerancia permitida para dicho error local.

Frecuentemente es recomendable alterar el algoritmo relajándolo en la forma

$$\mathbf{c}^{s+1} = \mathbf{c}^s + \omega\,\Delta\mathbf{c}^s \tag{12'}$$

para asegurar la convergencia del proceso iterativo. Aquí ω es el factor de relajación y en cada iteración se altera

$$\omega' = \rho\,\omega \qquad \rho < 1 \tag{14}$$

de manera de garantizar que dentro de esa iteración se cumpla que

$$S(\mathbf{c}^{s+1}) < S(\mathbf{c}^s) \tag{15}$$

Recuérdese que lo que se está buscando es el valor de $\mathbf{c}$ para que la función $S(\mathbf{c})$ se haga mínima.

Si la relación (15) se cumple en una iteración, entonces en la siguiente iteración se permite un incremento de ω en la forma

$$\omega' = \tau\,\omega \qquad \tau > 1 \tag{14'}$$

Normalmente se emplean los valores de $\rho = 0.5$ y $\tau = 2$, para producir el efecto de una búsqueda del ω óptimo mediante la bisección consecutiva de los intervalos $[c_k^s, c_k^s + \Delta c_k^s]$, comenzando con un $\omega = 1$.

Cuando las derivadas de las expresiones (11) se hacen complicadas de calcular, estas pueden ser obtenidas numéricamente de la siguiente forma

$$\Big(\frac{\partial F}{\partial c_j}\Big)_{\mathbf{x}_i}^s \cong \frac{F(\mathbf{x}_i, \mathbf{c}^s) - F(\mathbf{x}_i, \mathbf{c}_{(j)}^{s-1})}{c_j^s - c_j^{s-1}} \tag{16.a}$$

donde

$$F(\mathbf{x}_i, \mathbf{c}_{(j)}^{s-1}) = F(\mathbf{x}_i, c_1^s, c_2^s, c_3^s, \ldots, c_j^{s-1}, \ldots, c_n^s) \tag{16.b}$$

3.2.3. Método de Levenberg-Marquardt

La fórmula algorítmica del *método de Levenberg-Marquardt* es la siguiente [Levenberg,(1944)]

$$\sum_{k=1}^{n} (A_{jk}^s + \lambda D_{jk}^s)\,\Delta c_k^s = b_j^s \qquad\qquad [\mathbf{A}^s + \lambda\,\mathbf{D}^s].\,\Delta\mathbf{c}^s = \mathbf{b}^s \tag{17}$$

$$\mathbf{c}^{s+1} = \mathbf{c}^s + \Delta\mathbf{c}^s \tag{18}$$

donde el factor λ funciona similar a un factor de relajación y le da al método de Marquardt un carácter hibrido donde existe un compromiso entre el método del máximo descenso y el método de Gauss-Newton. Cuando $\lambda \to 0$, la dirección del método se dirije hacia el método de Gauss-Newton. Cuando $\lambda \to \infty$, la dirección del método se dirije hacia el método del máximo descenso.

Los estudios de Marquardt [(1963)] indican que el método posee un ángulo promedio entre los métodos de Gauss-Newton y Máximo Descenso de 90°. La selección de un λ entre 0 e ∞ produce una dirección intermedia.

Para efectos de garantizar la convergencia en cada iteración se altera el factor λ de la forma

$$\lambda' = \lambda/\rho \qquad \rho < 1 \tag{19}$$

hasta que se cumpla dentro de la misma iteración que

$$S(\mathbf{c}^{s+1}) < S(\mathbf{c}^s) \tag{15}$$

Una vez satisfecha la relación anterior en una iteración se puede disminuir λ en la siguiente iteración de manera que

$$\lambda' = \lambda/\tau \qquad \tau > 1 \tag{19'}$$

Nótese que incrementar λ en el método de Marquardt es equivalente a disminuir ω en el método de Gauss-Newton.

Normalmente, se toman los valores de $\lambda_{inicial} = 10^{-3}$, $\rho = 0.1$ y $\tau = 10$.

Cuando en varias iteraciones consecutivas el método mejora su convergencia, es decir se cumple la relación (15), entonces $\lambda \to 0$, y esencialmente se estará empleando el método de Gauss-Newton. Si la convergencia no mejora, por el contrario, λ se incrementa y se estará usando prácticamente el método del Máximo Descenso.

3.3. ALGORITMO BFGS

El algoritmo BFGS (Broyden-Fletcher-Goldfarb-Shanno), denominado así por Charles George Broyden, Roger Fletcher, Donald Goldfarb y David Shanno [Fletcher,1987], es un algoritmo de optimización que permite resolver el problema de minimización de una función escalar $f(\mathbf{x})$ diferenciable donde $\mathbf{x}$ es un vector en $\mathbb{R}^n$. No hay restricciones a los valores de $\mathbf{x}$. El algoritmo comienza con un estimado inicial $\mathbf{x}_0$ y procede iterativamente obteniendo mejores estimados en cada iteración k.

Se busca la dirección $\mathbf{z}_k$ en la iteración k mediante la solución análoga a la de Newton-Raphson (para la solución de la ecuación homogénea $\boldsymbol{\nabla} f(\mathbf{x}) = 0$)

$$[\mathbf{H}_k].\mathbf{z}_k = -\boldsymbol{\nabla} f(\mathbf{x}_k) \tag{1}$$

donde $[\mathbf{H}_k] = [\boldsymbol{\nabla}\boldsymbol{\nabla} f(\mathbf{x}_k)]$ es el tensor hessiano de $f(\mathbf{x})$, jacobiano simultáneamente de $\boldsymbol{\nabla} f(\mathbf{x})$, evaluadas en el iterado $\mathbf{x}_k$. Una línea de búsqueda $\mathbf{z}_k$ luego se usa para encontrar un nuevo punto $\mathbf{x}_{k+1} = \mathbf{x}_k + \omega\mathbf{z}_k$, minimizando la función $f(\mathbf{x}_k + \omega\mathbf{z}_k)$ sobre un escalar $\omega > 0$.

La condición cuasi-Newton impone una condición par la actualización de $[\mathbf{H}_k]$. Esta es

$$[\mathbf{H}_{k+1}].(\mathbf{x}_{k+1} - \mathbf{x}_k) = \boldsymbol{\nabla} f(\mathbf{x}_{k+1}) - \boldsymbol{\nabla} f(\mathbf{x}_k) \qquad [\mathbf{H}_{k+1}].\mathbf{s}_k = \mathbf{y}_k \tag{2}$$

que es la ecuación de la secante, donde $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k = \omega\mathbf{z}_k$ y $\mathbf{y}_k = \boldsymbol{\nabla} f(\mathbf{x}_{k+1}) - \boldsymbol{\nabla} f(\mathbf{x}_k)$. La condición de la curvatura $\mathbf{s}_k.\mathbf{y}_k > 0$ debe ser satisfecha. Si la función no es fuertemente convexa, entonces la condicón debe forzarse explícitamente.

En lugar de requerir el hessiano completo en el punto $\mathbf{x}_{k+1}$ sea calculado completamente, se hace un estimado a partir del punto $\mathbf{x}_k$ mediante la suma de dos matrices

$$[\mathbf{H}_{k+1}] = [\mathbf{H}_k] + [\mathbf{U}_k] + [\mathbf{V}_k] \tag{3}$$

Tanto $[\mathbf{U}_k]$ como $[\mathbf{V}_k]$ son matrices simétricas de rango uno, pero su suma es de rango dos en la matriz actualizada. Con la finalidad de mantener el carácter definido positivo y la simetría de $[\mathbf{H}_{k+1}]$, se escoge $[\mathbf{U}] = \alpha\,[\mathbf{uu}]$ y $[\mathbf{V}] = \beta\,[\mathbf{vv}]$. Imponiendo la ecuación secante $(2.b)$ y escogiendo $\mathbf{u} = \mathbf{y}_k$ y $\mathbf{v} = [\mathbf{H}_k].\mathbf{s}_k$ se obtiene

$$\alpha = \frac{1}{\mathbf{y}_k.\mathbf{s}_k} \qquad \beta = -\frac{1}{\mathbf{s}_k.[\mathbf{H}_k].\mathbf{s}_k} \tag{4}$$

Finalmente, cuando se substituyen α y β se obtiene el hessiano actualizado

$$[\mathbf{H}_{k+1}] = [\mathbf{H}_k] + \frac{\mathbf{y}_k\mathbf{y}_k}{\mathbf{y}_k.\mathbf{s}_k} - \frac{[\mathbf{H}_k].[\mathbf{s}_k\mathbf{s}_k].[\mathbf{H}_k]^{\mathrm{t}}}{\mathbf{s}_k.[\mathbf{H}_k].\mathbf{s}_k} \tag{5}$$

$f(\mathbf{x})$ denota la función objetivo a ser minimizada. La convergencia puede ser verificada observando la norma del gradiente $||\boldsymbol{\nabla} f(\mathbf{x}_k)||$. Prácticamente, $[\mathbf{H}_0]$ puede inicializarse con $[\mathbf{H}_0] = [\mathbf{I}]$, asi que en el primer

paso el método será equivalente al método del descenso gradiente, pero adelante los pasos adicionales son refinados por $[\mathbf{H}_k]$, la aproximación del hessiano.

la inversa de la matriz hessiana $[\mathbf{B}_k] = [\mathbf{H}_k]^{-1}$ se puede estimar utilizando la fórmula de ShermanMorrison [Barlett,(1951)]

$$[\mathbf{B}_{k+1}] = \left([\mathbf{I}] - \frac{[\mathbf{s}_k \mathbf{y}_k]}{\mathbf{y}_k . \mathbf{s}_k} \right) . [\mathbf{B}_k] . \left([\mathbf{I}] - \frac{[\mathbf{y}_k \mathbf{s}_k]}{\mathbf{s}_k . \mathbf{y}_k} \right) + \frac{[\mathbf{s}_k \mathbf{s}_k]}{\mathbf{y}_k . \mathbf{s}_k} \tag{6}$$

Esto puede ser computado eficientemente sin matrices temporales, reconociendo que $[\mathbf{B}_k]$ es simétrica y que $\mathbf{y}_k . [\mathbf{B}_k] . \mathbf{y}_k$ y $\mathbf{s}_k . \mathbf{y}_k$ son escalares, utilizando la expansión tal que

$$[\mathbf{B}_{k+1}] = [\mathbf{B}_k] + \frac{(\mathbf{s}_k . \mathbf{y}_k + \mathbf{y}_k . [\mathbf{B}_k] . \mathbf{y}_k)}{(\mathbf{s}_k . \mathbf{y}_k)^2} [\mathbf{s}_k \mathbf{s}_k] - \frac{[\mathbf{B}_k] . [\mathbf{y}_k \mathbf{s}_k] + [\mathbf{s}_k \mathbf{y}_k] . [\mathbf{B}_k]}{\mathbf{s}_k . \mathbf{y}_k} \tag{7}$$

Este método puede utilizarse para encontrar la solución al problema de mínimos cuadrados donde $f(\mathbf{x}) = S(\mathbf{c})$, la suma de las desviaciones al cuadrado, y las variables $\mathbf{x} = \mathbf{c}$ son los coeficientes de la regresión lineal o no-lineal.

3.4. EVALUACION

La evaluación del ajuste viene dada mediante el análisis de de ciertos factores que permiten, por un lado comparar cuan bueno es un ajuste en relación a otro, y por otro lado comparar cuando un ajuste reproduce bien el conjunto de puntos de datos.

Estas cantidades y coeficientes son los siguientes:

Suma de los cuadrados de las desviaciones con respecto a la función de aproximación o suma de las desviaciones con respecto a la función de aproximación

$$S(\mathbf{c}) = \sum_{i=1}^{p} \delta_i^2 \qquad \bar{S}(\mathbf{c}) = \sum_{i=1}^{p} \delta_i \qquad \delta_i = F(\mathbf{x}_i, \mathbf{c}) - f(\mathbf{x}_i) \tag{1}$$

Media de la variable dependiente o media de las desviaciones

$$f_m = \frac{1}{p} \sum_{i=1}^{p} f(\mathbf{x}_i) \qquad \delta_m = \frac{\bar{S}(\mathbf{c})}{p} \tag{2}$$

Suma de los cuadrados de las desviaciones con *respecto a la media* de la variable dependiente o desviaciones respecto a la desviación media

$$S_m = \sum_{i=1}^{p} [\, f(\mathbf{x}_i) - f_m \,]^2 \qquad \bar{S}_m = \sum_{i=1}^{p} (\delta_i - \delta_m)^2 \tag{3}$$

Desviación estándar (σ) ó *varianza* (σ^2) con respecto a la función de aproximación

$$\sigma = \sqrt{\frac{S}{p-n}} \tag{4}$$

Desviación estándar (σ_m) ó *varianza* (σ_m^2) con respecto a la media f_m o la media δ_m

$$\sigma_m = \sqrt{\frac{S_m}{p-1}} \qquad \bar{\sigma}_m = \sqrt{\frac{\bar{S}_m}{p-1}} \tag{5}$$

Coeficiente de determinación (R^2) ó *coeficiente de correlación* (R). R^2 indica el porcentaje de la incertidumbre inicial que ha sido disminuido usando la función de aproximación

$$R^2 = \frac{S_m - S}{S_m} \qquad \bar{R}^2 = \frac{S - \bar{S}_m}{S} \tag{6}$$

En algunas literaturas definen el coeficiente de determinación ó correlación (r) de la siguiente forma alternativa usando las desviaciones estándar

$$r = \frac{\sigma_m - \sigma}{\sigma_m} \qquad \bar{r} = \frac{\sigma - \bar{\sigma}_m}{\sigma} \tag{7}$$

Estos coeficientes ofrecen valores más sensible respecto a la aproximación y por lo tanto son valores más bajos. Los coeficientes y desviaciones con barra son calculados con respecto a la desviación media δ_m, y, por consiguiente, son más pequeños y su interpretaciones son distintas a las cantidades sin barra. Con barra significan que tan concentrado está la nube de puntos alrededor de la curva o superficie (mientras r es más cercano a uno significa que tan bajo es $\bar{\sigma}_m$ respecto a σ). Cabe mencionar que la definición de δ_i en (1) puede hacerse sin o con valor absoluto en el miembro de la derecha.

Coeficiente de variación.

$$C_v = \frac{\sigma_m}{f_m} \qquad \bar{C}_v = \frac{\bar{\sigma}_m}{\delta_m} \tag{8}$$

Desviación RMS (Root of the Mean Square).

$$\delta_{rms} = \sqrt{\frac{S}{p}} \tag{9}$$

Desviación máxima.

$$\delta_{max} = \max_{1 \le i \le p} |F(\mathbf{x}_i, \mathbf{c}) - f(\mathbf{x}_i)| = \max_{1 \le i \le p} |\delta_i| \tag{10}$$

En la desviación estándar σ, la cantidad S está dividida por $(p - n)$, debido a que n parámetros $(c_1, c_2, c_3, \ldots, c_n)$, derivados de los datos originales $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_p)$, fueron usados para computar $S(\mathbf{c})$. De aquí que se hallan perdido n grados de libertad en la probabilidad.

En la desviación estándar σ_m, la cantidad S_m está dividida por $(p - 1)$, debido a que la media de la variable dependiente, f_m, la cual se derivó de los datos originales $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_p)$, fué usada para computar S_m. De aquí que se halla perdido un grado de libertad. Lo mismo para las cantidades con barra.

La desviación estándar σ_m debe ser mayor que σ, de otra forma no se justifica el uso de la función de aproximación, y la media f_m da una mejor aproximación a los datos que la función de aproximación propuesta. Los análisis con las cantidades con barra son con respecto a la curva o superficie $F(\mathbf{x}, \mathbf{c})$ en sí y los sin barras son con respecto a la media f_m. Normalmente $\bar{\sigma}_m$ es mucho menor que σ_m (y que σ), pero cuando son comparables significa que los datos están muy dispersos y no siguen una tendencia marcada por alguna curva o superficie $F(\mathbf{x}, \mathbf{c})$ propuesta como modelo. En este caso, es conveniente pre-procesar los datos (filtrando o normalizando) para eliminar el ruido (noise) y realizar el ajuste a posteriori con un modelo, curva o superficie más aceptable. Las desviacione σ_m y $\bar{\sigma}_m$ son mayor y menor que σ, al igual que S_m y $\bar{S}_m$ respecto a S, lo que hace que los coeficiente R^2 y r, con y sin barras, sean siempre (levemente en los buenos ajustes) inferiores a la unidad. Menores, substancialmente más aún, los coeficientes con barra, pues son con respecto a la desviación media δ_m. Estos coeficientes con barra están más distanciado de la unidad que los con barra, pero en cantidad absoluta significan menos cantidad. En orden de magnitud, las cantidades mencionadas estarían ordenadas como $S_m > S > \bar{S}_m$ y $\sigma_m > \sigma > \bar{\sigma}_m$.

La función de aproximación que mejor se ajusta a los datos originales $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_p)$, no es aquella que ofrece un menor valor de S, sino aquella que brinda una menor desviación estándar σ, con respecto a la

función de aproximación. Esto implica que el coeficiente de determinación o correlación r, es el más adecuado para la evaluación del ajuste, mejor cuando sea más cercano a la unidad por debajo, normalmente expresado de forma porcentual multiplicando su valor por 100. Luego los coeficientes con barra se usan para saber que tan concentrados (cercanos) están los puntos alrededor de la curva o superficie usada como modelo.

El coeficiente de variación C_v nos brinda una medida normalizada de cual es la dispersión de los datos originales y normalmente se da en forma porcentual. Cuando la dispersión de los datos es muy grande significa que los puntos están muy dispersos y si se grafican formarán una nube ancha alrededor de cualquier correlación que se trate de hallar. En este caso, la mejor correlación la ofrecería la media f_m.

La desviación RMS y la desviación máxima dependen del ajuste particular que se está realizando. La desviación RMS se puede interpretar como una desviación promedio del ajuste, pero siempre es menor que el valor absoluto que la desviación media. La desviación máxima δ_{max} acota cuánto va a hacer el mayor error cometido con el ajuste. Entre mayor sea la diferencia entre estas dos desviaciones δ_{rms} y δ_{max}, mejor será el ajuste por sí mismo.

Una forma de optimizar el ajuste es descartar aquellos puntos para los cuales $|\delta_i| > \bar{\sigma}_m$. Este procedimiento aumenta los coeficientes de correlación R y r, con y sin barra.

BIBLIOGRAFIA

[1] Apostol, T. M. "An Elementary View of Euler's Summation Formula", **American Mathematical Monthly**, Vol.**106**, No.5, pp.409-418, (1999).

[2] Bartlett, M. S. "An Inverse Matrix Adjustment Arising in Discriminant Analysis". **Annals of Mathematical Statistics**, Vol.**22**, No.1, pp.107-111, (1951).

[3] Burden R. L.; Faires, J. D. **Numerical Analysis**. 3^{rd} Edition. PWS (Boston), 1985.

[4] Carnahan, B.; Luther, H. A.; Wilkes, J. O. **Applied Numerical Methods**. John Wiley & Sons (New York), 1969.

[5] Chapra, S. C.; Canale, R. P. **Numerical Methods for Engineers**, with Personal Computer Applications. McGraw-Hill Book Company, 1985.

[6] Fletcher, R. **Practical Methods of Optimization**, 2^{nd} Edition. John Wiley & Sons (New York), 1987.

[7] Gerald, C. F. **Applied Numerical Analysis**, 2^{nd} Edition. Addison-Wesley (New York), 1978.

[8] Granados M., A. L. **Nuevas Correlaciones para Flujo Multifásico.** INTEVEP S.A. Reporte Técnico No. INT-EPPR/322-91-0001. Los Teques, Febrero de 1991. Trabajo presentado en la Conferencia sobre: *Estado del Arte en Mecánica de Fluidos Computacional.* Auditorium de INTEVEP S.A. Los Teques, del 27 al 28 de Mayo de (1991).

[9] Granados M., A. L. **Free Order Polynomial Interpolation Algorithm.** INTEVEP S.A. Nota Técnica. Los Teques, Julio de 1991.

[10] Granados, A. L. "Criterios para Interpolar", Universidad Simón Bolívar, Jul. 2017.

[11] Hildebrand, F. B. **Introduction to Numerical Analysis**. McGraw-Hill (New York),1956.

[12] Levenberg, K. "A Method for the Solution of Certain Non-Linear Problems in Least Squares". **Quarterly of Applied Mathematics**, Vol.**2**, pp.164-168, (1944).

[13] Marquardt, D. "An Algorithm for Least Squares Estimation of Non-Linear Parameters". **SIAM J. Appl. Math.**, Vol.**11**, No.2, pp.431-441, (1963).

[14] Nocedal, J.; Wright, S. J. **Numerical Optimization,** 2^{nd} Edition. Springer (New York), 2006.

CAPITULO IV

ECUACIONES DIFERENCIALES ORDINARIAS

CONTENIDO

1. PROBLEMA DE VALOR INICIAL

Un problema de ecuación diferencial ordinaria (ODE - Ordinary Diferential Equation) con valor inicial de primer orden de define como

$$\frac{dy}{dx} = f(x,y) \qquad x = x_0 \qquad y(x_0) = y_0 \tag{1}$$

donde $y(x) : \mathbb{R} \longrightarrow \mathbb{R}$ es la solución que se desea encontrar, conocido su valor en un punto $y(x_0) = y_0$ denominado *valor inicial*. En teoría la solución se puede encontrar hacia adelante o hacia atrás del punto inicial.

Un sistema de ecuaciones diferenciales ordinarias con valor inicial de primer orden de igual manera se define como

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(x,\mathbf{y}) \qquad x = x_0 \qquad \mathbf{y}(x_0) = \mathbf{y}_0 \tag{2}$$

donde $\mathbf{y}(x) : \mathbb{R} \longrightarrow \mathbb{R}^M$ es la solución que se desea encontrar, conocido su valor en un punto $\mathbf{y}(x_0) = \mathbf{y}_0$ denominado igualmente valor inicial (aunque en realidad sean M valores definidos en un único punto $x = x_0$). Al igual que antes se puede encontrar hacia adelante o hacia atrás del punto inicial. Estos sistemas se tratarán más extensamente en la sección 3.

1.1. METODOS DE UN SOLO PASO

Los métodos de un sólo paso se basan en que, a partir de un valor inicial y_0, se encuentran valores consecutivos y_1, y_2, y_3, ..., tales que cada valor y_{n+1} se obtiene del inmediatamente anterior y_n, donde $x_{n+1} = x_n + h$, siendo h el *tamaño del paso*. Se hace un avance o integración en el paso a la vez, pudiéndose utilizar un tamaño del paso h igual o diferente en cada avance. Si se desea un avance hacia atrás basta con escoger un valor negativo para h.

Métodos más complejos como el método de Taylor o Runge-Kutta serán considerado también método de un solo paso.

1.1.1. Método de Euler

El método de Euler es el más sencillo de todos los métodos de un solo paso. Su fórmula algorítmica se basa en hallar una pendiente de recta apropiada para saltar cada paso.

• Simple

El método de *Euler simple* se basa en que el valor siguiente y_{n+1} se obtiene de y_n a partir de

$$y_{n+1} = y_n + h\, f(x_n, y_n) + O(h^2) \qquad \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \tag{3}$$

En el extremo derecho se ha colocado este método en la notación de Butcher que se verá en la sección 1.1.4.

El valor $f(x_n, y_n)$ es la pendiente de recta de la solución en el punto x_n, como indica (1). La solución numérica viene a ser polígono de tramos rectos cada uno con una pendiente diferente en el punto precedente y_n.

EJEMPLO:

De forma de ilustrar el uso del método, se presenta la solución para la siguiente ecuación diferencial

$$\frac{dy}{dx} = k\,y \qquad \text{con} \qquad y(0) = 1$$

cuya solución analítica es $y = \exp(kt)$. Utilizando el método de Euler verificar la exactitud de la solución propuesta, se evalua la expresión obtenida para un valor de $k = 1$ y los valores obtenidos se presentan en la siguiente tabla con distintos pasos de integración h

Resultados

	y_n			
x_n	$h = 0.2$	$h = 0.1$	$h = 0.05$	e^x
0.0	1.000	1.000	1.000	1.000
0.1	− − −	1.100	1.103	1.105
0.2	1.200	1.210	1.216	1.221
0.4	1.440	1.464	1.478	1.492
0.8	2.072	2.143	2.184	2.226
1.0	2.487	2.593	2.654	2.718

En la tabla es posible apreciar que cuanto menor es el paso de integración, menores son los errores.

• Modificado

El método de Euler se modifica de dos maneras distintas a saber. La primera forma (método de Heun) se formula en la siguientes dos expresiones

$$\begin{aligned} y_{n+1} &= y_n + h\,f(x_n, y_n) + O(h^2) \\ y_{n+1} &= y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right] + O(h^3) \end{aligned} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array} \tag{4}$$

La primera fórmula es la de Euler simple (3) y se usa como predictora, la segunda fórmula utiliza una pendiente de recta promedio entre los puntos de x_n y x_{n+1} y se usa como correctora. Del lado derecho se coloca en la notación de Butcher. Cuando se corrige una sóla vez, el método se considera explícito. Cuando se corrige más de una vez, el método se considera implícito y en la matriz de Butcher habría que cambiar el 1 de la posición a_{21} a a_{22}.

La segunda forma (método del polígono) se formula con las siguientes dos expresiones

$$\begin{aligned} y_{n+1/2} &= y_n + \frac{h}{2}\,f(x_n, y_n) + O(h^2) \\ y_{n+1} &= y_n + h\,f(x_{n+1/2}, y_{n+1/2}) + O(h^3) \end{aligned} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array} \tag{5}$$

La primera fórmula es la Euler simple (3) usada para estimar el punto medio $y_{n+1/2}$ con $h/2$, donde luego se utiliza la pendiente del punto medio $f(x_{n+1/2}, y_{n+1/2})$ para con Euler simple de nuevo calcular y_{n+1}. Del lado derecho está la notación de Butcher para este método.

La diferencia entre estos dos métodos en la forma como se calcula la pendiente usada. En la primera forma es la media de las pendientes (inicial y final), en la segunda forma es la pendiente del punto medio. Los órdenes de los errores locales son de h^3 para los métodos modificados, a diferencia del método de Euler simple que era de h^2.

1.1.2. Método de Taylor

El método de Taylor se basa en la expansión de series de Taylor

$$y_{n+1} = y_n + h\,y'(x_n) + \frac{1}{2!}\,h^2\,y''(x_n) + \frac{1}{3!}\,h^3\,y'''(x_n) + \cdots + \frac{1}{P!}\,h^P\,y^{(P)}(x_n) + O(h^{P+1}) \tag{6}$$

donde las diferentes derivadas se calculan aplicando la regla de la cadena, puesto que

$$y'(x) = f(x,y) \qquad y''(x) = f_x + f_y y' \qquad y'''(x) = f_{xx} + f_{yx}y' + f_y y'' + f_{yy}(y')^2 \tag{7}$$

etc. y deben evaluarse en $x = x_n$.

EJEMPLO:

De forma de ilustrar el uso del método, se presenta la solución para la siguiente ecuación diferencial

$$\frac{dy}{dx} = k\,y \qquad\qquad \text{con} \qquad y(0) = 1$$

cuya solución analítica es $y = \exp(kt)$. Utilizando el método de Taylor se deben evaluar las derivadas primera, segunda y sucesivas, por lo cual, derivando la ecuación diferencial, se obtiene

$$\frac{d^n y}{dx^n} = k^n\,y$$

Para verificar la exactitud de la solución propuesta, se evalua la expresión obtenida para un valor de $k = 1$ y los valores obtenidos se presentan en la siguiente tabla

Resultados

	y_n			
x_n	$P = 1$	$P = 3$	$P = 5$	e^x
0.0	1.000	1.000	1.000	1.000
0.1	1.100	1.105	1.105	1.105
0.2	1.200	1.221	1.221	1.221
0.3	1.300	1.350	1.350	1.350
0.5	1.500	1.646	1.649	1.649
1.0	2.000	2.667	2.717	2.718
2.0	3.000	6.333	7.267	7.389

En la tabla es posible apreciar dos características sumamente importantes del método de Taylor, una de ellas es que a medida que nos alejamos del centro de la serie para un valor de P fijo, los errores con

respecto a la solución exacta tienden a incrementarse; y la segunda es que a medida que el valor de P se incrementa para un mismo valor de x, la solucion obtenida se acerca rapidamente a la solución exacta.

1.1.3. Método Runge-Kutta

Un método Runge-Kutta al igual que los meétodos anteriores son métodos de un solo paso. Un método de N etapas y orden P tiene la siguiente fórmula algorítmica

$$y_{n+1} = y_n + h\,\varphi_N(x_n, y_n) + O(h^{P+1}) \tag{8}$$

donde $\varphi(x_n, y_n)$ es la ponderación de varias pendientes de recta k_s en el intervalo $[x_n, x_{n+1}]$

$$\varphi_N(x_n, y_n) = c_1k_1 + c_2k_2 + c_3k_3 + \cdots + c_Nk_N \qquad \sum_{s=1}^{N} c_s = 1 \tag{9}$$

y las variables auxiliares k_s se definen como

$$\begin{aligned}
k_1 &= f(x_n, y_n)\\
k_2 &= f(x_n + b_2h, y_n + h\,a_{21}k_1)\\
k_3 &= f(x_n + b_3h, y_n + h\,a_{31}k_1 + h\,a_{32}k_2)\\
k_4 &= f(x_n + b_4h, y_n + h\,a_{41}k_1 + h\,a_{42}k_2 + h\,a_{43}k_3)\\
&\vdots \qquad\qquad \vdots\\
x_N &= f(x_n + b_Nh, y_n + h\,a_{N1}k_1 + h\,a_{N2}k_2 + \cdots + h\,a_{N,N-1}k_{N-1})
\end{aligned} \tag{10}$$

No siempre el número de etapas coincide con el orden.

• Segundo Orden

Este método coincide con el método de Euler modificado tipo Heun con una sola correción (sección 1.1.1 • Modificado)

$$\begin{aligned}
k_1 &= f(x_n, y_n)\\
k_2 &= f(x_n + h, y_n + h\,k_1)\\
y_{n+1} &= y_n + \frac{h}{2}\,(k_1 + k_2) + O(h^3)
\end{aligned} \qquad \begin{array}{c|cc} 0 & 0 & 0\\ 1 & 1 & 0\\ \hline & 1/2 & 1/2 \end{array} \tag{11}$$

El método de Ralston es

$$\begin{aligned}
k_1 &= f(x_n, y_n)\\
k_2 &= f(x_n + 3h/4, y_n + h\,3k_1/4)\\
y_{n+1} &= y_n + \frac{h}{3}\,(k_1 + 2\,k_2) + O(h^3)
\end{aligned} \qquad \begin{array}{c|cc} 0 & 0 & 0\\ 3/4 & 3/4 & 0\\ \hline & 1/3 & 2/3 \end{array} \tag{12}$$

• Tercer Orden

El método de Ralston & Rabinowitz es el siguiente

$$\begin{aligned}
k_1 &= f(x_n, y_n)\\
k_2 &= f(x_n + h/2, y_n + h\,k_1/2)\\
k_3 &= f(x_n + h, y_n - h\,k_1 + 2\,h\,k_2)\\
y_{n+1} &= y_n + \frac{h}{6}\,(k_1 + 4\,k_2 + k_3) + O(h^4)
\end{aligned} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0\\ 1/2 & 1/2 & 0 & 0\\ 1 & -1 & 2 & 0\\ \hline & 1/6 & 2/3 & 1/6 \end{array} \tag{13}$$

• Cuarto Orden

El método de Kutta del primer tipo es

$$
\begin{aligned}
k_1 &= f(x_n, y_n) \\
k_2 &= f(x_n + h/2, y_n + h\,k_1/2) \\
k_3 &= f(x_n + h/2, y_n + h\,k_2/2) \\
k_4 &= f(x_n + h, y_n + h\,k_3) \\
y_{n+1} &= y_n + \frac{h}{6}\,(k_1 + 2\,k_2 + 2\,k_3 + k_4) + O(h^5)
\end{aligned}
\qquad
\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
1/2 & 1/2 & 0 & 0 & 0 \\
1/2 & 0 & 1/2 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
\hline
 & 1/6 & 1/3 & 1/3 & 1/6
\end{array}
\tag{14}
$$

EJEMPLO:

De forma de ilustrar el uso del método, se presenta la solución para la siguiente ecuación diferencial

$$\frac{dy}{dx} = k\,y \qquad \text{con} \qquad y(0) = 1$$

cuya solución analítica es $y = \exp(kt)$. Utilizando el método de Kutta del primer tipo.

Para verificar la exactitud de la solución propuesta, se evalúa la expresión obtenida para un valor de $k = 1$ y los valores obtenidos se presentan en la siguiente tabla

Resultados

	y_n		
x_n	$h = 0.1$	$h = 0.5$	e^x
0.0	1.0	1.0	1.0
0.1	1.10517	− − −	1.1057
0.2	1.22140	− − −	1.22140
0.5	− − −	1.64844	1.64872
1.0	− − −	2.71781	2.71828

En la tabla anterior se evidencia la precisión del método de Runge-Kutta de cuarto orden clásico de tipo explicito, que aún incrementando el paso cinco veces puede estimar valores con bastante precisión. Esta característica es propia de la mayoria de los métodos de esta familia, por lo cual son los mas populares para hallar la solución numérica de ecuaciones diferenciales ordinarias.

El método de Kutta del segundo tipo es

$$
\begin{aligned}
k_1 &= f(x_n, y_n) \\
k_2 &= f(x_n + h/3, y_n + h\,k_1/3) \\
k_3 &= f(x_n + 2h/3, y_n - h\,k_1/3 + h\,k_2) \\
k_4 &= f(x_n + h, y_n + h\,k_1 - h\,k_2 + h\,k_3) \\
y_{n+1} &= y_n + \frac{h}{8}\,(k_1 + 3\,k_2 + 3\,k_3 + k_4) + O(h^5)
\end{aligned}
\qquad
\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 0 \\
2/3 & -1/3 & 1 & 0 & 0 \\
1 & 1 & -1 & 1 & 0 \\
\hline
 & 1/8 & 3/8 & 3/8 & 1/8
\end{array}
\tag{15}
$$

Existen métodos con coeficientes exóticos como el método de Gill

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & \frac{-1+\sqrt{2}}{2} & \frac{2-\sqrt{2}}{2} & 0 & 0 \\ 1 & 0 & -\frac{\sqrt{2}}{2} & \frac{2+\sqrt{2}}{2} & 0 \\ \hline & 1/6 & \frac{2-\sqrt{2}}{6} & \frac{2+\sqrt{2}}{6} & 1/6 \end{array} \tag{16}$$

Los valores de los coeficientes están alrededor de los coeficientes del método de Kutta del primer tipo (13).

• **Quinto Orden**

Un método de 5 etapas pero también de quinto orden es el método de Merson

$$\begin{aligned} k_1 &= f(x_n, y_n) \\ k_2 &= f(x_n + h/3, y_n + h\,k_1/3) \\ k_3 &= f(x_n + h/3, y_n + h\,k_1/6 + h\,k_2/6) \\ k_4 &= f(x_n + h/2, y_n + h\,k_1/8 + h\,3k_3/8) \\ k_5 &= f(x_n + h, y_n + h\,k_1/2 - h\,3k_3/2 + h\,2k_4) \\ y_{n+1} &= y_n + \frac{h}{6}\,(k_1 + 4\,k_4 + k_5) + O(h^6) \end{aligned}$$

$$\begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 & 0 \\ 1/3 & 1/6 & 1/6 & 0 & 0 & 0 \\ 1/2 & 1/8 & 0 & 3/8 & 0 & 0 \\ 1 & 1/2 & 0 & -3/2 & 2 & 0 \\ \hline & 1/6 & 0 & 0 & 2/3 & 1/6 \end{array} \tag{17}$$

cuyo error de truncamiento local se puede estimar en función de las k_s

$$E_{n+1} = y_{n+1} - \tilde{y}_{n+1} = \frac{-h}{6}\,(2\,k_1 - 9\,k_3 + 8\,k_4 - k_5) + O(h^5) \tag{18.a}$$

$$\tilde{y}_{n+1} = y_n + \frac{h}{2}\,(k_1 - 3\,k_3 + 4\,k_4) + O(h^5) \tag{18.b}$$

donde $\tilde{y}_{n+1}$ es la solución de cuarto orden que se calcula con la última línea de la matriz de arriba para el punto de colocación en $x_n + h$. El siguiente factor

$$R = \frac{|E_{n+1}|}{5} = \frac{|h|}{30}\ \left|\ 2\,k_1 - 9\,k_3 + 8\,k_4 - k_5\ \right| + O(h^5) \tag{18.c}$$

sirve para controlar el tamaño h del paso. Cuando $R > \epsilon_{max}$, entonces el paso se debe reducir $h' = h/2$ a la mitad. Cuando $R \leq \epsilon_{max}/64$, entonces el paso se puede incrementar $h' = 2\,h$ al doble. Cuando $\epsilon_{max}/64 < R \leq \epsilon_{max}$ entonces el paso h es satisfactorio y se deja como está [Hazewinkel,1988].

El método de Butcher es el siguiente. Aqui se cumple que el número de etapas $N = 6$ y el orden $P = 5$ no coinciden.

$$\begin{aligned} k_1 &= f(x_n, y_n) \\ k_2 &= f(x_n + h/4, y_n + h\,k_1/4) \\ k_3 &= f(x_n + h/4, y_n + h\,k_1/8 + h\,k_2/8) \\ k_4 &= f(x_n + h/2, y_n - h\,k_2/2 + h\,k_3) \\ k_5 &= f(x_n + 3h/4, y_n + h\,3k_1/16 + h\,9k_4/16) \\ k_6 &= f(x_n + h, y_n - h\,3k_1/7 + h\,2k_2/7 + h\,12k_3/7 - h\,12k_4/7 + h\,8k_5/7) \\ y_{n+1} &= y_n + \frac{h}{90}\,(7\,k_1 + 32\,k_3 + 12\,k_4 + 32\,k_5 + 7\,k_6) + O(h^6) \end{aligned}$$

0	0	0	0	0	0	0
1/4	1/4	0	0	0	0	0
1/4	1/8	1/8	0	0	0	0
1/2	0	−1/2	1	0	0	0
3/4	3/16	0	0	9/16	0	0
1	−3/7	2/7	12/7	−12/7	8/7	0
	7/90	0	16/45	2/15	16/45	7/90

(19)

1.2. NOTACION DE BUTCHER

Como es bien conocido, todo sistema de ecuaciones diferenciales de cualquier orden, con un conveniente cambio de variables, puede ser transformado en un sistema de ecuaciones diferenciales de primer orden [Gerald,1979][Burden & Faires,1985]. Por esta razón, estos últimos sistemas son los que se estudiarán en esta parte.

Sea el siguiente sistema de M ecuaciones diferenciales de primer orden

$$\frac{dy^i}{dx} = f^i(x, \mathbf{y}) \qquad i = 1, 2, 3, \ldots, M \tag{1}$$

siendo $\mathbf{y}$ una función M-dimensional con cada una de sus componentes dependiendo de x. Esto es

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(x, \mathbf{y}) \tag{2}$$

donde

$$\mathbf{y} = \mathbf{y}(x) = (y^1(x), y^2(x), y^3(x), \ldots, y^M(x)) \tag{3}$$

Cuando cada función $f^i(x, \mathbf{y})$ depende sólo de la variable y^i, se dice que el sistema está desacoplado, de lo contrario se dice que está acoplado. Si el sistema está desacoplado, entonces cada una de las ecuaciones diferenciales se puede resolver separadamente.

Cuando las condiciones de las solución de $\mathbf{y}(x)$ son conocidas en un único punto, por ejemplo

$$x = x_o \qquad y^i(x_o) = y_o^i \tag{4}$$

las expresiones (1) y (4) se dicen que conforman un "problema de valor inicial", de lo contrario se dice que es un "problema de valor en la frontera".

En realidad, el sistema (1) es una caso particular del caso más general expresado de la siguiente forma [Burden & Faires,1985][Gear,1971]

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(\mathbf{y}) \equiv \begin{cases} dy^i/dx = 1 & \text{if } i = 1 \\ dy^i/dx = f^i(\mathbf{y}) & \text{if } i = 2, 3, \ldots, M+1 \end{cases} \tag{5}$$

pero con el adicional cambio de variable $y_o^1 = x_o$ en (4).

Tratando de hacer una formulación general, se puede plantear al método Runge-Kutta de orden P y equipado con N etapas con la siguiente expresión [Gear,1971]

$$y_{n+1}^i = y_n^i + c_r\, h\, k_r^i \tag{6.a}$$

donde las variables M-dimensionales auxiliares bfk_r son calculadas de la forma

$$k_r^i = f^i(x_n + b_r h, \mathbf{y}_n + a_{rs}\, h\, \mathbf{k}_s) \tag{6.b}$$

para

$$i = 1, 2, 3, \ldots, M \qquad r, s = 1, 2, 3, \ldots, N \tag{6.c}$$

Nótese que se ha usado la notación indicial, de manera que si un índice aparece dos veces (ó más) en un término, se debe realizar una sumatoria en todo su rango (en este contexto, no es importante el número de factores con el mismo índice en cada término).

Un método Runge-Kutta (6) tiene orden P, si para un problema lo suficientemente suave del tipo (2) y (4), se tiene que

$$\|\mathbf{y}(x_n + h) - \mathbf{y}_{n+1}\| \leq \Phi(\zeta)\ h^{P+1} = O(h^{P+1}) \qquad \zeta \in [x_n, x_n + h], \tag{7}$$

es decir, si la expansión en series de Taylor para la solución exacta $\mathbf{y}(x_n + h)$ del problema y la solución aproximada $\mathbf{y}_{n+1}$ coinciden hasta (e incluyendo) el término del orden de h^P [Lapidus & Seinfeld,1971].

El método Runge-Kutta antes definido se puede aplicar para resolver un problema de valor inicial y se usa recurrentemente. Dado un punto $(x_n, \mathbf{y}_n)$, el punto siguiente $(x_{n+1}, \mathbf{y}_{n+1})$ se obtiene usando la expresion (6), siendo

$$x_{n+1} = x_n + h \tag{8}$$

y h el paso del método. Cada vez que se hace este procedimiento, el método avanza hacia adelante (ó hacia atrás si h es negativo) un paso de integración h en x, ofreciendo la solución en puntos consecutivos, uno para cada salto. De esta forma, si el método comienza con el punto $(x_0, \mathbf{y}_0)$ definido por (4), entonces luego se pueden calcular $(x_1, \mathbf{y}_1), (x_2, \mathbf{y}_2), (x_3, \mathbf{y}_3), \ldots, (x_n, \mathbf{y}_n)$, y continuar de esta forma, hasta la frontera deseada en x. Cada integración o salto el método se reinicializa con la información del punto precedente inmediatamente anterior, por ello el método Runge-Kutta se considera dentro del grupo de métodos denominados de un sólo paso. No obstante, se debe notar que las variables auxiliares k_r^i son calculadas para todo r hasta N en cada paso. Estos cálculos no son más que evaluaciones de $f^i(x, \mathbf{y})$ para puntos intermedios $x + b_r h$ en el intervalo $[x_n, x_{n+1}]$ $(0 \leq b_r \leq 1)$, pero pre-multiplicadas por h (esta multiplicación por h puede hacerse al final, lo que hace al método más eficiente). La evaluación de cada variable M-dimensional auxiliar $\mathbf{k}_r$, representa una etapa del método.

