

Hans C. Müller S.C.

Una Introducción al Análisis Numérico

Hans C. Müller S.C.

Una Introducción al Análisis Numérico

Con 55 figuras

Hans C. Müller Santa Cruz
Departamento de Matemáticas
Facultad de Ciencias y Tecnología
Universidad Mayor de San Simón
Casilla 992, Cochabamba, Bolivia
e-mail hans@mat.umss.bo

Prólogo

La Facultad de Ciencias y Tecnología tiene 17 años de vida, período que ha sido fructífero en el ámbito universitario, porque se han creado carreras tecnológicas, como científicas, haciendo énfasis a una investigación científica seria como un mecanismo de garantizar una excelencia académica. La Carrera de Matemáticas, una de nuestras Carreras de reciente creación, está inscrita dentro este marco.

Ahora bien, la Universidad Mayor de San Simón consciente de esta situación, ha conseguido convenios de cooperación internacional, como es el caso del Programa MEMI, Programa de Mejoramiento de la Enseñanza de la Matemática y la Informática. De los objetivos principales de este programa es fortalecer el área de las matemáticas, incentivar la difusión de las diferentes ramas de las matemáticas en el medio universitario y fuera de éste. La Universidad y sus autoridades, dentro de sus políticas académicas y de investigación, han tenido la visión de conseguir los servicios de los mejores profesionales en esta disciplina y Hans Müller es uno de ellos.

El autor del libro, Hans Müller Santa Cruz, es un joven matemático boliviano que ha regresado a su país para compartir los conocimientos adquiridos en la Universidad de Ginebra, Suiza. Actualmente es el Coordinador del Programa Magister en Matemáticas, programa implementado de manera conjunta entre la Universidad Mayor de San Simón y la Universidad Católica del Norte, Chile. Ha dictado un curso de Elementos Finitos, destinado a los docentes en nuestra superior Casa de Estudios; en La Paz, bajo invitación ha impartido cursos tutoriales en la Academia de Ciencias en temas relacionados a las matemáticas aplicadas. El y otros profesionales de su área, están transformando la manera de hacer matemáticas en la Facultad de Ciencias y Tecnología.

Los tópicos tratados en este libro están muy relacionados con los cambios estructurales que ha ocasionado la introducción masiva de sistemas informáticos. En efecto, la utilización masiva del computador ha permitido al Hombre efectuar cálculos que en otra época hubiese sido impensable realizarlos. En la actualidad es muy corriente simular numéricamente experiencias, que en laboratorio son muy complicadas, costosas o peligrosas, o simplemente imposible de experimentarlas. Los problemas de optimización son abordables gracias a la utilización del ordenador y no faltan situaciones en las cuales el uso de estos dispositivos de cálculo, no solamente son de gran ayuda, sino indispensables. El Análisis Numérico es la rama de las Matemáticas, cuyo campo de acción es el estudio y análisis de los diferentes algoritmos

y métodos numéricos que se utilizan para resolver los problemas mediante computadoras. El libro “Una Intruducción al Análisis Numérico” presenta los temas básicos del Análisis Numérico de una manera rigurosa, permitiendo que el lector pueda adquirir los conocimientos matemáticos necesarios para profundizar en tópicos más especializados o simplemente pueda concebir e implementar métodos numéricos de una manera correcta y óptima.

Finalmente, mi esperanza es que este libro sea el inicio de una larga serie de otras publicaciones de alto nivel que ofrezca el autor y su unidad académica.

Cochabamba, septiembre de 1996

Ing. Alberto Rodríguez Méndez
Rector de la Universidad Mayor de San Simón

Prefacio

Este libro nace ante el vacío existente de una bibliografía en español que trate los temas capitales del Análisis Numérico. El nombre que lleva, “Una Introducción al Análisis Numérico”, se debe esencialmente al carácter que deseo que tenga este libro.

El Análisis Numérico es una disciplina de las Matemáticas en gran crecimiento gracias a la utilización masiva de medios informáticos. Día que pasa es más corriente el tratamiento numérico en las Ciencias, como en la Tecnología; el modelaje, la simulación numérica son moneda corriente. Ahora bien, toda persona que pretenda tener como herramienta de trabajo, los métodos numéricos, debe conocer los tópicos introductorios del Análisis Numérico que son: Sistemas Lineales, Interpolación, Resolución de Ecuaciones no Lineales, Cálculo de Valores Propios y Solución Numérica de Ecuaciones Diferenciales, porque tarde o temprano se topará con alguno de estos temas.

Siguiendo la línea trazada por este libro, éste contiene siete capítulos: el primero de carácter introductorio, donde se da los conceptos básicos de error y estabilidad, seguido de un ejemplo mostrando que la aritmética del punto flotante no es un impedimento para efectuar cálculos de precisión arbitraria; el segundo capítulo trata sobre los problemas lineales mas comunes y los métodos de solución de estos; el tercer capítulo aborda el tema de interpolación numérica y extrapolación, introduciendo el estudio de los *splines* cúbicos; el capítulo IV analiza los problemas no lineales y los métodos mas eficientes de resolución de estos; en el capítulo V se estudia el problema de valores propios y la implementación de métodos numéricos para el cálculo de valores propios; el capítulo sexto trata de la integración numérica y la transformada rápida de Fourier y finalmente el capítulo VII estudia los problemas diferenciales y los métodos numéricos de resolución mas usuales de estos problemas.

Prácticamente el contenido de este libro ven los estudiantes de segundo año de las Carreras de Matemáticas e Informática de la Universidad de Ginebra, Suiza, Universidad en la cual he sido formado. El pre-requisito para un buen aprovechamiento de este libro es conocer bien los principios básicos del Análisis Real y Complejo, como también tener una buena base de Algebra Lineal. Por consiguiente, este libro está destinado a estudiantes universitarios que siguen la asignatura de Análisis Numérico, como así mismo toda persona interesada en Análisis Numérico que tenga los pre-requisitos y que desea cimentar sus conocimientos en esta disciplina. Este libro puede

ser utilizado como texto base o bien como complemento bibliográfico.

Debo agradecer a mis profesores de la Universidad de Ginebra, E. Hairer y G. Wanner cuyos cursos han servido de base para la elaboración de este libro. Por otro lado, sin el apoyo en material de trabajo del Programa MEMI, Programa de Mejoramiento de la Enseñanza de las Matemáticas e Informática, de la Universidad Mayor de San Simón este libro no habría existido. Así mismo agradezco a mi familia, mis colegas y amigos que siguieron con interés la elaboración de este libro.

El libro ha sido transcrito en \TeX y las gráficas realizadas en las subrutinas gráficas **FORTTRAN** que G. Wanner me las cedió muy gentilmente. La transcripción, como la ejecución de los programas en sido realizados sobre una WorkStation HP-9000.

Posiblemente este libro contenga muchos errores, me gustaría que los hagan conocer, para que en una segunda edición estos sean corregidos.

Octubre, 1995

Hans C. Müller S.C.

Contenido

I. Preliminares

I.1 Algoritmos	2
Ejercicios	6
I.2 Estabilidad	8
Ejercicios	14
I.3 Un ejemplo: Cálculo de PI	15

II. Sistemas Lineales

II.1 Condición del Problema Lineal	25
Normas de Vectores y Matrices	25
La Condición de una Matriz	29
Ejercicios	33
II.2 Métodos Directos	35
El Algoritmo de Gauss	36
El Algoritmo de Cholesky	43
Ejercicios	47
II.3 Métodos Iterativos	48
Métodos de Jacobi y Gauss-Seidel	48
El Teorema de Perron-Frobenius	52
Método de Sobrerelajación SOR	56
Estudio de un Problema Modelo	59
Ejercicios	64
II.4 Métodos Minimizantes	66
Método del Gradiente	68
Método del Gradiente Conjugado	69
Polinomios de Chebichef	73
Método del Gradiente Conjugado Precondicionado	75
Resultados Numéricos	78
Ejercicios	81
II.5 Mínimos Cuadrados	83
La descomposición QR	87
La Pseudo-Inversa de una Matriz	92
Error del Método de los Mínimos Cuadrados	96
Ejercicios	101

III. Interpolación

III.1 Interpolación de Lagrange	104
Bases Teóricas	104
Construcción del Polinomio de Interpolación	106
El Error de Interpolación	111
Polinomios de Chebichef	113
Estudio de los Errores de Redondeo	115
Convergencia de la Interpolación	119
Ejercicios	129
III.2 Splines Cúbicos	131
Construcción del Spline Interpolante	133
El Error de la Aproximación Spline	136
Aplicación de Spline	142
Ejercicios	143
III.3 Extrapolación	145
Ejercicios	150

IV. Ecuaciones No Lineales

IV.1 Ecuaciones Polinomiales	152
Ecuaciones Resolubles por Radicales	152
Ecuaciones No Resolubles por Radicales	155
Localización de Ceros	155
Método de Newton	157
Sucesiones de Sturm	159
Ejercicios	161
IV.2 Métodos Iterativos	163
Posición del Problema	163
Método de la Falsa Posición	165
Sistema de Ecuaciones	168
Un Método Iterativo Simple	169
Ejercicios	173
IV.3 Método de Newton	174
El Teorema de Newton-Misovski	179
Método de Newton Simplificado	184
Método de Newton con Relajación	193
Aproximación de Broyden	197
Ejercicios	199
IV.4 Método de Gauss Newton	203
Convergencia del Método de Gauss-Newton	204
Modificaciones del Método de Gauss-Newton	207
El Método de Levenber-Marquandt	210
Ejercicios	211

V. Cálculo de Valores Propios

V.1 Teoría Clásica y Condición del Problema	214
La Condición del Problema a Valores Propios	217
Ejercicios	221
V.2 Determinación de Valores Propios	223
El Método de la Potencia	223
Formas Tridiagonales y Formas de Hessenberg	227
Teorema de Sturm y el Algoritmo de la Bisección	229
Generalización del Método de la Potencia	233
El Método QR	237
Ejercicios	241

VI. Integración Numérica

VI.1 Bases Teóricas	244
Fórmulas de Cuadratura	248
El Orden de una Fórmula de Cuadratura	249
Estimación del Error	250
Ejercicios	256
VI.2 Cuadraturas de Orden Elevado	258
Polinomios Ortogonales	259
Los Polinomios de Legendre	263
Las Fórmulas de Cuadratura de Gauss	264
Ejercicios	267
VI.3 Implementación Numérica	269
Tratamiento de Singularidades	273
Ejercicios	282
VI.4 Transformación de Fourier	284
Estudio del Error	287
Interpolación Trigonométrica	288
Transformación Rápida de Fourier FFT	290
Aplicaciones de FFT	292
Ejercicios	293

VII. Ecuaciones Diferenciales

VII.1 Generalidades	296
Teoremas de Existencia y Unicidad	297
Problemas con Valores en la Frontera	300
Diferenciabilidad respecto a los Valores Iniciales	300
<i>Simple Shooting</i>	303
<i>Shooting</i> Múltiple	307
Ejercicios	311

VII.2 Método de Euler	313
Efectos de los Errores de Redondeo	317
Estabilidad del Método de Euler	319
Método de Euler Impícito	321
Ejercicios	322
VII.3 Métodos de Runge-Kutta	323
Construcción de un Método de Orden 4	327
Métodos Encajonados	330
Soluciones Continuas	333
Convergencia de los Métodos de Runge-Kutta	335
Experiencias Numéricas	338
Ejercicios	340
VII.3 Métodos Multipasos	341
Métodos de Adams Explícitos	341
Métodos de Adams Implícitos	343
Métodos Predictor-Corrector	344
Métodos <i>BDF</i>	345
Estudio del Error Local	346
Estabilidad	348
Convergencia de los Métodos Multipaso	350
Ejercicios	353
Bibliografía	355
Índice de Símbolos	359
Índice Alfabético	361

Capítulo I

Preliminares

Con la utilización masiva de sistemas informáticos en la resolución de problemas a todo nivel, el Análisis Numérico ha tenido un gran desarrollo en las últimas décadas, como una rama integrante de las Matemáticas. En este primer capítulo se expondrá las nociones de base que sustentan el Análisis Numérico, para ser utilizadas en los capítulos posteriores.

La primera sección estará dedicada a estudiar los algoritmos como una composición de operaciones elementales, partiendo de un dispositivo ideal de cálculo. Se analizará las nociones de eficiencia y error de un algoritmo.

En la segunda sección se analizará el problema algorítmico a partir de los dispositivos materiales con que se cuenta en la actualidad, es decir ordenadores o computadoras. En esta sección se estudiará la representación en punto flotante de un número real y su redondeo en la computadora. Se introducirá la noción de precisión de una computadora y la relación de ésta con el concepto de estabilidad, en sus diferentes versiones, de un algoritmo.

En la tercera y última sección de este capítulo, se verá que las limitaciones impuestas por la representación en punto flotante, no es una restricción para calcular con la precisión que se requiera. Prueba de esto, π será calculado, como un ejemplo ilustrativo, con 10000 decimales de precisión.

I.1 Algoritmos

En esta sección se supondrá que se cuenta con un dispositivo ideal de cálculo, es decir una computadora de precisión exacta, situación que no sucede en la realidad. El cuerpo de base será \mathbb{R} , a menos que se especifique lo contrario. Este dispositivo está provisto de ciertas operaciones cuya acción se efectúa en un tiempo finito, por ejemplo es capaz de realizar las 4 operaciones aritméticas elementales, más algunas como la radicación, la exponenciación y las otras funciones elementales que se estudian en un curso de Cálculo I. Por consiguiente, se tiene la primera definición.

Definición I.1.1.- Una operación elemental es una operación matemática o función cuya acción se la efectúa en un tiempo finito y que además el dispositivo ideal la puede realizar.

Inicialmente se supondrá que las cuatro operaciones aritméticas como: la adición, la multiplicación, la sustracción y la división son posibles en este dispositivo de cálculo. De esta manera es posible evaluar toda función polinomial, toda función racional, en una cantidad finita de pasos o composición de operaciones elementales.

Definición I.1.2.- Un algoritmo es una sucesión finita de operaciones elementales. Si se denota por f_i una operación elemental, un algoritmo es la composición de n operaciones elementales, es decir

$$f = f_n \circ f_{n-1} \circ \cdots \circ f_2 \circ f_1. \quad (\text{I.1.1})$$

Por otro lado, el dispositivo de cálculo con el que se cuenta tiene la finalidad de evaluar la o las soluciones de un problema matemático, siempre que sea posible hacerlo. Ahora bien, lo que cuenta en este estudio es el resultado, de donde la:

Definición I.1.3.- Un problema es darse un dato $x \in \mathbb{R}^n$ y obtener un resultado $\mathcal{P}(x) \in \mathbb{R}$.

En consecuencia, son problemas todas las funciones cuyo dominio es un espacio vectorial real de dimensión finita, cuya imagen está incluida en la recta real. Una función cuya imagen está incluida en un espacio de dimensión más grande que 1, puede ser interpretada como un conjunto de problemas, donde cada problema es la función proyección correspondiente. Las ecuaciones pueden ser vistas como problemas, pero antes aclarando cual de las soluciones se está tomando.

Es necesario recalcar que problema y algoritmo son conceptos diferentes, aunque de alguna manera ligados. Considérese el problema siguiente,

$$\mathcal{P}(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0. \quad (\text{I.1.2})$$

$\mathcal{P}(x)$ puede ser obtenido de muchas maneras, en particular utilizando los 2 siguientes algoritmos: El primero consiste en evaluar el polinomio (I.1.2) tal como está definido, con la convención que:

$$\begin{aligned} x^1 &= x; \\ x^n &= x \cdot x^{n-1}, \quad \text{para } n \geq 2. \end{aligned} \quad (\text{I.1.3})$$

La segunda manera de evaluar el polinomio consiste en utilizar el algoritmo de Hörner, el cual está dado por:

$$\begin{aligned} q_0(x) &= a_n, \\ q_1(x) &= q_0(x)x + a_{n-1}, \\ q_2(x) &= q_1(x)x + a_{n-2}, \\ &\vdots \\ q_n(x) &= q_{n-1}(x)x + a_0. \end{aligned} \quad (\text{I.1.4})$$

Como puede observarse ambos algoritmos evalúan el polinomio $\mathcal{P}(x)$, sin embargo en el primero se requiere: $1 + \cdots + n - 1$ multiplicaciones para evaluar las potencias, n multiplicaciones para evaluar los términos de la forma $a_i x^i$ y finalmente n adiciones, lo cual hace un total de

$$\frac{n(n+3)}{2} \text{ operaciones elementales.} \quad (\text{I.1.5})$$

El algoritmo de Hörner requiere a cada paso de una multiplicación y una adición lo cual es igual a

$$2n \text{ operaciones elementales.} \quad (\text{I.2.6})$$

Por lo tanto, puede observarse claramente que el algoritmo de Hörner es más eficiente que el primero, pues la implementación de éste efectúa menos operaciones elementales.

El concepto de eficiencia de un algoritmo está ligado por consiguiente al costo en operaciones elementales que se requiere para ejecutar un algoritmo. Ahora bien, en la realidad una computadora requiere menos tiempo para evaluar una adición que para una multiplicación. Cada operación elemental toma cierto tiempo en efectuarse, que en general depende de la computadora y el lenguaje en el que está escrito el programa. Es

por eso, que es más conveniente medir la eficiencia en términos de tiempo ejecutado, en lugar del número de operaciones elementales ejecutadas. Con lo argumentado se puede formular una primera definición de eficiencia.

Definición I.1.4.- El costo del algoritmo $f_n \circ \dots \circ f_1$, está dado por

$$C = m_1 + m_2 + \dots + m_n, \quad (\text{I.1.7})$$

donde m_i es el tiempo de ejecución de f_i . Si C_1 y C_2 son los costos en tiempo de dos algoritmos que resuelven el mismo problema, se dirá que el primer algoritmo es más eficiente que el segundo si $C_1 \leq C_2$.

Tal como se dijo al inicio de esta sección, el dispositivo ideal con el que se cuenta, puede efectuar una cantidad finita de operaciones elementales. Suponiendo nuevamente, que solamente se cuenta con las cuatro operaciones aritméticas, las únicas funciones que pueden evaluarse utilizando un algoritmo son las funciones polinomiales y racionales. Ahora bien, existe una cantidad ilimitada de funciones y se puede mostrar que no existe ningún algoritmo que sea la composición de adiciones, multiplicaciones, sustracciones o divisiones, que permita calcular una raíz cuadrada; evaluar funciones trigonométricas, exponenciales o logarítmicas. Sin embargo, existen procedimientos matemáticos que permiten aproximar estas funciones de manera arbitraria. Los más comúnmente utilizados: son las series de Taylor, las fracciones continuas y algunos métodos iterativos. Todos estos métodos están sustentados en las nociones de límite, de continuidad, derivabilidad dados en los cursos de Cálculo y Análisis.

A continuación se verá un ejemplo ilustrativo, donde se introducirá la noción del error de truncación. Considérese la función exponencial, cuyo desarrollo en serie de Taylor está dada por

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}. \quad (\text{I.1.8})$$

Es evidente que e^x no puede evaluarse en un número finito de pasos, para un x dado, sin embargo en lugar de evaluar e^x , uno se puede contentar evaluando una cantidad finita de términos de la serie de Taylor, es decir

$$p(x) = \sum_{k=0}^n \frac{x^k}{k!}, \quad (\text{I.1.9})$$

para un n fijado de antemano. La diferencia entre e^x y $p(x)$ es el error de truncación de e^x respecto a $p(x)$. Este error de truncación es del orden $\mathcal{O}(x^{n+1})$ cuando x tiende a 0. El nombre de truncación proviene del hecho que para evaluar e^x se ha despreciado una parte de la serie. Por consiguiente, el error de truncación puede definirse como:

Definición I.1.5.- Sean $\mathcal{P}(x)$, $\mathcal{P}'(x)$ dos problemas. El error de truncación de $\mathcal{P}(x)$, respecto de $\mathcal{P}'(x)$ está dado por

$$\mathcal{P}(x) - \mathcal{P}'(x). \quad (\text{I.1.10})$$

El nombre que tiene este error, como se dijo más arriba, es debido a que la serie de Taylor es truncada a un número finito de términos. No obstante, existe una gran diversidad de métodos que aproximan a la solución del problema utilizando otro tipo de argumentos, en este caso es más conveniente utilizar el nombre de error de aproximación o error del método; de todas formas es cuestión de gusto.

El concepto de eficiencia ha sido definido para aquellos problemas donde se puede encontrar la solución mediante un algoritmo. Para aquellos problemas, donde no es posible encontrar un algoritmo que de la solución y suponiendo que es posible aproximar la solución de manera arbitraria mediante un método algorítmico, la eficiencia está ligada al costo en operaciones, como también al error del método. En esta situación la eficiencia es un concepto más subjetivo que depende de alguna manera del usuario, pues existen problemas donde el error de aproximación debe ser lo más pequeño posible, y otros problemas donde la exactitud no es un requisito primordial.

Ejemplos

1.- Considérese el problema, determinar π . Utilizando la identidad

$$\arctan 1 = \frac{\pi}{4}$$

y que la serie de Taylor en el origen está dada por

$$\arctan x = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1}, \quad (\text{I.1.11})$$

se puede obtener un método que aproxime π , pero el problema de este método es su convergencia demasiado lenta; es mas, para $x = 1$ la serie (I.1.11) no es absolutamente convergente.

Otro método que permitiría aproximar π , consiste en utilizar el hecho que

$$\arccos(1/2) = \frac{\pi}{3}$$

y desarrollando arccos en serie de Taylor, se obtiene un método cuya convergencia es mucho más rápida.

2.- Considérese el problema, determinar $\sqrt{2}$. La primera manera de determinar $\sqrt{2}$ es tomar un intervalo cuya extremidad inferior tenga un cuadrado inferior a 2 y la extremidad superior tenga un cuadrado más grande a 2. Se subdivide el intervalo en dos subintervalos de igual longitud, se elige aquel cuyas extremidades al cuadrado contengan 2. En la tabla siguiente se da algunas de las iteraciones de este método.

Iteración	Ext. Inferior	Ext. Superior
0	1.0	1.5
1	1.25	1.5
2	1.375	1.5
3	1.375	1.4375
4	1.40625	1.4375
5	1.40625	1.421875
16	1.41420745849609	1.41421508789063
17	1.41421127319336	1.41421508789063
18	1.41421318054199	1.41421508789063
19	1.41421318054199	1.41421413421631
20	1.41421318054199	1.41421365737915

La segunda posibilidad es utilizar la sucesión definida recursivamente por:

$$a_0 = 1,$$

$$a_{n+1} = \frac{1}{2}a_n + \frac{1}{a_n}.$$

Esta sucesión es convergente y se deja al lector la demostración para que verifique que el límite es $\sqrt{2}$. En la tabla siguiente se tiene los seis primeros términos de la sucesión. Efectuando unas cuantas iteraciones se obtiene:

Iteración	a_n
0	1.0
1	1.5
2	1.41666666666667
3	1.41421568627451
4	1.41421356237469
5	1.4142135623731

Puede observarse inmediatamente, que el segundo método para determinar $\sqrt{2}$ es más eficiente que el primer algoritmo.

Ejercicios

- Supóngase que, el dispositivo de cálculo, con el que se cuenta, puede efectuar la división con resto; es decir, para a, b enteros no negativos, con $b \neq 0$, el dispositivo calcula p, q satisfaciendo:

$$a = pb + r \quad \text{y} \quad 0 \leq r < b.$$

a) Implementar un algoritmo que calcule el máximo común divisor de a y b .

b) Utilizando el inciso precedente, implementar otro algoritmo que permita calcular $m, n \in \mathbb{Z}$, tales que

$$\text{mcd}(a, b) = ma + nb.$$

- c) Estudiar la situación más desfavorable, aquélla donde el costo en operaciones es el más alto. Deducir una estimación de la eficiencia del algoritmo.
- 2.- Suponiendo que, el dispositivo de cálculo puede efectuar la división con resto, con la particularidad siguiente:

$$a = pb + r \quad \text{y} \quad -\frac{|b|}{2} < r \leq \frac{|b|}{2}.$$

- a) Mostrar que el algoritmo implementado en el ejercicio 1a), puede implementarse para esta nueva división con resto.
- b) Verificar que, el nuevo algoritmo es más eficiente que aquél con división con resto normal. Encontrar una estimación de costo del algoritmo.
- 3.- Para el polinomio $p(x) = a_0 + a_1x + \cdots + a_nx^n$, el algoritmo de Hörner (I.1.9) está definido por:

$$\begin{aligned} b_n &= a_n; \\ b_i &= a_i + x_0b_{i+1}, \quad i = n-1, \dots, 1, 0. \end{aligned}$$

Se plantea $q(x) = b_1 + xb_2 + \cdots + x^{n-1}b_n$.

- a) Mostrar que $p'(x_0) = q(x_0)$. Verificar que $p(x) = b_0 + (x - x_0)q(x)$.
- b) Generalizar el algoritmo de Hörner, de tal forma que se pueda calcular $p(x_0)$ y $p'(x_0)$ al mismo tiempo.
- 4.- Una regla y compás constituyen un dispositivo de cálculo. Supóngase que se conocen dos puntos O y P tales que \overline{OP} sea de longitud 1.
- a) Mostrar que, además de las 4 operaciones aritméticas elementales, la radicación puede obtenerse a partir de un número finito de manipulaciones de regla y compás.
- b) Construir $\sqrt{1 + \sqrt{10}}$.
- c) Intentar construir $\sqrt[3]{2}$. ¿Es posible?

I.2 Estabilidad

En esta sección, a diferencia de la precedente, se analizará el problema de los algoritmos y métodos numéricos desde el punto de vista de un dispositivo material real, más conocido como computadora. Para comenzar existen esencialmente dos tipos de números en la aritmética de un ordenador: el tipo entero cuya aritmética es igual a la de los números enteros, con la diferencia que el tipo entero de un ordenador es un conjunto finito, motivo por el cual se tiene que tener cuidado con los *overflows* en las operaciones aritméticas; el segundo tipo de número es el real, en sus diferentes versiones, la idea de este tipo de número es la representación de un número real en punto flotante, es decir todo número real no nulo x se puede escribir de manera única como

$$x = a \cdot 10^b, \quad \text{donde } |a| < 1, \ b \in \mathbb{Z}; \quad (\text{I.2.1})$$

a se llama mantisa y b el exponente. Los números reales solo tienen una representación parcial en el tipo real de un ordenador por limitaciones físicas del dispositivo, por consiguiente un número real estará en una clase de equivalencia de tipo de número real. Esta representación de número real está dada por la precisión de la computadora.

Cuando se trabaja sobre un computador, se tiene a disposición un número finito de cifras, por ejemplo l , para la mantisa, Si \bar{a} denota el redondeado de a , se trabaja en realidad con

$$arr(x) = \bar{a} \cdot 10^b \quad (\text{I.2.2})$$

en lugar de x . De esta manera, el número $\sqrt{2} = 1.414213562\dots$ está representado por

$$arr(\sqrt{2}) = 0.141421356 \cdot 10^1,$$

si se calcula con $l = 8$ cifras en base 10.

Como se dijo más arriba, el redondeo está determinado por la precisión de la computadora. Esta precisión está dada por un número dado por la:

Definición I.2.1.- Se denota por eps , el número positivo más pequeño tal que

$$arr(1 + eps) > 1. \quad (\text{I.2.3})$$

Este número eps llamado precisión de la computadora está dado por los siguientes hechos: si los cálculos se hacen en base 10 y con l cifras en la

mantisa, se tiene

$$\begin{aligned} arr(0.\underbrace{10\dots 0}_{l}49\dots \cdot 10^1) &= 1, \\ arr(0.\underbrace{10\dots 0}_{l}50\dots \cdot 10^1) &= \underbrace{.10\dots 1}_{l} \cdot 10^1 > 1, \end{aligned}$$

deduciendose, por consiguiente

$$eps = 5 \cdot 10^{-l}; \quad (\text{I.2.4})$$

si se realiza el mismo cálculo en base 2, como todas las computadoras lo hacen, se obtiene

$$eps = 2^{-l}. \quad (\text{I.2.5})$$

Para el FORTRAN sobre una HP-9000, se tiene:

$$\begin{aligned} \text{REAL*4,} \quad eps &= 2^{-24} \approx 5.96 \cdot 10^{-8}; \\ \text{REAL*8,} \quad eps &= 2^{-55} \approx 1.39 \cdot 10^{-17}; \\ \text{REAL*16,} \quad eps &= 2^{-113} \approx 9.63 \cdot 10^{-35}. \end{aligned}$$

Ahora bien, un número real y su representación en punto flotante en una computadora están relacionados por el:

Teorema I.2.2.- Para $x \neq 0$, se tiene

$$\frac{|arr(x) - x|}{|x|} \leq eps. \quad (\text{I.2.6})$$

Demostración.- Sea $x = a \cdot 10^b$ y $arr(x) = \bar{a} \cdot 10^b$. Si se redondea a l cifras significativas se tiene

$$|\bar{a} - a| \leq 5 \cdot 10^{-l-1},$$

de donde

$$\frac{|arr(x) - x|}{|x|} = \frac{|\bar{a} - a| \cdot 10^b}{|a| \cdot 10^b} \leq \frac{5 \cdot 10^{-l-1}}{10^{-1}} = 5 \cdot 10^{-l} = eps$$

por que $|a| \geq 1/10$. □

La estimación dada por (I.2.6) puede ser escrita bajo la forma

$$arr(x) = x(1 + \epsilon), \quad \text{donde} \quad |\epsilon| \leq eps. \quad (\text{I.2.7})$$

Como se verá más adelante, la relación (I.2.7) es la base fundamental para todo estudio de los errores de redondeo.

Cuando se trabaja en un computador, se tiene que ser cuidadoso en la solución de los problemas, pues ya no se resuelve con los datos exactos, si no con los datos redondeados. Considérese la ecuación de segundo grado

$$x^2 - 2\sqrt{2}x + 2 = 0,$$

cuya solución $x = \sqrt{2}$ es un cero de multiplicidad 2. Resolviendo en simple precisión, la computadora da un mensaje de error, debido a que

$$\begin{aligned} arr(\sqrt{2}) &= 0.141421 \cdot 10, \\ arr(arr(\sqrt{2})^2 - 2) &= -0.119209 \cdot 10^{-6}. \end{aligned}$$

Como puede observarse, la manipulación de ciertos problemas utilizando la aritmética del punto flotante puede ocasionar grandes dificultades, como en el ejemplo anterior. Es por eso necesario introducir el concepto de condición de un problema, el cual está dado en la:

Definición I.2.3.- Sea $\mathcal{P}(x)$ un problema dado por $\mathcal{P} : \mathbb{R}^n \rightarrow \mathbb{R}$. La condición κ del problema \mathcal{P} es el número mas pequeño positivo κ , tal que

$$\frac{|\bar{x}_i - x_i|}{|x_i|} \leq eps \quad \Rightarrow \quad \frac{|\mathcal{P}(\bar{x}) - \mathcal{P}(x)|}{|\mathcal{P}(x)|} \leq \kappa eps. \quad (\text{I.2.8})$$

Se dice que el problema \mathcal{P} es bien condicionado si κ no es demasiado grande, sino el problema es mal condicionado.

En la definición, eps representa un número pequeño. Si eps es la precisión de la computadora entonces \bar{x}_i puede ser interpretado como el redondeo de x_i . Por otro lado, es necesario resaltar que κ depende solamente del problema, de x y no así del algoritmo con el que se calcula $\mathcal{P}(x)$. Para comprender más sobre la condición de un problema se analizará los siguientes dos ejemplos.

Ejemplos

1.- Multiplicación de dos números reales. Sean x_1 y x_2 reales, considérese el problema calcular $\mathcal{P}(x_1, x_2) = x_1 \cdot x_2$. Para los dos valores perturbados

$$\bar{x}_1 = x_1(1 + \epsilon_1), \quad \bar{x}_2 = x_2(1 + \epsilon_2), \quad |\epsilon_i| \leq eps; \quad (\text{I.2.9})$$

se obtiene

$$\frac{\bar{x}_1 \cdot \bar{x}_2 - x_1 \cdot x_2}{x_1 \cdot x_2} = (1 + \epsilon_1)(1 + \epsilon_2) - 1 = \epsilon_1 + \epsilon_2 + \epsilon_1 \cdot \epsilon_2.$$

Puesto que eps es un número pequeño, el producto $\epsilon_1 \cdot \epsilon_2$ es despreciable respecto a $|\epsilon_1| + |\epsilon_2|$, de donde

$$\left| \frac{\bar{x}_1 \cdot \bar{x}_2 - x_1 \cdot x_2}{x_1 \cdot x_2} \right| \leq 2 \cdot eps. \quad (\text{I.2.10})$$

Por consiguiente, $\kappa = 2$. El problema es bien condicionado.

2.- Adición de números reales. Para el problema $\mathcal{P}(x_1, x_2) = x_1 + x_2$, por analogía al ejemplo precedente, se obtiene

$$\left| \frac{(\bar{x}_1 + \bar{x}_2) - (x_1 + x_2)}{x_1 + x_2} \right| = \left| \frac{x_1 \epsilon_1 - x_2 \epsilon_2}{x_1 + x_2} \right| \leq \frac{|x_1| + |x_2|}{|x_1 + x_2|} \epsilon_{ps}. \quad (\text{I.2.11})$$

Si x_1 y x_2 son de signos iguales, se tiene $\kappa = 1$, de donde el problema es bien condicionado.

Pero, si $x_1 \approx -x_2$, la condición $\kappa = \frac{|x_1| + |x_2|}{|x_1 + x_2|}$ se convierte en una cantidad muy grande. Motivo por el cual el problema está mal condicionado. Para mejor ilustrar el efecto de condición muy grande, considérese el siguiente ejemplo numérico.

$$x_1 = \frac{1}{51}, \quad x_2 = -\frac{1}{52}, \quad \text{para el cual} \quad \kappa \approx \frac{2/50}{(1/50)^2} = 100.$$

Realizando el cálculo con 3 cifras significativas en base 10, se obtiene $\bar{x}_1 = .196 \cdot 10^{-1}$, $\bar{x}_2 = -.192 \cdot 10^{-1}$ y $\bar{x}_1 + \bar{x}_2 = .400 \cdot 10^{-1}$. Como las dos primeras cifras son las mismas para \bar{x}_1 y x_2 , la adición las ha hecho desaparecer y no hay más que una cifra que es significativa. El resultado exacto es $1/(51 \cdot 52) = 0.377 \cdot 10^{-3}$.

Respecto a la definición de condición, se debe observar dos situaciones. La primera, si uno de los $x_i = 0$, entonces se tiene $\bar{x}_i = 0$; la segunda sucede cuando $\mathcal{P}(x) = 0$, la condición se la calcula pasando al límite.

Una vez definida la condición de un problema, el siguiente paso es ver la incidencia que tienen los errores de redondeo en la implementación de un algoritmo para resolver un problema dado. Tal como se dijo en la sección precedente, un algoritmo es una sucesión finita de operaciones elementales, es decir

$$\mathcal{P}(x) = f_n(f_{n-1}(\dots f_2(f_1(x)) \dots)). \quad (\text{I.2.12})$$

La amplificación del error, efectuando la operación f_i , está dada por la condición $\kappa(f_i)$. Por consiguiente:

Proposición I.2.4.- *El algoritmo que resuelve el problema dado por (I.2.12), tiene la estimación siguiente*

$$\kappa(\mathcal{P}) \leq \kappa(f_1) \cdot \kappa(f_2) \cdots \kappa(f_n). \quad (\text{I.2.13})$$

Demostración.- Por inducción sobre n . Para $n = 1$, el resultado es trivial. Supóngase, que es cierto para n , por lo tanto, si el problema es de la forma

$$\mathcal{P}(x) = f_{n+1}(f_n(f_{n-1}(\dots f_2(f_1(x)) \dots))),$$

puede escribirse como

$$\mathcal{P}(x) = f_{n+1}(\mathcal{P}'(x)),$$

utilizando la definición de condición se tiene

$$\begin{aligned} \frac{|\mathcal{P}'(x) - \mathcal{P}'(\bar{x})|}{|\mathcal{P}'(x)|} &\leq \kappa(\mathcal{P}') \epsilon ps \\ \Rightarrow \frac{|f_{n+1}(\mathcal{P}'(x)) - f_{n+1}(\mathcal{P}'(\bar{x}))|}{|f_{n+1}(\mathcal{P}'(x))|} &\leq \kappa(f_{n+1}) \kappa(\mathcal{P}') \epsilon ps. \end{aligned}$$

Finalmente utilizando la hipótesis de inducción, se tiene (I.2.13). \square

Definición I.2.5.- Un algoritmo es numéricamente estable, en el sentido de *forward analysis*, si

$$\kappa(f_1) \cdot \kappa(f_2) \cdots \kappa(f_n) \leq Const \cdot \kappa(\mathcal{P}) \quad (\text{I.2.14})$$

donde *Const* no es demasiado grande.

La fórmula (I.2.14) expresa el hecho de que la influencia de los errores de redondeo durante el cálculo de $\mathcal{P}(x)$ no es mucho más grande que la influencia de los errores en los datos, lo que en realidad es inevitable.

Ejemplo

Sea $x = 10^4$ y considérese el problema de calcular $1/(x(1+x))$. Se examinará los dos algoritmos siguientes: El primero definido por

$$\begin{array}{ccc} & x & \\ \nearrow & & \searrow \\ x & & x(x+1) \longrightarrow \frac{1}{x(x+1)}. \\ \searrow & & \nearrow \\ & x+1 & \end{array}$$

Las operaciones efectuadas son muy bien condicionadas, por consiguiente, este algoritmo es numéricamente estable.

El segundo algoritmo definido por

$$\begin{array}{ccc} & 1/x & \\ \nearrow & & \searrow \\ x & & \frac{1}{x} - \frac{1}{x+1} = \frac{1}{x(x+1)}. \\ \searrow & & \nearrow \\ & x+1 \longrightarrow 1/(x+1) & \end{array}$$

En este algoritmo, solamente las tres primeras operaciones son bien condicionadas. Sin embargo la última operación, la sustracción, es muy

mal condicionada, porque $1/x \approx 1/(x+1)$. Entonces, este algoritmo es numéricamente inestable.

La verificación, si un algoritmo es estable en el sentido de *forward analysis*, sobre todo si el número de operaciones es elevado, es a menudo muy compleja y difícil. Por esta razón, Wilkinson introdujo otra definición de la estabilidad de un algoritmo.

Definición I.2.6.- Un algoritmo para resolver el problema $\mathcal{P}(x)$ es numéricamente estable en el sentido de *backward analysis* si el resultado numérico \bar{y} puede ser interpretado como un resultado exacto para los datos perturbados \bar{x} , es decir $\bar{y} = \mathcal{P}(\bar{x})$, y si

$$\frac{|\bar{x}_i - x_i|}{|x_i|} \leq \text{Const} \cdot \text{eps}, \quad (\text{I.2.15})$$

donde *Const* no es demasiado grande y *eps* es la precisión de la computadora.

De la definición I.2.6 y (I.2.15) se deduce que, en el estudio de este tipo de estabilidad no se requiere conocer de antemano la condición del problema.

Ejemplo

Considérese el problema de calcular el producto escalar $x_1 \cdot x_2 + x_3 \cdot x_4$. Para calcular éste, se utiliza el algoritmo

$$(x_1, x_2, x_3, x_4) \begin{array}{l} \nearrow x_1 \cdot x_2 \\ \searrow x_1 \cdot x_2 \end{array} \begin{array}{l} \searrow \\ \nearrow \end{array} x_1 \cdot x_2 + x_3 \cdot x_4. \quad (\text{I.2.16})$$

El resultado numérico bajo la influencia de los errores de redondeo es

$$(x_1(1 + \epsilon_1) \cdot x_2(1 + \epsilon_2)(1 + \eta_1) + x_3(1 + \epsilon_3) \cdot x_4(1 + \epsilon_4)(1 + \eta_2))(1 + \eta_3),$$

donde $|\epsilon_i|, |\eta_j| \leq \text{eps}$. Este resultado es igual a $\bar{x}_1 \cdot \bar{x}_2 + \bar{x}_3 \cdot \bar{x}_4$, si se plantea

$$\begin{aligned} \bar{x}_1 &= x_1(1 + \epsilon_1)(1 + \eta_1), & \bar{x}_3 &= x_3(1 + \epsilon_3)(1 + \eta_2), \\ \bar{x}_2 &= x_2(1 + \epsilon_2)(1 + \eta_3), & \bar{x}_4 &= x_4(1 + \epsilon_4)(1 + \eta_3). \end{aligned}$$

Despreciando los productos de la forma $\eta_i \cdot \epsilon_j$, la relación (I.2.15) es satisfecha con *Const* = 2. Por lo tanto, el algoritmo (I.2.16) siempre es numéricamente estable en el sentido de *backward analysis*.

El ejemplo precedente muestra que un algoritmo puede ser estable, incluso si el problema está mal condicionado. En consecuencia, es necesario bien distinguir las nociones de estabilidad y condición.

Ejercicios

- 1.- ¿Cual es la condición de la sustracción de dos números?
- 2.- Determinar la condición del problema $\mathcal{P}(x_1, x_2) = x_1/x_2$ con $x_2 \neq 0$.
- 3.- Hallar la condición del cálculo del producto escalar

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i.$$

- 4.- Mostrar que el sistema lineal

$$\begin{cases} ax + by = 1 \\ cx + dy = 0 \end{cases}, \quad \text{con } ad \neq bc$$

es numéricamente estable en el sentido de *backward analysis*.

- 5.- Las raíces del polinomio $x^2 - 2px - q = 0$ son:

$$\lambda_1 = p + \sqrt{p^2 + q}, \quad \lambda_2 = p - \sqrt{p^2 + q}.$$

Mostrar que para $p > 0$, grande, y $q \geq 0$, muy pequeño, este algoritmo es numéricamente inestable. Utilizando la relación $\lambda_1 \lambda_2 = q$, encontrar un algoritmo que es numéricamente estable en el sentido de *backward analysis*.

I.3 Un ejemplo: Cálculo de Pi

En la sección precedente se ha visto que el dispositivo material que efectúa los cálculos numéricos tiene en general dos tipos de números: los enteros que se asemejan mucho a los enteros matemáticos y el tipo real que es una representación del número real. Tal como se mostró, ambos tipos de número presentan ciertas particularidades, por ejemplo son en cantidad finita, para el tipo real existe el redondeo de los números. Por consiguiente si se utiliza el tipo real para determinar π , se tendrá una precisión de 8 decimales; para doble precisión se obtendrá 16 decimales de precisión; para cuádruple precisión, que algunos lenguajes de programación cuentan, se obtendrá 32 decimales de precisión. Por lo tanto, la precisión para obtener π está determinada por el lenguaje de programación y en algunos casos por el tipo de computadora con que se cuenta.

En esta sección se pretende mostrar que estas limitaciones no constituyen necesariamente un impedimento para determinar un número, en particular π , con la precisión que se desee.

Una de las maneras más usuales de determinar π es utilizar el desarrollo en serie de Taylor en el origen de $\arctan x$, el cual está dado por

$$\arctan x = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1}. \quad (\text{I.3.1})$$

De (I.3.1) se deduce, para $x = 1$, que

$$\pi = 4 \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1}. \quad (\text{I.3.2})$$

Como se recalcó en la sección I.2, utilizar la serie (I.3.2) no es conveniente, por que la convergencia de esta serie es muy lenta. Utilizando la relación

$$\tan(\alpha + \beta) = \frac{\tan \alpha + \tan \beta}{1 - \tan \alpha \tan \beta},$$

una verificación inmediata conduce a

$$\frac{\pi}{4} = 12 \arctan \frac{1}{18} + 8 \arctan \frac{1}{57} - 5 \arctan \frac{1}{239}. \quad (\text{I.3.3})$$

Remplazando (I.3.1) en (I.3.3), da como resultado

$$\pi = \underbrace{48 \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \left(\frac{1}{18}\right)^{2k+1}}_A + \underbrace{32 \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \left(\frac{1}{57}\right)^{2k+1}}_B - \underbrace{20 \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \left(\frac{1}{239}\right)^{2k+1}}_C. \quad (\text{I.3.4})$$

Ahora bien, se desea obtener una representación de π en notación decimal con N decimales de precisión. Si cada serie del lado derecho de (I.3.4) es calculada con una precisión de $10^{-(N+1)}$ se tiene largamente la precisión deseada. Denotando por \hat{A} , el valor calculado de A ; \hat{B} , el valor calculado de B y \hat{C} , el valor calculado de C , se tiene:

$$\begin{aligned} \hat{A} &= 48 \sum_{k=0}^{M_A} \frac{(-1)^k}{2k+1} \left(\frac{1}{18}\right)^{2k+1}, \\ \hat{B} &= 32 \sum_{k=0}^{M_B} \frac{(-1)^k}{2k+1} \left(\frac{1}{57}\right)^{2k+1}, \\ \hat{C} &= 20 \sum_{k=0}^{M_C} \frac{(-1)^k}{2k+1} \left(\frac{1}{239}\right)^{2k+1}. \end{aligned}$$

De donde:

$$\begin{aligned} |\hat{A} - A| &= 48 \left| \sum_{k=M_A+1}^{\infty} \frac{(-1)^k}{2k+1} \left(\frac{1}{18}\right)^{2k+1} \right| \leq 48 \frac{(1/18)^{2M_A+3}}{1 - (1/18)^2}, \\ |\hat{B} - B| &= 32 \left| \sum_{k=M_B+1}^{\infty} \frac{(-1)^k}{2k+1} \left(\frac{1}{57}\right)^{2k+1} \right| \leq 32 \frac{(1/57)^{2M_B+3}}{1 - (1/57)^2}, \\ |\hat{C} - C| &= 20 \left| \sum_{k=M_C+1}^{\infty} \frac{(-1)^k}{2k+1} \left(\frac{1}{239}\right)^{2k+1} \right| \leq 20 \frac{(1/239)^{2M_C+3}}{1 - (1/239)^2}. \end{aligned}$$

Por hipótesis, se tiene por ejemplo que $|A - \hat{A}| \leq 10^{-(N+1)}$, por lo tanto para asegurar esta precisión es suficiente que

$$48 \frac{(1/18)^{2M_A+3}}{1 - (1/18)^2} \leq 10^{-(N+1)},$$

reemplazando el signo \leq por $=$, se tiene

$$48 \frac{(1/18)^{2M_A+3}}{1 - (1/18)^2} = 10^{-(N+1)}, \quad (\text{I.3.5})$$

obteniendo de esta manera

$$\log_{10}(48) - (2M_A + 3) \log_{10}(18) - \log_{10}(1 - (1/18)^2) = -(N + 1).$$

Despejando M_A , se obtiene

$$M_A = \frac{N + 1 + \log_{10}(48) - 3 \log_{10}(18) + (1/18)^2}{2 \log_{10}(18)}. \quad (\text{I.3.6})$$

Del mismo modo se llega a:

$$M_B = \frac{N + 1 + \log_{10}(32) - 3 \log_{10}(57) + (1/57)^2}{2 \log_{10}(57)}, \quad (\text{I.3.7})$$

$$M_C = \frac{N + 1 + \log_{10}(20) - 3 \log_{10}(239) + (1/239)^2}{2 \log_{10}(239)}. \quad (\text{I.3.8})$$

Una vez determinada la cantidad de términos en cada serie para calcular \hat{A} , \hat{B} y \hat{C} , es necesario definir una representación de los números reales y su respectiva aritmética que permita calcular \hat{A} , \hat{B} y \hat{C} con la precisión requerida.

Sea $b > 1$, entonces todo número positivo x se escribe de manera única como

$$x = \sum_{k=0}^{\infty} a_k \frac{1}{b^k}, \quad (\text{I.3.9})$$

donde $0 \leq a_k < b$ para $k \geq 1$. Suponiendo que $b = 10^m$, para obtener una representación de x con N decimales de precisión, es suficiente conocer a_k para $0 \leq k \leq N/m + 2$. Se suma 2 a la cota anterior por seguridad. Por lo tanto, para todo número positivo x , el redondeado \hat{x} con N cifras decimales de precisión puede escribirse como

$$\hat{x} = \sum_{k=0}^{N/m+2} \bar{a}_k / 10^{mk}, \quad (\text{I.3.10})$$

con $0 \leq \bar{a}_k < 10^m$ para $k \geq 1$. Comparando (I.3.9) con (I.3.10) se tiene que $a_k = \bar{a}_k$ para $0 \leq k \leq N/m + 1$ y $|a_k - \bar{a}_k| \leq 1$ para $k = N/m + 2$, deduciéndose de esta manera la:

Proposición I.3.1.- Sea x un número no negativo y \hat{x} su representación dada por (I.3.10), entonces

$$|\hat{x} - x| \leq 10^{-(N+2m)}. \quad (\text{I.3.11})$$

Con los elementos presentados, se está en la capacidad de formular un algoritmo que permita calcular \hat{A} , \hat{B} , y \hat{C} ; por razones obvias el algoritmo será el mismo. Por consiguiente, es suficiente dar el método que permita calcular \hat{A} con la precisión requerida. De la fórmula (I.3.4) se puede observar que las operaciones efectuadas son adiciones o sustracciones, divisiones por expresiones de la forma $(2k+1)$ y divisiones por 18 o 18^2 . La aritmética será consiguientemente definida en tres pasos o etapas.