Ahora se introduce una representación condensada del método Runge-Kutta generalizado, originalmente desarrollada por Butcher [(1964)]. Esta representación matricial del método Runge-Kutta se presenta de forma sistemática en las referencias [Lapidus & Seinfeld,1971], [Hairer et al.,1987] y [Hairer & Wanner,1991], siendo las dos últimas un par de catálogos de todos los métodos Runge-Kutta imaginables. Después del artículo de Butcher [(1964)] se ha vuelto costumbre simbolizar un método Runge-Kutta (6) con valores ordenados de forma tabular. Con la finalidad de ilustrar la *notación de Butcher*, como se le denomina actualmente, considérese (6) aplicado a un método de cuatro etapas ($N = 4$). Acomodando los coeficientes a_{rs}, b_r y c_r de forma ordenada como en la siguiente tabla matricial

$$\begin{array}{c|cccc} b_1 & a_{11} & a_{12} & a_{13} & a_{14} \\ b_2 & a_{21} & a_{22} & a_{23} & a_{24} \\ b_3 & a_{31} & a_{32} & a_{33} & a_{34} \\ b_4 & a_{41} & a_{42} & a_{43} & a_{44} \\ \hline & c_1 & c_2 & c_3 & c_4 \end{array} \qquad \begin{array}{l} 0 \leq b_r \leq 1 \\ \displaystyle\sum_{s=1}^{N} a_{rs} = b_r \\ \displaystyle\sum_{r=1}^{N} c_r = 1 \end{array} \tag{9}$$

con valores particulares, se obtiene la notación de Butcher del método en particular.

La representación anterior permite hacer una distinción básica para los distintos métodos Runge-Kutta, de acuerdo a las características de la matriz a_{rs}: Si $a_{rs} = 0$ para $s \geq r$, entonces la matriz a_{rs} es triangular inferior, excluyendo la diagonal principal, y el método se clasifica como *completamente explícito*. Si, $a_{rs} = 0$ para $s > r$, entonces la matriz a_{rs} es triangular inferior, pero incluyendo la diagonal principal, y el método se clasifica como *semi-implícito* ó *simple-diagonalmente implícito*. Si la matriz a_{rs} es diagonal por bloques,

se dice que el método es *diagonalmente implícito* (por bloques). Si la primera fila de la matriz a_{rs} está llena de ceros, $a_{1,s} = 0$, y el método es diagonalmente implícito, entoces se denomina *método de Lagrange* [van der Houwen & Sommeijer,1991] (los coeficientes b_r pueden ser arbitrarios). Si un método de Lagrange tiene $b_N = 1$ y la última fila es el arreglo $a_{N,s} = c_s$, entonces el método se dice que es *rígidamente preciso*. Si, contrariamente, ninguna de las condiciones previas son satisfechas, el método se clasifica de implícito. Cuando ningún elemento de la matriz a_{rs} es nulo, se dice que el método es completamente implícito. En los casos de los métodos Runge-Kutta implícitos, se debe hacer notar que una variable auxiliar $\mathbf{k}_r$ puede depender de ella misma y de otras variables auxiliares no calculadas hasta el momento en la misma etapa. Es por ello, que estos métodos se llaman implícitos en estos casos.

Adicionalmente, la representación arriba descrita, permite verificar muy fácilmente las propiedades que los coeficientes a_{rs}, b_r, y c_r deben tener. En particular, se deben satisfacer las siguientes propiedades

$$0 \leq b_r \leq 1 \qquad a_{rs}\,\delta_s = b_r \qquad c_r\,\delta_r = 1 \tag{10.a,b,c}$$

donde el vector δ es unitario en todas sus componentes ($\delta_r = 1\ \forall r = 1, 2, 3, \ldots, N$). Las anteriores propiedades pueden interpretarse de la siguiente manera: La propiedad $(10.a)$ expresa que el método Runge-Kutta es un método de un sólo paso, y que las funciones $f^i(x, \mathbf{y}(x))$ en $(6.b)$ deben ser evaluadas para $x \in [x_n, x_{n+1}]$. La propiedad $(10.b)$ resulta de aplicar el método Runge-Kutta (6) a un sistema de ecuaciones diferenciales del tipo (5), donde $k_s^1 = 1\ \forall s = 1, 2, 3, \ldots, N$, y así la suma de a_{rs} en cada línea r ofrece el valor de b_r. La propiedad $(10.c)$ significa que en la expresión $(6.a)$, el valor de y_{n+1}^i es obtenido del valor de y_n^i, proyectando con h un promedio de las derivadas $dy^i/dx = f^i(x, \mathbf{y})$ en los puntos intermedio del paso. Este promedio se hace con los coeficientes de peso c_r, por lo que la suma obviamente debe ser la unidad.

Los coeficientes a_{rs}, b_r y c_r son determinados mediante la aplicación de las propiedades (10) y usando algunas relaciones que son deducidas de la siguiente manera:
Sea el siguiente sistema de ecuaciones diferenciales ordinarias de primer orden expresado de acuerdo a (5) como un problema de valor inicial del tipo

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(\mathbf{y}) \tag{5'}$$

$$x = x_0 \qquad \mathbf{y}(x_0) = \mathbf{y}_0 \tag{4'}$$

El método Runge-Kutta aplicado a este problema se formula como

$$\mathbf{y}_{n+1} = \mathbf{y}_n + c_r \mathbf{k}_r \tag{6.a'}$$

donde las variables auxiliares $\mathbf{k}_r$ se definen como

$$\mathbf{k}_r = h\,\mathbf{f}(\mathbf{y}_n + a_{rs}\mathbf{k}_s) \tag{6.b'}$$

Si ahora se hace una expansión en serie de Taylor a la componente k_r^i de $(6.b')$, alrededor del punto $(x_n, \mathbf{y}_n)$, siendo $\mathbf{y}_n = \mathbf{y}(x_n)$, resulta que

$$\begin{aligned} k_r^i =& h\,f^i\,[\delta_r] + h\,f_j^i\,[a_{rs}k_s^j] + \frac{h}{2}\,f_{jk}^i\,[a_{rs}k_s^j]\,[a_{rt}k_t^k] \\ &+ \frac{h}{6}\,f_{jkl}^i\,[a_{rs}k_s^j]\,[a_{rt}k_t^k]\,[a_{ru}k_u^l] \\ &+ \frac{h}{24}\,f_{jklm}^i\,[a_{rs}k_s^j]\,[a_{rt}k_t^k]\,[a_{ru}k_u^l]\,[a_{rv}k_v^m] + O(h^6) \end{aligned} \tag{11.a}$$

donde la regla del índice repetido y la siguiente notación ha sido usada

$$f^i = f^i(x_n) \qquad f_j^i = \frac{\partial f^i}{\partial y^j}\Big|_{\mathbf{y}_n} \qquad f_{jk}^i = \frac{\partial^2 f^i}{\partial y^j \partial y^k}\Big|_{\mathbf{y}_n} \qquad \cdots \tag{11.b}$$

Aquí las functiones se suponen del tipo C^∞ (funciones analíticas), y por consiguiente los índices en (11.b) son permutables.

La variable k_s^j en el segundo término del miembro de la derecha de (11.a) puede de nuevo ser expandida en serie de Taylor como

$$\begin{aligned} k_s^j =& h\, f^j\,[\delta_s] + h\, f_k^j\,[a_{s\alpha}k_\alpha^k] + \frac{h}{2}\, f_{kl}^j\,[a_{s\alpha}k_\alpha^k]\,[a_{s\beta}k_\beta^l] \\ &+ \frac{h}{6}\, f_{klm}^j\,[a_{s\alpha}k_\alpha^k]\,[a_{s\beta}k_\beta^l]\,[a_{s\gamma}k_\gamma^m] + O(h^5) \end{aligned} \tag{11.c}$$

De la misma manera k_α^k puede ser expandida como

$$k_\alpha^k = h\, f^k\,[\delta_\alpha] + h\, f_l^k\,[a_{\alpha\delta}k_\delta^l] + \frac{h}{2}\, f_{lm}^k\,[a_{\alpha\delta}k_\delta^l]\,[a_{\alpha\epsilon}k_\epsilon^m] + O(h^4) \tag{11.d}$$

y así sucesivamente

$$k_\delta^l = h\, f^l\,[\delta_\delta] + h\, f_m^l\,[a_{\delta\varphi}k_\varphi^m] + O(h^3) \tag{11.e}$$

hasta

$$k_\varphi^m = h\, f^m\,[\delta_\varphi] + O(h^2) \tag{11.f}$$

Si finalmente se hace una recurrente substitución regresiva, se obtiene que

$$\begin{aligned} k_r^i =& h\, f^i\,[\delta_r] + h^2\,[f_j^i f^j b_r] + h^3\,[f_j^i f_k^j f^k a_{rs} b_s + \frac{1}{2} f_{jk}^i f^j f^k b_r^2] \\ &+ h^4\,[f_j^i f_k^j f_l^k f^l a_{rs} a_{st} b_t + \frac{1}{2} f_j^i f_{kl}^j f^k f^l a_{rs} b_s^2 \\ &\quad + f_{jk}^i f_l^j f^k f^l b_r a_{rs} b_s + \frac{1}{6} f_{jkl}^i f^j f^k f^l b_r^3] \\ &+ h^5\,[f_j^i f_k^j f_l^k f_m^l f^m a_{rs} a_{st} a_{tu} b_u + \frac{1}{2} f_j^i f_k^j f_{lm}^k f^l f^m a_{rs} a_{st} b_t^2 \\ &\quad + f_j^i f_{kl}^j f_m^k f^l f^m a_{rs} b_s a_{st} b_t + \frac{1}{6} f_j^i f_{klm}^j f^k f^l f^m a_{rs} b_s^3 \\ &\quad + f_{jk}^i f_l^j f^k f_m^l f^m b_r a_{rs} a_{st} b_t + \frac{1}{2} f_{jk}^i f_{lm}^j f^k f^l f^m b_r a_{rs} b_s^2 \\ &\quad + \frac{1}{2} f_{jk}^i f_l^j f_m^k f^l f^m a_{rs} b_s a_{rt} b_t + \frac{1}{2} f_{jkl}^i f_m^j f^k f^l f^m b_r^2 a_{rs} b_s \\ &\quad + \frac{1}{24} f_{jklm}^i f^j f^k f^l f^m b_r^4] + O(h^6) \end{aligned} \tag{11.g}$$

Insertando esta última expresión de los componentes de $\mathbf{k}_r$ en la ecuación (6.a'), y comparando luego con la siguiente expansión en series de Taylor de $\mathbf{y}_{n+1}$ (Esta expansión se desarrolla alrededor del punto $\mathbf{y}_n$)

$$\begin{aligned} y_{n+1}^i &= y_n^i + h\, f^i + \frac{h^2}{2}(f_j^i f^j) + \frac{h^3}{6}(f_j^i f_k^j f^k + f_{jk}^i f^j f^k) \\ &+ \frac{h^4}{24}(f_j^i f_k^j f_l^k f^l + f_j^i f_{kl}^j f^k f^l + 3 f_{jk}^i f_l^j f^l f^k + f_{jkl}^i f^j f^k f^l) \\ &+ \frac{h^5}{120}(f_j^i f_k^j f_l^k f_m^l f^m + f_j^i f_k^j f_{lm}^k f^l f^m + 3 f_j^i f_{kl}^j f_m^k f^l f^m + f_j^i f_{klm}^j f^k f^l f^m + 4 f_{jk}^i f_l^j f^k f_m^l f^m \\ &\quad + 4 f_{jk}^i f_{lm}^j f^k f^l f^m + 3 f_{jk}^i f_l^j f_m^k f^l f^m + 6 f_{jkl}^i f_m^j f^k f^l f^m + f_{jklm}^i f^j f^k f^l f^m) + O(h^6) \end{aligned} \tag{11.h}$$

resultan las siguientes relaciones que deben satisfacerse por los coeficientes a_{rs}, b_r y c_r para un método Runge-Kutta de hasta quinto orden

$$
\begin{array}{c|c}
h & c_r\delta_r = 1 \\
\hline
h^2 & c_r b_r = 1/2 \\
\hline
h^3 & \begin{array}{c} c_r a_{rs} b_s = 1/6 \\ c_r b_r^2 = 1/3 \end{array} \\
\hline
h^4 & \begin{array}{c} c_r a_{rs} a_{st} b_t = 1/24 \\ c_r a_{rs} b_s^2 = 1/12 \\ c_r b_r a_{rs} b_s = 1/8 \\ c_r b_r^3 = 1/4 \end{array} \\
\hline
h^5 & \begin{array}{c} c_r a_{rs} a_{st} a_{tu} b_u = 1/120 \\ c_r a_{rs} a_{st} b_t^2 = 1/60 \\ c_r a_{rs} b_s a_{st} b_t = 1/40 \\ c_r a_{rs} b_s^3 = 1/20 \\ c_r b_r a_{rs} a_{st} b_t = 1/30 \\ c_r b_r a_{rs} b_s^2 = 1/15 \\ c_r a_{rs} b_s a_{rt} b_t = 1/20 \\ c_r b_r^2 a_{rs} b_s = 1/10 \\ c_r b_r^4 = 1/5 \end{array}
\end{array}
\qquad (12)
$$

En estas relaciones, b_r se ha definido de acuerdo a la propiedad (10.b). Nótese también que se han usado expansiones de las series de Taylor hasta el término de quinto orden (con h^5) en el desarrollo de las anteriores relaciones. Por consiguiente, las relaciones (12) son válidas para los métodos Runge-Kutta, tanto explícitos como implícitos, desde el primer orden (e.g. Método de Euler), pasando por los de segundo orden (e.g. Método de Euler modificado en sus variantes del paso medio ó del trapecio), los de tercer y cuarto órdenes (e.g. Métodos de Kutta), hasta el método de quinto orden (e.g. método Fehlberg y método de Cash & Karp) y de sexto orden (e.g. Métodos basados en las cuadraturas de Gauss-Legendre y de Lobatto). En todos los casos los índices r, s, t y u varían desde 1 hasta N, que es el número de etapas.

Gear [1971] presenta una deducción similar a (11), pero sólo para métodos explícitos. En Hairer et al. [1987], aparecen relaciones similares a (12), pero sólo para métodos explícitos hasta de cuarto orden. En esta última referencia aparece un teorema que resalta la equivalencia entre el método Runge-Kutta y los métodos de colocación ortogonal. El siguiente teorema [Hairer & Wanner,1991] resume los resultados de (12) de una forma más concisa:

Teorema [Butcher,(1964a)] [Hairer et al.,1987,pp.203-204]. Sea la siguiente condición definida como

$$
\begin{aligned}
&\mathcal{B}(P) && \sum_{i=1}^{N} c_i b_i^{q-1} = \frac{1}{q} && q = 1,2,\ldots,P \\
&\mathcal{C}(\eta) && \sum_{j=1}^{N} a_{ij} b_j^{q-1} = \frac{b_i^q}{q} && i = 1,2,\ldots,N \qquad q = 1,2,\ldots,\eta \\
&\mathcal{D}(\xi) && \sum_{i=1}^{N} c_i b_i^{q-1} a_{ij} = \frac{c_j^q}{q}\,(1 - b_j^q) && j = 1,2,\ldots,N \qquad q = 1,2,\ldots,\xi
\end{aligned}
\qquad (12')
$$

Si los coeficientes b_i, c_i y a_{ij} de un método Runge-Kutta satisfacen las condiciones $\mathcal{B}(P)$, $\mathcal{C}(\eta)$ y $\mathcal{D}(\xi)$, con $P \le \eta + \xi + 1$ and $P \le 2\eta + 2$, entonces el método es de orden P.

1.3. CONTROL DEL PASO

1.3.1. Análisis del Error

Sean los coeficientes de la cuadratura de Lobatto

$$
\begin{array}{ccccc|c}
0 & 0 & 0 & 0 & 0 & 0 \\
(5-\sqrt{5})/10 & (5+\sqrt{5})/60 & 1/6 & (15-7\sqrt{5})/60 & 0 & 0 \\
(5+\sqrt{5})/10 & (5-\sqrt{5})/60 & (15+7\sqrt{5})/60 & 1/6 & 0 & 0 \\
\hline
1 & 1/6 & (5-\sqrt{5})/12 & (5+\sqrt{5})/12 & 0 & 0 \\
\hline
 & 1/12 & 5/12 & 5/12 & 1/12 &
\end{array}
\qquad (1.a)
$$

Este método será denominado como el "principal" de sexto orden ($P = 6$).

Dentro de los coeficientes del método principal, pueden ser detectados una parte de ellos que forman otro método Runge-Kutta "secundario" empotrado en el primero. Este otro método es de tercer orden ($\tilde{P} = 3$), tiene tres etapas ($\tilde{N} = 3$) y en la notación de Butcher son

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 \\ (5-\sqrt{5})/10 & (5+\sqrt{5})/60 & 1/6 & (15-7\sqrt{5})/60 \\ (5+\sqrt{5})/10 & (5-\sqrt{5})/60 & (15+7\sqrt{5})/60 & 1/6 \\ \hline & 1/6 & (5-\sqrt{5})/12 & (5+\sqrt{5})/12 \end{array} \qquad (1.b)$$

Ambos métodos, el principal y el secundario, constituyen lo que se denomina la forma de la cuadratura de Lobatto empotrada de tercer y sexto órdenes con cuatro etapas (el método de Fehlberg [1971] posee una forma similar, pero es explícito). El método Runge-Kutta implícito de sexto orden y cuatro etapas definido por los coeficientes $(1.a)$, en realidad representa dos métodos: uno de tercer orden y tres etapas, empotrado en el otro de sexto orden y cuatro etapas. Es decir, los coeficientes $(1.b)$ están incluidos en $(1.a)$. Este aspecto es relevante para controlar el tamaño del paso. Resolviendo un sistema de ecuaciones diferenciales ordinarias para los mismos coeficientes, se obtienen con un sólo esfuerzo dos soluciones de diferentes órdenes en el error de truncamiento local, reduciendo a un mínimo el número de cálculos.

Fehlberg [1971] reportó este aspecto al diseñar un algoritmo del control del paso para su método Runge-Kutta-Fehlberg explícito de cuarto y quinto órdenes empotrado ó encapsulado completamente uno en el otro. Por ejemplo,

$$\begin{array}{c|ccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ \frac{3}{8} & \frac{3}{32} & \frac{9}{32} & 0 & 0 & 0 & 0 \\ \frac{12}{13} & \frac{1932}{2197} & -\frac{7200}{2197} & \frac{7296}{2197} & 0 & 0 & 0 \\ 1 & \frac{439}{216} & -8 & \frac{3680}{513} & -\frac{845}{4104} & 0 & 0 \\ \frac{1}{2} & -\frac{8}{27} & 2 & -\frac{3544}{20520} & \frac{1859}{4104} & -\frac{11}{40} & 0 \\ \hline \mathbf{4^{to}} & \frac{25}{216} & 0 & \frac{1408}{2565} & \frac{2197}{4104} & -\frac{1}{5} & 0 \\ \mathbf{5^{to}} & \frac{16}{135} & 0 & \frac{6656}{12825} & \frac{28561}{56430} & -\frac{9}{50} & \frac{2}{55} \end{array} \qquad (2.a)$$

$$\begin{array}{c|ccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 & 0 \\ \frac{3}{10} & \frac{3}{40} & \frac{9}{40} & 0 & 0 & 0 & 0 \\ \frac{3}{5} & \frac{3}{10} & -\frac{9}{10} & \frac{6}{5} & 0 & 0 & 0 \\ 1 & -\frac{11}{54} & \frac{5}{2} & -\frac{70}{27} & \frac{35}{27} & 0 & 0 \\ \frac{7}{8} & \frac{1631}{55296} & \frac{175}{512} & \frac{575}{13824} & \frac{44275}{110592} & \frac{253}{4096} & 0 \\ \hline \mathbf{4^{to}} & \frac{37}{378} & 0 & \frac{250}{621} & \frac{125}{594} & 0 & \frac{512}{1771} \\ \mathbf{5^{to}} & \frac{2825}{27648} & 0 & \frac{18575}{48384} & \frac{13525}{55296} & \frac{277}{14336} & \frac{1}{4} \end{array} \qquad (2.b)$$

donde existen dos juegos de coeficiente c_r, uno para el método de cuarto orden (línea de arriba) y otro para el método de quinto orden (línea de abajo). Debe observarse que los coeficientes expuestos antes en $(2.a)$ son

los originales de Fehlberg [1971]. El error en este caso entre el método de quinto orden $\tilde{y}^i_{n+1}$ y el de cuarto orden y^i_{n+1} en la tabla $(2.a)$ es

$$E^i_{n+1} = \tilde{y}^i_{n+1} - y^i_{n+1} = \frac{1}{752400}\,(\,2090\,k^i_1 - 22528\,k^i_3 - 21970\,k^i_4 + 15048\,k^i_5 + 27360\,k_6\,) + O(h^5_n) \qquad (3)$$

Los coeficientes particulares expuestos antes en $(2.b)$ fueron desarrollados por Cash & Karp [1990], y aunque están basados en la misma filosofía y orden, no son los originales de Fehlberg, pero algunos piensan que tiene un mejor comportamiento [Chapra & Canale,2007;p.759]. No obstante, los valores particulares encontrados por Cash & Karp hacen el método más eficiente que el método original de Fehlberg, con una mejora en las propiedades de los errores [Press et al.,1992]. Luego, siempre se continúa la integración con la solución y^i_{n+1}, en lugar de $\tilde{y}^i_{n+1}$ [Hairer et al.,1978;p.166], aunque sean de mayor $(P = \tilde{P} + 1)$ o menor $(P \geq \tilde{P} - 1)$ órdenes.

Sean $\mathbf{y}_n$ y $\tilde{\mathbf{y}}_{n+1}$ las soluciones del sistema de ecuaciones diferenciales ordinarias, ofrecidas por los métodos Runge-Kutta implícitos tipo Lobatto de sexto y tercer órdenes, respectivamente, empotrados en una sóla formulación como se describió antes en $(1.a)$. Esto es,

$$y^i_{n+1} = y^i_n + \frac{1}{12}(k^i_1 + 5k^i_2 + 5k^i_3 + k^i_4) \qquad (4)$$

$$\tilde{y}^i_{n+1} = y^i_n + \frac{1}{12}[2k^i_1 + (5 - \sqrt{5})k^i_2 + (5 + \sqrt{5})k^i_3] \qquad (5)$$

Las variables auxiliares $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_3$ y $\mathbf{k}_4$ son las mismas para ambas expresiones y son obtenidas usando el sistema de ecuaciones diferenciales ordinarias con los coeficientes $(1.a)$ y $(1.b)$.

Se denotará como E^i_{n+1} la diferencia entre la solución del método de sexto orden y el método de tercer orden, es decir, la ecuación (4) menos la ecuación (5). Esto es,

$$E^i_{n+1} = y^i_{n+1} - \tilde{y}^i_{n+1} = \frac{1}{12}[-k^i_1 + \sqrt{5}(k^i_2 - k^i_3) + k^i_4] + O(h^4_n) \qquad (6)$$

Si $\mathbf{y}(x_n)$ es la solución exacta de la ecuación diferencial en el valor $x = x_n$, entonces los errores de truncamiento local de las soluciones numéricas (4) y (5) son definidos respectivamente por

$$e^i_n = y^i_n - y^i(x_n) = O(h^7_{n-1}) \qquad (7)$$

$$\tilde{e}^i_n = \tilde{y}^i_n - y^i(x_n) = O(h^4_{n-1}) \qquad (8)$$

y luego

$$E^i_{n+1} = y^i_{n+1} - \tilde{y}^i_{n+1} = e^i_{n+1} - \tilde{e}^i_{n+1} = O(h^4_n) \qquad (9)$$

Recuérdese que, si el método Runge-Kutta es de orden P, el error de truncamiento local es de orden $P + 1$.

Si la expresión (9) se organiza de la siguiente forma

$$E^i_{n+1} = \Big[\frac{y^i_{n+1} - y^i(x_{n+1})}{y^i(x_{n+1})}\Big] y^i(x_{n+1}) - [\tilde{y}^i_{n+1} - y^i(x_{n+1})] \qquad (10)$$

se obtiene que

$$E^i_{n+1} = e^i_{(r),n+1} y^i(x_{n+1}) - \tilde{e}^i_{n+1} \qquad (10')$$

donde

$$e^i_{(r),n+1} = \Big[\frac{y^i_{n+1} - y^i(x_{n+1})}{y^i(x_{n+1})}\Big] \qquad (11)$$

es el error de truncamiento local relativo.

Si ahora se asume que $y^i(x_{n+1})$ es aproximado por y^i_{n+1} en el denominador de (11), se puede aplicar la desigualdad de Cauchy-Schwartz y la desigualdad triangular a la expresión (10′), y de esto resulta

$$|E^i_{n+1}| \leq |e^i_{(r),n+1}|\ |y^i(x_{n+1})| + |\tilde{e}^i_{n+1}| \leq e_{(r),max}\ |y^i_{n+1}| + \tilde{e}_{max} \tag{12}$$

donde $e_{(r),max}$ y $\tilde{e}_{max}$ son respectivamente las tolerancias para el error de truncamiento local relativo y absoluto de los métodos Runge-Kutta implícitos de sexto y tercer órdenes. La expresión (12) también significa que, para que la solución de la ecuación diferencial en un sólo paso sea aceptada, se debe verificar que

$$Q^i_n = \frac{|E^i_{n+1}|}{e_{(r),max}\ |y^i_{n+1}| + \tilde{e}_{max}} \leq 1 \tag{13}$$

siendo las tolerancias para los errores de truncamiento local relativo y absoluto propuestos por el usuario del algoritmo de control que se explicará a continuación.

Un método similar de cuarto ($\tilde{P} = 4$) y quinto orden ($P = 5$) empotrado, pero de siete etapas ($N = 7$), es el método DOPRI5 [Dormand & Prince,(1980)] [Hairer et al.,1978;p.178], cuyos coeficientes se emuestran a continuación

$$\begin{array}{c|ccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & 0 & 0 & 0 & 0 & 0 \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & 0 & 0 & 0 & 0 \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & 0 & 0 & 0 \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & 0 & 0 \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\hline
\mathbf{5^{to}} & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\mathbf{4^{to}} & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40}
\end{array} \tag{14}$$

y que da resultados muy buenos mostrados en la figura **1** de la sección 3.1 para la órbita de Arenstorff.

• 1.3.2. Algoritmo de Control

Sea h_{n+1} el tamaño del paso en el siguiente paso que tiende a hacer $Q^i_n \cong 1$. Teniendo en cuenta el orden de la diferencia E^i_{n+1} definida por (9), el parametro Q_n puede ser redefinido como

$$Q_n = \Big(\frac{h_n}{h_{n+1}}\Big)^{\tilde{P}+1} \qquad \tilde{P} = 3 \ \text{ó}\ 4 \tag{15}$$

donde ($\tilde{P} = \text{infimo}[P, \tilde{P}]$)

$$Q_n = \max_{1\leq i\leq M}\big(Q^i_n\big) \tag{16}$$

y así, resolviendo para h_{n+1}, se obtiene

$$h_{n+1} = h_n \Big(\frac{1}{Q_n}\Big)^{\alpha} = h_n S_n \tag{17}$$

con

$$S_n = \Big(\frac{1}{Q_n}\Big)^{\alpha} \qquad \alpha = \frac{1}{\tilde{P}+1} = 1/4 \ \text{ó}\ 1/5 \tag{18}$$

Para el método de Fehlberg descrito antes en (2), el exponente sería $\alpha = 1/5$, puesto que los errores más grandes provendrían del método con el menor orden, que en ese caso sería el de cuarto orden con $\tilde{P} = 4$.

Aquí es conveniente mencionar que Shampine et al.[1976] usan expresiones similares a (17) y (18) para controlar el tamaño del paso en el método Runge-Kutta de cuarto y quinto órdenes desarrollado originalmente por Fehlberg [1971], pero con algunas modificaciones, con la finalidad de garantizar que S_n siempre esté acotado en el intervalo $[S_{min}, S_{max}]$, y que h_{n+1} siempre sea más grande que el valor límite h_{min}. Adicionalmente, los mencionados autores multiplican S_n por un coeficiente C_q menor que la unidad para que h_{n+1} tienda a ser casi igual que h_n, y así hacer $Q_n \cong 1$, pero un poco menor. Todas las modificaciones descritas están resumidas a continuación

$$S_n = C_q \left(\frac{1}{Q_n}\right)^{\alpha} \qquad C_q = 0.9 \sim 0.99 \qquad \alpha = 1/4 \tag{19}$$

$$S'_n = \max(\min(S_n, S_{max}), S_{min}) \tag{20}$$

$$h_{n+1} = h_n S'_n \tag{21}$$

$$h'_{n+1} = \max(h_{n+1}, h_{min}) \tag{22}$$

Mientras que en [Shampine et al.,1976] el exponente α es $1/5$ en la expresión (19) para el método de Fehlberg, aquí dicho exponente es $1/4$ para el método de Cuadratura de Lobatto. En la mencionada referencia también se recomienda para los coeficientes y límites los valores $C_q = 0.9$, $S_{min} = 0.1$ and $S_{max} = 5$. El valor del mínimo paso de integración, h_{min}, se determina con la precisión del computador usado. En este trabajo se usaron los mismos valores antes citados para las expresiones de (19) a (22).

El procedimiento para calcular el valor óptimo del paso de integración, que permita satisfacer las tolerancias $e_{(r),max}$ y $\tilde{e}_{max}$, se describe a continuación:

- Estimado un tamaño de paso inicial h_n, el método Runge-Kutta implícito tipo Lobatto es utilizado para calcular las variables auxiliares k_1^i, k_2^i, k_3^i y k_4^i con la expresión del sistema de cuaciones diferenciales ordinarias, usando los coeficientes $(1.a)$ y con el proceso iterativo involucrado para resolver las $\mathbf{k}_s$, y usando los valores iniciales del problema.
- Las expresiones (4) y (5) permiten encontrar las soluciones y_{n+1}^i y $\tilde{y}_{n+1}^i$ de los métodos de sexto y tercer órdenes, respectivamente.
- La definición (6) permite calcular la diferencia E_{n+1}^i entre los dos métodos.
- Con la ecuación (14) se puede calcular los parámetros Q_n^i, y con la ecuación (16) se puede obtener el máximo de ellos.
- Las relaciones (19) a (22) determinan el valor del tamaño del paso siguiente h_{n+1}.
- Si $Q_n \leq 1$, la integración con el paso h_n (ó la aplicación del método Runge-Kutta desde x_n hasta x_{n+1}) se acepta y el paso h_{n+1} se considera el paso óptimo para la siguiente integración (ó la siguiente aplicación del método Runge-Kutta desde x_{n+1} hasta x_{n+2}).
- Si $Q_n > 1$, la integración con el paso h_n se rechaza y se repite todo el algoritmo de nuevo pero con $h_n = h'_{n+1}$ obtenido de (22).

Este procedimiento algunas veces incrementa el tamaño del paso, y otras veces lo disminuye, con la finalidad de garantizar que el error relativo $e_{(r),n+1}^i$ del método Runge-Kutta de sexto orden sea menor que la tolerancia $e_{(r),max}$, y que el error $\tilde{e}_{n+1}^i$ del meétodo Runge-Kutta de tercer orden sea menor que la tolerancia $\tilde{e}_{max}$. En cualquier caso, la solución del método Runge-Kutta será y_{n+1}^i, es decir, la solución con el método de sexto orden.

1.4. METODOS DE PASOS MULTIPLES

Las fórmulas de Adams-Bashforth (predictoras) y las fórmulas de Adams-Moulton (correctoras), debe utilizarse en parejas que tenga el mismo error de turncamiento local, para que el método predictor-corrector sea consistente. No obstante en el método de Euler modificado tipo Heun se han usado de órdenes h^2 y h^3 (sección 1.1.1 • Modificado).

1.4.1. Adams-Bashforth

Estas son las fórmulas predictoras

$$y_{n+1} = y_n + h\, f_n + \frac{1}{2}\, h^2 f'(\zeta) \tag{1.a}$$

$$y_{n+1} = y_n + \frac{h}{2}(3\, f_n - f_{n-1}) + \frac{5}{12}\, h^3 f''(\zeta) \tag{1.b}$$

$$y_{n+1} = y_n + \frac{h}{12}(23\, f_n - 16\, f_{n-1} + 5\, f_{n-2}) + \frac{3}{8}\, h^4 f'''(\zeta) \tag{1.c}$$

$$y_{n+1} = y_n + \frac{h}{24}(55\ f_n - 59\, f_{n-1} + 37\, f_{n-2} - 9\, f_{n-3}) + \frac{251}{720}\, h^5 f^{iv}(\zeta) \tag{1.d}$$

$$y_{n+1} = y_n + \frac{h}{720}(1901\, f_n - 2774\, f_{n-1} + 2616\, f_{n-2} - 1274\, f_{n-3} + 251\, f_{n-4}) + \frac{475}{1440}\, h^6 f^{v}(\zeta) \tag{1.e}$$

$$y_{n+1} = y_n + \frac{h}{720}(4277\, f_n - 7923\, f_{n-1} + 9982\, f_{n-2} - 7298\, f_{n-3} + 2877\, f_{n-4} - 475\, f_{n-5}) + \frac{19087}{60480}\, h^7 f^{vi}(\zeta) \tag{1.f}$$

La predicción se hace una sola vez, al inicio del proceso iterativo (con n constante).

1.4.2. Adams-Moulton

Estas son las fórmulas correctoras

$$y_{n+1} = y_n + h\,(f_{n+1}) - \frac{1}{2}\, h^2 f'(\zeta) \tag{2.a}$$

$$y_{n+1} = y_n + \frac{h}{2}\,(f_{n+1} + f_n) - \frac{1}{12}\, h^3 f''(\zeta) \tag{2.b}$$

$$y_{n+1} = y_n + \frac{h}{12}(5\, f_{n+1} + 8\, f_n - f_{n-1}) - \frac{1}{24}\, h^4 f'''(\zeta) \tag{2.c}$$

$$y_{n+1} = y_n + \frac{h}{24}(9\, f_{n+1} + 19\, f_n - 5\, f_{n-1} + f_{n-2}) - \frac{19}{720}\, h^5 f^{iv}(\zeta) \tag{2.d}$$

$$y_{n+1} = y_n + \frac{h}{720}(251\, f_{n+1} + 646\ f_n - 264\, f_{n-1} + 106\, f_{n-2} - 19\, f_{n-3}) - \frac{27}{1440}\, h^6 f^{v}(\zeta) \tag{2.e}$$

$$y_{n+1} = y_n + \frac{h}{1440}(475\, f_{n+1} + 1427\, f_n - 798\, f_{n-1} + 482\, f_{n-2} - 173\, f_{n-3} + 27\, f_{n-4}) - \frac{863}{60480}\, h^7 f^{vi}(\zeta) \tag{2.f}$$

Las correciones se pueden hacer las veces necesarias, hasta que el proceso iterativo converga, con cierta tolerancia. Una vez logrado un resultado, se salta en n al siguiente paso de integración $n+1$. No confundir "iteración" con "integración".

2. PROBLEMA DE VALOR EN LA FRONTERA

Un problema de valor en la frontera debe tener tantas condiciones como la suma de los órdenes de las ecuaciones diferenciales involucradas. Por ejemplo, si se tiene dos ecuaciones diferenciales ordinarias (una

sola variable independiente) de tercer y cuarto órdenes. Entonces, hacen falta siete condiciones para que el problema esté bien planteado, normalmente en derivadas menores a las superiores. Estas condiciones pueden darse en un solo punto para todas las variables, en cuyo caso estamos en la presencia de un problema de valor inicial. Pero eventualmente pueden darse las condiciones de forma mixta en la frontera. Si $x \in [a,b]$, entonces $x = a$ ó $x = b$ se denomina la frontera.

Sea la siguiente ecuación diferencial ordinaria de segundo orden

$$y'' = f(x, y, y') \qquad a \le x \le b \tag{1}$$

con condiciones en la frontera

$$y(a) = \alpha \qquad y(b) = \beta \tag{2}$$

Teorema de Unicidad. Sea al siguiente dominio

$$D = \begin{cases} a \le x \le b \\ -\infty < y < \infty \\ -\infty < y' < \infty \end{cases} \tag{3}$$

Si f, $\partial f/\partial y$ y $\partial f/\partial y'$ existen y son continuas en D y además:

1:) $\frac{\partial f}{\partial y} > 0$ en D.

2:) Existe un valor M, tal que $|\frac{\partial f}{\partial y'}| \le M$ en D.

Entonces el problema posee solución única.

2.1. TRANSFORMACION

Un problema de valor en la frontera tiene condiciones de valor o derivadas (de orden inferior a la mayor) en la frontera $x = a$ ó $x = b$, de forma mixta, algunas en $x = a$, algunas en $x = b$, o de formas combinadas, valores más derivadas.

Sea el siguiente problema de segundo orden de valor en la frontera

$$y'' = P(x)\, y' + Q(x)\, y + R(x) \tag{4}$$

$$y(a) = \alpha \qquad y(b) = \beta \tag{5}$$

Sea y_1 una solución, tal que

$$y_1'' = P(x)\, y_1' + Q(x)\, y_1 + R(x) \tag{6}$$

$$y_1(a) = \alpha \qquad y_1'(a) = 0 \tag{5}$$

entonces se postula

$$y(x) = y_1(x) + K\, y_2(x) \tag{6}$$

Substituyendo esto en la ecuación diferencial original, queda

$$K\, y_2'' = P(x)\, K\, y_2' + Q(x)\, K\, y_2 \tag{7}$$

$$y_2'' = P(x)\, y_2' + Q(x)\, y_2 \tag{8}$$

$$y_2(a) = 0 \qquad K\, y_2'(a) = y'(a) \tag{9}$$

Se fija $K = y'(a) \Longrightarrow y_2'(a) = 1$

$$\beta = y_1(b) + K\, y_2(b) \Longrightarrow K = \frac{\beta - y_1(b)}{y_2(b)} = y'(a) \tag{10}$$

$$y(x) = y_1(x) + \left[\frac{\beta - y_1(b)}{y_2(b)}\right] y_2(x) \tag{11}$$

Se ha transformado un problema de valor en la frontera en dos problemas de valor inicial, que combinados apropiadamente da la solución del problema original.

2.2. DISPARO

Sea la siguiente ecuación diferencial ordinaria de segundo orden

$$y'' = f(x, y, y') \qquad a \le x \le b \tag{1}$$

con condiciones en la frontera

$$y(a) = \alpha \qquad y(b) = \beta \tag{2}$$

Se formula el problema de valor inicial

$$y'' = f(x, y, y') \qquad a \le x \le b \tag{3}$$

$$y(a) = \alpha \qquad y'(a) = t \tag{4}$$

cuya solución $y = y(t, x)$ depende adicionalmente de t variable. Se define la función

$$g(t) = y(t, b) - \beta \tag{5}$$

Se desea hallar t, tal que $g(t) = 0$. Eta ecuación se puede resolver aplicando el método de la secante

$$t_{k+1} = t_k - \frac{g(t_k)\,(t_k - t_{k-1})}{g(t_k) - g(t_{k-1})} = t_k - \frac{[\,y(t_k, b) - \beta\,]\,(t_k - t_{k-1})}{y(t_k, b) - y(t_{k-1}, b)} \tag{6}$$

Para los primeros estimados, se puede tomar

$$t_0 = \frac{\beta - \alpha}{b - a} \qquad t_1 = \frac{\beta - \alpha - \delta}{b - a} \tag{7}$$

donde δ es la tolerancia con que se obtiene $y(b)$. El problema (1)-(2) de valor en en la frontera se ha convertido en un problema de valor inicial con t como valor inicial estimado e iterado (disparo con t) para hacer coincidir en la otra frontera ($x = b$) el otro valor $y(t, b) - \beta = 0$ (blanco).

2.3. DISCRETIZACION

Sea la siguiente ecuación diferencial ordinaria

$$y''(x) = P(x)\,y'(x) + Q(x)\,y(x) + R(x) \qquad a \le x \le b \tag{1}$$

$$y(a) = \alpha \qquad y(b) = \beta \tag{2}$$

Se pueden substiruir la aproximaciones obtenidas de la sección III.1.7

$$\begin{aligned} &y(x_i) y_i \\ &y'(x_i) = \frac{y_{i+1} - y_{i-1}}{2\,h} + O(h^2) \\ &y''(x_i) = \frac{y_{i+1} - 2\,y_i + y_{i-1}}{h^2} + O(h^2) \end{aligned} \tag{3}$$

Substituyendo en la ecuación diferencial

$$\frac{y_{i+1}-2\,y_i+y_{i-1}}{h^2}=P(x_i)\left(\frac{y_{i+1}-y_{i-1}}{2\,h}\right)+Q(x_i)\,y_i+R(x_i) \tag{4}$$

Reorganizando esta expresión queda

$$\left[\frac{h}{2}\,P(x_i)+1\right]y_{i-1}+\left[\,2+h^2\,Q(x_i)\,\right]y_i+\left[\frac{h}{2}\,P(x_i)-1\right]y_{i+1}=-h^2\,R(x_i) \tag{5}$$

Se obtienen n ecuaciones de este tipo con n incógnitas que son los valores y_i, $i=1,2,\ldots,n$ en intervalos regulares. En los puntos extremos considerar las condiciones de contorno $y_0=y(a)=\alpha$ y $y_{n+1}=y(b)=\beta$, $h=(b-a)/(n+1)$. La matriz de los coeficientes de este sistema de ecuaciones es una matriz tridiagonal (resolver con el algoritmo de Thomas sección II.1.1.7).

3. SISTEMAS DE ECUACIONES

Una ecuación homogénea del siguiente tipo

$$F\left(x,y,\frac{dy}{dx},\frac{d^2y}{dx^2},\frac{d^3y}{dx^3},\ldots,\frac{d^My}{dx^M}\right)=0 \tag{1}$$

se dice que es una ecuación diferencial ordinaria de orden M. Despejando la derivada de mayor orden se obtiene

$$\frac{d^My}{dx^M}=f\left(x,y,\frac{dy}{dx},\frac{d^2y}{dx^2},\ldots,\frac{d^{M-1}y}{dx^{M-1}}\right) \tag{2}$$

Haciendo el siguiente cambio de variables

$$y=y^1\qquad\cdots\qquad\frac{d^ky}{dx^k}=y^{k+1}\qquad\cdots\qquad\frac{d^{M-1}y}{dx^{M-1}}=f(x,y^1,y^2,\ldots,y^M) \tag{3}$$

a la final (1) se convierte en un sistema de ecuaciones diferenciales ordinarias de primer orden

$$\frac{d\mathbf{y}}{dx}=\mathbf{f}(x,\mathbf{y})\qquad\qquad\mathbf{y}(x):\mathbb{R}\longrightarrow\mathbb{R}^M \tag{4}$$

Si se especifican los valores de las distintas y^k para un único punto $x=x_0$, entonces se tiene un problema de valor inicial, donde se conoce $\mathbf{y}(x_0)=\mathbf{y}_0$. En cuanto a la función $\mathbf{f}(x,\mathbf{y}):\mathbb{R}\times\mathbb{R}^M\longrightarrow\mathbb{R}^M$ se puede decir que $f^k(x,\mathbf{y})=y^{k+1}$, $k=1,2,\ldots,M-1$, y $f^M(x,\mathbf{y})=f(x,y^1,y^2,\ldots,y^M)$ definido por el despeje (2).

3.1. FUNDAMENTOS

En esta parte, para hacer una presentación que se considera didáctica, se mostrarán las distintas técnicas numéricas aplicadas a resolver un único problema, las "Orbitas de Arenstorf", con dos métodos Runge-Kutta ambos implícitos de sexto orden, uno parcial implícito y el otro total implícito.

Uno de los métodos que se analizará es el método Runge-Kutta de sexto orden ($P=6$) y cuatro etapas ($N=4$), basado en la cuadratura de Lobatto [lobatto,1851-52]

0	0	0	0	0
$(5-\sqrt{5})/10$	$(5+\sqrt{5})/60$	$1/6$	$(15-7\sqrt{5})/60$	0
$(5+\sqrt{5})/10$	$(5-\sqrt{5})/60$	$(15+7\sqrt{5})/60$	$1/6$	0
1	$1/6$	$(5-\sqrt{5})/12$	$(5+\sqrt{5})/12$	0
	$1/12$	$5/12$	$5/12$	$1/12$

(5)

que realmente engloba dos métodos empotrado uno en el otro. El más pequeño adentro (separado con barras) es de tercer orden ($P = 3$). El método es implícito sólamente en la segunda y tercera etapa.

El otro método que se analizará será el método Runge-Kutta de sexto orden ($P = 6$) y tres etapas ($N = 3$), basado en la cuadratura de Gauss (Kuntzmann-Butcher) [Hairer et al.,1987]

$$\begin{array}{c|cccc} (5-\sqrt{15})/10 & 5/36 & (10-3\sqrt{15})/45 & (25-6\sqrt{15})/180 \\ 1/2 & (10+3\sqrt{15})/72 & 2/9 & (10-3\sqrt{15})/72 \\ (5+\sqrt{15})/10 & (25+6\sqrt{15})/180 & (10+3\sqrt{15})/45 & 5/36 \\ \hline & 5/18 & 4/9 & 5/18 \end{array} \tag{6}$$

totalmente implícito. Resultados numéricos impresionantes de la mecánica celeste con este método fueron reportados en la tesis de D. Sommer [Sommer,(1965)].

El único problema a resolver será el de las orbitas Arenstorf (1963), que ilustra un buen problema de la mecánica celestial, rígido por una parte y caótico por otro, completamente bien planteado, que es un caso particular del problema de tres cuerpos, con uno de ellos de masa despreciable. La orbita de este último es el de la órbita que se describe. Considérese dos cuerpos másicos de masas η y μ en traslación cuasi-circular en un plano y un tercer cuerpo de masa desprecible moviéndose alrededor en el mismo plano. Las ecuaciones diferenciales en variables relativas del caso Tierra y Luna son ($u = \{\mathbf{v}\}_x$, $v = \{\mathbf{v}\}_y$)

$$\begin{aligned} \frac{dx}{dt} &= u \\ \frac{du}{dt} &= x + 2v - \eta\left(\frac{x+\mu}{A}\right) - \mu\left(\frac{x-\eta}{B}\right) \\ \frac{dy}{dt} &= v \\ \frac{dv}{dt} &= y - 2u - \eta\left(\frac{y}{A}\right) - \mu\left(\frac{y}{B}\right) \end{aligned} \tag{7}$$

donde

$$\begin{aligned} A &= \sqrt{[(x+\mu)^2 + y^2]^3} \qquad \mu = 0.012277471 \\ B &= \sqrt{[(x-\eta)^2 + y^2]^3} \qquad \eta = 1 - \mu \end{aligned} \tag{8}$$

La figura 1 muestra el resultado de esta órbita con varios métodos [Hairer et al.,1987,pp.127-129].

(espacio vacío)

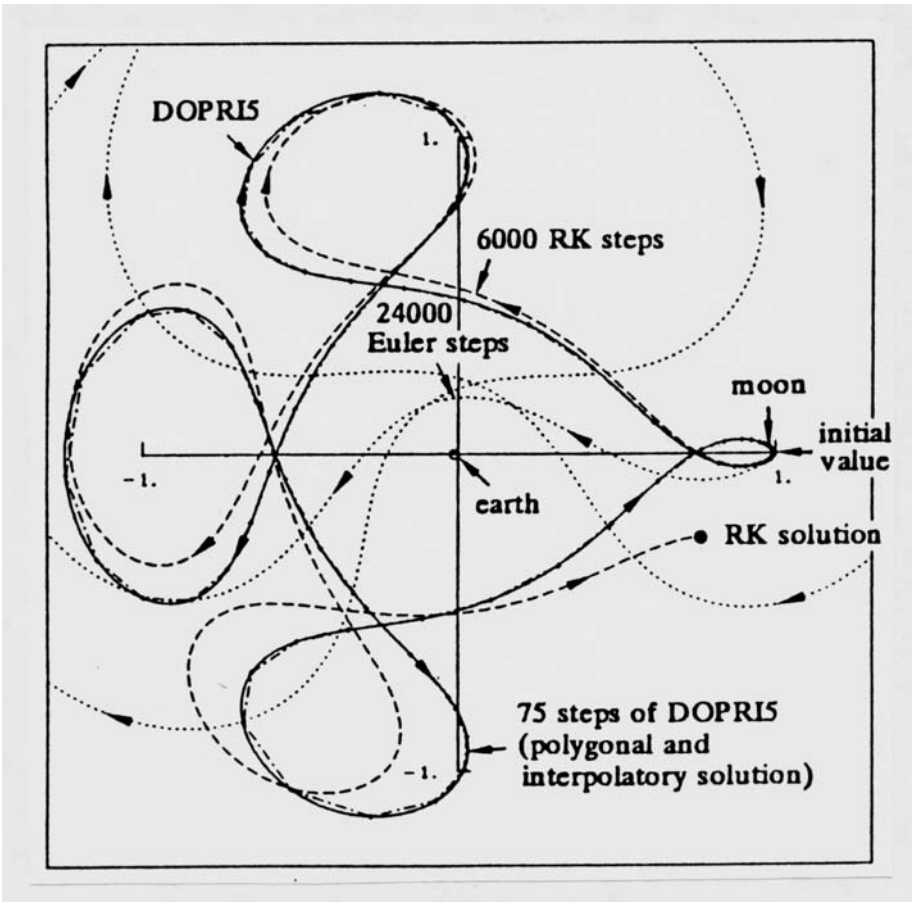


Figura 1. La órbita de Arenstorf computada con Euler equidistante, Runge-Kutta equidistante y paso variable con el método de Dormand y Prince (DOPRI5).

Las condiciones iniciales han sido cuidadosamente determinadas

$$\begin{aligned} x(0) &= 0.994 \qquad u(0) = 0 \qquad y(0) = 0 \\ v(0) &= -2.00158510637908252240537862224 \end{aligned} \tag{9}$$

para que la solución sea cíclica con período

$$T = 17.065216560157962558891720624 9 \tag{10}$$

Tales soluciones periódicas orbitales han facinado astrónomos y matemáticos por muchas décadas (Poincaré) y ahora frecuentemente llamadas "Arenstorf orbits" en honor a Arenstorf (1963), quien además hizo muchas simulaciones numéricas en computadoras electrónicas de alta velocidad. El problema es C^∞ con la excepción de dos puntos singulares $x = -\mu$ y $x = \eta$, para $y = 0$, por lo tanto la solución poligonal de Euler se sabe que converge a la solución exacta. Pero son numéricamente y realmente útiles aquí? Se ha escogido $n_s = 24000$ pasos de longitud $h = T/n_s$ para resolver el problema razonablemente regular. Con un método Runge-Kutta convencional de cuarto orden se han necesita 6000 pasos para resover el problema razonablemente bien.

Con la finalidad de simplificar el problema y los cálculos se han cambiado algunas variable y han

convertido el problema en este otro

$$\begin{aligned}
\frac{dx}{dt} &= u \\
\frac{du}{dt} &= x\,(1-\eta a^3-\mu b^3)+2v-\eta\mu\,(a^3-b^3) \\
\frac{dy}{dt} &= v \\
\frac{dv}{dt} &= y\,(1-\eta a^3-\mu b^3)-2u
\end{aligned} \tag{11}$$

donde

$$\begin{array}{lll}
a(x,y)=\sqrt[3]{1/A}=[(x+\mu)^2+y^2]^{-1/2} & \dfrac{\partial a^3}{\partial x}=-3\,(x+\mu)\,a^5 & \dfrac{\partial a^3}{\partial y}=-3\,y\,a^5 \\
b(x,y)=\sqrt[3]{1/B}=[(x-\eta)^2+y^2]^{-1/2} & \dfrac{\partial b^3}{\partial x}=-3\,(x-\eta)\,b^5 & \dfrac{\partial b^3}{\partial y}=-3\,y\,b^5
\end{array} \tag{12}$$

Si asignamos el siguiente orden a las variables

$$\mathbf{y}=\{x,u,y,v\} \qquad\qquad \frac{d\mathbf{y}}{dt}=\mathbf{f}(\mathbf{y}) \tag{13}$$

la matriz jacobiana del problema es

$$[\mathbf{J_f}(\mathbf{y})]=\begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{\partial f^2}{\partial x} & 0 & \frac{\partial f^2}{\partial y} & 2 \\ 0 & 0 & 0 & 1 \\ \frac{\partial f^4}{\partial x} & -2 & \frac{\partial f^4}{\partial y} & 0 \end{bmatrix} \tag{14.a}$$

donde algunos desarrollo ha sido colocado afuera por simplicidad

$$\frac{\partial f^2}{\partial x}=3\,[\,\eta\,(x+\mu)^2a^5+\mu\,(x-\eta)^2b^5\,]+1-\eta a^3-\mu b^3 \tag{14.b}$$

$$\frac{\partial f^4}{\partial x}=3\,y\,[\,\eta\,(x+\mu)\,a^5+\mu\,(x-\eta)\,b^5\,] \tag{14.c}$$

$$\frac{\partial f^2}{\partial y}=3\,y\,[\,\eta\,(x+\mu)\,a^5+\mu\,(x-\eta)\,b^5\,] \tag{14.d}$$

$$\frac{\partial f^4}{\partial y}=3\,y^2(\eta a^5+\mu b^5)+1-\eta a^3-\mu b^3 \tag{14.e}$$

El problema único planteado ya está en condiciones de ser resulto.

3.2. METODOS EXPLICITOS

Los método implícitos utilizados 3.3.(5) y 3.1.(6), requieren para su solución unos iterados iniciales en las variable $\mathbf{k}_s$ para cada paso de integración (n constante). Las técnicas que siguen permiten lograr este cometido.

3.2.1. Cuadratura de Kutta

En 1965 Ralston hizo un análisis similar a 1.2.(8), para obtener las relaciones de los coeficientes de un método Runge-Kutta explícito de cuarto orden y cuatro etapas, y encontró la siguiente familia de métodos en función de los coeficientes b_2 y b_3 (ver por ejemplo [Ralston & Rabinowitz,1978])

$$b_1 = 0 \qquad b_4 = 1 \qquad a_{rs} = 0 \quad (s \geq r) \tag{1.a-c}$$

$$a_{21} = b_2 \qquad a_{31} = b_3 - a_{32} \qquad a_{32} = \frac{b_3(b_3 - b_2)}{2\,b_2(1 - 2\,b_2)} \qquad a_{41} = 1 - a_{42} - a_{43} \tag{1.d-g}$$

$$a_{42} = \frac{(1 - b_2)[b_2 + b_3 - 1 - (2\,b_3 - 1)^2]}{2\,b_2(b_3 - b_2)[6\,b_2 b_3 - 4\,(b_2 + b_3) + 3]} \qquad a_{43} = \frac{(1 - 2\,b_2)(1 - b_2)(1 - b_3)}{b_3(b_3 - b_2)[6\,b_2 b_3 - 4\,(b_2 + b_3) + 3]} \tag{1.h,i}$$

$$c_1 = \frac{1}{2} + \frac{1 - 2(b_2 + b_3)}{12\,b_2 b_3} \qquad c_2 = \frac{2\,b_3 - 1}{12\,b_2(b_3 - b_2)(1 - b_2)} \tag{1.j,k}$$

$$c_3 = \frac{1 - 2\,b_2}{12\,b_3(b_3 - b_2)(1 - b_3)} \qquad c_4 = \frac{1}{2} + \frac{2\,(b_2 + b_3) - 3}{12\,(1 - b_2)(1 - b_3)} \tag{1.l,m}$$

Nótese que la substitución de los valores $b_2 = 1/2$ y $b_3 = 1/2$, ó los valores $b_2 = 1/3$ y $b_3 = 2/3$, permiten obtener los clásicos, bien conocidos y muy utilizados métodos de Kutta de cuarto orden y cuatro etapas del primer ó segundo tipo, y que se muestran a continuación

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 \\ 2/3 & 0 & -1/3 & 1 & 0 \\ 1 & 1 & -1 & 1 & 0 \\ \hline & 1/8 & 3/8 & 3/8 & 1/8 \end{array} \tag{2.a,b}$$

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & \frac{-1+\sqrt{2}}{2} & \frac{2-\sqrt{2}}{2} & 0 & 0 \\ 1 & 0 & -\frac{\sqrt{2}}{2} & \frac{2+\sqrt{2}}{2} & 0 \\ \hline & 1/6 & \frac{2-\sqrt{2}}{6} & \frac{2+\sqrt{2}}{6} & 1/6 \end{array} \tag{2.c}$$

El primer de los métodos arriba mostrados se basa en la cuadratura propuesta por Kutta originalmente. El segundo método en una variante del anterior, sólo que las evaluaciones intermedias son equidistantes (a veces se le denomina cuadratura de Kutta del segundo tipo). El tercer de los métodos es el método de Gill [1951], y no pertenece a la familia descrita por (1). No obstante, es una variante que mejora en cierta medida al método de Kutta del primer tipo [Carnahan et al.,1969], al cual se le parece mucho por la similitud de los coeficientes y por ser del mismo orden y tener el mismo número de etapas. Los coeficientes del método de Gill, por supuesto satisfacen las relaciones 1.2.(12) y 1.2.(12′). Todos los métodos de cuatro etapas (1) antes descritos son de cuarto orden en el error global y quinto orden en el error local.

Asi que, si $b_2 = (5 - \sqrt{5})/10$ y $b_3 = (5 + \sqrt{5})/10$ del método 3.1.5) son substituidos en la relaciones (1), se obtiene que

$$a'_{21} = \frac{(5 - \sqrt{5})}{10} \tag{3.a}$$

$$a'_{31} = -\frac{(5 + 3\,\sqrt{5})}{20} \tag{3.b}$$

$$a'_{32} = \frac{(3+\sqrt{5})}{4} \tag{3.c}$$

Estos son los coeficientes de un nuevo m'etodo Runge-Kutta explícito, que en la notación de Butcher puede ser expresado como

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ (5-\sqrt{5})/10 & (5-\sqrt{5})/10 & 0 & 0 & 0 \\ (5+\sqrt{5})/10 & -(5+3\sqrt{5})/20 & (3+\sqrt{5})/4 & 0 & 0 \\ 1 & (-1+5\sqrt{5})/4 & -(5+3\sqrt{5})/4 & (5-\sqrt{5})/2 & 0 \\ \hline & 1/12 & 5/12 & 5/12 & 1/12 \end{array} \tag{4}$$

El método Runge-Kutta explícito asi encontrado no está reportado en la literaturatura especializada y no corresponde a ninguna cuadratura en particular (aquí le hemos denominado cuadratura de Kutta), pero tiene los mismo puntos de colocación que el método de cuadratura de Lobatto, y pertenece a la familia de métodos Runge-Kutta explícito de cuarto orden de la solución (1). Las $\mathbf{k}'_2$ y $\mathbf{k}'_3$ que se obtienen con este método de coeficientes cambiados a'_{rs}

$$a'_{21} = \frac{(5-\sqrt{5})}{10} \tag{5.a}$$

$$a'_{31} = -\frac{(5+3\sqrt{5})}{20} \tag{5.b}$$

$$a'_{32} = \frac{(3+\sqrt{5})}{4} \tag{5.c}$$

son las estimaciones iniciales para el proceso iterativo

$$\mathbf{k}_{2,(0)} = \mathbf{f}(x_n + b_2 h\,,\, \mathbf{y}_n + a'_{21}\, h\, \mathbf{k}_1) \tag{6.a}$$

$$\mathbf{k}_{3,(0)} = \mathbf{f}(x_n + b_3 h\,,\, \mathbf{y}_n + a'_{31}\, h\, \mathbf{k}_1 + a'_{32}\, h\, \mathbf{k}_2) \tag{6.b}$$

Una vez substituida estas estimaciones, se espera una rápida convergencia del método implícito 3.1.(5) en cada paso (n constante). Sea que se implemente une esquema iterativo de punto fijo (sección 3.3.2) o de Newton-Raphson (sección 3.3.3) como se verá adelante.

3.2.2. Extrapolación de Lagrange

El proceso iterativo para resolver el método Runge-Kutta implícito 3.1.(6) comienza con un iterado inicial para las variable $\mathbf{k}_{r,(0)}$, $r = 1, 2, 3$, en cada paso. Para el caso en que estamos interesados, usaremos el método Runge-Kutta de quinto orden ($P = 5$), generado por la extrapolación de Lagrange, o el método Runge-Kutta de tercer orden ($P = 3$) [Ralston & Rabinowitz,1978], ambos explícitos

$$\begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ (5-\sqrt{15})/10 & (5-\sqrt{15})/10 & 0 & 0 & 0 & 0 \\ 1/2 & -(3+\sqrt{15})/4 & (5+\sqrt{15})/4 & 0 & 0 & 0 \\ (5+\sqrt{15})/10 & 3(4+\sqrt{15})/5 & -(35+9\sqrt{15})/10 & 2(4+\sqrt{15})/5 & 0 & 0 \\ 1 & -1 & 5/3 & -20/15 & 25/15 & 0 \\ \hline & 0 & 5/18 & 4/9 & 5/18 & 0 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}$$

(7.a, b)

El primero de estos Runge-Kutta, ec. (7.a), fué generado por la extrpolación de los polinomios de Lagrange

$$P_n(x) = \sum_{i=0}^{n} L_i(x)\ f(x_i) \qquad\qquad L_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)} \tag{8}$$

que pasa por las etapas previas $s = 1, 2, \ldots, r-1$, con puntos de colocación b_s como variables independientes y $\mathbf{k}_s$ como variables dependientes conocidas ($s = 1, 2, \ldots, r-1$, $r = 2, \ldots, N$). Entonces, mediante extrapolación de el polinomio $P_{r-2}(x)$ ($r \geq 2$), para la siguiente etapa b_r ($b_1 = 0$), los coeficientes a_{rs} son calculados como ($a_{1s} = 0$, $a_{21} = b_2$)

$$a_{rs} = b_r \alpha_s / \alpha \qquad\qquad \alpha_s = L_s(b_r) = \prod_{\substack{j=1 \\ j\neq s}}^{r-1} \frac{(b_r - b_j)}{(b_s - b_j)} \qquad\qquad \alpha = \sum_{s=1}^{r-1} \alpha_s \tag{9}$$

El valor de α es para normalizar α_s y satisface $a_{rs}\delta_s = \sum_{s=1}^{r-1} a_{rs} = b_r$ (entre líneas puntiadas en la ec.(7.a) están los valores en los que estamos interesados, pero la primera etapa es necesaria sólo para completar el esquema, las dos últimas filas son innecesarias). Para este caso en (7.a) $\alpha = 1$ siempre. Cuando el límite superior es menor que el límite inferior, el símbolo Π es 1. Los coeficientes c_r ($r = 1, 2, \ldots, N$) son calculados usando de nuevo (8), el polinomio de Lagrange $P_{N-1}(x)$, por integración de

$$c_r = \int_0^1 L_r(x)\ dx \qquad\qquad L_r(x) = \prod_{\substack{j=1 \\ j\neq r}}^{N} \frac{(x - b_j)}{(b_r - b_j)} \tag{10}$$

y debe satisfacer $\sum_{r=1}^{N} c_r = 1$ (para el caso de la ec.(7.a), $N = P = 5$). De hecho, no sólo (7.a), sino también (7.b), satisfacen (29) − (30). El método explícito (4) no pertenece al porceso de generación de coeficientes (9), sólo (10)

La segunda, ec. (7.b) equivalente a la cuadratura de Simpson 1/3, más simple, is un Runge-Kutta explícito de tercer orden que tiene los puntos de colocación cercanos a los requeridos. El valor $(5-\sqrt{15})/10 \approx 0.1127$ es cercano a 0 ($r = 1$ en Gauss), el valor $(5 + \sqrt{15})/10 \approx 0.8873$ es crecano a 1 ($r = 3$ en Gauss), y el valor 1/2 es exacto ($r = 2$ Gauss). Estas estimaciones de las variables $\mathbf{k}_r$ ($r = 1, 2, 3$) para los iterados iniciales son considerados suficientes para garantizar la convergencia en cada pasofor initial iterates are considered enough to guarantee convergence in each step. Cualquiera de los dos Runge-Kutta explícitos (7), seleccionado para los iterados iniciales de $\mathbf{k}_r$ es opcional.

3.3. METODOS IMPLICITOS

Ambos métodos utilizados para resolver el problema planteado son implícitos.

3.3.1. Cuadratura de Gauss

Los métodos de Runge-Kutta basados en la cuadratura de Gauss-Legendre son completamente implícitos (las matrices están totalmente llenas). En estos casos, los coeficientes satisfacen las siguientes relaciones

$$a_{rs} b_s^{\gamma-1} = \frac{b_r^{\gamma}}{\gamma} \qquad\qquad c_s b_s^{\gamma-1} = \frac{1}{\gamma} \qquad\qquad \gamma = 1, 2, 3, \ldots, N \tag{1}$$

donde los coeficientes b_r son las raíces del polinomio de Legendre de orden N, el número de etapas, es decir,

$$\mathcal{P}_N(2\, b_r - 1) = 0 \tag{2}$$

donde el orden del método Runge-Kutta que se origina es el doble del número de etapa ($P = 2N$). En la notación de Butcher, los tres primeros de estos métodos son

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} (3-\sqrt{3})/6 & 1/4 & (3-2\sqrt{3})/12 \\ (3+\sqrt{3})/6 & (3+2\sqrt{3})/12 & 1/4 \\ \hline & 1/2 & 1/2 \end{array} \tag{3.a,b}$$

$$\begin{array}{c|ccc} (5-\sqrt{15})/10 & 5/36 & (10-3\sqrt{15})/45 & (25-6\sqrt{15})/180 \\ 1/2 & (10+3\sqrt{15})/72 & 2/9 & (10-3\sqrt{15})/72 \\ (5+\sqrt{15})/10 & (25+6\sqrt{15})/180 & (10+3\sqrt{15})/45 & 5/36 \\ \hline & 5/18 & 4/9 & 5/18 \end{array} \tag{3.c}$$

de segundo ($P = 2$) orden, cuarto ($P = 4$) orden (Hammer-Hollingsworth) y sexto ($P = 6$) orden (Kuntzmann-Butcher), respectivamente.

3.3.2. Cuadratura de Lobatto

Los métodos Runge-Kutta explícitos son de aplicación directa, mientras que los métodos Runge-Kutta implícitos requieren la resolución de un sistema de ecuaciones con las variables auxiliares $\mathbf{k}_r$ en cada paso de integración de las ecuaciones diferenciales, como está sugerido por las ecuaciones 1.2.(6.b'). Este sistema de ecuaciones es generalmente no lineal, al menos que la función $\mathbf{f}(x, \mathbf{y})$ sea lineal, y puede ser resuelto aplicando un esquema iterativo del tipo punto fijo.

El método Runge-Kutta implícito que va a ser usado aquí, es un método de sexto orden ($P = 6$) con cuatro etapas ($N = 4$), desarrollado sobre las bases de la cuadratura de Lobatto [1851] (para más detalles ver [Butcher,1987] y [Lapidus & Seinfeld,1971]). Los coeficientes de este método organizados en la notación de Butcher son

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ (5-\sqrt{5})/10 & (5+\sqrt{5})/60 & 1/6 & (15-7\sqrt{5})/60 & 0 \\ (5+\sqrt{5})/10 & (5-\sqrt{5})/60 & (15+7\sqrt{5})/60 & 1/6 & 0 \\ \hline 1 & 1/6 & (5-\sqrt{5})/12 & (5+\sqrt{5})/12 & 0 \\ \hline & 1/12 & 5/12 & 5/12 & 1/12 \end{array} \tag{4.a}$$

Este método será denominado como el "principal".

Dentro de los coeficientes del método principal, pueden ser detectados una parte de ellos que forman otro método Runge-Kutta "secundario" empotrado en el primero. Este otro método es de tercer orden ($P = 3$), tiene tres etapas ($N = 3$) y en la notación de Butcher son

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ (5-\sqrt{5})/10 & (5+\sqrt{5})/60 & 1/6 & (15-7\sqrt{5})/60 \\ (5+\sqrt{5})/10 & (5-\sqrt{5})/60 & (15+7\sqrt{5})/60 & 1/6 \\ \hline & 1/6 & (5-\sqrt{5})/12 & (5+\sqrt{5})/12 \end{array} \tag{4.b}$$

Ambos métodos, el principal y el secundario, constituyen lo que se denomina la forma de la cuadratura de Lobatto empotrada de tercer y sexto órdenes con cuatro etapas (el método de Fehlberg [1971] posee una forma similar, pero es explícito). Nótese que esta forma Lobatto sólo es implícita en en las variables $\mathbf{k}_2$ and $\mathbf{k}_3$, y por lo tanto debe ser resuelto el sistema sólo en esas dos variables, lo que trae como consecuencia un incremento de la eficiencia de resolución, comparado con otros métodos implícitos. Las otras variables son de resolución directa (en la última etapa, una vez encontradas las anteriores.

Con la finalidad de aplicar un proceso iterativo para resolver el sistema de ecuaciones no lineales se requieren estimaciones iniciales de las variables auxiliares implícitas. La mejor forma de hacer esto es obtenerlas de un método Runge-Kutta explícito, donde las variables auxiliares $\mathbf{k}_r$ estén evaluadas en los mismos puntos intermedios en cada paso, o sea, que el método explícito tenga los mismos valores en los coeficientes b_r que el método implícito, o lo que es lo mismo que tenga los mismos puntos de colocación. Observando el método (15.a), está claro que el mencionado método explícito se puede obtener rápidamente de las relaciones (13) sugeridas por Gear, asumiendo los valores $b_1 = 0$, $b_2 = (5 - \sqrt{5})/10$, $b_3 = (5 + \sqrt{5})/10$ y $b_4 = 1$. Nótese que la selección de valores es consistente con las característica de un método explícito. Este último aspecto, casualmente hace que el método implícito (4.a) sea ideal para los propósitos deseados.

Así que, si los valores seleccionados para b_2 y b_3 son substituidos en las relaciones 3.2.(1), se obtienen los siguientes coeficientes

$$a'_{21} = \frac{(5 - \sqrt{5})}{10} \tag{5.a}$$

$$a'_{31} = -\frac{(5 + 3\sqrt{5})}{20} \tag{5.b}$$

$$a'_{32} = \frac{(3 + \sqrt{5})}{4} \tag{5.c}$$

Estos son los coeficientes de un nuevo método Runge-Kutta explícito, que en la notación de Butcher pueden globalmente ser expresados como

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ (5-\sqrt{5})/10 & (5-\sqrt{5})/10 & 0 & 0 & 0 \\ (5+\sqrt{5})/10 & -(5+3\sqrt{5})/20 & (3+\sqrt{5})/4 & 0 & 0 \\ 1 & 1/6 & (5-\sqrt{5})/12 & (5+\sqrt{5})/12 & 0 \\ \hline & 1/12 & 5/12 & 5/12 & 1/12 \end{array} \tag{6}$$

Este método, perteneciente a la familia de soluciones (13), será usado para obtener los estimados iniciales de $\mathbf{k}'_2$ y $\mathbf{k}'_3$ para el proceso iterativo de la siguiente forma

$$\mathbf{k}_{2,(0)} = \mathbf{f}(x_n + b_2 h\,,\, \mathbf{y}_n + a'_{21}\, h\, \mathbf{k}_1) \tag{7.a}$$

$$\mathbf{k}_{3,(0)} = \mathbf{f}(x_n + b_3 h\,,\, \mathbf{y}_n + a'_{31}\, h\, \mathbf{k}_1 + a'_{32}\, h\, \mathbf{k}_2) \tag{7.b}$$

con los coeficientes de (6). Una vez que estas estimaciones iniciales son usadas, se espera que exista una convergencia segura y rápida hacia la solución del sistema no lineal 1.2.(6.b'), con los coeficientes b_r de (4.a).

• Proceso Iterativo

Como se mencionó antes, el sistema de ecuaciones no lineales (6.b), que es originado por cualquier método Runge-Kutta implícito, puede ser resuelto en las variables auxiliares $\mathbf{k}_r$, aplicando un procedimiento iterativo de punto fijo (ver [Gear,1971])

$$k^i_{r,(m+1)} = f^i(x_n + b_r h\,,\, \mathbf{y}_n + a_{rs}\, h\, \mathbf{k}_{s,(m)}) \tag{8}$$

el cual es el más sencillo de usar debido a la forma de variable despejada que tiene el mencionado sistema de ecuaciones. Aquí el índice $m = 0, 1, 2, 3, \ldots$ es el número de la iteración en el proceso iterativo para cada paso de integración.

El error global durante el proceso iterativo se define como

$$\varepsilon^i_{r,(m)} = k^i_{r,(m)} - k^i_r \tag{9.a}$$

donde k_r^i es la solución exacta del sistema de ecuaciones no lineal.

El error local en cada iteración se define como

$$\epsilon^i_{r,(m)} = k^i_{r,(m+1)} - k^i_{r,(m)} \tag{9.b}$$

El procedimiento iterativo se detiene cuando se satisfaga

$$c_r \|\epsilon_{r,(m)}\| < \epsilon_{max} \tag{9.c}$$

donde ϵ_{max} es la tolerancia impuesta al error local para encontrar la solución y^i_{n+1} de las ecuaciones diferenciales en un paso de integración, y donde la norma del error local $\epsilon_{r,(m)}$ se supone euclidiana.

Si ahora la expresión 1.2.(6.b') se sustrae de la expresión (8), queda

$$k^i_{r,(m+1)} - k^i_r = f^i(x_n + b_r h\,,\, \mathbf{y}_n + a_{rs}\, h\, \mathbf{k}_{s,(m)}) - f^i(x_n + b_r h\,,\, \mathbf{y}_n + a_{rs}\, h\, \mathbf{k}_s)] \tag{10}$$

Luego, si se aplica la condición de Lipschitz (con $h > 0$ por conveniencia), resulta

$$|k^i_{r,(m+1)} - k^i_r| \le h\, l^i_j\, |a_{rs}|\; |k^j_{s,(m)} - k^j_s| \tag{11}$$

$$|\varepsilon^i_{r,(m+1)}| \le h\, l^i_j\, |a_{rs}|\; |\varepsilon^j_{s,(m)}| \tag{12}$$

donde l^i_j es el máximo del valor absoluto de cada elemento de la matriz jacobiana de $\mathbf{f}$. Esto es

$$|f^i_j| \le l^i_j \tag{13}$$

De manera que, si se satisface que

$$\varepsilon_{(m)} = \max_{1\le j\le M}\big(\max_{1\le s\le N} |\varepsilon^j_{s,(m)}|\big) \tag{14}$$

entonces

$$|\varepsilon^i_{r,(m+1)}| \le h\, l^i_j\, |a_{rs}|\; |\varepsilon^j_{s,(m)}| \le h\, l^i_j \delta_j\, |a_{rs}|\, \delta_s \varepsilon_{(m)} \tag{15}$$

$$\max_{1\le i\le M} \max_{1\le r\le N}\big(|\varepsilon^i_{r,(m+1)}|\big) \le \max_{1\le i\le M}\big[\max_{1\le r\le N}\big(h\, l^i_j \delta_j\, |a_{rs}|\, \delta_s \varepsilon_{(m)}\big)\big] \tag{16}$$

$$\varepsilon_{(m+1)} \le h\, \mathrm{L}\, \mathrm{A}\, \varepsilon_{(m)} \tag{17}$$

donde

$$\mathrm{L} = \max_{1\le i\le M}\big(l^i_j \delta_j\big) \qquad\qquad \mathrm{A} = \max_{1\le r\le N}\big(|a_{rs}|\, \delta_s\big) \tag{18}$$

La expresión (17) significa que, para un alto número de iteraciones, $m \to \infty$, el error global $\varepsilon_{(m)} \to 0$ cuando

$$h \le \frac{1}{\mathrm{L}\, \mathrm{A}} \tag{19}$$

y el proceso iterativo es convergente localmente (también globalmente) en la forma

$$c_r \|\epsilon_{r,(m+1)}\| < c_r \|\epsilon_{r,(m)}\| \tag{20}$$

(se suma en r). La expresión (19) es el límite del tamaño del paso para que el procedimiento iterativo de punto fijo descrito antes sea convergente. Ésta es la única restricción adicional de los métodos implícitos, frente a los explícitos. No obstante, los métodos implícitos son más estables que los explícitos, como se verá más adelante.

En la sección 3.7 se encontrará una formulación general del proceso iterativo, cuando se usa el método de Newton-Raphson, más costoso en cuanto al cómputo y con una convergencia más rápida, en lugar del sencillo y de lenta convergencia método de punto fijo (7).

Algunas veces el sistema de ecuaciones diferenciales no aparece en la forma de (1), sino de una forma implícita del tipo

$$\frac{dy^i}{dx} = f^i\Big(x, \mathbf{y}, \frac{d\mathbf{y}}{dx}\Big) \qquad i = 1, 2, 3, \ldots, M \tag{21}$$

En estos casos, el procedimiento iterativo se aplica en la forma descrita antes, pero se requiere estimaciones iniciales de $\mathbf{k}_r$ también para el método Runge-Kutta explícito. Estas estimaciones debe ser aceptables, o de otra forma el número de iteraciones puede volverse muy grande. Una forma de obtener tales estimaciones es mediante extrapolaciones con polinomios de Lagrange. esto es, para estimar $\mathbf{k}_2$, se hace una extrapolación con un polinomio de grado 0 comenzando en $\mathbf{k}_1$; para estimar $\mathbf{k}_3$, se hace una extrapolación con un polinomio de grado 1 definido por $\mathbf{k}_1$ and $\mathbf{k}_2$; Para estimar $\mathbf{k}_4$, se hace una extrapolación con un polinomio de grado 2 definido por $\mathbf{k}_1$, $\mathbf{k}_2$ y $\mathbf{k}_3$. El procedimiento descrito arroja los siguientes resultados

$$\mathbf{k}_2' = \mathbf{k}_1 \tag{22.a}$$

$$\mathbf{k}_3' = \frac{b_2 - b_3}{b_2}\mathbf{k}_1 + \frac{b_3}{b_2}\mathbf{k}_2 \tag{22.b}$$

$$\mathbf{k}_4' = \frac{(b_4 - b_2)(b_4 - b_3)}{b_2 b_3}\mathbf{k}_1 + \frac{b_4(b_4 - b_3)}{b_2(b_2 - b_3)}\mathbf{k}_2 + \frac{b_4(b_4 - b_2)}{b_3(b_3 - b_2)}\mathbf{k}_3 \tag{22.c}$$

Nótese que los coeficientes b_r han sido usado como los puntos de colocación de los correspondientes polinomios. Para el método Runge-Kutta Lobatto las expresiones (22) se particularizan como

$$\mathbf{k}_2' = \mathbf{k}_1 \tag{23.a}$$

$$\mathbf{k}_3' = -\frac{1 + \sqrt{5}}{2}\mathbf{k}_1 + \frac{3 + \sqrt{5}}{2}\mathbf{k}_2 \tag{23.b}$$

$$\mathbf{k}_4' = \mathbf{k}_1 - \sqrt{5}(\mathbf{k}_2 - \mathbf{k}_3) \tag{23.c}$$

En cualquier caso, la estimación de $\mathbf{k}_1$ se hace con la variable $\mathbf{k}_4$ del paso inmediatamente anterior.

3.3.3. Resuelto con Newton-Raphson

Esta sección explica como el método Newton-Raphson puede ser aplicado para resolver el sistema de acuaciones no lineales con las variables auxiliares k_r^i en los métodos Runge-Kutta implícitos en general.

Sea un sistema de ecuaciones diferenciales ordinarias expresado como

$$\frac{dy^i}{dx} = f^i(\mathbf{y}) \qquad\qquad \frac{d\mathbf{y}}{dx} = \mathbf{f}(\mathbf{y}) \tag{24.a}$$

con las condiciones iniciales

$$x = x_o \qquad\qquad y^i(x_o) = y_o^i \qquad\qquad \mathbf{y}(x_o) = \mathbf{y}_o \tag{24.b}$$

Para resolver el problema de valor inicial $(24.a, b)$ en un sistema autónomo, el método de Runge-Kutta implícito

$$\begin{aligned} y_{n+1}^i &= y_n^i + h\, c_r k_r^i + O(h^{P+1}) & \mathbf{y}_{n+1} &= \mathbf{y}_n + h\, c_r \mathbf{k}_r + O(h^{P+1}) \\ k_r^i &= f^i(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_s) & \mathbf{k}_r &= \mathbf{f}(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_s) \end{aligned} \tag{24.c}$$

puede ser utilizado con éxito acompañado del método de Newton-Raphson (del lado izquierdo se han colocado las expresiones en notación indicial, mientras que en el lado derecho se han escrito usando notación simbólica).

La expresión (24.c) (segunda línea) debe ser interpretada como un sistema de ecuaciones no lineales con las variables auxiliares k_r^i como incógnitas en cada paso (n constante). Por esta razón, es conveniente definir la función

$$g_r^i(\mathbf{k}) = f^i(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_s) - k_r^i = 0 \qquad\qquad \mathbf{g}_r(\mathbf{k}) = \mathbf{f}(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_s) - \mathbf{k}_r = \mathbf{0} \tag{25}$$

que debe ser cero en cada componente cuando la solución para cada k_r^i ha sido encontrada en cada paso.

Con la finalidad de resolver el sistema de ecuaciones no lineales (25), es más eficiente usar el método de Newton-Raphson que el método de punto fijo (como es sugerido por (10.b) y (25))

$$k_{r,(m+1)}^i = f^i(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_{s,(m)}) \qquad\qquad \mathbf{k}_{r,(m+1)} = \mathbf{f}(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_{s,(m)}) \tag{26}$$

el cual es más fácil de usar, pero tiene peor convergencia.

Sin embargo, para usar el método de Newton-Raphson method, la matriz jacobiana de la función $g_r^i(\mathbf{k})$, con respecto a las variables k_t^j tiene que ser calculada, y tiene que ser definida como

$$\begin{aligned}\frac{\partial g_r^i}{\partial k_t^j} &= h\,\frac{\partial f^i}{\partial y^k}\bigg|_{\mathbf{y}_n + h a_{rs}\mathbf{k}_s} a_{rs}\,\delta^{kj}\delta_{st} - \delta^{ij}\delta_{rt} \\ &= h\, J_{\mathbf{f}}^{ij}(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_s)\, a_{rt} - \delta^{ij}\delta_{rt}\end{aligned} \qquad\qquad \mathbf{J_g}(\mathbf{k}) = h\,\mathbf{J_f}(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_s) \otimes \mathbf{A} - \mathbf{I_f} \otimes \mathbf{I_A} \tag{27}$$

donde ha sido usada la regla de la cadena y, del lado derecho de (25) y (27), $\mathbf{k}$ contiene todas las $\mathbf{k}_r$, $r = 1, 2, \ldots, N$. Los superíndices significan las componentes del sistema de ecuaciones diferenciales y los subídices significan las correspondientes etapas. La notación $J_{\mathbf{f}}^{ij}$ se usa en lugar de $\partial f^i/\partial y^j$, para los elementos de la matriz jacobiana $\mathbf{J}_f(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_s)$, y r no suma aunque aparezca repetida dos veces. La matrices identidad $\mathbf{I_f}$ y $\mathbf{I_A}$ tienen las dimensiones de $\mathbf{f}$ y $\mathbf{A}$, respectivamente (rangos de los índices de las delta de Kronecker δ_{ij} y δ_{rt}). De ahora en adelante, reservaremos el uso de la letra $\mathbf{k}$ minúscula negrilla sin índice (excepto el índice m para las iteraciones internas) para aquellas variables donde se han agrupado en un solo arreglo todas las etapas.

Así, el método de Newton-Raphson puede ser aplicado de la siguiente manera algorítmica

$$k_{r,(m+1)}^i = k_{r,(m)}^i + \omega\,\Delta k_{r,(m)}^i \qquad\qquad \mathbf{k}_{(m+1)} = \mathbf{k}_{(m)} + \omega\,\Delta\mathbf{k}_{(m)} \tag{28.a}$$

donde las variables del error $\Delta k_{t,(m)}^i$ se encuentran de la resolución del siguiente sistema de ecuaciones lineales

$$\Big[\frac{\partial g_r^i}{\partial k_t^j}\Big]_{(m)} \Delta k_{t,(m)}^j = -g_r^i(\mathbf{k}_{(m)}) \qquad\qquad [\mathbf{J_g}(\mathbf{k}_{(m)})].\Delta\mathbf{k}_{(m)} = -\mathbf{g}(\mathbf{k}_{(m)}) \tag{28.b}$$

donde ω es el factor de relajación y (m) indica el número de la iteración interna (se mantiene $\mathbf{y}_n$ y h constantes). Los ídices j y t son los que se contraen en la operación " $\cdot$ " de la ecuación simbólica. El proceso iterativo descrito por (28) se aplica de forma sucesiva hasta satisfacer

$$h\, c_r\, \|\boldsymbol{\epsilon}_{r,(m)}\| = \omega\, h\, c_r\, \|\Delta\mathbf{k}_{r,(m)}\| < \epsilon_{max} \qquad \text{donde} \qquad \epsilon_{r,(m)}^i = k_{r,(m+1)}^i - k_{r,(m)}^i \tag{29}$$

El valor $\epsilon_{r,(m)}^i$ es el error local en la variable auxiliar k_r^i, mientras que $h\, c_r\, \boldsymbol{\epsilon}_{r,(m)}$ es el error local para $\mathbf{y}_{n+1}$, y ϵ_{max} es la tolerancia para el error de truncamiento local en la solución numérica de y_{n+1}^i. El índice r suma en (29.a)

Para funciones muy complicadas, es conveniente expresar las derivadas parciales en la matriz (tensor) jacobiana (27) usando diferencias finitas atrasada, como está indicado en la siguiente expresión

$$\Big[\frac{\partial g_r^i}{\partial k_t^j}\Big] \approx h\,\Big[\frac{f^i(\mathbf{y}_n + \Delta\mathbf{y}_n) - f^i(\mathbf{y}_n^{(j)})}{\Delta y_n^j}\Big]\, a_{rt} - \delta_j^i\,\delta_r^t \tag{30}$$

donde la perturbación para derivar es

$$\Delta y_n^j = h\, a_{rs} k_s^j \qquad \text{and} \qquad f^i(\mathbf{y}_n^{(j)}) = f^i(y_n^1 + \Delta y_n^1, y_n^2 + \Delta y_n^2, \ldots, y_n^j, \ldots, y_n^M + \Delta y_n^M) \tag{31}$$

y la evaluación de la derivada se hace mediante diferencias atrasadas siendo $f^i(\mathbf{y}_n^{(j)})$ la función perturbada (hacia atrás) únicamente en la componenten j.

Los valores iniciales $k_{r,(0)}^j$ para el proceso iterativo se pueden estimar con un método Runge-Kutta explícito, cuyos puntos de colocación b_r sean los mismos que los del método implícito.