El primer paso será definir la division por un entero p . Sea x un real positivo, su redondeo se escribe como en (I.3.10), por consiguiente x/q redondeado se escribe

$$\frac{\hat{x}}{q} = \sum_{k=0}^{N/m+2} b_k / 10^{mk}, \quad (\text{I.3.12})$$

donde los b_k se los define de manera recursiva, utilizando la división euclidiana, así:

$$\begin{aligned} a_0 &= b_0 q + r_0, \\ a_1 + r_0 \cdot 10^m &= b_1 q + r_1, \\ &\vdots \\ a_{N/m+2} + r_{N/m+1} \cdot 10^m &= b_{N/m+2} + r_{N/m+2}. \end{aligned} \quad (\text{I.3.13})$$

El segundo paso será definir la adición y la sustracción para dos reales positivos x y y . Si se denota los redondeados por \hat{x} y \hat{y} , sus representaciones en punto fijo están dadas por:

$$\hat{x} = \sum_{k=0}^{N/m+2} a_k / 10^{mk}, \quad \hat{y} = \sum_{k=0}^{N/m+2} b_k / 10^{mk}.$$

La suma redondeada se escribe por lo tanto

$$\widehat{\hat{x} \pm \hat{y}} = \sum_{k=0}^{N/m+2} c_k / 10^{mk}, \quad (\text{I.3.14})$$

donde los c_k están definidos recursivamente, utilizando la división euclidiana, por:

$$\begin{aligned} c_{N/m+2} + d_{N/m+2} 10^m &= a_{N/m+2} \pm b_{N/m+2}, \\ c_{N/m+1} + d_{N/m+1} 10^m &= a_{N/m+1} \pm b_{N/m+1} + d_{N/m+2}, \\ &\vdots \\ c_0 &= a_0 \pm b_0 + d_1. \end{aligned} \quad (\text{I.3.15})$$

El tercer paso será definir la multiplicación de un real positivo con un entero q . Sea el productovskip-3pt

$$\widehat{q\hat{x}} = \sum_{k=0}^{N/m+2} b_k / 10^{mk}, \quad (\text{I.3.16})$$

donde \hat{x} está dado por (I.3.10), los coeficientes b_k están definidos recursivamente, utilizando la división euclidiana, de la manera siguiente:

$$\begin{aligned} qa_{N/m+2} &= b_{N/m+2} + d_{N/m+2}10^m, \\ qa_{N/m+1} + d_{N/m+2} &= b_{N/m+1} + d_{N/m+1}10^m, \\ &\vdots \\ b_0 &= qa_0 + d_1. \end{aligned} \tag{I.3.17}$$

Habiendo definido las operaciones requeridas para calcular π con la precisión requerida, se puede definir el algoritmo que calculará \hat{A} , y por consiguiente $\hat{\pi}$.

Algoritmo

1.- Se define $x := 1/18$, $a = x$, $k = 0$.

2.- Hacer para $k = 1$ hasta $k = M_A$:

$$\begin{aligned} x &:= x/18^2, \\ y &:= x/(2k+1), \\ a; &= a + (-1)^k y. \end{aligned}$$

3.- $a := 48 \cdot a$.

Finalmente, se debe hacer un análisis de la propagación de errores de redondeo, cometidos en el cálculo de \hat{A} , para asegurar que el resultado que de la computadora corresponde con lo esperado. Este estudio se lo efectuará en tres partes. Sea \hat{x}^{2k+1} el resultado realizado por el algoritmo al calcular x^{2k+1} para $x = 1/18$. Utilizando la proposición I.3.1, se tiene,

$$\hat{x} = x + \epsilon, \quad |\epsilon| \leq 10^{-N-2m} : \tag{I.3.18}$$

deduciendose inmediatamente:

$$\begin{aligned} \hat{x}^3 &= (x + \epsilon)/18^2 + \eta, \\ \hat{x}^5 &= \hat{x}^3/18^2 + \delta, \\ &\vdots \end{aligned} \tag{I.3.19}$$

Utilizando desigualdades triangulares y considerando series geométricas, se obtiene

$$|\hat{x}^{2k+1} - x^{2k+1}| \leq \frac{1}{1 - (1/18)^2} 10^{-N-2m}. \tag{I.3.20}$$

La segunda operación, que aparece en el algoritmo, es la división por los enteros de la forma $2k+1$. Por el mismo procedimiento que antes, se llega a

$$\left| \widehat{x^{2k+1}/(2k+1)} - x^{2k+1}/(2k+1) \right| \leq 2 \cdot 10^{-N-2m}. \tag{I.3.21}$$

Por ultimo para obtener \hat{A} , se tiene una suma de $M_A + 1$ términos y un producto por 48. En cada suma se comete un error menor a $2 \cdot 10^{-N-2m}$, de donde se tiene como estimación del error acumulado

$$|\hat{A} - A| \leq 192(M_A + 1)10^{-N-2m} \quad (\text{I.3.22})$$

Ahora bien, si $192(M_A + 1)$ es más pequeño que 10^{2m-1} , entonces el objetivo es satisfecho largamente.

Como ilustración de lo expuesto, se ha calculado π con 10000 decimales de precisión, utilizando una HP-9000.

π con 10000 decimales

```

3.1415926535 8979323846 2643383279 5028841971 6939937510 E-00050
5820974944 5923078164 0628620899 8628034825 3421170679 E-00100
8214808651 3282306647 0938446095 5058223172 5359408128 E-00150
4811174502 8410270193 8521105559 6446229489 5493038196 E-00200
4428810975 6659334461 2847564823 3786783165 2712019091 E-00250
4564856692 3460348610 4543266482 1339360726 0249141273 E-00300
7245870066 0631558817 4881520920 9628292540 9171536436 E-00350
7892590360 0113305305 4882046652 1384146951 9415116094 E-00400
3305727036 5759591953 0921861173 8193261179 3105118548 E-00450
0744623799 6274956735 1885752724 8912279381 8301194912 E-00500
9833673362 4406566430 8602139494 6395224737 1907021798 E-00550
6094370277 0539217176 2931767523 8467481846 7669405132 E-00600
0005681271 4526356082 7785771342 7577896091 7363717872 E-00650
1468440901 2249534301 4654958537 1050792279 6892589235 E-00700
4201995611 2129021960 8640344181 5981362977 4771309960 E-00750
5187072113 4999999837 2978049951 0597317328 1609631859 E-00800
5024459455 3469083026 4252230825 3344685035 2619311881 E-00850
7101000313 7838752886 5875332083 8142061717 7669147303 E-00900
5982534904 2875546873 1159562863 8823537875 9375195778 E-00950
1857780532 1712268066 1300192787 6611195909 2164201989 E-01000
3809525720 1065485863 2788659361 5338182796 8230301952 E-01050
0353018529 6899577362 2599413891 2497217752 8347913151 E-01100
5574857242 4541506959 5082953311 6861727855 8890750983 E-01150
8175463746 4939319255 0604009277 0167113900 9848824012 E-01200
8583616035 6370766010 4710181942 9555961989 4676783744 E-01250
9448255379 7747268471 0404753464 6208046684 2590694912 E-01300
9331367702 9899152104 7521620569 6602405803 8150193511 E-01350
2533824300 3558764024 7496473263 9141992726 0426992279 E-01400
6782354781 6360093417 2164121992 4586315030 2861829745 E-01450
5570674983 8505494588 5869269956 9092721079 7509302955 E-01500
3211653449 8720275596 0236480665 4991198818 3479775356 E-01550
6369807426 5425278625 5181841757 4672890977 7727938000 E-01600
8164706001 6145249192 1732172147 7235014144 1973568548 E-01650
1613611573 5255213347 5741849468 4385233239 0739414333 E-01700
4547762416 8625189835 6948556209 9219222184 2725502542 E-01750
5688767179 0494601653 4668049886 2723279178 6085784383 E-01800
8279679766 8145410095 3883786360 9506800642 2512520511 E-01850
7392984896 0841284886 2694560424 1965285022 2106611863 E-01900
0674427862 2039194945 0471237137 8696095636 4371917287 E-01950
4677646575 7396241389 0865832645 9958133904 7802759009 E-02000
9465764078 9512694683 9835259570 9825822620 5224894077 E-02050
2671947826 8482601476 9909026401 3639443745 5305068203 E-02100
4962524517 4939965143 1429809190 6592509372 2169646151 E-02150
5709858387 4105978859 5977297549 8930161753 9284681382 E-02200
6868386894 2774155991 8559252459 5395943104 9972524680 E-02250
8459872736 4469584865 3836736222 6260991246 0805124388 E-02300
4390451244 1365497627 8079771569 1435997700 1296160894 E-02350
4169486855 5848406353 4220722258 2848864815 8456028506 E-02400
0168427394 5226746767 8895252138 5225499546 6672782398 E-02450
6456596116 3548862305 7745649803 5593634568 1743241125 E-02500
1507606947 9451096596 0940252288 7971089314 5669136867 E-02550
2287489405 6010150330 8617928680 9208747609 1782493858 E-02600
9009714909 6759852613 6554978189 3129784821 6829989487 E-02650
2265880485 7564014270 4775551323 7964145152 3746234364 E-02700

```

5428584447	9526586782	1051141354	7357395231	1342716610	E-02750
2135969536	2314429524	8493718711	0145765403	5902799344	E-02800
0374200731	0578539062	1983874478	0847848968	3321445713	E-02850
8687519435	0643021845	3191048481	0053706146	8067491927	E-02900
8191197939	9520614196	6342875444	0643745123	7181921799	E-02950
9839101591	9561814675	1426912397	4894090718	6494231961	E-03000
5679452080	9514655022	5231603881	9301420937	6213785595	E-03050
6638937787	0830390697	9207734672	2182562599	6615014215	E-03100
0306803844	7734549202	6054146659	2520149744	2850732518	E-03150
6660021324	3408819071	0486331734	6496514539	0579626856	E-03200
1005508106	6587969981	6357473638	4052571459	1028970641	E-03250
4011097120	6280439039	7595156771	5770042033	7869936007	E-03300
2305587631	7635942187	3125147120	5329281918	2618612586	E-03350
7321579198	4148488291	6447060957	5270695722	0917567116	E-03400
7229109816	9091528017	3506712748	5832228718	3520935396	E-03450
5725121083	5791513698	8209144421	0067510334	6711031412	E-03500
6711136990	8658516398	3150197016	5151168517	1437657618	E-03550
3515565088	4909989859	9823873455	2833163550	7647918535	E-03600
8932261854	8963213293	3089857064	2046752590	7091548141	E-03650
6549859461	6371802709	8199430992	4488957571	2828905923	E-03700
2332609729	9712084433	5732654893	8239119325	9746366730	E-03750
5836041428	1388303203	8249037589	8524374417	0291327656	E-03800
1809377344	4030707469	2112019130	2033038019	7621101100	E-03850
4492932151	6084244485	9637669838	9522868478	3123552658	E-03900
2131449576	8572624334	4189303968	6426243410	7732269780	E-03950
2807318915	4411010446	8232527162	0105265227	2111660396	E-04000
6655730925	4711055785	3763466820	6531098965	2691862056	E-04050
4769312570	5863566201	8558100729	3060659876	8611791045	E-04100
3348850346	1136576867	5324944166	8039626579	7877185560	E-04150
8455296541	2665408530	6143444318	5867697514	5661406800	E-04200
7002378776	5913440171	2749470420	5622305389	9456131407	E-04250
1127000407	8547332699	3908145466	4645880797	2708266830	E-04300
6343285878	5698305235	8089330657	5740679545	7163775254	E-04350
2021149557	6158140025	0126228594	1302164715	5097925923	E-04400
0990796547	3761255176	5675135751	7829666454	7791745011	E-04450
2996148903	0463994713	2962107340	4375189573	5961458901	E-04500
9389713111	7904297828	5647503203	1986915140	2870808599	E-04550
0480109412	1472213179	4764777262	2414254854	5403321571	E-04600
8530614228	8137585043	0633217518	2979866223	7172159160	E-04650
7716692547	4873898665	4949450114	6540628433	6639379003	E-04700
9769265672	1463853067	3609657120	9180763832	7166416274	E-04750
8888007869	2560290228	4721040317	2118608204	1900042296	E-04800
6171196377	9213375751	1495950156	6049631862	9472654736	E-04850
4252308177	0367515906	7350235072	8354056704	0386743513	E-04900
6222247715	8915049530	9844489333	0963408780	7693259939	E-04950
7805419341	4473774418	4263129860	8099888687	4132604721	E-05000
5695162396	5864573021	6315981931	9516735381	2974167729	E-05050
4786724229	2465436680	0980676928	2382806899	6400482435	E-05100
4037014163	1496589794	0924323789	6907069779	4223625082	E-05150
2168895738	3798623001	5937764716	5122893578	6015881617	E-05200
5578297352	3344604281	5126272037	3431465319	7777416031	E-05250
9906655418	7639792933	4419521541	3418994854	4473456738	E-05300
3162499341	9131814809	2777710386	3877343177	2075456545	E-05350
3220777092	1201905166	0962804909	2636019759	8828161332	E-05400
3166636528	6193266863	3606273567	6303544776	2803504507	E-05450
7723554710	5859548702	7908143562	4014517180	6246436267	E-05500
9456127531	8134078330	3362542327	8394497538	2437205835	E-05550
3114771199	2606381334	6776879695	9703098339	1307710987	E-05600
0408591337	4641442822	7726346594	7047458784	7787201927	E-05650
7152807317	6790770715	7213444730	6057007334	9243693113	E-05700
8350493163	1284042512	1925651798	0694113528	0131470130	E-05750
4781643788	5185290928	5452011658	3934196562	1349143415	E-05800
9562586586	5570552690	4965209858	0338507224	2648293972	E-05850
8584783163	0577775606	8887644624	8246857926	0395352773	E-05900
4803048029	0058760758	2510474709	1643961362	6760449256	E-05950
2742042083	2085661190	6254543372	1315359584	5068772460	E-06000
2901618766	7952406163	4252257719	5429162991	9306455377	E-06050
9914037340	4328752628	8896399587	9475729174	6426357455	E-06100
2540790914	5135711136	9410911939	3251910760	2082520261	E-06150
8798531887	7058429725	9167781314	9699009019	2116971737	E-06200
2784768472	6860849003	3770242429	1651300500	5168323364	E-06250
3503895170	2989392233	4517220138	1280696501	1784408745	E-06300
1960121228	5993716231	3017114448	4640903890	6449544400	E-06350

6198690754	8516026327	5052983491	8740786680	8818338510	E-06400
2283345085	0486082503	9302133219	7155184306	3545500766	E-06450
8282949304	1377655279	3975175461	3953984683	3936383047	E-06500
4611996653	8581538420	5685338621	8672523340	2830871123	E-06550
2827892125	0771262946	3229563989	8989358211	6745627010	E-06600
2183564622	0134967151	8819097303	8119800497	3407239610	E-06650
3685406643	1939509790	1906996395	5245300545	0580685501	E-06700
9567302292	1913933918	5680344903	9820595510	0226353536	E-06750
1920419947	4553859381	0234395544	9597783779	0237421617	E-06800
2711172364	3435439478	2218185286	2408514006	6604433258	E-06850
8856986705	4315470696	5747458550	3323233421	0730154594	E-06900
0516553790	6866273337	9958511562	5784322988	2737231989	E-06950
8757141595	7811196358	3300594087	3068121602	8764962867	E-07000
4460477464	9159950549	7374256269	0104903778	1986835938	E-07050
1465741268	0492564879	8556145372	3478673303	9046883834	E-07100
3634655379	4986419270	5638729317	4872332083	7601123029	E-07150
9113679386	2708943879	9362016295	1541337142	4892830722	E-07200
0126901475	4668476535	7616477379	4675200490	7571555278	E-07250
1965362132	3926406160	1363581559	0742202020	3187277605	E-07300
2772190055	6148425551	8792530343	5139844253	2234157623	E-07350
3610642506	3904975008	6562710953	5919465897	5141310348	E-07400
2276930624	7435363256	9160781547	8181152843	6679570611	E-07450
0861533150	4452127473	9245449454	2368288606	1340841486	E-07500
3776700961	2071512491	4043027253	8607648236	3414334623	E-07550
5189757664	5216413767	9690314950	1910857598	4423919862	E-07600
9164219399	4907236234	6468441173	9403265918	4044378051	E-07650
3338945257	4239950829	6591228508	5558215725	0310712570	E-07700
1266830240	2929525220	1187267675	6220415420	5161841634	E-07750
8475651699	9811614101	0029960783	8690929160	3028840026	E-07800
9104140792	8862150784	2451670908	7000699282	1206604183	E-07850
7180653556	7252532567	5328612910	4248776182	5829765157	E-07900
9598470356	2226293486	0034158722	9805349896	5022629174	E-07950
8788202734	2092222453	3985626476	6914905562	8425039127	E-08000
5771028402	7998066365	8254889264	8802545661	0172967026	E-08050
6407655904	2909945681	5065265305	3718294127	0336931378	E-08100
5178609040	7086671149	6558343434	7693385781	7113864558	E-08150
7367812301	4587687126	6034891390	9562009939	3610310291	E-08200
6161528813	8437909904	2317473363	9480457593	1493140529	E-08250
7634757481	1935670911	0137751721	0080315590	2485309066	E-08300
9203767192	2033229094	3346768514	2214477379	3937517034	E-08350
4366199104	0337511173	5471918550	4644902636	5512816228	E-08400
8244625759	1633303910	7225383742	1821408835	0865739177	E-08450
1509682887	4782656995	9957449066	1758344137	5223970968	E-08500
3408005355	9849175417	3818839994	4697486762	6551658276	E-08550
5848358845	3142775687	9002909517	0283529716	3445621296	E-08600
4043523117	6006651012	4120065975	5851276178	5838292041	E-08650
9748442360	8007193045	7618932349	2292796501	9875187212	E-08700
7267507981	2554709589	0455635792	1221033346	6974992356	E-08750
3025494780	2490114195	2123828153	0911407907	3860251522	E-08800
7429958180	7247162591	6685451333	1239480494	7079119153	E-08850
2673430282	4418604142	6363954800	0448002670	4962482017	E-08900
9289647669	7583183271	3142517029	6923488962	7668440323	E-08950
2609275249	6035799646	9256504936	8183609003	2380929345	E-09000
9588970695	3653494060	3402166544	3755890045	6328822505	E-09050
4525564056	4482465151	8754711962	1844396582	5337543885	E-09100
6909411303	1509526179	3780029741	2076651479	3942590298	E-09150
9695946995	5657612186	5619673378	6236256125	2163208628	E-09200
6922210327	4889218654	3648022967	8070576561	5144632046	E-09250
9279068212	0738837781	4233562823	6089632080	6822246801	E-09300
2248261177	1858963814	0918390367	3672220888	3215137556	E-09350
0037279839	4004152970	0287830766	7094447456	0134556417	E-09400
2543709069	7939612257	1429894671	5435784687	8861444581	E-09450
2314593571	9849225284	7160504922	1242470141	2147805734	E-09500
5510500801	9086996033	0276347870	8108175450	1193071412	E-09550
2339086639	3833952942	5786905076	4310063835	1983438934	E-09600
1596131854	3475464955	6978103829	3097164651	4384070070	E-09650
7360411237	3599843452	2516105070	2705623526	6012764848	E-09700
3084076118	3013052793	2054274628	6540360367	4532865105	E-09750
7065874882	2569815793	6789766974	2205750596	8344086973	E-09800
5020141020	6723585020	0724522563	2651341055	9240190274	E-09850
2162484391	4035998953	5394590944	0704691209	1409387001	E-09900
2645600162	3742880210	9276457931	0657922955	2498872758	E-09950
4610126483	6999892256	9596881592	0560010165	5256375678	E-10000

Capítulo II

Sistemas Lineales

Una gran cantidad de problemas implica la resolución de sistemas lineales de la forma

$$Ax = b,$$

donde

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

con coeficientes reales o complejos. Por esta necesidad es importante la construcción de algoritmos que permitan su resolución en un tiempo razonable y con el menor error posible. Cada problema tiene su particularidad, motivo por el cual no existe un método general mediante el cual se pueda resolver eficazmente todos los problemas. Es verdad que existen métodos casi generales, que con pequeñas modificaciones pueden servir para encontrar la solución de un problema particular. A lo largo de este capítulo se expondrán estos métodos, dando criterios de estabilidad y estimaciones de error.

En la primera sección se tratará la condición del problema lineal, para este efecto será necesario introducir elementos de la teoría de normas en las matrices, para luego introducir las nociones relativas a la condición del problema lineal.

En la segunda sección se abordará los métodos de resolución como son: el Algoritmo de Eliminación de Gauss, la Descomposición de Cholesky y algunas modificaciones de estos métodos. En esta sección se estudiará la implementación de tales métodos, como también la estabilidad de estos.

La tercera sección tratará, la teoría e implementación de los métodos iterativos lineales como: Jacobi, Gauss-Seidel y SOR. Así mismo se analizarán algunos problemas tipos y se harán comparaciones de tales metodos.

En la cuarta sección, se verá los métodos de tipo gradiente cuyo enfoque es diferente a los métodos de las dos secciones precedentes, en efecto, son métodos que encuentran la solución a partir de problemas de minimización. Se resolverán ejemplos tipos y se comparará la eficiencia de éstos con otros métodos.

La última sección describirá el Método de los Mínimos Cuadrados, como una generalización de lo anteriormente expuesto, introduciendo como corolario la noción de Pseudo-Inversa. Así mismo se analizará la implementación del método QR , incluyendo una estimación del error de tal método.

II.1 Condición del Problema lineal

Normas de Vectores y Matrices

La noción de norma y espacio normado es un instrumento matemático muy útil en el estudio de las magnitudes que se manipulan, como también un instrumento en el estudio de la convergencia y los límites. Se empezará definiendo el concepto de norma en espacios \mathbb{R}^n , para luego definir en el álgebra de las matrices a coeficientes reales.

Definición II.1.1.- Una norma sobre \mathbb{R}^n es una aplicación

$$\| \cdot \| : \mathbb{R}^n \longrightarrow \mathbb{R},$$

con las siguientes propiedades:

- i) $\|x\| \geq 0$, $\|x\| = 0 \iff x = 0$;
- ii) $\|\alpha x\| = |\alpha| \|x\|$, donde $\alpha \in \mathbb{R}$;
- iii) $\|x + y\| \leq \|x\| + \|y\|$.

La primera condición implica que una norma siempre es positiva y es nula siempre y cuando el vector sea el vector nulo. La segunda propiedad es la homogeneidad de la norma y la tercera condición es más conocida como desigualdad del triángulo. Las normas más usuales en \mathbb{R}^n son las siguientes:

$$\|x\|_1 = \sum_{i=1}^n |x_i|, \quad \text{norma ciudad-bloque;}$$

$$\|x\|_2 = \left(\sum_{i=1}^n |x_i|^2 \right)^{\frac{1}{2}}, \quad \text{norma euclidiana;}$$

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}, \quad p > 1;$$

$$\|x\|_\infty = \max_{i=1, \dots, n} |x_i|, \quad \text{norma de la convergencia uniforme.}$$

Estas normas, que son las usualmente utilizadas, tienen algunas propiedades en común. La más importante es que, si se aumenta en valor absoluto una de las componentes, la norma se incrementa. Es necesario formalizar este hecho, motivo por el cual, se tiene la:

Definición II.1.2.- Si $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, se define el valor absoluto de x como $|x| = (|x_1|, \dots, |x_n|)$. Se dice que $|x| \leq |y|$, si $|x_i| \leq |y_i|$ para todo $i = 1, \dots, n$. Una norma $\| \cdot \|$ sobre \mathbb{R}^n se dice que es:

- (a) Monótona, si $|x| \leq |y|$ implica que $\|x\| \leq \|y\|$ para todo $x, y \in \mathbb{R}^n$.
- (b) Absoluta, si $\|x\| = \||x|\|$ para todo $x \in \mathbb{R}^n$.

Proposición II.1.3.- Una norma $\| \cdot \|$ sobre \mathbb{R}^n es monótona, si y solamente si es absoluta.

Demostración.- Si la norma $\| \cdot \|$ es monótona, sea $x \in \mathbb{R}^n$, llamemos $y = |x|$. Como $|x| \leq |y|$, y $|y| \leq |x|$ se tiene inmediatamente, porque la norma es monótona, que $\|x\| = \|y\|$.

Si la norma $\| \cdot \|$ es absoluta, sea $x \in \mathbb{R}^n$, considérese $\bar{x} = (x_1, \dots, x_{k-1}, \alpha x_k, x_{k+1}, \dots, x_n)$, con $\alpha \in [0, 1]$. Utilizando el hecho que la norma sea absoluta, desigualdad del triángulo y efectuando cálculos algebraicos se tiene:

$$\begin{aligned} \|\bar{x}\| &= \left\| \frac{1}{2}(1-\alpha)(x_1, \dots, x_{k-1}, -x_k, x_{k+1}, \dots, x_n) + \frac{1}{2}(1-\alpha)x + \alpha x \right\| \\ &\leq \frac{1}{2}(1-\alpha) \|(x_1, \dots, x_{k-1}, -x_k, x_{k+1}, \dots, x_n)\| + \frac{1}{2}(1-\alpha) \|x\| + \alpha \|x\| \\ &= \frac{1}{2}(1-\alpha) \|x\| + \frac{1}{2}(1-\alpha) \|x\| + \alpha \|x\| = \|x\|. \end{aligned}$$

Ahora bien, si $x = (x_1, \dots, x_k, \dots, x_n)$ y $y = (x_1, \dots, x_{k-1}, y_k, x_{k+1}, \dots, x_n)$, con $|y_k| \geq |x_k|$, utilizando la desigualdad anterior se tiene $\|x\| \leq \|y\|$. Para demostrar que $|x| \leq |y|$ implica que $\|x\| \leq \|y\|$, se repite el anterior paso n veces, es decir una vez por cada componente. \square

Una matriz de orden $m \times n$ puede ser vista como un vector que pertenece al espacio \mathbb{R}^{mn} , de esta manera definir la norma de una matriz como la de un vector, pero se perdería así muchas de las propiedades que tiene una aplicación lineal. Es por eso la:

Definición II.1.4.- Sea A una matriz de $m \times n$, se define su norma como

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|. \quad (\text{II.1.1})$$

La definición de la norma de una matriz depende evidentemente de las normas elegidas para $\|x\|$ y $\|Ax\|$. Sin embargo puede ser verificado sin ningún problema, que la norma de una matriz, verifica las condiciones de norma de un vector. La demostración es una simple verificación de estas condiciones, utilizando la definición de supremo. Además si las norma de los espacios \mathbb{R}^n y \mathbb{R}^m son monótonas o absolutas, es facil verificar que la norma

de matriz inducida por éstas, es todavía monótona; es suficiente utilizar la definición para probar esta afirmación. Por otro lado $\|A\|$ es el número positivo más pequeño α que satisface $\|Ax\| \leq \alpha \|x\|$, por lo tanto

$$\|Ax\| \leq \|A\| \|x\|, \quad \forall x \in \mathbb{R}^n. \quad (\text{II.1.2})$$

Una norma sobre el espacio de matrices verifica las siguientes propiedades, dada por la:

Proposición II.1.5.- *Cualquier norma sobre el espacio de las matrices $M_m(\mathbb{R})$ satisface las propiedades adicionales siguientes:*

$$\|I\| = 1, \quad (\text{II.1.3})$$

$$\|AB\| \leq \|A\| \|B\|. \quad (\text{II.1.4})$$

Demostración.- La relación (II.1.3) de la proposición es consecuencia inmediata de la definición de la norma de una matriz.

La relación (II.1.4) es consecuencia de las observaciones hechas después de la definición, en efecto

$$\begin{aligned} \|ABx\| &\leq \|A\| \|Bx\| \leq \|A\| \|B\| \|x\|, \\ \frac{\|ABx\|}{\|x\|} &\leq \|A\| \|B\|, \\ \|AB\| &\leq \|A\| \|B\|. \end{aligned}$$

□

Se ha dado las propiedades esenciales de la norma de matrices, pero es necesario conocer algunas de éstas por su utilización frecuente. Utilizando la misma notación que en las normas de los vectores definidas al inicio de la sección, se puede utilizar la misma notación en los índices de las normas de las matrices, con la convención que las normas de los vectores tienen los mismos índices.

Teorema II.1.6.- *Sea A una matriz de $n \times m$, entonces:*

$$\|A\|_1 = \max_{j=1, \dots, m} \sum_{i=1}^n |a_{ij}|, \quad (\text{II.1.5})$$

$$\|A\|_2 = \sqrt{\text{valor propio más grande de } A^t A}, \quad (\text{II.1.6})$$

$$\|A\|_\infty = \max_{i=1, \dots, n} \sum_{j=1}^m |a_{ij}|. \quad (\text{II.1.7})$$

Demostración.- Se comenzará por $\|A\|_1$, se tiene:

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1}^n \left| \sum_{j=1}^m a_{ij} x_j \right| \leq \sum_{i=1}^n \sum_{j=1}^m |a_{ij}| |x_j| \\ &\leq \sum_{j=1}^m \left(\sum_{i=1}^n |a_{ij}| \right) |x_j| \leq \left(\max_{j=1, \dots, m} \sum_{i=1}^n |a_{ij}| \right) \|x\|_1,\end{aligned}$$

por lo tanto $\|A\|_1 \leq \max_{j=1, \dots, m} \sum_{i=1}^n |a_{ij}|.$

Se mostrará, que la igualdad se cumple, en efecto, sea j_o tal que:

$$\sum_{i=1}^n |a_{ij_o}| = \max_{j=1, \dots, m} \sum_{i=1}^n |a_{ij}|; \text{ y } \bar{x} \text{ tal que } \bar{x}_{j_o} = 1, \ x_i = 0 \text{ si } i \neq j_o;$$

de esta manera $\|A\bar{x}\| = \left\| \begin{pmatrix} a_{1j_o} \\ \vdots \\ a_{mj_o} \end{pmatrix} \right\|_1 = \sum_{i=1}^n |a_{ij_o}| \|\bar{x}\|_1.$

Para la $\|\cdot\|_2$ se tiene:

$$\|Ax\|_2^2 = \langle Ax, Ax \rangle = x^t A^t A x,$$

ahora bien $A^t A$ es una matriz simétrica definida positiva, de donde los valores propios son reales no negativos, además es posible formar una base de vectores propios ortonormales. Sea $\{e_1, \dots, e_m\}$ una base de vectores propios

ortonormales, entonces $x = \sum_{i=1}^m \alpha_i e_i$ y $Ax = \sum_{i=1}^m \lambda_i \alpha_i e_i$, donde los $\lambda_i \geq 0$ son los valores propios de A . Por lo tanto, se tiene

$$\|Ax\|_2^2 = \sum_{i=1}^m \lambda_i^2 \alpha_i^2 \leq \max_{i=1, \dots, m} \lambda_i \sum_{i=1}^m \alpha_i^2 = \max_{i=1, \dots, m} \lambda_i \|x\|_2^2.$$

Para obtener la igualdad, es suficiente tomar $x = e_{j_o}$, donde λ_{j_o} es el autovalor más grande.

Para la $\|\cdot\|_\infty$ se tiene:

$$\begin{aligned}\|Ax\|_\infty &= \max_{i=1, \dots, n} \left| \sum_{j=1}^m a_{ij} x_j \right| \leq \max_{i=1, \dots, n} \left(\sum_{j=1}^m |a_{ij}| |x_j| \right) \\ &\leq \max_{i=1, \dots, n} \left(\sum_{j=1}^m |a_{ij}| \max_{j=1, \dots, m} |x_j| \right) \leq \left(\max_{i=1, \dots, n} \sum_{j=1}^m |a_{ij}| \right) \|x\|_\infty,\end{aligned}$$

así $\|A\|_\infty \leq \max_{j=1, \dots, m} \sum_{i=1}^n |a_{ij}|.$

Para obtener la igualdad es suficiente tomar $x = \mathbf{1}$, donde $\mathbf{1} = (1, \dots, 1)^t$. \square

La Condición de una Matriz

La resolución de un sistema lineal de ecuaciones debe hacer meditar sobre muchos aspectos relacionados, como ser la existencia de la solución, si el problema a resolver está bien condicionado, conocer una estimación del error cometido en la solución numérica, etc. Se tiene inicialmente el sistema de ecuaciones

$$Ax = b, \quad \text{donde } A \in M_n(\mathbb{R}), \text{ y } b \in \mathbb{R}^n; \quad (\text{II.1.8})$$

el problema es encontrar $x \in \mathbb{R}$ que sea solución del sistema (II.1.8), esta situación puede escribirse formalmente como

$$\mathcal{P}(A, b) = x.$$

De esta manera la condición del problema está dada por

$$\frac{|\mathcal{P}(A, b) - \mathcal{P}(\bar{A}, \bar{b})|}{|\mathcal{P}(A, b)|} \leq \text{cond} \cdot \text{eps}, \quad (\text{II.1.9})$$

donde:

$$\bar{a}_{ij} = a_{ij}(1 + \epsilon_{ij}), \quad |\epsilon_{ij}| \leq \text{eps}; \quad (\text{II.1.10})$$

$$\bar{b}_i = b_i(1 + \epsilon_i), \quad |\epsilon_i| \leq \text{eps}. \quad (\text{II.1.10}')$$

Si se plantea $\mathcal{P}(\bar{A}, \bar{b}) = \bar{x}$, se tiene $\frac{\|x - \bar{x}\|}{\|x\|} \leq \text{cond} \cdot \text{eps}$. Por otro lado, considerando que solamente normas monótonas son utilizadas, se tiene:

$$\left\| \begin{pmatrix} a_{11}\epsilon_{11} & \cdots & a_{1n}\epsilon_{n1} \\ \vdots & \vdots & \vdots \\ a_{n1}\epsilon_{n1} & \cdots & a_{nn}\epsilon_{nn} \end{pmatrix} \right\| \leq \text{eps} \|A\|,$$

de esta manera

$$\|\bar{A} - A\| \leq \text{eps} \|A\| \quad \text{y} \quad \|\bar{b} - b\| \leq \text{eps} \|b\|. \quad (\text{II.1.11})$$

Suponiendo que la matriz sea inversible, se puede enunciar el siguiente teorema, pero antes una definición es necesaria.

Definición II.1.7.- La condición de una matriz A inversible se define como

$$\text{cond}(A) = \|A\| \|A^{-1}\|. \quad (\text{II.1.12})$$

Teorema II.1.8.- Sea A una matriz con $\det A \neq 0$, supóngase que $\|\bar{A} - A\| \leq \|A\| \epsilon_A$, $\|\bar{b} - b\| \leq \|b\| \epsilon_b$. Si $\epsilon_A \text{cond}(A) < 1$, entonces

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \epsilon_A \text{cond}(A)} (\epsilon_A + \epsilon_b). \quad (\text{II.1.13})$$

Si además se tiene $\epsilon_A \text{cond}(A) < \frac{1}{2}$, entonces la condición del problema resolver $Ax = b$ es $\leq 2\text{cond}(A)$.

Demostración.- Se tiene $Ax = b$ y $\bar{A}\bar{x} = \bar{b}$, de donde:

$$Ax - \bar{A}\bar{x} = b - \bar{b} \quad \text{y} \quad Ax - A\bar{x} + \bar{A}\bar{x} - \bar{A}\bar{x} = b - \bar{b},$$

$$A(x - \bar{x}) = (\bar{A} - A)\bar{x} + (b - \bar{b}), \quad x - \bar{x} = A^{-1}(\bar{A} - A)\bar{x} + A^{-1}(b - \bar{b}),$$

introduciendo las desigualdades en las normas se obtiene:

$$\begin{aligned} \|x - \bar{x}\| &\leq \|A^{-1}\| \|\bar{A} - A\| \|\bar{x}\| + \|A^{-1}\| \|b - \bar{b}\| \\ &\leq \|A^{-1}\| (\|A\| \epsilon_A (\|x\| + \|\bar{x} - x\|) + \|b\| \epsilon_b) \\ &\leq \|A^{-1}\| \|A\| (\epsilon_A (\|x\| + \|\bar{x} - x\|) + \|x\| \epsilon_b), \end{aligned}$$

$$\text{de esta manera} \quad \frac{\|\bar{x} - x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \epsilon_A \text{cond}(A)} (\epsilon_A + \epsilon_b). \quad \square$$

Como la condición del problema lineal está íntimamente ligada a la condición de la matriz, es importante conocer algunas de sus propiedades, dadas por la:

Proposición II.1.9.- La condición de una matriz satisface las siguientes propiedades:

$$\text{cond}(A) \geq 1; \quad (\text{II.1.14})$$

$$\text{cond}(I) = 1; \quad (\text{II.1.15})$$

$$\text{cond}(\alpha A) = \text{cond}(A), \quad \alpha \in \mathbb{R}. \quad (\text{II.1.16})$$

Demostración.- Verificación inmediata. \square

Ejemplos

- 1.- Q ortogonal, es decir $Q^t Q = I$, la condición respecto a la norma $\|\cdot\|_2$ esta dada por

$$\text{cond}_2(Q) = 1. \quad (\text{II.1.17})$$

- 2.- Sea la matriz A dada por

$$A = \frac{1}{h} \begin{pmatrix} 4 & 1 & 0 & \cdots & 0 \\ 1 & 4 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & 1 & 4 & 1 \\ \cdots & \cdots & & 1 & 4 \end{pmatrix},$$

entonces $\|A\|_\infty = \frac{6}{h}$, además $A = \frac{4}{h}(I + N)$ donde

$$N = \begin{pmatrix} 0 & \frac{1}{4} & 0 & \cdots & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \cdots & 0 \\ \vdots & \ddots & \ddots & \cdots & \vdots \\ 0 & \cdots & 0 & \frac{1}{4} & 0 \end{pmatrix}, \quad \|N\|_\infty = \frac{1}{2} < 1.$$

Se deduce que

$$\begin{aligned} A^{-1} &= \frac{h}{4}(I + N)^{-1} = \frac{h}{4}(I - N + N^2 - N^3 + \cdots), \\ \|A^{-1}\|_\infty &\leq \frac{h}{4}(1 + \|N\| + \|N^2\| + \cdots) \\ &\leq \frac{h}{4} \left(\frac{1}{1 - \|N\|} \right) \\ &\leq \frac{h}{2}, \end{aligned}$$

entonces $\text{cond}_\infty(A) \leq 3$, por lo tanto, la matriz es bien condicionada.

3.- El siguiente ejemplo muestra la existencia de una matriz mal condicionada.

Sea H la matriz de $n \times n$, llamada matriz de Hilbert, definida por

$$h_{ij} = \frac{1}{i + j - 1}, \quad i, j = 1, \dots, n.$$

H es una matriz simétrica definida positiva, motivo por el cual la condición respecto a la norma euclidiana está dada por

$$\text{cond}_2 H = \frac{\lambda_{\max}}{\lambda_{\min}},$$

donde los λ son valores propios de H . Se puede mostrar que

$$\text{cond}_2 H \sim ce^n. \quad (\text{II.1.18})$$

Se puede observar claramente que las matrices de Hilbert son mal condicionadas, inclusive para n bastante pequeño.

4.- Finalmente, este ejemplo muestra la existencia de otra matriz mal condicionada, como ser las matrices de Vandermonde. Sea V la matriz de $n \times n$ definida por

$$V = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ c_1 & c_2 & \cdots & c_n \\ \vdots & \vdots & \cdots & \vdots \\ c_1^{n-1} & c_2^{n-1} & \cdots & c_n^{n-1} \end{pmatrix},$$

donde los c_i son diferentes. Se puede mostrar que la $\text{cond}_2 V \sim b^n$, donde $b > 1$.

Ahora bien, la estimación $\frac{\|\bar{x}-x\|}{\|x\|} \leq 2\text{cond}(A)\text{eps}$ puede ser demasiada pesimista, para ver esto, considérese el siguiente ejemplo,

$$\begin{pmatrix} 1 & 1 \\ 0 & 10^8 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

Si se llama A a la matriz del sistema, se tiene $\|A\|_2 = 10^8$ y $\|A^{-1}\| \approx 1$. El sistema con los errores de redondeo incorporados, está dado por:

$$\begin{pmatrix} 1+\epsilon_1 & 1+\epsilon_2 \\ 0 & 10^8(1+\epsilon_3) \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} 2(1+\epsilon_4) \\ 1+\epsilon_5 \end{pmatrix},$$

$$\bar{y} = 10^8 \frac{1+\epsilon_5}{1+\epsilon_3} \simeq 10^{-8}(1+\epsilon_5-\epsilon_3),$$

de donde $\frac{|y-\bar{y}|}{|y|} \leq 2\text{eps}$.

Calculando \bar{x} , se tiene

$$\begin{aligned} \bar{x} &= \frac{2(1+\epsilon_4) - (1+\epsilon_2)\bar{y}}{1+\epsilon_1} \\ &= \frac{2 - 10^{-8} + 2\epsilon_4 - 10^{-8}(\epsilon_2 + \epsilon_5 - \epsilon_3)}{1+\epsilon_1} \\ &= \frac{x + 2\epsilon_1 - 10^{-8}(\epsilon_2 + \epsilon_5 - \epsilon_3)}{1+\epsilon_1} \\ &\simeq x(1+4\text{eps}), \end{aligned}$$

por lo tanto, $\frac{|x-\bar{x}|}{|x|} \leq 4\text{eps}$.

El problema es bien condicionado, aunque la matriz A tenga una gran condición. Si se multiplica el sistema de ecuaciones por una matriz diagonal D se obtiene, el nuevo problema dado por

$$DAx = Db,$$

por el teorema II.1.8, se tiene

$$\frac{\|x-\bar{x}\|}{\|x\|} \leq 2\text{cond}(DA)\text{eps}, \quad \text{si} \quad \text{cond}(DA)\text{eps} < \frac{1}{2}.$$

En el ejemplo anterior, se plantea

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-8} \end{pmatrix}, \quad \text{así} \quad DA = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

obteniendo el:

Corolario II.1.10.- *Con las misma hipótesis del teorema II.1.8, y además si $\text{cond}(DA)\epsilon < \frac{1}{2}$, se tiene*

$$\text{La condición del problema} \leq 2 \inf_{D \text{ diagonal}} \text{cond}(DA). \quad (\text{II.1.19})$$

Ejercicios

1.- a) Sea $\| \cdot \|$ definida en \mathbb{R}^n . La bola unitaria cerrada se define como

$$\bar{B} = \{x \in \mathbb{R}^n \mid \|x\| \leq 1\},$$

mostrar que la bola unitaria es un conjunto convexo.

b) Sea \mathcal{D} un conjunto convexo acotado, en cuyo interior está 0. Si se supone que \mathcal{D} es equilibrado, es decir si $x \in \mathcal{D}$ implica que $-x \in \mathcal{D}$. Mostrar que se puede definir una norma cuya bola unitaria cerrada sea precisamente \mathcal{D} .

2.- ¿Es la función $f(x) = |x_1 - x_2| + |x_2|$ una norma sobre \mathbb{R}^2 ? Si lo es, ¿es monótona? Dibujar la bola unitaria.

3.- Dar las condiciones para que una norma sea monótona, observando su bola unitaria cerrada.

4.- Para una matriz A se define la norma de Frobenius como

$$\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2}.$$

a) Mostrar que $\|A\|_F$ es una norma sobre $\mathbb{R}^{n \times n}$.

b) Verificar la desigualdad

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n} \|A\|_2.$$

5.- Verificar la desigualdad

$$\max_{i,j} |a_{ij}| \leq \|A\|_2 \leq n \cdot \max_{i,j} |a_{ij}|.$$

6.- Sea A una matriz con $\det A \neq 0$. Mostrar que

$$\|A^{-1}\| = \left(\min_{\|x\|=1} \|Ax\| \right).$$

7.- Sea R una matriz triangular inversible. Mostrar que:

$$|r_{ii}| \leq \|R\|_p, \quad |r_{ii}|^{-1} \leq \|R^{-1}\|_p; \quad \text{para } p = 1, 2, \infty.$$

Deducir que

$$\text{cond}_p(R) \geq \max_{i,k} \frac{|r_{ii}|}{|r_{kk}|}.$$

8.- Sea

$$A = \begin{pmatrix} 2\left(\frac{1}{h_0} + \frac{1}{h_1}\right) & 0 & \cdots & \cdots & 0 \\ \frac{1}{h_1} & 2\left(\frac{1}{h_1} + \frac{1}{h_2}\right) & \frac{1}{h_2} & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{h_{n-2}} & 2\left(\frac{1}{h_{n-2}} + \frac{1}{h_{n-1}}\right) \end{pmatrix}$$

la matriz, que se encuentra en la interpolación *spline*. Mostrar:

a) $\text{cond}_\infty(A)$ puede ser arbitrariamente grande.

(Si $\max h_i / \min h_i \rightarrow \infty$).

b) Existe una matriz diagonal D tal que $\text{cond}_\infty(DA) \leq 3$.

II.2 Métodos Directos

En esta sección se desarrollará los algoritmos de resolución directa de sistemas lineales más comunes en la actualidad. El problema a resolver es

$$Ax = b, \quad (\text{II.2.1})$$

donde $A \in M_n(\mathbb{R})$, $x, b \in \mathbb{R}^n$. Por los resultados obtenidos en la teoría del Algebra Lineal, este sistema de ecuaciones lineales tiene una sola solución, si y solamente si $\det A \neq 0$. Para mostrar la importancia de contar con algoritmos cuyo costo en operaciones sea mínimo, vale la pena dar el siguiente ejemplo de resolución de este sistema de ecuaciones lineales. El ejemplo consiste en la utilización de la regla de Cramer, que en si misma constituye uno de los resultados teóricos más hermosos que se ha obtenido. Para ilustrar el método de Cramer, se debe hacer algunas convenciones sobre la notación. Para comenzar una matriz A de $n \times n$ se puede escribir como

$$A = (A_1, \dots, A_n)$$

donde A_i es la i -ésima columna de la matriz A . Dada una matriz A y un vector $b \in \mathbb{R}^n$, se denota por

$$A^{(j)} = (A_1, \dots, A_{j-1}, b, A_{j+1}, \dots, A_n,$$

la matriz obtenida remplazando la j -ésima columna por el vector b . Ahora bien, la solución del sistema (II.2.1) está dada por

$$x_i = \frac{\det A^{(i)}}{\det A},$$

donde $x = (x_1, \dots, x_n)$. Por lo tanto, la resolución del sistema implica el cálculo de $n+1$ determinantes. Si para encontrar el valor de los determinantes se utiliza la siguiente relación

$$\sum_{\sigma \in \mathcal{S}_n} \text{signo}(\sigma) \prod_{i=1}^n a_{i\sigma(i)},$$

donde \mathcal{S}_n es el grupo de permutaciones de n elementos, se debe efectuar $n!$ sumas. Si se desea resolver un sistema de 69 ecuaciones con el mismo número de incógnitas, suponiendo que cada suma se efectúa en 10^{-9} segundos, se tardaría aproximadamente 6.75×10^{49} años en llegar a la solución del

problema. Como conclusión se puede decir que existen métodos cuyo valor teórico es importante, pero su ejecución numérica es desastrosa, razón por la cual es imprescindible implementar algoritmos cuyo costo no sea muy elevado.

El Algoritmo de Gauss

Uno de los algoritmos más utilizados, precisamente por su relativo bajo costo, es el Algoritmo de Eliminación de Gauss. Considérese el sistema de ecuaciones dado por

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases}.$$

El Algoritmo de Gauss consiste en:

Primer paso Si $a_{11} \neq 0$, $l_{i1} = \frac{a_{i1}}{a_{11}}$, para $i = 2, \dots, n$.
 Se calcula $\text{linea}_i - l_{i1} * \text{linea}_1$, para $i = 2, \dots, n$.
 Si $a_{11} = 0$ se intercambia líneas.
 Obteniéndose el sistema equivalente dado por

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{22}^{(1)}x_2 + \cdots + a_{2n}^{(1)}x_n = b_2^{(1)} \\ \vdots \\ a_{n2}^{(1)}x_2 + \cdots + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases}.$$

Paso 2 Si $a_{22}^{(1)} \neq 0$, $l_{i2} = \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}$, para $i = 3, \dots, n$.
 Se calcula $\text{linea}_i - l_{i2} * \text{linea}_2$, para $i = 3, \dots, n$.
 Si $a_{22}^{(1)} = 0$ se intercambia líneas.

Se repite el procedimiento hasta obtener un sistema triangular de ecuaciones como el siguiente

$$\begin{cases} r_{11}x_1 + \cdots + r_{1,n-1}x_{n-1} + r_{1n}x_n = c_1 \\ \vdots \\ r_{n-1,n-1}x_{n-1} + r_{n-1,n}x_n = c_{n-1} \\ r_{nn}x_n = c_n \end{cases}.$$

De donde se tiene:

$$\begin{aligned}
 x_n &= \frac{c_n}{r_{nn}}, \\
 x_{n-1} &= \frac{c_{n-1} - r_{n-1,n}x_n}{r_{n-1,n-1}}, \\
 &\vdots \\
 x_1 &= \frac{c_1 - r_{12}x_2 - \cdots - r_{1n}x_n}{r_{11}}.
 \end{aligned} \tag{II.2.2}$$

Si se utiliza la notación matricial, se obtiene el siguiente esquema del algoritmo de Gauss, con las matrices aumentadas:

$$\begin{aligned}
 &\begin{matrix} (A, b) & & (A^{(1)}, b^{(1)}) & & (A^{(2)}, b^{(2)}) \end{matrix} \\
 &\begin{pmatrix} * & * & * & \cdots & * \\ * & * & * & \cdots & * \\ & & \vdots & & \\ * & * & * & \cdots & * \\ * & * & * & \cdots & * \end{pmatrix} \rightarrow \begin{pmatrix} * & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ & \vdots & & & \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \end{pmatrix} \rightarrow \begin{pmatrix} * & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & & & \\ 0 & 0 & * & \cdots & * \end{pmatrix} \\
 &\begin{matrix} (A^{(n-1)}, b^{(n-1)}) \\ \dots \rightarrow \begin{pmatrix} * & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & 0 & \cdots & 0 & * \end{pmatrix} \end{matrix}
 \end{aligned}$$

Teorema II.2.1.- Sea $\det A \neq 0$. El algoritmo de Gauss da la descomposición siguiente:

$$PA = LR, \tag{II.2.3}$$

donde:

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix}, \quad L = \begin{pmatrix} 1 & & & 0 \\ l_{21} & 1 & & \\ \vdots & & \ddots & \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix}, \tag{II.2.4}$$

y P es una matriz de permutación.

Demostración.- Supóngase, que las permutaciones necesarias han sido efectuadas al principio, es decir se aplica el Algoritmo de Gauss a la matriz

PA . Para no complicar la notación se escribe A , en vez de PA . Utilizando el mismo esquema dado más arriba, se obtiene:

$$A \longrightarrow A^{(1)} \longrightarrow \cdots \longrightarrow A^{(n-1)} = R,$$

donde:

$$A^{(1)} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -l_{21} & 1 & 0 & \cdots & 0 \\ -l_{31} & 0 & 1 & 0 \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \\ -l_{n1} & 0 & \cdots & 0 & 1 \end{pmatrix} = L_1 A,$$

$$A^{(2)} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -l_{32} & 1 & 0 \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & -l_{n2} & 0 & \cdots & 1 \end{pmatrix} = L_2 A,$$

por lo tanto

$$R = L_{n-1} L_{n-2} \cdots L_2 L_1 A.$$

Lo único que falta mostrar, es que

$$L^{-1} = L_{n-1} L_{n-2} \cdots L_2 L_1,$$

y para eso, se tiene:

$$L_i = I - V_i, \quad \text{donde} \quad V_i = \begin{pmatrix} 0 & & & & 0 \\ & 0 & & & \\ & & \ddots & & \\ & & & 0 & \\ & & & l_{i+1,i} & 0 \\ & & & \vdots & \vdots & \ddots \\ & & & l_{ni} & 0 & \cdots & 0 \end{pmatrix}.$$

Se puede verificar facilmente que $V_i V_j = 0$ para $i = 1, \dots, n$, de donde se obtiene finalmente:

$$L_i^{-1} = I + V_i,$$

$$L = I + V_1 + V_2 + \cdots + V_{n-1}.$$

□

Muchas veces, es necesario calcular sistemas de la forma:

$$\begin{aligned} Ax_1 &= b_1, \\ Ax_2 &= b_2. \end{aligned} \tag{II.2.5}$$

Se calcula una vez la descomposición LR y se resuelve de la manera siguiente:

$$\begin{aligned} Ly &= b, \\ Rx &= y. \end{aligned} \quad (\text{II.2.6})$$

Teorema II.2.2.- La descomposición LR da el siguiente resultado,

$$\det A = \pm r_{11} r_{22} \cdots r_{nn}, \quad (\text{II.2.7})$$

donde los r_{ii} son coeficientes de la diagonal de R .

Demostración.- Utilizando identidades en los determinantes se tiene

$$\det P \det A = \det L \det R.$$

□

El costo de la descomposición LR .

Para evaluar cuantas operaciones son necesarias para llegar a la descomposición LR de una matriz $A \in M_n(\mathbb{R})$, se procede de la manera siguiente:

$$A \longrightarrow A^{(1)} \quad \begin{array}{l} \text{Cálculo de los } l_{i1}: n-1 \text{ divisiones} \\ \text{Para cada fila } i, \text{ es necesario efectuar} \\ n-1 \text{ multiplicaciones mas adiciones, lo} \\ \text{que hace un total de } (n-1)^2 \text{ multipli-} \\ \text{caciones y adiciones.} \end{array}$$

Por lo tanto, contando el número de operaciones en cada etapa del algoritmo, se tiene

$$\begin{aligned} \# \text{ operaciones} &\approx (n-1)^2 + (n-2)^2 + \cdots + 2^2 + 1^2 \\ &= \sum_{i=1}^{n-1} i^2 \\ &\approx \int_0^n x^2 dx \\ &= \frac{n^3}{3}. \end{aligned} \quad (\text{II.2.8})$$

La resolución de $LRx = b$, implica aproximadamente $\frac{n^2}{2}$ multiplicaciones más adiciones en la solución de $Ly = b$. Igual número de operaciones se tiene para la resolución de $Rx = y$, lo que hace un total aproximado de n^2 operaciones.

La elección del pivote

Definición II.2.3.- Sea A una matriz, se llama pivote de la matriz A al coeficiente a_{11} .

Para ilustrar la necesidad en elegir un buen pivote, considérese el siguiente ejemplo. La precisión de los cálculos tienen tres cifras significativas en base 10. Sea el sistema de ecuaciones dado por

$$\begin{cases} 10^{-4}x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases} \quad (\text{II.2.9})$$

La solución exacta de este sistema de ecuaciones está dada por:

$$\begin{aligned} x_1 &= 1,000100010001000\dots, \\ x_2 &= 0,999899989998999\dots \end{aligned} \quad (\text{II.2.10})$$

Ahora bien, aplicando el algoritmo de Gauss con 10^{-4} como pivote, se tiene:

$$l_{21} = \frac{a_{21}}{a_{11}} = 0,100 \cdot 10^5,$$

La segunda línea del sistema se convierte en

$$-0,100 \cdot 10^5 x_2 = -0,100 \cdot 10^5,$$

de donde

$$x_2 = 0,100 \cdot 10^1,$$

resolviendo x_1 se tiene

$$0,100 \cdot 10^{-3} x_1 = 0,100 \cdot 10^1 - 0,100 \cdot 10^1,$$

por lo tanto

$$x_1 = 0. \quad (\text{II.2.11})$$

Ahora, aplíquese el algoritmo de Gauss con 1 como pivote, es decir intercambiando la primera ecuación con la segunda, se tiene:

$$l_{21} = \frac{a_{21}}{a_{11}} = 0,100 \cdot 10^{-5},$$

La segunda línea del sistema se convierte en

$$-0,100 \cdot 10^1 x_2 = 0,100 \cdot 10^1,$$

de donde

$$x_2 = 0,100 \cdot 10^1,$$

resolviendo x_1 , se tiene

$$0,100 \cdot 10^1 x_1 = 0,200 \cdot 10^1 - 0,100 \cdot 10^1,$$

por lo tanto

$$x_1 = 0,100 \cdot 10^1. \quad (\text{II.2.12})$$

La explicación de este fenómeno consiste en que la sustracción es una operación mal condicionada, cuando las cantidades a restar son muy parecidas; en efecto, considérese el siguiente sistema

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \end{cases}, \quad (\text{II.2.13})$$

al aplicar el algoritmo de Gauss se obtiene:

$$\begin{aligned} a_{22}^{(1)} &= a_{22} - l_{21}a_{12}, \\ l_{21} &= \frac{a_{21}}{a_{11}} \quad b_2^{(1)} = b_2 - l_{21}b_1, \\ a_{22}^{(1)}x_2 &= b_2^{(1)}. \end{aligned}$$

Si $|l_{21}| \gg 1$ se tiene:

$$\begin{aligned} a_{22}^{(1)} &\approx -l_{21}a_{12}, \quad b_2^{(1)} \approx -l_{21}b_1, \\ x_2 &\approx \frac{b_1}{a_{12}}, \\ x_1 &= \frac{1}{a_{11}}(b_1 - a_{12}x_2) \approx \frac{1}{a_{11}}(b_1 - b_1). \end{aligned}$$

Para solucionar este problema, se realiza una búsqueda de pivote parcial, que consiste en:

Se escoge el pivote a_{i1} tal que

$$|a_{i1}| \geq |a_{j1}| \quad j = 1, \dots, n.$$

Se procede de la misma manera para cada paso del algoritmo de Gauss.

La Estabilidad del Algoritmo de Gauss

Sea A una matriz cuyo determinante es diferente de 0. Se desea saber cual es el error cometido por el algoritmo de Gauss para encontrar la descomposición $A = LR$. No se considera las permutaciones, puesto que no se comete error de redondeo.

Si se aplica el algoritmo de Gauss, se obtiene las matrices \hat{L} y \hat{R} , llámese

$$\hat{A} = \hat{L}\hat{R}, \quad \text{la descomposición exacta de } \hat{A}.$$

Para saber, si el algoritmo es numéricamente estable, se utiliza la noción de *backward analysis* dada en capítulo I. Es decir encontrar una constante que verifique:

$$\frac{|\hat{a}_{ij} - a_{ij}|}{|a_{ij}|} \leq C \cdot \text{eps}. \quad (\text{II.2.14})$$

Por consiguiente, es necesario estimar la diferencia $|A - \hat{A}|$ elemento por elemento. Se tiene el siguiente:

Teorema II.2.4.- *Wilkinson.* Sea $\det A \neq 0$; \hat{L}, \hat{R} el resultado numérico de la descomposición LR con búsqueda de pivote $|l_{ij}| \leq 1$. Entonces

$$|A - \hat{L}\hat{R}| \leq 2a \, eps \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 2 \\ 1 & 2 & 3 & \cdots & 3 \\ \vdots & \vdots & \vdots & & \\ 1 & 2 & 3 & \cdots & n-1 \end{pmatrix}, \quad (\text{II.2.15})$$

donde $a = \max_{i,j,k} |a_{ij}^{(k)}|$ y

$$A^{(0)} \longrightarrow A^{(1)} \longrightarrow \cdots \longrightarrow A^{(n-1)}.$$

Como resultado de este teorema, el algoritmo de Gauss es numéricamente estable, siempre y cuando n no sea demasiado grande. La experiencia numérica indica que es aconsejable utilizar la descomposición de Gauss para sistemas no mayores a 1000 ecuaciones.

Demostración.- Al efectuar la descomposición LR , considerando los errores de redondeo, se tiene el siguiente esquema:

$$\hat{A}^{(0)} \longrightarrow \hat{A}^{(1)} \longrightarrow \cdots \longrightarrow \hat{A}^{(n-1)} = \hat{R}$$

Sin considerar los errores de redondeo se tiene $L_1 A = A^{(1)}$, de donde

$$\hat{L}_1 = \begin{pmatrix} 1 & & & \\ -\hat{l}_{21} & 1 & & \\ -\hat{l}_{31} & 0 & 1 & \\ \vdots & & & \vdots \\ -\hat{l}_{n1} & 0 & & 1 \end{pmatrix}, \quad \text{es la matriz } L_1 \text{ con los errores de redondeo.}$$

Por otro lado, se tiene

$$\hat{L} = \hat{L}_1^{-1} \hat{L}_2^{-1} \cdots L_{n-1}^{-1},$$

obteniendo

$$\begin{aligned} A - \hat{L}\hat{R} = & \hat{L}_1^{-1} \left(\hat{L}_1 A - \hat{A}^{(1)} \right) + \hat{L}_1^{-1} \hat{L}_2^{-1} \left(\hat{L}_2 \hat{A}^{(1)} - \hat{A}^{(2)} \right) \\ & + \cdots \left(L_1^{-1} \hat{L}_2^{-1} \cdots L_{n-1}^{-1} \right) \left(\hat{L}_{n-1} \hat{A}^{(n-2)} - \hat{A}^{(n-1)} \right). \end{aligned}$$

Ahora bien, los coeficientes de la matriz obtenida en el primer paso, están dadas por

$$\hat{a}_{ij}^{(1)} = \left(a_{ij} - \hat{l}_{i1} a_{1j} (1 + \epsilon_1) \right) (1 + \epsilon_2), \quad i \geq 1,$$

que da como consecuencia

$$\begin{aligned} \left(\hat{L}_1 A - \hat{A}^{(1)} \right)_{ij} &= \hat{l}_{i1} a_{1j} - \hat{a}_{ij} \\ &= \hat{l}_{i1} a_{1j} - \left(a_{ij} - \hat{l}_{i1} a_{1j} (1 + \epsilon_1) \right) (1 + \epsilon_2) \\ &= -a_{ij} \epsilon_2 + \hat{l}_{i1} a_{1j} \epsilon_2 + \hat{l}_{i1} a_{1j} \epsilon_1 + \hat{l}_{i1} a_{1j} \epsilon_1 \epsilon_2 \\ &= -\hat{a}_{ij}^{(1)} \epsilon_2 + \hat{l}_{i1} a_{1j} \epsilon_1 + \hat{l}_{i1} a_{1j} \epsilon_1 \epsilon_2, \end{aligned}$$

obteniendo así:

$$\left| \hat{L}_1 A - \hat{A}^{(1)} \right| \leq 2a \, eps \quad \text{donde } a = \max_{i,j,k} \left| a_{ij}^{(k)} \right|, \quad i \geq 2.$$

Bajo forma matricial, el resultado anterior está dado por

$$\left| \hat{L}_1 A - \hat{A}^{(1)} \right| \leq 2a \, eps \begin{pmatrix} 0 & \cdots & 0 \\ 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix}.$$

Continuando con el mismo procedimiento en la demostración, se obtiene

$$\left| \hat{L}_2 \hat{A}^{(1)} - \hat{A}^{(2)} \right| \leq 2a \, eps \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 1 \end{pmatrix},$$

resultados similares tambien se obtienen para los demás pasos del algoritmo de Gauss con lo que se obtiene el resultado deseado. \square

El Algoritmo de Cholesky

Un caso particular de sistema de ecuaciones lineales, es donde la matriz A es:

Definición II.2.5.- Una matriz $A \in M_n(\mathbb{R})$ es simétrica y definida positiva, si cumple las siguientes dos condiciones:

$$(II.2.16) \quad A^t = A;$$

$$(II.2.17) \quad x^t A x > 0, \quad \forall x \in \mathbb{R}^n, \quad x \neq 0.$$

Teorema II.2.6.- Sea A simétrica y definida positiva, entonces:

- a) El algoritmo de Gauss es posible sin búsqueda de pivote.
- b) La descomposición $A = LR$ satisface

$$R = DL^t, \quad \text{con } D = \text{diag}(r_{11}, r_{22}, \dots, r_{nn}). \quad (\text{II.2.17})$$

Demostración.- Sea A simétrica y definida positiva, dada por

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}.$$

El coeficiente a_{11} de la matriz A es diferente de 0, porque la matriz A es definida positiva y $a_{11} = e_1^t A e_1$ donde $e_1^t = (1, 0, \dots, 0)$. Después del primer paso del algoritmo de Gauss, se obtiene:

$$A^{(1)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & & & \\ \vdots & & C^{(1)} & \\ 0 & & & \end{pmatrix}, \quad l_{i1} = \frac{a_{i1}}{a_{11}},$$

de donde se tiene

$$c_{ij}^{(1)} = a_{ij} - l_{i1}a_{1j} = a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}}.$$

Por lo tanto es suficiente mostrar que $C^{(1)}$ es simétrica y definida positiva. En efecto, expresando la matriz A como

$$\begin{pmatrix} a_{11} & z^t \\ z & C \end{pmatrix},$$

se tiene

$$C^{(1)} = C - \frac{1}{a_{11}} z z^t.$$

Hay que mostrar que

$$y^t C^{(1)} y = y^t C y - \frac{1}{a_{11}} (y^t z)^2 > 0, \quad \forall y \neq 0.$$

Por hipótesis la matriz A es definida positiva, de donde

$$\begin{pmatrix} x_1 & y^t \end{pmatrix} \begin{pmatrix} a_{11} & z^t \\ z & C \end{pmatrix} \begin{pmatrix} x_1 \\ y \end{pmatrix} = a_{11}x_1^2 + x_1z^ty + y^tzx_1 + y^tCy > 0,$$

para $x_1 \in \mathbb{R}$ y $y \in \mathbb{R}^{n-1}$ los dos no nulos al mismo tiempo.

Planteando $x_1 = -\frac{y^t z}{a_{11}}$, se tiene

$$\frac{(y^t z)^2}{a_{11}} - \frac{2(z^t y)^2}{a_{11}} + y^t C y > 0,$$

por consiguiente

$$y^t C y - \frac{1}{a_{11}} (y^t z)^2 > 0.$$

La descomposición LR es única, si ésta existe, en efecto, si

$$A = L_1 R_1 = L_2 R_2,$$

dos descomposiciones de A . Se tiene

$$L_2^{-1} L_1 = R_2 R_1^{-1};$$

las matrices del tipo L , como aquéllas de tipo R forman subgrupos dentro el grupo de las matrices inversibles, deduciendose que

$$L_2^{-1} L_1 = I,$$

por lo tanto $L_1 = L_2$ y $R_1 = R_2$. Para demostrar la parte b) del teorema, se define la matriz L_1 , como

$$L_1^t = D^{-1} R,$$

donde $D = \text{diag}(r_{11}, \dots, r_{nn})$, hay verificar que $L_1 = L$. Las siguientes identidades se cumplen:

$$\begin{aligned} A &= LR = LD L_1^t, \\ A^t &= L_1 D L^t, \end{aligned}$$

como A es simétrica y por la unicidad de la descomposición LR , se deduce $L = L_1$. \square .

Definición II.2.7.- Sea D una matriz diagonal a coeficientes no negativos, entonces:

$$D^{\frac{1}{2}} = \text{diag} \left(\sqrt{d_{11}}, \dots, \sqrt{d_{nn}} \right). \quad (\text{II.2.18})$$

Si se define $\bar{L} = LD^{\frac{1}{2}}$, se tiene

$$A = \bar{L} \bar{L}^t,$$

que es la descomposición de Cholesky de la matriz A simétrica y definida positiva. Para simplificar notación, se escribe L , en lugar de \bar{L} . Entonces los

coeficientes de la matriz L de la descomposición de Cholesky de A , están dados por:

para $k = 1, \dots, n$:

$$l_{kk} = \sqrt{a_{kk} - l_{k1}^2 - l_{k2}^2 - \dots - l_{k,k-1}^2}; \quad (II.2.19)$$

$$l_{ik} = \frac{a_{ik} - l_{i1}l_{k1} - \dots - l_{i,k-1}l_{k,k-1}}{l_{kk}}, \quad i = 1, \dots, k-1.$$

El costo en operaciones, despreciando el cálculo de las raíces cuadradas, para obtener la descomposición de Cholesky, está dado por

$$\sum_{k=1}^n k(n-k) \approx \int_0^n x(n-x)dx = \frac{n^3}{6}. \quad (II.2.20)$$

La resolución de la ecuación $Ax = b$, donde A es una matriz simétrica y definida positiva, se puede efectuar en dos pasos utilizando la descomposición de Cholesky:

$$Ly = b,$$

$$L^t x = y,$$

los cuales necesitan un total aproximado, lo mismo que en el algoritmo de Gauss,

$$n^2 \text{ operaciones.}$$

La estabilidad de la descomposición de Cholesky está dada por el siguiente:

Teorema II.2.8.- Sea A una matriz simétrica y definida positiva. \hat{L} el resultado numérico de la descomposición de Cholesky, entonces

$$\left| A - \hat{L}\hat{L}^t \right| \leq a \text{ eps} \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 2 & \dots & 2 \\ 1 & 2 & 3 & \dots & 3 \\ \vdots & \vdots & \vdots & & \\ 1 & 2 & 3 & \dots & n \end{pmatrix}, \quad (II.2.21)$$

donde $a = \max_{ij} |a_{ij}|$.

Demostración.- Se demuestra de la misma manera que el teorema II.2.4 referente a la estabilidad de la descomposición LR .

□

Ejercicios

- 1.- Escribir una subrutina $\text{DEC}(\mathbf{N}, \mathbf{NDIM}, \mathbf{A}, \mathbf{B}, \mathbf{IP}, \mathbf{IER})$ que calcule la descomposición LR , tomando en cuenta la búsqueda parcial de pivote. Luego escribir una subrutina $\text{SOL}(\mathbf{N}, \mathbf{NDIM}, \mathbf{A}, \mathbf{B}, \mathbf{IP})$ que permita resolver $Ax = b$ utilizando la descomposición obtenida por DEC.

a) Resolver

$$\begin{pmatrix} 5 & 2 & -1 & 3 \\ 1 & 20 & 3 & 4 \\ 0 & 1 & 1 & 30 \\ 2 & 8 & -25 & 4 \end{pmatrix} x = \begin{pmatrix} 9 \\ 28 \\ 32 \\ -11 \end{pmatrix}.$$

b) Calcular la inversa de la matriz de Hilbert dada por

$$H = \left(\frac{1}{i+j} \right)_{i,j=1}^n \quad \text{para } n = 2, 3, 4, 5.$$

- 2.- a) Calcular la descomposición de Cholesky LL^t para la matriz de Hilbert

$$H = \left(\frac{1}{i+j} \right)_{i,j=1}^n, \quad n = 15.$$

b) Comparar el resultado numérico con los valores exactos,

$$l_{jk} = \frac{\sqrt{2k-1} \cdot (j-1)! \cdot (j-1)!}{(j-k)!(j+k-1)!}. \quad (\text{II.2.22})$$

¿Cuántas cifras son exactas?

c) Si \hat{L} es el resultado numérico, calcular el residuo

$$A - \hat{L}\hat{L}^t,$$

calcular también el residuo $A - LL^t$ para la matriz L dada por (II.2.22).

- 3.- Calcular $\text{cond}_\infty(A_n)$ para las matrices:

a) de Hilbert

b) de Vandermonde $(a_{ij}) = (c_i^{j-1})$, $i, j = 1 \cdots n$ con $c_i = \frac{i}{n}$,

para $n = 1, 2, 3, \dots, 15$. Calcular también $\frac{1}{n} \log_{10}(\text{cond}_\infty(A_n))$ y encontrar una fórmula que aproxime $\text{cond}_\infty(A_n)$.

- 4.- Sea A una matriz-banda, simétrica y definida positiva, p el grosor de la banda. Mostrar que, L de la descomposición de Cholesky es también una matriz-banda. Si n es el orden de la matriz, sabiendo que $n \gg p$, ¿Cuántas multiplicaciones son necesarias para calcular L ?

II.3 Métodos Iterativos

En la sección II.2, se analizó dos métodos para resolver directamente sistemas de ecuaciones lineales. Un método directo de resolución de un sistema de ecuaciones debería dar el resultado numérico igual a la solución exacta, si no se considerase los errores de redondeo, es decir, si se contase con un dispositivo de cálculo con una precisión infinita, desgraciadamente éste no es el caso. Si bien, un método directo da una solución que se aproxima a la exacta, la principal desventaja de utilizarlos, reside en el hecho en que cuando se debe resolver grandes sistemas lineales, la propagación del error de redondeo es muy grande, desvirtuando su valor; además, en muchos casos no se toma en cuenta muchas de las particularidades que pudiese tener la matriz del sistema lineal, o finalmente no es necesario obtener una solución exacta, si no una aproximación de ésta. Los métodos que serán estudiados en esta sección, son iterativos en el sentido en que se utilizan las soluciones anteriores para obtener la siguiente. Entre los métodos que serán analizados se tiene: Jacobi, Gauss-Seidel, SOR.

Métodos de Jacobi y Gauss-Seidel

Estos métodos son utilizados para resolver el sistema de ecuaciones, dado por

$$u = Au + b. \quad (\text{II.3.1})$$

El método de Jacobi consiste en la utilización de la solución anterior para calcular la nueva solución, es decir

$$u^{(k+1)} = Au^{(k)} + b. \quad (\text{II.3.2})$$

Obviamente las soluciones obtenidas por el método de Jacobi no son exactas, pero son aproximaciones de éstas.

Para la formulación del método de Gauss-Seidel, considérese la matriz A como

$$A = L + U, \quad (\text{II.3.3})$$

donde:

$$L = \begin{pmatrix} 0 & & \cdots & 0 \\ a_{21} & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{pmatrix}, \quad U = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}.$$

El método de Gauss-Seidel está dado por:

$$u^{(k+1)} = Lu^{(k+1)} + Uu^{(k)} + b, \quad (\text{II.3.4})$$

cuya formulación equivalente es

$$u^{(k+1)} = (I - L)^{-1} Uu^{(k)} + (I - L)^{-1} b. \quad (\text{II.3.5})$$

Una vez formulados estos métodos, es importante, saber que condiciones tiene que cumplir la matriz A , para que estos sean convergentes. Si se denota por u^* la solución exacta de (II.3.1), se define

$$e^{(k)} = u^{(k)} - u^*, \quad (\text{IV.3.6})$$

el error cometido en la k -ésima iteración. Por consiguiente, $e^{(k)}$ son los resultados obtenidos en las iteraciones de uno de los métodos definidos más arriba del problema

$$e = Ae, \quad (\text{IV.3.7})$$

de donde el método será convergente, siempre y cuando

$$\lim_{n \rightarrow \infty} e^{(n)} = 0. \quad (\text{IV.3.8})$$

Existe una clase de matrices, para las cuales estos métodos son convergentes. Una de sus características está dada por la:

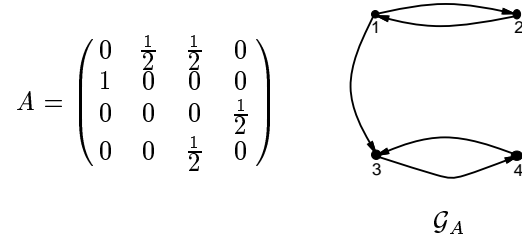
Definición II.3.1.- Una matriz A es irreducible si para cada (i, j) , existe una sucesión $l_0 = i, l_1, \dots, l_m = j$ tal que

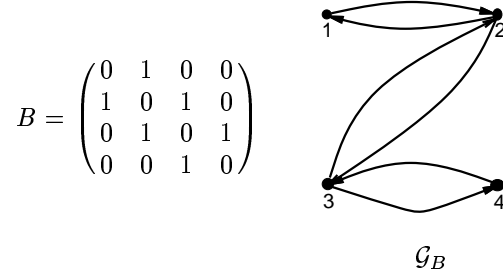
$$a_{l_k, l_{k+1}} \neq 0.$$

Gráficamente se puede visualizar mediante la noción de grafo dirigido, en Grimaldi se puede encontrar una explicación bastante detallada sobre las aplicaciones de la teoría de grafos. Considérese el grafo \mathcal{G} compuesto del conjunto de vértices \mathcal{V} y el conjunto de aristas \mathcal{A} , definido por:

- i) $\mathcal{V} = \{1, 2, \dots, n\}$,
- ii) $(i, j) \in \mathcal{A} \iff a_{ij} \neq 0$.

Para comprender esta definición, considérese los dos siguientes ejemplos, sean:





Se puede observar facilmente, que la matriz A no es irreducible, mientras que la matriz B si es irreducible.

Con la definición anterior se puede formular el siguiente teorema que da una condición suficiente, para que los métodos de Jacobi, como de Gauss-Seidel, sean convergentes.

Teorema II.3.2.- Sea A una matriz no-negativa ($a_{ij} \geq 0$), con las siguientes propiedades:

- i) $\sum_{j=1}^n a_{ij} \leq 1, \quad i = 1, \dots, n;$
- ii) Existe i_o , tal que $\sum_{j=1}^n a_{i_o j} < 1;$
- iii) A es irreducible;

entonces los métodos de Jacobi y Gauss-Seidel convergen hacia una solución única, además existe $\rho < 1$, tal que

$$\|e^{(k)}\|_{\infty} \leq C \rho^{k-1} \|e^{(0)}\|_{\infty}. \quad (\text{II.3.9})$$

Demostración.- Se mostrará para el método de Jacobi. Se tiene

$$e^{(k+1)} = A e^{(k)},$$

sea $\epsilon = \max |e_i^{(0)}|$, de donde

$$|e_i^{(1)}| \leq \sum_j a_{ij} \underbrace{|e_j^{(0)}|}_{\leq \epsilon} \leq \epsilon,$$

con desigualdad estricta para $i = i_o$.

Puesto que la matriz A es irreducible, entonces los coeficientes de A^n son todos no nulos, en efecto

$$(A^n)_{ij} = \sum_{\underbrace{k_1 \cdots k_{n-1}}_{n-1 \text{ veces}}} a_{ik_1} a_{k_1 k_2} \cdots a_{k_{n-1} j}$$

tiene un término no nulo. Además para l fijo, se tiene

$$\begin{aligned} \sum_k (A^2)_{lk} &= \sum_k \sum_j a_{lj} a_{jk} \\ &= \sum_j a_{lj} \left(\sum_k a_{jk} \right) \\ &\leq \sum_j a_{lj} \leq 1. \end{aligned}$$

Por inducción matemática, se deduce la desigualdad anterior para todas las potencias de A , e incluso

$$\begin{aligned} \sum_k (A^{n+1})_{lk} &= \sum_k \sum_j a_{lj} (A^n)_{jk} \\ &= \sum_j a_{lj} \left(\sum_k (A^n)_{jk} \right) \\ &< \sum_j a_{lj} \leq 1. \end{aligned}$$

Por otro lado, utilizando la desigualdad (II.1.2), se tiene:

$$\begin{aligned} \|Ae^{(0)}\|_\infty &\leq \|A\|_\infty \|e^{(0)}\|_\infty \leq \epsilon, \\ \|e^{(k)}\|_\infty &= \|Ae^{(k-1)}\|_\infty \leq \epsilon. \end{aligned}$$

Estas desigualdades se vuelven estrictas para N_o bastante grande, por ejemplo $N_o = n + 1$, por lo tanto

$$\|e^{(N_o)}\|_\infty \leq \rho \|e^{(0)}\|_\infty,$$

planteando

$$\rho = \left(\frac{\|e^{(N_o)}\|_\infty}{\|e^{(0)}\|_\infty} \right)^{\frac{1}{N_o}},$$

se tiene que $\rho < 1$, de donde la convergencia. El mismo procedimiento se sigue para mostrar el método de Gauss-Seidel. \square

El Teorema de Perron-Frobenius

Indudablemente por las consecuencias que implica el teorema de Perron-Frobenius, es que vale la pena estudiarlo como un tema aparte. Es un teorema cuya demostración ha llevado mucho esfuerzo de parte de muchos matemáticos. Se enunciará este teorema sin dar una demostración rigurosa, sino los pasos de ésta.

Teorema II.3.3.- *Sea A una matriz no negativa irreducible, entonces:*

- i) *Existe un vector propio u , con $u_i > 0$ respecto a un valor propio $r > 0$.*
- ii) *Todos los otros valores propios λ de A son $|\lambda| < r$. Sola excepción posible $\lambda = re^{\frac{i2\pi}{n}}$ valor propio simple.*
- iii) *r depende de manera estrictamente monótona de todos los elementos de A .*

Demostración.- Se comenzará por el punto i). Sea $u \in \mathbb{R}^n$, tal que $u_i > 0$; entonces

$$v = Au,$$

definiendo

$$r_i = \frac{v_i}{u_i},$$

si los r_i son todos iguales, el punto i) está demostrado; sino es posible variar el vector u de manera que

$$\underbrace{[r_{\min}, r_{\max}]}_{\text{nuevo}} \subsetneq \underbrace{[r_{\min}, r_{\max}]}_{\text{antiguo}},$$

obteniendo una sucesión de intervalos estrictamente encajonados, siendo el límite r .

No hay otro vector propio con $u_i > 0$. Sea el cono

$$\mathcal{K} = \{(u_1, \dots, u_n) | u_i \geq 0\}.$$

Como la matriz A es no negativa, se tiene

$$A(\mathcal{K}) \subset \mathcal{K}.$$

No es posible tener:

$$\left. \begin{array}{l} Au = \lambda u, \\ Av = \mu v, \end{array} \right\} \begin{array}{l} u \in \mathcal{K}; \\ v \in \mathcal{K}; \end{array} \quad \text{con} \quad \lambda \neq \mu.$$

En efecto, supóngase que $\lambda < \mu$ y sea $w = -u + tv$, tal que t sea lo más pequeño posible, de manera que $w \in \mathcal{K}$, entonces

$$Aw = -\lambda u + t\mu v \notin \mathcal{K},$$

por consiguiente, no puede haber más de un valor propio cuyo vector propio esté en el interior del cono.

Además, r es un valor propio simple, ya que si no lo fuera, una pequeña perturbación en la matriz A llevaría al caso de los valores propios simples, caso que ha sido ya estudiado más arriba.

Por otro lado, no hay vectores propios en el borde del cono \mathcal{K} . Efectivamente, sea $u \in \partial\mathcal{K}$, vector propio. Por hipótesis, algunos de las componentes de u son nulos, pero al menos una de las componentes es diferente de 0. De donde

$$u + Au + \cdots + A^n u = (1 + \lambda + \cdots + \lambda^n)u.$$

Como A es irreducible, existe $N_o \leq n$ con

$$a_{ij}^{N_o} > 0, \quad \forall i, j;$$

por lo tanto

$$(A^n u)_j > 0, \quad \forall j;$$

conduciendo a una contradicción.

ii) Sean λ otro valor propio de A , v su vector propio respectivo, supóngase que $\lambda \in \mathbb{R}$, se define

$$w = u + tv,$$

donde $u \in \mathcal{K}$ vector propio, t sea lo más grande posible de manera que $w \in \mathcal{K}$, ver la figura IV.3.1.

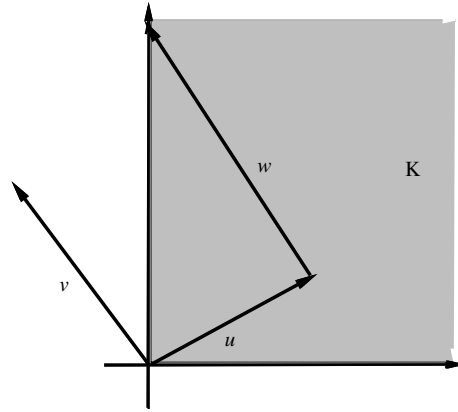


Figura IV.3.1. Demostración, Teorema Perron-Frobenius.

Si $|\lambda| > r$, Aw sale de \mathcal{K} , lo que es imposible.

Para λ complejo se tiene

$$Av = (\alpha + i\beta)v,$$

donde $\lambda = \alpha + i\beta$, $v = v_1 + iv_2$ siendo los v_j a coeficientes reales. Por lo tanto:

$$Av_1 = \alpha v_1 - \beta v_2,$$

$$Av_2 = \beta v_1 + \alpha v_2,$$

el mismo análisis que para el caso real, da el mismo resultado.

iii) La monotonía de r se muestra, después de hacer un análisis sobre el polinomio característico. \square

Las consecuencias del teorema de Perron-Frobenius son muchas, entre las cuales, la de mayor utilidad para el tema expuesto, reside sobre el teorema II.3.2. En efecto, como la matriz A es no negativa e irreducible existe un valor propio $r > 0$ más grande en modulo que los otros. Si la suma de las lineas de la matriz A fuese igual a 1 se tendría que $r = 1$, pero existe una fila cuya suma es inferior a 1, de donde por el punto iii) del teorema de Perron-Frobenius, se tiene $r < 1$. Por consiguiente, todos los valores propios en valor absoluto son inferiores a 1, dando la consiguiente convergencia de los métodos de Jacobi y Gauss-Seidel.

Continuando el estudio de los métodos de Jacobi y Gauss-Seidel se tiene otro resultado importante como el:

Teorema II.3.4.- Sea λ el valor propio de A mas grande en módulo y μ el valor propio más grande en módulo de

$$(I - L)^{-1} U. \quad (\text{II.3.10})$$

Si $\lambda < 1$, entonces $\mu < \lambda$, para toda matriz A irreducible no negativa.

Demostración.- Se tiene

$$(I - L)^{-1} = I + L + L^2 + \cdots + L^m,$$

de donde $(I - L)^{-1} U$ es una matriz no negativa.

Supóngase que x es vector propio respecto a μ , por consiguiente

$$(I - L)^{-1} Ux = \mu x,$$

$$Ux = \mu x - L\mu x,$$

$$(\mu L + U)x = \mu x,$$

por el teorema de Perron-Frobenius, $\mu \geq 0$; si $\mu = 0$ no hay nada que demostrar, si no

$$\left(L + \frac{U}{\mu} \right) x = x.$$

Utilizando nuevamente el teorema de Perron-Frobenius, es necesario que $\mu < 1$, por que de lo contrario algunos coeficientes de $L + \frac{U}{\mu}$ serán mas pequeños que de $L + U$ y por lo tanto $1 < \lambda$

Nuevamente por el teorema de Perron-Frobenius, se tiene que el valor propio maximal de $\mu L + U$ es estrictamente menor al valor propio maximal de $L + U$. \square

Como consecuencia de este teorema se tiene que el método de Gauss-Seidel converge más rápidamente a la solución que el método de Jacobi, si la matriz A es no negativa, irreducible y con valor propio maximal más pequeño que 1.

Una propiedad muy importante de algunas matrices, está dada por la:

Definición II.3.5.- Una matriz $A = L + U$, como en (II.3.3), posee la *property A* de Young, si los valores propios de la matriz definida por

$$\alpha L + \frac{1}{\alpha} U = \begin{pmatrix} & \frac{1}{\alpha} U \\ \alpha L & \end{pmatrix} \quad (\text{II.3.11})$$

son independientes de α .

Ejemplos

1.- La matriz A definida por

$$A = \begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix},$$

donde B y C son matrices cuadradas del mismo orden. A posee la *property A*. En efecto, si $\begin{pmatrix} x \\ y \end{pmatrix}$ es un vector propio de A se tiene:

$$\begin{pmatrix} & B \\ C & \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix},$$

$$\begin{pmatrix} & \frac{1}{\alpha} B \\ \alpha C & \end{pmatrix} \begin{pmatrix} x \\ \alpha y \end{pmatrix} = \lambda \begin{pmatrix} x \\ \alpha y \end{pmatrix}.$$

2.- La matriz tridiagonal con coeficientes diagonales nulos, dada por

$$\begin{pmatrix} 0 & a & & & \\ b & 0 & c & & \\ & d & 0 & e & \\ & & \ddots & & \ddots \end{pmatrix}$$

posee la *property A*. Pues, se tiene:

$$\begin{pmatrix} 0 & a & & & \\ b & 0 & c & & \\ & d & 0 & e & \\ & & \ddots & & \ddots \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ \vdots \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \\ z \\ \vdots \end{pmatrix},$$

$$\begin{pmatrix} 0 & \frac{1}{\alpha}a & & & \\ \alpha b & 0 & \frac{1}{\alpha}c & & \\ & \alpha d & 0 & \frac{1}{\alpha}e & \\ & & \ddots & & \ddots \end{pmatrix} \begin{pmatrix} x \\ \alpha y \\ \alpha^2 z \\ \vdots \end{pmatrix} = \lambda \begin{pmatrix} x \\ \alpha y \\ \alpha^2 z \\ \vdots \end{pmatrix}.$$

Teorema II.3.6.- Sea A una matriz no negativa, irreducible, con valor propio maximal $\lambda < 1$ y con la *property A*, entonces

$$\mu = \lambda^2, \quad (\text{II.3.12})$$

donde μ es le valor propio más grande de la matriz inducida por el método de Gauss-Seidel.

Demostración.- Sea x el vector propio respecto a μ , de donde

$$(\mu L + U)x = \mu x,$$

dividiendo esta expresión por $\sqrt{\mu}$ y aplicando la *property A*, se tiene

$$\left(\sqrt{\mu}L + \frac{1}{\sqrt{\mu}}U \right) x = \sqrt{\mu}x,$$

de donde $\sqrt{\mu}$ es el valor propio maximal de A . □

Si la matriz A es irreducible, no negativa, con valor propio maximal menor a 1 y además con la *property A*, se tiene:

- Método de Gauss-Seidel converge 2 veces más rápido que el de Jacobi.
- Método de Gauss-Seidel ocupa la mitad de sitio de memoria.
- Método de Gauss-Seidel ocupa la mitad de plaza de cálculo.
- Pero Jacobi puede ser paralelizado, en cambio Gauss-Seidel no.

Método de Sobrerelajación SOR

Las siglas SOR significan en inglés Successive over relaxations. Es una modificación del método de Gauss-Seidel. Consiste en utilizar un factor de sobrerelajación ω . Primero se efectua una iteración de Gauss-Seidel para luego corregir con ω , en síntesis el método SOR está dado por:

$$\begin{aligned} u^{(k+\frac{1}{2})} &= Lu^{(k+1)} + Uu^{(k)} + b, \\ u^{(k+1)} &= u^{(k)} + \omega \left(u^{(k+\frac{1}{2})} - u^{(k)} \right), \\ \omega &> 1. \end{aligned} \quad (\text{II.3.13})$$

Haciendo el cálculo de u , componente por componente, se tiene:

$$\begin{aligned} u \text{aux}_i &= \sum_{j < i} a_{ij} u_j^{(k+1)} + \sum_{j \geq i} a_{ij} u_j^{(k)} + b_i, \\ u_i^{(k+1)} &= u_i^{(k)} + \omega \left(u \text{aux}_i - u_i^{(k)} \right). \end{aligned} \quad (\text{II.3.14})$$

Para ver la velocidad de convergencia de este método, es necesario hacer un estudio sobre la convergencia. Una iteración de SOR puede ser escrita como

$$u^{(k+1)} = u^{(k)} + \omega L u^{(k+1)} + (\omega U - \omega I) u^{(k)} + \omega b,$$

es decir

$$u^{(k+1)} = \omega L u^{(k+1)} + (\omega U + (1 - \omega)I) u^{(k)} + \omega b,$$

por lo tanto

$$u^{(k+1)} = (I - \omega L)^{-1} \left[(\omega U + (1 - \omega)I) u^{(k)} + \omega b \right]. \quad (\text{II.3.15})$$

Sea μ un valor propio de $u^{(k+1)} = (I - \omega L)^{-1} (\omega U + (1 - \omega)I) u^{(k)}$ y x el vector propio asociado, entonces:

$$\begin{aligned} (I - \omega L)^{-1} (\omega U + (1 - \omega)I) x &= \mu x, \\ (\omega U + (1 - \omega)I) x &= \mu (I - \omega L) x, \\ \omega U x + (1 - \omega)x &= \mu x - \mu \omega L x, \\ \omega U x &= (\mu - 1 + \omega)x - \mu \omega L x, \\ (\mu L + U) x &= \frac{\mu - 1 + \omega}{\omega} x, \end{aligned}$$

dividiendo por $\sqrt{\mu}$ se obtiene

$$\left(\sqrt{\mu} L + \frac{1}{\sqrt{\mu}} U \right) x = \frac{\mu - 1 + \omega}{\omega \sqrt{\mu}} x,$$

de donde, se ha mostrado el siguiente:

Teorema II.3.7.- Sea A una matriz irreducible, no negativa y con la property A. Si μ es un valor propio de la matriz obtenida por SOR, entonces

$$\lambda = \frac{\mu - 1 + \omega}{\omega \sqrt{\mu}} \quad (\text{II.3.16})$$

es valor propio de la matriz A .

El problema ahora, es determinar ω , de manera que los valores propios μ sean lo más pequeños posibles. Se tiene:

$$\begin{aligned} \mu - 1 + \omega &= \lambda \omega \sqrt{\mu}, \\ (\mu + (\omega - 1))^2 &= \lambda^2 \omega^2 \mu, \end{aligned} \quad (\text{II.3.17})$$

dando la ecuación de segundo grado

$$\mu^2 + (2(\omega - 1) - \lambda^2 \omega^2) \mu + (\omega - 1)^2 = 0. \quad (\text{II.3.18})$$

Si μ_1, μ_2 , las dos raíces de esta ecuación, son complejas; entonces ambas son conjugadas, de donde

$$|\mu| = |\omega - 1|.$$

Por lo tanto, condición necesaria para la convergencia es

$$-1 < \omega < 2. \quad (\text{II.3.19})$$

Si μ_1, μ_2 son raíces reales de (II.3.18), se tiene que uno de los raíces es más grande que la otra, a menos que $\mu_1 = \mu_2$. Esto sucede cuando la parábola $\lambda \omega \sqrt{\mu}$ corta tangencialmente con la recta $\mu - 1 + \omega$. Ver figura II.3.2, por consiguiente, el discriminante de la ecuación (II.3.18) es nulo, es decir:

$$(2(\omega - 1) - \lambda^2 \omega^2)^2 = 4(\omega - 1)^2,$$

$$\lambda^2 \omega^2 = 4(\omega - 1),$$

$$\lambda^2 \omega^2 - 4\omega + 4 = 0.$$

De donde, se ha demostrado el:

Teorema II.3.8.- *El ω optimal está dado por*

$$w = \frac{2}{1 + \sqrt{1 - \lambda^2}}, \quad (\text{II.3.20})$$

y el radio espectral de la matriz SOR correspondiente es

$$\mu_{\max} = w - 1 = \frac{1 - \sqrt{1 - \lambda^2}}{1 + \sqrt{1 - \lambda^2}}. \quad (\text{II.3.21})$$

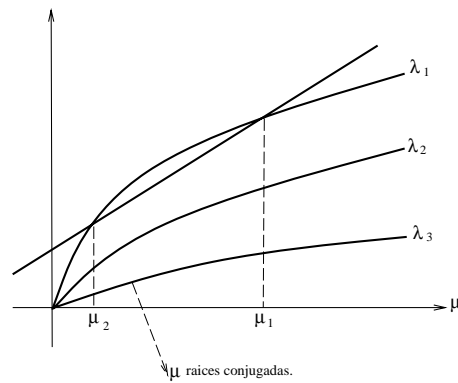


Figura IV.3.2 Determinación de w optimal.

Estudio de un problema modelo

Una de las aplicaciones más importantes de la utilización de métodos iterativos para resolver sistemas lineales, consiste en la resolución numérica de ecuaciones con derivadas parciales de tipo elíptico. Considérese el siguiente problema tipo:

$$\begin{aligned} -\Delta u &= f, \quad \text{sobre } \Omega = [0, 1] \times [0, 1]; \\ f|_{\partial\Omega} &= 0. \end{aligned} \quad (\text{II.3.22})$$

Utilizando el método de diferencias finitas, se discretiza la ecuación de derivadas parciales con un esquema de diferencias centradas. El cuadrado Ω es dividido en una malla uniforme de tamaño

$$h = \frac{1}{n+1}, \quad (\text{II.3.23})$$

se plantea:

$$\begin{aligned} x_i &= ih, \quad i = 0, \dots, n+1; \\ y_j &= jh, \quad j = 0, \dots, n+1; \end{aligned} \quad (\text{II.3.24})$$

denotando

$$u_{ij} = u(x_i, y_j), \quad f_{ij} = f(x_i, y_j). \quad (\text{II.3.25})$$

Discretizando la ecuación para esta malla, se obtiene las siguientes ecuaciones:

$$\begin{aligned} -\frac{u_{i,j+1} + u_{i,j-1} + u_{i+1,j} + u_{i-1,j} - 4u_{i,j}}{h^2} &= f_{ij}, \quad \begin{cases} i = 1, \dots, n, \\ j = 1, \dots, n; \end{cases} \\ u_{0j} &= 0, \quad j = 0, \dots, n+1; \\ u_{n+1,j} &= 0, \quad j = 0, \dots, n+1; \\ u_{i0} &= 0, \quad i = 0, \dots, n+1; \\ u_{i,n+1} &= 0, \quad i = 0, \dots, n+1. \end{aligned}$$

Cambiando la forma de las ecuaciones, se tiene

$$u_{ij} = \frac{1}{4} (u_{i,j+1} + u_{i,j-1} + u_{i+1,j} + u_{i-1,j} + h^2 f_{ij}), \quad \begin{cases} i = 1, \dots, n, \\ j = 1, \dots, n; \end{cases}$$

que en notación matricial, tiene la forma

$$u = Au + \frac{1}{4}h^2 f, \quad (\text{II.3.26})$$

donde:

$$u = (u_{11}, \dots, u_{1n}, u_{21}, \dots, u_{2n}, \dots, u_{n1}, \dots, u_{nn})^t,$$

$$f = (f_{11}, \dots, f_{1n}, f_{21}, \dots, f_{2n}, \dots, f_{n1}, \dots, f_{nn})^t,$$

$$A = \frac{1}{4} \begin{pmatrix} B & I & & & \\ I & B & I & & \\ & & \ddots & \ddots & \ddots \\ & & & I & B & I \\ & & & & I & B \end{pmatrix},$$

$$B = \begin{pmatrix} 0 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & 0 \end{pmatrix}, \quad I = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}.$$

Definición II.3.9.- El producto tensorial de dos matrices P de orden $n \times m$ y Q de orden $l \times k$ se define como la matriz

$$P \otimes Q = \begin{pmatrix} p_{11}Q & \cdots & p_{1m}Q \\ \vdots & & \vdots \\ p_{n1}Q & \cdots & p_{nm}Q \end{pmatrix}, \quad (\text{II.3.27})$$

de orden $nl \times mk$.