El problema planteado en (25) puede ser re-escrito de la siguiente forma

$$\mathbf{g}(\mathbf{k}) = \mathbf{F}(\mathbf{k}) - \mathbf{k} = \mathbf{0} \qquad \mathbf{F}(\mathbf{k}) = \left\{ \begin{array}{c} \mathbf{f}(\mathbf{y}_n + h\, a_{1s}\, \mathbf{k}_s) \\ \vdots \\ \mathbf{f}(\mathbf{y}_n + h\, a_{rs}\, \mathbf{k}_s) \\ \vdots \\ \mathbf{f}(\mathbf{y}_n + h\, a_{Ns}\, \mathbf{k}_s) \end{array} \right\} \qquad \mathbf{g}(\mathbf{k}) = \left\{ \begin{array}{c} \mathbf{f}(\mathbf{y}_n + h\, a_{1s}\, \mathbf{k}_s) - \mathbf{k}_1 \\ \vdots \\ \mathbf{f}(\mathbf{y}_n + h\, a_{rs}\, \mathbf{k}_s) - \mathbf{k}_r \\ \vdots \\ \mathbf{f}(\mathbf{y}_n + h\, a_{Ns}\, \mathbf{k}_s) - \mathbf{k}_N \end{array} \right\} \tag{32}$$

La función $\mathbf{F}$ y la variable $\mathbf{k}$ tienen dimensiones $M \times N$ (caso autónomo). El algoritmo del método Newton-Raphson (28) se puede exprear como

$$\mathbf{k}_{m+1} = \mathbf{k}_m - \omega\, [\mathbf{J}_\mathbf{g}(\mathbf{k}_m)]^{-1}.\, \mathbf{g}(\mathbf{k}_m) = \mathbf{k}_m - \omega\, [\mathbf{J}_\mathbf{F}(\mathbf{k}_m) - \mathbb{I}]^{-1}.\, (\mathbf{F}(\mathbf{k}_m) - \mathbf{k}_m) \qquad [\mathbf{J}_\mathbf{g}(\mathbf{k})] = [\mathbf{J}_\mathbf{F}(\mathbf{k})] - [\,\mathbb{I}\,] \tag{33}$$

con $\mathbb{I} = \mathbf{I}_\mathbf{f} \otimes \mathbf{I}_\mathbf{A}$ o lo que es en resumen lo mismo

$$\mathbf{k}_{m+1} = \mathbf{k}_m + \omega\, \Delta\mathbf{k}_m \qquad [\mathbf{J}_\mathbf{g}(\mathbf{k}_m)].\Delta\mathbf{k}_m = -\mathbf{g}(\mathbf{k}_m) \qquad [\mathbf{J}_\mathbf{F}(\mathbf{k}_m) - \mathbb{I}].\Delta\mathbf{k}_m = -(\mathbf{F}(\mathbf{k}_m) - \mathbf{k}_m) \tag{33'}$$

donde $[\mathbf{J}_\mathbf{F}(\mathbf{k})]$ es el jacobiano de la función $\mathbf{F}(\mathbf{k})$, y el jacobiano $[\mathbf{J}_\mathbf{g}(\mathbf{k})]$ de la función $\mathbf{g}(\mathbf{k})$ en (33.b) puede ser calculada de (27).

El método de punto fijo (26) tiene convergencia lineal en la proximidad de la solución cuando

$$\mathbf{k}_{m+1} = \mathbf{F}(\mathbf{k}_m) \qquad [\mathbf{J}_\mathbf{F}(\boldsymbol{\zeta})]_{\boldsymbol{\zeta} \in \mathbb{B}_\rho(\mathbf{k}_*)} < 1 \qquad [\mathbf{J}_\mathbf{F}(\mathbf{k})]_{rt} = h\, a_{rt}\, \mathbf{J}_\mathbf{f}(\mathbf{y}_n + h\, a_{rs}\, \mathbf{k}_s) \tag{34.a,b,c}$$

$$[\mathbf{J}_\mathbf{F}(\mathbf{k})] = \left[\begin{array}{ccccc} h\, a_{11}\, [\, \mathbf{J}_\mathbf{f}(\mathbf{y}_n + h\, a_{1s}\mathbf{k}_s)\,] & \cdots & h\, a_{1t}\, [\, \mathbf{J}_\mathbf{f}(\mathbf{y}_n + h\, a_{1s}\mathbf{k}_s)\,] & \cdots & h\, a_{1N}\, [\, \mathbf{J}_\mathbf{f}(\mathbf{y}_n + h\, a_{1s}\mathbf{k}_s)\,] \\ \vdots & & \vdots & & \vdots \\ h\, a_{r1}\, [\, \mathbf{J}_\mathbf{f}(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_s)\,] & \cdots & h\, a_{rt}\, [\, \mathbf{J}_\mathbf{f}(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_s)\,] & \cdots & h\, a_{rN}\, [\, \mathbf{J}_\mathbf{f}(\mathbf{y}_n + h\, a_{rs}\mathbf{k}_s)\,] \\ \vdots & & \vdots & & \vdots \\ h\, a_{N1}\, [\, \mathbf{J}_\mathbf{f}(\mathbf{y}_n + h\, a_{Ns}\mathbf{k}_s)\,] & \cdots & h\, a_{Nt}\, [\, \mathbf{J}_\mathbf{f}(\mathbf{y}_n + h\, a_{Ns}\mathbf{k}_s)\,] & \cdots & h\, a_{NN}\, [\, \mathbf{J}_\mathbf{f}(\mathbf{y}_n + h\, a_{Ns}\mathbf{k}_s)\,] \end{array} \right] \tag{34.c'}$$

con el bloque de la fila y la columna r, t indicado (no suma en r y si suma en s), $r, t, s = 1, \ldots, N$. De forma similar, es bien conocido que, en la proximidad de la solución, el método de Newton-Raphson (33) tiene una convergencia cuadrática cuando

$$\|\mathbf{J}_\mathbf{h}(\boldsymbol{\zeta})\|_{\boldsymbol{\zeta} \in \mathbb{B}_\rho(\mathbf{k}_*)} < 1 \qquad \text{with} \qquad \begin{array}{c} \mathbf{h}(\mathbf{k}) = \mathbf{k} - \omega\, [\mathbf{J}_\mathbf{g}(\mathbf{k})]^{-1}.\, \mathbf{g}(\mathbf{k}) \\ [\mathbf{J}_\mathbf{h}(\mathbf{k})] = [\,\mathbf{I}\,] - \omega\, [\mathbf{J}_\mathbf{g}]^{-1}.\, [\mathbf{J}_\mathbf{g}]^{\mathrm{t}} + \omega\, [\mathbf{J}_\mathbf{g}]^{-1}.\, \big\{\, [\mathbb{H}_\mathbf{g}]\,.\,[\mathbf{J}_\mathbf{g}]^{-1}.\, \mathbf{g}\, \big\}^{\mathrm{t}} \end{array} \tag{35}$$

donde los jacobianos son $[\mathbf{J}_\mathbf{g}(\mathbf{k})] = [\mathbf{J}_\mathbf{F}(\mathbf{k})] - [\,\mathbb{I}\,]$, los hessianos son iguales, $[\mathbb{H}_\mathbf{g}(\mathbf{k})] = [\mathbb{H}_\mathbf{F}(\mathbf{k})]$, y $\mathbb{B}_\rho(\mathbf{K}_*)$ es la bola cerrada de radio $\rho = \|\mathbf{k} - \mathbf{k}_*\| < \rho_*$ con centro en $\mathbf{k}_*$, la solución de (32.a). La norma usada es la norma infinita $\|\cdot\|_\infty$ [Burden & Faires,1985].

Cuando la condición (35.a) es apropiadamente aplicada al método Runge-Kutta, esta impone una restricción al valor del tamao del paso h. Esta restricción nunca debe ser confundida con la restricción

impuesta por los criterios de estabilidad, ni con el control del tamaño del paso. Para un análisis de estos últimos aspectos en el caso propuesto como ejemplo, referirse a [Granados,1996].

• Implícito Parcial

El proceso iterativo será ilustrado para el método Runge-Kutta de la cuadratura de lobato, implícito sólo en k_2 y k_3 del ejemplo $(4.a)$, en el caso the una ecuación diferencial ordinaria autónoma $\dot{y} = f(y)$. Así, la función de la ecuación homogénea es

$$\{\mathbf{g}(\mathbf{k})\} = \left\{ \begin{array}{c} g_2(\mathbf{k}) \\ g_3(\mathbf{k}) \end{array} \right\} = \left\{ \begin{array}{c} f(y_n + h\, a_{2s} k_s) - k_2 \\ f(y_n + h\, a_{3s} k_s) - k_3 \end{array} \right\} = \mathbf{0} \tag{36}$$

y tiene un jacobiano igual a

$$[\mathbf{J_g}(\mathbf{k})] = \left[\begin{array}{cc} h\, a_{22}\, f'(y_n + h\, a_{2s} k_s) - 1 & h\, a_{23}\, f'(y_n + h\, a_{2s} k_s) \\ h\, a_{32}\, f'(y_n + h\, a_{3s} k_s) & h\, a_{33}\, f'(y_n + h\, a_{3s} k_s) - 1 \end{array} \right] \tag{37}$$

Entonces el proceso iterativo se implementa como

$$[\mathbf{J_g}(\mathbf{k})]_m \left\{ \begin{array}{c} \Delta k_2 \\ \Delta k_3 \end{array} \right\}_m = - \left\{ \begin{array}{c} g_2(\mathbf{k}) \\ g_3(\mathbf{k}) \end{array} \right\}_m \qquad \left\{ \begin{array}{c} k_2 \\ k_3 \end{array} \right\}_{m+1} = \left\{ \begin{array}{c} k_2 \\ k_3 \end{array} \right\}_m + \omega \left\{ \begin{array}{c} \Delta k_2 \\ \Delta k_3 \end{array} \right\}_m \tag{38}$$

las iteraciones en m se realizan para cada paso de tamaño h (n constante) hasta que la condición (29) se satisfaga. Después de sto se calcula k_4 y y_{n+1}, y se puede intentar realizar la integraciópn en otro paso (siguiente n) con el mismo u otro tamaño.

• Implícito Total

Un método Runge-Kutta con tres etapas $N = 3$, como el ejemplo $(3.c)$, para un sistema autónomo de ecuaciones diferenciales ordinarias

$$\mathbf{k}_r = \mathbf{f}(\mathbf{y}_n + h\, a_{rs} \mathbf{k}_s) \qquad r, s = 1, 2, \ldots, N \tag{39}$$

encuentra la solución del paso siguiente $\mathbf{y}_{n+1}$, con un error local del orden de h^{P+1}, como

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\, c_r\, \mathbf{k}_r + O(h^{P+1}) \tag{40}$$

El error global es del orden de h^P. Para resolver las incógnitas $\mathbf{k}_s$ en cada paso, se establece el sistema de ecuaciones no lineales, así se puede aplica el método de Newton-Raphson a la función $\mathbf{g}(\mathbf{k})$

$$\mathbf{g}(\mathbf{k}) = \mathbf{F}(\mathbf{k}) - \mathbf{k} = \mathbf{0} \qquad \mathbf{F}(\mathbf{k}) = \left\{ \begin{array}{c} \mathbf{f}(\mathbf{y}_n + h\, a_{1s}\, \mathbf{k}_s) \\ \mathbf{f}(\mathbf{y}_n + h\, a_{2s}\, \mathbf{k}_s) \\ \mathbf{f}(\mathbf{y}_n + h\, a_{3s}\, \mathbf{k}_s) \end{array} \right\} \qquad \mathbf{g}(\mathbf{k}) = \left\{ \begin{array}{c} \mathbf{f}(\mathbf{y}_n + h\, a_{1s}\, \mathbf{k}_s) - \mathbf{k}_1 \\ \mathbf{f}(\mathbf{y}_n + h\, a_{2s}\, \mathbf{k}_s) - \mathbf{k}_2 \\ \mathbf{f}(\mathbf{y}_n + h\, a_{3s}\, \mathbf{k}_s) - \mathbf{k}_3 \end{array} \right\} \tag{41}$$

El jacobiano de dicha función es

$$[\mathbf{J_g}(\mathbf{k})] = \left[\begin{array}{ccc} h\, a_{11}\, [\, \mathbf{J_f}(\mathbf{y}_n + h\, a_{1s}\mathbf{k}_s)\,] - [\,\mathbf{I}\,] & h\, a_{12}\, [\, \mathbf{J_f}(\mathbf{y}_n + h\, a_{1s}\mathbf{k}_s)\,] & h\, a_{13}\, [\, \mathbf{J_f}(\mathbf{y}_n + h\, a_{1s}\mathbf{k}_s)\,] \\ h\, a_{21}\, [\, \mathbf{J_f}(\mathbf{y}_n + h\, a_{2s}\mathbf{k}_s)\,] & h\, a_{22}\, [\, \mathbf{J_f}(\mathbf{y}_n + h\, a_{2s}\mathbf{k}_s)\,] - [\,\mathbf{I}\,] & h\, a_{23}\, [\, \mathbf{J_f}(\mathbf{y}_n + h\, a_{2s}\mathbf{k}_s)\,] \\ h\, a_{31}\, [\, \mathbf{J_f}(\mathbf{y}_n + h\, a_{3s}\mathbf{k}_s)\,] & h\, a_{32}\, [\, \mathbf{J_f}(\mathbf{y}_n + h\, a_{3s}\mathbf{k}_s)\,] & h\, a_{33}\, [\, \mathbf{J_f}(\mathbf{y}_n + h\, a_{3s}\mathbf{k}_s)\,] - [\,\mathbf{I}\,] \end{array} \right] \tag{42}$$

Un procedimiento iterativo se aplica después para obtener siguientes iterados

$$[\mathbf{J_g}(\mathbf{k}_m)] \left\{ \begin{array}{c} \Delta\mathbf{k}_1 \\ \Delta\mathbf{k}_2 \\ \Delta\mathbf{k}_3 \end{array} \right\}_m = -\mathbf{g}(\mathbf{k}_m) \qquad \left\{ \begin{array}{c} \mathbf{k}_1 \\ \mathbf{k}_2 \\ \mathbf{k}_3 \end{array} \right\}_{m+1} = \left\{ \begin{array}{c} \mathbf{k}_1 \\ \mathbf{k}_2 \\ \mathbf{k}_3 \end{array} \right\}_m + \omega \left\{ \begin{array}{c} \Delta\mathbf{k}_1 \\ \Delta\mathbf{k}_2 \\ \Delta\mathbf{k}_3 \end{array} \right\}_m \tag{43}$$

dentro de un único paso h (n constante). Las iteraciones se calculan (índice m) hasta que se asegure la convergencia en (29).

3.4. ESTABILIDAD

La estabilidad de los métodos Runge-Kutta se establecen mediante el análisis del problema

$$\frac{dy}{dx} = f(y) \qquad\qquad f(y) = \lambda\, y \tag{1}$$

La aplicación del método Runge-Kutta el problema (1) da

$$k_r = f(y_n + h\, a_{rs} k_s) = \lambda\,(y_n + h\, a_{rs} k_s) = \lambda y_n + h\,\lambda\, a_{rs} k_s \tag{2}$$

Agrupando las k_s y extrayendo el factor común, se obtiene

$$(\delta_{rs} - h\lambda\, a_{rs}) k_s = \lambda y_n\, \delta_r \qquad\qquad [\,\mathbf{I} - h\lambda \mathbf{A}\,].\{\mathbf{k}\} = \{\mathbf{1}\}\,\lambda y_n \tag{3}$$

La matriz $[\mathbf{A}]$ contiene los coeficientes a_{rs} del método. El vector $\mathbf{k}$ representa las k_s para todas las etapas en un arreglo. El vector $\{\mathbf{1}\}$ es un vector lleno de 1 en todas sus componentes, y $[\,\mathbf{I}\,]$ es la matriz identidad. La dimensión del sistema (3) es el número de etapas N. Resolviendo el sistema de ecuaciones para el vector $\mathbf{k}$ se obtiene

$$\mathbf{k} = [\,\mathbf{I} - h\lambda\,\mathbf{A}\,]^{-1}.\{\mathbf{1}\}\,\lambda y_n \tag{4}$$

y computando la nueva integración y_{n+1}

$$\begin{aligned} y_{n+1} &= y_n + h\, c_r k_r \qquad\qquad y_{n+1} = y_n + h\,\mathbf{c}.\mathbf{k} \\ &= y_n + h\,\mathbf{c}\,.\,[\,\mathbf{I} - h\lambda\,\mathbf{A}\,]^{-1}.\{\mathbf{1}\}\,\lambda y_n = \left\{\, 1 + \mathbf{c}\,.\,[\,\mathbf{I} - h\lambda\,\mathbf{A}\,]^{-1}.\{\mathbf{1}\}\, h\lambda \,\right\} y_n \\ &= \mu(h\lambda)\, y_n \end{aligned} \tag{5}$$

donde la función involucrada $\mu(z)$, $z = h\lambda$, es la denominada *raíz característica* del método y puede ser expresada por

$$\mu(z) = 1 + \mathbf{c}\,.\,[\,\mathbf{I} - z\,\mathbf{A}\,]^{-1}.\{\mathbf{1}\}\, z \tag{6}$$

El método se dice que es "estable" en los rangos de $z = h\lambda$ donde $\mu(z)$ es menor en valor absoluto que la unidad. Si $|\mu(h\lambda)|$ es menor que la unidad, entonces se satisface que $|y_{n+1}|$ es menor que $|y_n|$ y la estabilidad es garantizada.

Para el método 3.3.(3.c), método de Cuadratura de Gauss, la función de la raíz característica es [Lapidus & Seinfeld,1971,p.135]

$$\mu(z) = \frac{1 + \frac{1}{2}z + \frac{1}{10}z^2 + \frac{1}{120}z^3}{1 - \frac{1}{2}z + \frac{1}{10}z^2 - \frac{1}{120}z^3} \tag{7}$$

una aproximación de Padé para e^z en la posición diagonal de la tabla [Burden & Faires,1985] [Lapidus & Seinfeld,1971]. In este caso $|\mu(h\lambda)| < 1$ para la parte real $\Re(\lambda) < 0$, por lo tanto el método se dice que es *A-stable* o absolutamente estable.

Para el método 3.3.(4.a), método de Cuadratura de Lobatto, la función de la raíz característica es [Granados,(1996)]

$$\mu(z) = \frac{1 + \frac{2}{3}z + \frac{1}{5}z^2 + \frac{1}{30}z^3 + \frac{1}{360}z^4}{1 - \frac{1}{3}z + \frac{1}{30}z^2} \tag{8}$$

y también es una aproximación de Padé de e^z. La expresión (8) es siempre positiva y menor que la unidad en el intervalo $z \in (\,-9.648495248\,,\,0.0\,) = (-a, 0)$, donde $a = 4 + b - 4/b$ y $b = \sqrt[3]{124 + 4\sqrt{965}} = 6.284937532$. Se puede notar que la función $\mu(z)$ se aproxima relativamente bien a la función $y = \mathrm{e}^z$ para el rango $z \gtrsim -4$,

donde cercanamente tiene un mínimo ($z = -3.827958538$). Aunque existe un máximo local y un mínimo local alrededor de $z = 6.276350$ y $z = 12.278646$, respectivamente (no se ven en la gráfica de la figura 1).

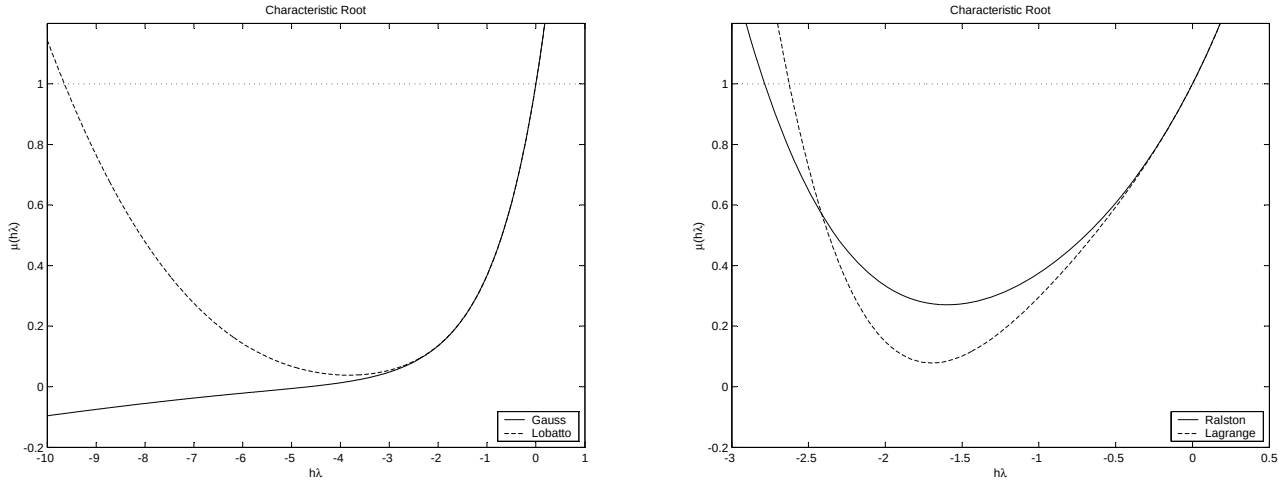


Figura 1. Raíz Característica para cuadraturas de Gauss y de Lobatto implícitos (izquierda). Raíz Caracter. para los coeficientes del método de Lobatto explícito, calc. con Ralston y Lagrange (derecha).

La figura 1 (izquierda) muestra las funciones de las raíces características para los métodos de cuadratura de Lobatto 3.3.(4.a) con la ec. (8), y el método de cuadratura de Gauss 3.3.(3.c) con la ec. (7) en el intervalo $h\lambda \in [-10, 1]$. Un aspecto especial es el hecho de que, siendo irracionales la mayor parte de los coeficientes de los métodos, las funciones de las raíces características obtenidas con (6) son completamentes racionales.

La figura 1 (derecha) compara las funciones de la raíces características en el intervalo $h\lambda \in [-3, 0.5]$, de la familia de métodos de Ralston ec. 3.2.(1), R-K explícito ($P = 4$) eq. 3.3.(6), y mediante la generación con extrapolación de Lagrange ec. 3.2.(9), R-K explicito ($P = 4$) ($a_{21} = (5 - \sqrt{5})/10$, $a_{31} = -(5 + 3\sqrt{5})/10$, $a_{32} = (5 + 2\sqrt{5})/5$, $a_{41} = 1$, $a_{42} = -\sqrt{5}$, $a_{43} = \sqrt{5}$, $a_{rs} = 0$ $s \geq r$), ambos con los mismos puntos de colocación ($b_1 = 0$, $b_2 = (5 - \sqrt{5})/10$, $b_3 = (5 + \sqrt{5})/10$, $b_4 = 1$) del método de cuadratura de Lobatto. Las curvas están muy cercanas entre sí con mínimos en ($h\lambda = -1.5961$, $\mu = 0.2704$) y ($h\lambda = -1.6974$, $\mu = 0.0786$) y límites de estabilidad en ($h\lambda = -2.7853$, $\mu = 1$) y ($h\lambda = -2.625253$, $\mu = 1$), para los tipos de Ralston y Lagrange, respectivamente. Esto significa que las características de estabilidad para ambos métodos son similares.

Para el caso del método Runge-Kutta implícito tipo Lobatto de tercer orden ($P = 3$), resumido en los coeficientes 3.3.(4.b), se obtiene

$$\tilde{\mu}(z) = \left[\frac{1 + \frac{2}{3}z + \frac{1}{5}z^2 + \frac{1}{30}z^3}{1 - \frac{1}{3}z + \frac{1}{30}z^2} \right] \tag{9}$$

Expresión que ya no es una aproximación de Padé.

Los límites de estabilidad de los métodos de cuadratura de Lobatto son los siguientes

$$h \leq \frac{6.8232}{|\lambda|} \qquad \text{(Método Implícito de 3er orden)} \tag{10}$$

$$h \leq \frac{9.6485}{|\lambda|} \qquad \text{(Método Implícito de 6to orden)} \tag{11}$$

Las condiciones (10) y (11) revelan que los métodos Runge-Kutta implícitos tipo Lobatto de tercer y sexto órdenes son más estables que el método Runge-Kutta explícito tipo Fehlberg de cuarto y quinto

órdenes, los cuales poseen las siguientes condiciones de estabilidad

$$h \leq \frac{2.785}{|\lambda|} \qquad \text{(Método Explícito 4to orden)} \tag{12}$$

$$h \leq \frac{3.15}{|\lambda|} \qquad \text{(Método Explícito 5to orden)} \tag{13}$$

Una consideración importante es que los método implícitos son mucho más estables que los métodos explícitos, con rangos de estabilidad mucho más amplios. Esto permite escoger tamaños de pasos mucho más grandes para los métodos implícitos, lo que reduce el tiempo de cómputo substancialmente, aunque los algoritmos numéricos seam más complejos como se acaba de ver.

3.5. RESULTADOS

La figura 1 muestra los resultados de comparar los métodos de cuadratura de Lobatto implícitos de tercer y sexto órdenes (RKI36) y Fehlberg de cuarto y quinto órdenes (RKF45). Ambos con control de paso y con $n_s = 2000$ pasos. Se observa que RKF45 luego de realizar 4 órbitas se vuelve inestable, mientras que RKI36 continua siendo estable.

(espacio vacío)

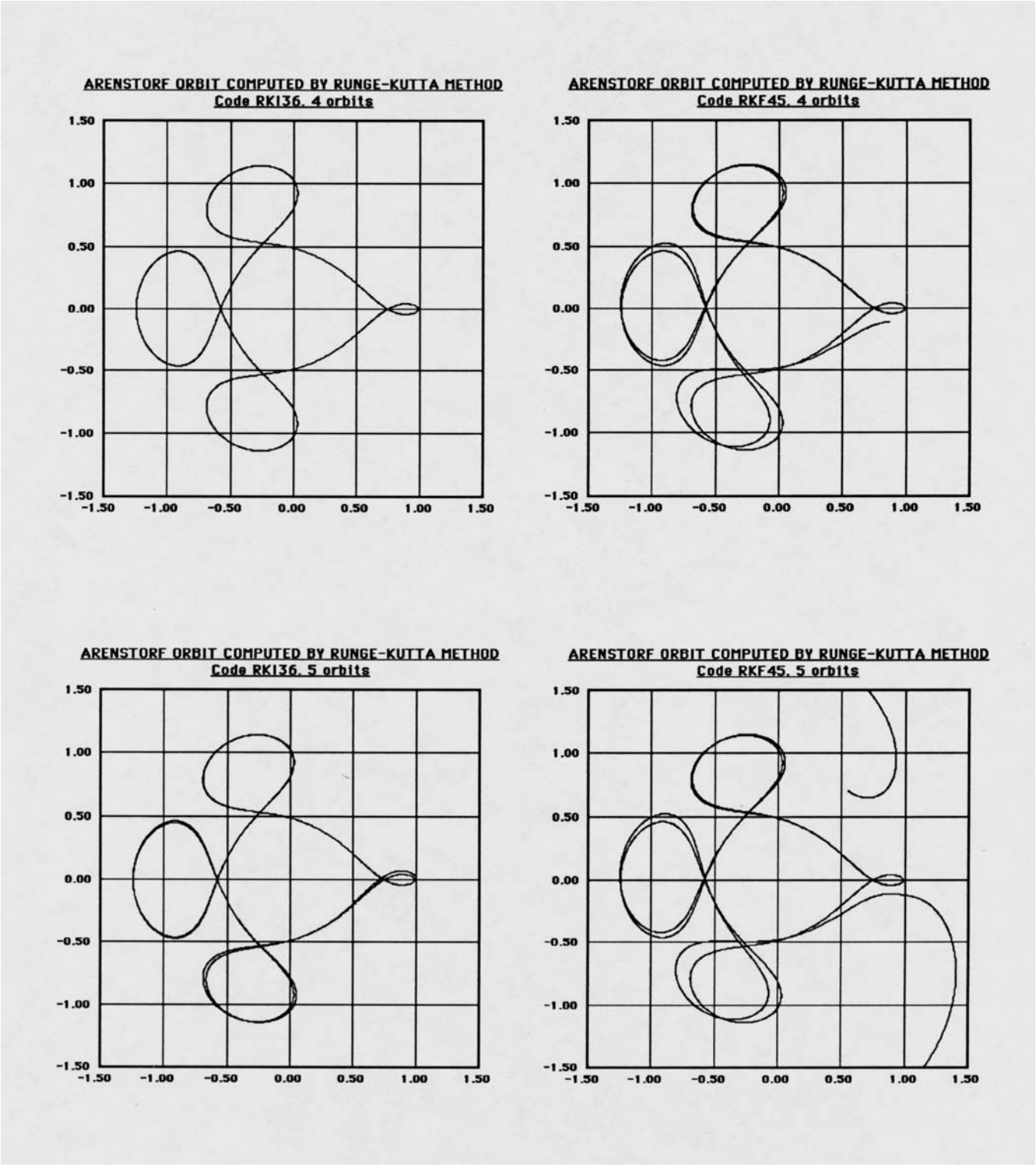


Figura 1. Orbitas de Arenstorf computadas con métodos Runge-Kutta implícitos (RKI36) y explícitos (RKF45) con $n_s = 2000$ pasos por órbita (para impresión) con control de paso automático (4 y 5 órbitas).

La figura 2 muestra los resultados para el cómputo de las órbitas de Arenstorf con el método de cuadratura de Gauss 3.3.(3.c), con tres tipos de procedimientos para resolver las variables auxiliares $\mathbf{k}_r$.

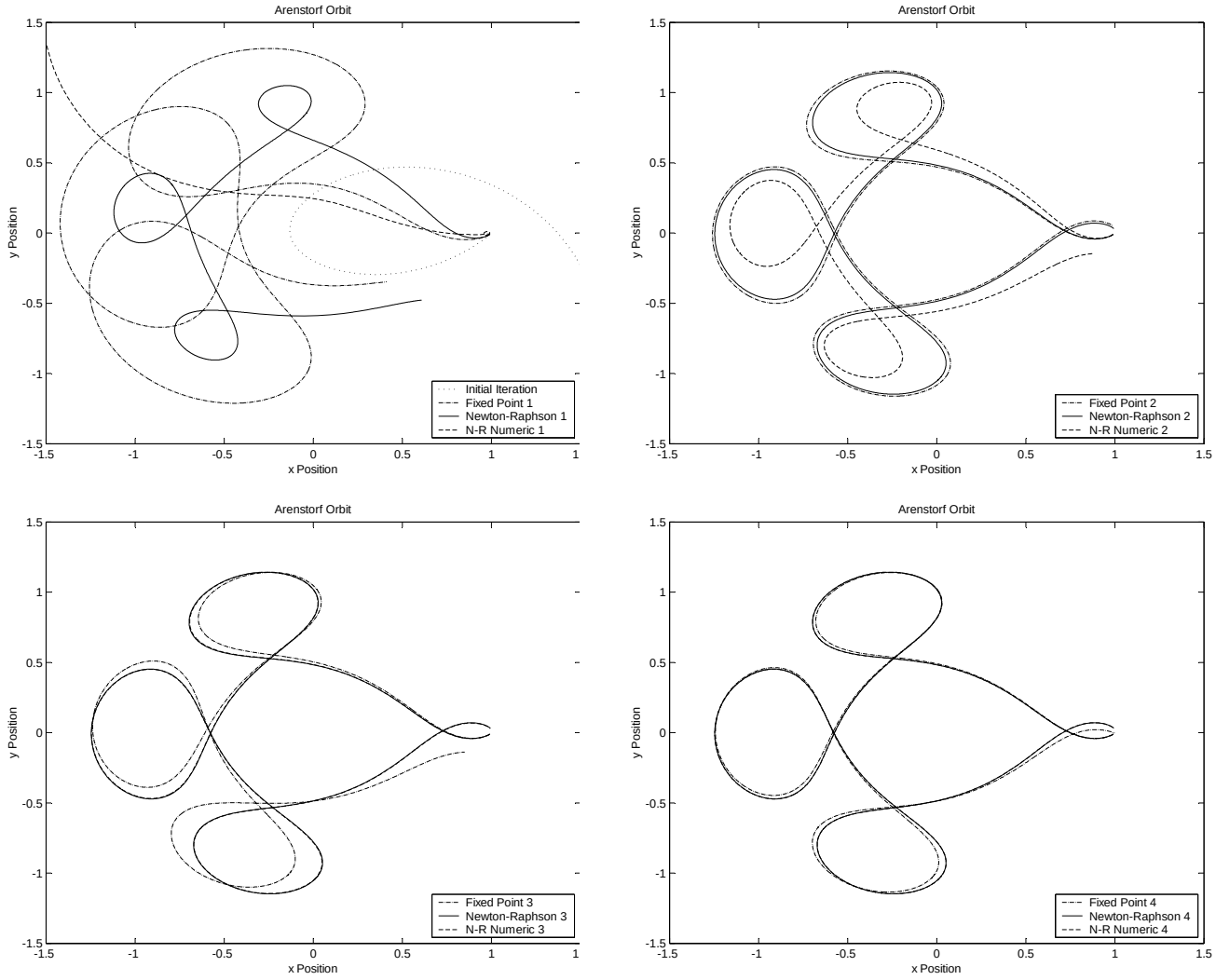


Figura 2. Una órbita de Arenstorf computada con Runge-Kutta implícito equidistante ($n_s = 3000$ pasos). Punto Fijo, Newton-Raphson Analítico, y Newton-Raphson Numérico. Una, dos, tres, y cuatro iteraciones.

El problema 3.1.(7) − (10) fué resuelto para $n_s = 3000$ pasos en un período T ($h = T/n_s$), que es una órbita completa. El problema fué reformulado en 3.1.(11) − (14) con la finalidad de calcular el jacobiano $[\mathbf{J_f}]$ más fácilmente. Con los valores iniciales estimados con la precisión mostrada en las ecuaciones, un proceso iterativo (en el índice m para cada paso) se implementó con tres variantes: Método de Punto Fijo, Método de Newton-Raphson, y Método de Newton-Raphson con el jacobiano $[\mathbf{J_g}]$ cálculado numéricamente. De una a cuatro iteraciones fueron ejecutadas en cada procedimiento, identificados en la leyenda de las gráficas en la esquina inferior derecha con los dígitos 1, 2, 3 ó 4. Para obtener una solución razonable con el métodos explícito 3.2.(7.a), linea punteada en la primera pantalla (arriba-izquierda) en la figura 2, identificada con 'Iteración Inicial', fué necesario más de $n_s = 7.3 \times 10^5$ pasos. Con $n_s = 3000$ este método produce un espiral incremental que se sale de pantalla en el sentido del reloj con centro en (1,0). Lo mismo para el Método de Newton-Raphson Numérico, una iteración, pero con una espiral más grande.

Para este caso ($n_s = 3000$), dos iteraciones son suficientes para el Método Newton-Raphson Analítico

y cuatro iteraciones para el Método de Punto Fijo para obtener una buena precisión. Con tres iteraciones, tanto los Métodos Newton-Raphson Analítico como el Numérico son equivalentes en precisión.

La figura 3 muestra los resultados para $n_s = 2000$, cerca de los límites de estabilidad. El Método Newton-Raphson Analítico necesita dos iteraciones para tener una buena ejecución. En cambio, el Método Newton-Raphson Numérico y Método de Punto Fijo necesitan cuatro iteraciones para estabilizarse, pero el segundo se comportó mejor.

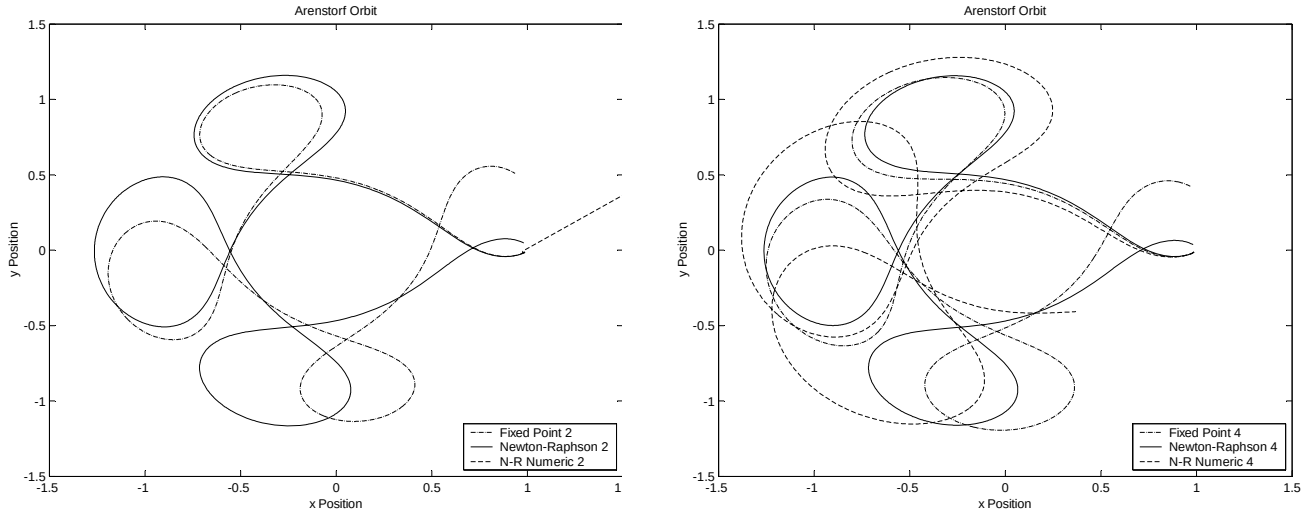


Figura 3. Una órbita de Arenstorf computada con Runge-Kutta implícito equidistante ($n_s = 2000$ pasos). Punto Fijo, Newton-Raphson Analítico, y Newton-Raphson Numérico. Dos y cuatro iteraciones.

La conclusión es que los métodos se han ordenados del mejor al peor: Newton-Raphson Analítico, Newton-Raphson Numérico y Punto Fijo, para alto número de iteraciones ($m \geq 4$). Para bajo número de iteraciones ($m = 2$), Newton-Raphson Numérico y Punto fijo son competitivos por razones numéricas al calcular estimaciones de las derivadas en la matriz jacobiana, pero esto declina con el incremento del número de iteraciones o la disminución del tamaño del paso, cuando los métodos Newton-Raphson Numérico y Analítico se vuelven equivalentes.

BIBLIOGRAFIA

[1] Burden R. L.; Faires, J. D. **Numerical Analysis**, 3^{rd} Edition. PWS (Boston), 1985.

[2] Butcher, J. C. "Implicit Runge-Kutta Processes", **Math. Comput.**, Vol.**18**, pp.50-64, (1964a).

[3] Butcher, J. C. "On Runge-Kutta Processes of High Order", **J. Austral. Math. Soc.**, Vol.**IV**, Part 2, pp.179-194, (1964b).

[4] Butcher, J. C. **The Numerical Analysis of Ordinary Differential Equations, Runge-Kutta and General Linear Methods**. John Wiley (New York), 1987.

[5] Butcher, J. C. **Numerical Methods for Ordinary Differential Equations**, $2^{nd}/3^{rd}$ Editions. John Wiley & Sons (New York), 2008/2016.

[6] Cash, J. R.; Karp, A. H. **ACM Transactions on Mathematical Software**, Vol.**16**, pp.201-222, (1990).

[7] Chapra S. C.; Canale, R. P. **Métodos Numéricos para Ingenieros**, Quinta Edición. McGraw-Hill Interamericana Editores (México), 2007.

[8] Dormand, J. R.; Prince, P. J. "A Family of Embedded Runge-Kutta Formulae", **J. Comp. Appl. Math.**, Vol.**6**, pp.19-26, (1980).

[9] Fehlberg, E. "Low-Order Classical Runge-Kutta Formulas with Stepsize Control", **NASA Report No.** TR R-315, 1971.

[10] Gear, C. W. **Numerical Initial Value Problems in Ordinary Differential Equations**. Prentice-Hall (Englewood Cliffs-New Jersey), 1971.

[11] Gerald, C. F. **Applied Numerical Analysis**, 2^{nd} Edition. Addison-Wesley (New York), 1978.

[12] Granados M., A. L. "Lobatto Implicit Sixth Order Runge-Kutta Method for Solving Ordinary Differential Equations With Stepsize Control", **Mecánica Computacional, Vol.XVI**, compilado por G. Etse y B. Luccioni (AMCA, Asociación Argentina de Mecánica Computacional), pp.349-359, (1996).

[13] Granados M., A. L. "Implicit Runge-Kutta Algorithm Using Newton-Raphson Method". **Simulación con Métodos Numéricos: Nuevas Tendencias y Aplicaciones**, Editores: O. Prado, M. Rao y M. Cerrolaza. Memorias del *IV CONGRESO INTERNACIONAL DE METODOS NUMERICOS EN INGENIERIA Y CIENCIAS APLICADAS, CIMENICS'98*. Hotel Intercontinental Guayana, 17-20 de Marzo de 1998, Puerto Ordaz, Ciudad Guayana. Sociedad Venezolana de Métodos Numéricos en Ingeniería (SVMNI), pp.TM9-TM16. Corregido y ampliado Abril, 2016. https://www.academia.edu/11949052/Implicit_Runge-Kutta_Algorithm_Using_Newton-Raphson_Method

[14] Granados M., A. L. "Implicit Runge-Kutta Algorithm Using Newton-Raphson Method". *Fourth World Congress on Computational Mechanics*, realizado en el Hotel Sheraton, Buenos Aires, Argentina, 29/Jun/98 al 2/Jul/98. International Association for Computational Mechanics, **Abstracts**, Vol.**I**, p.37, (1998).

[15] Hairer, E.; Nørsett, S. P.; Wanner, G. **Solving Ordinary Differential Equations I. Nonstiff Problems**. Springer-Verlag (Berlin), 1987.

[16] Hairer, E.; Wanner, G. **Solving Ordinary Differential Equations II: Stiff and Differential - Algebraic Problems**. Springer-Verlag (Berlin), 1991.

[17] Hazewinkel, M. **Encyclopaedia of Mathematics**. Kluwer Academic Publishers (Dordrecht), 1988.

[18] Lapidus, L.; Seinfeld, J. H. **Numerical Solution of Ordinary Differential Equations**. Academic Press (New York), 1971.

[19] Lobatto, R. **Lessen over Differentiaal- en Integraal-Rekening**. 2 Vols. (La Haye), 1851-52.

[20] Ralston, A.; Rabinowitz, P. **A First Course in Numerical Analysis**, 2^{nd} Edition. McGraw-Hill (New York), 1978.

[21] Shampine, L. F.; Watts, H. A.; Davenport, S. M. "Solving Non-Stiff Ordinary Differential Equations - The State of the Art". **SANDIA** Laboratories, Report No. SAND75-0182, 1975. **SIAM Review**, Vol.**18**, No.3, pp.376-411, (1976).

[22] Sommer, D. "Numerische Anwendung Impliziter Runge_Kutta-Formeln", ZAMM, Vol.**45** (Sonderheft), pp.T77-T79, (1965).

[23] van der Houwen, P. J.; Sommeijer, B. P. "Iterated Runge-Kutta Methods on Parallel Computers". **SIAM J. Sci. Stat. Comput.**, Vol.**12**, No.5, pp.1000-1028, (1991).

CAPITULO V

ECUACIONES EN DERIVADAS PARCIALES

CONTENIDO

1. INTRODUCCION

1.1. FUNDAMENTOS

Las ecuaciones diferenciales en derivadas parciales son todas aquellas que poseen términos en función de derivadas parciales. Pueden ser lineales con coeficientes constantes, variables (función de x y y y sus derivadas parciales en 2D) y no lineales (potencias o funciones trascendentes de derivadas parciales).

1.2. CLASIFICACION DE LAS ECUACIONES

Sea $f(x,y)$ una función de x y y, entonces la ecuación diferential en derivadas parciales de segundo orden con coeficientes constantes o dependientes de x y y

$$a\,\frac{\partial^2 f}{\partial x^2} + b\,\frac{\partial^2 f}{\partial x \partial y} + c\,\frac{\partial^2 f}{\partial y^2} + d\,\frac{\partial f}{\partial x} + e\,\frac{\partial f}{\partial y} + h\,f + g = 0 \tag{1}$$

válida en un dominio $\mathcal{D}$ y con condiciones de contorno en la frontera $\partial\mathcal{D}$ de tipo valor (Dirichlet), derivada (Neumann) o mixtas, se dice que es *lineal*. Si adicionalmente a, b y c dependenden de f, $\partial f/\partial x$ y $\partial f/\partial y$, y d y e dependen de f, se dice que es *cuasi-lineal*. En caso contrario, se dice que es *no-lineal*.

Si definimos el *discriminante*

$$\Delta = b^2 - 4\,a\,c \tag{2}$$

entonces la ecuación diferencial (1) se clasifica como:

$\Delta < 0$ Ecuación elíptica e.g.

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = S \qquad \Delta = -4 \tag{3.a}$$

$\Delta = 0$ Ecuación parabólica e.g.

$$\frac{\partial U}{\partial x} = \Gamma\,\frac{\partial^2 U}{\partial y^2} \qquad \Delta = 0 \tag{3.b}$$

$\Delta > 0$ Ecuación hiperbólica e.g.

$$\frac{\partial^2 U}{\partial x^2} = C^2\,\frac{\partial^2 U}{\partial y^2} = 0 \qquad \Delta = 4\,C^2 \tag{3.c}$$

La ecuación (3.a) recibe el nombre de la ecuación de *Laplace* si $S = 0$ o de *Poisson* si $S \neq 0$. La ecuación (3.b) recibe el nombre de *difusión transitoria* ($x = t$). La ecuación (3.c) recibe el nombre de la *ecuación de onda* ($x = t$).

1.3. CONSISTENCIA, ESTABILIDAD y CONVERGENCIA

Si designamos por:

$U(x_i)$ Solución analítica de la ecuación diferencial.

U_i Solución exacta del esquema numérico.

u_i Solución numérica hallada por la computadora.

Vamos a definir los siguientes términos:

Error de Truncamiento. $E_i^t = |U(x_i) - U_i|$ Es el error causado por el método numérico y se debe al hecho de que el método en cuestión se origina de una serie truncada.

Error de Redondeo. $E_i^r = |U_i - u_i|$ Es el error causado por el uso de un número limitado de dígitos en los cálculos que realiza el computador.

Error Total. $E_i = |U(x_i) - u_i| \leq E_i^t + E_i^r$ Es el error global causado por los dos aspectos anteriormente mencionados, pero que no son simplemente aditivos.

Consistencia. Un esquema numérico es consistente si el error e truncamiento tiende a anularse cuando Δx tiende a cero ($\Delta x \to 0 \Longrightarrow E_i^t \to 0$).

Estabilidad. Un esquema numérico es estable si el error de redondeo tiende a anularse cuando Δx tiende a cero ($\Delta x \to 0 \Longrightarrow E_i^r \to 0$).

Convergencia. (Teorema de Lax) Si un esquema numérico es consistente y estable, entonces converge ($\Delta x \to 0 \Longrightarrow E_i \to 0$).

2. METODO DE DIFERENCIAS FINITAS

2.1. ECUACIONES ELIPTICAS

2.1.1. Discretización del Laplaciano

El *laplaciano* se discretiza con diferencias centrales (III.1.7.(2.b)) para intervalos regulares como

$$(\nabla^2 \mathbf{u})_i = I\!L(U_i) + O(h^2) = \frac{U_{i-1} - 2\,U_i + U_{i+1}}{h^2} + O(h^2) \tag{1}$$

Para intervalos irregulares como

$$I\!L(U_i) = 2\,U[x_{i-1}, x_i, x_{i+1}] \tag{2}$$

hallada con la parábola que pasa por los puntos x_{i-1}, x_i y x_{i+1}. Se ha usado la notación de los polinomios de Newton en diferencias divididas.

Para dos dimensiones cartesianas la expresión (1) se convierte en

$$(\nabla^2 \mathbf{u})_{i,j} = I\!L(U_{i,j}) + O(h^2) = \frac{U_{i-1,j} - 2\,U_{i,j} + U_{i+1,j}}{h_x^2} + \frac{U_{i,j-1} - 2\,U_{i,j} + U_{i,j+1}}{h_y^2} + O(h^2) \tag{3}$$

siendo $h^2 = h_x^2 + h_y^2$. Para dimensión tres, el resultado es similar para $I\!L(U_{i,j,k})$. En los tres casos los términos con U_i, $U_{i,j}$ y $U_{i,j,k}$ se acumulan teniendo al final un coeficiente -2^D (D = dimensión).

En el caso unidimensional en intervalos regulares, la substitución de la ecuación (1) para cada punto $i = 1, 2, \ldots, N$ discretizados para la ecuación de Laplace

$$\nabla^2 \mathbf{u} = 0 \qquad u(a) = \alpha \qquad u(b) = \beta \tag{4}$$

resulta en un sistema de ecuaciones lineales en las incógnitas U_i, con una matriz tridiagonal, donde en los extremos se aplica las condiciones de valor en la frontera $U_0 = u(a) = \alpha$ y $U_{N+1} = u(b) = \beta$.

2.1.2. Término de Fuente

En las discretizaciones antes realizadas, para la ecuación de Poisson, el término de fuente debe evaluarse para el término central S_i, $S_{i,j}$ y $S_{i,j,k}$, según el caso.

2.1.3. Término Adventivo

En algunas ecuaciones semi-elípticas, como la ecuación de Burgers o de Navier-Stokes, contiene un término *adventivo* $I\!H(\mathbf{u}) = \mathbf{u}.\boldsymbol{\nabla}\mathbf{u}$, que se discretiza según tenga una influencia aguas-arriba o aguas-abajo. Aunque el término adventivo tiene carácter hiperbólico, aunque este tipo de ecuaciones no cae dentro de la clasificación dada en la sección 1.2.

El *adventivo* se discretiza con diferencias no-centrales (III.1.7.(1.c)) para intervalos regulares como

$$\mathbf{u}.(\boldsymbol{\nabla}\mathbf{u})_i = I\!H(U_i) + O(h^2) = U_i \, \frac{U_{i-2} - 4\,U_{i-1} + 3\,U_i}{2\,h} + O(h^2) \tag{5.a}$$

$$\mathbf{u}.(\boldsymbol{\nabla}\mathbf{u})_i = I\!H(U_i) + O(h^2) = U_i \, \frac{-3\,U_i + 4\,U_{i+1} - U_{i+2}}{2\,h} + O(h^2) \tag{5.b}$$

la expresión (5.a) para aguas-abajo y la expresión (5.b) para aguas-arriba.

Para intervalos irregulares como

$$I\!H(U_i) = U(x_i) \, \{ \, U[x_{i-1}, x_i] + (x_i - x_{i-1}) \, U[x_{i-2}, x_{i-1}, x_i] \, \} \tag{6.a}$$

$$I\!H(U_i) = U(x_i) \, \{ \, U[x_i, x_{i+1}] + (x_i - x_{i+1}) \, U[x_i, x_{i+1}, x_{i+2}] \, \} \tag{6.b}$$

halladas con la parábola que pasa por los puntos x_{i-2}, x_{i-1} y x_i o los puntos x_i, x_{i+1} y x_{i+2}.

2.2. ECUACIONES PARABOLICAS

Estos métodos los vamos a explicar con el ejemplo de la ecuación de difusión transitoria unidimensional

$$\frac{\partial \mathbf{u}}{\partial t} = \Gamma \, \frac{\partial^2 \mathbf{u}}{\partial x^2} \qquad \begin{cases} u(0,x) = u_o(x) \\ u(a,t) = \alpha(t) \\ u(b,t) = \beta(t) \end{cases} \tag{1}$$

donde del lado derecho se han colocado las condiciones iniciales y las condiciones de contorno de valor en la frontera.

2.2.1. Método de Euler

El término transitorio $\partial \mathbf{u}/\partial t$ se discretiza de dos formas: explícita e implícita.

La discretización explícita es

$$\left. \frac{\partial \mathbf{u}}{\partial t} \right]_i^t = \frac{U_i^{t+1} - U_i^t}{\Delta t} + O(\Delta t) = \Gamma \, I\!L(U_i^t) + O(h^2) = \Gamma \, \frac{U_{i-1}^t - 2\,U_i^t + U_{i+1}^t}{h^2} + O(h^2) \tag{2}$$

Se acostumbra a agrupar los órdenes del error de truncamiento bajo un mismo símbolo $O(\Delta t+h^2)$. Despejando U_i^{t+1} se obtiene

$$U_i^{t+1} = \frac{\Gamma\Delta t}{h^2}\,U_{i-1}^t + \left(1 - \frac{2\,\Gamma\Delta t}{h^2}\right) U_i^t + \frac{\Gamma\Delta t}{h^2}\,U_{i+1}^t \tag{3}$$

Este esquema iterativo es similar al de Jacobi (sección II.1.2.1), luego es convergente si la matriz es diagonalmente dominante. El método es estable si

$$CFL = \frac{\Gamma\Delta t}{h^2} \le \frac{1}{2} \tag{4}$$

El parámetro del lado izquierdo de la desigualdad anterior es lo que se denomina el factor CFL (Courant-Friedrichs-Lewy). Esta condición se llama así en honor a Richard Courant, Kurt Friedrichs y Hans Lewy que la describieron en un artículo en 1928.

La discretización implícita es

$$\left.\frac{\partial \mathbf{u}}{\partial t}\right]_i^{t+1} = \frac{U_i^{t+1} - U_i^t}{\Delta t} + O(\Delta t) \tag{5.a}$$

$$\frac{U_i^{t+1} - U_i^t}{\Delta t} = \Gamma\, I\!\!L(U_i^{t+1}) + O(\Delta t + h^2) = \Gamma\,\frac{U_{i-1}^{t+1} - 2\,U_i^{t+1} + U_{i+1}^{t+1}}{h^2} + O(\Delta t + h^2) \tag{5.b}$$

La primera derivada $\partial\mathbf{u}/\partial t$ en (2) y (5.a) se ha discretizado según III.1.7.(1.a).

Reorganizando la ecuación queda

$$\frac{\Gamma\Delta t}{h^2}\,U_{i-1}^{t+1} - \left(1 + \frac{2\,\Gamma\Delta t}{h^2}\right) U_i^{t+1} + \frac{\Gamma\Delta t}{h^2}\,U_{i+1}^{t+1} = -U_i^t \tag{6}$$

que aplicada a todos los puntos $i = 1, 2, \ldots, N$ forma un sistema de ecuaciones lineales en las incógnitas U_i^{t+1} con matriz tridiagonal. Esta se puede resolver con el algoritmo de Thomas. El método planteado es incondicionalmente estable.

2.2.2. Método de Crank-Nicolson

Si hacemos un promedio de los métodos anteriores explícitos e implícito se obtiene

$$\begin{aligned}\frac{U_i^{t+1} - U_i^t}{\Delta t} &= \frac{\Gamma}{2}\left[\,\frac{U_{i-1}^t - 2\,U_i^t + U_{i+1}^t}{h^2} + \frac{U_{i-1}^{t+1} - 2\,U_i^{t+1} + U_{i+1}^{t+1}}{h^2}\,\right] + O(\Delta t^2 + h^2)\\ &= \frac{\Gamma}{2}\left[\, I\!\!L(U_i^t) + I\!\!L(U_i^{t+1})\,\right] + O(\Delta t^2 + h^2)\end{aligned} \tag{7}$$

Reorganizando los términos U^{t+1} de una lado y los términos U^t del otro, queda

$$\frac{\Gamma\Delta t}{2\,h^2}\,U_{i-1}^{t+1} - \left(1 + \frac{\Gamma\Delta t}{h^2}\right) U_i^{t+1} + \frac{\Gamma\Delta t}{2\,h^2}\,U_{i+1}^{t+1} = -\frac{\Gamma\Delta t}{2\,h^2}\,U_{i-1}^t - \left(1 - \frac{\Gamma\Delta t}{h^2}\right) U_i^t - \frac{\Gamma\Delta t}{h^2}\,U_{i+1}^t \tag{8}$$

Se obtiene un sistema de ecuaciones en U_i^{t+1}, con matriz tridiagonal.

Cuando el método se aplica en dos dimensiones x y y con dos mallados de tamaños $h = \Delta x$ y $k = \Delta y$, entonces la estabilidad del método

$$\frac{U_{i,j}^{t+1} - U_{i,j}^t}{\Delta t} = \Gamma\left[\,(1-\eta)\, I\!\!L(U_{i,j}^t) + \eta\, I\!\!L(U_{i,j}^{t+1})\,\right] \tag{9}$$

viene determinada por $\eta > \eta_1$ estable, $\eta_1 > \eta > \eta_2$ estable oscilante y $\eta < \eta_2$ inestable, donde

$$\eta_1 = 1 - \frac{1}{4\lambda} \qquad \eta_2 = \frac{1}{2}\left(1 - \frac{1}{2\lambda}\right) \qquad \lambda = \frac{\Gamma\,\Delta t}{h^2 + k^2} \tag{10}$$

Los métodos (7) y (9) son aplicaciones particulares del método Runge-Kutta (Euler modificado tipo Heun) de segundo orden en Δt, para el sistema de ecuaciones diferenciales ordinarias planteado en dichas ecuaciones, como se verá en la sección siguiente [Crank & Nicolson,(1947)].

2.2.3. Método de Las Líneas

El método de las líneas consiste en discratizar sólamente en aquella direcciones donde la ecuación diferencial es elíptica y, en una dirección diferente, se hace la integración del sistemas de ecuaciones diferenciales ordinarias de primer orden que se origina, por otro método. Al hacer esta discretización se obtiene, por ejemplo,

$$\frac{dU_i}{dy} = I\!L(U_i) \qquad \frac{dU_{i,j}}{dz} = I\!L(U_{i,j}) \qquad \frac{dU_{i,j,k}}{dt} = I\!L(U_{i,j,k}) \tag{11}$$

donde $I\!L(U)$ es la discretización del laplaciano de U en U_i ó $U_{i,j}$ ó $U_{i,j,k}$, en el dominio de 1, 2 ó 3 dimensiones. El sistema de ecuaciones diferenciales ordinarias resultante se puede resolver con un método Runge-Kutta.

2.3. ECUACIONES HIPERBOLICAS

Considérese la ecuación parcial-diferencial de segundo orden en las dos variables x y t

$$a\,u_{xx} + b\,u_{xt} + c\,u_{tt} + e = 0 \tag{1}$$

Aquí hemos usado la notación de subíndices para representa las derivadas parciales. Los coeficientes a, b, c y e pueden ser funciones de x, t, u_x, u_t y u, así que la ecuación planteada es muy general. Cuando los coeficiente son independientes de u o sus derivadas, se dice que es *lineal*. Si son funciones de u, u_x o u_t (pero no u_{xx} o u_{tt}), se dice que es *cuasi-lineal* [Gerald,1970,p.442].

Asumimos $u_{xt} = u_{tx}$ por ser continuas (teorema de Clairaut). Para facilitar la manipulación, sean

$$p = \frac{\partial u}{\partial x} = u_x \qquad q = \frac{\partial u}{\partial t} = u_t \tag{2}$$

Escribimos los diferenciales de p y q

$$\begin{aligned} dp &= \frac{\partial p}{\partial x}\,dx + \frac{\partial p}{\partial t}\,dt = u_{xx}\,dx + u_{xt}\,dt \\ dq &= \frac{\partial q}{\partial x}\,dx + \frac{\partial q}{\partial t}\,dt = u_{tx}\,dx + u_{tt}\,dt \end{aligned} \tag{3}$$

Despejando estas ecuaciones para u_{xx} y u_{tt}, respectivamente, tenemos

$$\begin{aligned} u_{xx} &= \frac{dp - u_{xt}\,dt}{dx} = \frac{dp}{dx} - u_{xt}\,\frac{dt}{dx} \\ u_{tt} &= \frac{dq - u_{tx}\,dt}{dt} = \frac{dq}{dt} - u_{tx}\,\frac{dx}{dt} \end{aligned} \tag{4}$$

Substituyendo en (1) y re-arreglando la ecuación, obtenemos

$$a\,u_{xt}\,\frac{dt}{dx} - b\,u_{xt} + c\,u_{xt} - a\,\frac{dp}{dx} - c\,\frac{dq}{dt} + e = 0 \tag{5}$$

Ahora, multiplicando por dt/dx, finalmente nos queda

$$u_{xt}\left[a\left(\frac{dt}{dx}\right)^2 - b\left(\frac{dt}{dx}\right) + c\right] - \left[a\,\frac{dp}{dx}\,\frac{dt}{dx} + c\,\frac{dq}{dx} + e\,\frac{dt}{dx}\right] \tag{6}$$

Suponga que, en el plane $x - t$, definimos las curvas tales que la expresión entre los primeros corchetes se anulan. Sobre tales curvas, la ecuación diferencial original es equivalente a anular la segunda expresión entre corchetes. Esto es,

$$a\,m^2 - b\,m + c = 0 \tag{7}$$

donde $m = dt/dx = 1/C$, define la pendiente inversa de la *curva característica* antes mencionada. La solución de la ecuación diferencial (1) se obtiene de integrar

$$a\,m\,dp + c\,dq + e\,dt = 0 \tag{8}$$

Obviamente, el discriminante $\Delta = b^2 - 4ac$ de la ecuación (7), coincidente con el discriminante de (1) según sección 1.2 (clasificación de las ecuaciones), debe ser positivo para que (1) sea *hiperbólica* y este enfoque sea exitoso.

Sean dos líneas caracteísticas C^+ y C^-, dadas por las dos soluciones de la cuadrática (7). Sobre la línea característica C^+ hallamos la solución de (8) entre los puntos A inicial y P final. Sobre la línea característica C^- hallamos la solución de (8) entre los puntos B inicial y P final. Estas dos soluciones, obtenidas a partir de los puntos iniciales A y B donde es conocida la solución, permite obtener las condiciones para el punto único final P

$$\begin{aligned} a\,m^+\,(p_P - p_A) + c\,(q_P - q_A) + e\,\Delta t = 0 \\ a\,m^-\,(p_P - p_B) + c\,(q_P - q_B) + e\,\Delta t = 0 \end{aligned} \tag{9}$$

Esto conforma un sistema lineal de dos ecuaciones con dos incógnitas p_P y q_P. Los coeficientes a, c y e deben tomarse en promedio entre los puntos A y P o entre los puntos B y P, según el recorrido seguido.

Resolviendo el sistema da

$$p_P = \left(\frac{a_{AP}m^+}{c_{AP}} - \frac{a_{BP}m^-}{c_{BP}}\right)^{-1}\left[\left(\frac{a_{AP}m^+p_A}{c_{AP}} - \frac{a_{BP}m^-p_B}{c_{BP}}\right) + (q_A - q_B) - \left(\frac{e_{AP}}{c_{AP}} - \frac{e_{BP}}{c_{BP}}\right)\Delta t\right] \tag{10.a}$$

$$q_P = q_A - \frac{a_{AP}m^+}{c_{AP}}(p_P - p_A) - \frac{e_{AP}}{c_{AP}}\Delta t \tag{10.b}$$

Una vez resuelto para todos los puntos A y B regulares distanciados entre sí $2\Delta x$ se tienen las soluciones de p y q para los diferentes puntos P, desplazados en el tiempo Δt y ubicados en la mitad entre cada A y B. Se deben recorrer todos los puntos $x_i = x_0 + i\,\Delta x$, $i = 1, 2, \ldots, N$, y los puntos $x_0 = a$ y x_{N+1} ($x_N = b$, $\Delta x = (b-a)/N$) se utilizan para formular las condiciones de contorno en la frontera $u(a) = \alpha(t)$ y $x(b) = \beta(t)$. Todo el proceso iterativo comienza con las condiciones iniciales $u(0, x) = u_o(x)$, $x \in [a, b]$.

Luego que tenemos las soluciones de p y q, encontramos las soluciones de u mediante la integración de

$$du = p\,dx + q\,dt \qquad \begin{aligned} u_P &= u_A + \Delta u|_A^P & u_P &= u_B + \Delta u|_B^P \\ \Delta u|_A^P &= p_{AP}\Delta x + q_{AP}\Delta t & \Delta u|_B^P &= p_{BP}\Delta x + q_{BP}\Delta t \end{aligned} \tag{11}$$

Para que el problema esté bien formulado las condiciones iniciales y de borde de $p = u_x$ y $q = u_t$ deben ser conocidas.

3. METODO DE VOLUMENES FINITOS

Los métodos de *volúmenes Finitos* fueron diseñados básicamente para resolver problemas de transporte y su evolución. Por ello sus fundamentos se basan en la discretización de la ecuación de transporte de cantidades como: densidad, cantidad de movimiento lineal, cantidad de movimiento angular, temperatura, entalpía, entropía, concentración, especie, energía, disipación, vorticidad, etc. Su concepto fundamental es que el mallado divide el dominio en diminutos volúmenes, donde es cada uno y sus alrederores se satisfacen los mismos principios de conservación que en la totalidad del dominio.

3.1. FUNDAMENTOS

El método de volúmenes finitos de basan en unas pocas reglas que iremos mencionando en las siguientes secciones.

3.1.1. Cuatro Reglas Básicas

Las cuatro reglas básicas de este método son.

• Regla 1. Consistencia del Volumen de Control

Cada volumen de control en el que se divide el dominio está identificado por un punto central o nodo cuyo valor de la propiedad transportada es $\varphi_{_P}$. Los vúmenes vecinos igualmente se identifican con letras mayúsculas como en una brújula N, S, W, E, T y B (T-top, B-bottom). Entre cada par de volúmenes existe una superficie inteface que se identifica con las letras minúsculas n, s, w, e, t y b. En cada interfaz existe un flujo único, sea este calculado con, por ejemplo, los tres de los valores $\varphi_{_W}$, $\varphi_{_P}$ y $\varphi_{_E}$ que la contiene, u otro grupo de tres valores que también la contenga (en la misma dirección). El flujo en cada interfaz viene gobernado por la difusividad Γ_i en la interfaz, dependiente de la difusividades de los nodos vecinos $\Gamma_{_L}$ y $\Gamma_{_R}$. También interviene una velocidad u_i perpendicular a la interfaz i, la cual determina un número de Peclet $I\!P_i = \rho_i u_i \delta x_i/\Gamma_i$ único (Peclet de malla). Por consiguiente, la interpolación de la cuadrática que pasa por tres nodos, típica de las diferencias finitas, no es consistente. La interfaz tiene una identidad propia independiente del volumen precedente o el volumen posterior y determinada por las condiciones de flujo dependientes del número de Peclet y el gradiente de φ en dicha interfaz.

• Regla 2. Coeficientes Positivos

Los coeficiente $a_{_J}$ que acompaña a cada variable en la discretización, por ejemplo $a_{_J}\varphi_{_J}$, son tales que

$$a_{_P}\varphi_{_P} = \sum_{J\in N} a_{_J}\varphi_{_J} + b \tag{1}$$

El coeficiente del nodo central tiene signo igual que los signos de los coeficientes de los nodos vecinos (N-Vecindad). Todos positivos. Esto garantiza que, si el valor en un nodo vecino se incrementa, también lo harán los valores en los demás nodos.

• Regla 3. Pendiente Negativa en el Término de Fuente

La linealización del término de fuente del tipo

$$S = S_o + S_{_P}\,(\varphi_{_P} - \varphi_o) = S_c + S_{_P}\,\varphi_{_P} \qquad\qquad S_c = S_o - S_{_P}\,\varphi_o \tag{2}$$

donde la pendiente $S_{_P}$ debe ser negativa o nula. Esto garantiza en parte que la solución sea estable, debida a este término.

• Regla 4. Suma de Los Coeficientes Vecinos

La suma de los coeficientes de los nodos vecinos $J \in \partial P$ suman igual que el coeficiente del nodo central P

$$a_{_P} = \sum_{J\in\partial P} a_{_J} \tag{3}$$

Excluyendo el término de fuente y el término transitorio. De esta forma, problemas con soluciones igualmente válidas más una constante, satisfacen también la ecuación diferencial.

3.1.2. Difusividad en la Interfaz

Consideremos dos nodos vecinos P y E tales que alrededor de cada uno domina una difusividad distintas. Sean dichas difusividades $\Gamma_{_P}$ y $\Gamma_{_E}$. Surge la duda de cuales de ambas aplica para la interfaz e intermedia mostrada en la figura 1 ($\mathbf{J} = -\Gamma\,\boldsymbol{\nabla}\varphi$).

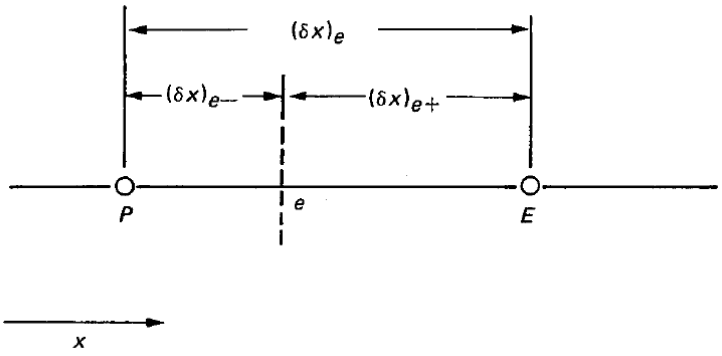


Figura 1. Distancias asociadas con la difusión en la interfaz e.

Un análisis sencillo del flujo J_e, por difusión lineal de la variable φ en la interfaz e, nos da que

$$J_e = -\Gamma_e \frac{\varphi_E - \varphi_P}{\delta x_e} = -\frac{\varphi_E - \varphi_P}{\delta x_e^-/\Gamma_P + \delta x_e^+/\Gamma_E} \qquad \Gamma_P \frac{\varphi_e - \varphi_P}{\delta x_e^-} = \Gamma_E \frac{\varphi_E - \varphi_e}{\delta x_e^+} = -J_e \tag{4}$$

donde el último miembro de (4.a) se ha obtenido de eliminar φ_e del balance del flujo a uno y otro lado de la interfaz en (4.b) (Intensidad del flujo J = diferencial del potencial difusivo $-\Delta\varphi$ entre la suma de las resistencias difusivas $\delta x/\Gamma$ en serie). Esto se reduce a tener una difusividad intermedia equivalente igual a

$$\Gamma_e = \left(\frac{1-f_e}{\Gamma_P} + \frac{f_e}{\Gamma_E}\right)^{-1} \qquad f_e = \frac{\delta x_e^+}{\delta x_e} \tag{5}$$

siendo f_e la fracción de la distancia que separa la interfaz e del nodo E. Se observa claramente que esto difiere de una simple interpolación lineal de la difusividad Γ, como lo hubiese sugerido el sentido común (ver por ejemplo Apéndice B en [Versteeg & Malalasekera,1995]), y tiene un fundamento físico mayormente justificable [Patankar,1980]. Cuando $f_e = 0.5$, es decir con la interfaz justamente en la mitad entre los nodos, la relación (5) se convierte en la "media armónica", más que en la media aritmética obtenida mediante una interpolación lineal. La expresión (5) será usada en aquellas interfases ubicadas entre dos secciones del dominio con difusividades diferente ($\Gamma \to \infty \Longrightarrow -\Delta\varphi \to 0$).

3.1.3. Discretización de Dos Puntos

La *discretización de dos puntos* de basa en la solución de la ecuación

$$\frac{d}{dx}(\rho\, u\, \varphi) = \frac{d}{dx}\left(\Gamma \frac{d\varphi}{dx}\right) \qquad \frac{dJ}{dx} = 0 \qquad J = \rho\, u\, \varphi - \Gamma \frac{d\varphi}{dx} \tag{6}$$

$$\begin{aligned} x &= 0 \qquad & \varphi &= \varphi_L \\ x &= \delta x \qquad & \varphi &= \varphi_R \end{aligned} \tag{7}$$

donde J es el flujo neto convección + difusión = constante. La solución de (6) con las condiciones (7) es

$$\frac{\varphi(x) - \varphi_L}{\varphi_R - \varphi_L} = \frac{\exp(I\!P x/\delta x) - 1}{\exp(I\!P) - 1} \qquad I\!P = \frac{\rho\, u\, \delta x}{\Gamma} \tag{8}$$

con $I\!P$ siendo el *número de Peclet*. La velocidad u y el gradiente $d\varphi/dx$ son perpendiculares a la interfaz.

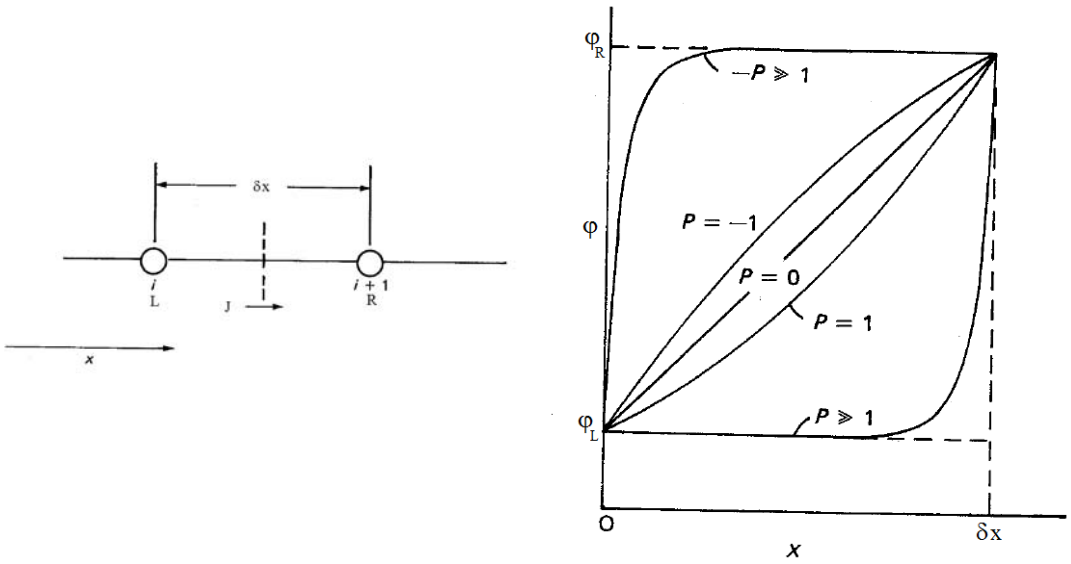


Figura 2. Flujo total entre dos puntos del mallado i (L-Left) e $i+1$ (R-Right).
Derecha. Solución exacta para el problema de convección-difusión uni-dimensional $x \in [0, \delta x]$.

La figura 2 muestra el dominio de integración identificando el nodo L con el nodo i y el nodo R con el nodo $i+1$. La interfaz, donde quiera que esté, deja pasar un flujo J constante. Del lado derecho de la figura se observa como es el perfil de φ en la solución exacta para el problema de difusión-convección uni-dimensional, donde el número de Peclet $I\!P$, adimensional, es un modulador de la solución (a veces denominado "Peclet de malla").

El flujo J constante adimesionalizado es J^*

$$J^* = \frac{J\,\delta x}{\Gamma} = I\!P\,\varphi - \frac{d\,\varphi}{d(x/\delta x)} \tag{9}$$

El valor de φ en la interfaz i entre los nodos L y R es una promedio ponderado de $\varphi_{_L}$ y $\varphi_{_R}$, mientras que el gradiente es un múltiplo de $(\varphi_{_R} - \varphi_{_L})$. Así, se propone la expresión

$$J^* = I\!P\,[\,\alpha\,\varphi_{_L} + (1-\alpha)\,\varphi_{_R}\,] - \beta\,(\varphi_{_R} - \varphi_{_L}) \tag{10}$$

donde α y β son multiplicadores adimensionales que dependen de $I\!P$. De manera que, J^* también puede ser expresado como

$$J^* = B(I\!P)\,\varphi_{_L} - A(I\!P)\,\varphi_{_R} \qquad\qquad J = \frac{\Gamma}{\delta x}\big[\,B(I\!P)\,\varphi_{_L} - A(I\!P)\,\varphi_{_R}\,\big] \tag{11}$$

La subtitución de (8), para una interfaz intermedia en un x cualquiera, es finalmente independiente de x, puesto que J^* es constante. Esto da que α y β en (10) se asocien de la forma

$$A(I\!P) = \beta - I\!P\,(1-\alpha) = \frac{I\!P}{\exp(I\!P) - 1} \qquad\qquad B(I\!P) = \beta + I\!P\,\alpha = \frac{I\!P\,\exp(I\!P)}{\exp(I\!P) - 1} \tag{12}$$

mostrados en la figura 3.

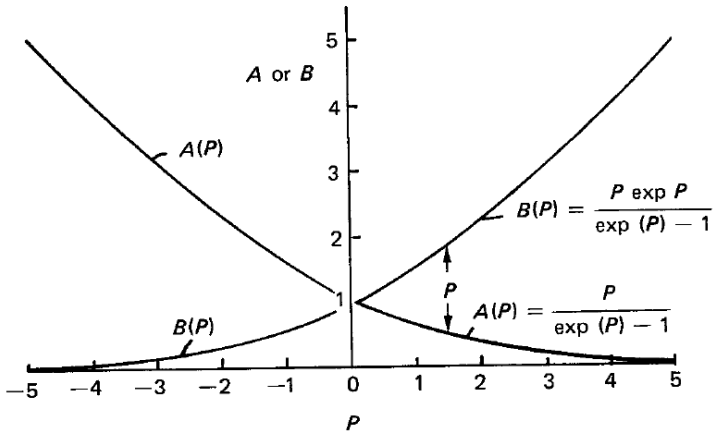


Figura 3. Variación de A y B con el número de Peclet $I\!P$.

Los coeficientes A y B tienen ciertas propiedades que es menester mostrar. Primero, en el caso donde φ_L y φ_L son iguales, el flujo por difusión es cero y J^* es simplemente el flujo por convección $J^* = I\!P\,\varphi_L = I\!P\,\varphi_R$ (u se asume constante). Bajo estas condiciones y comparando con (11.a) da que

$$B(I\!P) = A(I\!P) + I\!P \tag{13}$$

Propiedad también mostrada en la figura 3, donde la diferencia entre las curvas es justamente $I\!P$. El mismo resultado se obtiene colocando A y B en función de α y β.

La segunda propiedad de A y B tiene que ver con su simetría. Si cambiamos el sistema de coordenadas y lo revertimos, entonces $I\!P$ debería aparecer como $-I\!P$, y A y B intercambian sus roles. Así $A(I\!P)$ y $B(I\!P)$ se relacionan mediante

$$A(-I\!P) = B(I\!P) \qquad\qquad B(-I\!P) = A(I\!P) \tag{14}$$

Propiedad iguamente mostrada en la figura 3, con la simetris de las curvas respecto al eje central vertical.

Estas propiedades producen que finalmente A y B pueden expresarse únicamente en función de $A(|I\!P|)$ de la forma

$$A(I\!P) = A(|I\!P|) + [\![-I\!P, 0]\!] \qquad\qquad B(I\!P) = A(|I\!P|) + [\![I\!P, 0]\!] \tag{15}$$

donde el símbolo $[\![\,\cdot\,]\!]$ significa el mayor valor, que en este caso es comparando con 0.

Debido a que la forma (12) con el exponencial de $A(|I\!P|)$ es computacionalmente costosa desde el punto de vista de su cálculo, se ha ideado una forma aproximada más conveniente desde ese punto de vista, que se denomina "la ley de potencia" y se expresa como

$$A(|I\!P|) = [\![(1 - 0.1\,|I\!P|)^5, 0]\!] \tag{16}$$

la tabla siguiente muestra una comparación con este esquema y otros esquemas provenientes de diversos tipos de discretizaciones.

Tabla. La función $A(|I\!P|)$ para diferentes esquemas.

Esquema	Fórmula $A(\|I\!P\|)$
Diferencia Central	$(\,1 - 0.5\,\|I\!P\|\,)$
Aguas Arriba	1
Híbrido	$[\![\,(1 - 0.5\,\|I\!P\|), 0\,]\!]$
Ley de Potencia	$[\![\,(1 - 0.1\,\|I\!P\|)^5, 0\,]\!]$
Exponencial (exacta)	$\|I\!P\| \,/\, [\,\exp(\|I\!P\|) - 1\,]$

Es de hacer notar que la ecuación diferencial (9) se resolvió haciendo $I\!P\varphi(x) - \varphi'(x) = a + bx + cx^2$ ó $I\!P\varphi(x) - \varphi'(x) = \exp(a + bx + cx^2)$, de manera de hacerla depender de dos parámetros adicionales b y c, que permitiese a aplicación de trazadores rectilíneos o parabólicos en $x = 0$ y tener continuidad en la primera derivada de dos curvas definidas antes y después de dicho nodo central ($x = 0$). El resultado final fué que $\varphi'(0)$ dió independiente de b y c y $J^*(0) = a$ ó $J^*(0) = \exp(a)$ (siempre constante), lo que impidió aplicar este procedimiento.

3.2. DISCRETIZACION GENERAL

La discretización general de cualquier ecuación de transporte se hace mediante la aplicación de la discretización de dos puntos 3.1.(11.b) a todas las parejas de punto vecino - punto central que aparecen en cada configuración. La figura 4 muestra un ejemplo de coordenadas cilíndricas indicando la localización de los nodos y las interfaces. Los flujos J_i (convección + difusión) son perpendiculares a las interfaces i.

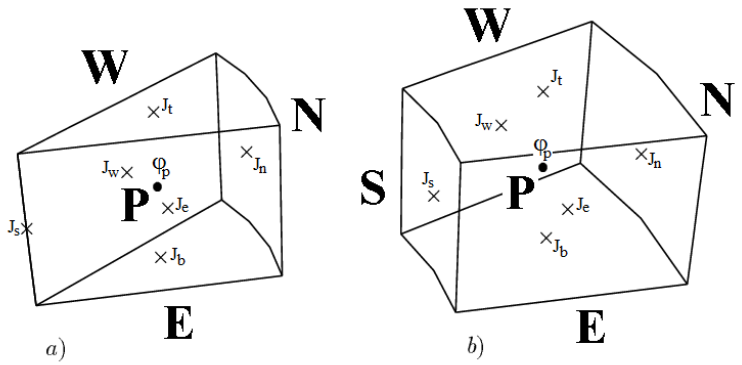


Figura 4. Volumen finito en el caso de coordenadas cilíndricas.

3.2.1. Ecuación Estacionaria

Sea la ecuación diferencial

$$\boldsymbol{\nabla}.(\,\rho\,\mathbf{u}\,\varphi\,) = \boldsymbol{\nabla}.(\,\Gamma\,\boldsymbol{\nabla}\varphi\,) + S \qquad\qquad \boldsymbol{\nabla}.\mathbf{J} = S \qquad\qquad \mathbf{J} = \rho\,\mathbf{u}\,\varphi - \Gamma\,\boldsymbol{\nabla}\varphi \tag{1}$$

Nótese la similitud de esta ecuación global y la ecuación unidimensional para dos puntos 3.1.(6).

La integración de la equación diferencial (1) para un volumen finito $\Delta\mathcal{V}_{_P} = \Delta x\,\Delta y\,\Delta z$, cuyo nodo central es el nodo P. Los nodos vecinos son designados con las letras mayúsculas N, S, E, W, T y B. Las interfaces que rodean al volumen infinito son $\delta\mathcal{A}_n$, $\delta\mathcal{A}_s$, $\delta\mathcal{A}_e$, $\delta\mathcal{A}_w$, $\delta\mathcal{A}_t$ y $\delta\mathcal{A}_b$. La aplicación del teorema de Gauss a la integral de (1.b) sobre el volumen de control finito $\Delta\mathcal{V}_{_P}$ da

$$J_n\,\delta\mathcal{A}_n - J_s\,\delta\mathcal{A}_s + J_e\,\delta\mathcal{A}_e - J_w\,\delta\mathcal{A}_w + J_t\,\delta\mathcal{A}_t - J_b\,\delta\mathcal{A}_b = S\,\Delta\mathcal{V}_{_P} \tag{2}$$

Los flujos J_n, J_s, J_e, J_w, J_t, J_b (positivos saliendo del volumen finito, negativos entrando), son perpendiculares a las respectivas caras (interfaces) del volumen finito.

Substituyendo la discretización para dos puntos 3.1.(11.b) en la ecuación anterior, se obtiene

$$a_{_P}\,\varphi_{_P} = a_{_N}\,\varphi_{_N} + a_{_S}\,\varphi_{_S} + a_{_E}\,\varphi_{_E} + a_{_W}\,\varphi_{_W} + a_{_T}\,\varphi_{_T} + a_{_B}\,\varphi_{_B} + b \tag{3}$$

con coeficientes

$$\begin{aligned} a_{_N} &= \frac{\Gamma_n\delta\mathcal{A}_n}{\delta y_n}\,A(I\!P_n) & a_{_E} &= \frac{\Gamma_e\delta\mathcal{A}_e}{\delta x_e}\,A(I\!P_e) & a_{_T} &= \frac{\Gamma_t\delta\mathcal{A}_t}{\delta z_t}\,A(I\!P_t) \\ a_{_S} &= \frac{\Gamma_s\delta\mathcal{A}_s}{\delta y_s}\,B(I\!P_s) & a_{_W} &= \frac{\Gamma_w\delta\mathcal{A}_w}{\delta x_w}\,B(I\!P_w) & a_{_B} &= \frac{\Gamma_b\delta\mathcal{A}_b}{\delta z_b}\,B(I\!P_b) \end{aligned} \tag{4.a–f}$$

$$a_P = a_N + a_S + a_E + a_W + a_T + a_B - S_P \Delta \mathcal{V}_P \qquad b = S_c \Delta \mathcal{V}_p \qquad (4.g, h)$$

donde la linealización 3.1.(2), $S = S_c + S_P\, \varphi_P$, del término de fuente se ha aplicado. En $(4.g)$ se ha aplicado la regla 4 (ec. 3.1.(3)). La equivalencia $a_P = \sum a_I$ en esta regla proviene de aplicar la ecuación de continuidad en los números de Peclet $\sum I\!P_i = 0$ (algebraicamente) cuando substituimos $B = A + I\!P$.

La figura 5 muestra un volumen de control en el borde y un volumen de control típico en el medio del dominio. El tamaño del volumen de control es Δx_P y los nodos vecinos W y E están ubicados a distancias δx_w y δx_e del nodo central P (central no necesariamente significa que está en el centro), respectivamente. En el volumen de control en el borde, el nodo central coincide con la frontera del dominio.

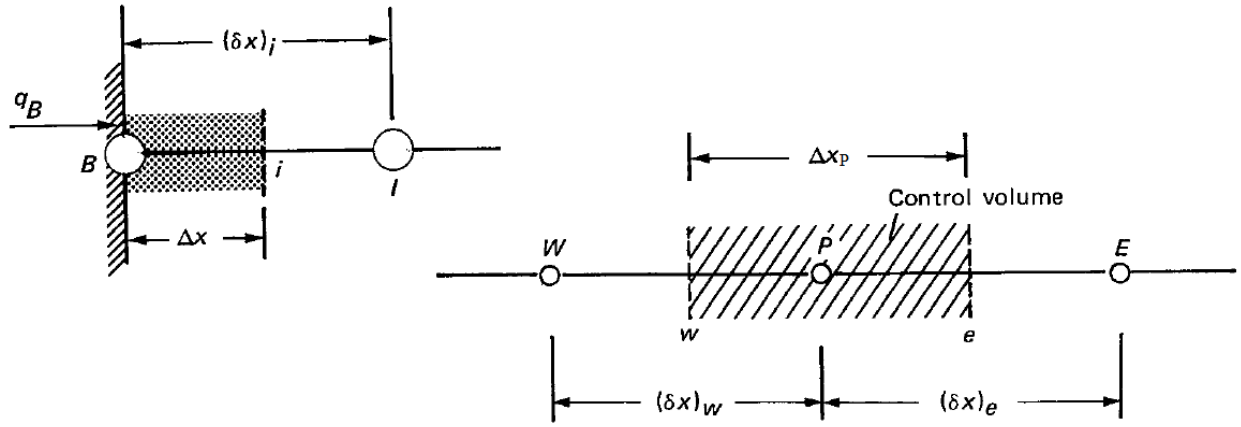


Figura 5. Volúmenes finitos mostrando un volumen de borde y otro típico en el medio del dominio.

Se puede escoger entre ubicar los nodos en la mitad de las interfaces o ubicar las interfaces en la mitad entre los nodos.