Por lo tanto, la matriz A del problema se escribe como

$$A = \frac{1}{4} (B \otimes I + I \otimes B).$$

Para hacer estimaciones del costo de operaciones, al encontrar la solución con un error fijado de antemano, es necesario determinar la condición de la matriz A , como así mismo el radio espectral de A para determinar la velocidad de convergencia de los métodos de Jacobi, Gauss-Seidel y SOR. Para tal efecto, se tiene la:

Proposición II.3.10.- El producto tensorial de matrices verifica la siguiente propiedad

$$(B \otimes C) (D \otimes E) = BD \otimes CE. \quad (\text{II.3.28})$$

Demostración.- Se deja como ejercicio □

Sean, y_k y ν_k el vector y el escalar respectivamente, definidos por:

$$y_k = \left(\sin\left(\frac{k i \pi}{n+1}\right) \right)_{i=1}^n, \quad (\text{II.3.29})$$

$$\nu_k = 2 \cos\left(\frac{k \pi}{n+1}\right),$$

puede verificarse como ejercicio que y_k es un vector propio de B asociado al valor propio ν_k . Se tiene, utilizando (II.3.28)

$$\begin{aligned} A(y_k \otimes y_l) &= \frac{1}{4} ((B \otimes I)(y_k \otimes y_l) + (I \otimes B)(y_k \otimes y_l)) \\ &= \frac{1}{4} (\nu_k + \nu_l)(y_k \otimes y_l), \end{aligned}$$

de donde se ha demostrado el:

Teorema II.3.11.- *Los valores propios de A están dados por:*

$$\frac{1}{4}(\nu_k + \nu_l) = \frac{1}{2} \left(\cos\left(\frac{k\pi}{n+1}\right) + \cos\left(\frac{l\pi}{n+1}\right) \right). \quad (\text{II.3.30})$$

La matriz B es tridiagonal con los coeficientes de la diagonal nulos, por consiguiente tiene la *property A*, verificar el segundo ejemplo sobre esta propiedad. Como la matriz A es igual a la suma de dos productos tensoriales entre B y I , ésta también verifica la *property A*. Conociendo los valores propios de A , se está en la posibilidad de aplicar los teoremas relativos a la convergencia de los métodos de Jacobi, Gauss-Seidel y SOR.

Utilizando el anterior teorema, se tiene que el valor propio maximal de A es igual a

$$\lambda_{\max} = \cos\left(\frac{\pi}{n+1}\right). \quad (\text{II.3.31})$$

Recordando el método de Jacobi se tiene:

$$\begin{aligned} u^{(k+1)} &= Au^{(k)} + f, \\ U^* &= AU^* + f, \\ e^{(k)} &= u^* - u^{(k)}, \\ e^{(k+1)} &= Ae^{(k)}, \end{aligned}$$

donde u^* es la solución exacta del problema lineal, $e^{(k)}$ el error en k -ésima iteración.

Existe una base de vectores propios de la matriz A , para la cual el primer vector está asociado a λ_{\max} , de donde se tiene

$$e_i^{(k+1)} = \lambda_i e_i^{(k)}, \quad i = 1, \dots, n^2;$$

donde $e_i^{(k)}$ es la i -ésima componente respecto a la base. Por consiguiente

$$e_1^{(N)} = \lambda_1^N e_1^{(0)},$$

obteniéndose así el número de iteraciones necesarias para conseguir una precisión relativa de 10^{-m} . Por las siguientes relaciones:

$$\begin{aligned}\lambda_1^n &\leq 10^{-m}, \\ N \ln \lambda_{\max} &\leq -m \ln 10, \\ N \ln \left(1 - \frac{\pi^2}{2(n+1)^2}\right) &\leq -m \ln 10, \\ -N \frac{\pi^2}{2(n+1)^2} &\leq -m \cdot 2.3,\end{aligned}$$

por lo tanto

$$N = \frac{2m \cdot 2.3}{\pi^2} (n+1)^2 \approx \frac{1}{2} mn^2. \quad (\text{II.3.32})$$

Ahora bien, una iteración equivale más o menos $4n^2$ operaciones, de donde el trabajo necesario es $2mn^4$ operaciones.

Considerando, que el método de Gauss-Seidel se realiza en la mitad de operaciones respecto al de Jacobi, y por el teorema II.3.6, el número de iteraciones para llegar a una precisión relativa de 10^{-m} es la mitad de Jacobi. Las siguiente tabla indica el número de operaciones, tiempo para diferentes precisiones requeridas.

Tabla II.3.1. Valores para el método de Jacobi.

n	Precisión	# operaciones	Tiempo de Cálculo
10	10^{-2}	10^5	0.1sec
100	10^{-4}	10^9	$10^3 \text{sec} \approx 20 \text{min}$
1000	10^{-6}	10^{13}	$10^7 \text{sec} \approx 7, 7 \text{años}$

Para el método de Gauss-Seidel los valores, se obtienen de la tabla II.3.1, dividiendo por 4 los valores para número de operaciones y tiempo de cálculo.

En este tipo de problema es donde se ve la verdadera potencia del método SOR, respecto a los métodos de Jacobi y Gauss-Seidel. Para estimar el tiempo de cálculo, como el número de operaciones es necesario estimar ω optimal, como también μ_{\max} del teorema II.3.8. Utilizando desarrollos de Taylor para la raíz cuadrada, como para el cociente, se obtiene:

$$\begin{aligned}u_{\max} &\approx 1 - \frac{2\pi}{n+1}, \\ \omega_{\text{op}} &\approx 2 \left(1 - \frac{\pi}{n+1}\right).\end{aligned} \quad (\text{II.3.33})$$

Por lo tanto, el número de iteraciones para obtener una precisión relativa de 10^{-m} es aproximadamente igual a

$$N \approx 0.4mn. \quad (\text{II.3.34})$$

Para una precisión de 10^{-6} se tiene la siguiente tabla:

Tabla II.3.2. Valores para el método SOR.

n	# Operaciones	Tiempo de Cálculo	ω_{op}
10	10^4	10^{-2}sec	1,42
100	10^7	10sec	1,93
1000	10^{10}	$10^4\text{sec} \approx 3\text{horas}$	1,9937

Para el problema con

$$f = \begin{cases} 1, & \text{sobre } [1/4, 3/4] \times [1/4, 3/4]; \\ -1, & \text{sino;} \end{cases} \quad (\text{II.3.34})$$

se obtiene utilizando SOR, los valores de u graficados en figura II.3.3.

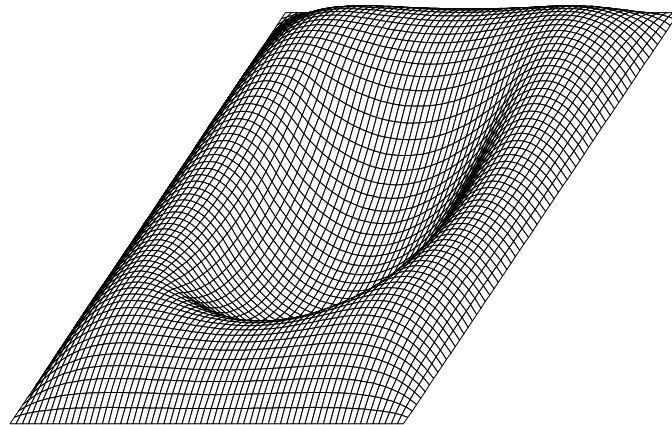


Figura II.3.3. Gráfica de la solución del problema (II.3.34).

Ejercicios

- 1.- Para resolver iterativamente el problema $Ax = b$, se considera la iteración

$$Mx^{(k+1)} = Nx^k + b, \quad \text{con } A = M - N.$$

Determinar M y N para los métodos de Jacobi y Gauss-Seidel.
Programar los dos métodos y aplicarlos al problema

$$\begin{pmatrix} 4 & -1 & 0 & -1 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ -1 & 0 & -1 & 4 \end{pmatrix} x = \begin{pmatrix} 3 \\ -3 \\ 7 \\ 1 \end{pmatrix},$$

la solución exacta es $x^* = (1 \ 0 \ 2 \ 1)$. Estudiar la velocidad de convergencia

$$\frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|}$$

para los dos métodos.

- 2.- (Estudio de la demostración del Teorema de Perron-Frobenius). Sea

$$A = \begin{pmatrix} 0 & 1/2 & & \\ 1/2 & 0 & 1/2 & \\ & 1/2 & 0 & 1/2 \\ & & 1 & 0 \end{pmatrix}.$$

Para un vector positivo u dado, se calcula $v = Au$ y se define $r_i = v_i/u_i$.
Para el vector $u = (1, 1, 1, 1)^t$ se obtiene de esta manera $r_1 = 1/2$,
 $r_2 = r_3 = r_4 = 1$.

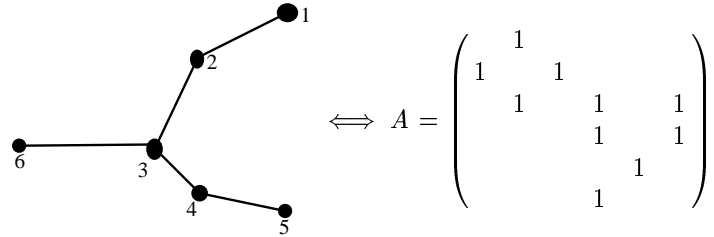
Modificar este vector con pequeñas perturbaciones, para llegar a

$$\frac{1}{2} < r_i < 1 \quad \text{para todo } i.$$

- 3.- (Un teorema de la teoría de grafos). Se da un grafo conexo de n nudos y se define una matriz A por

$$a_{ij} = \begin{cases} 1 & \text{si } i \neq j \text{ y el nudo } i \text{ está ligado al nudo } j, \\ 0 & \text{si } i = j \text{ o el nudo } i \text{ no está ligado al nudo } j, \end{cases}$$

Ejemplo:



Demostrar: Sea r el valor propio maximal de A , entonces se tiene siempre

$$r > \sqrt{3},$$

excepto en los casos:

$$\begin{array}{ll} r=1, & \text{para un grafo;} \\ r=\sqrt{2}, & \text{para un grafo;} \\ r=(1+\sqrt{5}/2), & \text{para un grafo;} \\ r=\sqrt{3}, & \text{para dos grafos.} \end{array}$$

4.- Demostrar que la matriz bloque

$$A = \begin{pmatrix} 0 & B & \\ C & 0 & D \\ & E & 0 \end{pmatrix}$$

verifica la *property A*.

5.- (*Producto de Kronecker*). Sea $B = (b_{ij})$ una matriz de $n \times n$ y $C = (c_{ij})$ una matriz $m \times m$. Se define $A = B \otimes C$, una matriz $nm \times nm$, por

$$A = B \otimes C = \begin{pmatrix} b_{11}C & \cdots & b_{1n}C \\ \vdots & & \vdots \\ b_{n1}C & \cdots & b_{nn}C \end{pmatrix}.$$

a) Mostrar: si x es un vector propio de dimensión n e y es un vector de dimensión m , entonces

$$(Bx) \otimes (Cy) = (B \otimes C)(x \otimes y)$$

donde $x \otimes y$ está definido similarmente.

b) Deducir que: si

x es vector propio de B , con valor propio μ ;

y es vector propio de C , con valor propio ν ;

entonces

$x \otimes y$ es vector propio de $B \otimes C$, con valor propio $\mu \cdot \nu$.

II.4 Métodos Minimizantes

Otra clase de métodos comúnmente utilizados en la resolución de grandes sistemas lineales, son los métodos de tipo gradiente que consisten en la resolución un problema de minimización equivalente. Esta clase de métodos se aplica a la resolución de

$$Ax = b, \quad (\text{II.4.1})$$

donde A es una matriz simétrica definida positiva.

La equivalencia con el problema de minimización, está dada por la siguiente:

Proposición II.4.1.- *Si A es simétrica y definida positiva, los dos problemas siguientes son equivalentes:*

$$f(x) = \frac{1}{2}x^t Ax - x^t b + c \rightarrow \min, \quad (\text{II.4.2})$$

$$Ax = b. \quad (\text{II.4.2})$$

Demostración.- El primer problema puede ser escrito como

$$\frac{1}{2} \sum_{i,j} x_i a_{ij} x_j - \sum_j x_j b_j + c \rightarrow \min. \quad (\text{II.4.4})$$

El procedimiento para encontrar el mínimo, consiste en derivar f , de donde

$$\frac{\partial f}{\partial x_k} = \frac{1}{2} \sum_j a_{kj} x_j + \frac{1}{2} \sum_i x_i a_{ik} - b_k = 0.$$

Puesto que A es simétrica, se tiene

$$\sum_j a_{kj} x_j - b_k = 0.$$

Si x es solución de (II.4.2), se tiene que x es solución del problema (II.4.3).

Ahora bien, si x es solución de (II.4.3), es un punto crítico de la función f , pero como A es definida positiva, se tiene que f posee x como un mínimo. \square

Método del Gradiente

El gradiente de f , utilizando la demostración de la anterior proposición, está dado por

$$\nabla f(x) = Ax - b. \quad (\text{II.4.5})$$

Sea x_o un punto de partida, se tiene:

$$\begin{aligned} f(x) &= f(x_o) + \langle \nabla f(x_o), x - x_o \rangle + \frac{1}{2}(x - x_o)^t A(x - x_o), \\ &= f(x_o) + \|x - x_o\| \|\nabla f(x_o)\| \cos \theta + \frac{1}{2}(x - x_o)^t A(x - x_o), \end{aligned}$$

donde θ es el ángulo entre $\nabla f(x_o)$ y $x - x_o$. Como se busca que $f(x)$ sea mínimo, $\nabla f(x_o)$ y $x - x_o$ deben tener la misma dirección y el mismo sentido.

Supóngase que se ha encontrado este x , que se lo denota por x_1 , se puede plantear para $k \geq 0$, obteniendo así, el método del gradiente

$$x^{k+1} = x^k - \tau_k g^k, \quad (\text{II.4.6})$$

donde $g^k = \nabla f(x^k)$. El problema, por consiguiente, es determinar τ_k teniendo en cuenta las observaciones anteriores, se define la función $G(\tau)$ por

$$G(\tau) = \frac{1}{2}(x^k - \tau g^k)^t A(x^k - \tau g^k) - (x^k - \tau g^k)^t b, \quad (\text{II.4.7})$$

de donde, el mínimo de G está en τ_k , dado por

$$\tau_k = \frac{g^{k^t} g^k}{g^{k^t} A g^k}. \quad (\text{II.4.8})$$

Una ilustración de las iteraciones del método del gradiente está dada en la figura II.4.1.

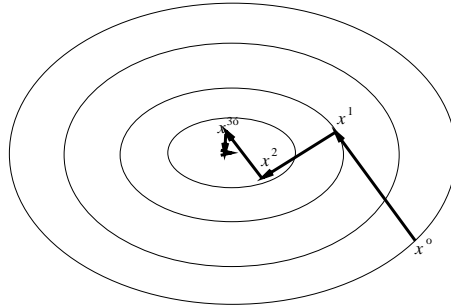


Figura II.4.1 Ilustración método del gradiente.

Planteando $h^k = Ag^k$, se tiene la versión algorítmica del método del gradiente:

```

1       $x := x_o$ ;
2       $g := Ax - b$ ;
3      Si  $|g| \leq TOL$ , entonces FIN;
4       $h := Ag$ ;
5       $\tau := \frac{\langle g, g \rangle}{\langle g, h \rangle}$ ;
6       $x := x - \tau g$ ;
7      retornar al paso 2.
```

La velocidad de convergencia del método del gradiente, está dada por el teorema siguiente, que será enunciado sin demostración.

Teorema II.4.2.- Sean, A simétrica y definida positiva, λ_{\max} el valor propio más grande, λ_{\min} el valor propio más pequeño de A , por lo tanto

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \text{cond}_2 A;$$

entonces

$$\|x^k - x^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^k \|x^0 - x^*\|, \quad (\text{II.4.9})$$

donde x^* es la solución exacta.

Además, se tiene el:

Teorema II.4.3.- Existe, x^0 tal que

$$\|x^k - x^*\| = \left(\frac{\kappa - 1}{\kappa + 1} \right)^k \|x^0 - x^*\|. \quad (\text{II.4.10})$$

Demostración.- En un referencial con un origen adecuado y una base de vectores propios de A , el problema (II.4.2) puede formularse como

$$\frac{1}{2} \hat{x}^t \bar{A} \hat{x} - c \rightarrow \min,$$

es decir

$$\bar{A} \hat{x} = 0, \quad (\text{II.4.11})$$

donde $\bar{A} = \text{diag}(\lambda_{\min}, \dots, \lambda_{\max})$. Dividiendo (II.4.11) por λ_{\max} , se obtiene el sistema equivalente

$$\hat{A} \hat{x} = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & \kappa \end{pmatrix} \hat{x} = 0. \quad (\text{II.4.12})$$

Por lo tanto, se puede suponer que A tiene la forma de \hat{A} . Planteando

$$x_i^0 = 0, \quad \text{para } i = 2, \dots, n-1;$$

el problema se reduce a resolver

$$\begin{pmatrix} 1 & 0 \\ 0 & \kappa \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (\text{II.4.13})$$

El siguiente paso, es encontrar $(x^0, y^0)^t$, que al aplicar el método del gradiente, se tenga

$$\begin{pmatrix} x^1 \\ y^1 \end{pmatrix} = \begin{pmatrix} qx^0 \\ -qy^0 \end{pmatrix}.$$

Ahora bien, una iteración del método del gradiente da:

$$g^0 = \begin{pmatrix} x^0 \\ \kappa y^0 \end{pmatrix},$$

$$\begin{pmatrix} x^1 \\ y^1 \end{pmatrix} = \begin{pmatrix} x^0 \\ y^0 \end{pmatrix} - \tau \begin{pmatrix} x^0 \\ \kappa y^0 \end{pmatrix} = \begin{pmatrix} (1-\tau)x^0 \\ (1-\kappa\tau)y^0 \end{pmatrix},$$

de donde:

$$q = 1 - \tau,$$

$$-q = 1 - \kappa\tau;$$

despejando q , se obtiene

$$q = \frac{\kappa - 1}{\kappa + 1}.$$

Puesto que la oscilación encontrada no depende del punto inicial, las iteraciones siguientes presentarán el mismo fenómeno. Con lo que se ha mostrado el teorema. \square

Método del Gradiente Conjugado

La última demostración muestra que en determinadas ocasiones el método del gradiente conduce a una oscilación en torno a la solución, haciendo que la convergencia no sea tan rápida como se espera, pudiendo mejorar esta situación, tomando aparte de la dirección del gradiente otra dirección alternativa. El método del gradiente conjugado consiste en tomar también la dirección conjugada al gradiente. Para poder enunciar tal método es necesario dar los elementos teóricos para su comprensión.

Proposición II.4.4.- Sea $f(x) = \frac{1}{2}x^tAx - x^tb$, donde A es una matriz simétrica y definida positiva. Sea d una dirección, entonces los mínimos de $f(x)$ sobre las rectas paralelas a d se encuentran en el hiperplano

$$d^t(Ax - b) = 0. \quad (\text{II.4.14})$$

Demostración.- Si x pertenece a una recta de dirección d , es de la forma

$$x = a + \lambda d, \quad \text{donde } \lambda \in \mathbb{R}.$$

Se define la función g , como $g(t) = f(a + td)$, por lo tanto esta función tiene un mínimo en t^0 , si

$$g'(t^0) = d^t f'(a + t^0 d) = 0,$$

es decir, si se cumple

$$d^t (A(a + t^0 d) - b) = 0.$$

□

Definición II.4.5.- Dos direcciones d^1 y d^2 son conjugadas respecto a A si

$$d^{1^t} A d^2 = 0. \quad (\text{II.4.15})$$

Con lo enunciado anteriormente, se puede formular el método del gradiente conjugado.

Sea x^0 un punto de partida, el gradiente inicial está dado por

$$g^0 = A x^0 - b,$$

se define

$$d^0 = -g^0.$$

Supóngase, que se ha efectuado k iteraciones, es decir, se conoce d^k , g^k y x^k ; entonces se define:

$$\begin{aligned} \tau_k &= -\frac{\langle d^k, g^k \rangle}{\langle d^k, A d^k \rangle}, \\ x^{k+1} &= x^k + \tau_k d^k, \\ g^{k+1} &= A x^{k+1} - b = g^k + \tau_k A d^k, \\ d^{k+1} &= -g^{k+1} + \beta_k d^k. \end{aligned} \quad (\text{II.4.16})$$

Para determinar β_k , se impone la condición que d^{k+1} y d^k sean conjugados, por consiguiente

$$g^{k+1^t} A d^k = \beta_k d^{k^t} A d^k,$$

de donde

$$\beta = \frac{\langle g^{k+1}, A d^k \rangle}{\langle d^k, A d^k \rangle}. \quad (\text{II.4.17})$$

La versión algorítmica está dada por:

```

1       $x :=$  punto de partida;
2       $g := Ax - b$ ;
3       $d := -g$ ;
4       $h := Ad$ ;
5       $\tau := \frac{\langle d, g \rangle}{\langle d, h \rangle}$ ;
6       $x := x + \tau d$ ;
7       $g := g + \tau h$ ;
8      si  $|g| \leq 10^{-6}$ , entonces fin algoritmo;
9       $\beta := \frac{\langle g, h \rangle}{\langle d, h \rangle}$ ;
10      $d := -g + \beta d$ ;
11     Retornar al paso 4.
```

Teorema II.4.6.- Con el algoritmo del Gradiente Conjugado se tiene:

$$\begin{aligned} \langle d^k, Ad^l \rangle &= 0, & l < k; \\ \langle g^k, g^l \rangle &= 0, & l < k. \end{aligned} \quad (\text{II.4.18})$$

Demostración.- Por inducción sobre k .

Para $k = 1$ se tiene $\langle d^0, Ad^1 \rangle = 0$ por construcción del método,

$$\begin{aligned} \langle g^1, g^0 \rangle &= \langle g^0, g^0 \rangle - \langle \tau_0 Ag^0, g^0 \rangle \\ &= 0. \end{aligned}$$

Supóngase cierto para k , por la construcción de g^{k+1} , se tiene $\langle g^{k+1}, d^k \rangle = 0$, entonces:

$$\begin{aligned} \langle g^{k+1}, g^k \rangle &= \underbrace{\langle g^{k+1}, d^k \rangle}_0 + \beta_{k-1} \langle g^{k+1}, d^{k-1} \rangle \\ &= \beta_{k-1} \underbrace{\langle g^k, d^{k-1} \rangle}_0 + \tau_k \beta_{k-1} \underbrace{\langle Ad^k, d^{k-1} \rangle}_{\text{o hip. ind}}; \end{aligned}$$

$$\begin{aligned} \langle g^{k+1}, g^l \rangle &= \underbrace{\langle g^k, g^l \rangle}_0 + \tau_k \langle Ad^k, d^l \rangle \\ &= 0; \end{aligned}$$

$$\begin{aligned} \langle d^{k+1}, Ad^l \rangle &= c_1 \langle g^{k+1}, Ad^l \rangle + c_2 \underbrace{\langle d^k, Ad^l \rangle}_{\text{o hip. ind}} \\ &= c_1 \langle g^{k+1}, g^l \rangle = 0. \end{aligned}$$

En el ejercicio 4, se tiene otras fórmulas para τ_k, β_k □

Definición II.4.7.- Se define la norma natural del problema por

$$\|x\|_A = \sqrt{x^t A x}. \quad (\text{II.4.19})$$

Teorema II.4.8.- El error del método del Gradiente Conjugado, después de l iteraciones satisface

$$\|x^* - x^l\|_A \leq M \|x^* - x^0\|_A, \quad (\text{II.4.20})$$

donde

$$M = \inf_{\mu_i \in \mathbb{R}} \left(\sup_{\lambda \text{ v.p. } A} |1 + \mu_1 \lambda + \mu_2 \lambda^2 + \cdots + \mu_l \lambda^l| \right). \quad (\text{II.4.21})$$

Corolario II.4.9.- Si A tiene solamente l valores propios diferentes, entonces el método del Gradiente Conjugado da la solución exacta después de l iteraciones.

Demostración del teorema.- Se puede suponer por traslación, que $b = 0$, de donde la solución exacta es $x^* = 0$, por lo tanto el error cometido es $e^l = x^l$. Se tiene $d^0 = -g^0$, definiendo E_1 por:

$$\begin{aligned} E_1 &= \{x = x^0 + \tau_1 d^0\} \\ &= \{x = x^0 + \gamma_1 A x^0\} \\ &= \{x = (I + \gamma_1 A) x^0\}. \end{aligned}$$

x^1 es el punto de E_1 que minimiza $f(x) = \frac{1}{2} x^t A x$.

$$\begin{aligned} E_2 &= \{x | x = x^0 + \tau_1 d^0 + \tau_2 d^1\} \\ &= \{x | x = (I + \gamma_1 A + \gamma_2 A^2) x^0\}. \end{aligned}$$

x^2 minimiza $f(x)$ en E_2 . Continuando se llega a

$$E_l = \{x | x = (I + \gamma_1 A + \cdots + \gamma_l A^l) x^0\},$$

x^l minimiza $f(x)$ en E_l .

Sean v_1, \dots, v_n vectores propios normalizados de A , por lo tanto forman una base. Por consiguiente:

$$\begin{aligned} x^0 &= \sum_{i=1}^n \alpha_i v_i, \\ x^{0^t} A x^0 &= \sum_{i,j} \alpha_i v_i^t A v_j \alpha_j \\ &= \sum_i \alpha_i^2 \lambda_i \\ &= \|x^0\|_A^2. \end{aligned}$$

Por otro lado, se tiene

$$\begin{aligned} x^k &= (I + \gamma_1 A + \cdots + \gamma_l A^l) \sum_{i=1}^n \alpha_i v_i \\ &= \sum_{i=1}^n \alpha_i (1 + \gamma_1 \lambda_i + \cdots + \gamma_l \lambda_i^l) v_i, \\ x^l A x^l &= \sum_{i=1}^n \alpha_i^2 \lambda_i (1 + \gamma_1 \lambda_i + \cdots + \gamma_l \lambda_i^l), \end{aligned}$$

de donde

$$\|x^l\|_A^2 \leq \max_{\lambda_1, \dots, \lambda_n} |1 + \gamma_1 \lambda_i + \cdots + \gamma_l \lambda_i^l|^2 \|x^0\|_A^2.$$

□

Polinomios de Chebichef

Supóngase, que se conoce los valores propios de la matriz A simétrica y definida positiva, es decir:

$$0 < \lambda_{\min} \leq \lambda_i \leq \lambda_{\max}.$$

El polinomio que minimiza el polinomio del teorema II.4.8 tiene la propiedad siguiente:

- es igual a ϵ en λ_{\min} ,
- admite los valores $-\epsilon, +\epsilon$ como valores maximos en el intervalo $[\lambda_{\min}, \lambda_{\max}]$ exactamente $l + 1$ veces,
- es igual a $\pm\epsilon$ en λ_{\max} .

La razón puede apreciarse en la figura II.4.2, para el caso $l = 2$.

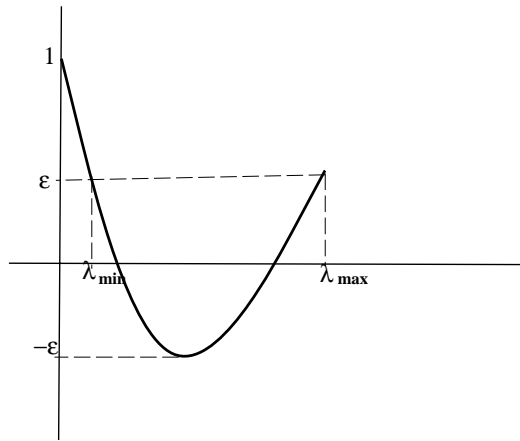


Figura II.4.2 Polinomio minimizante para $l = 2$.

Definición II.4.10.- El n -simo polinomio de Chebichef, es el polinomio de grado n , $T_n(x)$ tal que:

$$T_n(1) = 1,$$

$$T_n(-1) = (-1)^n,$$

Todos los mínimos y máximos relativos son ± 1 y pertenecen a $(-1, 1)$.

Teorema II.4.11.- *Chebichef.* $T_n(x)$ está dado por

$$T_n(x) = \frac{1}{2} \left(\left(x + \sqrt{x^2 - 1} \right)^n + \left(x - \sqrt{x^2 - 1} \right)^n \right). \quad (\text{II.4.22})$$

Demostración.- Este teorema será abordado en el Capítulo IV, motivo por el cual la demostración no es necesaria por el momento. Lo único que puede decirse, es que la definición moderna de los polinomios de Chebichef está dada por

$$T_n(x) = \cos n\theta, \quad \text{donde } x = \cos \theta. \quad (\text{II.4.23})$$

□

Teorema II.4.12.- *La constante M del teorema II.4.8, está dada por*

$$M = \frac{1}{T_l \left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} \right)}. \quad (\text{II.4.24})$$

Demostración.- Se utiliza la transformación afín dada por

$$x = \alpha\lambda + \beta,$$

donde:

$$\alpha = \frac{2}{\lambda_{\max} - \lambda_{\min}},$$

$$\beta = -\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}.$$

por consiguiente tomando $1/M = T_l(\beta)$ se tiene el resultado deseado. □

Teorema II.4.13.- *Para el método del Gradiente Conjugado, se tiene la estimación*

$$\|x^l - x^*\|_A \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^l \|x^0 - x^*\|_A, \quad (\text{II.4.25})$$

donde $\kappa = \text{cond}_2(A)$, x^* la solución exacta.

Demostración.- Para $x \geq 1$, se tiene

$$|T_l(x)| \geq \frac{1}{2} \left(x + \sqrt{x^2 - 1} \right)^l,$$

de donde

$$M \leq \frac{2}{\left(x + \sqrt{x^2 - 1} \right)^l}, \quad \text{para } x \geq 1; \quad (\text{II.4.26})$$

particularmente para

$$x = \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} = \frac{\kappa + 1}{\kappa - 1},$$

reemplazando en (II.4.26), se obtiene (II.4.25). \square

Con los resultados obtenidos respecto a los errores de convergencia, tanto para el método del gradiente, como para el método del gradiente conjugado, se puede estimar el número de iteraciones necesarias para conseguir una precisión relativa de 10^{-b} . Para el método del gradiente por (II.4.6), se tiene:

$$\begin{aligned} 10^{-b} &\approx \left(\frac{\kappa - 1}{\kappa + 1} \right)^l, \\ -2.3b &\approx l \ln(1 - 1/\kappa) - l \ln(1 + 1/\kappa), \\ -2.3b &\approx -2l/\kappa, \\ l &\approx b\kappa. \end{aligned} \quad (\text{II.4.27})$$

Con el mismo procedimiento que para el método del gradiente, el método del gradiente conjugado necesita aproximadamente l iteraciones, de donde

$$l \approx b\sqrt{\kappa}. \quad (\text{II.4.28})$$

Método del Gradiente Conjugado Precondicionado

El corolario II.4.9 indica claramente, que el método del Gradiente Conjugado converge a la solución exacta después de l iteraciones, siendo l el número de valores propios diferentes de la matriz A . Se puede mejorar la velocidad de convergencia, haciendo un cambio de variables en el problema original

$$f(x) = \frac{1}{2} x^t A x - x^t b.$$

Se define \hat{x} como

$$\hat{x} = E^t x,$$

donde E es una matriz inversible con ciertas condiciones a cumplir que serán dadas más adelante. Para no recargar la notación, se denota

$$E^{t-1} = E^{-t}.$$

Por consiguiente, el problema se convierte en

$$f(\hat{x}) = \frac{1}{2} \hat{x}^t E^{-1} A E^{-t} \hat{x} - \hat{x} E^{-1} b. \quad (\text{II.4.29})$$

El problema es encontrar E , de manera $EE^t \approx \mu A$, $\mu > 0$ con el menor gasto de operaciones posibles. Si $EE^t = \mu A$, aplicando el método del gradiente conjugado a (II.4.29) se tiene la solución exacta después de una sola iteración, en este caso, la matriz E es la descomposición de Cholesky de μA . Ahora bien, realizar la descomposición de Cholesky significa que se puede calcular directamente la solución, lo que implica un gran número de operaciones a ejecutar. La determinación de E se la realiza utilizando dos métodos que serán analizados más adelante.

El algoritmo del gradiente conjugado para el problema condicionado está dado por:

- 1 $\hat{x} :=$ punto de partida;
- 2 $\hat{g} := E^{-1} A E^{-t} \hat{x} - E^{-1} b$;
- 3 $\hat{d} := -\hat{g}$;
- 4 $\hat{h} := E^{-1} A E^{-t} \hat{d}$;
- 5 $\hat{\tau} := \frac{\langle \hat{d}, \hat{g} \rangle}{\langle \hat{d}, \hat{h} \rangle}$;
- 6 $\hat{x} := \hat{x} + \hat{\tau} \hat{d}$;
- 7 $\hat{g} := \hat{g} + \hat{\tau} \hat{h}$;
- 8 si $|\hat{g}| \leq 10^{-6}$, entonces fin algoritmo;
- 9 $\hat{\beta} := \frac{\langle \hat{g}, \hat{h} \rangle}{\langle \hat{d}, \hat{h} \rangle}$;
- 10 $\hat{d} := -\hat{g} + \hat{\beta} \hat{d}$;
- 11 Retornar al paso 4.

La implementación del método del gradiente conjugado al problema condicionado (II.4.29), tiene dos inconvenientes: El primero que se debe conocer explícitamente la matriz E y el segundo que se efectúan muchas operaciones con la matriz E .

Planteando $C = EE^t$, se puede formular el algoritmo, comparando con el método del gradiente conjugado para el problema condicionado y utilizando las identidades para τ_k y β_k dadas en el ejercicio 4, uno se

dará cuenta de las sutilezas empleadas. Por consiguiente, la formulación algorítmica está dada por:

```

1       $x :=$  punto de partida;
2       $g := Ax - b$ ;
3       $\chi := C^{-1}g$ ;
4       $\delta_0 := \langle g, \chi \rangle$ ;
5       $d := -\chi$ ;
6       $h := Ad$ ;
7       $\tau := \delta_0 / \langle d, h \rangle$ ;
8       $x := x + \tau d$ ;
9       $g := g + \tau h$ ;
10     si  $|g| \leq 10^{-6}$ , entonces fin algoritmo;
11      $\chi := C^{-1}g$ ;
12      $\delta_1 := \langle g, \chi \rangle$ ;
13      $\beta := \delta_1 / \delta_0$ ;
14      $d := -\chi + \beta d$ ;
15      $\delta_0 := \langle d, g \rangle$ ;
16     Retornar al paso 6.
```

Esta formulación presenta algunas novedades respecto a la formulación del método del gradiente conjugado. La más importante es la aparición del vector χ . El gasto en memoria es prácticamente el mismo, pues se puede utilizar el mismo lugar de memoria para h y χ . Por las observaciones anteriores la matriz C debe ser lo más parecida a la matriz A o a un múltiplo de ésta. Además, se debe resolver la ecuación

$$C\chi = g,$$

motivo por el cual la matriz C debe ser la más simple posible. Actualmente existen dos métodos muy utilizados para determinar C , dependiendo sobre todo de las características de A . Estos son:

1.- Factorización Incompleta de Cholesky

La descomposición de Cholesky incompleta de la matriz A , está dada por EE^t . Los coeficientes de E son nulos en el lugar donde los coeficientes de A son nulos; para los otros coeficientes, los valores están dados por las fórmulas obtenidas para la descomposición de Cholesky dada en la sección II.2, es decir:

$$e_{ii} = \sqrt{a_{ii} - \sum_{j=1}^{i-1} e_{ji}^2}, \quad (\text{II.4.30a})$$

$$e_{ki} = \frac{\left(a_{ki} - \sum_{j=1}^{k-1} e_{kj}e_{ji} \right)}{e_{ii}}. \quad (\text{II.4.30b})$$

Luego se resuelve:

$$\begin{aligned} E\hat{g} &= g; \\ E^t\chi &= \hat{g}. \end{aligned}$$

2.- SSOR preconditionado (*Axelsson*, 1973)

La matriz A puede ser descompuesta en tres partes, la diagonal, la parte inferior y la parte superior. El método SSOR consiste en definir la matriz E , como una matriz triangular inferior, cuyos coeficientes de la diagonal son iguales a la raíz cuadrada de los coeficientes de la diagonal de A . Los coeficientes de la parte inferior de la matriz E son iguales a los coeficientes de la parte inferior de la matriz A multiplicados por un factor de relajación positivo ω , es decir:

$$e_{ii} = \sqrt{a_{ii}}; \quad (\text{II.4.31a})$$

$$e_{ki} = \omega a_{ki}, \quad \omega \geq 0. \quad (\text{II.4.31b})$$

Luego se resuelve:

$$\begin{aligned} E\hat{g} &= g, \\ E^t\chi &= \hat{g}. \end{aligned}$$

La determinación del ω optimal se realiza al tanteo, dependiendo del problema a resolver, pues por el momento no existe un criterio analítico para determinar este valor optimal.

Resultados Numéricos

El estudio teórico y la formulación de métodos de resolución de sistemas lineales por si solo constituye un hermoso ejercicio mental. Un método no es bueno, hasta que ha sido probado numéricamente en problemas concretos, por eso es importante poder comparar la eficiencia de los diferentes métodos formulados en esta sección, en diferentes tipos de problemas.

Al igual que la sección II.3, se considerará una ecuación a derivadas parciales de tipo elíptico, pues este tipo de problema, se presta mucho a la resolución de grandes sistemas lineales. Sea,

$$\begin{aligned} \Delta u &= -1, \quad \text{sobre } \Omega = [0, 1] \times [0, 1]; \\ u|_{\partial\Omega} &= 0. \end{aligned} \quad (\text{II.4.32})$$

De la misma manera, que la sección precedente, se subdivide el dominio Ω en una malla uniforme de longitud $1/n + 1$, planteando:

$$\left. \begin{aligned} x_i &= ih \\ y_j &= jh \end{aligned} \right\}, \quad h = \frac{1}{n+1}; \quad (\text{II.4.33})$$

se obtiene el sistema de ecuaciones dado por:

$$\begin{aligned} u_{i,j+1} + u_{i,j-1} + u_{i+1,j} + u_{i-1,j} - 4u_{ij} &= -h^2, & i, j &= 1, \dots, n; \\ u_{kj} &= u_{il} = 0, & k &= 0 \text{ o } k = n+1, \quad l = 0 \text{ o } l = n+1. \end{aligned}$$

Planteando $u = (u_{11}, \dots, u_{1n}, u_{21}, \dots, u_{2n}, \dots, u_{n,1}, \dots, u_{nn})^t$, el sistema lineal se escribe como

$$Au = h^2 \mathbf{1}, \quad (\text{II.4.34})$$

donde $A = 4I + B \otimes I_n + I \otimes B$, con \otimes producto tensorial de matrices definido en la anterior sección, I_n la matriz identidad de orden $n \times n$ y

$$B = \begin{pmatrix} 0 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 0 \end{pmatrix}.$$

A es una matriz definida positiva y simétrica de orden $n^2 \times n^2$.

El problema (II.4.34) será resuelto tomando valores aleatorios para u^1 . Se verá el costo en iteraciones para diferentes valores de n . Una vez efectuados los experimentos numéricos, las gráficas de las iteraciones del método del gradiente conjugado preconditionado SSOR, pueden ser observadas en la figura II.4.3.

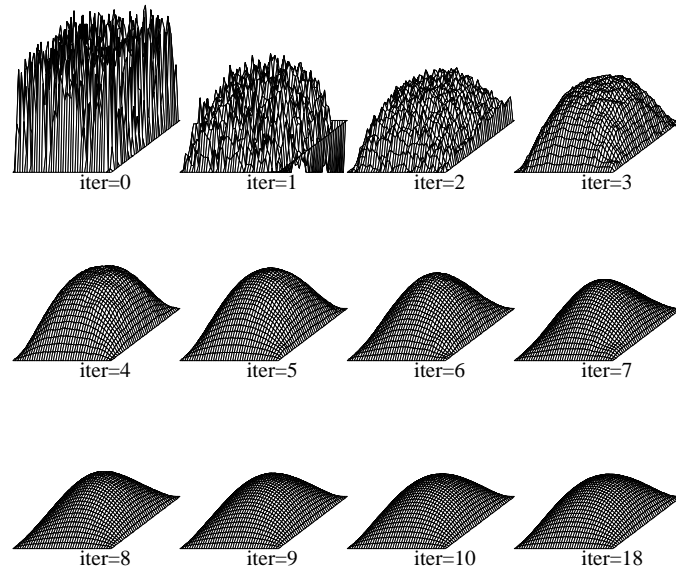


Figura II.4.3. Solución del problema (II.4.34).

En la figura II.4.4 se observa en una gráfica el costo en iteraciones para alcanzar una precisión de 10^{-6} para los diferentes métodos de tipo gradiente.

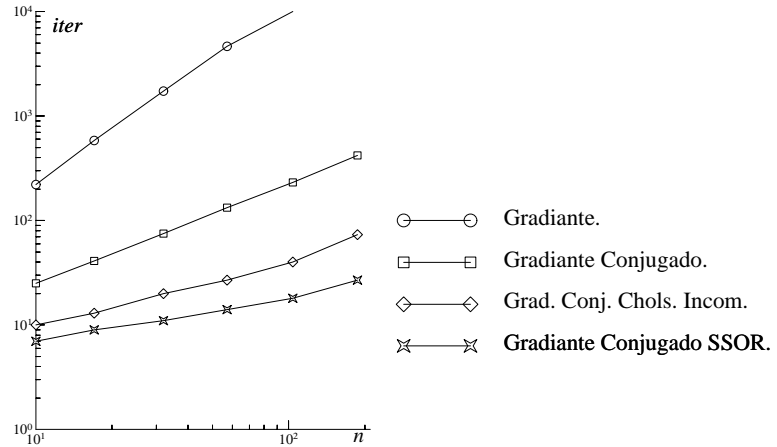


Figura II.4.4. Número de iteraciones vs n .

Finalmente la tabla II.4.1, da el número de iteraciones para alcanzar una precisión de 10^{-6} utilizando el método del gradiente conjugado precondicionado para diferentes valores de n y el factor de relajación ω . En la última columna se agrega, ω optimal en función de n .

Tabla II.4.1. Determinación de ω optimal.

$n \setminus \omega$	0.6	.65	.7	.75	.8	.85	.9	.95	ω_{op}
10	10	9	8	7	8	9	11	12	.75
20	12	12	11	10	9	10	11	14	.8
50	27	24	22	18	17	15	14	16	.9
100	43	35	37	31	27	24	20	18	.95
180	72	59	55	46	40	41	33	27	.95

Ejercicios

- 1.- Mostrar que una matriz simétrica A es definida positiva si y solamente si

$$\det(A_k) > 0 \quad k = 1, 2, \dots, n \quad \text{donde } A_k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix}$$

son los menores principales.

Indicación: Intentar con la descomposición de Cholesky.

- 2.- Calcular el mínimo de

$$f(x) = \frac{1}{50}(x_1, x_2) \begin{pmatrix} 93 & 24 \\ 24 & 107 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \frac{1}{5}(42, 21) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

por el método del gradiente conjugado partiendo del valor inicial $x^0 = 0$. Hacer un buen gráfico de los puntos x^0, x^1, x^2 , de los vectores g^0, g^1, d^0, d^1 , y de las curvas de nivel de la función $f(x)$.

- 3.- Aplicar el método del gradiente conjugado al sistema lineal

$$\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} x = \begin{pmatrix} 3 \\ 2 \\ 3 \end{pmatrix}$$

partiendo de $x^0 = 0$. El método converge en 2 pasos. ¿Por qué?

- 4.- Mostrar, utilizando las fórmulas y las relaciones de ortogonalidad del teorema II.4.6, las expresiones siguientes para β_k y τ_k :

$$\beta_k = \frac{\langle g^{k+1}, g^{k+1} \rangle}{\langle g^k, g^k \rangle}, \quad \tau_k = \frac{\langle g^k, g^k \rangle}{\langle d^k, Ad^k \rangle}.$$

- 5.- Demostrar que la función de la forma

$$f(x) = x^n + \alpha_1 x^{n-1} + \alpha_2 x^{n-2} + \cdots + \alpha_n$$

se separa lo menos posible de cero entre los límites $x = -1$ y $x = +1$, está dada por

$$f(x) = \frac{T_n(x)}{2^n}.$$

6.- Demostrar las relaciones de ortogonalidad

$$\int_{-1}^{+1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}dx = \begin{cases} 0 & \text{si } n \neq m, \\ \frac{\pi}{2} & \text{si } n = m \neq 0, \\ \pi & \text{si } n = m = 0. \end{cases}$$

7.- (Modificación del método del gradiente conjugado para matrices no simétricas o no definidas positivas). Sea

$$Ax = b \quad (\text{II.4.35})$$

un sistema lineal, donde A es una matriz arbitraria inversible. El sistema

$$A^t Ax = A^t b \quad (\text{II.4.36})$$

tiene las mismas soluciones que (II.4.35), pero la matriz $C = A^t A$ es simétrica y definida positiva. Aplicar el método del gradiente conjugado al sistema (II.4.36) y describir el algoritmo obtenido de manera que el producto $A^t A$ no aparezca mas. Cual es la desventaja del nuevo algoritmo si, por mala suerte, se aplica al sistema (II.4.35) que inicialmente era simétrico y definido positivo.

II.5 Mínimos Cuadrados

Hasta la sección precedente, se estudiaron problemas inherentes a la resolución de sistemas de ecuaciones lineales donde la solución existe y es única. En esta sección se estudiarán problemas más generales, donde la existencia y unicidad no juegan un rol tan preponderante.

El problema que se estudiará consiste básicamente en el siguiente. Se tiene como datos:

$$(t_j, y_j), \quad j = 1, \dots, m, \quad (m \text{ grande}), \quad (\text{II.5.1})$$

donde $t_j \in \mathbb{R}$, $y_j \in \mathbb{R}$; y una función de modelo

$$y = \varphi((t, x_1, \dots, x_n)), \quad n \leq m, \quad (\text{II.5.2})$$

donde $t \in \mathbb{R}$, $y \in \mathbb{R}$ y $x_i \in \mathbb{R}$.

Un ejemplo de función de modelo, es la siguiente función:

$$\varphi((t, x_1, x_2)) = x_1 + tx_2 + t^2x_1.$$

El problema consiste en encontrar x_1, \dots, x_n , tales que

$$\varphi((t, x_1, \dots, x_n)) \approx y_j, \quad j = 1, \dots, m. \quad (\text{II.5.3})$$

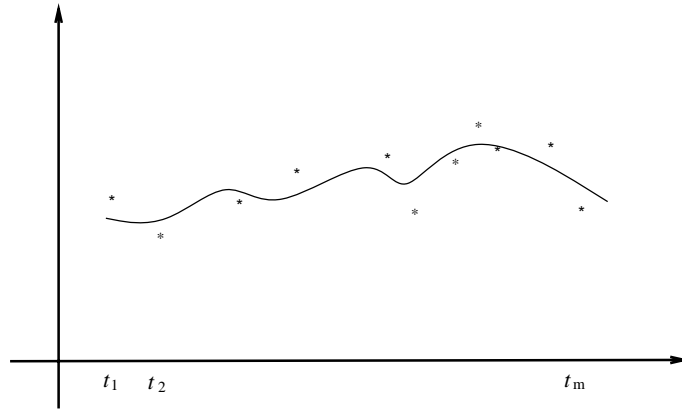


Figura II.5.1 Problema de los mínimos cuadrados.

Supóngase que, φ es lineal respecto a los x_i , es decir

$$\varphi((t, x_1, x_2)) = \sum_{i=1}^n c_i(t)x_i, \quad (\text{II.5.4})$$

de donde el problema se reduce a resolver el siguiente sistema lineal

$$\underbrace{\begin{pmatrix} c_1(t_1) & c_2(t_1) & \cdots & c_n(t_1) \\ c_1(t_2) & c_2(t_2) & \cdots & c_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ c_1(t_m) & c_2(t_m) & \cdots & c_n(t_m) \end{pmatrix}}_A \underbrace{\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}}_x \approx \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}}_b, \quad (\text{II.5.5})$$

por consiguiente, resolver

$$Ax \approx b. \quad (\text{II.5.6})$$

Planteando:

$$e_j = \varphi(t_j, x) - y_j, \quad j = 1, \dots, m; \quad (\text{II.5.7})$$

la solución del problema (II.5.6), mediante el método de los mínimos cuadrados, será aquélla en la que

$$\sum_{i=1}^m e_i^m \longrightarrow \min, \quad (\text{II.5.8})$$

es decir

$$\|Ax - b\| \longrightarrow \min. \quad (\text{II.5.9})$$

La justificación del método de los mínimos cuadrados, está dada por dos interpretaciones. La primera:

Interpretación estadística

Las componentes y_i del vector b en (II.5.5), pueden ser vistas como medidas experimentales. Ahora bien, toda medida experimental lleva consigo un error, por lo tanto es completamente razonable considerar éstas como variables aleatorias.

Puesto que, la mayor parte de las medidas experimentales siguen leyes normales, se tiene las dos hipótesis siguientes, para determinar la solución del problema (II.5.6):

- H1: El valor y_i de (II.5.5) es la realización de un suceso para la variable aleatoria Y_i , $i = 1, \dots, m$. Se supone que los Y_i son independientes y que obedecen la ley normal de esperanza μ_i y varianza σ_i . Para simplificar el problema, se supondrá que los σ_i son iguales y conocidos de antemano.
- H2: El sistema sobredeterminado (II.5.6) posee una solución única, si se reemplaza los y_i por los números μ_i ; es decir, que existe $\xi \in \mathbb{R}^n$ único tal que $A\xi = \mu$ donde $\mu = (\mu_1, \dots, \mu_m)^t$.

La función de distribución de Y_i , en $Y_i = y_i$, es igual a

$$f_i(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2\right).$$

Como los Y_i son independientes, el vector aleatorio $Y = (Y_1, \dots, Y_m)^t$ tiene la función de distribución

$$f(y) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2\right),$$

donde $y = (y_1, \dots, y_m)^t$.

Efectuando cálculos, se obtiene

$$f(y) = C \exp\left(-\frac{1}{2} \sum_{i=1}^m \left(\frac{y_i - \mu_i}{\sigma}\right)^2\right).$$

Aplicando el principio de máxima presunción, la probabilidad de medir y_i en lugar de μ_i , para $i = 1, \dots, m$; está dada por

$$f(y) \rightarrow \max. \quad (\text{II.5.10})$$

Por lo tanto, (II.5.10) sucede si

$$\sum_{i=1}^m \left(\frac{y_i - \mu_i}{\sigma}\right)^2 \rightarrow \min, \quad (\text{II.5.11})$$

es decir

$$\sum_{i=1}^m \left(y_i - \sum_{j=1}^n a_{ij} x_j\right)^2 \rightarrow \min. \quad (\text{II.5.12})$$

con lo que se ha justificado (II.5.7).

Interpretación geométrica

Se define el subespacio vectorial de \mathbb{R}^m de dimensión n dado por

$$E = \{Ax | x \in \mathbb{R}^n\} \subset \mathbb{R}^m,$$

la solución de (II.5.9) está dada por el vector x^0 , tal que $\|Ax^0 - b\|$ es mínima, es decir Ax^0 es el vector más próximo perteneciente a E al vector b , de donde el vector $Ax^0 - b$ es ortogonal al espacio E , ver figura II.5.2.

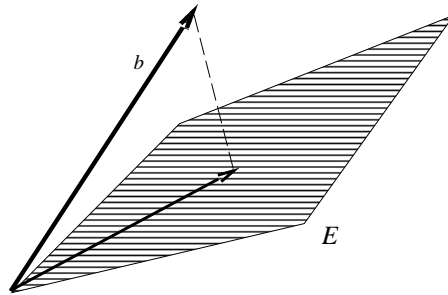


Figura II.5.2 Interpretación geométrica.

Teorema II.5.1.- Sea A una matriz de orden $m \times n$, $b \in \mathbb{R}^m$, entonces:

$$\|Ax - b\|_2 \rightarrow \min \iff A^t Ax = A^t b. \quad (\text{II.5.13})$$

Las ecuaciones de la parte derecha de la equivalencia, son las ecuaciones normales del problema de mínimos cuadrados.

Demostración.- Utilizando la interpretación geométrica del método de mínimos cuadrados se tiene:

$$\begin{aligned} & \|Ax - b\|_2 \longrightarrow \min \\ \iff & b - Ax \perp Ay, \quad \forall y \in \mathbb{R}^n \\ \iff & \langle Ay, b - Ax \rangle = 0, \quad \forall y \in \mathbb{R}^n \\ \iff & y^t A^t (b - Ax) = 0, \quad \forall y \in \mathbb{R}^n \\ \iff & A^t (b - Ax) = 0 \\ \iff & A^t Ax = A^t b. \end{aligned}$$

□

Se remarca inmediatamente que la matriz $A^t A$ es simétrica y semi-definida positiva, es decir $x^t A^t Ax \geq 0$. La matriz $A^t A$ será definida positiva si y solamente las columnas de A son linealmente independientes. Si éste es el caso, se puede utilizar la descomposición de Cholesky para resolver

$$A^t Ax = A^t b,$$

pero en la mayoría de los casos $A^t A$ es una matriz mal condicionada. Como ejemplo vale la pena citar el siguiente ejemplo:

Se tiene, la función de modelo dada por

$$\varphi(t) = \sum_{i=1}^n x_i t^{i-1},$$

es decir, se trata de determinar el polinomio de grado n que mejor ajusta a los puntos (t_i, y_i) , $i = 1, \dots, m$ y $0 = t_1 \leq t_2 \leq \dots \leq t_n = 1$. Por consiguiente, se debe encontrar $x \in \mathbb{R}^n$, tal que

$$\|Ax - b\|_2 \longrightarrow \min,$$

donde:

$$A = \begin{pmatrix} 1 & t_1 & \dots & t_1^{n-1} \\ 1 & t_2 & \dots & t_2^{n-1} \\ \vdots & & & \vdots \\ 1 & t_m & \dots & t_m^{n-1} \end{pmatrix}, \quad b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

Por lo tanto

$$\begin{aligned}
 A^t A &= \begin{pmatrix} m & \sum t_i & \cdots & \sum t_i^{n-1} \\ \sum t_i & \sum t_i^2 & \cdots & \sum t_i^n \\ \vdots & \vdots & \ddots & \vdots \\ \sum t_i^{n-1} & \sum t_i^n & \cdots & \sum t_i^{2n-1} \end{pmatrix} \\
 &= \frac{1}{m} \begin{pmatrix} 1 & 1/m \sum t_i & \cdots & 1/m \sum t_i^{n-1} \\ 1/m \sum t_i & 1/m \sum t_i^2 & \cdots & 1/m \sum t_i^n \\ \vdots & \vdots & \ddots & \vdots \\ 1/m \sum t_i^{n-1} & 1/m \sum t_i^n & \cdots & 1/m \sum t_i^{2n-1} \end{pmatrix} \\
 &\approx m \begin{pmatrix} 1 & 1/2 & \cdots & 1/n \\ 1/2 & 1/2 & \cdots & 1/(n+1) \\ \vdots & \vdots & \ddots & \vdots \\ 1/n & 1/(n+1) & \cdots & 1/2n \end{pmatrix},
 \end{aligned}$$

puesto que

$$\int_0^1 t^k dt \approx \frac{1}{m} \sum t_i^k.$$

De donde, la matriz $A^t A$ se aproxima a una matriz de tipo Hilbert, la cual es una matriz mal condicionada, resultado visto en la primera sección de este capítulo. Por lo tanto se debe formular un algoritmo donde la matriz $A^t A$ no aparezca directamente.

La descomposición QR

Dada una matriz A de orden $m \times n$, se busca las matrices: Q de orden $m \times m$ y la matriz R de orden $m \times n$, tales que la matriz Q sea ortogonal, es decir

$$Q^t Q = I;$$

y la matriz R triangular superior, es decir

$$R = \left(\begin{array}{ccc} \overbrace{\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \\ 0 & & r_{nn} \end{pmatrix}}^n \\ O \end{array} \right) \Bigg\} m, \quad m \geq n;$$

con

$$A = QR. \quad (\text{II.5.14})$$

Puesto que la matriz Q es ortogonal, para todo vector $c \in \mathbb{R}^m$, se tiene

$$\|Q^t c\|_2 = \|c\|,$$

de donde

$$\begin{aligned} \|Ax - b\|_2 &= \|Q^t(Ax - b)\|_2 \\ &= \|Rx - Q^t b\|_2 \longrightarrow \min. \end{aligned} \quad (\text{II.5.15})$$

Por otro lado, la matriz R y el vector $Q^t b$ pueden escribirse como:

$$R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}, \quad Q^t b = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

donde R_1 es una matriz triangular de orden $n \times n$ y $c_1 \in \mathbb{R}^n$. Por consiguiente, resolver el problema (II.5.6) es resolver la ecuación

$$R_1 x = c_1. \quad (\text{II.5.16})$$

Ahora bien, el principal problema es encontrar un algoritmo simple y rapido que permita descomponer la matriz A en un producto de la forma QR . Precisamente uno de estos métodos consiste en la utilización de matrices de *Householder*.

Matrices de Householder (1958)

Sea $u \in \mathbb{R}^m$, con $\|u\|_2 = 1$. Se define la matriz Q por

$$Q = I - 2uu^t, \quad (\text{II.5.17})$$

matriz que es simétrica y ortogonal. En efecto:

$$\begin{aligned} Q^t &= (I - 2uu^t)^t \\ &= (I - 2uu^t), \\ Q^t Q &= (I - 2uu^t)^t (I - 2uu^t) \\ &= I - 2uu^t - 2uu^t + 4uu^t uu^t \\ &= I. \end{aligned}$$

La manera como actúa Q sobre \mathbb{R}^m es la siguiente, sea $x \in \mathbb{R}^m$,

$$Qx = x - 2uu^t x,$$

si $u^t x = 0$, se tiene inmediatamente que $Qx = x$, de donde Q es una simetría respecto al hiperplano $\{x | u^t x = 0\}$, en efecto

$$Qu = (I - 2uu^t)u = -u.$$

En base a estas matrices se construirá un método que permita descomponer la matriz A en QR en $n - 1$ pasos a lo máximo.

Se busca una matriz $Q_1 = I - 2u_1u_1^t$ con $\|u_1\|_2 = 1$, tal que

$$Q_1 A = \begin{pmatrix} \alpha_1 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{pmatrix}.$$

Denotando por e_1 a $(1, 0, \dots, 0)^t$ y A_1 la primera columna de la matriz A , hay que determinar Q_1 , tal que

$$Q_1 A_1 = \alpha_1 e_1, \quad \alpha \in \mathbb{R}. \quad (\text{II.5.18})$$

Por las propiedades de norma se tiene $|\alpha_1| = \|A_1\|_2$, por consiguiente

$$\alpha_1 = \pm \|A_1\|_2, \quad (\text{II.5.19})$$

Q_1 es una matriz de Householder, por lo tanto:

$$\begin{aligned} A_1 - 2u_1u_1^t A_1 &= \alpha_1 e_1, \\ 2u_1 \underbrace{(u_1^t A_1)}_{\in \mathbb{R}} &= A_1 - \alpha_1 e_1, \end{aligned}$$

de donde u_1 tiene la misma dirección que $A_1 - \alpha_1 e_1$. En consecuencia,

$$u_1 = \frac{A_1 - \alpha_1 e_1}{\|A_1 - \alpha_1 e_1\|_2}. \quad (\text{II.5.20})$$

Sabiendo que la sustracción es una operación mal condicionada cuando las cantidades a restar son cercanas, se plantea

$$\alpha_1 = -\text{signo}(a_{11}) \|A_1\|_2. \quad (\text{II.5.21})$$

El siguiente paso es determinar $Q_1 A_j$, donde A_j es la j -ésima columna de la matriz A , definiendo $v_1 = A_1 - \alpha_1 e_1$, se tiene:

$$\begin{aligned} \frac{v_1^t v_1}{2} &= \frac{1}{2} (A_1^t - \alpha_1 e_1^t) (A_1 - \alpha_1 e_1) \\ &= \frac{1}{2} \left(\underbrace{\|A_1\|_2^2}_{\alpha_1^2} - \alpha_1 a_{11} - \alpha_1 a_{11} + \alpha_1^2 \right) \\ &= \alpha_1 (\alpha_1 - a_{11}), \\ Q_1 A_j &= A_j - 2u_1 u_1^t A_j \\ &= A_j - \frac{2}{v_1^t v_1} v_1 v_1^t A_j \\ &= A_j - \frac{1}{\alpha_1 (\alpha_1 - a_{11})} v_1 v_1^t A_j. \end{aligned} \quad (\text{II.5.22})$$

Después del primer paso de la descomposición QR , se ha obtenido

$$Q_1 A = \begin{pmatrix} \alpha_1 & * & \cdots & * \\ 0 & & & \\ \vdots & \bar{A}^{(1)} & & \\ 0 & & & \end{pmatrix}.$$

El segundo paso es determinar \bar{Q}_2 matriz de Householder, como en el primer paso, de manera que

$$Q_2 Q_1 A = \begin{pmatrix} \alpha_1 & * & \cdots & * \\ 0 & & & \\ \vdots & \bar{Q}_2 \bar{A}^{(1)} & & \\ 0 & & & \end{pmatrix} = \begin{pmatrix} \alpha_1 & * & * & \cdots & * \\ 0 & \alpha_2 & * & \cdots & * \\ 0 & 0 & & & \\ \vdots & \vdots & & \bar{A}^{(2)} & \\ 0 & 0 & & & \end{pmatrix},$$

donde

$$Q_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & \bar{Q}_2 & & \\ 0 & & & \end{pmatrix}.$$

Costo de la descomposición QR

La descomposición se la realiza en $n - 1$ etapas,

$$\underbrace{Q_{n-1} \cdots Q_2 Q_1}_{Q^t} A = R,$$

para el primer paso contando el producto escalar se efectúan:

$$m + (n - 1)2m \text{ operaciones,}$$

por lo tanto, el costo es aproximadamente igual

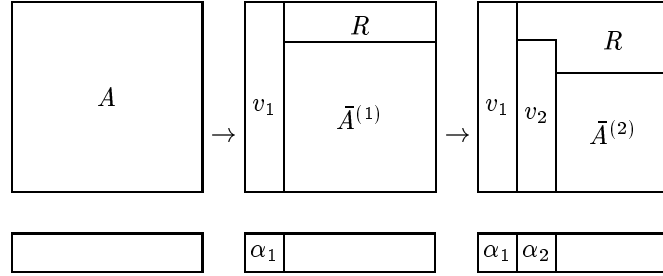
$$2(mn + (m - 1)(n - 1) + \cdots (m - n + 1)),$$

si $n \approx m$ se tiene $\approx \frac{2}{3}n^3$ operaciones; si $n \ll m$ se tiene $\approx mn^2$ operaciones.

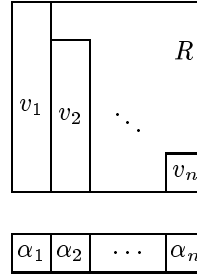
Programación

La descomposición QR presenta la ventaja que se puede utilizar las plazas ocupadas por los coeficientes de la matriz A , con los coeficientes de la matriz

R y los vectores v_1 . En lo que se sigue, se tiene un esquema del método de descomposición QR .



Al final de la descomposición se obtiene la siguiente matriz



Hay que observar que QR es aplicable, si las columnas de A son linealmente independientes. Si no fuesen linealmente independientes se llegaría a la situación de obtener un $\alpha_i = 0$, y la descomposición en este caso sería numéricamente inestable. Para evitar esta situación, se puede modificar la descomposición QR de la manera siguiente. Sea A la matriz a descomponer, se considera

$$\|A_{j_o}\|_2^2 = \max_{j=1, \dots, n} \|A_j\|_2^2, \quad (\text{II.5.23})$$

se intercambia la primera columna A_1 con A_{j_o} y se procede el primer paso de la descomposición QR , para el segundo paso se procede de la misma manera y así sucesivamente. Por consiguiente, con esta modificación se obtiene la sucesión decreciente,

$$\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_k > 0. \quad (\text{II.5.24})$$

Si hay vectores linealmente dependientes se llega por lo tanto al siguiente resultado

$$R = \begin{pmatrix} R_1 & R_2 \\ 0 & 0 \end{pmatrix} \quad \text{con } R_1 = \underbrace{\begin{pmatrix} \alpha_1 & \dots & r_{1n} \\ & \ddots & \vdots \\ & & \alpha_k \end{pmatrix}}_{k = \text{rang} A}. \quad (\text{II.5.25})$$

La descomposición QR es un instrumento numérico que permite determinar el rango de la matriz A . Ahora bien, debido a los errores de redondeo, el principal problema consiste en decidir cuando un α_j es nulo en la sucesión decreciente (II.5.24). Se remarca inmediatamente que, si $\alpha_{k+1} = 0$, entonces $\alpha_{k+i} = 0$ para $i > 2$.

Se define la matriz \hat{R} el resultado numérico obtenido despues de k pasos como

$$\hat{R} = \begin{pmatrix} \alpha_1 & & \\ & \ddots & \hat{R}_2 \\ & 0 & \alpha_k \\ & & & 0 \end{pmatrix}. \quad (\text{II.5.26})$$

Planteando

$$\hat{A} = Q\hat{R}, \quad (\text{II.5.27})$$

se tiene la:

Definición II.5.2.- α_{k+1} es despreciable, si y solamente si

$$\|A - \hat{A}\|_2 \leq \text{eps} \|A\|_2. \quad (\text{II.5.28})$$

Se tiene, por consiguiente:

$$\begin{aligned} A - \hat{A} &= Q(R - \hat{R}), \\ \|A\|_2 &= \|R\|_2, \\ \|A - \hat{A}\|_2 &= \|R - \hat{R}\|_2, \end{aligned}$$

de donde

$$\alpha_{k+1} \text{ despreciable} \iff \|R - \hat{R}\|_2 \leq \text{eps} \|R\|_2,$$

α_{k+1} es una buena aproximación de $\|R - \hat{R}\|_2$, y α_1 es una buena aproximación de $\|R\|_2$, obtenido el siguiente resultado

$$\alpha_{k+1} \text{ despreciable} \iff \alpha_{k+1} \leq \text{eps} \alpha_1. \quad (\text{II.5.29})$$

La pseudo-inversa de una matriz

El problema (II.5.6) tiene una solución única, si el rango de la matriz A es igual a n y está dada por

$$x = (A^t A)^{-1} A^t b. \quad (\text{II.5.30})$$

La matriz

$$A^+ = (A^t A)^{-1} A^t, \quad (\text{II.5.31})$$

será la matriz inversa para el problema (II.5.6).

Para el problema más general, donde el rango de A es $\leq n$, la descomposición QR da el siguiente resultado

$$R = \begin{pmatrix} R_1 & R_2 \\ 0 & 0 \end{pmatrix}, \quad \text{con } R_1 = \begin{pmatrix} \alpha_1 & \cdots & r_{1k} \\ & \ddots & \\ 0 & & \alpha_k \end{pmatrix};$$

planteando $x = (x_1, x_2)^t$, con $x_1 \in \mathbb{R}^k$, se tiene

$$\begin{aligned} \|Ax - b\|_2^2 &= \|Rx - Q^t b\|_2^2 \\ &= \left\| \begin{pmatrix} R_1 x_1 + R_2 x_2 - c_1 \\ -c_2 \end{pmatrix} \right\|_2^2 \\ &= \|R_1 x_1 + R_2 x_2 - c_1\|_2^2 + \|c_2\|_2^2, \end{aligned}$$

de donde el:

Teorema II.5.3.- $x = (x_1, x_2)^t$ es solución de $\|Ax - b\|_2 \rightarrow \min$, si y solamente si

$$R_1 x_1 + R_2 x_2 = c_1. \quad (\text{II.5.32})$$

Si el rango de A es igual a n , la solución es única; si no las soluciones del problema (II.5.6) constituyen un espacio afín. Sea por consiguiente

$$\mathcal{F} = \{x | x \text{ es solución de (II.5.6)}\},$$

de donde, para obtener una solución única, el problema (II.5.6) se convierte en

$$\begin{aligned} \|Ax - b\|_2 &\rightarrow \min \\ \|x\|_2 &\rightarrow \min, \end{aligned} \quad (\text{II.5.33})$$

Definición II.5.4.- La pseudo-inversa de la matriz A de orden $m \times n$ es la matriz A^+ de orden $n \times m$, que expresa la solución x^* para todo $b \in \mathbb{R}^m$ del problema (II.5.33) como

$$x^* = A^+ b. \quad (\text{II.5.34})$$

Si la matriz A es de rango n , entonces la pseudo inversa está dada por la fórmula (II.5.31). Para el caso más general se tiene los siguientes resultados:

$$\begin{aligned} \mathcal{F} &= \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \left| \begin{array}{l} x_2 \text{ arbitrario} \\ x_1 = R_1^{-1}(c_1 - R_2 x_2) \end{array} \right. \right\}, \\ \|x\|_2^2 &= \|x_1\|_2^2 + \|x_2\|_2^2 \\ &= \|R_1^{-1}(c_1 - R_2 x_2)\|_2^2 + \|x_2\|_2^2 \rightarrow \min. \end{aligned} \quad (\text{II.5.35})$$

Derivando (II.5.35), se obtiene

$$\underbrace{(I + (R_2^t R_1^{-t}) R_1^{-1} R_2)}_{\text{simétrica y definida positiva}} x_2 = R_2^t R_1^{-t} R_1^{-1} c_1. \quad (\text{II.3.36})$$

Se ha definido la pseudo-inversa de la matriz A , su existencia está asegurada por el hecho que el problema (II.5.33) tiene solución única, en lo que sigue se determinará de manera explícita esta pseudo-inversa con algunas de sus propiedades más interesantes.

Teorema II.5.5.- Sea A una matriz de $m \times n$, entonces existe dos matrices ortogonales U y V tales que

$$U^t A V = \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_n & \\ & & & 0 \end{pmatrix} \quad \text{con: } \begin{matrix} \sigma_1 \geq \cdots \geq \sigma_k > 0, \\ \sigma_{k+1} = \cdots = \sigma_n = 0. \end{matrix} \quad (\text{II.3.37})$$

(II.3.37) se llama la descomposición a valores singulares de la matriz A y $\sigma_1, \dots, \sigma_n$ son los valores singulares de la matriz A .

Demostración.- La matriz $A^t A$ es simétrica y semi definida positiva, por consiguiente existe una matriz V ortogonal tal que

$$V^t A^t A V = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{pmatrix},$$

con $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. Se busca una matriz U ortogonal tal que

$$A V = U \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & & 0 \end{pmatrix},$$

suponiendo $\sigma_k > 0$ y $\sigma_{k+1} = 0$, se define D la matriz diagonal de $k \times k$ con coeficientes σ_i , $i = 1, \dots, k$; por consiguiente si

$$A V = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} U_1 D & 0 \end{pmatrix},$$

donde U_1 está constituida por las primeras k columnas de U . Sean

$(AV)_1$ las primeras k columnas de AV ,

$(AV)_2$ las otras $n - k$ columnas de AV ,

$(AV)_2 = 0$, en efecto, sea $x = (\underbrace{0, \dots, 0}_k, \hat{x})^t$, donde $\hat{x} \in \mathbb{R}^{n-k}$, obteniendo:

$$x^t V^t A^t A V x = x^t \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{pmatrix} x = 0;$$

$$\|AVx\|_2^2 = \|(AV)_2 \hat{x}\|_2^2, \quad \forall \hat{x}.$$

Se define U_1 por

$$U_1 = (AV)_1 D^{-1},$$

obteniendo columnas ortonormales, ya que

$$U_1^t U_1 = D^{-1} ((AV)_1^t (AV)_1) D^{-1} = I$$

Finalmente se construye U_2 , de manera que U sea ortogonal. \square

Teorema II.5.6.- Sea A una matriz de $m \times n$, siendo

$$U^t A V = \Sigma = \begin{pmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_k & \\ & & & 0 \end{pmatrix}, \quad (\text{II.5.38})$$

la descomposición en valores singulares. Entonces la pseudo-inversa de A está dada por

$$A^+ = V \begin{pmatrix} \sigma_1^{-1} & & & 0 \\ & \ddots & & \\ & & \sigma_k^{-1} & \\ & & & 0 \end{pmatrix} U^t. \quad (\text{II.5.39})$$

Demostración.- Se tiene

$$\begin{aligned} \|Ax - b\|_2^2 &= \|U \Sigma V^t x - b\|_2^2 \\ &= \|U(\Sigma V^t x - U^t b)\|_2^2 = \left\| \Sigma \underbrace{V^t x}_y - \underbrace{U^t b}_c \right\|_2^2 \\ &= \|\Sigma y - c\|_2^2 = \left\| \begin{pmatrix} Dy_1 \\ 0 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right\|_2^2 \\ &= \|Dy_1 - c_1\|_2^2 + \|c_2\|_2^2. \end{aligned}$$

Para que la expresión sea minimal es necesario que $y_1 = D^{-1}c_1$. Por otro lado se tiene que $y = V^t x$, de donde $\|y\|_2 = \|x\|_2$, de manera, que para que x sea minimal es necesario que $y_2 = 0$. De donde

$$\begin{aligned} x &= V \begin{pmatrix} D^{-1}c_1 \\ 0 \end{pmatrix} \\ &= V \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \\ &= V \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^t b \\ &= A^+ b, \quad \forall b. \end{aligned}$$

□

Finalmente se tiene:

Teorema II.5.7.- *La pseudo-inversa de una matriz verifica las siguientes propiedades:*

- a) $(A^+ A)^t = A^+ A,$
- b) $(A A^+)^t = A A^+ ,$
- c) $A^+ A A^+ = A^+ ,$
- d) $A A^+ A = A ,$

llamados axiomas de Moore-Penrose, además la pseudo-inversa está únicamente determinada por estos axiomas.

Demostración.- Verificación inmediata utilizando (II.5.39).

□

Error del Método de los Mínimos Cuadrados

Retomando la interpretación estadística del método de los mínimos cuadrados. Se han dado dos hipótesis de partida, para determinar la solución: la primera sobre los y_i de la ecuación (II.5.5); la segunda concerniente a la compatibilidad de la función modelo (II.5.2) con los datos (II.5.1).

Suponiendo válidas estas dos hipótesis, se tiene

$$A\xi = \mu, \tag{II.5.40}$$

donde $\mu = (\mu_1, \dots, \mu_m)^t$ es la esperanza del vector aleatorio Y . Ahora bien, el método de los mínimos cuadrados da como solución $(x_1, \dots, x_n)^t$, que por (II.5.30), es igual a

$$x = (A^t A)^{-1} A^t b.$$

Por lo tanto

$$x - \xi = (A^t A)^{-1} A^t (b - \mu) \quad \text{o} \quad x_i - \xi_i = \sum_{j=1}^m \alpha_{ij} (b_j - \mu_j), \quad (\text{II.5.41})$$

donde α_{ij} es el elemento (i, j) de la matriz $(A^t A)^{-1} A^t$.

Se supondrá, por lo tanto, que x_i es una realización de una variable aleatoria X_i definida por

$$X_i - \xi_i = \sum_{j=1}^m \alpha_{ij} (Y_j - \mu_j). \quad (\text{II.5.42})$$

Teorema II.5.8.- Sean X_1, X_2, \dots, X_m variables aleatorias independientes, con esperanza μ_i y varianza $\sigma_i = 1$. Entonces, la variable aleatoria X_i , definida por (II.5.42), satisface

$$E(x_i) = \xi_i \quad \text{y} \quad \text{Var}(X_i) = \epsilon_{ii}, \quad (\text{II.5.43})$$

donde ϵ_{ii} es el i -ésimo elemento de la diagonal de $(A^t A)^{-1}$. Los otros elementos de $(A^t A)^{-1}$ son las covarianzas de X_i .

Demostración.- Se tiene, utilizando la linealidad de la esperanza,

$$\begin{aligned} E(X_i) &= \sum_{j=1}^m \alpha_{ij} E(Y_j - \mu_j) + \xi_i \\ &= \xi_i. \end{aligned}$$

Para calcular la varianza de X_i , se utiliza el hecho que $\text{Var}(Z_1 + Z_2) = \text{Var}(Z_1) + \text{Var}(Z_2)$, si Z_1 y Z_2 son independientes. De donde, con e_i el i -ésimo vector de la base canónica, se tiene

$$\begin{aligned} \text{Var}(X_i) &= \text{Var}(X_i - \xi_i) \\ &= \sum_{j=1}^m \alpha_{ij}^2 \text{Var}(Y_j - \mu_j) \\ &= \sum_{j=1}^m \alpha_{ij}^2 \\ &= \|(A^t A)^{-1} A^t e_i\|_2^2 \\ &= \|e_i^t (A^t A)^{-1} A^t\|_2^2 \\ &= e_i^t (A^t A)^{-1} A^t A (A^t A)^{-1} e_i \\ &= e_i^t (A^t A)^{-1} e_i = \epsilon_{ii}. \end{aligned}$$

□

Por consiguiente, si los y_i son realizaciones de una variable aleatoria Y_i que sigue $N(0, 1)$, suponiendo que una solución exacta ξ existe, entonces se tiene una probabilidad de 95% que

$$\xi_i = x_i \pm 2\sigma_{X_i}. \quad (\text{II.5.44})$$

La fórmula (II.5.43) da los valores de σ_{X_i} , sin embargo se ha visto que la determinación de $A^t A$ puede representar una cantidad grande de cálculo. Utilizando la descomposición QR , se tiene

$$A^t A = R^t Q^t Q R = R^t R. \quad (\text{II.5.45})$$

Como las últimas filas de ceros no incide en el cálculo del producto $R^t R$, se puede considerar R como una matriz de $n \times n$, obteniendo así la descomposición de Choleski de $(A^t A)$.

Una vez determinados los parámetros x de la función modelo (II.5.2) corresponde saber, si los (II.5.1) son compatibles con tal modelo. Por la descomposición QR , el problema (II.5.6) se convierte, ver (II.5.15), en

$$\begin{pmatrix} R \\ 0 \end{pmatrix} x = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}, \text{ donde } \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = Q^t b. \quad (\text{II.5.46})$$

Denotando los elementos de Q^t por q_{ij} , los elementos del vector C , satisfacen

$$c_i = \sum_{j=1}^m q_{ij} b_j,$$

y las componentes del vector C_2 satisfacen también

$$c_i = \sum_{j=1}^m q_{ij} (b_j - \mu_j).$$

Es razonable considerar las variables aleatorias

$$Z_i = \sum_{j=1}^m q_{ij} (Y_j - \mu_j), \quad i = n+1, \dots, m. \quad (\text{II.5.47})$$

A continuación, se da la proposición siguiente, sin demostración.

Proposición II.5.9.- Sean Y_1, \dots, Y_m variables aleatorias independientes siguiendo $N(0, 1)$. Entonces las variables aleatorias Z_{n+1}, \dots, Z_m , definidas por (II.5.47) son independientes y siguen también una ley normal $N(0, 1)$.

El error cometido, por el método de los mínimos cuadrados, es equivalente a estudiar $\|C_2\|_2^2$ de donde el teorema, sin demostración:

Teorema II.5.10.- Pearson. Sean Z_1, Z_2, \dots, Z_n variables aleatorias independientes siguiendo la ley normal $N(0, 1)$. Entonces, la distribución de la variable aleatoria

$$Z_1^2 + \dots + Z_n^2, \quad (\text{II.5.48})$$

está dada por

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \quad x > 0; \quad (\text{II.5.49})$$

y por $f_n(x) = 0$ para $x \leq 0$. Es la ley χ^2 con n grados de libertad. La esperanza de esta variable es n y su varianza es $2n$.

Ejemplo

A partir de la función $f(t) = t + 0.5 \sin(\pi t/2)$, se ha elaborado la tabla II.5.1, que tiene como elementos (i, b_i) con $0 \leq i \leq 9$, donde $b_i = f(i) + \epsilon_i$. ϵ_i es la realización de una variable aleatoria siguiendo una ley normal $N(0, \sigma)$ con $\sigma = 0.1$.

Tabla II.5.1. Valores de b_i en función de i .

i	b_i	i	b_i
0	0.11158	5	5.5158
1	1.2972	6	6.0119
2	2.07201	7	6.6802
3	2.4744	8	7.9294
4	3.963	9	9.4129

Se considera, la función de modelo

$$\varphi(t) = xt + y \sin(\pi t/2). \quad (\text{II.5.50})$$

Obteniendo por la descomposición QR :

$$\begin{aligned} x &= 1.00074, \\ y &= 0.413516. \end{aligned}$$

El siguiente paso será estimar el intervalo de confianza para x y y . Para tal efecto, se considera el problema

$$x \frac{t}{\sigma} + y \frac{\sin(\pi t/4)}{\sigma} = \frac{b_i}{\sigma}, \quad i = 0, \dots, 9;$$

pues, los b_i/σ deben ser realizaciones de una variable normal con varianza igual a 1. De donde la matriz de covarianza está dada por

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} 3.57143 \cdot 10^{-5} & -3.57143 \cdot 10^{-5} \\ -3.57143 \cdot 10^{-5} & 2.03571 \cdot 10^{-3} \end{pmatrix}.$$

Por consiguiente:

$$x = 1.00074 \pm 0.012,$$

$$y = 0.41352 \pm 0.09.$$

El método QR , con la notación precedente da

$$\|C_2\|^2 = 0.0693103;$$

corrigiendo, para el problema normalizado, se tiene

$$\frac{\|C_2\|_2^2}{\sigma_2} = 6.93103.$$

Se tiene $10 - 2$ grados de libertad. Viendo en una tabla de la distribución χ^2 , se deduce que este valor es lo suficientemente pequeño para ser probable.

Ahora bien, si se hubiese considerado la función de modelo

$$\varphi(t) = xt + y, \quad (\text{II.5.51})$$

se hubiera encontrado

$$\frac{\|C_2\|_2^2}{\sigma_2} = 90.5225,$$

valor demasiado grande para ser probable.

La conclusión es que, para los valores de la tabla II.5.1, la ley (II.5.50) es más probable que la ley (II.5.51).

En la figura II.5.3, puede observarse en línea continua la gráfica de la función modelo dada por (II.5.50) y con líneas segmentadas la gráfica de la función modelo (II.5.1).

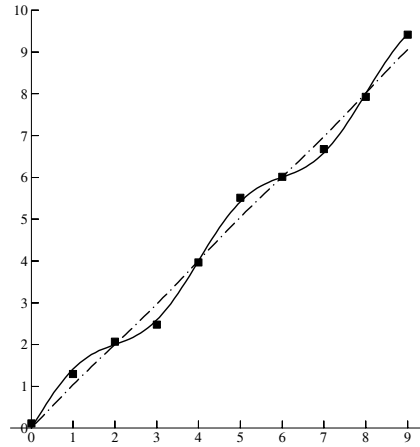


Figura II.5.3. Resultados del método de los mínimos cuadrados.

Ejercicios

- 1.- Sea A una matriz invertible de $n \times n$. Mostrar que la descomposición QR es única, si se supone que $r_{jj} > 0$ para $j = 1, \dots, n$.
- 2.- Aplicar el algoritmo de Golub-Householder a la matrice de rotación

$$A = \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}.$$

Dar una interpretación geométrica.

- 3.- Sea Q una matriz ortogonal de $n \times n$. Mostrar que Q puede escribirse como el producto de n matrices de Householder, de donde cada transformación ortogonal de \mathbb{R}^n es una sucesión de al menos n reflexiones.
- 4.- Escribir una subrutina $\text{DECQR}(N, M, \text{MDIM}, A, \text{ALPH})$, $(A(\text{MDIM}, N), \text{ALPH}(N))$, que calcula la descomposición QR de una matriz $m \times n$, $m \geq n$.
Escribir también una subrutina $\text{SOLQR}(N, M, \text{MDIM}, A, \text{ALPH}, B)$ que calcula la solución del problema

$$\|AX - b\|_2 \longrightarrow \min.$$

Determinar la parábola $x_1 + x_2 t + x_2 t^2$, que ajusta lo mejor posible:

t_i	0	0,2	0,4	0,6	0,8	1.0
y_i	0,10	0,15	0,23	0,58	0,45	0,60

- 5.- Sea A una matriz $m \times n$. Mostrar que $A^t A + \rho I$ es no singular para $\rho > 0$ y

$$A^+ = \lim_{\rho \rightarrow 0_+} (A^t A + \rho I)^{-1} A^t.$$

Capítulo III

Interpolación

Uno de los mayores problemas con que se tropieza, radica en la evaluación de las funciones en determinados puntos. Las causas principales de estas dificultades están esencialmente en la difícil manipulación de estas funciones; una función tan inocente como la logaritmo presenta dificultades en su evaluación y se debe recurrir a útiles matemáticos como series, etc. Por otro lado, la sola información que se conoce de esta función, son determinados puntos, e incluso suponiendo que ésta es lo bastante regular, la evaluación en otros puntos es de gran dificultad. De donde un método de fácil manipulación consiste en la aproximación de las funciones a evaluar por polinomios. Existen dos enfoques diferentes, el primero radica en la aproximación mediante polinomios de interpolación y el otro mediante la mejor aproximación polinomial respecto a una norma.

Este capítulo abordará sobre todo la aproximación mediante polinomios. La primera parte tratará sobre la interpolación de Lagrange y Hermite, la formulación de algoritmos para calcular estos polinomios será hecha, las estimaciones de error cometido serán también estudiadas a fondo, como así mismo los problemas de estabilidad inherentes a la interpolación de Lagrange serán abordados con resultados y ejemplos, como elementos teóricos los polinomios de Chebichef serán vistos dando como consecuencia resultados interesantes sobre la elección de los puntos de interpolación. Un segundo tópico a ver será los *splines*, polinomios que operan como serchas, por su fácil manipulación, como por su gran utilidad serán estudiados en la formulación de los métodos, el cálculo de error y sus diferentes aplicaciones. Un tercer tema de gran interés por las grandes utilidades que puede dar, consiste en los métodos de extrapolación tanto como en su valor teórico, como también por su propiedad de acelerar la convergencia de muchas sucesiones.

III.1 Interpolación de Lagrange

En la introducción del capítulo, se mencionó la interpolación como un instrumento de aproximación de una función determinada. Sea f una función de la que se conoce las imágenes en una cantidad finita de puntos, es decir

$$f(x_i) = y_i \quad i = 1, \dots, n;$$

la función interpolante p será por definición igual a $f(x_i)$ para $i = 1, \dots, n$. Se hablará de interpolación cuando se evalúa en el punto x con

$$x \in [\min x_i, \max x_i],$$

y de extrapolación si no. Por su fácil manipulación, pues solo se efectúan multiplicaciones y adiciones, es conveniente que la función interpolante p sea un polinomio.

Definición III.1.1.- Dados $(k+1)$ puntos diferentes: x_0, x_1, \dots, x_k de $[a, b]$, los enteros no negativos $\alpha_0, \dots, \alpha_k$ y los reales y_{il_i} con $0 \leq i \leq k$ y $0 \leq l_i \leq \alpha_i$. El Polinomio de Hermite p_n de grado $n = k + \alpha_0 + \dots + \alpha_k$, respecto a los x_i, α_i y los y_{il_i} verifica

$$p_n^{(l_i)}(x_i) = y_{il_i}. \quad (\text{III.1.1})$$

Si los $\alpha_i = 0$ para $i = 0, \dots, k$, p_n se llama Polinomio de Lagrange.

Definición III.1.2.- Cuando los puntos y_{il_i} satisfacen

$$y_{il_i} = f_n^{(l_i)}(x_i), \quad (\text{III.1.2})$$

para una determinada función f , p_n es el polinomio de interpolación de Hermite, y si los α_i son nulos, se hablará del polinomio de interpolación de Lagrange.

Estas dos definiciones dan las condiciones que deben cumplir los polinomios de interpolación, sin embargo la existencia, como la unicidad no están dadas, podría suceder que no existiesen en algún caso. Por eso es necesario insistir en la base teórica que asegure la existencia y la unicidad de estos, además, para que la construcción de estos pueda ser relativamente fácil.

Bases Teóricas

Sin querer dar un tratado sobre la teoría de los polinomios, existen resultados que deben formularse para la mejor comprensión de la interpolación polinomial. Aunque en Álgebra se hace distinción entre polinomio y su función

polinomial asociada, no se hará distinción entre estos, por que son polinomios a coeficientes reales o complejos los que se utilizan en la mayor parte de los problemas a resolverse numéricamente.

Definición III.1.3.- Sea $p(x) \in \mathbb{R}[x]$, a es un cero de multiplicidad m de $p(x)$, si

$$p^{(k)}(a) = 0 \quad m = 0, \dots, m-1, \quad \text{y } p^{(m)} \neq 0.$$

Proposición III.1.4.- a es un cero de multiplicidad m de $p(x)$, si y solamente si, existe un polinomio $q(x)$ con $q(a) \neq 0$ tal que

$$p(x) = (x - a)^m q(x).$$

Demostración.- Ver en cualquier libro de Algebra. □

Corolario III.1.5.- Un polinomio $p_n(x)$ de grado n tiene a lo más n ceros contando con su multiplicidad.

Teorema III.1.6.- Existe a lo sumo un polinomio de Hermite de grado n respecto $x_0, \dots, x_k, \alpha_0, \dots, \alpha_k$, con $k + \alpha_0 + \dots + \alpha_k = n$, tal que

$$p_n^{(l_i)}(x_i) = y_{il_i},$$

donde $y_{il_i} \in \mathbb{R}$, $0 \leq l_i \leq \alpha_i$.

Demostración.- Sean p_n y q_n dos polinomios de Hermite que satisfacen las hipótesis del teorema. $p_n - q_n$ es un polinomio nulo o de grado $\leq n$. Ahora bien, $p_n - q_n$ tiene ceros de al menos multiplicidad $\alpha_i + 1$ en x_i para $i = 0, \dots, k$. Por la proposición anterior se tiene

$$p_n - q_n = \left(\prod_{i=0}^k (x - x_i)^{\alpha_i + 1} \right) r(x),$$

sumando los grados, la única posibilidad es que $r(x) = 0$. □

Teorema III.1.7.- Existe al menos un polinomio de Hermite de grado n respecto $x_0, \dots, x_k, \alpha_0, \dots, \alpha_k$, con $k + \alpha_0 + \dots + \alpha_k = n$, tal que

$$p_n^{(l_i)}(x_i) = y_{il_i},$$

donde $y_{il_i} \in \mathbb{R}$, $0 \leq l_i \leq \alpha_i$.

Demostración.- Es suficiente mostrar que existe un polinomio de Hermite de grado n tal que para $i \in \{0, \dots, k\}$ y $l \in \{0, \dots, \alpha_i\}$ se tenga:

$$p^{(l)}(x_i) = 1, \quad p^{(m)}(x_j) = 0 \text{ para } m \neq l \text{ o } j \neq i,$$

denótese por $p_{i,l}$ este polinomio. x_j para $j \neq i$ es un cero de multiplicidad $\alpha_j + 1$, de donde:

$$p_{i,l} = r(x) \left(\prod_{j \neq i} (x - x_j)^{\alpha_j + 1} \right),$$

$$p_{i,\alpha_i} = C_{i,\alpha_i} (x - x_i)^{\alpha_i} \left(\prod_{j \neq i} (x - x_j)^{\alpha_j + 1} \right),$$

la constante C_{i,α_i} es escogida de manera que $p_{i,\alpha_i}^{(\alpha_i)}(x_i) = 1$. Los polinomios $p_{i,l}$ se los define recursivamente de la siguiente manera

$$p_{i,l} = C_{i,l} \left((x - x_i)^{\alpha_i} \left(\prod_{j \neq i} (x - x_j)^{\alpha_j + 1} \right) - \sum_{m > l} c_{l,m}^i p_{i,m} \right),$$

donde las constantes $c_{l,m}^i$ son escogidas de manera que $p_{i,l}^{(m)}(x_i) = 0$ para $m > l$ y $C_{i,l}$ de manera que $p_{i,l}^{(l)} = 1$. \square

Para construir el polinomio de interpolación no debe utilizarse los polinomios definidos en la demostración, uno porque el costo en operaciones es demasiado elevado y otro por la sensibilidad de estos polinomios así definidos a los errores de redondeo.

Construcción del Polinomio de Interpolación

Inicialmente se considerará polinomios de interpolación del tipo de Lagrange. Por lo tanto, dados los puntos x_0, \dots, x_n todos diferentes, y los valores y_0, \dots, y_n , el problema se reduce a encontrar un polinomio de grado n que satisfaga

$$p(x_i) = y_i \quad 0 \leq i \leq n. \quad (\text{III.1.3})$$

Indudablemente el caso más sencillo es para $n = 1$ y luego para $n = 2$. Para $n = 1$ la recta de interpolación está dada por

$$p(x) = y_0 + \left(\frac{y_1 - y_0}{x_1 - x_0} \right) (x - x_0),$$

para $n = 2$ la parábola de interpolación está dada por

$$p(x) = y_0 + \left(\frac{y_1 - y_0}{x_1 - x_0} \right) (x - x_0) + a(x - x_0)(x - x_1),$$

este último polinomio verifica $p(x_0) = y_0$ y $p(x_1) = y_1$, por consiguiente es necesario determinar a de manera que $p(x_2) = y_2$. Unos simples cálculos algebraicos dan

$$a = \frac{1}{x_2 - x_0} \left[\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_2} \right].$$

Para $n = 3$ se puede continuar con el mismo esquema, pero antes es necesario dar la noción de diferencias divididas.

Definición III.1.8.- (*Diferencias Divididas*) Sean $(x_0, y_0), \dots, (x_n, y_n)$, los x_i diferentes entre si. Entonces las diferencias divididas de orden k se definen de manera recursiva por:

$$y[x_i] = y_i, \quad (\text{III.1.4a})$$

$$y[x_i, x_j] = \frac{y_j - y_i}{x_j - x_i}, \quad (\text{III.1.4b})$$

$$y[x_{i_0}, \dots, x_{i_k}] = \frac{y[x_{i_1}, \dots, x_{i_k}] - y[x_{i_0}, \dots, x_{i_{k-1}}]}{x_{i_k} - x_{i_0}}. \quad (\text{III.1.4c})$$

Teorema III.1.9.- (*Fórmula de Newton*). El polinomio de interpolación de grado n que pasa por (x_i, y_i) , $i = 0, \dots, n$, está dado por

$$p(x) = \sum_{i=0}^n y[x_0, \dots, x_i] \prod_{k=0}^{i-1} (x - x_k), \quad (\text{III.1.5})$$

con la convención $\prod_{k=0}^{-1} (x - x_k) = 1$.

Demostración.- Por inducción sobre n . Para $n = 1$ y $n = 2$ la fórmula de Newton es correcta, ver más arriba. Se supone que la fórmula sea correcta para $n - 1$.

Sea, $p(x)$ el polinomio de grado n que pasa por (x_i, y_i) , $i = 0, \dots, n$; de donde

$$p(x) = p_1(x) + a(x - x_0)(x - x_1) \cdots (x - x_{n-1}),$$

con $p_1(x)$ el polinomio de interpolación de grado $n - 1$ que pasa por (x_i, y_i) $i = 0, \dots, n - 1$. Por hipótesis de inducción se tiene

$$p_1(x) = \sum_{i=0}^{n-1} y[x_0, \dots, x_i] \prod_{k=0}^{i-1} (x - x_k),$$

por lo tanto hay que demostrar que

$$a = y[x_0, x_1, \dots, x_n].$$

Sea, $p_2(x)$ el polinomio de interpolación de grado $n-1$ que pasa por (x_i, y_i) , $i = 1, \dots, n$; definiendo el polinomio $q(x)$ por

$$q(x) = p_2(x) \frac{(x - x_0)}{(x_n - x_0)} + p_1(x) \frac{(x_n - x)}{(x_n - x_0)},$$

$q(x)$ es un polinomio de grado a lo sumo n , además $q(x_i) = y_i$, $i = 0, \dots, n$. Por la unicidad del polinomio de interpolación se tiene que $p(x) = q(x)$. Comparando los coeficientes del término de grado n se obtiene

$$a = \frac{y[x_1, \dots, x_n] - y[x_0, \dots, x_{n-1}]}{x_n - x_0} = y[x_0, \dots, x_n]$$

□

La fórmula de Newton es muy simple y facil de implementar en un programa para determinar el polinomio de interpolación de tipo Lagrange, ver la figura III.1.1 donde se muestra un esquema del algoritmo de Newton.