3.2.2. Ecuación Transitoria

Sea la ecuación diferencial

$$\frac{\partial \rho\varphi}{\partial t} + \boldsymbol{\nabla}.(\,\rho\,\mathbf{u}\,\varphi\,) = \boldsymbol{\nabla}.(\,\Gamma\,\boldsymbol{\nabla}\varphi\,) + S \qquad \frac{\partial \rho\varphi}{\partial t} + \boldsymbol{\nabla}.\mathbf{J} = S \qquad \mathbf{J} = \rho\,\mathbf{u}\,\varphi - \Gamma\,\boldsymbol{\nabla}\varphi \qquad (5)$$

El término transitorio se discretiza aplicando el método Euler implícito (sección 2.2.1, ec. 2.2.(5.a))

$$\frac{\partial \rho\varphi}{\partial t} = \frac{\rho_P \varphi_P - \rho_P^o \varphi_P^o}{\Delta t} + O(\Delta t) \qquad (6)$$

Con el procedimiento de la sección anterior y agregando la parte transitoria, se obtiene

$$\frac{(\rho_P \varphi_P - \rho_P^o \varphi_P^o)\,\Delta \mathcal{V}_P}{\Delta t} + J_n\,\delta\mathcal{A}_n - J_s\,\delta\mathcal{A}_s + J_e\,\delta\mathcal{A}_e - J_w\,\delta\mathcal{A}_w + J_t\,\delta\mathcal{A}_t - J_b\,\delta\mathcal{A}_b = (S_c + S_P\,\varphi_P)\,\Delta\mathcal{V}_P \qquad (7)$$

donde $\rho^o\varphi_P^o$ son las condiciones en el paso anterior. La ecuación (7) es un caso particular de la ecuación más general

$$\rho_P\,\Delta\mathcal{V}_P\,\frac{d\varphi_P}{dt} + J_n\,\delta\mathcal{A}_n - J_s\,\delta\mathcal{A}_s + J_e\,\delta\mathcal{A}_e - J_w\,\delta\mathcal{A}_w + J_t\,\delta\mathcal{A}_t - J_b\,\delta\mathcal{A}_b = S\,\Delta\mathcal{V}_P \qquad (8)$$

aplicando el método de Euler implícito (ρ_P= constante, $S = S_c + S_P\,\varphi_P$). Esta última aplicada a todos los nodos centrales P conforma un sistema de ecuaciones diferenciales ordinarias de primer orden, el cual se puede resolver con cualquier método Runge-Kutta.

Substituyendo la discretización para dos puntos 3.1.(11.b) en la ecuación anterior, se obtiene

$$a_{_P}\,\varphi_{_P} = a_{_N}\,\varphi_{_N} + a_{_S}\,\varphi_{_S} + a_{_E}\,\varphi_{_E} + a_{_W}\,\varphi_{_W} + a_{_T}\,\varphi_{_T} + a_{_B}\,\varphi_{_B} + b \tag{9}$$

con coeficientes

$$\begin{aligned} a_{_N} &= \frac{\Gamma_n \delta\mathcal{A}_n}{\delta y_n}\,A(\mathbb{P}_n) & a_{_E} &= \frac{\Gamma_e \delta\mathcal{A}_e}{\delta x_e}\,A(\mathbb{P}_e) & a_{_T} &= \frac{\Gamma_t \delta\mathcal{A}_t}{\delta z_t}\,A(\mathbb{P}_t) \\ a_{_S} &= \frac{\Gamma_s \delta\mathcal{A}_s}{\delta y_s}\,B(\mathbb{P}_s) & a_{_W} &= \frac{\Gamma_w \delta\mathcal{A}_w}{\delta x_w}\,B(\mathbb{P}_w) & a_{_B} &= \frac{\Gamma_b \delta\mathcal{A}_b}{\delta z_b}\,B(\mathbb{P}_b) \end{aligned} \tag{10.a–f}$$

$$a_{_P} = a_{_P}^o + a_{_N} + a_{_S} + a_{_E} + a_{_W} + a_{_T} + a_{_B} - S_{_P}\Delta\mathcal{V}_{_P} \qquad a_{_P}^o = \frac{\rho_{_P}^o\,\Delta\mathcal{V}_{_P}}{\Delta t} \qquad b = a_{_P}^o\varphi_{_P}^o + S_c\Delta\mathcal{V}_p \;\; (10.g,h,i)$$

Los coeficientes en (10.a–f) son exactamente los mismos que en (4.a–f). Los cambios han sido en las ecuaciones (10.g,h,i) para $a_{_P}$ y b.

Una forma más general de plantear (9) es

$$\rho_{_P}\,\Delta\mathcal{V}_{_P}\,\frac{d\varphi_{_P}}{dt} = -a_{_P}\,\varphi_{_P} + a_{_N}\,\varphi_{_N} + a_{_S}\,\varphi_{_S} + a_{_E}\,\varphi_{_E} + a_{_W}\,\varphi_{_W} + a_{_T}\,\varphi_{_T} + a_{_B}\,\varphi_{_B} + b \tag{11}$$

$$a_{_P} = a_{_N} + a_{_S} + a_{_E} + a_{_W} + a_{_T} + a_{_B} - S_{_P}\Delta\mathcal{V}_{_P} \qquad b = S_c\Delta\mathcal{V}_p \tag{12.a,b}$$

donde la ecuaciones (12.a,b) substituyen a las ecuaciones (10.g,h,i), formando como ya se dijo un sistema de ecuaciones diferenciales ordinarias de primer orden.

3.2.3. Método ADI

Cuando el problema es transitorio, se acostumbra a usar el esquema ADI (Alternate Direction Implicit), en el cual se resuelve el problema con Euler implícito en una sóla dirección y con Euler explícito en las direcciones restantes (ver sección 2.2.1). Lo que garatiza que el sistema siempre tiene una matriz tridiagonal. Estas direcciones se van alternando de manera secuencial en cada oportunidad (cada integración en t).

4. FLUJO GENERAL INCOMPRESIBLE VISCOSO

La metodología planteada para la ecuación de transporte φ, aplica también cuando $\phi = \mathbf{v}$, la ecuación de transporte de la cantidad de movimiento lineal. A esta ecuación en el caso de los fluidos newtoniano incompresibles se le denomina la ecuación de Navier-Stokes. La diferencia con los métodos para el transporte de ϕ antes planteados, es que el mallado para la velocidad tiene los nodos justo en el medio de las caras de los volúmenes finitos para ϕ, que es donde se necesita la información de la velocidad (mallas desplazadas).

En esta parte se ha hecho el replanteamiento de los problemas de flujo incompresible viscoso llevando la ecuación de Navier-Stokes a formularse como un sistema de ecuaciones diferenciales ordinarias de primer orden de dimensión infinita en el caso analítico y de dimensión finita en el caso discretizado. Las condiciones de frontera se ven reflejadas en la vecindad de la misma y las soluciones dentro del conjunto abierto del dominio están subordinadas a ellas. Las condiciones en la frontera no forma parte del sistema de ecuaciones diferenciales ordinarias, sino a través de las ecuaciones de los puntos vecinos. Modernamente se están usando métodos que se denominan de pasos fraccionados (e.g. [Kim & Moin,(1985)] y [Orlandi,2000]) que no son más que métodos Runge-Kutta de varias etapas. Con esta formulación se hace adecuado el planteamiento para usar cualquiera de estos métodos Runge-Kutta.

4.1. ECUACIONES FUNDAMENTALES

Las ecuaciones fundamentales para el estudio del flujo incompresible son la ecuación de conservación de masa ó continuidad

$$\boldsymbol{\nabla}.\mathbf{v} = 0 \tag{1}$$

y la ecuación de conservación de cantidad de movimiento lineal ó Navier-Stokes

$$\rho\left(\frac{\partial \mathbf{v}}{\partial t}+\mathbf{v}.\boldsymbol{\nabla}\mathbf{v}\right)=\rho\,\mathbf{g}-\boldsymbol{\nabla}P+\mu\,\nabla^2\mathbf{v} \tag{2}$$

Para eliminar la densidad de esta última expresión, se divide por ρ, resultando

$$\frac{\partial \mathbf{v}}{\partial t}=-\boldsymbol{\nabla}\tilde{P}-\mathbf{v}.\boldsymbol{\nabla}\mathbf{v}+\nu\,\nabla^2\mathbf{v} \qquad \frac{\partial \mathbf{v}}{\partial t}+\boldsymbol{\nabla}.\,\mathbf{J}=-\boldsymbol{\nabla}\tilde{P} \qquad \mathbf{J}=\mathbf{v}\mathbf{v}-\nu\boldsymbol{\nabla}\mathbf{v} \tag{3}$$

La ecuación de transporte transitoria para $\varphi=\mathbf{v}$ con densidad uno, difusividad ν (viscosidad cinemática) y fuente menos gradiente de presión. Las fuerzas másicas son conservativas, por lo que $\mathbf{g}=-\boldsymbol{\nabla}\varphi$ se genera de una función potencial φ (e.g. la fuerza de gravedad $\mathbf{g}=-g\,\mathbf{e}_z$ se genera a partir del potencial $\varphi=g\,z$). La cantidad $\tilde{P}=(P-P_o)/\rho+(\varphi-\varphi_o)$ es la presión equivalente o reducida. Los valores P_o y φ_o son dos valores de referencia arbitrarios que no alteran la ecuación original (3). Finalmente, tomando la divergencia de la ecuación (3.a), se obtiene la ecuación de Poisson para la presión

$$\nabla^2\tilde{P}=-\boldsymbol{\nabla}\mathbf{v}:\boldsymbol{\nabla}\mathbf{v}=-\mathbf{G}:\mathbf{G} \qquad \mathbf{G}=[\boldsymbol{\nabla}\mathbf{v}]^{\mathrm{t}} \tag{4}$$

Se ha usado la identidad $\boldsymbol{\nabla}.(\mathbf{T}.\mathbf{a})=(\boldsymbol{\nabla}.\mathbf{T}).\mathbf{a}+\mathbf{T}:(\boldsymbol{\nabla}\mathbf{a})^{\mathrm{t}}$ y la conmutatividad de la divergencia y el gradiente. En esta última parte se ha supuesto que los operadores de la divergencia y el laplaciano conmutan, y de igual manera la divergencia conmuta con la derivación parcial con respecto al tiempo. Donde al conmutar, aparece la divergencia de $\mathbf{v}$, el término se anula. Para conmutar, las derivadas mixtas se han supuesto continuas en su dominio (Teorema de Clairaut).

4.2. APROXIMACIONES DISCRETAS DIRECTAS

Haciendo un abuso de la notación, se han designado los siguientes operadores como aproximaciones discretas de las operaciones diferenciales de los miembros de la derecha

$$\mathbb{G}(\tilde{P})\approx\boldsymbol{\nabla}\tilde{P} \qquad \mathbb{D}(\mathbf{v})\approx\boldsymbol{\nabla}.\mathbf{v} \qquad \mathbb{H}(\mathbf{v})\approx\mathbf{v}.\boldsymbol{\nabla}\mathbf{v}=\boldsymbol{\nabla}.(\mathbf{v}\mathbf{v}) \qquad \mathbb{L}(\mathbf{v})\approx\nabla^2\mathbf{v} \tag{1}$$

El operador discreto aplicado a un punto se calcula tomando en consideración los valores de los puntos vecinos, utlizando cualquiera de los métodos de discretización de ecuaciones diferenciales en derivadas parciales (diferencias finitas, volúmenes finitos, elementos finitos, etc.) y sus variantes. Con la definición de los operadores, la ecuación 4.1.(3) en derivadas parciales de funciones continuas se convierte en un sistema de ecuaciones diferenciales ordinarias de la forma

$$\frac{d\mathbf{v}}{dt}\approx-\mathbb{G}(\tilde{P})-\mathbb{H}(\mathbf{v})+\nu\,\mathbb{L}(\mathbf{v}) \qquad \mathbb{D}(\mathbf{v})=0 \tag{2}$$

El problema original que era un problema de valor en la frontera con condiciones iniciales, se convierte en un problema exclusivamente de valores iniciales.

Involucrando la ecuación 4.1.(4), el sistema de ecuaciones diferenciales (2) se puede reformular en el siguiente sistema

$$\begin{cases}\mathbf{F}(\mathbf{v})=-\mathbb{H}(\mathbf{v})+\nu\,\mathbb{L}(\mathbf{v})\\ \mathbb{L}(\tilde{P})=-\mathbb{D}[\mathbb{H}(\mathbf{v})]=-\mathbb{G}(\mathbf{v}):\mathbb{G}(\mathbf{v})\\ \dfrac{d\mathbf{v}}{dt}=\mathbf{f}(\mathbf{v})=\mathbf{F}(\mathbf{v})-\mathbb{G}(\tilde{P})\end{cases} \tag{3}$$

donde se ha tenido en cuenta que $\mathbb{D}[\mathbb{L}(\mathbf{v})]=0$. El operador diferencial discreto $\mathbb{G}(\,\cdot\,)$ se utiliza de manera indistinta para campos escalares y campos vectoriales, debido a que es lineal y no actúa sobre la base del espacio vectorial.

En cuanto a las condiciones de frontera para la velocidad, se tienen dos circunstancias. La primera, la condición de Dirichlet $\mathbf{v} = \mathbf{v}_w + \mathbf{v}_o$, donde se tiene que el fluido sobre una pared adquiere su velocidad $\mathbf{v}_w$, más la velocidad de transpiración $\mathbf{v}_o$, si la hubiese. La segunda, la condición de Neumann $\boldsymbol{\nabla}_n \mathbf{v} = \boldsymbol{\mathcal{T}}_w/\mu$, donde el gradiente de la velocidad en la dirección normal a la pared es conocida. En cualquiera de estas circunstancias, la condición de la frontera introducida en la ecuación de movimiento 4.1.(3), da como resultado la condición de la frontera de tipo Neumann $\boldsymbol{\nabla}_n \tilde{P} = -d\mathbf{v}_n/dt + \nu \nabla^2 \mathbf{v}_n$ para la presión, en caso que no se conozca la condición de tipo Dirichlet $\tilde{P} = \tilde{P}_w$, siendo $\mathbf{v}_n = (\mathbf{v}.\mathbf{n})\,\mathbf{n}$ y $\mathbf{n}$ la normal exterior al fluido en la frontera.

4.3. APROXIMACIONES DISCRETAS PROYECTADAS

A priori, conociendo el campo de velocidades, se puede obtener el campo de presiones resolviendo la ecuación de Poisson 4.1.(4). Sin embargo, para conocer el campo de velocidades, se requiere a priori conocer el campo de presiones. Este círculo vicioso se puede romper, si en lugar de usar la ecuación 4.2.(2), se elimina de la misma el gradiente de la presión, de manera que ahora la ecuación

$$\frac{d\hat{\mathbf{v}}}{dt} \approx -I\!H(\hat{\mathbf{v}}) + \nu\, I\!L(\hat{\mathbf{v}}) \qquad\qquad I\!D(\hat{\mathbf{v}}) \neq 0 \tag{1}$$

permite obtener un campo de velocidades, sin conocer a priori el campo de presiones. No obstante, dicho campo de velocidades ya no será solenoidal, como se indica en la segunda parte de (1). Consideremos que tanto el campo de velocidades solenoidal y el no solenoidal parten de las mismas condiciones iniciales y con condiciones de borde siempre siendo las mismas, tal como se indica a continuación

$$\begin{aligned} &\text{c.i.} \qquad \mathbf{v}_o = \hat{\mathbf{v}}_o = \mathbf{v}(t_o, \mathbf{x}) \qquad \boldsymbol{\nabla}.\mathbf{v}_o = 0 \quad \text{para} \quad t = t_o \quad \text{y} \quad \mathbf{x} \in \bar{\Omega} \\ &\text{c.b.} \qquad \mathbf{v} = \hat{\mathbf{v}} = \mathbf{h}(t, \mathbf{x}) \quad \text{para} \quad \mathbf{x} \in \partial\Omega \end{aligned} \tag{2}$$

Si ahora a la ecuación 4.2.(2) le restamos la ecuación (1), resulta la siguiente ecuación diferencial

$$\frac{d}{dt}(\mathbf{v} - \hat{\mathbf{v}}) \approx -\not{G}(\tilde{P}) - I\!H(\mathbf{v}) + I\!H(\hat{\mathbf{v}}) + \nu\, I\!L(\mathbf{v} - \hat{\mathbf{v}}) \tag{3}$$

Con el siguiente cambio de variables

$$\frac{d}{dt}(\mathbf{v} - \hat{\mathbf{v}}) = -\boldsymbol{\nabla}\phi \qquad\qquad \mathbf{v} - \hat{\mathbf{v}} = -\boldsymbol{\nabla}\Phi \qquad\qquad \frac{d\Phi}{dt} = \phi \tag{4}$$

formulado bajo el supuesto que las diferencias de velocidades se originan de una función potencial Φ, y asumiendo que, cerca del instante inicial, los términos no lineales son muy parecidos

$$I\!H(\mathbf{v}) - I\!H(\hat{\mathbf{v}}) \approx \mathbf{0} \tag{5}$$

entonces, aplicando la divergencia a (3) y (4), se obtiene que

$$\nabla^2 \phi \approx \frac{d}{dt}[I\!D(\hat{\mathbf{v}})] \qquad\qquad \tilde{P} \approx \phi - \nu\, I\!L(\Phi) \approx \phi - \nu\, \boldsymbol{\nabla}.\hat{\mathbf{v}} \tag{6}$$

Este planteamiento permite formular el siguiente sistema de ecuaciones diferenciales

$$\begin{cases} \dfrac{d\hat{\mathbf{v}}}{dt} = -I\!H(\hat{\mathbf{v}}) + \nu\, I\!L(\hat{\mathbf{v}}) \\ \nabla^2 \phi = \dfrac{d}{dt}[I\!D(\hat{\mathbf{v}})] \\ \dfrac{d\mathbf{v}}{dt} = \dfrac{d\hat{\mathbf{v}}}{dt} - \not{G}(\phi) \end{cases} \tag{7}$$

Geométricamente, el sistema anterior se puede interpretar como que el campo de velocidades $\mathbf{v}$, se pueden obtener a partir del campo de velocidades $\hat{\mathbf{v}}$, proyectándolo de tal forma que, el complemento ortogonal sea justamente el gradiente del campo escalar Φ. De una forma más estructurada, el sistema (7) se puede expresar como

$$\begin{cases} \mathbf{F}(\hat{\mathbf{v}}) = -I\!H(\hat{\mathbf{v}}) + \nu\, I\!L(\hat{\mathbf{v}}) \\ I\!L(\phi) = I\!D[\mathbf{F}(\hat{\mathbf{v}})] \\ \dfrac{d\hat{\mathbf{v}}}{dt} = \mathbf{F}(\hat{\mathbf{v}}) \\ \dfrac{d\mathbf{v}}{dt} = \mathbf{f}(\mathbf{v}) = \mathbf{F}(\hat{\mathbf{v}}) - \not{G}(\phi) \end{cases} \tag{8}$$

usando la función auxiliar $\mathbf{F}$. Aunque en la segunda ecuación se tiene que $I\!D[\mathbf{F}(\hat{\mathbf{v}})] = -I\!D[I\!H(\hat{\mathbf{v}})]$, se ha preferido dejarlo así, para poder aplicar adecuadamente el método Runge-Kutta.

4.4. METODO DE PASO FRACCIONADO

Para el sistema 4.3.(8) también se puede usar el método Runge-Kutta de la siguiente forma

$$\begin{array}{lll} \hat{\mathbf{v}}^{n+1} = \hat{\mathbf{v}}^n + c_r\,\Delta t\,\mathbf{K}_r & \mathbf{K}_r = \mathbf{F}(\hat{\mathbf{v}}^n + a_{rs}\,\Delta t\,\mathbf{K}_s) & I\!L(\phi_r) = I\!D(\mathbf{K}_r) \\ \mathbf{v}^{n+1} = \mathbf{v}^n + c_r\,\Delta t\,\mathbf{k}_r & \mathbf{k}_r = \mathbf{f}(\mathbf{v}^n + a_{rs}\,\Delta t\,\mathbf{k}_s) = \mathbf{K}_r - \not{G}(\phi_r) & \hat{\mathbf{v}}^n = \mathbf{v}^n \end{array} \tag{1}$$

donde para cada paso de integración en el tiempo se parte de un campo de velocidades solenoidal, que es el campo de velocidades actual $\mathbf{v}^n$ para dicho instante $t^n = t_o + n\,\Delta t$.

El método de paso fraccionado, a diferencia del método Runge-Kutta, se expresa mediante la siguente fórmulas algorítmicas [Kim & Moin,1985]

$$\begin{cases} \hat{\mathbf{v}}^{n+1} = \mathbf{v}^n + \Delta t\,[\, -\gamma_n\, I\!H(\mathbf{v}^n) - \zeta_n\, I\!H(\mathbf{v}^{n-1}) + 0.5\,\alpha_n\,\nu\, I\!L(\hat{\mathbf{v}}^{n+1} + \mathbf{v}^n)\,] \\ \gamma_n\, I\!L(\phi^{n+1}) + \zeta_n\, I\!L(\phi^n) = I\!D(\hat{\mathbf{v}}^{n+1})/\Delta t \qquad \alpha_n = \gamma_n + \zeta_n \\ \mathbf{v}^{n+1} = \hat{\mathbf{v}}^{s+1} - \Delta t\,[\,\gamma_n \not{G}(\phi^{n+1}) + \zeta_n \not{G}(\phi^n)\,] \end{cases} \tag{2}$$

El factor de 0.5 se debe a que se está usando un esquema del tipo Crank-Nicolson para la parte implícita en las derivaciones de segundo orden en el operador $\mathbf{L}$.

Los valores de los coeficientes, que con cierta frecuencia se usan, son:

$$\begin{array}{lll} \gamma_1 = \dfrac{8}{15} & \zeta_1 = 0 & \alpha_1 = \dfrac{8}{15} \\ \gamma_2 = \dfrac{5}{12} & \zeta_2 = -\dfrac{17}{60} & \alpha_2 = \dfrac{2}{15} \\ \gamma_3 = \dfrac{3}{4} & \zeta_3 = -\dfrac{5}{12} & \alpha_3 = \dfrac{1}{3} \end{array} \tag{3}$$

Tratando de hacer una analogía con el método Runge-Kutta de tercer orden, en la notación de Butcher, los coeficientes del método de paso fraccionado se pueden expresar, para el operador $I\!H$ como

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \gamma_1 & \gamma_1 & 0 & 0 & 0 \\ \gamma_1+\gamma_2+\zeta_2 & \gamma_1+\zeta_2 & \gamma_2 & 0 & 0 \\ \gamma_1+\gamma_2+\gamma_3+\zeta_2+\zeta_3 & \gamma_1+\zeta_2 & \gamma_2+\zeta_3 & \gamma_3 & 0 \\ \hline & 0 & 0 & 0 & 1 \end{array} \tag{4.a}$$

y para el operador $I\!L$ como

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \alpha_1 & 0.5\,\alpha_1 & 0.5\,\alpha_1 & 0 & 0 \\ \alpha_1+\alpha_2 & 0.5\,\alpha_1 & 0.5\,(\alpha_1+\alpha_2) & 0.5\,\alpha_2 & 0 \\ \alpha_1+\alpha_2+\alpha_3 & 0.5\,\alpha_1 & 0.5\,(\alpha_1+\alpha_2) & 0.5\,(\alpha_2+\alpha_3) & 0.5\,\alpha_3 \\ \hline & 0 & 0 & 0 & 1 \end{array} \qquad (4.b)$$

Teniendo en cuenta la relación $\alpha_s = \gamma_s + \zeta_s$, con $\zeta_1 = 0$, se puede observar que los puntos de colocación de ambas matrices de Butcher son los mismos. No obstante, el esquema es explícito para el operador no lineal $I\!H$ y semi-implícito para el operador lineal $I\!L$. Sacadas las cuentas con los valores particulares antes mencionados en (2), las dos matrices $(4.a, b)$ quedan como

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 8/15 & 8/15 & 0 & 0 & 0 \\ 2/3 & 1/4 & 5/12 & 0 & 0 \\ 1 & 1/4 & 0 & 3/4 & 0 \\ \hline & 0 & 0 & 0 & 1 \end{array} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 8/15 & 4/15 & 4/15 & 0 & 0 \\ 2/3 & 4/15 & 1/3 & 1/15 & 0 \\ 1 & 4/15 & 1/3 & 7/30 & 1/6 \\ \hline & 0 & 0 & 0 & 1 \end{array} \qquad (5)$$

Dos aspectos diferencian al método de paso fraccionado con el método Runge-Kutta. Primero, que en el método de paso fraccionado se usa en donde sea posible el campo de velocidades solenoidal $\mathbf{v}$, en lugar de $\hat{\mathbf{v}}$ para la evaluación de la función $\mathbf{F}(\hat{\mathbf{v}}) = -I\!H(\hat{\mathbf{v}}) + \nu\, I\!L(\hat{\mathbf{v}})$. Esto hace que en el método de paso fraccionado, el campo de velocidades obtenido en cada paso $\hat{\mathbf{v}}^{s+1}$ esté más cerca del campo solenoidal, y por lo tanto, haga que el campo escalar $\phi = \gamma_s \phi^{s+1} + \zeta_s \phi^s$ también esté más cerca del campo de presiones $\tilde{P}$ (valor de $I\!L(\Phi) \approx \boldsymbol{\nabla}.\hat{\mathbf{v}}$ pequeño). Segundo, en el método de paso fraccionado el campo escalar ϕ se descompone en la forma antes mencionada, para obtener valores de ϕ^{s+1} más pequeños, y así reducir los errores $\Delta t \not{G}(\phi)$ de la velocidad (en realidad, lo que se reduce es el error parcial por cada componente de ϕ en cada paso).

Las raíces características de los métodos Runge-Kutta (5) se encuentran de igual forma que antes, resolviendo el sistema de ecuaciones lineales IV.3.4.(3), con lo cual se obtienen

$$\tilde{\Gamma}(z) = \left[\, 1 + z + z^2 + \frac{1}{2}z^3 + \frac{1}{6}z^4 \,\right] \qquad (6.a)$$

$$\Gamma(z) = \left[\, \frac{1 - \frac{23}{30}z + \frac{191}{150}z^2 - \frac{586}{3375}z^3 - \frac{56}{2025}z^4 + \frac{128}{50625}z^5}{1 - \frac{23}{30}z + \frac{93}{450}z^2 - \frac{76}{3375}z^3 + \frac{8}{10125}z^4} \,\right] \qquad (6.b)$$

respectivamente para el operador $I\!H$ y el operador $I\!L$. Luego la estabilidad de los diferentes métodos se establece imponiendo que las raíces características sean menores que la unidad. Esto da los siguientes límites para el avance del tiempo Δt

$$\tilde{\Delta} t \leq \frac{1.596}{|\lambda|} \qquad \Delta t \leq \frac{7.243}{|\lambda|} \qquad (7)$$

respectivamente para los dos operadores. Como se tiene que el límite CFL$\geq U_m\, \Delta t/\Delta x$, si consideramos que el autovalor $|\lambda| = U_m/\Delta x$, entonces los valores en los numeradores de (7) son los valores CFL máximos necesarios para que las diferentes partes del método de pasos escalonados sea estable (El método de Euler explícito, con $b_1 = 1$, $a_{11} = 0$ y $c_1 = 1$ y $\Gamma(z) = 1 + z$ requiere de un CFL=1). Esto permite relajar un poco el procedimiento (2) de integración en el tiempo con valores de Δt más grandes, de manera de obtener un avance más rápido en el algoritmo numérico. En la tesis [Granados,2003] se ha recomendado y utilizado el valor CFL=1.7, levemente superior al menor de los valores de CFL en (7), contando que la parte implícita del método mejore en cierta medida a su parte explícita.

Cuando el método Runge-Kutta se hace muy pesado y se desea un avance más rápido en el tiempo, se puede usar un método de paso múltiple del tipo Adams-Bashforth de segundo orden (semi-implícito). Para este método los valores de los coeficientes son:

$$\gamma_1 = \frac{3}{2} \qquad \zeta_1 = -\frac{1}{2} \qquad \alpha_1 = 1 \tag{8}$$

respectivamente para los operadores $\mathit{I\!H}$ y $\mathit{I\!L}$. Particularmente en este método, por ser tan sólo de dos pasos, no se hace la descomposición de ϕ. Con los coeficientes (8) se obtienen las siguientes matrices de Butcher

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 3/2 & 3/2 & 0 \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 0 & 1 \end{array} \tag{9}$$

y las siguientes raíces características

$$\tilde{\Gamma}(z) = \left[\, 1 + z + \frac{3}{2} z^2 \,\right] \qquad \Gamma(z) = \left[\, \frac{1 + \frac{1}{2} z + \frac{1}{2} z^2}{1 - \frac{1}{2} z} \,\right] \tag{10}$$

La estabilidad de estos métodos queda establecida con los dos límites siguientes

$$\tilde{\Delta} t \leq \frac{0.666}{|\lambda|} \qquad \Delta t \leq \frac{2}{|\lambda|} \tag{11}$$

Como se podrá observar, la estabilidad del método para la parte explícita es peor que el método de Euler, no obstante, la estabilidad se mejora notablemente con la parte implícita.

Es conveniente expresar la primera ecuación del método de paso fraccionado como

$$[\, \mathit{I\!I} - 0.5\, \alpha_s\, \Delta t\, \nu\, \mathit{I\!L} \,] \, (\hat{\mathbf{v}}^{s+1} - \mathbf{v}^s) = \Delta t \, [\, -\gamma_s\, \mathit{I\!H}(\mathbf{v}^s) - \zeta_s\, \mathit{I\!H}(\mathbf{v}^{s-1}) + \alpha_s\, \nu\, \mathit{I\!L}(\mathbf{v}^s) \,] \tag{12}$$

($\mathit{I\!I}$ es el operador identidad) debido a que el operador diferencial se puede ahora factorizar de la siguiente forma aproximada

$$[\, \mathbf{I} - 0.5\, \alpha_s\, \Delta t\, \mathbf{L} \,] \approx [\, (\mathbf{I} - \mathbf{L}_1)\, (\mathbf{I} - \mathbf{L}_2)\, (\mathbf{I} - \mathbf{L}_3) \,] \qquad \begin{array}{l} \mathbf{L}_i \approx 0.5\, \alpha_s\, \Delta t\, (\nu\, \nabla_i^2) \\ \displaystyle\sum_i \mathbf{L}_i = 0.5\, \alpha_s\, \Delta t\, \mathbf{L} \end{array} \tag{13}$$

(siendo $i = 1, 2, 3$ tres direcciones ortogonales) lo que permite que las matrices a resolver sean tridiagonales, en lugar de grandes matrices de banda dispersa. Esto resulta en una significante reducción del costo de cómputo y de memoria. Finalmente la ecuación (12) del método de paso fraccionado queda en la forma

$$[\, (\mathbf{I} - \mathbf{L}_1)\, (\mathbf{I} - \mathbf{L}_2)\, (\mathbf{I} - \mathbf{L}_3) \,] \, (\hat{\mathbf{v}}^{s+1} - \mathbf{v}^s) = \Delta t \, [\, -\gamma_s\, \mathbf{H}(\mathbf{v}^s) - \zeta_s\, \mathbf{H}(\mathbf{v}^{s-1}) + \alpha_s\, \mathbf{L}(\mathbf{v}^s) \,] \tag{12'}$$

($\mathbf{H}(\mathbf{v}) = \mathbf{v}.\boldsymbol{\nabla}\mathbf{v}$) que se aplica en direcciones alternadas (ADI - Alternating Direction Implicit) para hacer más eficiente el algoritmo.

4.5. METODO DE LOS PUNTOS DE COLOCACION

La ecuación de continuidad y de Navier-Stokes para flujo incompresible

$$\boldsymbol{\nabla}.\mathbf{v} = 0 \qquad \rho \left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}.\boldsymbol{\nabla}\mathbf{v} \right) = \rho\, \mathbf{g} - \boldsymbol{\nabla} P + \mu\, \nabla^2 \mathbf{v} \tag{1}$$

se pueden resolver más eficientemente utilizando un mallado no ortogonal y no estructurado formados por celdas con forma poliédrica con todos los valores estipulados en un nodo central para cada celda y teniendo cada una de estas una frontera superficial cerrada formada una retícula de polígonos planos, cada uno de los cuales es a su vez también una parte de la frontera de otra celda contigua, y así llenar todo el dominio con la totalidad de las celdas sin dejar algún vacío.

4.5.1. Fundamentos

Para realizar la formulación de este método es conveniente expresar las ecuaciones anteriores de una forma conservativa similar a 4.1.(3), resultando

$$\rho\frac{\partial \mathbf{v}}{\partial t} + \boldsymbol{\nabla}.\,\mathbf{J} = -\boldsymbol{\nabla}\tilde{P} \qquad\qquad \mathbf{J} = \rho\mathbf{v}\mathbf{v} - \mu\boldsymbol{\nabla}\mathbf{v} \tag{2}$$

con una presión reducida $\tilde{P} = P + \rho\phi$ que incluye la fuerza $\mathbf{g} = -\boldsymbol{\nabla}\phi$ proveniente de un potencial $\phi = gh$.

Integrando esta ecuación conservativa para una celda genérica de centro el nodo "P", con volumen $\Delta\mathcal{V}_{_P}$ y rodeado de nodos vecinos designados en su totalidad como $I \in$ NB y fronteras vecinas correspondientes designadas con minúsculas como $i \in$ nb, se obtiene

$$\rho_{_P}\,\Delta\mathcal{V}_{_P}\,\frac{d\mathbf{v}_{_P}}{dt} + \sum_{i\in\mathrm{nb}} J_i\,\delta\mathcal{A}_i = S_{_P}\,\Delta\mathcal{V}_{_P} \qquad\qquad \partial\Delta\mathcal{V}_{_P} = \bigcup_{i\in\mathrm{nb}}\mathcal{A}_i \tag{3}$$

con $S_{_P}$ como $S = -\boldsymbol{\nabla}\tilde{P}$ en el nodo "P".

Aplicando la discretización de los dos puntos explicado en la sección 3.1.3. resulta en las siguientes formulaciones

$$\rho_{_P}\,\Delta\mathcal{V}_{_P}\,\frac{d\mathbf{v}_{_P}}{dt} = -a_{_P}\,\mathbf{v}_{_P} + \sum_{I\in\mathrm{NB}} a_{_I}\,\mathbf{v}_{_I} + b_{_P} \tag{4.a}$$

$$a_{_P} = \sum_{I\in\mathrm{NB}} a_{_I} \qquad\qquad b_{_P} = S_{_P}\Delta\mathcal{V}_p \qquad\qquad a_{_I} = \frac{\mu_i\,\delta\mathcal{A}_i}{\delta r_i}\,A(I\!P_i) \tag{4.b}$$

Siguen siendo válidas las observaciones hechas después de la ecuación 3.2.$(4.g,h)$.

4.5.2. Interpolación de Rhie-Chow

El mallado desplazado tipicamente usado en la resolución de la ecuación de cantidad de movimiento asegura la estabilidad del procedimiento numérico. Este mismo objetivo puede lograrse con el método de los puntos de colocación usando un único mallado con los mismos nodos para las velocidades y las presiones (y sus gradientes). Para ellos es necesario calcular la velocidad normal v_i en las interfases de las celdas y que interviene en la formulación a través del número de Peclet de mallas $I\!P_i$ en los coeficientes $a_{_I}$.

La forma propuesta por Rhie & Chow en (1983) fué realizar la interpolación lineal sobre el campo de velocidades no solenoidal y luego corregir para hacerlo solenoidal [Moukalled, Mangani & Darwish,2016]. Esto formulado en ecuaciones se reduce a hacer

$$\mathbf{v}_i = \overline{\mathbf{v}}_i - (\boldsymbol{\nabla}\tilde{P}_i - \overline{\boldsymbol{\nabla}\tilde{P}_i}) \tag{5}$$

donde al promedio lineal $\overline{\mathbf{v}_i}$

$$\overline{\mathbf{v}_i} = (1-f_i)\,\mathbf{v}_{_P} + f_i\,\mathbf{v}_{_I} \qquad\qquad f_i = \frac{\delta r_i^{+}}{\delta r_i} \tag{6}$$

se le suma el gradiente de presión $\overline{\boldsymbol{\nabla}\tilde{P}_i}$ promediado de igual forma para hacer el campo de velocidades no solenoidal (ver sección 4.3). Luego el resultado se corrige restándole el gradiente de presión $\boldsymbol{\nabla}\tilde{P}_i$ para hacer

de nuevo el campo de velocidades solenoidal. Esta última cantidad se calcula como la diferencia finita entre los puntos nodales adyacentes a la interfase. Esto es

$$\boldsymbol{\nabla}\tilde{P}_i - \overline{\boldsymbol{\nabla}\tilde{P}_i} = \left[\frac{\tilde{P}_I - \tilde{P}_P}{\delta r_i} - \overline{\boldsymbol{\nabla}\tilde{P}_i} \,.\, \mathbf{e}_i \right] \mathbf{e}_i \qquad\qquad \mathbf{e}_i = \frac{\mathbf{r}_I - \mathbf{r}_P}{\|\mathbf{r}_I - \mathbf{r}_P\|} \tag{7}$$

pudiéndose calcular finalmente el número Peclet de malla $I\!P_i$ como

$$I\!P_i = \frac{\rho\, v_i\, \delta r_i}{\mu} \qquad\qquad v_i = \mathbf{v}_i .\, \mathbf{n}_i \qquad\qquad \delta r_i = |\overline{\mathrm{PI}}| = \|\mathbf{r}_I - \mathbf{r}_P\| \tag{8}$$

donde $\mathbf{e}_i$ es el versor que une la dirección del punto "P" al punto "I", distanciados δr_i, y $\mathbf{n}_i$ es el vector unitario exterior a la celda "P" sobre la interfase "i". Los vectores unitarios $\mathbf{e}_i$ y $\mathbf{n}_i$ no necesariamente coinciden. La cantidad δr_i^+ es la porción exterior a la celda "P" de la cantidad δr_i. La velocidad v_i del flujo es la proyección normal al área $\delta\mathcal{A}_i$ de la velocidad $\mathbf{v}_i$ interpolada en la interfase "i".

Una vez realizada la interpolación, el gradiente de presión en el nodo central de la celda $\boldsymbol{\nabla}\tilde{P}_P$, se calcula promediando con el volumen de la forma

$$\Delta\mathcal{V}_P\, \boldsymbol{\nabla}\tilde{P}_P = \sum_{i\in \mathrm{nb}} \Delta\mathcal{V}_i\, \boldsymbol{\nabla}\tilde{P}_i \qquad\qquad \Delta\mathcal{V}_P = \bigcup_{i\in \mathrm{nb}} \Delta\mathcal{V}_i \tag{9}$$

Esto completa el procedimiento. El haber eliminado la influencia del campo de presiones en el campo de velocidades durante la interpolación produce el aspecto deseado de estabilidad que produciría el mallado desplazado en las velocidades.

5. METODOS VARIACIONALES

5.1. METODO DE LOS RESIDUOS PONDERADOS

Sea la ecuación diferencial en 1D para $\varphi(x)$

$$\frac{d}{dx}\left(\Gamma\, \frac{d\varphi}{dx}\right) + S(x) = 0 \qquad\qquad a \le x \le b \tag{1}$$

con valor en la frontera

$$\varphi(a) = \alpha \qquad\qquad \varphi(b) = \beta \tag{2}$$

El coeficiente Γ puede depender de x. Este problema es el mismo problema de difusión pura (sin convección) con fuente propuesto antes, con coeficiente de difusión $\Gamma(x)$ dependiente.

Denominamos el *residuo* a la función

$$R(x) = \begin{cases} \hat{\varphi}(a) - \alpha & \text{si } x = a \\ \dfrac{d}{dx}\Big(\Gamma\, \dfrac{d\hat{\varphi}}{dx}\Big) + S(x) & a < x < b \\ \hat{\varphi}(b) - \beta & \text{si } x = b \end{cases} \tag{3}$$

donde $\hat{\varphi} = \hat{\varphi}(x)$ es la solución aproximada.

El promedio del residuo es

$$\bar{R} = \frac{1}{b-a} \int_a^b R(x)\, dx \tag{4}$$

y el promedio ponderado sería

$$\bar{R}_w = \frac{1}{b-a}\int_a^b w(x)\,R(x)\;dx \tag{5}$$

donde $w(x)$ es la función de ponderación. Finalmente el *residuo ponderado* se define como

$$R_w = \int_a^b w(x)\,R(x)\;dx \tag{6}$$

El método de los residuos ponderados consiste en determinar "a priori" la estructura matemática de $\hat{\varphi}(x)$, que hace $R_w = 0$.

Lema (Lema fundamental del cálculo d variaciones). Sea $C^1[a,b]$ un conjunto de funciones $w(x)$ continuas y con derivadas continuas en el intervalo cerrado $[a,b]$, tales $w(a) = w(b) = 0$. Si $R(x)$ es continua en $[a,b]$ y se cumple que

$$\int_a^b w(x)\,R(x)\;dx = 0 \qquad\qquad \forall w \in C^1[a,b] \tag{7}$$

entonces $R(x) \equiv 0$ en $[a,b]$.

Demostración. Suponer que $R(x_o) > 0$ para un $x_o \in [a,b]$ en su entorno $x_o - \delta$ y $x_o + \delta$. Si existe un $w(x_o) > 0$ en $[x_o - \delta, x_o + \delta]$ y $w(x) = 0$ fuera de este entorno, entonces

$$\int_a^b w(x)\,R(x)\;dx = \int_{x_o-\delta}^{x_o+\delta} w(x)\,R(x)\;dx > 0 \tag{8}$$

y esto contradice el lema. △

Corolario. Sea el residuo ponderado R_w dado por (6). Si $R_w = 0$ para toda $w(x) \in C^1[a,b]$, entonces $\hat{\varphi}(x) = \varphi(x)$.

Encontrar $\varphi(x)$ que satisfaga (1), con las condiciones de contorno (2) (formulación diferencial), se convierte en un problema equivalente a encontrar $\hat{\varphi}(x)$ en (3), que substituida en (6) satisfaga $R_w = 0$ para toda $w(x) \in C^1[a,b]$ (formulación variacional).

5.2. METODO DE COLOCACION

La función impulso o delta de Dirac se define como

$$\delta(x) = \begin{cases} 0 & \text{si } x \neq 0 \\ \infty & \text{si } x = 0 \end{cases} \qquad\qquad \int_{-\infty}^{\infty} \delta(x)\;dx = 1 \tag{1}$$

Esta función es la derivada de la función escalón o de Heaviside definida como

$$h(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases} \qquad\qquad \delta(x) = \frac{dh}{dx} \tag{2}$$

Esta función se puede desplazar de la forma

$$\delta(x-a) = \begin{cases} 0 & \text{si } x \neq a \\ \infty & \text{si } x = a \end{cases} \qquad\qquad \int_{-\infty}^{\infty} \delta(x-a)\,f(x)\;dx = f(a) \tag{3}$$

La integral del lado derecho se deduce del teorema del valor medio

$$\int_{a-\varepsilon}^{a+\varepsilon} f(x)\,\delta(x-a)\;dx = f(\zeta)\int_{a-\varepsilon}^{a+\varepsilon} \delta(x-a)\;dx = f(\zeta) \qquad\qquad \zeta \in [a-\varepsilon, a+\varepsilon] \tag{4}$$

Tomando el límite, cuando $\varepsilon \longrightarrow 0$ entonces $\zeta \longrightarrow a$.

5.2.1. Colocación Determinística

Proponemos una solución aproximada

$$\hat{\varphi}(x) = \sum_{j=1}^{n} c_j \, \phi_j(x) \tag{5}$$

donde $\phi_j(x)$ son las funciones bases (especificadas "a priori") y c_j son los coeficientes indeterminados, cuyo cálculo será el objetivo del método. Cuando la función de ponderación de escoge como la función delta de Dirac, entonces

$$R_k = \int_a^b \delta(x - x_k)\, R(x)\, dx = R(x_k) \qquad\qquad w(x) = \delta(x - x_k) \tag{6}$$

Los valores x_k donde se conoce el residuo, se denominan puntos de colocación.

Imponiendo la condición de que el residuo sea nulo en cada uno de los puntos de colocación x_i, $i = 1, 2, \ldots, p$, tenemos

$$R_i = R(x_i) = 0 \qquad\qquad R(x_i) = \frac{d}{dx}\left(\Gamma \frac{d\varphi}{dx}\right)_{x_i} + S(x_i) = 0 \qquad\qquad i = 1, 2, \ldots, p \tag{7}$$

Substituyendo la solución aproximada (5), obtenemos

$$\sum_{j=1}^{n} c_j \, \frac{d}{dx}\left(\Gamma \frac{d\phi_j}{dx}\right)_{x_i} = -S(x_i) \qquad\qquad [\mathbf{A}].\mathbf{c} = \mathbf{b} \qquad\qquad A_{ij} = \frac{d}{dx}\left(\Gamma \frac{d\phi_j}{dx}\right)_{x_i} \qquad b_i = -S(x_i) \tag{8}$$

Se deben definir las $\phi_j(x)$ de tal forma que $\hat{\varphi}(x)$ satisfaga las condiciones de borde. En caso contrario, si se substituye $\hat{\varphi}(x)$ en las condiciones de borde, se agregan dos ecuaciones más al sistema de ecuaciones, que debe coincidir ($p = n$) con el número de incógnitas c_j, $j = 1, 2, \ldots, n$. En este caso, los bordes se convierten en puntos de colocación.

5.2.2. Colocación Sobre Especificada

En este caso el número de ls puntos de colocación p supera al número de coeficientes indeterminados n. Se define un error cuadrático global

$$E = \sum_{i=1}^{p} [R(x_i)]^2 \qquad\qquad R(x_i) = \sum_{j=1}^{n} c_j \, \frac{d}{dx}\left(\Gamma \frac{d\phi_j}{dx}\right)_{x_i} + S(x_i) \tag{9}$$

y en el valor de coeficientes c_k, donde este error se minimiza $E = E_{min}$, se cumplen las ecuaciones normales

$$\frac{\partial E}{\partial c_k} = 0 \qquad\qquad \frac{\partial E}{\partial c_k} = \sum_{i=1}^{p} 2\, R(x_i)\, \frac{\partial R(x_i)}{\partial c_k} = 0 \tag{10}$$

$$\sum_{i=1}^{p} \left\{ \sum_{j=1}^{n} c_j \left[\frac{d}{dx}\left(\Gamma \frac{d\phi_j}{dx}\right) \right]_{x_i} + S(x_i) \right\} \left[\frac{d}{dx}\left(\Gamma \frac{d\phi_k}{dx}\right) \right]_{x_i} = 0 \tag{11}$$

donde se ha eliminado el factor común 2. Intercambiando las sumatorias sobre p y sobre n se obtiene

$$\sum_{j=1}^{n} \left\{ \sum_{i=1}^{p} \left[\frac{d}{dx}\left(\Gamma \frac{d\phi_k}{dx}\right) \right]_{x_i} \left[\frac{d}{dx}\left(\Gamma \frac{d\phi_j}{dx}\right) \right]_{x_i} \right\} c_j = -\sum_{i=1}^{p} \left[\frac{d}{dx}\left(\Gamma \frac{d\phi_k}{dx}\right) \right]_{x_i} S(x_i) \tag{12}$$

o lo que es equivalente

$$\sum_{j=1}^{n} A_{kj} c_j = b_k \qquad A_{kj} = \sum_{i=1}^{p} Q_{ki} Q_{ij}^{\mathrm{t}} \qquad b_k = -\sum_{i=1}^{p} Q_{ki} S(x_i) \qquad Q_{ki} = \left[\frac{d}{dx}\left(\Gamma \frac{d\phi_k}{dx} \right) \right]_{x_i} \tag{13}$$

Todo se reduce a resolver un sistema de n ecuaciones lineales con las incógnitas a_j.

EJEMPLO:

Resolver la ecuación diferencial

$$\frac{d}{dx}\left(\Gamma \frac{d\varphi}{dx} \right) + S(x) = 0 \qquad S(x) = a\, x^2 \qquad \varphi(0) = \varphi(l) = 0$$

con tres puntos de colocación

$$x_1 = \frac{l}{4} \qquad x_2 = \frac{l}{2} \qquad x_3 = \frac{3l}{4} \qquad (p = 3)$$

y dos funciones bases

$$\phi_1 = \operatorname{sen} \frac{\pi x}{l} \qquad \phi_2 = \operatorname{sen} \frac{2\pi x}{l} \qquad (n = 2)$$

Los resultados son

$$[\mathbf{Q}] = \frac{\Gamma}{l^2} \begin{bmatrix} -6.979 & -9.870 & -6.979 \\ -39.478 & 0 & 39.478 \end{bmatrix}$$

$$[\mathbf{A}] = [\mathbf{Q}][\mathbf{Q}]^{\mathrm{t}} = \frac{\Gamma^2}{l^4} \begin{bmatrix} 194.83 & 0 \\ 0 & 3117.0 \end{bmatrix} \qquad \mathbf{b} = -[\mathbf{Q}]\mathbf{S} = \Gamma a \begin{bmatrix} 6.829 \\ -19.739 \end{bmatrix}$$

5.2.3. Colocación Ortogonal

Mediante el cambio de variable propuesto en III.2.2.(7), se puede cambiar el dominio del problema (1) y llevarlo de $x \in [a, b]$ a $Z \in [-1, 1]$. Haciendo esto el residuo ponderado se transforma en ($z = [2x - (a + b)]/(b - a)$, $x = [(b - a)z + (a + b)]/2$)

$$R_w = \int_{-1}^{1} w(z)\, R(z)\, dz \approx \sum_{i=0}^{n} \omega_i\, w(z_i)\, R(z_i) \approx 0 \tag{14}$$

Como $\omega_i\, w(z_i) \neq 0$, para que $R_w = 0$, se debe satisfacer que

$$R(z_i) = 0 \qquad i = 0, 1, 2, \ldots, n \tag{15}$$

por lo que se escogen estos puntos como los puntos de colocación determinística, siendo $z_{\hat{\imath}}$, $\hat{\imath} = i + 1 = 1, 2, \ldots, p$, las raíces del polinomio de Legendre $P_p(z)$ de grado $p = n + 1$, transformadas las variables x al intervalo [-1,1] de z. Esto es

$$\sum_{j=1}^{p} c_j \left[\frac{d}{dz}\left(\Gamma \frac{d\phi_j}{dz} \right) \right]_{z_{\hat{\imath}}} + S(z_{\hat{\imath}}) = 0 \tag{16}$$

donde las funciones bases se pueden escoger como los p polinomios de Legendre $\phi_j(z) = P_{j-1}(z)$, $j = 1, 2, \ldots, p$, si satisfacen las condiciones de borde. En este caso, el $n + 1$ del grado del polinomio de donde obtener la raíces, no tiene que ver con el número de coeficientes incógnitas c_j, $j = 1, 2, \ldots, p$. Los p puntos de colocación $x_{\hat{\imath}}$ se escogen interiores al intervalo $[a, b]$. Para los puntos extremos se utilizan las condiciones de borde.

La selección de $\hat{\varphi}(x)$ se puede hacer de la siguiente manera, si se toma en cuenta III.2.2.(4.b)

$$\hat{\varphi}(x) = \psi(x) + \sum_{j=1}^{p} c_j \, (x-a) \, (x-b)^j \tag{17.a}$$

o alternativamente

$$\hat{\varphi}(x) = \psi(x) + \sum_{j=1}^{p} c_j \, (x-a)^j \, (x-b) \tag{17.b}$$

donde

$$\psi(x) = \frac{\beta - \alpha}{b-a} \, (x-a) + \alpha \tag{18}$$

para que se satisfagan las condiciones de borde $\hat{\varphi}(a) = \psi(a) = \alpha$ y $\hat{\varphi}(b) = \psi(b) = \beta$. Si se conoce el comportamiento de la ecuación diferencial, se pueden escoger otras funciones bases $\phi_j(x)$ que no sean necesariamente polinómicas.

Como $\hat{\varphi}[x(z)]$ son polinomios de grado $p+1$, se puede expresar como combinación lineal de los polinomios de Legendre $P_k(z)$

$$\hat{\varphi}[x(z)] = \sum_{k=0}^{p+1} \alpha_k \, P_k(z) \qquad \alpha_k = \frac{\langle \hat{\varphi}, P_k \rangle}{\langle P_k, P_k \rangle} \qquad \langle f, g \rangle = \int_{-1}^{1} f(z) \, g(z) \, dz \tag{19}$$

donde se ha usado la ortogonalidad de los polinomios de Legendre III.2.2.(5) (sección 2.2.2).

5.3. METODO DE GALERKIN

En el método de *Galerkin* se escoge como función de ponderación $w(x)$ para los residuos 5.1.(6), las mismas funciones bases $\phi_j(x)$. Esto es, el residuo ponderado

$$R_\phi = \int_a^b \phi_k(x) \, R(x) \, dx = 0 \qquad \forall \phi_k \in C^0[a,b] \tag{1}$$

se establece para todas las funciones bases ϕ_k, $k = 1, 2, \ldots, n$.

Para el mismo problema planteado en 5.1.(1) − (2), substituyendo la solución aproximada 5.2.(5), se obtiene

$$\sum_{j=1}^{n} c_j \int_a^b \phi_k(x) \, \frac{d}{dx}\left(\Gamma \, \frac{d\phi_j}{dx} \right) dx = - \int_a^b \phi_k(x) \, S(x) \, dx \tag{2}$$

Lo que resulta en un sistema de ecuaciones lineales

$$\sum_{j=1}^{n} A_{kj} \, c_j = b_k \qquad A_{kj} = \int_a^b \phi_k(x) \, \frac{d}{dx}\left(\Gamma \, \frac{d\phi_j}{dx} \right) dx \qquad b_k = - \int_a^b \phi_k(x) \, S(x) \, dx \tag{3}$$

que permite el cálculo de los coeficientes c_j.

Cuando el término de fuente depende de la variable dependiente, igualmente se substituye en (3.c) $S[\hat{\varphi}(x)]$. Observación que también es válida para los métodos de colocación.

Si la integración en (3) se realiza de forma numérica, con cuadratura de Newton-Cotes (colocación regular, sección III.2.2.1.) o cuadratura de Gauss-Legendre (colocación ortogonal, sección III.2.2.2.), el método es igualmente válido.

La integral $(3.b)$ puede hacerse por partes, aplicando el teorema de Green, con lo cual

$$A_{kj} = \int_a^b \phi_k(x)\,\frac{d}{dx}\left(\Gamma\,\frac{d\phi_j}{dx}\right)\,dx = \Gamma\,\phi_k(x)\,\frac{d\phi_j}{dx}\,\Big|_a^b - \int_a^b \Gamma\,\frac{d\phi_k}{dx}\,\frac{d\phi_j}{dx}\,dx \tag{4}$$

lo que facilita más aún la resolución. La formulación (4) se denomina "formulación débil", en contraposición de la "formulación fuerte" (1), porque las restricciones sobre la continuidad de las funciones y sus derivadas son menores, como puede observarse ahora en $(1.b)$ (En 5.1.$(7.b)$ la restricción era $\forall w \in C^1[a,b]$).

5.4. METODO DE ELEMENTOS FINITOS

Es un procedimiento sistemático aplicando la formulación de Galerkin, pero en donde se ha usado unas funciones bases muy particulares. Estas funciones bases reciben el nombre de *funciones de forma* tales que

$$\phi_i(\mathbf{x}_j) = \delta_{ij} \qquad \mathbf{x} \in \Omega_m \qquad m = 1, 2, \ldots, M \tag{1}$$

Cada elemento Ω_m, del total de M elementos en el dominio Ω, tiene varios nodos $i = 1, 2, \ldots$. Para cada uno de estos nodos existen varias funciones de forma en los elementos vecinos que comparten dichos nodos, cada una con las mismas característica (1). Fuera de cada elemento donde la función de forma actúa, tiene valor nulo. Dentro de cada elemento tiene una dependencia, que en el caso más simple, es lineal. De manera que

$$\hat{\varphi}(\mathbf{x}) = \sum_{j=1}^{N} \varphi_j\,\phi_j(\mathbf{x}) \tag{2}$$

donde φ_j es el valor de la variable resuelta φ en el nodo j, de un total de N nodos. El número de nodos N y el número de elementos M no son lo mismo, porque un nodo es compartido por varios elementos.

En el caso unidimensional estos elementos m tienen forma de segmentos con nodos m_1 y m_2 en los extremos y en el caso bidimensional pueden tener forma de triángulos o rectángulos con nodos m_1, m_2, m_3 y hasta m_4 en los vértices, en el sentido anti-horario.

5.4.1. Unidimensional

En el caso unidimensional el dominio $\Omega = [a,b]$ se divide en M elementos que son segmentos que van del nodo $m_1 = i-1, i$ al nodo $m_2 = i, i+1$ para cada elemento $m = i+1 = 1, 2, \ldots, M$. Las funciones de forma tienen la siguiente expresión lineal

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{\Delta x_{i-1}} & \text{si } x_{i-1} \le x \le x_i \\ \dfrac{x_{i+1} - x}{\Delta x_i} & \text{si } x_i \le x \le x_{i+1} \\ 0 & \text{si } x \le x_{i-1} \text{ ó } x \ge x_{i+1} \end{cases} \qquad \phi_k(x) = \begin{cases} \dfrac{x - x_{m_1}}{\Delta x_m} & \text{si } k = m_2 \\ \dfrac{x_{m_2} - x}{\Delta x_m} & \text{si } k = m_1 \\ 0 & \text{si } k \ne m_1 \text{ y } k \ne m_2 \end{cases} \quad x \in [x_{m_1}, x_{m_2}] \tag{3}$$

donde $\Delta x_i = x_{i+1} - x_i$ es el tamaño del elemento $m = i+1$ y $\Delta x_m = x_{m_2} - x_{m_1}$ es el tamaño del mismo elemento m. Vemos que dentro de un mismo elemento existen parcialmente dos funciones de forma distintas para los nodos extremos de dicho elemento. De hecho se cruzan en cada elemento. La figura 1 muestra la gráfica de estas funciones de forma ϕ_i para todo el dominio $[a,b]$. Las alturas de todos los triángulos en dicha figura son la unidad.

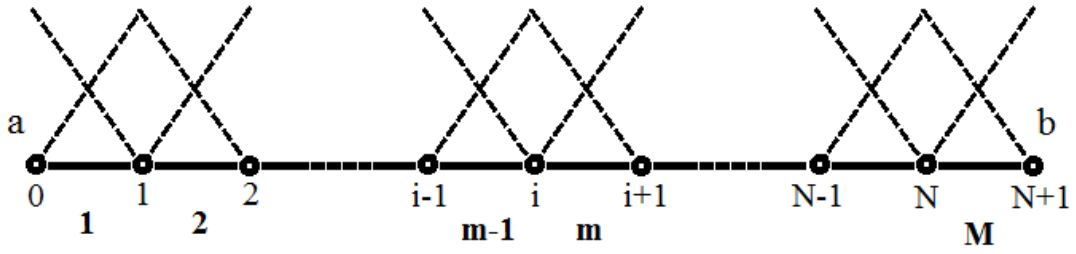


Figura 1. Funciones de forma $\phi_i(x)$ para elementos finitos unidimensionales.

La solución aproximada $\hat{\varphi}$ es

$$\hat{\varphi}(x) = \sum_{j=1}^{N} \varphi_j \, \phi_j(x) \qquad\qquad \phi_j(x_k) = \begin{cases} 1 & \text{si } j = k \\ 0 & \text{si } j \neq k \end{cases} \tag{4}$$

de donde $\varphi_j = \hat{\varphi}(x_j)$. Para el problema planteado 5.1.(1) – (2) el residuo ponderado de Galerkin es

$$R_\phi = \int_a^b \phi_k(x) \, R(x) \, dx = \int_a^b \phi_k(x) \left[\frac{d}{dx}\Big(\Gamma \frac{d\hat{\varphi}}{dx}\Big) + S(x) \right] dx = 0 \qquad\qquad \forall \phi_k \in C^0[a,b] \tag{5}$$

Substituyendo la solución aproximada (4), obtenemos

$$\sum_{j=1}^{N} \varphi_j \int_a^b \phi_k(x) \frac{d}{dx}\Big(\Gamma \frac{d\phi_j}{dx}\Big) \, dx + \int_a^b \phi_k(x) \, S(x) \, dx = 0 \qquad\qquad \sum_{j=1}^{N} A_{kj} \, \varphi_j = b_k \qquad k = 1, 2, \ldots, N \tag{6}$$

Aunque esta expresión es elegante a la hora de explicar el método de Galerkin, como se coloca a continuación

$$A_{kj} = \int_a^b \phi_k(x) \frac{d}{dx}\Big(\Gamma \frac{d\phi_j}{dx}\Big) \, dx = \Gamma \, \phi_k(x) \frac{d\phi_j}{dx} \Big|_a^b - \int_a^b \Gamma \frac{d\phi_k}{dx} \frac{d\phi_j}{dx} \, dx \qquad\qquad b_k = -\int_a^b \phi_k(x) \, S(x) \, dx \tag{7}$$

donde la integración sobre todo el dominio $[a, b]$ se ha hecho sumando los resultados por elementos m

$$A_{kj} = \Gamma \, \phi_k(x) \frac{d\phi_j}{dx} \Big|_a^b + \sum_{m=1}^{M} A_{kj}^m \qquad\qquad A_{kj}^m = -\int_{x_{m1}}^{x_{m2}} \Gamma \frac{d\phi_k}{dx} \frac{d\phi_j}{dx} \, dx = \begin{cases} = 0 & \text{si } k, j \neq m_1 \text{ y } k, j \neq m_2 \\ \neq 0 & \text{si } k, j = m_1 \text{ o } k, j = m_2 \end{cases} \tag{8}$$

es más conveniente para realizar los cálculos, mostrar el sistema de ecuaciones lineales con una matriz tridiagonal, donde los coeficientes no nulos acompañan sólo a la variable φ de los nodos vecinos al nodo central k para cada fila k

$$A_{k,k-1} \, \varphi_{k-1} + A_{k,k} \, \varphi_k + A_{k,k+1} \, \varphi_{k+1} = b_k \tag{9}$$

y se calculan los coeficientes individualmente de forma

$$A_{k,k-1} = \frac{\Gamma_{k-1}}{\Delta x_{k-1}} \qquad A_{k,k} = -\frac{\Gamma_{k-1}}{\Delta x_{k-1}} - \frac{\Gamma_k}{\Delta x_k} \qquad A_{k,k+1} = \frac{\Gamma_k}{\Delta x_k} \qquad b_k = -\frac{\Delta x_{k-1} + \Delta x_k}{2} \, \bar{S}(x_k) \tag{10}$$

Los coeficiente de difusividad Γ_k es el coeficiente promedio en cada elemento k. Los resultados anteriores se han obtenido fijando $\phi_k(x)$ y extendiendo el resultado no negativo de las integrales (7.a) y (8.b) a los nodos vecinos para $\phi_j(x)$, con $j = k-1, k, k+1$. El valor $\bar{S}(x_k)$ es el valor medio del término de fuente alrededor del nodo k, ponderado con las distancias relativas $\Delta x_{k-1}/(\Delta x_{k-1} + \Delta x_k)$ y $\Delta x_k/(\Delta x_{k-1} + \Delta x_k)$.

Para los nodos inicial y final se aplican las condiciones de borde como se muestra en (8.a) y se alteran los coeficientes b_1 y b_N

$$b'_1 = b_1 - A_{1,0}\,\varphi_0 + \Gamma\,\phi_1(x)\,\frac{d\phi_0}{dx}\,\Big|_a\,\varphi_0 \qquad\qquad b'_N = b_N - A_{N,N+1}\,\varphi_{N+1} - \Gamma\,\phi_N(x)\,\frac{d\phi_{N+1}}{dx}\,\Big|_b\,\varphi_{N+1} \tag{11}$$

aunque los últimos términos de las expresiones anteriores son nulos debido a que $\phi_1(a) = \phi_N(b) = 0$. Las condiciones de borde se establecen como $\varphi(a) = \varphi_0 = \alpha$ y $\varphi(b) = \varphi_{N+1} = \beta$.

5.4.2. Bidimensional

Se la siguiente ecuación diferencial en 2D para $\varphi(\mathbf{x})$

$$\boldsymbol{\nabla}.\left(\,[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\varphi\,\right) + S(\mathbf{x}) = 0 \qquad\qquad \mathbf{x} \in \Omega \subset \mathbb{R}^2 \tag{12}$$

con valor en la frontera

$$\varphi(\mathbf{x}) = \alpha \qquad \mathbf{x} = \mathbf{a} \in \partial_1\Omega \qquad\qquad \mathbf{n}.[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\varphi(\mathbf{x}) = \beta \qquad \mathbf{x} = \mathbf{b} \in \partial_2\Omega \tag{13}$$

donde la frontera $\partial\Omega$ de Ω se ha dividido en dos partes, siendo la primera $\partial_1\Omega$ con valor especificado (Dirichlet), y la segunda $\partial_2\Omega$ con gradiente perpendicular especificado (Neumann). El coeficiente de difusividad tensorial $[\boldsymbol{\Gamma}(\mathbf{x})]$ tiene componentes

$$[\boldsymbol{\Gamma}] = \begin{bmatrix} \Gamma_{xx} & \Gamma_{xy} \\ \Gamma_{yx} & \Gamma_{yy} \end{bmatrix} \tag{14}$$

y puede depender de la posición $\mathbf{x}$. El vector $\mathbf{n}$ es la normal unitaria exterior a Ω.

El residuo de la equación diferencial (12) con la solución aproximada $\hat{\varphi}(\mathbf{x})$ es

$$R(\mathbf{x}) = \begin{cases} \hat{\varphi}(\mathbf{x}) - \alpha & \text{si } \mathbf{x} = \mathbf{a} \in \partial_1\Omega \\ \boldsymbol{\nabla}.\left(\,[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\hat{\varphi}\,\right) + S(\mathbf{x}) & \text{si } \mathbf{x} \in \overset{\circ}{\Omega} \\ \mathbf{n}.[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\hat{\varphi}(\mathbf{x}) - \beta & \text{si } \mathbf{x} = \mathbf{b} \in \partial_2\Omega \end{cases} \tag{15}$$

donde $\overset{\circ}{\Omega} = \Omega - \partial\Omega$ es el interior de Ω (también se le denomina *abierto* de Ω). El residuo ponderado se define como

$$R_w = \int_\Omega w(\mathbf{x})\,R(\mathbf{x})\;d\mathcal{A} \tag{16}$$

Para la formulación de Galerkin, la solución aproximada y su residuo ponderado con las funciones bases $\phi_j(\mathbf{x})$ son

$$\hat{\varphi}(\mathbf{x}) = \sum_{j=1}^{N} \varphi_j\,\phi_j(\mathbf{x}) \qquad\qquad R_\phi = \int_\Omega \phi_k(\mathbf{x})\,R(\mathbf{x})\;d\mathcal{A} = 0 \qquad\qquad \forall\phi_k \in C^0(\Omega) \tag{17}$$

Substituída estas expresiones, se obtiene el siguiente sistema de ecuaciones lineales

$$\sum_{j=1}^{N} \varphi_j \int_\Omega \phi_k(\mathbf{x})\left[\,\boldsymbol{\nabla}.\left(\,[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\phi_j\,\right) + S(\mathbf{x})\,\right] d\mathcal{A} = 0 \qquad\qquad \sum_{j=1}^{N} A_{kj}\,\varphi_j = b_k \qquad k = 1, 2, \ldots, N \tag{18}$$

donde los coeficientes del sistema lineal son

$$\begin{aligned} A_{kj} &= \int_\Omega \phi_k(\mathbf{x})\,\boldsymbol{\nabla}.\left(\,[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\phi_j\,\right) d\mathcal{A} = \oint_{\partial\Omega} \phi_k(\mathbf{x})\,\mathbf{n}.[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\phi_j\;d\mathcal{C} - \int_\Omega \boldsymbol{\nabla}\phi_k(\mathbf{x}).[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\phi_j(\mathbf{x})\;d\mathcal{A} \\ b_k &= -\int_\Omega \phi_k(\mathbf{x})\,S(\mathbf{x})\;d\mathcal{A} \end{aligned} \tag{19}$$

Fíjese que se ha aplicado el teorema de Green a los coeficientes A_{kj}. El primer término del tercer miembro de (19.a) es realmente la integral cerrada de línea $\oint_{\partial\Omega} d\mathcal{C}$, donde $\mathcal{C} = \partial_1\Omega \cup \partial_2\Omega$ es la curva de la frontera.

La función de forma ϕ_i, $i = m_1, m_2, m_3$, para la interpolación en un elemento triangular m son [Reddy,2005] [Reddy & Gartling,2000]

$$\phi_i = \frac{1}{2\mathcal{A}_m}(\alpha_i^m + \beta_i^m\, x + \gamma_i^m\, y) \qquad \begin{aligned} \alpha_i^m &= x_j y_k - x_k y_j \\ \beta_i^m &= y_j - y_k \\ \gamma_i^m &= -(x_j - x_k) \end{aligned} \qquad A_{kj}^m = \frac{-1}{4\mathcal{A}_m} \left\{ \beta_k^m \quad \gamma_k^m \right\} \begin{bmatrix} \Gamma_{xx} & \Gamma_{xy} \\ \Gamma_{yx} & \Gamma_{yy} \end{bmatrix}^m \begin{Bmatrix} \beta_j^m \\ \gamma_j^m \end{Bmatrix} \tag{20}$$

Los índices i, j, k en (20.a, b) son cualquier permutación derecha de m_1, m_2, m_3. Los índices k, j en (20.c) se refieren a $\boldsymbol{\nabla}\phi_k$ y $\boldsymbol{\nabla}\phi_j$ en (21.b) abajo. El signo menos de (21.b) no se cancela, como en el caso unidimensional, aunque estos dos gradientes tengan valores opuestos, pero dichos valores están contenidos en los coeficientes β y γ de cada lado k y j. Uno de los $\mathcal{A}_m^2$ en el denominador de (20.c) se ha cancelado durante la integración de (21.b) con gradientes constantes. El super-índice m en el tensor $[\boldsymbol{\Gamma}]^m$ se refiere a que dicho tensor se establece promedio para el elemento m.

La integración en (19) sobre todo el dominio Ω se ha hecho sumando los resultados por elementos Ω_m

$$\begin{aligned} A_{kj} &= \oint_{\partial\Omega} \phi_k(\mathbf{x})\, \mathbf{n}.[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\phi_j \; d\mathcal{C} + \sum_{m=1}^{M} A_{kj}^m \qquad\qquad b_k = -\frac{1}{3}\sum_{m,k} \mathcal{A}_{m,k}\, \bar{S}_m(\mathbf{x}_k) \\ A_{kj}^m &= -\int_{\Omega_m} \boldsymbol{\nabla}\phi_k(\mathbf{x}).[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\phi_j(\mathbf{x})\; d\mathcal{A} = \begin{cases} = 0 & \text{si } k,j \neq m_1 \text{ y } k,j \neq m_2 \text{ y } k,j \neq m_3 \\ \neq 0 & \text{si } k,j = m_1 \text{ o } k,j = m_2 \text{ o } k,j = m_3 \end{cases} \end{aligned} \tag{21}$$

El elemento b_k se calcula como la media de los términos de fuente en los elementos alrededor del nodo k, ponderado con los volúmenes $\frac{1}{3}\mathcal{A}_{m,k}$ de las distintas funciones de forma ϕ_k en los elementos m vecinos al nodo k.

La matriz A_{kj} se rellena de la siguiente manera:

Para cada k fijo, se fija también la fila en la matriz y se fija un nodo k correspondiente en el dominio. Se rellena los elementos de los nodos vecinos integrando los resultados (21.b) en los elementos vecinos a los que pertenece, cuyos resultados para cada uno son (20.c). Luego el elemento del nodo central $j = k$ es menos la suma de los coeficientes de los elementos de la matriz para los nodos vecinos.

Finalmente se establecen las condiciones de contorno. Todos los nodos j de la porción $\partial_1\Omega$, se le suma al elemento independiente k (índice del nodo próximo a la frontera) de dicho nodo, el valor $-A_{kj}\,\varphi_j$, donde $\varphi_j = \varphi_j(\mathbf{x}) = \alpha$, con $\mathbf{x} = \mathbf{a} \in \partial_1\Omega$. Para este borde, el término con $\phi_k(\mathbf{a})$ siempre se anula en (21.a) en la porción $\partial_1\Omega$ de $\mathcal{C}$.

Todos los nodos j de la porción $\partial_2\Omega$, se le suma al elemento independiente k (índice del nodo próximo a la frontera) de dicho nodo, el valor $-\mathbf{n}.[\boldsymbol{\Gamma}].\boldsymbol{\nabla}\varphi(\mathbf{x})\big|_{\mathbf{b}}\, \Delta l_k = -\beta\, \Delta l_k$, donde $\mathbf{x} = \mathbf{b} \in \partial_2\Omega$. La cantidad Δl_k es el tamaño del lado opuesto al nodo k, vértice en el triángulo del borde, donde el segmento Δl_k forma parte de la frontera aproximada poligonal en $\partial_2\Omega$. Para este borde, el término con $\phi_k(\mathbf{b})$ siempre se anula en (21.a) en la porción $\partial_2\Omega$ de $\mathcal{C}$. Los valores the φ_j para los nodos de la porción de la frontera $\partial_2\Omega$ de $\mathcal{C}$ también son incógnitas. Estos nodos son los vértices de los elementos triangulares con una sola punta (o varias) en la frontera estrellada, una vez eliminados los segmentos (frontera $\partial_2\Omega$ = frontera poligonal (segmentos) + frontera estrellada (nodos)).

Al final se contará con un sistema de N ecuaciones lineales con N incógnitas, los valores de φ_k, $k = 1, 2, \ldots, N$, indeterminados en los nodos, excluyendo los nodos en la frontera en la porción $\partial_1\Omega$, que conforma una matriz diagonal en bloques.

5.4.3. Transitorio

Sea la ecuación diferencial en $\varphi(t, \mathbf{x})$

$$\frac{\partial \varphi}{\partial t} = \boldsymbol{\nabla}. \left([\boldsymbol{\Gamma}]. \boldsymbol{\nabla} \varphi \right) + S \tag{22}$$

con condiciones iniciales $\varphi^o = \varphi(0, \mathbf{x})\ \forall \mathbf{x} \in \Omega$ conocidas en $t = 0$, y condiciones de borde (13) conocidas para todo instante t. El coeficiente de difusividad tensorial $\boldsymbol{\Gamma}(t, \mathbf{x})$ y el término de fuente $S(t, \mathbf{x})$ pueden depender también del tiempo t.

Una vez substituída la solución aproximada

$$\hat{\varphi}(t, \mathbf{x}) = \sum_{j=1}^{N} \varphi_j(t)\, \phi_j(\mathbf{x}) \tag{23}$$

y aplicado el método de Galerkin (17)-(18) y discretizado el problema en los elementos finitos (19)-(21), se obtiene el siguiente sistema de ecuaciones diferenciales ordinaria de primer orden en $\boldsymbol{\varphi}(t) = \{\varphi_j(t)\}$

$$B_{kj}\, \frac{d\varphi_j(t)}{dt} = A_{kj}\, \varphi_j(t) - b_k \qquad\qquad [\mathbf{B}]. \frac{d\boldsymbol{\varphi}}{dt} = [\mathbf{A}]. \boldsymbol{\varphi} - \mathbf{b} \tag{24}$$

donde B_{kj} se calcula como

$$B_{kj} = \int_{\Omega} \phi_k(\mathbf{x})\, \phi_j(\mathbf{x})\, d\mathcal{A} = \sum_{m=1}^{M} B_{kj}^m \qquad\qquad B_{kj}^m = \int_{\Omega_m} \phi_k(\mathbf{x})\, \phi_j(\mathbf{x})\, d\mathcal{A} \tag{25}$$

Los valores de A_{kj} y b_k son los mismos que en (19), con las mismas observaciones que allí se han hecho. El sistema se resuelve con cualquiera de los métodos expuestos en el capítulo IV, una vez despejado el vector $d\boldsymbol{\varphi}/dt$ al multiplicar la ecuación (24) por $[\mathbf{B}]^{-1}$. También se pueden emplear los esquemas de Euler implícito de la sección 2.2.1. o el esquema de Crank-Nicolson de la sección 2.2.2., diseñados para ecuaciones diferenciales parabólicas, como lo es la ecuación (22). Los valores B_{kj}^m son distintos de cero sólamente para los elementos m vecinos al nodo k. El índice j indica la variable $\varphi_j(t)$ sobre la que actúa B_{kj}^m y A_{kj}^m, para cada elemento m vecino del nodo k, instantáneamente.

Como los coeficientes de (24.b) son matriciales, pero no necesariamente constantes en t, se puede aplicar el factor integrante

$$[\boldsymbol{\mu}(t)] = \exp\left(- \int_0^t [\mathbf{B}]^{-1}. [\mathbf{A}]\ dt \right) \tag{26}$$

obteniéndose

$$\frac{d}{dt} \left\{ [\boldsymbol{\mu}]. \boldsymbol{\varphi} \right\} = -[\boldsymbol{\mu}]\, .\, [\mathbf{B}]^{-1}. \mathbf{b} \qquad\qquad \varphi(t, \mathbf{x}) = [\boldsymbol{\mu}]^{-1}. \left[\boldsymbol{\varphi}^o - \int_0^t [\boldsymbol{\mu}]. [\mathbf{B}]^{-1}. \mathbf{b}\ dt \right] \tag{27}$$

Cuando los coeficientes $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{b}$ son constantes en el tiempo, la solución (27.b) es fácilmente obtenible sin necesidad de utilizar ningún método numérico adicional para el sistema de ecuaciones diferenciales ordinarias.

BIBLIOGRAFIA

[1] Anderson, D. A.; Tannehill, J. C.; Pletcher, R. H. **Computational Fluid Mechanics and Heat Transfer**. Hemisphere Publishing Corporation, 1984.

[2] Bathe, K.-J. **Finite Element Procedures**. Prentice-Hall, 1982 - Simon & Schuster (New Jersey), 1996.

[3] Burden R. L.; Faires, J. D. **Numerical Analysis**. 3^{rd} Edition. PWS. Boston, 1985.

[4] Ciarlet, Ph. G. **The Finite Element Method for Elliptic Problems**. North-Holland (Amsterdam), 1978. Siam (Philadelphia), 2002.

[5] Crank, J.; Nicolson, P. "A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of The Heat-Conduction Type". **Proc. Camb. Phil. Soc.**, Vol.**43**, pp.50-67, (1947). **Advances in Computational Mathematics**, Vol.**6**, pp.207-226, (1996).

[6] Donea, J.; Huerta, A. **Finite Element Methods for Flow Problems**. John Wiley & Sons (West Sussex, UK), 2003.

[7] Finlayson, B. A. **The Method of Weighted Residuals and Variational Principles**, with Application in Fluid Mechanics, Heat and Mass Transfer. Academic Press (New York), 1972.

[8] Gerald, C. F. **Applied Numerical Analysis**, 2^{nd} Edition. Addison-Wesley (New York), 1978.

[9] Granados, A. L. **Flujo Turbulento Cargado con Partículas Sólidas en una Tubería Circular**, Tesis Doctoral, Univ. Politécnica de Madrid, E. T. S. Ing. Industriales, 2003.

[10] Hughes, T. J. R. **The Finite Element Method**, Linear Static and Dynamic Finite Element Analysis. Prentice-Hall (Englewood Cliff, N. J.), 1987. Dover Publications (New York), 2000.

[11] Kim, J.; Moin, P. "Application of a Fractional-Step Method to Incompresible Navier-Stokes Equations", J. Comp. Physics, Vol.**59**, pp.308-323, (1985).

[12] Moukalled, F.; Mangani, L.; Darwish, M. **The Finite Volume Method in Computational Fluid Dynamics**. An Advanced Introduction with OpenFOAM® and Matlab® . Springer International Publishing (Switzerland), 2016.

[13] Orlandi, P. **Fluid Flow Phenomena: A Numerical Toolkit**. Kluwer Academic Publishers (Dordrecht, The Netherlands), 2000.

[14] Özişik, M. Necati **Finite Difference Methods in Heat Transfer**. CRC Press, 1994.

[15] Patankar, S.V. **Numerical Heat Transfer and Fluid Flow**. Hemisphere Publishing Corporation (New York), 1980.

[16] Reddy, J. N. **An Introduction to the Finite Element Method**, Third Edition. McGraw-Hill, 2005.

[17] Reddy, J. N. **Energy Principles and Variational Methods in Applied Mechanics**, 2^{nd} Edition. John Wiley & Sons (New Jersey), 2002.

[18] Reddy, J. N.; Gartling, D. K. **The Finite Element Method in Heat Transfer and Fluid Dynamics**, Second Edition. CRC Press, 2000.

[19] Rhie, C. M.; Chow, W. L. "Numerical Study of The Turbulent Flow Past an Airfoil with Trailing Edge Separation", **AIAA Journal**, Vol.**21**, pp.15251532, (1983).

[20] Thomas, J. W. **Numerical Partial Differential Equations: Finite Difference Method**. Springer Science+Business Media (New York), 1995.

[21] Versteeg, H. K.; Malalasekera, W. **An Introduction to Computational Fluid Dynamics: The Finite Volume Method**. Pearson Education, 1995. Second Edition, 2007.

[22] Zienkiewicz, O. C.; Taylor, R. L.; Nithiarasu, P. **The Finite Element Method for Fluid Dynamics**, Sixth Edition. Elsevier - Butterworth-Heinemann (Boston, MA), 2005.

APENDICE. SERIES DE TAYLOR.

TAYLOR SERIES FOR MULTI-VARIABLE FUNCTIONS

Andrés L. Granados M.
Department of Mechanics
SIMON BOLIVAR UNIVERSITY
Valle de Sartenejas, Estado Miranda
Apdo.89000, Caracas 1080A, Venezuela. **e-mail:** agrana@usb.ve

ABSTRACT

This paper intends to introduce the Taylor series for multi-variable real functions. More than a demostration of the teorema, it shows how to expose the series in a compact notation. Generalization of the jacobian of any order of a function with multiple dependence is defined. For this we use the differential operator $\boldsymbol{\nabla}$ with multiple tensor products. Also is established a special multiplication of the derivatives with the displacement of independent variables. This multiplicaction is identified as the contraction of indexes. At the end there is a new proof of the Taylor's Theorem for vectorial and tensorial functions. Also it is included the multi-index notation version of the series.

PRELIMINARS

We shall go here step by step. First we define the different operators performed on escalar, vectorial and tensorial function. The funtions may be on escalar or vectorial variable. Second we define the operations, different types of multiplications, between them or with functions or variables.

Escalars

Let be $f(\mathbf{x})\colon \mathbb{R}^M \longrightarrow \mathbb{R}$ a continuous escalar function, with continuous partial derivative. The *gradient* is the following operation

$$\operatorname{grad} f = \boldsymbol{\nabla} f \qquad\qquad \boldsymbol{\nabla} = \hat{\mathbf{e}}^i\, \partial_i \tag{1}$$

Although the both notation are common, the second is more used. The operator $\boldsymbol{\nabla}$, denominated "nabla", is defined in (1.b), with $\partial_i = \partial/\partial x^i$ and $\hat{\mathbf{e}}^i$ is the constant base. There is not confusion about the ordering of the operator and the operated. We follow the summation convention for repeated indexes (dummy index).

Vectors

Let be $\mathbf{f}(\mathbf{x})\colon \mathbb{R}^M \longrightarrow \mathbb{R}^N$ a continuous vectorial function, with continuous partial derivative of their components. The *gradient* and the *divergence* are the following operations

$$\operatorname{grad} \mathbf{f} = (\boldsymbol{\nabla}\mathbf{f})^{\mathrm{t}} = \mathbf{J_f} \qquad\qquad \operatorname{div} \mathbf{f} = \boldsymbol{\nabla}.\mathbf{f} = \partial_i f^i \tag{2}$$

In the case of gradient we operate with $\boldsymbol{\nabla}$, but then we transpose. That is the correct ordering. Thus the jacobian $\mathbf{J_f}$ has component $J^i_{\cdot j} = \partial_j f^i$. In a matrix, this componente will be in the row i and the column j. That is why we transpose. With the divergence there is not confusion. The operator makes a escalar product with the vectorial function. This product is commutative, but this is not necessary because the result is a escalar.

Tensors

Let be $\mathbf{F}(\mathbf{x})\colon \mathbb{R}^M \longrightarrow \mathbb{R}^N \times \mathbb{R}^N$ a continuous second order tensorial function, with continuous partial derivative of their components. The *gradient* and the *divergence* are the following operations

$$\operatorname{grad} \mathbf{F} = (\boldsymbol{\nabla}\mathbf{F})^{\mathrm{t}} = \mathbf{J_F} \qquad\qquad \operatorname{div} \mathbf{F} = \boldsymbol{\nabla}.\mathbf{F}^{\mathrm{t}} = \hat{\mathbf{e}}_i\, \partial_j F^{ij} \tag{3}$$

The gradient needs a transposition with the operator nabla because the variable which is derivated has the first indice in the array (free index). For the divergence the double transposition is necessary because the dummy index (repeated index by summation convention) contracted by the operation " $\cdot$ " corresponds to the last index of $\mathbf{F}$ components and the index of ∂_j ($\hat{\mathbf{e}}^i.\hat{\mathbf{e}}_j = \delta^i_j$). The difference of operators, between grad or div and $\boldsymbol{\nabla}$ or $\boldsymbol{\nabla}.$, is the ordering of derivation. That is why we eventually need the transpositions, as for "rot" and $\boldsymbol{\nabla}\times$ in the *rotational* operator, when is applied to tensors.

Operators

Instead "grad" and "div", we shall use the following operators that have some especial properties

$$\boldsymbol{\nabla} = \boldsymbol{\nabla} \otimes \qquad \boldsymbol{\nabla}. \qquad \Delta = \nabla^2 = \boldsymbol{\nabla}^2 = \boldsymbol{\nabla}.\boldsymbol{\nabla} \qquad (\quad \boldsymbol{\nabla}\mathbf{x} = \mathbf{I} \qquad \boldsymbol{\nabla}.\mathbf{x} = N \quad) \tag{4}$$

The first operator is the gradient. When applied on a vectorial function forms a diadic. Frequently, the symbol $\otimes$ is avoided for simplicity, as in $(2.a)$ and $(3.a)$. The second operator is the divergence and one has to take care over which part acts the contraction to produce a dummy index. The third operator is the well known *laplacian*. The last two properties between parenthesis are obvious, resulting in identity tensor $\mathbf{I}$ and the dimension of $\mathbf{x}$.

CALCULUS

Two aspects are involves in the following notation: the multiplicactions and the derivatives.

Multiplications

There are two forms of multiplicactions. The first of them is called the *tensorial multiplication*. At the left is shown how is the exponentiation of a vector by an exponent k with this multiplication.

$$\mathbf{v}^{k^{\otimes}} = \overbrace{\mathbf{v} \otimes \mathbf{v} \otimes \cdots \otimes \mathbf{v}}^{k \text{ times}} \qquad \boldsymbol{\nabla}^{k^{\otimes}} = \overbrace{\boldsymbol{\nabla} \otimes \boldsymbol{\nabla} \otimes \cdots \otimes \boldsymbol{\nabla}}^{k \text{ times}} \tag{5}$$

At the right it is shown how is the same exponentiation but with the differential operator nabla. The permutation of factors in $(5.b)$ may be in any manner due to the interchageable of the ordering of derivation by the continuity of derivatives. To be consistent with, $\mathbf{v}^{0^{\otimes}} = 1$ and $\boldsymbol{\nabla}^{0^{\otimes}} =$ non-derivatives.