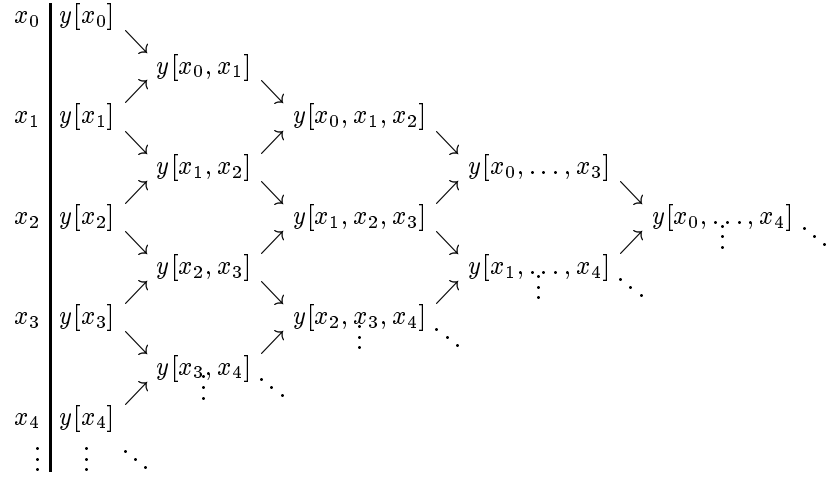


Figura III.1.1. Esquema de la Fórmula de Newton.

La programación del polinomio de interpolación es muy facil, además que se puede utilizar la plaza ocupada por y_i . En efecto la primera columna del esquema es utilizada por los y_i , se remarca inmediatamente que y_n aparece una vez en los calculos, por consiguiente las últimas $n-1$ lugares de la columna son utilizados por los $n-1$ valores de las diferencias divididas que aparecen en la segunda columna, quedando el primer valor de y_0 , y se continua de esta manera hasta obtener $y[x_0, \dots, x_n]$.

Un caso particular en la determinación del polinomio de interpolación ocurre cuando la subdivisión de los puntos es uniforme, es decir

$$x_i = x_0 + ih, \quad i = 0, \dots, n; \quad (\text{III.1.6})$$

donde $h = (x_n - x_0)/n$. Para poder implementar el algoritmo de Newton con esta subdivisión es necesario, las siguientes dos definiciones.

Definición III.1.10.- (*Diferencias Finitas Progresivas*) Sea y_0, y_1, \dots , una sucesión de números reales, se define el operador de diferencias finitas progresivas de orden k de manera recursiva como sigue:

$$\nabla^0 y_i = y_i, \quad (\text{III.1.7a})$$

$$\nabla y_i = y_{i+1} - y_i, \quad (\text{III.1.7b})$$

$$\nabla^{k+1} y_i = \nabla^k y_{i+1} - \nabla^k y_i. \quad (\text{III.1.7b})$$

Definición III.1.11.- (*Diferencias Finitas Retrógradas*) Sea y_0, y_1, \dots , una sucesión de números reales, se define el operador de diferencias finitas retrógrada de orden k de manera recursiva como sigue:

$$\bar{\nabla}^0 y_i = y_i, \quad (\text{III.1.8a})$$

$$\bar{\nabla} y_i = y_i - y_{i-1}, \quad (\text{III.1.8b})$$

$$\bar{\nabla}^{k+1} y_i = \bar{\nabla}^k y_i - \bar{\nabla}^k y_{i-1}. \quad (\text{III.1.8b})$$

Ahora bien, tomando una subdivisión uniforme $x_0 < \dots < x_n$ se tiene los dos resultados equivalentes para el polinomio de interpolación de Lagrange que pasa por los puntos (x_i, y_i) . Utilizando las diferencias finitas progresivas se tiene

$$p(x) = \sum_{i=0}^n \frac{\nabla^i y_0}{i! h^i} \prod_{k=0}^{i-1} (x - x_k), \quad (\text{III.1.9})$$

donde $h = (x_n - x_0)/n$ y con la convención $\prod_{k=0}^{-1} (x - x_k) = 1$. Utilizando las diferencias retrógradas se tiene

$$p(x) = \sum_{i=0}^n \frac{\bar{\nabla}^i y_n}{i! h^i} \prod_{k=0}^{i-1} (x - x_{n-k}), \quad (\text{III.1.10})$$

con la convención $\prod_{k=0}^{-1} (x - x_{n-k}) = 1$. La programación es la misma que para el caso general, con la única diferencia que no se efectúan divisiones. Por

consiguiente, se tiene el mismo esquema de resolución para una subdivisión uniforme. En la figura III.1.2 se observa los esquemas para las diferencias finitas progresivas y las diferencias finitas retrógradas.

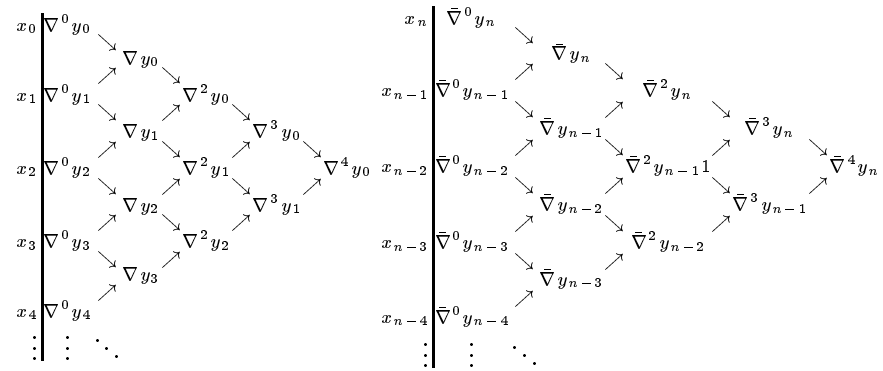


Figura III.1.2. Esquemas para Diferencias Finitas

La evaluación del polinomio de interpolación en un punto $x \in [x_0, x_n]$ se realiza utilizando el algoritmo de Horner, ver ejemplo sección I.1. Para ver la simplicidad de la determinación del polinomio de interpolación utilizando diferencias divididas, y si la subdivisión es uniforme, diferencias finitas; se tiene el siguiente:

Ejemplo

Se desea evaluar la raíz cuadrada de un número positivo, por ejemplo 1, 4. Para mostrar la eficacia y simplicidad, se va a construir un polinomio de interpolación de grado 3, y con tal efecto se utiliza los puntos donde la raíz cuadrada es exacta como una expresión decimal.

x_i	1,00	1,21	1,44	1,69
y_i	1,00	1,10	1,20	1,30

Por consiguiente, el esquema de diferencias divididas para el polinomio de interpolación, está dado por

1,00	1,00			
		,4762		
1,21	1,10		,0941	
		,4348		,0313
1,44	1,20		,0725	
		,400		
1,69	1,30			

de donde el polinomio de interpolación es igual a

$$p(x) = 1 + 0,476(x-1) - 0,094(x-1)(x-1,21) \\ + 0,031(x-1)(x-1,21)(x-1,44),$$

y $p(1,4) = 1,183$.

El Error de Interpolación

Sea $f(x)$ una función que cumple ciertas condiciones, se supone que se hace pasar un polinomio de interpolación por los puntos $(x_i, f(x_i))$, la pregunta natural es saber que error se comete con el cálculo del polinomio de interpolación, es decir $p(x) - f(x)$ vale cuanto, para un x determinado. Es por esta razón que los siguientes teoremas serán enunciados para tener una idea del error cometido.

Teorema III.1.12.- Sea $f(x)$ n -veces continuamente diferenciable y $y_i = f(x_i)$, $i = 0, \dots, n$; los x_i todos diferentes. Entonces existe $\xi \in (\min x_i, \max x_i)$ tal que

$$y[x_0, x_1, \dots, x_n] = \frac{f^{(n)}(\xi)}{n!}. \quad (\text{III.1.11})$$

Demostración.- Se define la función $r(x)$ por

$$r(x) = f(x) - p(x),$$

donde $p(x)$ es el polinomio de interpolación que pasa por $(x_i, f(x_i))$, $i = 0, \dots, n$. Se observa inmediatamente que $r(x_i) = 0$ para $i = 0, \dots, n$; por el teorema de Rolle se deduce que $r'(x)$ se anula al menos una vez en cada subintervalo $[x_i, x_{i+1}]$, es decir existe $\xi_{i1} \in (x_i, x_{i+1})$, para $i = 0, \dots, n$. Aplicando una vez más Rolle se deduce que $r''(x)$ tiene $n-1$ ceros en el intervalo $[x_0, x_n]$, finalmente aplicando el teorema de Rolle las veces que sean necesarias se llega a la conclusión que $r^{(n)}$ tiene al menos un cero en el intervalo requerido. Por otro lado

$$r^{(n)}(x) = f^{(n)}(x) - n!y[x_0, \dots, x_n].$$

□

Teorema III.1.13.- Sea, f $n+1$ veces continuamente diferenciable y $p(x)$ el polinomio de interpolación de grado n definido por $(x_i, f(x_i))$, $i = 0, \dots, n$ y $x_i \neq x_j$ si $i \neq j$. Entonces $\forall x, \exists \xi \in (\min(x_0, \dots, x_n, x), \max(x_0, \dots, x_n, x))$, tal que

$$f(x) - p(x) = (x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(\xi)}{(n+1)!}. \quad (\text{III.1.12})$$

Demostración.- Se deja x fijo, se plantea $x_{n+1} = x$. Sea $\bar{p}(x)$ el polinomio de interpolación de grado $n + 1$ que pasa por $(x_i, f(x_i))$, $i = 0, \dots, n + 1$. Por consiguiente

$$\bar{p}(x) = p(x) + \prod_{k=0}^n (x - x_k) y[x_0, \dots, x_{n+1}],$$

por el teorema anterior, se tiene

$$\bar{p}(x) = p(x) + \prod_{k=0}^n (x - x_k) \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

$\bar{p}(x) = f(x)$, de donde el teorema queda demostrado. \square

Ejemplo

Para el ejemplo de la implementación del método de diferencias divididas para calcular el valor $\sqrt{1,4}$, la función interpolada es \sqrt{x} . El intervalo de estudio del polinomio de interpolación de grado 3 está dado por $[1; 1,69]$, $x_0 = 1$, $x_1 = 1,21$, $x_2 = 1,44$ y $x_3 = 1,69$. Por lo tanto

$$\begin{aligned} f(x) &= \sqrt{x}, \\ f'(x) &= -\frac{1}{2}x^{-\frac{1}{2}}, \\ f''(x) &= \frac{1}{4}x^{-\frac{3}{2}}, \\ f^{(3)}(x) &= -\frac{3}{8}x^{-\frac{5}{2}}, \\ f^{(4)}(x) &= \frac{15}{16}x^{-\frac{7}{2}}, \end{aligned}$$

por consiguiente $|f^{(4)}(x)| \leq \frac{15}{16}$ para $x \in [1; 1,69]$, lo cual implica que el error cometido en el cálculo de \sqrt{x} está dado por:

$$\begin{aligned} |\sqrt{x} - p(x)| &\leq \frac{15}{16 \cdot 4!} |(x-1)(x-1,21)(x-1,44)(x-1,69)|, \\ |\sqrt{1,4} - p(1,4)| &\leq 3,45 \cdot 10^{-5}. \end{aligned}$$

El teorema II.1.13 da un resultado sobre el error cometido durante el proceso de interpolación. Este error depende de la derivada número $n + 1$ de la función f , éste depende por lo tanto de la función a interpolar y está

fuera de alcance. Pero también el error depende de la subdivisión x_0, \dots, x_n , por lo tanto en cierta manera es controlable. De ahí una pregunta natural surge: Sea $[a, b]$ un intervalo. ¿Cómo escoger x_0, x_1, \dots, x_n , tales que

$$\max_{x \in [a, b]} |(x - x_0)(x - x_1) \cdots (x - x_n)| \longrightarrow \min? \quad (\text{III.1.13})$$

Polinomios de Chebichef

La respuesta de la anterior interrogante, está en el estudio de los polinomios de Chebichef, que ya fueron tratados en la sección II.4.

Definición III.1.14.- El n -simo polinomio de Chebichef está definido en el intervalo $[-1, 1]$ por la siguiente relación

$$T_n(x) = \cos(n \arccos x). \quad (\text{III.1.14})$$

Por lo tanto $T_0(x) = 1$, $T_1(x) = x$, etc.

Los polinomios de Chebichef tienen las siguientes propiedades, que vale la pena enunciarlas en la siguiente proposición.

Proposición III.1.15.- Los polinomios de Chebichef satisfacen:

a) $T_0(x) = 1$, $T_1(x) = x$ y

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \quad (\text{III.1.15})$$

b) Se tiene para n entero no negativo:

$$|T_n(x)| \leq 1, \quad \text{para } x \in [-1, 1]; \quad (\text{III.1.16})$$

$$T_n\left(\cos\left(\frac{k\pi}{n}\right)\right) = (-1)^k; \quad (\text{III.1.17})$$

$$T_n\left(\cos\left(\frac{2k+1}{2n}\pi\right)\right) = 0; \quad (\text{III.1.18})$$

para $k = 0, 1, \dots, n-1$.

Demostración.- El inciso a) se demuestra utilizando el hecho que

$$\cos((n+1)\varphi) = 2\cos\varphi \cos(n\varphi) - \cos((n-1)\varphi).$$

La primera parte del inciso b) es consecuencia de la definición del n -simo polinomio de Chebichef. Las dos últimas relaciones provienen de las ecuaciones $\cos n\varphi = 0$ y $|\cos n\varphi| = 1$. \square

Finalmente, es facil observar que el coeficiente del término dominante de T_n para $n > 0$ es igual a 2^{n-1} . Los polinomios de Chebichef T_0, T_1, T_2 y T_3 pueden observarse en la figura III.1.3.

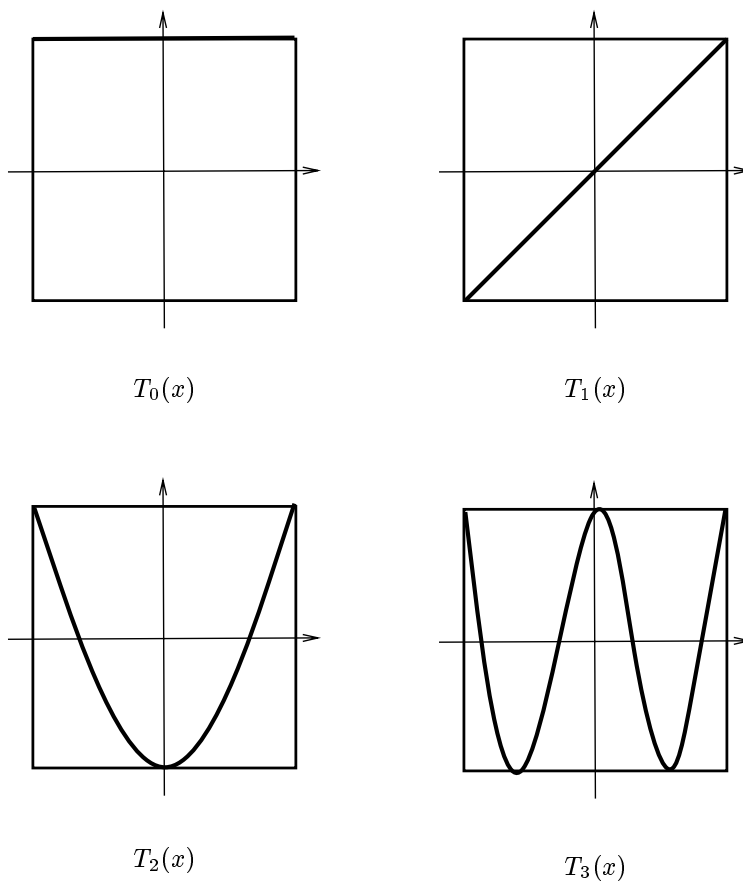


Figura III.1.3. Los cuatro primeros polinomios de Chebichef.

Teorema III.1.16.- Sea $q(x)$ un polinomio de grado n con coeficiente dominante 2^{n-1} para $n \geq 1$ y $q(x)$ diferente a $T_n(x)$. Entonces

$$\max_{x \in [-1, 1]} |q(x)| > \max_{x \in [-1, 1]} |T_n(x)| = 1. \quad (\text{III.1.19})$$

Demostración.- La demostración se la realiza por el absurdo. Se supone que existe un polinomio $q(x)$ de grado n , tal que

$$|q(x)| \leq 1 \quad \text{para } |x| \leq 1,$$

de donde $r(x) = q(x) - T_n(x)$ polinomio no nulo de grado al menos $n-1$, pero $r(x)$ posee al menos n raíces, lo que contradice que $r(x)$ sea un polinomio de grado menor o igual $n-1$. \square

Teorema III.1.17.- Si $x_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right)$, entonces

$$\max_{x \in [-1, 1]} |(x - x_0) \cdots (x - x_n)|$$

es minimal respecto a todas las divisiones $x_0 < x_1 < \dots < x_n$ con $x_i \in [-1, 1]$.

Demostración.- Se tiene:

$$\begin{aligned}(x - x_0) \cdots (x - x_n) &= T_{n+1}(x) 2^{-n}, \\ (x - \bar{x}_0) \cdots (x - \bar{x}_n) &= q(x) 2^{-n}.\end{aligned}$$

\square

Para construir una división óptima para cualquier intervalo $[a, b]$, se toma los puntos \hat{x}_k definidos por

$$\hat{x}_k = \frac{a+b}{2} + \frac{b-a}{2} x_k, \quad (\text{III.1.20})$$

donde los x_k son los ceros del $n+1$ -simo polinomio de Chebichef.

Estudio de los Errores de Redondeo

En esta subsección, se estudiará la incidencia de los errores de redondeo en la determinación del polinomio de interpolación, más precisamente en los polinomios de interpolación de tipo Lagrange. Hay que recalcar que los errores de redondeo en el cálculo del polinomio interpolante está muy relacionado con el concepto de estabilidad del algoritmo de determinación del polinomio de interpolación.

El problema es el siguiente: Dada una subdivisión x_0, \dots, x_n , determinar $p(x)$ de grado n tal que $p(x_i) = y_i$, $i = 0, \dots, n$. En el cálculo se cometen errores de redondeo por un lado, y por otro lado los datos iniciales tienen una parte de error de redondeo. Para simplificar el estudio de la estabilidad del polinomio de interpolación, los errores de redondeo a considerar son aquellos presentes en los y_i , y se supone que no se comete errores de redondeo en los cálculos propiamente dichos y que los x_i son considerados por su valor exacto.

El teorema III.1.7 afirma la existencia de un polinomio de interpolación de grado n de tipo Lagrange para (x_i, y_i) , $i = 0, \dots, n$ y los x_i diferentes,

durante la demostración de este teorema se vio de manera explícita la forma que debía tener este polinomio, recordando se enuncia el siguiente teorema:

Teorema III.1.18.- Sean (x_i, y_i) , $i = 0, 1, \dots, n$, los x_i diferentes entre si, $p(x)$ el polinomio de grado n que satisface $p(x_i) = y_i$. Entonces

$$p(x) = \sum_{i=0}^n y_i l_i(x), \quad (\text{III.1.21})$$

$$\text{donde} \quad l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{(x - x_j)}{(x_i - x_j)}. \quad (\text{III.1.22})$$

Como se dijo al inicio de esta subsección, se considerará los errores de redondeo cometidos en los y_i , es decir

$$\bar{y}_i = y_i(1 + \epsilon_i), \quad \text{donde } |\epsilon_i| \leq \epsilon ps.$$

Definiendo el polinomio $\bar{p}(x)$ por $\bar{p}(x_i) = \bar{y}_i$, $i = 0, \dots, n$. Realizando cálculos y mayoraciones se obtiene:

$$\begin{aligned} p(x) - \bar{p}(x) &= \sum_{i=0}^n \underbrace{(y_i - \bar{y}_i)}_{-\epsilon_i y_i} l_i(x), \\ |p(x) - \bar{p}(x)| &\leq \sum_{i=0}^n |\epsilon_i| |y_i| |l_i(x)| \\ &\leq \epsilon ps |y|_{\max} \sum_{i=0}^n |l_i(x)|, \\ \frac{|p(x) - \bar{p}(x)|}{|y|_{\max}} &\leq \epsilon ps \sum_{i=0}^n |l_i(x)|. \end{aligned}$$

Definición III.1.19.- La constante de Lebesgue asociada a la subdivisión $x_0 < x_1 < \dots < x_n$, está dada por

$$\Lambda_n = \max_{[x_0, x_n]} \sum_{i=0}^n |l_i(x)|. \quad (\text{III.1.23})$$

Ejemplos

- a) Para la división equidistante $x_j = -1 + \frac{2j}{n}$, $j = 0, \dots, n$. Se puede mostrar que la constante de Lebesgue se comporta asintóticamente, cuando $n \rightarrow \infty$, como

$$\Lambda_n \sim \frac{2^{n+1}}{en \log n}, \quad (\text{III.1.24})$$

En el ejercicio 7 de esta sección, se indica cómo se calcula numéricamente estas constantes, en la tabla III.1.1 están los resultados de estos cálculos.

- b) Para la división con puntos de Chebichef $x_j = -\cos\left(\frac{2j+1}{2n+2}\pi\right)$, se puede mostrar que la constante de Lebesgue se comporta asintóticamente, cuando $n \rightarrow \infty$, como

$$\Lambda_n \sim \frac{2}{\pi} \log n. \quad (\text{III.1.25})$$

Tabla III.1.1. Constantes de Lebesgue

n	División Equidistante	División de Chebichef
10	29.900	2.0687
20	10987.	2.4792
30	$6.6011 \cdot 10^6$	2.7267
40	$4.6925 \cdot 10^9$	2.9044
50	$3.6398 \cdot 10^{12}$	3.0432
60	$2.9788 \cdot 10^{15}$	3.1571
70	$2.5281 \cdot 10^{18}$	3.2537
80	$2.2026 \cdot 10^{21}$	3.3375
90	$1.9575 \cdot 10^{24}$	3.4115
100	$1.7668 \cdot 10^{27}$	3.4779

Por los valores dados en la tabla, no se debe utilizar polinomios de interpolación de grado superior a 20, si la división utilizada es equidistante. Tratar en lo posible de utilizar los puntos de Chebichef para interpolación.

En la figura III.1.4 se tienen las graficas de l_{11} para $n = 18$ para las divisiones equidistantes y de Chebichef. En la figura III.1.5, se ve claramente los errores cometidos por redondeo utilizando puntos equidistantes al interpolar la función $\sin(\pi x)$.

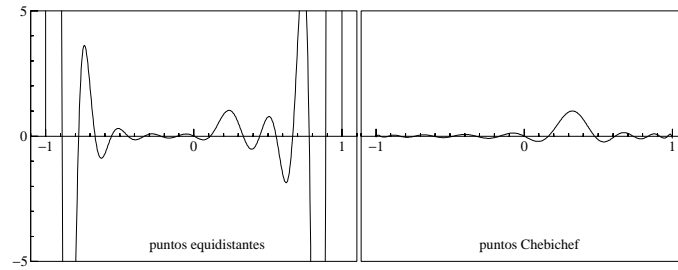


Figura III.1.4. Gráfica de un Polinomio de Lagrange.

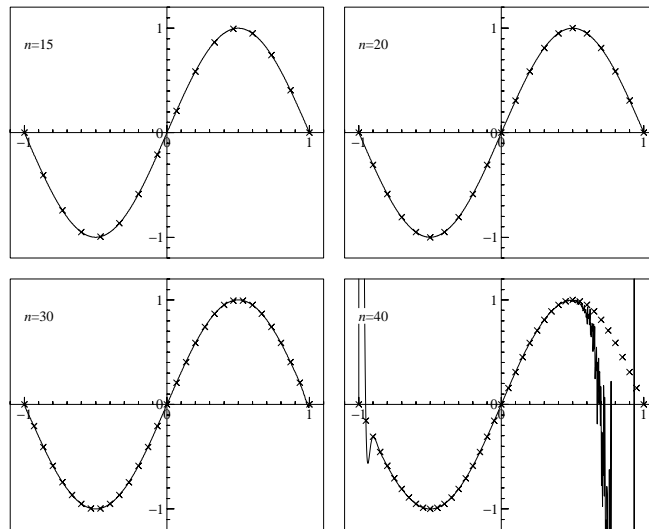


Figura III.1.5. Interpolación de $\sin(\pi x)$.

Convergencia de la Interpolación

En esta subsección, se analizará el problema de la convergencia de la interpolación. Sea $f : [a, b] \rightarrow \mathbb{R}$, para cada entero natural se considera una subdivisión de $[a, b]$ dada por

$$x_0^{(n)} < x_2^{(n)} < \dots < x_n^{(n)},$$

$p_n(x)$ el polinomio de interpolación respecto a esta subdivisión. Se desearía saber que sucede cuando $n \rightarrow \infty$, es decir

$$|f(x) - p(x)| \xrightarrow{n \rightarrow \infty} ?.$$

Teorema III.1.20.- Sea $f \in \mathcal{C}^\infty[a, b]$ tal que $|f^{(n)}(x)| \leq M$ para todo $x \in [a, b]$ y para todo $n \in \mathbb{N}$, sea $\{x_0^{(n)} < x_2^{(n)} < \dots < x_n^{(n)}\}$ una sucesión de subdivisiones de $[a, b]$. Entonces

$$\max_{x \in [a, b]} |f(x) - p_n(x)| \xrightarrow{n \rightarrow \infty} 0, \quad (\text{III.1.25})$$

donde $p_n(x)$ es el polinomio de interpolación de grado n , respecto a las subdivisiones dadas.

Demostración.- Utilizando el teorema III.1.13 y la hipótesis que las derivadas de cualquier orden son acotadas se tiene

$$\begin{aligned} |f(x) - p_n(x)| &= \left| (x - x_0^{(n)}) \dots (x - x_n^{(n)}) \right| \frac{|f^{(n+1)}(\xi)|}{(n+1)!} \\ &\leq M \frac{(b-a)^{n+1}}{(n+1)!} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

□

Las funciones exponencial, senos y cosenos son acotadas y sus derivadas lo son también, motivo por el cual estas funciones pueden ser aproximadas por polinomios de interpolación teniendo asegurada la convergencia de la interpolación. El teorema III.1.20 da condiciones suficientes para asegurar la convergencia, no obstante existen muchas otras funciones cuyas derivadas de orden superior no están acotadas por una sola constante M , por ejemplo funciones racionales. Por lo tanto es necesario poder enunciar otras condiciones para asegurar convergencia, o de lo contrario para decidir que sucede divergencia en el proceso de interpolación.

Para poder comprender más aspectos de la convergencia o divergencia de la interpolación es necesario introducir algunas definiciones y resultados importantes en el estudio de la aproximación de funciones por polinomios.

Considérese la aplicación L_n que a la función f asocia su polinomio de interpolación en los puntos x_0, x_1, \dots, x_n : $L_n(f) = p_n$. Se remarca inmediatamente que esta aplicación es lineal, además si \mathcal{P}_n es el espacio de los polinomios de grado igual o menor a n se tiene

$$\forall q \in \mathcal{P}_n, \quad L_n(q) = q.$$

Se puede mostrar, ver ejercicios, que

$$\Lambda_n = \max_{\substack{g \in \mathcal{C}^0[a,b] \\ g \neq 0}} \frac{\|L_n(g)\|_\infty}{\|g\|_\infty},$$

de manera que la constante Λ_n representa la norma del operador L_n respecto a la norma de la convergencia uniforme en $\mathcal{C}^0[a, b]$. Por otro lado se obtiene el siguiente teorema de mayoración sobre el error de interpolación.

Teorema III.1.21.- *Dados una función $f \in \mathcal{C}^0[a, b]$ y su polinomio de interpolación de Lagrange en los puntos x_0, \dots, x_n , se tiene*

$$\|f - p_n\|_\infty \leq (1 + \Lambda_n)E_n(f), \quad (\text{III.1.26})$$

donde $E_n(f) = \inf_{q \in \mathcal{P}_n} \|f - q\|_\infty$.

Demostración.- Para todo $q \in \mathcal{P}_n$, se tiene

$$f - p_n = (f - q) - (p_n - q) = (f - q) - L_n(f - q),$$

por consiguiente

$$\|f - p_n\|_\infty \leq \|f - q\|_\infty + \|L_n(f - q)\|_\infty \leq (1 + \Lambda_n) \|f - q\|_\infty,$$

el resultado se obtiene tomando la cota inferior sobre los $q \in \mathcal{P}_n$. □

Definición III.1.22.- La cantidad $E_n(f)$ se llama el grado de aproximación de la función f por los polinomios de grado $\leq n$, respecto a la norma de la convergencia uniforme.

Examinando los diferentes factores de la mayoración dada por el anterior teorema, la cantidad Λ_n depende solamente de la subdivisión tomada del intervalo $[a, b]$, mientras que $E_n(f)$ depende solamente de la función f . En el estudio de la estabilidad se vio Λ_n para dos tipos de subdivisión, los puntos de Chebichef y los equidistantes, se ha podido demostrar que Λ_n es minimal cuando se toman los puntos de Chebichef, pero en todas las sucesiones de subdivisiones dadas al principio de este paragrafo, se tiene

$$\lim_{n \rightarrow \infty} \Lambda_n = \infty.$$

Como Λ_n es la norma del operador L_n , consecuencia del teorema de Banach-Steinhaus, se tiene que, cualquiera sea la sucesión de subdivisiones $x_0^{(n)}, \dots, x_n^{(n)}$, existe una función continua f , tal que $L_n(f)$ no converge uniformemente hacia f cuando $n \rightarrow \infty$.

El siguiente paso es estudiar el comportamiento de $E_n(f)$, cuando $n \rightarrow \infty$; esto está intimamente ligado a la regularidad de la función f y más particularmente a su módulo de continuidad, denotado por $\omega(f, h)$, definido por

$$\omega(f, h) = \max_{\substack{t, t' \in [a, b] \\ |t - t'| \leq h}} |f(t) - f(t')|. \quad (\text{III.1.27})$$

Se verifica fácilmente las propiedades siguientes del módulo de continuidad:

- i) la función $h \rightarrow \omega(f, h)$ definida sobre \mathbb{R}^+ es positiva, creciente y subaditiva, es decir: $\forall h_1, h_2 \geq 0, \omega(f, h_1 + h_2) \leq \omega(f, h_1) + \omega(f, h_2)$,
- ii) si $n \in \mathbb{N}$, $\omega(f, nh) \leq n\omega(f, h)$; si $\lambda \in \mathbb{R}^+$, $\omega(f, \lambda h) \leq (1 + \lambda)\omega(f, h)$,
- iii) si $f \in \mathcal{C}^0[a, b]$, $\lim_{h \rightarrow 0} \omega(f, h) = 0$ y la función $h \rightarrow \omega(f, h)$ es continua sobre \mathbb{R}^+ ,
- iv) si $f \in \mathcal{C}^1[a, b]$, $\omega(f, h) \leq h \|f'\|_\infty$.

Teorema III.1.23.- Existe un real M , independientemente de n , a y b , tal que para todo $n \geq 1$ y toda $f \in \mathcal{C}^0[a, b]$, se tiene

$$E_n(f) \leq M\omega\left(f, \frac{b-a}{n}\right). \quad (\text{III.1.28})$$

Demostración. El cambio de variable $s = (x - a)/(b - a)$ permite de trasladarse al caso en que $a = 0$, $b = 1$, $f \in \mathcal{C}^0[0, 1]$, lo que se supondrá a continuación.

Considérese primero el caso en que $n = 2p$ es par, y planteando

$$j_n(t) = \frac{1}{\alpha_n} \left(\frac{\sin\left(\frac{(p+1)t}{2}\right)}{\sin \frac{t}{2}} \right)^4 = \frac{1}{\alpha_n} \left(\sum_{k=0}^n \cos \frac{(p-2k)t}{2} \right)^4,$$

donde α_n es escogido de manera que $\int_{-\pi}^{\pi} j_n(t) dt = 1$, es evidente que j_n puede escribirse bajo la forma

$$j_n(t) = a_{0n} + a_{1n} \cos t + \dots + a_{nn} \cos nt. \quad (\text{III.1.29})$$

Planteando $g(t) = f(|\cos t|)$, está claro que g es una función par, continua sobre \mathbb{R} y periódica de periodo π , además $\omega(g, h) \leq \omega(f, h)$ porque $|t - t'| \leq$

$h \Rightarrow ||\cos t| - |\cos t|| \leq h$. Continuando con la demostración se introduce la función

$$\varphi_n(s) = \int_{-\pi}^{\pi} g(t) j_n(s-t) dt = \int_{-\pi}^{\pi} g(s-t) j_n(t) dt,$$

se tiene

$$g(s) - \varphi_n(s) = \int_{-\pi}^{\pi} (g(s) - g(s-t)) j_n(t) dt,$$

de donde

$$\begin{aligned} |g(s) - \varphi_n(s)| &\leq \int_{-\pi}^{\pi} |g(s) - g(s-t)| j_n(t) dt \\ &\leq \int_{-\pi}^{\pi} \omega(g, |t|) j_n(t) dt = 2 \int_0^{\pi} \omega(g, t) g_n(t) dt \\ &\leq 2 \int_0^{\pi} (1+nt) j_n(t) dt \cdot \omega(g, 1/n), \quad \text{por la propiedad ii.} \end{aligned}$$

Utilizando las mayoraciones $\frac{2}{\pi} \leq \frac{\sin x}{x} \leq 1$ para $x \in [0, \pi/2]$, se obtiene

$$\int_0^{\pi} t j_n(t) dt \leq \frac{C}{n},$$

ver ejercicio 12, por lo tanto se tiene

$$|g(s) - \varphi_n(s)| \leq 2(1+C)\omega(g, \frac{1}{n}).$$

Se remarca que

$$\begin{aligned} \int_{-\pi}^{\pi} g(t) \cos(k(s-t)) dt &= \cos ks \int_{-\pi}^{\pi} g(t) \cos(kt) dt + \sin ks \int_{-\pi}^{\pi} g(t) \sin(kt) dt \\ &= \cos ks \int_{-\pi}^{\pi} g(t) \cos(kt) dt, \end{aligned}$$

puesto que g es una función par.

Resulta de (III.1.29) y de la definición de φ_n que existe $p_n \in \mathcal{P}_n$ tal que para todo $s \in \mathbb{R}$, $\varphi_n(s) = p_n(\cos s)$, por consiguiente se tiene

$$\begin{aligned} \max_{x \in [0,1]} |f(x) - p_n(x)| &\leq \max_{s \in \mathbb{R}} |f(|\cos s|) - p_n(\cos s)| = \max_{s \in \mathbb{R}} |g(s) - \varphi_n(s)| \\ &\leq 2(1+C)\omega(g, 1/n) \leq 2(1+C)\omega(f, 1/n). \end{aligned}$$

El caso en que $n = 2p+1$ es impar se deduce del precedente remarcando que $E_{2p+1}(f) \leq E_{2p}(f)$ y que $\omega(f, (b-a)/n) \leq 2\omega(f, (b-a)/(n+1))$. \square

Corolario III.1.24.- *Teorema de Weirstrass.* Sea $[a, b]$ un intervalo cerrado y acotado de \mathbb{R} , entonces el conjunto de las funciones polinomiales es denso en $\mathcal{C}^0[a, b]$ para la topología de la convergencia uniforme

Demostración.- En efecto, por la propiedad iii) del módulo de continuidad, se tiene que

$$\lim_{n \rightarrow \infty} \omega \left(f, \frac{b-a}{n} \right) = 0,$$

por consiguiente $\lim_{n \rightarrow \infty} E_n(f) = 0$. para toda función $f \in \mathcal{C}^0[a, b]$. \square

Corolario III.1.25.- Se supone que $f \in \mathcal{C}^p[a, b]$, $p \geq 0$, entonces para todo $n > p$ se tiene

$$E_n(f) \leq M^{p+1} \frac{(b-a)^p}{n(n-1) \cdots (n-p+1)} \omega \left(f^{(p)}, \frac{b-a}{n-p} \right). \quad (\text{III.1.30})$$

Demostración.- Por el teorema III.1.23, el corolario es cierto para $p = 0$, supóngase cierto el enunciado del corolario para p y sea $f \in \mathcal{C}^{p+1}[a, b]$, se tiene por lo tanto $f' \in \mathcal{C}^p[a, b]$, de donde $\forall n > p+1$

$$E_{n-1}(f') \leq M^{p+1} \frac{(b-a)}{(n-1) \cdots (n-p)} \omega \left(f^{(p+1)}, \frac{b-a}{n-1-p} \right).$$

Se puede mostrar, ver Crouzeix, Capítulo I.3, que existe $q \in \mathcal{P}_{n-1}$ tal que $\|f' - q\|_\infty = E_{n-1}(f')$. Planteando por lo tanto $p(x) = \int_a^x q(t)dt$ y $\varphi(x) = f(x) - p(x)$, se tiene $p \in \mathcal{P}_n$, entonces $E_n(f) = E_n(\varphi)$, además $\|\varphi'\|_\infty = \|f' - q\|_\infty = E_{n-1}(f')$. Aplicando el teorema III.1.23 a la función φ y la propiedad iv) del módulo de continuidad, se obtiene

$$E_n(f) = E_n(\varphi) \leq M \omega \left(\varphi, \frac{b-a}{n} \right) \leq M \frac{b-a}{n} \|\varphi'\|_\infty,$$

de donde

$$E_n(f) \leq M^{p+2} \frac{(b-a)^{p+1}}{n(n-1) \cdots (n-p)} \omega \left(f^{(p+1)}, \frac{b-a}{n-p-1} \right).$$

\square

Utilizando las estimaciones dadas para las constantes de Lebesgue y el teorema III.1.21, se obtiene las mayoraciones siguientes del error de interpolación de Lagrange si $f \in \mathcal{C}^p[a, b]$:

a) En el caso que los x_i son equidistantes,

$$\|f - p_n\|_\infty \leq C_p \frac{(b-a)^p 2^n}{n^{p+1} \log n} \omega \left(f^{(p)}, \frac{b-a}{n-p} \right),$$

b) En el caso de los puntos de interpolación de Chebichef

$$\|f - p_n\|_{\infty} \leq C_p(b-a)^p \frac{\log n}{n^p} \omega\left(f^{(p)}, \frac{b-a}{n-p}\right),$$

este resultado con $p = 0$, da el resultado de Bernstein que dice que el interpolante de Lagrange con los puntos de Chebichef converge hacia f del momento en que $\lim_{h \rightarrow 0} \omega(f, h) \log h = 0$, que es el caso si $f \in \mathcal{C}^1[a, b]$.

Teorema III.1.26.- Si la función $f \in \mathcal{C}^1[a, b]$, si los puntos utilizados durante el proceso de interpolación son puntos de Chebichef, entonces

$$\lim_{n \rightarrow \infty} p_n = f.$$

La utilización de los puntos de Chebichef en la aproximación de una función continuamente diferenciable en un intervalo $[a, b]$, no implica necesariamente convergencia numérica, pues el cálculo del polinomio interpolante por medio de computadoras está imbuido de errores de redondeo. Por lo tanto hay que tener mucho cuidado cuando se utilizan muchos puntos de interpolación.

Fenómeno de Runge

Hasta ahora, se ha visto condiciones suficientes para asegurar la convergencia de la interpolación de una función por polinomios. También se ha mostrado que la utilización de puntos de Chebichef en la interpolación de una función continuamente derivable permitía la convergencia. En esta parte, se verá que tener una función continuamente derivable y una división equidistante no es una condición suficiente de convergencia. Con el siguiente ejemplo se ilustrará el conocido fenómeno de Runge. En la figura III.1.6, puede apreciarse que la interpolación de Lagrange diverge cuando la división es equidistante. Mientras, que en la figura III.1.7. utilizando subdivisiones de Chebichef se tiene convergencia. En líneas punteadas se tiene la gráfica de la función a interpolar.

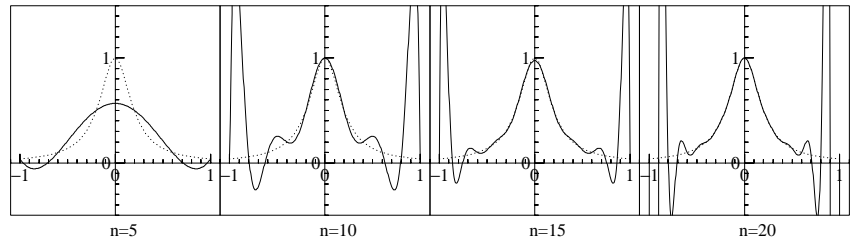


Figura III.1.6. Interpolación con puntos equidistantes.

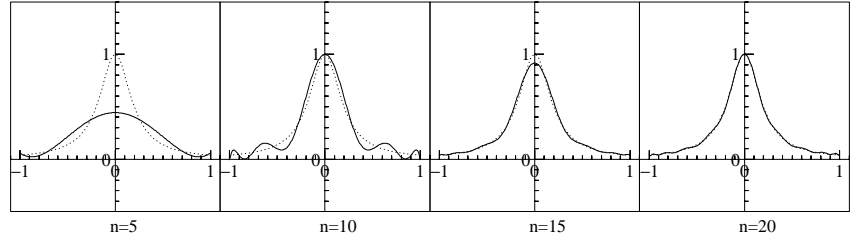
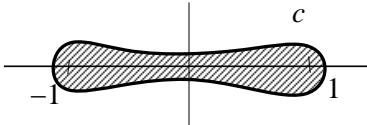


Figura III.1.7. Interpolación con puntos de Chebichef.

Sea $f : [-1, 1] \rightarrow \mathbb{R}$, definida por

$$f(x) = \frac{1}{1 + 25x^2},$$

función que es indefinidamente derivable. Sean $x_0^{(n)}, \dots, x_n^{(n)}$ la división equidistante de $[-1, 1]$, $p_n(x)$ el polinomio de interpolación de grado n de la función f respecto a la subdivisión dada anteriormente. Prolongando $f(x)$ al plano complejo, se tiene que $f(z)$ tiene polos simples en $z = \frac{1}{5}i$ y en $z = -\frac{1}{5}i$. Se considera un camino cerrado simple \mathcal{C} tal que $[-1, 1] \subset \text{interior de } \mathcal{C}$, y los polos estén en el exterior de \mathcal{C} . Por consiguiente se puede aplicar la fórmula de Cauchy para integrales complejas, de donde

$$f(x) = \frac{1}{2i\pi} \int_{\mathcal{C}} \frac{f(z)}{z - x} dz.$$


Se define el polinomio $\Pi_n(x)$ de grado n por

$$\Pi_n(x) = (x - x_0^{(n)})(x - x_1^{(n)}) \cdots (x - x_n^{(n)}),$$

Por otro lado se comprueba fácilmente que la expresión $(\Pi_n(z) - \Pi_n(x))/(z - x)$ es un polinomio de grado n respecto a x , definiendo el polinomio de grado n , $q(x)$ por

$$q(x) = \frac{1}{2i\pi} \int_{\mathcal{C}} \frac{f(z)}{z - x} \frac{(\Pi_n(z) - \Pi_n(x))}{\Pi_n(z)} dz, \quad (\text{III.1.31})$$

se verifica que $q(x_i^{(n)}) = f(x_i^{(n)})$, de donde se ha mostrado el teorema siguiente.

Teorema III.1.27.- Si f es una función racional con polos fuera de \mathcal{C} . Entonces el error de interpolación está dado por

$$f(x) - p_n(x) = \frac{1}{2i\pi} \int_{\mathcal{C}} \frac{f(z)}{z-x} \frac{\Pi_n(x)}{\Pi_n(z)} dz, \quad (\text{III.1.32})$$

donde p_n es el polinomio de interpolación.

Introduciendo valores absolutos, se obtiene

$$|f(x) - p_n(x)| \leq \frac{1}{2\pi} \int_{\mathcal{C}} \frac{|f(z)|}{|z-x|} \left| \frac{\Pi_n(x)}{\Pi_n(z)} \right| |dz|,$$

por consiguiente se debe analizar el comportamiento cuando $n \rightarrow \infty$ de la expresiones que aparecen en la integral precedente. Se tiene convergencia si

$$\left| \frac{\Pi_n(x)}{\Pi_n(z)} \right| < \kappa^n,$$

con $\kappa < 1$, de lo contrario cuando $n \rightarrow \infty$ $|f(x) - p_n(x)|$ diverge. Por consiguiente hay convergencia en $x \in [-1, 1]$ si y solamente si

$$\lim_{n \rightarrow \infty} \sqrt[n]{\left| \frac{\Pi_n(x)}{\Pi_n(z)} \right|} < 1. \quad (\text{III.1.33})$$

En lugar de la expresión (III.1.33), se puede utilizar

$$\begin{aligned} \ln \sqrt[n]{|\Pi_n(z)|} &= \frac{1}{n} \ln (|\Pi_n(z)|) \\ &= \frac{2}{2n} \sum_{j=0}^n \ln |z - x_j|, \end{aligned}$$

como los x_j son equidistantes, cuando $n \rightarrow \infty$ se tiene

$$\begin{aligned} \lim_{n \rightarrow \infty} \ln \sqrt[n]{|\Pi_n(z)|} &= \frac{1}{2} \int_{-1}^1 \ln |z - t| dt \\ &= \frac{1}{2} \Re \int_{-1}^1 \log(z - t) dt \\ &= \frac{1}{2} \Re \{ (z+1) \log(z+1) + (1-z) \log(z-1) \} - 1. \end{aligned}$$

Definiendo $G(z)$ por

$$G(z) = \exp \left(\frac{1}{2} \Re \{ (z+1) \log(z+1) + (1-z) \log(z-1) \} - 1 \right), \quad (\text{III.1.34})$$

se obtiene que

$$\lim_{n \rightarrow \infty} \sqrt[n]{|\Pi_n(z)|} = G(z).$$

Si $x \in \mathbb{R}$, la función G es igual a

$$\begin{aligned} G(x) &= \exp \left(\frac{1}{2} \ln(x+1)^{(x+1)} + \frac{1}{2} \ln(1-x)^{(1-x)} - 1 \right) \\ &= \sqrt{(1+x)^{1+x} (1-x)^{1-x}} / e. \end{aligned}$$

Para decidir si la interpolación converge para un x dado se debe tener necesariamente $G(z)/G(x) < 1$, esto sucede solamente si $|x| < 0,72668$. La figura III.1.8 muestra una grafica de $G(x)$ y en la figura III.1.7 están dadas las curvas de nivel para $G(z)$.

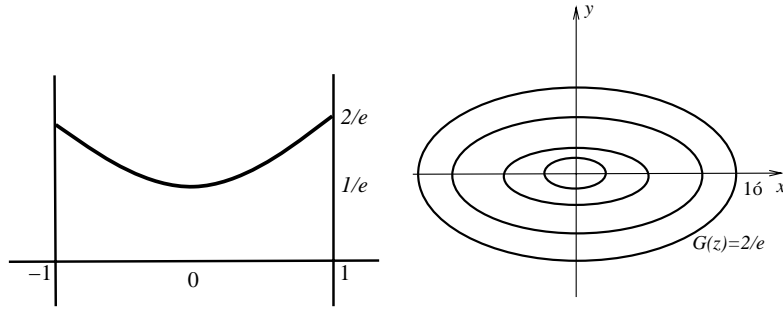


Fig. III.1.8. Gráfica de $G(x)$. **Fig. III.1.9.** Curvas de nivel de $G(z)$.

Ejercicios

1.- Demostrar por inducción

$$\begin{aligned} y[x_0, \dots, x_n] &= \sum_{j=0}^n y_j \prod_{i \neq j} \frac{1}{x_j - x_i}, \\ \Delta^n y_0 &= \sum_{j=0}^n \binom{n}{j} y_j (-1)^{n-j}. \end{aligned}$$

2.- Supóngase conocidos los valores de la función de Bessel

$$J_0(x) = \frac{1}{\pi} \int_0^\pi \cos(x \sin t) dt.$$

para los puntos equidistantes $x_i = x + 0 + ih$, $i \in \mathbb{Z}$.

a) ¿Para que h , el error de la interpolación lineal es menor a 10^{-11} ?

b) La misma pregunta para la interpolación cuadrática.