The second form of multiplication is the *escalar product or interior multiplication*

$$\mathbf{u}.\mathbf{v} = u_i v^i \qquad \mathbf{U} : \mathbf{V} = U_{ij} V^{ji} \qquad \begin{aligned} (\mathbf{A} \overset{k}{\odot} \mathbf{B})^{::rs}_{ij..} &= A_{ij..lmn}\, B^{nml..rs} \\ \mathbf{T} \overset{k}{\odot} \mathbf{v}^{k^{\otimes}} &= [\cdots[\mathbf{T}.\overbrace{\mathbf{v}].\mathbf{v}\cdots].\mathbf{v}}^{k \text{ times}} \end{aligned} \tag{6}$$

Between vectors is the escalar multiplication. Repeated twice between two second order tensor is the escalar multiplication of tensors (some mathematicians use only one point for this multiplication). In general, in the extreme right, it means the number of contraction of the adjacent index in each part, at one side and the other side of the point, to form dummy indexes. In the example $(6.c)$, k times products contract indexes $nml..$ in that ordering (from inside to outside), thus this number coincides (at last) with the number of repeated indexes. Normally, this ocurres to mixed index. In $(6.d)$, order of tensor $\mathbf{T} \geq k$.

Particularly, the notation in (5) may be extended to another kind of multiplication. This is the case of the potency or the exponentiation of a second order tensor $\mathbf{A}$ where should be interpreted

$$\mathbf{A}^{k^{\odot}} = \overbrace{\mathbf{A}.\mathbf{A}.\cdots.\mathbf{A}}^{k \text{ times}} \equiv \mathbf{A}^k \tag{5'}$$

as in matrix exponentiation (matrixes are arrays of the components of second order tensors in a particular basis, and their exponentiation is with conventional matrix multiplication where $[\mathbf{A}.\mathbf{B}] = [\mathbf{A}]\,[\mathbf{B}]$). Also, this has been naively used for vectors in scalar multiplication such as $\mathbf{v}^2 = \mathbf{v}.\mathbf{v}$ in $(6.a)$ or $\boldsymbol{\nabla}.\boldsymbol{\nabla} = \nabla^2$ in $(4.c)$. Obviously, exponentiations with respect to $k^{\odot}$ and $k^{\otimes}$ exponents are substantially different.

Derivatives

As two examples of gradient derivatives of vectorial functions ($\mathbf{x} = x^i\,\hat{\mathbf{e}}_i$), we have the jacobian matrix and the hessian tensor, whose definitions are shown below

$$\mathbf{J_f}(\mathbf{x}) = [\boldsymbol{\nabla}\mathbf{f}(\mathbf{x})]^{\mathrm{t}} \qquad\qquad \mathbf{H_f}(\mathbf{x}) = \mathbf{J}_\mathbf{f}^2(\mathbf{x}) = [\boldsymbol{\nabla}[\boldsymbol{\nabla}\mathbf{f}(\mathbf{x})]^{\mathrm{t}}]^{\mathrm{t}} = [\boldsymbol{\nabla}^{2^\otimes}\mathbf{f}(\mathbf{x})]^{\mathrm{t}} \tag{7}$$

The necessary transposition are patent in the ordering of the indexes of the components (i=row, j=column and k=layer)

$$J^i_{\cdot j} = \frac{\partial f^i}{\partial x^j} \qquad\qquad H^i_{\cdot jk} = \frac{\partial^2 f^i}{\partial x^j \partial x^k} \tag{8}$$

A generalization of this concept of derivation, is the k order jacobian defined as follows

$$\mathbf{J}_\mathbf{f}^k(\mathbf{x}) = \overbrace{[\boldsymbol{\nabla}[\boldsymbol{\nabla}\cdots[\boldsymbol{\nabla}\mathbf{f}(\mathbf{x})]^{\mathrm{t}}\cdots]^{\mathrm{t}}]^{\mathrm{t}}}^{k\text{ times}} = [\boldsymbol{\nabla}^{k^\otimes}\mathbf{f}(\mathbf{x})]^{\mathrm{t}} \qquad\qquad \mathbf{J}_\mathbf{f}^{k+1}(\mathbf{x}) = [\boldsymbol{\nabla}\mathbf{J}_\mathbf{f}^k(\mathbf{x})]^{\mathrm{t}} \tag{9}$$

See the particular cases $k = 1, 2$ in (7), for the jacobian and the hessian. The number of transpositions and tensorial multiplications are the same, k times. Here the symbol $\otimes$ has been partially omitted for simplicity as in $(4.a)$. The expression is briefly defined with symbols of $(5.b)$ at the end of the expression (9). The transposition is for the global factor. Obviously, $\mathbf{J}_\mathbf{f}^0(\mathbf{x}) = \mathbf{f}(\mathbf{x})$.

TAYLOR SERIES

There are shown two forms of Taylor series, the escalar and the vectorial or tensorial. The tensorial form is the same to the vectorial form, changing $\mathbf{f}$ by $\mathbf{F}$, a slight modification of equation (9) (see $(2.a)$ and $(3.a)$). All the rest remains equal.

Escalar Series

The escalar form of the Taylor series [1,2] is the following

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_o)}{k!}\,(x - x_o)^k + R_n(x) \tag{10.a}$$

The remainder term $R_n(x)$ is

$$R_n(x) = \int_{x_o}^{x} \frac{f^{(n+1)}(t)}{n!}\,(x - t)^n\,dt = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,(x - x_o)^{(n+1)} \qquad\qquad \xi \in [x_o, x] \tag{10.b}$$

The second member is the integral form used recurrently, with integrations by parts, to obtain the serie $(10.a)$. The third member is the form of Lagrange for the residual or remainder $R_n(x)$, which may be demonstrated by the Theorem of Mean-Value [3,4], but also by the Theorem of Rolle [5,6]. Remember that $0! = 1$ and $f^{(0)} = f$.

Vectorial Series

The vectorial form of the Taylor series is the following

$$\mathbf{f}(\mathbf{x}) = \sum_{k=0}^{n} \frac{\mathbf{J}_\mathbf{f}^k(\mathbf{x}_o)}{k!} \overset{k}{\odot} (\mathbf{x} - \mathbf{x}_o)^{k^\otimes} + \mathbf{R}_n(\mathbf{x}) \tag{11.a}$$

The remainder term $\mathbf{R}_n(\mathbf{x})$ is with $\boldsymbol{\xi} \in \mathbb{B}(\mathbf{x}_o, \|\mathbf{x}-\mathbf{x}_o\|)$

$$\mathbf{R}_n(\mathbf{x}) = \int_0^1 \frac{\mathbf{J}_\mathbf{f}^{n+1}(\mathbf{r}(t))}{n!} \overset{n+1}{\odot} (\mathbf{x} - \mathbf{x}_o)^{(n+1)^\otimes}\,(1 - t)^n\,dt = \frac{\mathbf{J}_\mathbf{f}^{n+1}(\boldsymbol{\xi})}{(n+1)!} \overset{n+1}{\odot} (\mathbf{x} - \mathbf{x}_o)^{(n+1)^\otimes} \tag{11.b}$$

where $\mathbb{B}(\mathbf{x}_o, \|\mathbf{x}-\mathbf{x}_o\|)$ is the $\mathbb{R}^N$ close ball of center in $\mathbf{x}_o$ and radius $\|\mathbf{x}-\mathbf{x}_o\|$. The topological structures of (11) and (10) are the same. Next section we shall show why the second member of (11.b) has such expression.

Some solutions use what is explained in continuation. Parametrize the line segment between $\mathbf{x}_o$ and $\mathbf{x}$ by $\mathbf{r}(t) = \mathbf{x}_o + t\,(\mathbf{x}-\mathbf{x}_o)$ $(t \in [0,1])$. Then we apply the one-variable version of Taylor's theorem to the function $\mathbf{g}(t) = \mathbf{f}(\mathbf{r}(t))$, where $\mathbf{g}'(t) = [\boldsymbol{\nabla}\mathbf{f}(\mathbf{r})]^{\mathrm{t}}.\,\mathbf{r}'(t)$. Results are the same as [6,7], but the notations are different.

In [6] it is suggested to put under a unique exponent k the factors " $\mathbf{J}^k_{\mathbf{f}}(\mathbf{x}_o)$ " and " $(\mathbf{x}-\mathbf{x}_o)^{k^{\otimes}}$ ", and the multiple-operation " $\overset{k}{\odot}$ " in between, although the operation (without exponent), comprehended as a 'escalar product', is not exposed explicitly with symbol. The nabla ∇ operator is used for generalized jacobian $\mathbf{J}^{\alpha}_{f}(\mathbf{a}) = \nabla^{\alpha} f(\mathbf{a})$, with transposition included, implicitly understood. However, $\mathbf{J}^k_{\mathbf{f}}(\mathbf{x})$ should be seen as k-times compositions of a differential operator $\boldsymbol{\nabla}$ over $\mathbf{f}$ (transposition included), rather than a simple power k of $\boldsymbol{\nabla}\mathbf{f}$ (see equation (9)). One may be tempted to enclose the superindex of $\mathbf{J}$ with parenthesis, but this will over-recharge the notation innecessarily (besides, there is no the confusion as in f^k and $f^{(k)}$). In [7] is used $D^{\alpha} f(\mathbf{a})$ instead $\nabla^{\alpha} f(\mathbf{a})$, and no explicit operation is mentioned between the factors $D^{\alpha} f$ and $(\mathbf{x}-\mathbf{a})^{\alpha}$. The remainder term is consistent.

Tensorial Series

This form is exactly the same as the vectorial form without any particularity, except as mentioned. The demostration of vectorial (11) or tensorial form of series is similar to the escalar (10) form of series [3,4], taking into account the particularity of the operations (5) and (6), and the definition (9).

TAYLOR'S THEOREM

Taylor's theorem establish the existence of the corresponding series and the remainder term, under already mentioned conditions. We present now two proof of Taylor's theorem based on integration by parts, one for escalar functions [4], the other for vectorial function, similar in context, but different in scope. The first will guide the second. Both are based on a recurrent relationship that starts with an initial expression.

Escalar Proof

Integration by parts states that

$$\int u\,dv = uv - \int v\,du \qquad\qquad \int_a^b u(t)\,v'(t)\,dt = u(t)\,v(t)\,\big|_a^b - \int_a^b v(t)\,u'(t)\,dt \tag{12}$$

If we select $a = x_o$, $b = x$ and

$$u(t) = f^{(k)}(t) \qquad du = f^{(k+1)}(t)\,dt \qquad v = -\frac{(x-t)^k}{k!} \qquad dv = \frac{(x-t)^{k-1}}{(k-1)!}\,dt \tag{13}$$

it is obtained

$$\int_{x_o}^{x} \frac{f^{(k)}(t)}{(k-1)!}\,(x-t)^{k-1}\,dt = \frac{f^{(k)}(x_o)}{k!}\,(x-x_o)^k + \int_{x_o}^{x} \frac{f^{(k+1)}(t)}{k!}\,(x-t)^k\,dt \tag{14}$$

The recurrent expression (14) permits to obtain (10.a) series, begining with $k = 1$ and

$$\int_{x_o}^{x} f'(t)\,dt = f(x) - f(x_o) \tag{15}$$

Including its remainder term (10.b), with $k = n$, in its first form (second member), which becomes in the second form (last member) via the mean-value theorem

$$\int_a^b g(t)\,h(t)\,dt = g(c)\int_a^b h(t)\,dt \qquad\qquad c \in [a,b] \tag{16}$$

for continuous functions $g(t)$ and $h(t)$ in the interval.

Vectorial Proof

We now parametrize the line segment between $\mathbf{x}_o$ and $\mathbf{x}$ by a function $\mathbf{r}(t)\colon \mathbb{R} \longrightarrow \mathbb{R}^N$ defined as

$$\mathbf{r}(t) = \mathbf{x}_o + t\,(\mathbf{x} - \mathbf{x}_o) \qquad\qquad t \in [0,1] \tag{17}$$

Then we apply the one-variable version of Taylor's theorem to the function $\mathbf{g}(t)\colon \mathbb{R} \longrightarrow \mathbb{R}^M$ with

$$\mathbf{g}(t) = \mathbf{f}(\mathbf{r}(t)) \qquad\qquad \text{and} \qquad\qquad \mathbf{g}'(t) = [\boldsymbol{\nabla}\mathbf{f}(\mathbf{r})]^{\mathrm{t}}\,.\,\mathbf{r}'(t) = \mathbf{J}_{\mathbf{f}}(\mathbf{r})\,.\,\mathbf{r}'(t) \tag{18}$$

where $\mathbf{r}'(t) = \mathbf{x} - \mathbf{x}_o$ is a constant in t, therefore

$$\mathbf{g}^{(k)}(t) = \frac{d\mathbf{g}^{(k-1)}}{dt} = \left\{\, [\boldsymbol{\nabla}\mathbf{J}_{\mathbf{f}}^{k-1}(\mathbf{r})]^{\mathrm{t}} \overset{k-1}{\odot} [\mathbf{r}'(t)]^{(k-1)^{\otimes}} \right\}.\mathbf{r}'(t) = \mathbf{J}_{\mathbf{f}}^{k}(\mathbf{r}) \overset{k}{\odot} [\mathbf{r}'(t)]^{k^{\otimes}} = \mathbf{J}_{\mathbf{f}}^{k}(\mathbf{r}) \overset{k}{\odot} (\mathbf{x} - \mathbf{x}_o)^{k^{\otimes}} \tag{19}$$

Note that, in the third member of (19), the operations " $\overset{k-1}{\odot}$ " and " . " combine in one operation " $\overset{k}{\odot}$ ". Application of (14) to $\mathbf{g}(t)$ function, instead $f(t)$; with $a = 0$ and $b = 1$ in (12), instead $a = x_o$ and $b = x$, produces

$$\int_0^1 \frac{\mathbf{J}_{\mathbf{f}}^{k}(\mathbf{r}(t))}{(k-1)!} \overset{k}{\odot} (\mathbf{x} - \mathbf{x}_o)^{k^{\otimes}}\,(1-t)^{k-1}\,dt = \frac{\mathbf{J}_{\mathbf{f}}^{k}(\mathbf{x}_o)}{k!} \overset{k}{\odot} (\mathbf{x} - \mathbf{x}_o)^{k^{\otimes}} + \int_0^1 \frac{\mathbf{J}_{\mathbf{f}}^{k+1}(\mathbf{r}(t))}{k!} \overset{k+1}{\odot} (\mathbf{x} - \mathbf{x}_o)^{(k+1)^{\otimes}}\,(1-t)^{k}\,dt \tag{20}$$

The equivalent of (14).

The recurrent expression (20) permits to obtain (11.a) series, begining with $k = 1$ and

$$\int_0^1 \mathbf{g}'(t)\,dt = \mathbf{g}(1) - \mathbf{g}(0) = \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_o) = \int_0^1 \mathbf{J}_{\mathbf{f}}(\mathbf{r}(t))\,.\,(\mathbf{x} - \mathbf{x}_o)\,dt \tag{21}$$

Including its remainder term (11.b), with $k = n$, in its first form (second member), which becomes in the second form (last member) via the mean-value theorem (16), applied to a vectorial function $\mathbf{g}(t)$

$$\int_0^1 \mathbf{g}(t)\,h(t)\,dt = \mathbf{g}(\tau)\int_0^1 h(t)\,dt \qquad\qquad \tau \in [0,1] \tag{22}$$

what means that $\boldsymbol{\xi} = \mathbf{x}_o + \tau\,(\mathbf{x} - \mathbf{x}_o)$ in (11.b). As $\mathbf{x}$ is in the close ball spherical cap of center $\mathbf{x}_o$, then $\boldsymbol{\xi}$ is inside the ball. All said here in this proof for vectorial functions is valid also for tensorial functions changing $\mathbf{f}$ by $\mathbf{F}$ and $\mathbf{g}$ by $\mathbf{G}$.

MULTI-INDEX FORM

An m-dimensional multi-index is an ordered m-tuple [8]

$$\alpha = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_m) \tag{23}$$

of non-negative integers $\mathbb{Z}^+$ (natural numbers $\mathbb{N}$) $\alpha_i \in \mathbb{N}$. They have the properties:

• Sum of components

$$|\alpha| = \alpha_1 + \alpha_2 + \alpha_3 + \cdots + \alpha_m = \sum_{i=1}^{m} \alpha_i \tag{24}$$

• Factorial

$$\alpha! = \alpha_1!\ \alpha_2!\ \alpha_3!\ \cdots\ \alpha_m! = \prod_{i=1}^{m} \alpha_i! \tag{25}$$

With this notation, the Taylor series will be expressed as

$$\mathbf{f}(\mathbf{x}) = \sum_{|\alpha|\geq 0}^{|\alpha|=n} \frac{\mathbf{J}_{\mathbf{f}}^{|\alpha|}(\mathbf{x}_o)}{\alpha!} \overset{|\alpha|}{\odot} (\mathbf{x}-\mathbf{x}_o)^{|\alpha|^{\otimes}} + \mathbf{R}_n(\mathbf{x}) \tag{26.a}$$

The remainder term $\mathbf{R}_n(\mathbf{x})$ is

$$\mathbf{R}_n(\mathbf{x}) = \sum_{|\beta|=n+1} \int_0^1 \frac{(n+1)\,\mathbf{J}_{\mathbf{f}}^{|\beta|}(\mathbf{r}(t))}{\beta!} \overset{|\beta|}{\odot} (\mathbf{x}-\mathbf{x}_o)^{|\beta|^{\otimes}} (1-t)^n\, dt = \sum_{|\beta|=n+1} \frac{\mathbf{J}_{\mathbf{f}}^{|\beta|}(\boldsymbol{\xi})}{\beta!} \overset{|\beta|}{\odot} (\mathbf{x}-\mathbf{x}_o)^{|\beta|^{\otimes}} \tag{26.b}$$

with $\boldsymbol{\xi} \in \mathbb{B}(\mathbf{x}_o, \|\mathbf{x}-\mathbf{x}_o\|)$ and n ($|\alpha| = n \neq 0$) a multi-index limit. Where it must be interpreted

$$\mathbf{J}_{\mathbf{f}}^{|\alpha|}(\mathbf{x}_o) = \left.\frac{\partial^{|\alpha|}\mathbf{f}}{\partial x_1^{\alpha_1}\cdots\partial x_m^{\alpha_m}}\right|_{\mathbf{x}=\mathbf{x}_o} \qquad (\mathbf{x}-\mathbf{x}_o)^{|\alpha|^{\otimes}} = (x_1-x_{o1})^{\alpha_1}\cdots(x_m-x_{om})^{\alpha_m} \tag{27}$$

The order of derivatives and powers are the same, Term by term, which guarantees the contraction factor by factor. Some factors for the derivatives, others factors for the powers, in each term. The order of derivations, the exponent of powers and the number of contractions concide. The derivative notation has a natural way to include the transposition of the operator implicitly (last derivates are respect to the first variables), which makes the transposition unnecessary.

This form means that the variability of a vectorial function, that depend on various variables, are additive in multiple directions for several terms, mutiplied the directions for each term (powers) on corresponding variable, with the same directions and order of derivations. The factorial in the denominator of (26), as a multi-index, takes into account the number of permutations of the same variable in a power (or a mixed derivative), and simplify them (Clairaut Theorem) [7]. Contrary to (11), which contains all the possible permutations of the powers, and thus may have repeated terms. However, both are equivalent. The same global power of variables may be repeated in different terms, but in different ways.

Example

For example, the third order Taylor polynomial of a scalar function $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$, denoting $\mathbf{v} = \mathbf{x}-\mathbf{x}_o$, is

$$\begin{aligned} P_3(\mathbf{x}) = f(\mathbf{x}_o) &+ \frac{\partial f(\mathbf{x}_o)}{\partial x_1} v_1 + \frac{\partial f(\mathbf{x}_o)}{\partial x_2} v_2 \\ &+ \frac{\partial^2 f(\mathbf{x}_o)}{\partial x_1^2} \frac{v_1^2}{2!} + \frac{\partial^2 f(\mathbf{x}_o)}{\partial x_1 \partial x_2} v_1 v_2 + \frac{\partial^2 f(\mathbf{x}_o)}{\partial x_2^2} \frac{v_2^2}{2!} \\ &+ \frac{\partial^3 f(\mathbf{x}_o)}{\partial x_1^3} \frac{v_1^3}{3!} + \frac{\partial^3 f(\mathbf{x}_o)}{\partial x_1^2 \partial x_2} \frac{v_1^2 v_2}{2!} + \frac{\partial^3 f(\mathbf{x}_o)}{\partial x_1 \partial x_2^2} \frac{v_1 v_2^2}{2!} + \frac{\partial^3 f(\mathbf{x}_o)}{\partial x_2^3} \frac{v_2^3}{3!} \end{aligned} \tag{28}$$

where it can be observed the mentioned characteristic [7]. The central term of second order appear twice in (11.a). As $v_1 v_2$ and $v_2 v_1$, that is why when they are divided by 2, disappear the factorial for this term. The two central terms of third order appear three times each one in (11.a). As $v_1^2 v_2$, $v_1 v_2 v_1$ and $v_2 v_1^2$ and as $v_1 v_2^2$, $v_2 v_1 v_2$ and $v_2^2 v_1$, respectively, that is why when they are divided by 3!, disappear 3 and appear 2! for those terms. This occurs only in the mixed terms. Finally, the polynomial (28) has the form (26.a), with (27) up to $|\alpha| = n = 3$, but can also be obtained with (11.a) for $n = 3$, and the posterior consolidation of terms.

REFERENCES

[1] Taylor, B. "Methodus Incrementorum Directa et Inversa", **Phylosophycal Transactions of the Royal Society** (London), (1715).

[2] Taylor, B. **Contemplatio Philosophica**. Published by his nephew Sir William Young, 1793.

[3] Apostol, T. M. **Calculus. Volume 2: Multivariable Calculus and Linear Algebra, with Applications to differential Equations and probability**, 2^{nd} Edition. John Wiley & Sons (New York), 1969.

[4] Thomas, G. B. **Calculus and Analytic Geometry**, 4^{th} Edition. Addison-Wesley (Massachusetts), 1968.

[5] Thomas, G. B. **Thomas' Calculus**, 12^{th} Edition. Addison-Wesley (Massachusetts), 2010.

[6] https://es.wikipedia.org/wiki/Teorema_de_Taylor

[7] https://en.wikipedia.org/wiki/Taylor%27s_theorem

[8] Saint Raymond, X. **Elementary Introduction to The Theory of Pseudodifferential Operators**. Chap 1.1. CRC Press, 1991.

BIBLIOGRAFIA GENERAL

- Abramowitz, M.; Stegun, I. A. **Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables**. Dover Publications, 1965. Ninth Printing, 1970.
- Anderson, D. A.; Tannehill, J. C.; Pletcher, R. H. **Computational Fluid Mechanics and Heat Transfer**. Hemisphere Publishing Corporation, 1984.
- Apostol, T. M. "An Elementary View of Euler's Summation Formula", **American Mathematical Monthly**, Vol.**106**, No.5, pp.409-418, (1999).
- Atkinson, K.E. **An Introduction to Numerical Analysis**. 2nd Edition. John Wiley & Sons, 1989.
- Bakhvalov, N. S. bf Numerical Methods. MIR Publishers-Moscow, 1977.
- Bathe, K.-J. **Finite Element Procedures**. Prentice-Hall, 1982 - Simon & Schuster (New Jersey), 1996.
- Bonnans, J. F.; Gilbert J. Ch.; Lemaréchal, C.; Sagastizábal, C. A. **Numerical Optimization - Theoretical and Practical Aspects**, Second Edition, Springer-Verlag (Berlin), 2006.
- Brent, R. P. **Algorithms for Minimization without Derivatives**. Prentice-Hall, 1973.
- Broyden, C. G. "A Class of Methods for Solving Non-Linear Simultaneous Equations", **Mathematics of Computation**, Vol.**19**, pp.577-593, (1965).
- Burden R. L.; Faires, J. D. **Numerical Analysis**. 3^{rd} Edition. PWS. Boston, 1985.
- Butcher, J. C. "Implicit Runge-Kutta Processes". **Math. Comput.**, Vol.**18**, pp.50-64, (1964).
- Butcher, J. C. "On Runge-Kutta Processes of High Order". **J. Austral. Math. Soc.**, Vol.**IV**, Part 2, pp.179-194, (1964).
- Butcher, J. C. **The Numerical Analysis of Ordinary Differential Equations, Runge-Kutta and General Linear Methods**. John Wiley & Sons (New York), 1987.
- Butcher, J. C. **Numerical Methods for Ordinary Differential Equations**, $2^{nd}/3^{rd}$ Editions. John Wiley & Sons (New York), 2008/2016.
- Carnahan, B.; Luther, H. A.; Wilkes, J. O. **Applied Numerical Methods**. John Wiley & Sons (New York), 1969.
- Cash, J. R.; Karp, A. H. **ACM Transactions on Mathematical Software**, Vol.**16**, pp.201-222, 1990.
- Chapra, S. C.; Canale, R. P. **Numerical Methods for Engineers**, with Personal Computer Applications. McGraw-Hill Book Company, 1985.
- Chapra S. C.; Canale, R. P. **Métodos Numéricos para Ingenieros**, Tercera Edición. McGraw-Hill Interamericana Editores (México), 1999.
- Ciarlet, Ph. G. **The Finite Element Method for Elliptic Problems**. North-Holland (Amsterdam), 1978. Siam (Philadelphia), 2002.
- Collatz, L. **The Numerical Treatment of Differential Equations**. Third Edition. Springer-Verlag, 1960. Second Printing, 1966.
- Conte, S.D.; deBoor, C. **Elementary Numerical Analysis**. McGraw-Hill (New York), 1972.
- Conte, S.D.; Carl de Boor. **Análisis Numérico**. 2da Edición. McGraw-Hill (México), 1974.
- Cottle, R. W.; Thapa, M. N. **Linear and Nonlinear Optimization**. Springer Science+Business Media (New York), 2017.
- Crank, J.; Nicolson, P. "A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of The Heat-Conduction Type". **Proc. Camb. Phil. Soc.**, Vol.**43**, pp.50-67, (1947). **Advances in Computational Mathematics**, Vol.**6**, pp.207-226, (1996).
- Dahlquist, G.; Björck, Å. **Numerical Methods**. Dover Publications (New York), 2003. Prentice-Hall, 1974.

- Dantzig, G. B.; Thapa, M. N. **Linear Programming**. Vol.**1**: "Introduction". Vol.**2**: "Theory and Extension". Springer (New York), 1997/2003.
- Dennis, J. E. Jr.; Moré, J. J. "Cuasi-Newton Methods, Motivation and Theory", **SIAM Review**, Vol.**19**, No.1, pp.46-89, (1977).
- Devaney, R. L. **An Introduction to Chaotic Dynamical Systems**. Addison-Wesley, 1987.
- Donea, J.; Huerta, A. **Finite Element Methods for Flow Problems**. John Wiley & Sons (West Sussex, UK), 2003.
- Fehlberg, E. "Low-Order Classical Runge-Kutta Formulas with Stepsize Control". **NASA** Report No. TR R-315, 1971.
- Finlayson, B. A. **The Method of Weighted Residuals and Variational Principles**, with Application in Fluid Mechanics, Heat and Mass Transfer. Academic Press (New York), 1972.
- Fletcher, R. **Practical Methods of Optimization**, 2^{nd} Edition. John Wiley & Sons (New York), 1987.
- Ford, J. A. **Improved Algorithms of Illinois-type for the Numerical Solution of Nonlinear Equations**, Technical Report CSM-257, University of Essex Press, 1995.
- Gear, C. W. **Numerical Initial Value Problems in Ordinary Differential Equations**. Prentice-Hall, 1971.
- Gerald, C. F. **Applied Numerical Analysis**. 2^{nd} Edition. Addison-Wesley, 1978.
- Gilbert, W. J. "Generalizations of Newton's Method". **Fractals**. Vol.**9**, No.3, pp.251-262, (2001).
- Granados M., A. L. **Nuevas Correlaciones para Flujo Multifásico.** INTEVEP S.A. Reporte Técnico No. INT-EPPR/322-91-0001. Los Teques, Febrero de 1991. Trabajo presentado en la Conferencia sobre: *Estado del Arte en Mecánica de Fluidos Computacional*. Auditorium de INTEVEP S.A. Los Teques, del 27 al 28 de Mayo de (1991).
- Granados M., A. L. **Second Order Methods for Solving Non-Linear Equations**, INTEVEP, S. A. (Research Institute for Venezuelan Petroleum Industry), Tech. Rep. No.INT-EPPR/322-91-0002, Los Teques, Edo. Miranda, Jun, 1991, págs. 14-36.
- Granados M., A. L. **Free Order Polynomial Interpolation Algorithm.** INTEVEP S.A. Nota Técnica. Los Teques, Julio de 1991.
- Granados M., A.L. **Lobatto Implicit Sixth Order Runge-Kutta Method for Solving Ordinary Differential Equations with Stepsize Control**. INTEVEP S.A. Reporte Técnico No. INT-EPPR/3-NT-92-003. Los Teques, Marzo de 1992.
- Granados M., A.L. "Fractal Technics to Measure the Numerical Instability of Optimization Methods". **Mecánica Computacional Vol.XV**: Anales del "9° CONGRESO SOBRE METODOS NUMERICOS Y SUS APLICACIONES, ENIEF'95". Hotel Amancay, 6-10 de Noviembre de 1995, San Carlos de Bariloche, Argentina. Compilado por Larreteguy, A. E. y Vénere, M. J. Asociación Argentina de Mecánica Computacional (AMCA), pp.369-374,1995.
- Granados M., A. L. "Fractal Techniques to Measure the Numerical Instability of Optimization Methods". **Numerical Methods in Engineering Simulation**: *Proceedings of The Third International Congress on Numerical Methods in Engineering and Applied Sciences, CIMENICS'96*. Cultural Centre Tulio Febres Cordero, March 25-29, 1996. Mérida, Venezuela. Editors: M. Cerrolaza, C. Gajardo, C. A. Brebbia. Computational Mechanics Publications of the Wessex Institute of Technology (UK), pp.239-247, (1996).
- Granados M. A. L. "Lobatto Implicit Sixth Order Runge-Kutta Method for Solving Ordinary Differential Equations with Stepsize Control". **Mecánica Computacional Vol.XVI**: *Anales del V Congreso Argentino de Mecánica Computacional, MECOM'96*. Universidad Nacional de Tucumán, Residencia Universitaria Horco Molle, Comuna de Yerba Buena, 10-13 de Septiembre de (1996). San Miguel de Tucumán, Argentina. Compilado por: Etse, G. y Luccioni, B. Asociación Argentina de Mecánica Computacional (AMCA), pp.349-359, (1996).

- Granados M., A. L. "Implicit Runge-Kutta Algorithm Using Newton-Raphson Method". **Simulación con Métodos Numéricos: Nuevas Tendencias y Aplicaciones**, Editores: O. Prado, M. Rao y M. Cerrolaza. Memorias del *IV CONGRESO INTERNACIONAL DE METODOS NUMERICOS EN INGENIERIA Y CIENCIAS APLICADAS, CIMENICS'98*. Hotel Intercontinental Guayana, 17-20 de Marzo de 1998, Puerto Ordaz, Ciudad Guayana. Sociedad Venezolana de Métodos Numéricos en Ingeniería (SVMNI), pp.TM9-TM16. Corregido y ampliado Abril, 2016. https://www.academia.edu/11949052/Implicit_Runge-Kutta_Algorithm_Using_Newton-Raphson_Method
- Granados M., A. L. "Implicit Runge-Kutta Algorithm Using Newton-Raphson Method". *Fourth World Congress on Computational Mechanics*, realizado en el Hotel Sheraton, Buenos Aires, Argentina, 29/Jun/98 al 2/Jul/98. International Association for Computational Mechanics, **Abstracts**, Vol.**I**, p.37, (1998).
- Granados, A. L. **Flujo Turbulento Cargado con Partículas Sólidas en una Tubería Circular**, Tesis Doctoral, Univ. Politécnica de Madrid, E. T. S. Ing. Industriales, 2003.
- Granados, A. L. "Numerical Taylor's Methods for Solving Multi-Variable Equations", Universidad Simón Bolívar, Mayo, 2015. https://www.academia.edu/12520473/Numerical_Taylors_Methods_for_Solving_Multi-Variable_Equations
- Granados, A. L. "Taylor Series for Multi-Variable Functions", Universidad Simón Bolívar, Dic. 2015. https://www.academia.edu/12345807/Taylor_Series_for_Multi-Variables_Functions
- Granados, A. L. "Criterios para Interpolar", Universidad Simón Bolívar, Jul. 2017.
- Granados, A. L. **Mecánica y Termodinámica de Sistemas Materiales Continuos**. Universidad Simón Bolívar, 2022. ISBN 980-07-2428-1.
- Granados, A. L. "Métodos Numéricos para Redes", Universidad Simón Bolívar, Mar. 2023.
- Granados M., A. L. "Algoritmo Simplex Lineal y No Lineal". Universidad Simón Bolívar, Departamento de Mecánica, Sep., 2024. https://www.academia.edu/123412738/Algoritmo_Simplex_Lineal_y__No_Lineal
- Gundersen, T. "Numerical Aspects of the Implementation of Cubic Equations of State in Flash Calculation Routines". **Computer and Chemical Engineering**. Vol.**6**, No.3, pp.245-255., (1982).
- Hageman, L. A.; Young, D. M. **Applied Iterative Methods**. Academic Press, 1981.
- Hairer, E.; Nørsett, S. P.; Wanner, G. **Solving Ordinary Differential Equations I: Nonstiff Problems**. Springer-Verlag, 1987.
- Hairer, E.; Wanner, G. **Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems**. Springer-Verlag, 1991.
- Hämmerlin, G.; Hoffmann, K.-H. **Numerical Mathematics**. Springer-Verlag (New York), 1991.
- Hamming, R. W. **Numerical Methods for Scientists and Engineers**. Second Edition. McGraw-Hill, 1973. Dover Publications, 1986.
- Hazewinkel, M. **Encyclopaedia of Mathematics**. Kluwer Academic Publishers (Dordrecht), 1988.
- Hildebrand, F. B. **Introduction to Numerical Analysis**, 2^{nd} Edition. Dover Publications (New York), 1974.
- Hillier, F. S.; Lieberman, G. J. **Introducción a La Investigación de Operaciones**, Novena Edición. McGraw-Hill (México), 2010.
- Hoffman, K.; Kunze, R. **Linear Algebra**, 2^{nd} Edition. Prentice-Hall (Englewood Cliff-New Jersey), 1971.
- Horner, W. G. "A new method of solving numerical equations of all orders, by continuous approximation". **Philosophical Transactions of the Royal Society of London**, pp.308-335, july (1819).
- Householder, A. S. **The Numerical Treatment of a Single Nonlinear Equation**. McGraw-Hill (New York), 1970.
- Householder, A. S. **The Theory of Matrices in Numerical Analysis**. Blaisdell Publishing Company (New York), 1964. Dover Publications (new York), 1975.

- Hughes, T. J. R. **The Finite Element Method,** Linear Static and Dynamic Finite Element Analysis. Prentice-Hall (Englewood Cliff, N. J.), 1987. Dover Publications (New York), 2000.
- Isaacson, E.; Keller, H.B. **Analysis of Numerical Methods**. John Wiley & Sons (New York), 1966.
- King, R. F. "Anderson-Bjorck for Linear Sequences". **Mathematics of Computation**, Vol.**41**, No.164, pp.591596, October, (1983).
- Lapidus, L. **Digital Computation for Chemical Engineers**. McGraw-Hill (New York), 1962.
- Lapidus, L.; Seinfeld, J. H. **Numerical Solution of Ordinary Differential Equations**. Academic Press (New York), 1971.
- Layton, W.; Sussman, M. **Numerical Linear Algebra**. World Scientific Publishing (New Jersey), 2020.
- Linz, P. **Theoretical Numerical Analysis**, An Introductuion to Advanced Techniques. John Wiley & Sons, 1979.
- Levenberg, K. "A Method for the Solution of Certain Non-Linear Problems in Least Squares". **Quarterly of Applied Mathematics**, Vol.**2**, pp.164168, (1944).
- Lobatto, R. **Lessen over Differentiaal- en Integraal-Rekening**. 2 Vol. La Haye, 1851-52.
- Luenberger, D. G. **Optimization by Vector Space Methods**. John Wiley & Sons, 1969.
- Luenberger, D. G.; Ye, Y. **Linear and Nonlinear Programming**, 5^{th} Edition. Springer Nature (Switzerland), 2021.
- Mandelbrot, B. B. **The Fractal Geometry of Nature**, Updated and Augmented Edition. W. H. Freeman and Company (New York), 1983.
- Marquardt, D. "An Algorithm for Least Squares Estimation of Non-Linear Parameters". **SIAM J. Appl. Math.**, Vol.**11**, No.2, pp.431-441, (1963).
- Méndez, M. V. **Tuberías a Presión.** En Los Sistemas de Abastecimiento de Agua. Fundación Polar & Universidad Católica Andrés Bello, 1995.
- Miranker, W. L. **Numerical Methods for Stiff Equations, and Singular Perurbation Problems**. D. Reidel Publishing Company, 1981.
- Moheuddin, Mir Md.; Jashim Uddin, Md.; Kowsher, Md. "A New Study to ind Out The Best Computational Method for Solving The Nonlinear Equation". **Applied Mathematics and Sciences**, Vol.**6**, No.2/3, pp.15-31, (2019).
- Moukalled, F.; Mangani, L.; Darwish, M. **The Finite Volume Method in Computational Fluid Dynamics**. An Advanced Introduction with OpenFOAM® and Matlab® . Springer International Publishing (Switzerland), 2016.
- Müller, D. E. "An Algorithm for Least Squares Estimation of Non-Linear Parameters". **Mathematical Tables and Other Aids to Computation (MTAC)**. Vol.**10**, pp.208-215, (1956).
- Nakamura, S. **Métodos Numéricos Aplicados con Software**. Prentice-Hall, 1992.
- Newton, I. bf De metodis fluxionum et serierum infinitarum, Méthod of Fluxions and Infinite Series. John Colson, 1736.
- Nocedal, J.; Wright, S. J. **Numerical Optimization,** 2^{nd} Edition. Springer (New York), 2006.
- Oliveira, I. F. D.; Takahashi, R. H. C. "An Enhancement of The Bisection Method Average Performance Preserving Minmax Optimality". **ACM Transactions on Mathematical Software**, Vol.**47**, No.1, pp.5:15:24, (2021).
- Ortega, J. M. **Numerical Analysis**, A Second Course. SIAM, 1990.
- Ortega, J. M.; Rheinboldt, W. C. **Iterative Solution of Nonlinear Equations in Several Variables**. Academic Press, 1970.
- [Ostrowski, A. M. **Solution of Equations and Systems of Equations**, 2^{nd} Edition. Academic Press (New York), 1966.

- Özişik, M. Necati **Finite Difference Methods in Heat Transfer**. CRC Press, 1994.
- Pachner, J. **Handbook of Numerical Analysis Applications**, With Programs for Engineers and Scientists. McGraw-Hill, 1984.
- Peitgen, H.-O.; Richter, P. H. **The Beauty of Fractals. Images of Complex Dynamical Systems**. Springer-Verlag, 1986.
- Pennington, R. H. **Introductory Computer Methods and Numerical Analysis**, 2^{nd} Edition. Collier Macmillan Ltd., 1970.
- Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. **Numerical Recipes**, The Art of Scientific Computing. Cambridge University Press, 1986. 4th Printing, 1988.
- Pundir, S. K. **Applied Numerical Analysis**. CBS Publisher & Distributors Pvt. Ltd., 2023.
- Rabinowitz, Ph.; Ed. **Numerical Methods for Nonlinear Algebraic Equations**. Gordon and Breach Science Publishers, 1970.
- Ralston, A.; Rabinowitz, P. **A First Course in Numerical Analysis**. 2^{nd} Edition. McGraw-Hill, 1978.
- Raphson, J. **Aequationum Universalis**. Royal Society, 1690.
- Reddy, J. N. **An Introduction to the Finite Element Method**, Third Edition. McGraw-Hill, 2005.
- Reddy, J. N. **Energy Principles and Variational Methods in Applied Mechanics**, 2^{nd} Edition. John Wiley & Sons (New Jersey), 2002.
- Reddy, J. N.; Gartling, D. K. **The Finite Element Method in Heat Transfer and Fluid Dynamics**, Second Edition. CRC Press, 2000.
- Rhie, C. M.; Chow, W. L. "Numerical Study of The Turbulent Flow Past an Airfoil with Trailing Edge Separation", **AIAA Journal**, Vol.**21**, pp.15251532, (1983).
- Richmond, H. W. "On Certain Formulae for Numerical Approximation", **J. London Math. Soc.**, Vol.**19**, Issue 73 Part 1, pp.31-38, January (1944).
- Samarski, A. A. **Introducción a los Métodos Numéricos**. Editorial MIR-Mocú, 1986.
- Samarski, A. A.; Andréiev, V. B. **Métodos en Diferencias para las Ecuaciones Elípticas**. Editorial MIR-Moscú, 1979.
- Schatzman, M. **Numerical Analysis**, A Mathematical Introduction. Oxford University Press, 2002.
- Scheid, F.; Di Costanzo, R.E. **Métodos Numéricos**, 2da Edición. McGraw-Hill, 1991.
- Shampine, L. F.; Watts, H. A.; Davenport, S. M. "Solving Non-Stiff Ordinary Differential Equations - The State of the Art". **SANDIA** Laboratories, Report No. SAND75-0182, 1975. **SIAM Review**, Vol.**18**, No.3, pp. 376-411, (1976).
- Stewart, G. W. **Introduction to Matrix Computations.** Academic Press (New York), 1973.
- Stoer, J.; Bulirsch, R. **Introduction to Numerical Analysis**. Springer-Verlag, 1980.
- Szidarovszky, F.; Yakowitz, S. **Principles and Procedures of Numerical Analysis**. Plenum Press, 1978.
- Taylor, C.; Hughes, T. G. **Finite Element Programming of the Navier-Stokes Equations**. Pineridge Press, 1981.
- Thomas, J. W. **Numerical Partial Differential Equations: Finite Difference Method**. Springer Science+Business Media (New York), 1995.
- Versteeg, H. K.; Malalasekera, W. **An Introduction to Computational Fluid Dynamics: The Finite Volume Method**. Pearson Education, 1995. Second Edition, 2007.
- Wood, D. J.; Charles, C. O. A. "Hydraulic Network Analysis Using Linear Theory". **Journal of The Hydraulics Division**, ASCE, Vol.**98**, No.HY7, July (1972).
- Zienkiewicz, O. C.; Taylor, R. L.; Nithiarasu, P. **The Finite Element Method for Fluid Dynamics**, Sixth Edition. Elsevier - Butterworth-Heinemann (Boston, MA), 2005.

ACERCA DEL AUTOR

Nació en Valencia, Edo. Carabobo, Venezuela, el 11 de junio de 1959. Graduado USB Ingeniero Mecánico 1982, USB Magister en Ingeniería Mecánica 1988, UPM-ETSII Doctor Ingeniero Industrial 2003 (Cum Laude). Profesor Titular de la Universidad Simón Bolívar (USB), Departamento de Mecánica, Sept/1985 - Ene/2011 (jubilado). Ha dictado los cursos: Mecánica de Fluidos I, II & III, Mecánica Computacional I & II, Mecánica de Medios Continuos, Métodos Numéricos, Mecánica de Fluidos Avanzada, etc. Trabajó en prestigiosas empresas como: Vepica, Inelectra, Intevep (PDVSA). Tiene en su haber más de 80 publicaciones entre libros, artículos en revistas arbitradas y presentaciones en congresos o conferencias y publicaciones académicas. Actualmente vive retirado en Madrid, España.

Enlaces:

E-Mails: agrana@usb.ve granados.al@gmail.com
https://www.researchgate.net/profile/Andres-Granados/publications
https://usb.academia.edu/AndrésGranados