3.- Calcular el polinomio de interpolación de grado $n = 10$ y $n = 20$, que pasa por $(x_i, f(x_i))$, $i = 0, 1, \dots, n$, donde $f(x) = 1/(1 + 25x^2)$.

a) $x_i = -1 + \frac{2i}{n}$.

b) $x_i = \cos(\frac{2i+1}{2n+2}\pi)$.

Hacer gráficas, estudiar los errores.

4.- Sea $p(x)$ el polinomio de interpolación de grado 1 de una función f dos veces continuamente derivable. Mostrar que el error satisface

$$f(x) - p(x) = \int_{x_0}^{x_1} G(x, t) f''(t) dt \quad (*)$$

con

$$G(x, t) = \begin{cases} \frac{1}{h}(x - x_0)(t - x_1) & \text{si } x \leq t \\ \frac{1}{h}(t - x_0)(x - x_1) & \text{si } x \geq t \end{cases}.$$

Dibujar la función $G(x, \cdot)$. Deducir de (*) el resultado del teorema III.1.13.

5.- Sean $x_0 < x_1 < \dots < x_n$ y $l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{(x - x_j)}{(x_i - x_j)}$.

Verificar que la función

$$\lambda_n(x) = \sum_{i=0}^n |l_i(x)|$$

tiene un solo máximo local sobre cada $[x_{i-1}, x_i]$

Indicación: Sobre $[x_{j-1}, x_j]$ se tiene a $\lambda_n(x) = \sum_{i=0}^n \epsilon_i l_i(x)$ con $\epsilon = \pm 1$.

6.- Si la función $f : [x_0, x_1] \rightarrow \mathbb{R}$ tiene solamente un máximo local y si $x_0 < a < b < x_1$, entonces

$$f(a) \leq f(b) \implies x^* \in [a, x_1],$$

$$f(a) \geq f(b) \implies x^* \in [x_0, b],$$

donde $f(x^*) = \max_{x \in [x_0, x_1]} f(x)$.

7.- Calcular las constantes de Lebesgue

$$\Lambda_n = \max_{x \in [-1, 1]} \sum_{i=0}^n |l_i(x)|, \quad n = 10, 20, \dots, 100,$$

- a) para la división equidistante $x_i = -1 + 2i/n$, $i = 0, 1, \dots, n$;
 b) para los puntos de Chebichef.

Calcular el maximo de $f(x) = \sum_{i=0}^n |l_i(x)|$ sobre $[x_{j-1}, x_j]$ con la *búsqueda de Fibonacci*:

$$x_1 = x_j, \quad x_0 = x_{j-1}, \quad \gamma = (\sqrt{5} - 1)/2;$$

$$d = \gamma(x_1 - x_0);$$

$$a = x_1 - d, \quad b = x_0 + d;$$

$$10 \quad d = \gamma d;$$

si $(f(a) \leq f(b))$ entonces

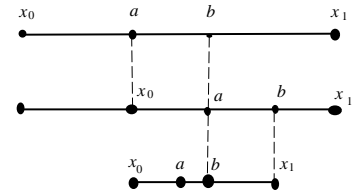
$$x_0 = a, \quad a = b, \quad b = x_0 + d$$

si no

$$x_1 = b, \quad b = a, \quad a = x_1 - d;$$

si $(x_1 - x_0 \geq 10^{-14})$ vaya a 10

si no, final.



8.- Sean dados (x_i, y_i, y'_i) , $i = 0, 1, \dots, n$, los x_i distintos entre si.

Mostrar que existe un único polinomio $q(x)$ de grado $2n + 1$ tal que

$$q(x_i) = y_i, \quad q'(x_i) = y'_i \quad (i = 0, \dots, n).$$

Utilizar la fórmula de Newton con (x_0, y_0) , $(x_0 + \epsilon, y_0 + \epsilon y'_0)$, \dots y estudiar su límite cuando $\epsilon \rightarrow 0$.

¿Qué fórmula se obtiene para $n = 2$?, utilizar $y[x_i, x_i] = y'_i$, Generalizar el ejercicio para cualquier polinomio de Hermite y obtener un esquema de cálculo semejante a la figura III.1.1.

9.- Se considera la función $f(s) = |s(s-1) \cdots (s-n)|$ definida para $s \in [0, n]$.

a) Mostrar que f alcanza su maximo en un punto $s_n \in (0, 1/2)$ y que $s_n \sim \frac{1}{\ln n}$ cuando $n \rightarrow \infty$.

b) Mostrar que $\lim_{n \rightarrow \infty} [f(1/\ln n)(\ln n/n!)] = 1/e$, deducir que existe $c_1 > 0$ tal que $\forall n > 1$, $f(s_n) \geq \frac{c_1 n!}{\ln n}$.

Sean $\{x_0, x_1, \dots, x_n\}$ $n+1$ puntos equidistantes sobre el intervalo $[a, b]$ con $x_0 = a$ y $x_n = b$. Mostrar que existe una constante C_2 tal que

$$\forall n > 1. \quad \max_{x \in [a, b]} \left| \prod_{i=0}^n (x - x_i) \right| \leq \frac{C_2 e^{-n}}{\sqrt{n} \ln n} (b - a)^{(n+1)}.$$

10.- Conservando las notaciones, construir una función $f \in \mathcal{C}^0[a, b]$ tal que $\|L_n(f)\|_\infty = \Lambda \|f\|_\infty$ y $\|f - L_n(f)\|_\infty = (1 + \Lambda_n) \|f\|_\infty$.

11.- Se considera el conjunto $\{x_0, \dots, x_n\}$ de $n+1$ puntos diferentes del intervalo $[a, b]$ y Λ_n la constante de Lebesgue correspondiente a la interpolación de Lagrange en estos puntos. Se plantea

$$\hat{l}_i(\theta) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{\cos \theta - \cos \theta_j}{\cos \theta_i - \cos \theta_j} \quad \text{con } \theta_i = \arccos((2x_i - (a+b))/(b-a)),$$

$i = 0, \dots, n$.

a) Mostrar que:

$$\Lambda_n = \max_{\theta \in \mathbb{R}} \left(\sum_{i=0}^n |\hat{l}_i(\theta)| \right);$$

$$\cos k(\theta - \varphi) + \cos k(\theta + \varphi) = \sum_{i=0}^n [\cos k(\theta_i - \varphi) + \cos k(\theta_i + \varphi)] \hat{l}_i(\theta),$$

$$0 \leq k \leq n.$$

b) Planteando $\varphi_n(\theta) = \sum_{i=0}^n [\hat{l}_i(\theta + \theta_i) + \hat{l}_i(\theta - \theta_i) - \hat{l}_i(\theta)]$, mostrar que:

$$\varphi_n(\theta) = 1 + 2 \sum_{k=1}^n \cos k\theta = \frac{\sin((n+1/2)\theta)}{\sin(\theta/2)};$$

$$\text{si } F(\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \text{signo}(\varphi_n(\theta - s)) \varphi_n(s) ds, \text{ entonces}$$

$$3\Lambda_n \geq \|F\|_\infty = \frac{1}{\pi} \int_{-\pi}^{\pi} |\varphi_n(s)| d\theta \leq \frac{2}{\pi} \int_0^{(n+1/2)\pi} \frac{|\theta|}{\theta} d\theta = \gamma_n,$$

$$\gamma_n - \gamma_{n-1} = \frac{4}{n\pi^2} + \mathcal{O}\left(\frac{1}{n^3}\right) \text{ cuando } n \rightarrow \infty,$$

$$\gamma_n \sim \frac{4}{\pi^2} \ln n \text{ cuando } n \rightarrow \infty.$$

12.- Sea j_n la función definida en la demostración del teorema III.1.23. con $n = 2p$. Mostrar que $\forall p \geq 1$,

$$\int_0^\pi t^k j_n(t) dt \leq c_k / n^k,$$

para $k = 1$ y 2 .

III.2 Splines Cúbicos

En la sección precedente, se ha visto cómo determinar el polinomio de interpolación. Existen dos problemas de gran magnitud cuando se requiere encontrar un polinomio de interpolación dado un conjunto (x_i, y_i) , $i = 0, \dots, n$; el primero consiste en la inestabilidad de la interpolación de Lagrange cuando n es grande, por ejemplo para $n \geq 20$ no es aconsejable tomar puntos equidistantes. El segundo problema es que aun obteniendo el polinomio de interpolación, éste no refleja la verdadera forma de la función a interpolar, por ejemplo, para una función racional, se desearía que el interpolante tenga la menor curvatura promedio, tal como se grafica con una sercha. Por consiguiente, dados los puntos $a = x_0 < x_1 < \dots < x_n = b$ y los reales y_i con $i = 0, \dots, n$ se busca una función *spline* s , tal que:

- (i) $s(x_i) = y_i \quad i = 0, \dots, n$
- (ii) $s \in \mathcal{C}^2([a, b])$,
- (iii) s de curvatura pequeña.

La curvatura de s está dada por

$$\kappa(x) = \frac{(s''(x))}{(1 + (s'(x))^2)^{\frac{3}{2}}},$$

suponiendo que $s'(x)$ es despreciable, se obtiene $s''(x) = \kappa(x)$, por consiguiente la condición, se expresa de la siguiente forma

$$\int_{x_0}^{x_n} (s''(x))^2 dx \longrightarrow \min. \quad (\text{III.2.1})$$

En la figura III.2.1 se observa dos gráficas: la de la izquierda es de un polinomio de interpolación de tipo Lagrange; mientras que en la derecha es la gráfica del *spline* interpolante que pasa por los mismos puntos.

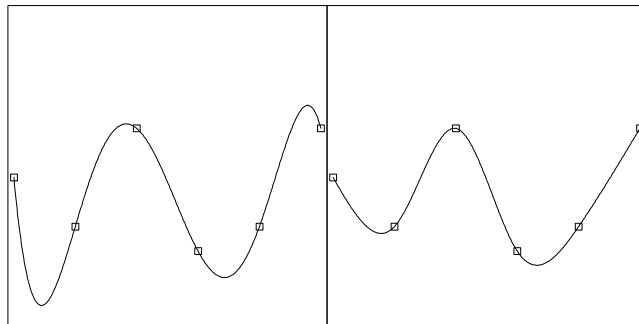


Figura III.2.1. *Spline* de Interpolación

Definición III.2.1.- Un *spline* cúbico es una función $s : [a, b] \rightarrow \mathbb{R}$ con $a = x_0 < x_1 < \dots < x_n = b$, que satisface:

$$s \in \mathcal{C}^2[a, b], \quad (\text{III.2.2a})$$

$$s|_{[x_{j-1}, x_j]} \text{ polinomio de grado 3.} \quad (\text{III.2.2b})$$

Teorema III.2.2.- Dados (x_i, y_i) , $i = 0, \dots, n$, con $a = x_0 < x_1 < \dots < x_n = b$. Sean:

$$s : [a, b] \rightarrow \mathbb{R} \text{ un spline con } s(x_i) = y_i,$$

$$f : [a, b] \rightarrow \mathbb{R} \text{ una función con } f(x_i) = y_i,$$

si

$$s''(b)[f'(b) - s'(b)] = s''(a)[f'(a) - s'(a)], \quad (\text{III.2.3})$$

entonces

$$\int_a^b [s''(x)]^2 dx \leq \int_a^b [f''(x)]^2 dx.$$

La condición (III.2.3) es satisfecha si por ejemplo $s''(a) = s''(b) = 0$, o por ejemplo si $f'(a) = s'(a)$ y $f'(b) = s'(b)$.

Definición III.2.3.- Si $s''(a) = s''(b) = 0$, el *spline* se llama *spline* natural. Si $f'(a) = s'(a)$ y $f'(b) = s'(b)$ el *spline* se dice fijo en los bordes.

Demostración.- Calculando las integrales, se obtiene

$$\begin{aligned} \int_a^b [f''(x)]^2 dx - \int_a^b [s''(x)]^2 dx &= \underbrace{\int_a^b [f''(x) - s''(x)]^2 dx}_{\geq 0} \\ &\quad + 2 \int_a^b s''(x)[f''(x) - s''(x)] dx. \end{aligned}$$

Para demostrar la afirmación del teorema, es suficiente mostrar que la segunda integral del miembro derecho de la ecuación es nula. En efecto, integrando por partes, se tiene

$$\begin{aligned} \int_a^b s''(x)[f''(x) - s''(x)] dx &= \underbrace{s''(x)[f'(x) - s'(x)]}_a^b \\ &= 0 \text{ por (III.2.2)} \\ &\quad - \int_a^b s'''(x)[f'(x) - s'(x)] dx \\ &= - \sum_{j=1}^n \kappa_j \int_{x_{j-1}}^{x_j} [f'(x) - s'(x)] dx \\ &= - \sum_{j=1}^n \kappa_j [f(x) - s(x)]_{x_{j-1}}^{x_j} = 0. \end{aligned}$$

□

Construcción del Spline Interpolante

En este paragrafo, se determinará explícitamente el *spline* interpolante. Para tal efecto sea $a = x_0 < x_1 < \dots < x_n = b$ una división cualquiera del intervalo $[a, b]$; y sean y_0, y_1, \dots, y_n números reales. Se desea construir $s : [a, b] \rightarrow \mathbb{R}$, el *spline* tal que $s(x_i) = y_i$. Llamando $s_i = s|_{[x_{i-1}, x_i]} : [x_{i-1}, x_i] \rightarrow \mathbb{R}$. Por consiguiente se busca una representación de s_i , utilizando los siguientes datos:

$$y_{i-1}, \quad y_i, \quad p_{i-1} = s'_i(x_{i-1}), \quad p_i = s'_i(x_i). \quad (\text{III.2.4})$$

Por lo tanto, $s_i(x)$ es igual a

$$s_i(x) = y_{i-1} + y[x_{i-1}, x_i](x - x_{i-1}) + (x - x_{i-1})(x - x_i)[a(x - x_{i-1}) + b(x - x_i)], \quad (\text{III.2.5})$$

faltando determinar los valores de a y b . Planteando $h_i = x_i - x_{i-1}$ y derivando (III.2.5) en los puntos x_{i-1} y x_i , se obtiene:

$$p_{i-1} = s'(x_{i-1}) = y[x_{i-1}, x_i] + h_i^2 b, \quad (\text{III.2.6})$$

$$p_i = s'(x_i) = y[x_{i-1}, x_i] + h_i^2 a, \quad (\text{III.2.7})$$

de donde

$$s_i(x) = y_{i-1} + y[x_{i-1}, x_i](x - x_{i-1}) + \frac{(x - x_{i-1})(x - x_i)}{h_i^2} [(p_i - y[x_{i-1}, x_i])(x - x_{i-1}) + (p_{i-1} - y[x_{i-1}, x_i])(x - x_i)]. \quad (\text{III.2.8})$$

Ahora bien, se debe determinar los p_i , $i = 0, \dots, n$ de manera que $s \in \mathcal{C}^2[a, b]$. Por consiguiente:

$$s''_i(x_i) = s''_{i+1}(x_i) \quad i = 1, \dots, n-1,$$

Utilizando la regla de Leibnitz

$$(fg)''(x) = f''(x)g(x) + 2f'(x)g'(x) + f(x)g''(x),$$

se obtiene:

$$\begin{aligned} \frac{d^2}{dx^2} ((x - x_{i-1})^2(x - x_i)) \Big|_{x=x_i} &= 4h_i, \\ \frac{d^2}{dx^2} ((x - x_{i-1})(x - x_i)^2) \Big|_{x=x_i} &= 2h_i, \end{aligned}$$

de donde:

$$s_i''(x_i) = \frac{1}{h_i} [4(p_i - y[x_{i-1}, x_i]) + 2(p_{i-1} - y[x_{i-1}, x_i])],$$

$$s_{i+1}''(x_i) = \frac{1}{h_{i+1}} [2(p_{i+1} - y[x_i, x_{i+1}]) + 4(p_i - y[x_i, x_{i+1}])]$$

por lo tanto,

$$\frac{p_{i-1}}{h_i} + 2p_i \left(\frac{1}{h_i} + \frac{1}{h_{i+1}} \right) + \frac{p_{i+1}}{h_{i+1}} = 3 \left[\frac{y[x_{i-1}, x_i]}{h_i} + \frac{y[x_i, x_{i+1}]}{h_{i+1}} \right], \quad (\text{III.2.9})$$

para $i = 1, \dots, n-1$. Es así que se ha obtenido $n-1$ ecuaciones, teniendo $n+1$ ecuaciones, las dos ecuaciones que faltan para obtener un sistema completo, se obtienen de las condiciones de borde para x_0 y x_n .

Si el *spline* es natural, se tiene $s''(a) = s''(b) = 0$, obteniendo:

$$2\frac{p_0}{h_1} + \frac{p_1}{h_1} = 3\frac{y[x_0, x_1]}{h_0} \quad (\text{III.2.10a})$$

$$2\frac{p_{n-1}}{h_n} + \frac{p_n}{h_n} = 3\frac{y[x_{n-1}, x_n]}{h_n} \quad (\text{III.2.10b})$$

Si el *spline* está fijado en los bordes, se tiene $s'(a) = f'(a)$ y $s'(b) = f'(b)$, obteniendo:

$$p_0 = f'(a), \quad (\text{III.2.11a})$$

$$p_n = f'(b). \quad (\text{III.2.11b})$$

Existen muchos otros tipos de *spline*, como por ejemplo periódicos, de Bezier, etc. Las condiciones de borde son diferentes, pero las ecuaciones dadas por (III.2.9) son esencialmente las mismas, estos aspectos serán desarrollados con más detalle posteriormente o en la parte de los ejercicios.

Por otro lado, las ecuaciones que determinan los p_i pueden escribirse de manera matricial. A continuación, se darán las notaciones matriciales para los *splines* natural y fijo.

$$\begin{pmatrix} \frac{2}{h_1} & \frac{1}{h_1} & 0 & & \\ \frac{1}{h_1} & 2(\frac{1}{h_1} + \frac{1}{h_2}) & \frac{1}{h_2} & & \\ 0 & \frac{1}{h_2} & 2(\frac{1}{h_2} + \frac{1}{h_3}) & \frac{1}{h_3} & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{h_{n-1}} & 2(\frac{1}{h_{n-1}} + \frac{1}{h_n}) & \frac{1}{h_n} \\ & & & \frac{1}{h_n} & \frac{2}{h_n} \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_{n-1} \\ p_n \end{pmatrix} = \begin{pmatrix} \frac{3}{h_1} y[x_0, x_1] \\ \frac{3}{h_1} y[x_0, x_1] + \frac{3}{h_2} y[x_1, x_2] \\ \frac{3}{h_2} y[x_1, x_2] + \frac{3}{h_3} y[x_2, x_3] \\ \vdots \\ \frac{3}{h_{n-1}} y[x_{n-2}, x_{n-1}] + \frac{3}{h_n} y[x_{n-1}, x_n] \\ \frac{3}{h_n} y[x_{n-1}, x_n] \end{pmatrix}$$

Matriz del *spline* natural.

$$\begin{pmatrix} 2(\frac{1}{h_1} + \frac{1}{h_2}) & \frac{1}{h_2} & & \\ \frac{1}{h_2} & 2(\frac{1}{h_2} + \frac{1}{h_3}) & \frac{1}{h_3} & \\ & \ddots & \ddots & \ddots \\ & & \frac{1}{h_{n-1}} & 2(\frac{1}{h_{n-1}} + \frac{1}{h_n}) \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{n-1} \end{pmatrix} = \begin{pmatrix} \frac{3}{h_1} y[x_0, x_1] + \frac{3}{h_2} y[x_1, x_2] - \frac{p_0}{h_1} \\ \frac{3}{h_2} y[x_1, x_2] + \frac{3}{h_3} y[x_2, x_3] \\ \vdots \\ \frac{3}{h_{n-1}} y[x_{n-2}, x_{n-1}] + \frac{3}{h_n} y[x_{n-1}, x_n] - \frac{p_n}{h_n} \end{pmatrix}$$

Matriz del *spline* fijo en los bordes.

Denotando A , la matriz del *spline* natural como fijo en los bordes, se tiene el siguiente teorema.

Teorema III.2.4.- *La matriz A es inversible, es decir*

$$\det A \neq 0.$$

Demostración.- Se mostrará que la siguiente implicación es cierta

$$Ap = 0 \implies p = 0.$$

Supóngase que $Ap = 0$, entonces

$$\frac{p_{i-1}}{h_i} + 2p_i \left(\frac{1}{h_i} + \frac{1}{h_{i+1}} \right) + \frac{p_{i+1}}{h_{i+1}} = 0,$$

se denota por p_{i_o} a

$$|p_{i_o}| = \max_i |p_i|,$$

por consiguiente

$$2|p_{i_o}| \left(\frac{1}{h_{i_o}} + \frac{1}{h_{i_o+1}} \right) \leq \frac{|p_{i_o}|}{h_{i_o}} + \frac{|p_{i_o}|}{h_{i_o+1}},$$

de donde

$$2|p_{i_o}| \leq |p_{i_o}|.$$

□

En uno de los ejercicios de la sección II.1, se muestra que el problema de encontrar los p_i es un problema bien condicionado. Por otra lado la resolución de este sistema se procede facilmente con el algoritmo de Eliminación de Gauss, si n no es muy grande; en el caso contrario con un método iterativo utilizando el hecho que A es simétrica y definida positiva.

El Error de la Aproximación Spline

Se ha visto en el paragrafo anterior, que la construcción del *spline* interpolante es muy simple, solo se requiere resolver un sistema tridiagonal de ecuaciones lineales. En esta parte, se insistirá en las estimaciones del error cometido en el proceso de interpolación *spline*.

Sean $f : [a, b] \longrightarrow \mathbb{R}$ una función, $a = x_0 < x_1 < \dots < x_n = b$ una subdivisión del intervalo $[a, b]$ y finalmente $s(x)$ el *spline* fijo en los bordes que interpola f , es decir

$$\begin{aligned} s(x_i) &= f(x_i), \quad i = 0, \dots, n; \\ s'(x_0) &= f'(x_0); \\ s'(x_n) &= f'(x_n); \end{aligned}$$

se desearía saber que error se comete, en otras palabras

$$|f(x) - s(x)| \leq ?$$

Teorema III.2.5.- Si $f \in \mathcal{C}^4[a, b]$, $a = x_0 < x_1 < \dots < x_n = b$ una subdivisión, $h_i = x_i - x_{i-1}$, los p_i derivadas en los nudos del *spline* fijo en los bordes. Entonces

$$|f'(x_i) - p_i| \leq \frac{h^3}{24} \|f^{(4)}\|_{\infty}, \quad (\text{III.2.12})$$

donde $h = \max h_i$. Si además, la división es equidistante, se tiene

$$|f'(x_i) - p_i| \leq \frac{h^4}{60} \|f^{(5)}\|_{\infty}. \quad (\text{III.2.13})$$

Además del interés del resultado del teorema que servirá para la estimación del error, se tiene un medio simple para calcular las derivadas de f en los puntos x_i .

Demostración.- La demostración del caso general es muy similar a la del caso de los puntos equidistantes, aquí se mostrará el caso equidistante, dejando al lector el caso general. La construcción del *spline* está dada por las ecuaciones

$$\frac{1}{h} (p_{i-1} + 4p_i + p_{i+1}) - \frac{3}{h^2} (f(x_{i+1}) - f(x_{i-1})) = 0. \quad (\text{III.2.14})$$

Se define q_i , $i = 1, \dots, n-1$, por

$$q_i = \frac{1}{h} (f'(x_{i-1}) + 4f'(x_i) + f'(x_{i+1})) - \frac{3}{h^2} (f(x_{i+1}) - f(x_{i-1})) \quad (\text{III.2.15})$$

Por otra lado, utilizando el teorema de Taylor con resto en el punto x_i se tiene:

$$\begin{aligned} f(x_{i+1}) &= f(x_i) + hf'(x_i) + \frac{h^2}{2!}f''(x_i) + \frac{h^3}{3!}f^{(3)}(x_i) \\ &\quad + \frac{h^4}{4!}f^{(4)}(x_i) + h^5 \int_0^1 \frac{(1-t)^4}{4!}f^{(5)}(x_i + th)dt, \\ f'(x_{i+1}) &= f'(x_i) + hf''(x_i) + \frac{h^2}{2!}f^{(3)}(x_i) + \frac{h^3}{3!}f^{(4)}(x_i) \\ &\quad + h^4 \int_0^1 \frac{(1-t)^3}{3!}f^{(5)}(x_i + th)dt, \end{aligned}$$

de donde, introduciendo en (III.2.15), se obtiene

$$\begin{aligned} q_i &= h^3 \int_0^1 \left[\frac{(1-t)^3}{3!} - 3\frac{(1-t)^4}{4!} \right] f^{(5)}(x_i + th)dt \\ &\quad + h^3 \int_0^1 \left[\frac{(1-t)^3}{3!} - 3\frac{(1-t)^4}{4!} \right] f^{(5)}(x_i - th)dt. \end{aligned}$$

Utilizando el teorema del valor medio y calculando la integral

$$\int_0^1 \left[\frac{(1-t)^3}{3!} - 3\frac{(1-t)^4}{4!} \right] dt = \frac{1}{60},$$

se obtiene finalmente que

$$q_i = \frac{1}{60}h^3 \left(f^{(5)}(\xi_i) + f^{(5)}(\eta_i) \right),$$

con $\xi_i \in [x_{i-1}, x_i]$ y $\eta_i \in [x_i, x_{i+1}]$, por consiguiente

$$q_i \leq \frac{h^3}{20} \|f^{(5)}\|_{\infty}. \quad (\text{III.2.16})$$

Restando (III.2.14) con (III.2.15), se tiene

$$e_i = f'(x_i) - p_i,$$

que es solución del sistema de ecuaciones

$$\frac{1}{h}[e_{i-1} + 4e_i + e_{i+1}] = q_i, \quad i = 1, \dots, n-1.$$

Sea i_o tal que

$$|e_{i_o}| = \max_{i=1, \dots, n} |e_i|,$$

en consecuencia, se tiene:

$$\begin{aligned} 4e_{i_o} &= hq_{i_o} - e_{i_o-1} - e_{i_o+1}, \\ 4|e_{i_o}| &\leq h|q_{i_o}| + |e_{i_o-1}| + |e_{i_o+1}| \\ &\leq h|q_{i_o}| + |e_{i_o}| + |e_{i_o}|, \end{aligned}$$

de donde finalmente

$$|e_i| \leq |e_{i_o}| \leq \frac{h}{2} |q_{i_o}|.$$

□

Antes de poder estimar el error cometido por el polinomio de interpolación *spline*, es necesario el siguiente teorema.

Teorema III.2.6.- Si $x_i = x_0 + ih$, con $h = (b - a)/n$, $f \in \mathcal{C}^5[a, b]$, $s(x)$ el *spline* fijo en los bordes, y $p(x)$ el polinomio de interpolación de Hermite sobre el intervalo $[x_{i-1}, x_i]$, tal que

$$p(x_j) = s(x_j), \quad p'(x_j) = f'(x_j), \quad j = i-1, i.$$

Entonces

$$|p(x) - s(x)| \leq \frac{h^5}{240} \|f^{(5)}\|_\infty \quad \text{para } x \in [x_{i-1}, x_i]. \quad (\text{III.2.17})$$

Demostración.- El polinomio de Hermite, utilizando (III.2.8), está dado por

$$\begin{aligned} p(x) &= y_{i-1} + y[x_{i-1}, x_i](x - x_{i-1}) \\ &\quad + \frac{(x - x_{i-1})(x - x_i)}{h_i^2} [(f'(x_i) - y[x_{i-1}, x_i])(x - x_{i-1}) \\ &\quad + (f'(x_{i-1}) - y[x_{i-1}, x_i])(x - x_i)], \quad (\text{III.2.18}) \end{aligned}$$

restando (III.2.18) con (III.2.8), se obtiene

$$\begin{aligned} p(x) - s(x) &= \frac{(x - x_{i-1})(x - x_i)}{h_i^2} [(f'(x_i) - p_i)(x - x_{i-1}) \\ &\quad + (f'(x_{i-1}) - p_{i-1})(x - x_i)], \end{aligned}$$

por lo tanto

$$\begin{aligned} |p(x) - s(x)| &\leq |x - x_{i-1}| |x - x_i| \frac{h^4 \|f^{(5)}\|_\infty}{60} [|x - x_i| + |x - x_{i-1}|] \\ &\leq \frac{h^5}{240} \|f^{(5)}\|_\infty. \end{aligned}$$

□

Ahora bien, para tener una estimación del error de la interpolación *spline*, solo falta conocer el error cometido por el polinomio de interpolación de Hermite, estimación que está dada en el siguiente teorema.

Teorema III.2.7.- Sean, $f \in [x_0, x_1]$, $h = x_1 - x_0$, $p(x)$ el polinomio de interpolación de Hermite que satisface:

$$\begin{aligned} p(x_i) &= f(x_i), & i &= 0, 1; \\ p'(x_i) &= f'(x_i), & i &= 0, 1; \end{aligned}$$

entonces

$$|f(x) - p(x)| \leq \frac{h^4}{384} \|f^{(4)}\|_{\infty}. \quad (\text{III.2.19})$$

Demostración.- Sean $\epsilon > 0$ suficientemente pequeño, $p_{\epsilon}(x)$ el polinomio de interpolación que satisface:

$$\begin{aligned} p_{\epsilon}(x_0) &= f(x_0), & p_{\epsilon}(x_0 + \epsilon) &= f(x_0 + \epsilon), \\ p_{\epsilon}(x_1) &= f(x_1), & p_{\epsilon}(x_1 - \epsilon) &= f(x_1 - \epsilon). \end{aligned}$$

Por el teorema (III.1.13), se tiene para todo ϵ .

$$|f(x) - p_{\epsilon}(x)| \leq |(x - x_0)(x - x_0 - \epsilon)(x - x_1 + \epsilon)(x - x_1)| \frac{\|f^{(4)}\|}{4!},$$

haciendo tender ϵ hacia 0, se obtiene

$$\begin{aligned} |f(x) - p(x)| &\leq |(x - x_0)^2(x - x_1)^2| \frac{\|f^{(4)}\|}{4!} \\ &\leq \frac{h^4}{384} \|f^{(4)}\|. \end{aligned}$$

□

Finalmente, con los anteriores teoremas se puede dar la siguiente estimación del error del *spline* fijo en los bordes.

Teorema III.2.8.- Sean, $f \in \mathcal{C}^5[a, b]$, $x_i = x_0 + ih$, $i = 0, \dots, n$, con $h = (b - a)/n$, $s(x)$ el *spline* fijo en los bordes de la función f respecto a la subdivisión equidistante, entonces

$$|f(x) - s(x)| \leq \frac{h^4}{384} \max_{x \in [x_{i-1}, x_i]} |f^{(4)}(x)| + \frac{h^5}{240} \|f^{(5)}\|_{\infty}. \quad (\text{III.2.20})$$

Demostración.- Suficiente utilizar la desigualdad del triángulo en los dos anteriores teoremas, es decir

$$|f(x) - s(x)| \leq |f(x) - p(x)| + |p(x) - s(x)|.$$

□

Los resultados obtenidos en los teoremas de esta subsección han sido demostrados para el caso de divisiones equidistantes, no obstante modificando las demostraciones para divisiones más generales se puede obtener resultados similares en la mayoración del error de la interpolación *spline*, no hay que sorprenderse que las mayoraciones sean muy parecidas al del caso equidistante, con la diferencia que h esté elevado a una potencia de un grado menor.

Por otro lado, los teoremas enunciados consideran el caso del *spline* fijo en los bordes, para los otros tipos de *spline*, las mayoraciones del error de interpolación se deben efectuar utilizando los resultados anteriores más un error debido al cambio de tipo de *spline*.

Teorema III.2.9.- Sean, $f \in \mathcal{C}^5[a, b]$, $x_i = x_0 + ih$, $i = 0, \dots, n$; con $h = (b - a)/n$, $\hat{s}(x)$ el *spline* fijo en los bordes, $s(x)$ es el *spline* natural. Entonces

$$|\hat{s}(x) - s(x)| \leq \frac{h^2}{8} \max_{x \in I_1 \cup I_n} |f''(x)| + \frac{h^5}{240} \|f^{(5)}\|_\infty, \quad (\text{III.2.21})$$

donde $I_1 = [x_0, x_n]$ y $I_n = [x_{n-1}, x_n]$.

Demostración.- Sustrayendo los sistemas lineales que determinan \hat{p}_i y p_i en la construcción de los *splines* fijo en los bordes y natural, se obtiene:

$$\begin{aligned} (\hat{p}_{i-1} - p_{i-1}) + 4(\hat{p}_i - p_i) + (\hat{p}_{i+1} - p_{i+1}) &= 0 \quad i = 0, \dots, n-1; \\ 2(\hat{p}_0 - p_0) + (\hat{p}_1 - p_1) &= C_0; \\ 2(\hat{p}_n - p_n) + (\hat{p}_{n-1} - p_{n-1}) &= C_n; \end{aligned}$$

donde:

$$C_0 = 3y[x_0, x_1] - 2\hat{p}_0 - \hat{p}_1, \quad C_n = 3y[x_{n-1}, x_n] - 2\hat{p}_n - \hat{p}_{n-1}.$$

Sea $i_o \in \{0, 1, \dots, n\}$, tal que

$$|\hat{p}_{i_o} - p_{i_o}| = \max_i |\hat{p}_i - p_i|.$$

Ahora bien, $i_o = 0$, o bien $i_o = n$, por que de lo contrario se tendría

$$4|\hat{p}_{i_o} - p_{i_o}| \leq |\hat{p}_{i_o-1} - p_{i_o-1}| + |\hat{p}_{i_o+1} - p_{i_o+1}| \leq 2|\hat{p}_{i_o} - p_{i_o}|.$$

Suponiendo que $i_o = 0$, se tiene la siguiente mayoración

$$|\hat{p}_i - p_i| \leq |\hat{p}_0 - p_0| \leq C_0;$$

en el otro caso, se obtiene

$$|\hat{p}_i - p_i| \leq |\hat{p}_n - p_n| \leq C_n.$$

Utilizando la fórmula de Taylor, con resto en forma de integral, se obtiene para la función f en el punto x_0 , los siguientes resultados:

$$f(x_1) - f(x_0) = hf'(x_0) + h^2 \int_0^1 (1-t)f''(x_0+th)dt, \quad (\text{III.3.22})$$

$$f'(x_1) = f'(x_0) + h \int_0^1 f''(x_0+th)dt; \quad (\text{III.3.23})$$

y en el punto x_n los siguientes resultados:

$$f(x_{n-1}) - f(x_n) = -hf'(x_n) + h^2 \int_0^1 (1-t)f''(x_n-th)dt,$$

$$f'(x_{n-1}) = f'(x_n) - h \int_0^1 f''(x_n-th)dt.$$

La fórmula (III.3.13) del teorema III.3.5, conduce a las estimaciones:

$$\hat{p}_1 = f'(x_1) + \epsilon \frac{h^4}{60} \|f^{(5)}\|_\infty, \quad \text{con } |\epsilon| \leq 1; \quad (\text{III.3.24})$$

$$\hat{p}_{n-1} = f'(x_{n-1}) + \epsilon' \frac{h^4}{60} \|f^{(5)}\|_\infty, \quad \text{con } |\epsilon'| \leq 1.$$

Suponiendo que $i_o = 0$, introduciendo (III.3.22-24) en C_0 , se obtiene

$$\begin{aligned} C_0 &= 3f'(x_0) + h \int_0^1 (1-t)f''(x_0+th)dt - 2f'(x_0) - f'(x_0) \\ &\quad - h \int_0^1 f''(x_0+th)dt - \epsilon \frac{h^4}{60} \|f^{(5)}\|_\infty \\ &= -h \int_0^1 tf''(x_0+th)dt - \epsilon \frac{h^4}{60} \|f^{(5)}\|_\infty, \end{aligned}$$

de donde

$$|C_0| \leq \frac{1}{2}h \max_{x \in I_1} |f''(x)| + \frac{h^4}{60} \|f^{(5)}\|_\infty.$$

Si $i_o = n$, se obtiene similarmente

$$|C_n| \leq \frac{1}{2}h \max_{x \in I_n} |f''(x)| + \frac{h^4}{60} \|f^{(5)}\|_{\infty}.$$

La diferencia de ambos *splines* está mayorada por lo tanto por

$$\begin{aligned} |\hat{p}(x) - p(x)| &= \left| \frac{(x - x_{i-1})(x - x_i)}{h^2} [(\hat{p}_i - p_i)(x - x_{i-1}) + (\hat{p}_{i-1} - p_{i-1})(x - x_i)] \right| \\ &\leq \frac{h}{4} \left(\frac{1}{2} \max_{x \in I_1 \cup I_n} |f''(x)| + \frac{h^5}{60} \|f^{(5)}\|_{\infty} \right) (|x - x_{i-1}| + |x - x_i|) \\ &\leq \frac{h^2}{8} \max_{x \in I_1 \cup I_n} |f''(x)| + \frac{h^5}{240} \|f^{(5)}\|_{\infty}. \end{aligned}$$

□

El *spline* fijo en los bordes es más preciso que el natural, siempre y cuando se conozca los valores de la derivada de la función interpolada en los bordes. En la demostración del último teorema puede observarse que la diferencias de los p_i verifican una ecuación de diferencias finitas con valores en la frontera, resolviendo esta ecuación puede observarse que la diferencia entre los p_i es mucho menor en los nudos centrales, que en aquellos que están cerca de los bordes. Por consiguiente la estimación del teorema (III.2.9) es muy pesimista, en la práctica la diferencia de los errores es menor.

Aplicación de spline

Al inicio de esta sección, los *splines* fueron abordados como un instrumento numérico de aproximación, como una alternativa a la interpolación polinomial de tipo Lagrange. Esto debido a sus propiedades de no presentar problemas de inestabilidad numérica, por sus cualidades de convergencia y por la poca curvatura que presentan.

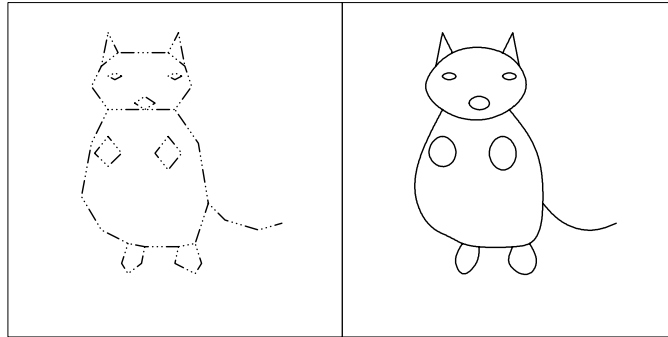


Figura III.2.1. Aplicación gráfica de los *splines*

Ahora bien, los *splines* son utilizados también como un instrumento gráfico. Una gran cantidad de los algoritmos desarrollados en grafismo asistido por computadora, utilizan ampliamente los *splines* en sus diferentes versiones. En la figura III.2.2, se observa con claridad la potencia de la interpolación *spline*. En el dibujo de la izquierda, se tiene los puntos unidos por segmentos rectilíneos, mientras que en el dibujo de la derecha unidos por *splines*.

Ejercicios

- 1.- Considerar puntos equidistantes $x_i = x_0 + ih$, $h > 0$. Mostrar que existe un único *spline* llamado B-spline, tal que para un j fijo se tiene:

$$\begin{aligned} s(x_j) &= 1, \\ s(x) &= 0 \quad \text{si } |x - x_j| \geq 2h. \end{aligned}$$

Dibujar.

- 2.- Desarrollar una fórmula para los *splines* periódicos.
Mostrar la existencia y la unicidad del *spline* que pasa por (x_i, y_i) , $i = 0, \dots, n$, tal que

$$s'(x_0) = s'(x_n), \quad s''(x_0) = s''(x_n).$$

- 3.- Escribir una subrutina `SPLICO(N,X,Y,DY,P,A,B)`.
Los datos son N , $X(0:N)$, $Y(0:N)$ y $P(0)$, $P(N)$. La subrutina debe dar como valores $DY(0:N-1)$ diferencias divididas, $P(0:N)$ los p_i , $A(0:N-1)$ los $(p_i - y[x_{i-1}, x_i])$ y $B(0:N-1)$ los $(p_{i-1} - y[x_{i-1}, x_i])$.
- 4.- Escribir una función `SPLIVA(N,X,Y,DY,A,B,T)` que calcula el valor del *spline* en el punto T . Examinarla sobre varias funciones de su elección.
- 5.- Sea $x_i = x_0 + ih$, $i = 0, \dots, n$, una división equidistante y $l_j(x)$ el *spline* de tipo Lagrange que satisface

$$\begin{cases} 1, & \text{si } i = j; \\ 0, & \text{sino;} \end{cases}$$

y $l'_j(x_0) = 0$, $l'_j(x_n) = 0$. Para las pendientes $p_i = l'_j(x_i)$ mostrar:

$$\begin{aligned} -\frac{3}{h} &< p_j < \frac{2}{h}; \\ \dots, p_{j-3} &> 0, \quad p_{j-2} < 0, \quad p_{j-1} > 0, \quad p_{j+1} < 0, \quad p_{j+2} > 0, \dots \end{aligned}$$

- 6.- Los *splines* $l_j(x)$ del anterior ejercicio no cambian de signo en ningún intervalo $[x_{i-1}, x_i]$.
- 7.- Sobre el intervalo $[x_{i-1}, x_i]$

$$\sum_{i=0}^n |l_j(x)| = t_i(x)$$

es el *spline* que satisface

$$\begin{cases} 1 & \text{si } j = i + 2k \text{ o } j = i - 2k - 1 \text{ con } k = 0, 1, \dots, \\ -1 & \text{sino.} \end{cases}$$

- 8.- Utilizando **SPLIC0** y **SPLIVA** de los ejercicios 3 y 4. Calcular para $x_i = i/n$, $n = 20$:

a) el *spline* $l_7(x)$,

b) la función $\sum_{i=0}^n |l_i(x)|$,

hacer dibujos. Verificar las propiedades de los ejercicios 5, 6 y 7.

III.3 Extrapolación

En las dos primeras secciones, se ha abordado la interpolación polinomial y la interpolación *spline* como un medio de aproximación dentro un intervalo cerrado y acotado $[a, b]$. Sin embargo estos procedimientos no sirven cuando se trata de determinar el valor de una función o el límite de una función definida en un intervalo abierto (a, b) o mas aun si la función está definida en un intervalo no compacto.

En esta sección, se estudiará la extrapolación por medio de funciones polinomiales, como una consecuencia natural de los resultados obtenidos en la primera sección de este capítulo. Por otro lado, se verá que la extrapolación polinomial, no es solamente un instrumento para determinar valores fuera de cierto rango, si no también como un medio para acelerar la convergencia de sucesiones convergentes.

Sea $f : (0, 1] \rightarrow \mathbb{R}$, se desea determinar

$$\lim_{x \rightarrow 0_+} f(x), \quad (\text{III.3.1})$$

sabiendo que se puede determinar sin ningún problema $f(h)$ para $h > 0$. Los otros casos de determinación de límite se convierten el primer caso por traslación afín, o planteando $h = 1/x$ si se desea determinar el límite cuando $x \rightarrow \infty$. Por consiguiente se supondrá que f tiene un desarrollo asintótico por derecha de la forma

$$f(h) = a_0 + a_1 h + \cdots + a_n h^n + \mathcal{O}(h^{n+1}), \quad h > 0. \quad (\text{III.3.2})$$

Sea $1 \geq h_0 > h_1 > \cdots > h_n > 0$ una subdivisión decreciente del intervalo $(0, 1]$, por lo tanto

$$\lim_{x \rightarrow 0_+} f(x) \approx p(0), \quad (\text{III.3.3})$$

donde $p(x)$ es el polinomio de interpolación que pasa por $(h_i, f(h_i))$, $i = 0, \dots, n$. Escogiendo adecuadamente estos h_i se puede obtener un algoritmo facil y sencillo de implementar, para este efecto es necesario la siguiente proposición.

Proposición III.3.1.- Sean (x_i, y_i) , $i = 0, 1, \dots, n$, los x_i diferentes entre si, $p_1(x)$ el polinomio de interpolación de grado $n - 1$ que pasa por (x_i, y_i) , $i = 0, \dots, n - 1$, y $p_2(x)$ el polinomio de interpolación de grado $n - 1$ que pasa por (x_i, y_i) , $i = 1, \dots, n$: entonces el polinomio $p(x)$ de grado n que pasa por (x_i, y_i) , $i = 0, \dots, n$; está dado por

$$p(x) = p_1(x) \frac{x_n - x}{x_n - x_0} + p_2(x) \frac{x - x_0}{x_n - x_0}. \quad (\text{III.3.4})$$

Demostración.- Se puede ver fácilmente que $p(x)$ es un polinomio de grado menor o igual a n , En los nudos, excepto para $i = 0$ o $i = n$, se tiene

$$\begin{aligned} p(x_i) &= \frac{x_n - x_i}{x_n - x_0} p_1(x_i) + \frac{x_i - x_0}{x_n - x_0} \\ &= y_i \left(\frac{x_n - x_i}{x_n - x_0} + \frac{x_i - x_0}{x_n - x_0} \right), \end{aligned}$$

para $i = 0$ y $i = n$ verificación inmediata. \square

Corolario III.3.2.- Con las mismas hipótesis del teorema precedente, se tiene

$$\begin{aligned} p(0) &= \frac{p_1(0)x_n - p_2(0)x_0}{x_n - x_0} \\ &= p_1(0) + \frac{p_1(0) - p_2(0)}{x_n/x_0 - 1}. \end{aligned} \quad (\text{III.3.5})$$

Demostración.- Verificación inmediata. \square

Ahora bien, volviendo al problema de extrapolación, sea $f : (0, 1] \rightarrow \mathbb{R}$; $1 \geq h_0 > h_1 > \dots > h_n > 0$ una subdivisión decreciente del intervalo $(0, 1]$. Suponiendo que f admite el desarrollo asintótico por derecha, dado por

$$f(h) = a_0 + a_1 h + \dots + a_n h^n + \mathcal{O}(h^{n+1}),$$

$p(x)$ el polinomio de interpolación que pasa por $(x_i, f(x_i))$, $i = 0, \dots, n$; entonces se puede suponer

$$\lim_{h \rightarrow 0_+} f(x) \approx p(0).$$

Definiendo $p_{j,k}$ el polinomio de interpolación de grado $\leq k$ que pasa por $(h_i, f(h_i))$, $j - k, j$; se tiene por consiguiente

$$p(x) = p_{n,n}(x). \quad (\text{III.3.6})$$

Utilizando la proposición III.3.1 y el corolario III.3.2 se tiene las siguientes relaciones recursivas para determinar $p(0)$:

$$p_{j0}(0) = f(h_j), \quad (\text{III.3.7a})$$

$$p_{j,k+1}(0) = p_{j,k}(0) + \frac{p_{j,k}(0) - p_{j-1,k}(0)}{h_{j-k-1}/h_j - 1}. \quad (\text{III.3.7b})$$

La práctica en la extrapolación numérica muestra que es práctico y conveniente utilizar sucesiones de la forma $h_j = 1/n_j$, donde $\{n_j\}_{j=0}^{\infty}$ es una sucesión creciente de números naturales, por consiguiente la fórmula (III.3.7) se convierte en

$$\begin{aligned} p_{j0}(0) &= f(h_j), \\ p_{j,k+1}(0) &= p_{j,k}(0) + \frac{p_{j,k}(0) - p_{j-1,k}(0)}{n_{j-k-1}/n_j - 1}. \end{aligned} \quad (\text{III.3.8})$$

Las tres sucesiones de enteros positivos crecientes son:

- i) Armónica: 1, 2, 3, 4, 5, ...;
- ii) Romberg: 1, 2, 4, 8, 16, ...;
- iii) Bulirsch: 1, 2, 3, 4, 6, 8, 12, 16, ...,
es decir los enteros de la forma 2^i y $3 \cdot 2^i$.

Teorema III.3.3.- Sea $f : (0, 1] \rightarrow \mathbb{R}$, supóngase que f tiene el desarrollo asintótico por derecha en $x = 0$, para $k = 0, 1, 2, \dots, k_o$; dado por

$$f(x) = a_0 + a_1 x + \dots + a_k x^k + R_{k+1}(x),$$

con $|R_{k+1}(x)| \leq C_{k+1} x^{k+1}$, para $x > 0$. Sea $h_j = 1/n_j$, con n_j una sucesión creciente de naturales, entonces el procedimiento de extrapolación definido por (III.3.8) verifica la siguiente propiedad:

Si para todo $j > 0$ se tiene $h_{j+1}/h_j \leq r < 1$, entonces

$$\forall k \text{ con } 0 \leq k \leq k_o, \quad T_{m,k} = a_0 + \mathcal{O}(h_{m-k}^{k+1}). \quad (\text{III.3.9})$$

Demostración.- Escribiendo el polinomio de interpolación de grado $\leq k$, se tiene

$$f(x) = p_{mk}(x) + R_{k+1}(x),$$

de donde

$$p_{mk}(x) - f(x) = \sum_{i=m-k}^m R_{k+1}(h_i) l_{m,k,i}(x), \quad (\text{III.3.10})$$

con $l_{m,k,i}(h) = \prod_{\substack{j=m-k \\ j \neq i}}^m \frac{h - h_j}{h_i - h_j}$. Para $x = 0$, se deduce que

$$T_{mk} - a_0 = \sum_{i=m-k}^m \prod_{\substack{j=m-k \\ j \neq i}}^m \frac{h_j/h_i}{h_j/h_i - 1} R_{k+1}(h_i).$$

Si $j < i$, se mayor $\left| \frac{h_j/h_i}{h_j/h_i - 1} \right|$ por $\frac{1}{1 - r^{i-j}}$, mientras que
 si $i < j$, se mayor $\left| \frac{h_j/h_i}{h_j/h_i - 1} \right|$ por $\frac{r^{j-i}}{1 - r^{j-i}}$, finalmente $|R_{k+1}(h_i)|$ por
 $C_{k+1} (r^{i-m+k} h_{m-k})^{k+1}$, deduciendo

$$|T_{m,k} - a_0| \leq C(k, r) C_{k+1} h_{m-k}^{k+1}$$

donde $C(k, r)$ es una constante independiente de m . \square

La convergencia hacia a_0 es obtenida en cada una de las columnas del tablero, para la primera columna la convergencia es $\mathcal{O}(h_m)$, para la segunda columna la velocidad de convergencia está dada por $\mathcal{O}(h_{m-1}^2)$, y para la k -ésima columna la velocidad de convergencia es igual a $\mathcal{O}(h_{m-k+1}^{k+1})$. Por consiguiente, si $a_1 \neq 0$ la convergencia de la k -ésima columna es k veces mas rápida que de la primera columna.

Las sucesiones de Romberg y Bulirsch satisfacen las condiciones del teorema III.3.3 con $r = 1/2$ y $r = 3/4$. Sin embargo la otra sucesión utilizada ampliamente en las extrapolaciones numéricas como ser la armónica no cumple el requisito que $h_{j+1}/h_j \leq r < 1$. No obstante se tiene a disposición el siguiente teorema.

Teorema III.3.4.- *Mismas hipótesis del teorema III.3.3 para la función f , se define $h_n = 1/n$. Entonces se tiene:*

$$\forall k \text{ con } 0 \leq 2k \leq k_o \quad T_{mk} = a_0 + \mathcal{O}(h_m^{k+1}). \quad (\text{III.3.11})$$

Demostración.- Se tiene

$$\begin{aligned} R_{k+1}(h) &= q_k(h) + R_{2k+1}(h) \text{ con} \\ q_k(h) &= a_{k+1}h^{k+1} + \dots + a_{2k}h^{2k}. \end{aligned}$$

Utilizando (III.3.10), se deduce

$$T_{mk} - a_0 = \sum_{i=m-k}^m (q_k(h_i) l_{m,k,i}(0) + R_{2k+1}(h_i) l_{m,k,i}(0)).$$

Por el teorema III.1.13, se tiene

$$q_k(0) - \sum_{i=m-k}^m q_k(h_i) l_{m,k,i}(0) = \frac{1}{(k+1)!} \prod_{j=m-k}^m (-h_j) q_k^{(k+1)}(\xi),$$

lo que muestra que

$$\sum_{i=m-k}^m q_k(h_i) l_{m,k,i}(0) = \mathcal{O}(h_m^{k+1}).$$

Por otro lado

$$l_{m,k,i}(0) = \prod_{\substack{j=m-k \\ j \neq i}}^m \frac{h_j}{h_j - h_i} = \prod_{\substack{j=m-k \\ j \neq i}}^m \frac{i}{i - j},$$

de donde $|R_{2k+1}(h_i) l_{m,k,i}(0)| \leq C_{2k+1} h_i^{2k+1} i^k \leq C_{2k+1} y_0^k y_i^{k+1}$,
deduciendo fácilmente el teorema. \square

Debido al error de redondeo que se comete en el cálculo del polinomio de interpolación, y por consiguiente en el proceso de extrapolación, se debe evitar la utilización de polinomios de interpolación de grado muy elevado, la práctica aconseja no pasar de 10.

Por otro lado, la extrapolación numérica es muy utilizada en la construcción de métodos de integración, como también en la construcción de métodos de resolución de ecuaciones diferenciales.

Ejercicios

- 1.- Sea $f : (0, 1] \longrightarrow \mathbb{R}$ una función cuyo desarrollo asintótico por derecha en $x = 0$ está dado por

$$f(x) = a_0 + a_1 x^2 + \cdots + a_k x^{2k} + \mathcal{O}(x^{2k+2}).$$

Modificar el algoritmo de Aitken-Naville, es decir el proceso de extrapolación, de manera que el número de operaciones se reduzca ostensiblemente.

- 2.- En esta sección, se da un ejemplo de cómo acelerar la convergencia de una sucesión para calcular π . Suponiendo que el desarrollo asintótico es par, como en el ejercicio 1, recalcule y compare.
- 3.- Utilizando el procedimiento de la sección III.1 para determinar la influencia de los errores de redondeo en la determinación del polinomio de interpolación. Estudie la influencia de los errores de redondeo en la extrapolación, suponiendo que los únicos errores de redondeo que se comete son aquellos relativos a la evaluación de la función f a ser extrapolada.
- 4.- Calcule $\sqrt{2}$, utilizando un procedimiento de extrapolación, por ejemplo $\sqrt{1} = 1$, $\sqrt{1,21} = 1,1$, $\sqrt{1,44} = 1,2$, $\sqrt{1,69} = 1,3$, $\sqrt{1,96} = 1,4$, $\sqrt{1,9881} = 1,41$, ...

Capítulo IV

Ecuaciones No Lineales

En el capítulo II, se vio diferentes métodos para resolver problemas lineales, pero también existen problemas no lineales, como por ejemplo ecuaciones cuadráticas, polinomiales, trigonométricas y muchas otras más. La Historia de las Matemáticas muestra que los problemas lineales fueron estudiados y resueltos en tiempos inmemoriales, pero luego el hombre confrontó otros tipos de problemas cuya característica no era precisamente lineal. En el caso lineal la solución de un problema puede darse de manera explícita, es suficiente determinar la inversa de la matriz del sistema y multiplicar para obtener la solución de este problema. En el caso no lineal se continuó con esta óptica, lo cual es siempre posible bajo ciertas condiciones, como ejemplo se tiene el teorema de la función inversa, pero lastimosamente en la mayor parte de los casos esta función inversa no puede ser expresada como una combinación de funciones elementales, entendiéndose como función elemental aquellas que uno estudia en colegio y los primeros semestres de universidad.

En este capítulo, se intentará de cierta manera seguir este desarrollo histórico, dejando sobrentendido que el estudio de los problemas lineales en todas sus variantes es conocido por el lector. Por consiguiente, una primera parte será dedicada al estudio de las soluciones de ecuaciones polinomiales, teniendo como epílogo el Teorema de Galois. Después se abordará los métodos iterativos para encontrar la solución o soluciones de una ecuación, dando condiciones para asegurar la existencia, la unicidad local y la convergencia de tales métodos. Como una clase de método iterativo, se estudiará el Método de Newton, enunciando: teoremas de convergencia, existencia y unicidad de soluciones, algunos problemas frecuentes que se encuentran en la implementación de tal método; como también modificaciones para simplificar su utilización en determinadas situaciones. Finalmente, se verá el equivalente del Método de los Mínimos Cuadrados, en los problemas no lineales, que es el Método de Gauss-Newton.

IV.1 Ecuaciones Polinomiales

Una de la clase de ecuaciones que se confronta a diario son las ecuaciones polinomiales, que son de la forma

$$x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n = 0, \quad (\text{IV.1.1})$$

donde los a_i son números reales o complejos.

Como \mathbb{C} es una extensión del cuerpo \mathbb{R} , se puede suponer que el problema consiste en determinar las raíces o ceros de un polinomio $p(x) \in \mathbb{C}[x]$. La existencia de las soluciones de una ecuación polinomial está dada por el Teorema Fundamental del Algebra que se lo enuncia sin demostración.

Teorema IV.1.1.- *Fundamental del Algebra.* Sea $p(x) \in \mathbb{C}[x]$ de grado n , entonces $p(x)$ tiene exactamente n ceros contando con su multiplicidad.

Comentando este teorema, se puede agregar que un polinomio a coeficientes reales puede no tener ceros reales; pero por el teorema Fundamental del Algebra, éste tiene exactamente n raíces contando con su multiplicidad.

Por otro lado, $p(x)$ puede ser visto como una función de \mathbb{C} en \mathbb{C} , ahora bien toda función polinomial es holomorfa, y si el polinomio es no nulo, entonces los ceros forman un conjunto discreto, es decir que para todo cero x_0 , existe un $r > 0$ tal que $p(x) \neq 0$ para todo $x \in \mathbb{C}$ tal que $0 < |x| < r$. Además, la utilización del teorema de Rouché permite de determinar los discos donde se encuentran los ceros y complementando esta información el número exacto de ceros.

Ecuaciones Resolubles por Radicales

Tal como se dijo en la introducción de este capítulo, se ha buscado inicialmente la manera de expresar la solución de una ecuación por medio de funciones elementales. En esta subsección, se verá precisamente esta manera de expresar las soluciones. Se comenzará con el caso mas simple, las ecuaciones cuadráticas.

Ecuaciones cuadráticas

Una ecuación cuadrática o ecuación polinomial de segundo grado es de la forma

$$ax^2 + bx + c = 0, \quad \text{con } a \neq 0;$$

ecuación que es equivalente a

$$x^2 + \bar{b}x + \bar{c} = 0, \quad (\text{IV.1.2})$$

completando cuadrados se tiene

$$\left(x + \frac{\bar{b}}{2}\right)^2 + \bar{c} - \frac{\bar{b}^2}{4} = 0,$$

obteniendo, así dos raíces:

$$x_1 = -\frac{\bar{b}}{2} + \sqrt{\frac{\bar{b}^2}{4} - \bar{c}}, \quad x_2 = -\frac{\bar{b}}{2} - \sqrt{\frac{\bar{b}^2}{4} - \bar{c}}. \quad (\text{IV.1.3})$$

Si $\frac{\bar{b}^2}{4} - \bar{c} \geq 0$, se utiliza la determinación usual de la raíz cuadrada. En caso contrario, una determinación donde las raíces cuadradas de números negativos esté definida.

Ecuaciones Cúbicas

Las ecuaciones cúbicas o ecuaciones polinomiales de tercer grado son de la forma

$$ax^3 + bx^2 + cx + d = 0, \quad \text{con } a \neq 0; \quad (\text{IV.1.4})$$

esta ecuación dividiendo por a se reduce a una ecuación de la forma

$$x^3 + \bar{a}x^2 + \bar{b}x + \bar{c} = 0, \quad (\text{IV.1.5})$$

planteando $\bar{x} = x + a/3$, la ecuación se convierte en una ecuación de la forma

$$\bar{x}^3 - 3p\bar{x} - 2q = 0. \quad (\text{IV.1.6})$$

Ahora bien, si se plantea $\bar{x} = u + v$, se obtiene

$$u^3 + 3u^2v + 3uv^2 + v^3 - 3p(u + v) - 2q = 0,$$

de donde

$$u^3 + v^3 + (u + v)(3uv - 3p) - 2q = 0,$$

deduciendose, dos ecuaciones para u y v :

$$\begin{cases} uv = p, \\ u^3 + v^3 = 2q; \end{cases}$$

por consiguiente $u^3v^3 = p^3$, mostrando así que u^3, v^3 son las raíces del polinomio de segundo grado

$$\lambda^2 - 2q\lambda + p^3, \quad (\text{IV.1.7})$$

por lo tanto:

$$u^3 = q + \sqrt{q^2 - p^3}, \quad v^3 = q - \sqrt{q^2 - p^3};$$

para obtener finalmente la fórmula de *Cardano*

$$\bar{x} = \sqrt[3]{q + \sqrt{q^2 - p^3}} + \sqrt[3]{q - \sqrt{q^2 - p^3}}. \quad (\text{IV.1.8})$$

teniendo cuidado de verificar $uv = p$.

Ecuaciones polinomiales de grado cuarto

Son ecuaciones que pueden escribirse de la forma

$$x^4 + ax^3 + bx^2 + cx + d = 0, \quad (\text{IV.1.9})$$

planteando $\bar{x} = x + a/4$, esta ecuación se convierte en una ecuación de la forma

$$\bar{x}^4 - p\bar{x}^2 - 2q\bar{x} - r = 0, \quad (\text{IV.1.10})$$

introduciendo $z \in \mathbb{C}$ en la anterior ecuación, se obtiene la siguiente ecuación equivalente

$$(\bar{x}^2 + z)^2 = (2z + p)\bar{x}^2 + 2q\bar{x} + z^2 + r,$$

se elige z de manera que el segundo miembro de la ecuación precedente sea un cuadrado perfecto, y eso ocurre, si y solamente si

$$(2z + p)(z^2 + r) - q^2 = 0, \quad (\text{IV.1.11})$$

de donde z es raíz de una ecuación de tercer grado, que ya ha sido resuelta anteriormente. Habiendo determinado z , se tiene:

$$\begin{aligned} (\bar{x}^2 + z)^2 &= (\alpha\bar{x} + \beta)^2, \quad \text{con } \alpha = \sqrt{2z + p}, \beta = \sqrt{z^2 + r}; \\ (\bar{x}^2 + \alpha\bar{x} + z + \beta)(\bar{x}^2 - \alpha\bar{x} + z - \beta) &= 0, \end{aligned}$$

por consiguiente, las cuatro raíces de la ecuación están dadas por:

$$x_1 = \frac{-\alpha + \sqrt{\alpha^2 - 4(z + \beta)}}{2}, \quad (\text{IV.1.12a})$$

$$x_2 = \frac{-\alpha - \sqrt{\alpha^2 - 4(z + \beta)}}{2}, \quad (\text{IV.1.12b})$$

$$x_3 = \frac{-\alpha + \sqrt{\alpha^2 - 4(z - \beta)}}{2}, \quad (\text{IV.1.12c})$$

$$x_4 = \frac{-\alpha - \sqrt{\alpha^2 - 4(z - \beta)}}{2}. \quad (\text{IV.1.12d})$$

Ecuaciones no Resolubles por Radicales

Se ha visto que las ecuaciones polinomiales de grado igual o menor a 4 tienen soluciones que pueden ser expresadas como una composición de radicales. Las fórmulas desarrolladas en el anterior paragrafo fueron estudiadas y comprendidas hasta mediados del siglo XVI, posteriormente se atacó al problema de las ecuaciones de grado superior, en particular a las ecuaciones de grado quinto. Todos los intentos iban en la dirección de encontrar una fórmula general que estuviese expresada con radicales. Pero a principios del siglo pasado Galois mostró su famoso teorema que trastorno en cierta manera el mundo matemático de aquella época. Se enunciará este teorema sin dar la demostración.

Teorema IV.1.2.- *Galois. No existe una fórmula general en forma de radicales para las ecuaciones polinomiales de grado superior o igual a 5.*

Este resultado lejos de descorazonar el estudio de las ecuaciones polinomiales constituyó un aporte, ya que se atacó el problema utilizando métodos analíticos que estaban emergiendo con gran fuerza a principios y mediados del siglo pasado. Por otro lado se abandonó la idea de encontrar una fórmula general, para insistir sobre la posible ubicación de los ceros de un polinomio. Se desarrollaron métodos iterativos para la solución de estos problemas cuyo alcance es mas vasto que el de las ecuaciones polinomiales.

Localización de Ceros

Como se dijo anteriormente, al no poder calcular explícitamente las raíces de un polinomio, es importante determinar en que región se encuentran estas raíces de manera de poder aplicar un método iterativo.

Sea, $p(x)$ un polinomio a coeficientes complejos de grado n , es decir

$$p(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_n,$$

si x es una raíz, se tiene

$$x^n = -\frac{a_1}{a_0} x^{n-1} - \frac{a_2}{a_0} x^{n-2} + \cdots - \frac{a_n}{a_0}.$$

Introduciendo la matriz de *Frobenius* de esta ecuación, se tiene

$$x \begin{pmatrix} x^{n-1} \\ x^{n-2} \\ \vdots \\ x \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} -\frac{a_1}{a_0} & -\frac{a_2}{a_0} & \cdots & -\frac{a_n}{a_0} \\ 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 & 0 \end{pmatrix}}_{F \text{ matriz de Frobenius}} \begin{pmatrix} x^{n-1} \\ x^{n-2} \\ \vdots \\ x \\ 1 \end{pmatrix},$$

de donde $p(x) = 0$, si y solamente si x es un valor propio de F . Por lo tanto la localización de ceros de $p(x)$ es equivalente a localizar los valores propios de la matriz F . A continuación se da un teorema cuya demostración será hecha en el capítulo 5.

Teorema IV.1.3.- Gerschgorin. Sean A una matriz, λ un valor propio de A , entonces existe i tal que

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|. \quad (\text{IV.1.13})$$

Aplicando este teorema a la matriz F se tiene la siguiente estimación:

Corolario IV.1.4.- La raíces del polinomio $p(x) = a_0x^n + \cdots + a_n$, con $a_0 \neq 0$ verifican

$$|x| \leq \max \left(1, \sum_{i=1}^n \left| \frac{a_i}{a_0} \right| \right). \quad (\text{IV.1.14})$$

Uno de los resultados prácticos e importantes en el álgebra de valores propios, es que la propiedad de valor propio es invariante por trasposición, por consiguiente, el:

Corolario IV.1.5.- Las mismas hipótesis del corolario precedente, dan la estimación siguiente

$$|x| \leq \max_{0 \leq i \leq n} \left(1 + \left| \frac{a_i}{a_0} \right| \right). \quad (\text{IV.1.15})$$

Utilizando el teorema de Rouché y otros, se puede encontrar mas estimaciones sobre la localización de los ceros de un polinomio. Por ejemplo, en los ejercicios se debe mostrar la siguiente estimación

$$|x| \leq 2 \max_{1 \leq i \leq n-1} \left(\sqrt[i]{\left| \frac{a_i}{a_0} \right|}, \sqrt[n]{\left| \frac{an}{2a_0} \right|} \right). \quad (\text{IV.1.16})$$

Con el siguiente ejemplo se verá que existen mejores estimaciones que otras, esto depende sobre todo del tipo de polinomio que se tiene.

Ejemplo

Se considerará la ecuación

$$x^n - 2 = 0,$$

la primera estimación, como la segunda dan $|x| \leq 2^n$, mientras que, la tercera da $|x| \leq 2$.

Debido a la aritmética de los ordenadores, el tipo REAL es un conjunto finito, que se asemeja mucho mas a los números racionales que los reales,

por eso que se puede suponer que $p(x) \in \mathbb{Q}[x]$. Por otro lado, cuando los polinomios tienen raíces de multiplicidad mayor a 1, la mayor parte de los algoritmos no funcionan, debido a las inestabilidades que se pueden presentar, o a las características propias de los métodos empleados.

Sea $p(x) \in \mathbb{Q}[x]$, $p'(x)$ su derivada, se define el polinomio $q(x)$ utilizando el algoritmo de Euclides, por

$$q(x) = \text{mcd}(p, p'), \quad (\text{IV.1.17})$$

de donde el polinomio $r(x) = p(x)/q(x)$ tiene las mismas raíces que $p(x)$ pero todas simples. Por consiguiente, se ha visto un procedimiento simple para reducir un polinomio $p(x)$ a otro polinomio simple. En lo que sigue, se puede suponer que los polinomios considerados tienen raíces simples.

Método de Newton

Uno de los métodos comúnmente utilizados en la determinación de raíces de un polinomio, consiste en el uso del método de Newton. Aquí se formulará este método, sin estudiar la convergencia y el cálculo de error, dejando esto para el siguiente parágrafo. Por lo tanto, el método está dado por la relación recursiva siguiente

- x_0 arbitrario, próximo a una solución.
- $x_{k+1} = x_k - \frac{p(x_k)}{p'(x_k)}$, donde $p(x)$ es el polinomio a determinar sus raíces.

El problema consiste en determinar todas las raíces reales de $p(x) = 0$, suponiendo que se ha encontrado la primera raíz, se podría utilizar nuevamente el método de Newton utilizando otro valor, pero existe la posibilidad de encontrar nuevamente la primera solución. Para evitar esta situación se puede proceder básicamente de dos maneras. La primera, una vez que se ha encontrado la raíz ξ_1 , se define el polinomio

$$q(x) = \frac{p(x)}{x - \xi_1}, \quad (\text{IV.1.18})$$

aplicando Newton sobre $q(x)$. El principal inconveniente de este procedimiento radica en el hecho que ξ puede no ser racional, motivo por el cual $r(x)$ ya no es exacto y la propagación de errores de redondeo puede dar resultados completamente alejados de la realidad. El segundo procedimiento consiste en aplicar el método de Newton a $q(x)$, pero sin calcular los coeficientes de $q(x)$. Sea $p(x)$ el polinomio a determinar sus raíces, suponiendo que se conoce de antemano ξ_1, \dots, ξ_k raíces de $p(x)$, de donde

$$q(x) = \frac{p(x)}{(x - \xi_1) \cdots (x - \xi_k)}. \quad (\text{IV.1.19})$$

Introduciendo el logaritmo y derivando se obtiene:

$$\begin{aligned}\log q(x) &= \log p(x) - \sum_{i=1}^k \log(x - \xi_i), \\ \frac{q'(x)}{q(x)} &= \frac{p'(x)}{p(x)} - \sum_{i=1}^k \frac{1}{(x - \xi_i)}.\end{aligned}\tag{IV.1.20}$$

Aplicando el método de Newton a (IV.1.19), utilizando (IV.1.20), se obtiene como m -sima iteración.

$$x_{m+1} = x_m - \frac{p(x_m)}{p'(x_m) - p(x_m) \sum_{i=1}^k \frac{1}{(x_m - \xi_i)}}.\tag{IV.1.21}$$

Este último procedimiento se conoce como el método de *Maehly*. Este método es numéricamente estable, en efecto considerando el siguiente ejemplo dado por *Stoer*.

Ejemplo

Se considera el polinomio $p(x)$ definido por

$$p(x) = \prod_{j=0}^{12} (x - 2^{-j}) = x^{13} + a_1 x^{12} + \cdots + a_{12}.$$

Calculando los coeficientes en simple precisión e utilizando el método de Newton en sus dos variantes se obtiene la tabla IV.1.1.

Así mismo, el método de Newton permite encontrar las raíces complejas, mas precisamente las raíces no reales, de una ecuación polinomial. Para tal efecto, el punto de partida debe ser un número complejo no real, pues puede observarse que partiendo de un punto real, los valores obtenidos por el método de Newton son reales, ya sea con la primera variante, o con la mejora de Maschley. Puede suceder que una vez encontradas todas las raíces reales del polinomio, siendo éstas no todas las raíces; si no se tiene cuidado, se producirá un fenómeno de oscilación con el método de Newton. Por lo tanto hay que agregar al algoritmo que a partir de un número determinado de iteraciones, si no se ha alcanzado la convergencia hacia una raíz, se detenga el proceso, quedando por lo tanto dos alternativas: la primera verificar si todas las raíces reales han sido calculadas, para luego comenzar con un número complejo no real; la segunda alternativa consiste en cambiar simplemente de punto inicial.

Tabla IV.1.1. Error en el cálculo de raíces para un caso extremo.

	METODO DIVISION		METODO MAEHLI	
SOL EXAC	SOL NUM	ERROR	SOL NUM	ERROR
1/1	1.00	0.0	1.000	0.0
1/2	0.500	$0.950 \cdot 10^{-4}$	0.5000	$0.894 \cdot 10^{-7}$
1/2 ²	-0.15	0.400	0.2500	$0.894 \cdot 10^{-7}$
1/2 ³	-0.426	0.551	0.1250	$0.820 \cdot 10^{-7}$
1/2 ⁴	0.560	0.498	$0.6250 \cdot 10^{-1}$	$0.138 \cdot 10^{-6}$
1/2 ⁵	-0.132	0.163	$0.671 \cdot 10^{-7}$	$0.140 \cdot 10^{-7}$
1/2 ⁶	0.266	0.250	$0.1562 \cdot 10^{-1}$	$0.419 \cdot 10^{-8}$
1/2 ⁷	0.157	0.149	$0.7813 \cdot 10^{-2}$	$0.168 \cdot 10^{-7}$
1/2 ⁸	$-0.237 \cdot 10^{-1}$	0.028	$0.1953 \cdot 10^{-2}$	$0.168 \cdot 10^{-7}$
1/2 ⁹	-0.740	0.742	$0.3906 \cdot 10^{-2}$	$0.291 \cdot 10^{-9}$
1/2 ¹⁰	0.865	0.864	$0.9766 \cdot 10^{-3}$	$0.291 \cdot 10^{-10}$
1/2 ¹¹	0.140	0.140	$0.4883 \cdot 10^{-3}$	$.146 \cdot 10^{-10}$
1/2 ¹²	$-0.179 \cdot 10^{-1}$	$0.181 \cdot 10^{-1}$	$0.2441 \cdot 10^{-3}$	$0.728 \cdot 10^{-10}$

Sucesiones de Sturm

La utilización de sucesiones de Sturm en la resolución de ecuaciones polinomiales permite determinar las raíces reales y no así las raíces no reales. En la mayor parte de los problemas, es solo importante conocer los ceros reales, motivo por el cual los procedimientos que utilizan sucesiones de Sturm son muy comunes en la resolución de ecuaciones polinomiales

Definición IV.1.6.- Una sucesión $\{f_0, f_1, \dots, f_n\}$ de funciones definidas en \mathbb{R} con valores reales, continuamente derivables es una sucesión de Sturm, si:

- a) $f'_0(x^*)f_1(x^*) < 0$ si $f_0(x^*) = 0$, $x^* \in \mathbb{R}$.
- b) $f_{j-1}(x^*)f_{j+1}(x^*) < 0$ si $f_j(x^*) = 0$, $1 \leq j \leq n-1$.
- c) $f_n(x)$ no cambia de signo sobre \mathbb{R} y es positiva

Teorema IV.1.7.- Sea $\{f_0, f_1, \dots, f_n\}$ una sucesión de Sturm, $w(x)$ se denota al número de cambios de signo de $\{f_0(x), f_1(x), \dots, f_n(x)\}$.

Entonces la función $f_0(x)$ tiene exactamente $w(b) - w(a)$ ceros en el intervalo $[a, b]$.

Demostración.- $w(x)$ puede cambiar de valor solamente si uno de los $f_j(x) = 0$. Por consiguiente, sea x^* un tal número y supóngase que $1 \leq j \leq n-1$. Estudiando en un vecindario de x^* , la función w estará determinada por los valores de las siguientes tablas:

	$x^* - \epsilon$	x^*	$x^* + \epsilon$		$x^* - \epsilon$	x^*	$x^* + \epsilon$
f_{j-1}	+	+	+	f_{j-1}	-	-	-
f_j	\pm	0	\pm	f_j	\pm	0	\pm
f_{j+1}	-	-	-	f_{j+1}	+	+	+

	$x^* - \epsilon$	x^*	$x^* + \epsilon$
f_{n-1}	\pm	\pm	\pm
f_n	+	0	+

de donde $w(x^* + \epsilon) = w(x^* - \epsilon)$ cuando $f_j(x^*) = 0$ para $1 \leq j \leq n$.

Ahora bien, si $f_0(x^*) = 0$, estudiando alrededor de un vecindario de x^* se tiene las tablas:

	$x^* - \epsilon$	x^*	$x^* + \epsilon$		$x^* - \epsilon$	x^*	$x^* + \epsilon$
f_0	+	0	-	f_0	-	0	+
f_1	-	-	-	f_1	+	+	+

de donde $w(x^* + \epsilon) = w(x^* - \epsilon) + 1$. □

Volviendo al problema original de encontrar los ceros reales de un polinomio $p(x)$ con raíces simples, a coeficientes reales (rationales), el objetivo es construir una sucesión de Sturm de tal manera que $f_0 = p$. Se define la siguiente sucesión de polinomios de manera recursiva utilizando la división con resto, por

$$\left\{ \begin{array}{l} f_0(x) := p(x), \\ f_1(x) := -p'(x), \\ f_0(x) = g_1(x)f_1(x) - \gamma_1^2 f_2(x), \\ f_1(x) = g_2(x)f_2(x) - \gamma_2^2 f_3(x), \\ \vdots \\ f_{n-1}(x) = g_n(x)f_n(x). \end{array} \right. \quad (\text{IV.1.22})$$

La definición de la sucesión (IV.1.22) es el algoritmo de Euclides para determinar el $\text{mcd}(p(x), p'(x))$. Se tiene el siguiente teorema.

Teorema IV.1.8.- *Supóngase que $p(x)$ solamente tiene raíces simples, entonces la sucesión definida por (IV.1.22) es una sucesión de Sturm.*

Demostración.- Puesto que $p(x)$ tiene raíces simples, si $p'(x^*) = 0$, se tiene $p(x^*) \neq 0$, de donde la condición a) es satisfecha,

- a) $f'_0(x^*)f_1(x^*) = -(f'_0(x^*))^2 < 0$.
- b) Se cumple esta condición por construcción de la sucesión.
- c) $f_n = \text{mcd}(p, p') = C$.

□

Algoritmo

Con el algoritmo de bisección se separa las raíces reales de $p(x)$, es decir se divide en la mitad un intervalo original $[a, b]$ y se continua hasta tener todas las raíces localizadas en subintervalos $[a_i, b_i]$. Se puede continuar con el algoritmo de bisección hasta lograr la precisión deseada, o si no se mejora el resultado utilizando el método de Newton.

Ejercicios

- 1.- El polinomio $x^4 - 8x^3 + 24x^2 - 32x + a_0$, $a_0 = 16$ posee una raíz de multiplicidad 4 en el punto $x = 2$. ¿Cómo cambian las raíces del polinomio si se reemplaza a_0 por $\tilde{a}_0 = a_0(1 + \epsilon)$ con $|\epsilon| \leq \epsilon_{ps}$? Calcular las raíces de este polinomio, ¿es un problema bien condicionado?
- 2.- Sea $p(x)$ un polinomio de grado n a coeficientes reales. Supóngase que todas las raíces $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n$ sean reales.
 - a) Mostrar que el método de Newton converge hacia α_1 , si $x_0 > \alpha_1$.
Indicación.- Para $x > \alpha_1$ los valores de $p(x), p'(x), p''(x)$ tienen el mismo signo que $\lim_{x \rightarrow \infty} p(x)$. Mostrar que $\{x_k\}$ es una sucesión monótona.
 - b) Si x_0 es mucho mas grande que α_1 , entonces el método de Newton converge muy lentamente. ($x_{k+1} \sim x_k(1 - 1/h)$)
- 3.- Considérese el polinomio

$$f(x) = x^6 - 3x^5 + 6x^3 - 3x^2 - 3x + 2.$$

Sin calcular las raíces de este polinomio, determinar el número de raíces distintas de $f(x)$.

- 4.- Encontrar un algoritmo que, para un polinomio arbitrario $f \in \mathbb{Q}[x]$, permita calcular la factorización

$$f(x) = f_1(x) \cdot ((f_2(x))^2 \cdot ((f_3(x))^3 \cdots ((f_k(x))^k,$$

donde f_1, \dots, f_k son primos entre si y donde cada polinomio $f_j(x)$ solo tiene raices simples.

5.- Para un polinomio

$$f(x) = a_0 x^n + a_1 x^{n-1} + \dots + a_{n-1} x + a_n, \quad a_0 \neq 0;$$

con coeficientes en \mathbb{C} , mostrar que todas las raices ξ satisfacen la estimación

$$|\xi| \leq 2 \max \left\{ \left| \frac{a_1}{a_0} \right|, \sqrt{\left| \frac{a_1}{a_0} \right|}, \dots, \sqrt[n-1]{\left| \frac{a_{n-1}}{a_0} \right|}, \sqrt[n]{\left| \frac{a_n}{2a_0} \right|} \right\}. \quad (\text{IV.1.23})$$

Encontrar, para cada n , un polinomio tal que se tenga igualdad en (IV.1.23)

6.- Considérese el polinomio

$$f(x) = x^5 - 5x^4 + 3x^3 + 3x^2 + 2x + 8.$$

a) Determinar el número de raices reales.

b) ¿Cuántas raices son complejas?

c) ¿Cuántas raices son reales y positivas?

Indicación.- Utilizar una sucesión de Sturm.

7.- Sea f $(p+1)$ veces continuamente diferenciable en un vecindario de $\xi \in \mathbb{R}$ y sea ξ una raíz de multiplicidad p . Mostrar que la sucesión $\{x_k\}$, definida por

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

converge hacia ξ , si x_0 es suficientemente proximo de ξ , y que se tiene

$$\frac{x_{k+1} - \xi}{(x_k - \xi)^2} \rightarrow \frac{f^{(p+1)}(\xi)}{p(p+1)f^{(p+2)}(\xi)}.$$

8.- El método de Newton, aplicado a $z^3 - 1 = 0$, da la iteración

$$z_{k+1} = \Phi(z_k), \quad \text{con} \quad \Phi(z) = z - \frac{z^3 - 1}{3z^2} = \frac{2z^3 + 1}{3z^2}.$$

Denótese por $A_j = \{z_0 \in \mathbb{C}; z_k \rightarrow e^{2i\pi j/3}\}$, $j = 0, 1, 2$; los canales de atracción. Mostrar:

a) $A_0 \cap \mathbb{R} = \mathbb{R} - \{\xi_0, \xi_1, \dots\}$ donde $\xi_0 = 0$ y $\xi_{k-1} = \Phi(\xi_k)$,

b) $A_{j+1} = e^{2\pi i/3} A_j$,

c) Calcular $\Phi^{-1}(0)$ y $\Phi^{-2}(0)$ con la fórmula de Cardano. Esto muestra que el sector $\{z \in \mathbb{C} | |\arg(z)| < \pi/3\}$ contiene puntos que no convergen hacia 1.

d) Supóngase que $\Phi^k(\bar{z}) = 0$ para un $k \in \mathbb{N}$. Mostrar que $U \cap A_j \neq \emptyset$ ($j = 0, 1, 2$) para cada vecindario U de \bar{z} .

IV.2 Métodos Iterativos

En la sección precedente, se ha visto que la mayor parte de las ecuaciones a resolver no pueden ser resueltas de manera explícita, es decir mediante una fórmula mágica. Es por esta razón, que el estudio de los métodos iterativos es una necesidad imperiosa para poder encontrar las soluciones de una ecuación. Para tal efecto es también importante conocer el tipo de funciones con las que se está trabajando. En lo que sigue, se estudiará varios aspectos que son fundamentales para la comprensión del tema.

Posición del Problema

Uno de los problemas más comunes, se trata del siguiente. Supóngase, que se tiene la función

$$f : I \longrightarrow \mathbb{R},$$

donde $I \subset \mathbb{R}$ intervalo. El problema, se trata de obtener con una aproximación arbitraria las raíces de la ecuación

$$f(x) = 0. \tag{IV.2.1}$$

Está claro, que sería ridículo esperar obtener, en general, fórmulas de resolución de (IV.2.1), del tipo de fórmulas clásicas resolviendo las ecuaciones de segundo grado. Se ha visto que, es imposible en el caso en que f es un polinomio de grado ≥ 5 . Incluso para el problema más razonable planteado aquí, no existe un método general conduciendo al resultado buscado, los procedimientos teóricos de aproximación pueden conducir, sin hipótesis adecuadas para la función f a cálculos numéricos inextricables, incluso para los ordenadores más potentes que existen en la actualidad.

Por otro lado, incluso si el intervalo I es acotado, puede suceder que la ecuación (IV.2.1) tenga una infinidad de raíces, por ejemplo, cuando f es idénticamente nula en un subintervalo. El ejemplo

$$f(x) = x^3 \sin \frac{1}{x}, \tag{IV.2.2}$$

muestra que puede existir una infinidad de raíces, aun cuando f no sea idénticamente nula en un subintervalo.

Por consiguiente, un primer paso consiste en descomponer el intervalo I , por un número finito o infinito de puntos de subdivisión, en subintervalos en cada uno de los cuales se sepa que, la ecuación no tiene raíz, o la ecuación tenga una y una sola solución. Se estará seguro, que es así cuando la función

es monótona para el primer caso si $f(x)$ no cambia de signo y en el segundo caso si $f(x)$ cambia de signo. Para asegurar la existencia en el segundo caso, se exige como hipótesis suplementaria la continuidad en tal subintervalo.

Ejemplo

La función

$$f(x) = \frac{b_1}{x - a_1} + \frac{b_2}{x - a_2} + \cdots + \frac{b_n}{x - a_n}, \quad (\text{IV.2.3})$$

donde $a_1 < a_2 < \cdots < a_n$ y los b_j son todos diferentes a 0 y del mismo signo, está definida en cada uno de los subintervalos abiertos $]\infty, a_1[,]a_1, a_2[, \dots,]a_{n-1}, a_n[,]a_n, +\infty[$; en cada uno de estos subintervalos es monótona, ya que su derivada tiene el signo opuesto de los b_j . Por otro lado cuando x tiende por derecha a uno de los a_j , $f(x)$ tiende en valor absoluto a ∞ y cuando x tiende por izquierda a uno de los a_j , $f(x)$ tiende en valor absoluto a ∞ , pero con signo opuesto que el límite por derecha. Por lo tanto, se concluye que la función f tiene exactamente una raíz en cada subintervalo.

Para obtener una descomposición de I en intervalos del tipo considerado anteriormente, es decir para separar las raíces, se necesitaría teóricamente estudiar el sentido de variación de la función f , agregando otra hipótesis suplementaria respecto a f , como que f sea continuamente derivable, esto significaría estudiar el signo de $f'(x)$, por consiguiente se estaría obligado a encontrar las raíces de la ecuación $f'(x) = 0$, que salvo algunos casos, su resolución presenta las mismas dificultades que la ecuación original.

Por lo expuesto más arriba, se está en la capacidad de enunciar el siguiente teorema de existencia y unicidad de la solución de una ecuación con su algoritmo de resolución incluido.

Teorema IV.2.1.- Sea $f : I \rightarrow \mathbb{R}$ continua y monótona, entonces

$$\begin{array}{l} \text{la ecuación } f(x) = 0 \text{ tiene} \\ \text{una y una sola solución} \end{array} \iff \begin{array}{l} \text{existen } a < b \in I, \text{ tales} \\ \text{que } f(a)f(b) < 0. \end{array}$$

Además si existiese la solución, ésta puede ser aproximada de manera arbitraria con el algoritmo de la bisección.

La primera observación que se puede hacer al algoritmo de la bisección consiste en que este algoritmo construye subintervalos encajonados reduciendo la longitud de cada subintervalo de la mitad en cada iteración. Pero una de las interrogantes que uno se plantea, es cuando detener el algoritmo. En general, esto sucede cuando el valor absoluto de la función f evaluada en las extremidades es menor a un valor de tolerancia prefijado con anterioridad. Ahora bien, si solamente se exige que f sea monótona y continua puede pasar situaciones, como las del siguiente ejemplo.

Ejemplo

Sea $f : (-1, 1) \rightarrow \mathbb{R}$ definida por

$$f(x) = x^{100}g(x), \quad (\text{IV.2.4})$$

donde g es una función continua que no se anula en $(-1, 1)$. Es evidente que, $x = 0$ es una raíz de la ecuación $f(x) = 0$. Aplicando el algoritmo de la bisección con una tolerancia de 10^{-6} y tomando como puntos de partida $a = -0,9$ y $b = 0,9$ puede suceder que $|f(x)| < TOL$, dejando una gran elección para escoger la raíz de $f(x)$.

Por lo tanto, si no se tiene la hipótesis que $f(x)$ sea derivable y que $f'(x) \neq 0$ para x raíz de la ecuación $f(x) = 0$, se debe tener mucho cuidado en la determinación de la solución numérica del problema, para no cometer grandes errores.

Método de la Falsa Posición

En este paragrafo, se supondrá que

$$f : [a, b] \rightarrow \mathbb{R}$$

es dos veces continuamente derivable, que $f'(x)$ no se anula en el intervalo $]a, b[$ y además $f(a)f(b) < 0$.

La idea para obtener una valor aproximado de la raíz ξ_0 de la ecuación $f(x) = 0$ en el intervalo $I =]a, b[$, consiste en remplazar f por un polinomio que toma los valores de f en determinados puntos de I . Como se ha visto en la primera sección de este capítulo, las ecuaciones polinomiales más simples de resolver son las ecuaciones de primer grado y las ecuaciones cuadráticas. Por razones de simplicidad en la formulación, el método de la falsa posición será estudiado tomando como polinomio, uno de primer grado. Sea $L(x)$ el polinomio de interpolación de primer grado, tal que:

$$L(a) = f(a), \quad L(b) = f(b), \quad (\text{IV.2.5})$$

ver en la figura IV.2.1.

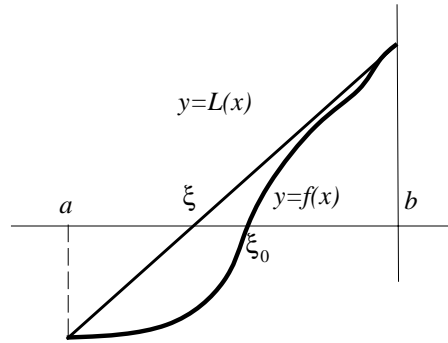


Figura IV.2.1 Método de la Falsa Posición.

Las hipótesis implican que $L(x)$ se anula en un punto $\xi \in I$, que se toma como valor aproximado de ξ_0 . Por consiguiente se trata de mayorar el error cometido $|\xi - \xi_0|$; para tal efecto se recordará el teorema III.1.13 que da el siguiente resultado

$$f(x) - L(x) = \frac{1}{2}f''(\zeta)(x - x_0)(x - x_1) \quad (\text{IV.2.6})$$

con $\zeta \in I$, deduciendo la siguiente proposición.

Proposición IV.2.2.- *Si f' no se anula en el intervalo I , y si $f(\xi_0) = 0$, $L(\xi) = 0$ para $\xi, \xi_0 \in I$, entonces se tiene*

$$\xi - \xi_0 = \frac{1}{2} \frac{f''(\zeta)}{f'(\zeta')} (\xi - a)(\xi - b), \quad (\text{IV.2.7})$$

donde ζ y ζ' son números que pertenecen a I .

Demostración.- Por (IV.2.6), se tiene

$$f(\xi) = \frac{1}{2}f''(\zeta)(\xi - a)(\xi - b),$$

por el teorema del valor medio se tiene

$$f(\xi) = f'(\zeta')(\xi - \xi_0)$$

con ζ' que pertenece al intervalo cuyas extremidades son ξ_0 y ξ . Combinando ambos resultados se obtiene el resultado de la proposición. \square

Corolario IV.2.3.- *Mismas hipótesis de la proposición precedente, y además $|f'(x)| \geq m > 0$, y $|f''(x)| \leq M$ para todo $x \in I$, entonces*

$$|\xi - \xi_0| \leq \frac{M}{8m}(b - a)^2. \quad (\text{IV.2.8})$$

Si el error $|\xi - \xi_0|$ evaluado por la estimación (IV.2.8) no es lo suficientemente pequeño, se puede repetir el procedimiento: se calcula $f(\xi)$ y dependiendo de su signo, la raíz ξ_0 se encuentra en el intervalo $[a, \xi]$ o $[\xi, b]$, al cual se aplica el mismo método obteniendo así un segundo valor aproximado ξ' . Teóricamente, se puede aplicar indefinidamente este procedimiento, y se muestra fácilmente que la sucesión de los números obtenidos converge hacia ξ , ver ejercicios.

Como se dijo al abordar el método de la falsa posición, se puede aproximar la raíz de la ecuación $f(x) = 0$ con un polinomio de segundo grado. Con la misma notación que antes ξ_0 es la raíz de la ecuación, suponiendo

ademas que $f(x)$ es tres veces continuamente diferenciable sobre el intervalo I . Se denota por c el punto medio del intervalo I . Puesto que $f'(x)$ no cambia de signo en este intervalo, $f(c)$ es positivo o negativo. Sea $L(x)$ el polinomio de interpolación de grado 2, tal que

$$L(a) = f(a) \quad L(c) = f(c) \quad L(b) = f(b); \quad (\text{IV.2.9})$$

utilizando la fórmula de Newton del capítulo III, se tiene

$$\begin{aligned} L(x) = f(a) + \frac{f(c) - f(a)}{c - a}(x - a) \\ + 2 \frac{f(b) - 2f(c) + f(a)}{(b - a)^2}(x - a)(x - c), \end{aligned} \quad (\text{IV.2.10})$$

resolviendo esta ecuación, sea $\xi \in I$ la raíz del polinomio de segundo grado, la cual puede ser determinada utilizando el método de resolución de ecuaciones de segundo grado, teniendo el cuidado de escoger la raíz que se encuentra en $[a, b]$. Suponiendo que sobre el intervalo $[a, b]$, f es tres veces continuamente derivable, el teorema III.1.13 da la siguiente estimación

$$f(x) - L(x) = \frac{1}{6} f'''(\zeta)(x - a)(x - c)(x - b), \quad (\text{IV.2.11})$$

donde $\zeta \in [a, b]$. deduciendo la siguiente proposición.

Proposición IV.2.4.- Si f' no se anula en el intervalo I , y si $f(\xi_0) = 0$, $L(\xi) = 0$ para $\xi, \xi_0 \in [a, b]$, entonces se tiene

$$\xi - \xi_0 = \frac{1}{6} \frac{f'''(\zeta)}{f'(\zeta')}(\xi - a)(\xi - c)(\xi - b) \quad (\text{IV.2.12})$$

donde ζ y ζ' son números que pertenecen a $[a, b]$.

Demostración.- Por (IV.2.11), se tiene

$$f(\xi) = \frac{1}{6} f'''(\zeta)(\xi - a)(\xi - c)(\xi - b),$$

por el teorema del valor medio, se tiene

$$f(\xi) = f'(\zeta')(\xi - \xi_0),$$

con ζ' que pertenece al intervalo cuyas extremidades son ξ_0 y ξ . Combinando ambos resultados se obtiene el resultado de la proposición. \square

Corolario IV.2.5.- *Mismas hipótesis de la proposición precedente, y además $|f'(x)| \geq m > 0$, y $|f'''(x)| \leq M$ para todo $x \in [a, b]$, entonces*

$$|\xi - \xi_0| \leq \frac{M}{24m}(b-a)^3. \quad (\text{IV.2.13})$$

Si el error $|\xi - \xi_0|$ evaluado por la estimación (IV.2.13) no es lo suficientemente pequeño, se puede repetir el procedimiento: se calcula $f(\xi)$ y dependiendo de su signo, la raíz ξ_0 se encuentra en el intervalo $[a, \xi]$ o $[\xi, c]$, etc, al cual se aplica el mismo método obteniendo así un segundo valor aproximado ξ' . Teóricamente, se puede aplicar indefinidamente este procedimiento, y se muestra fácilmente que la sucesión de los números obtenidos converge hacia ξ_0 , ver ejercicios.

El método de la Falsa Posición es un claro ejemplo de un método iterativo, pues utiliza soluciones anteriormente obtenidas para calcular una nueva solución. En general, se dirá que un método iterativo es de orden k o de k pasos, si la iteración puede expresarse de la forma:

$$x_{n+1} = \Phi(x_n, x_{n-1}, \dots, x_{n-k+1}). \quad (\text{IV.2.14})$$

Por lo tanto, los métodos descritos anteriormente son métodos iterativos de dos pasos, pues se sirven de 2 valores para calcular el valor de la iteración siguiente.

Sistemas de Ecuaciones

Los problemas que han sido estudiados más arriba estaban relacionados a funciones de una sola variable con valores reales. Ahora bien, existe una gran variedad de problemas donde se deben resolver sistemas de ecuaciones, que en general no son lineales. La posición del problema es por consiguiente:

Dada $f : \mathcal{U} \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$, encontrar $\xi \in \mathcal{U}$ tal que

$$f(\xi) = 0. \quad (\text{IV.2.15})$$

Reiterando lo que se dijo al inicio de esta sección no se puede esperar de obtener una fórmula milagrosa para determinar ξ , lo que se debe buscar es por consiguiente una aproximación de esta solución con la precisión que se desee. La continuidad no es una condición suficiente para determinar la existencia de ceros, hipótesis primordial en el caso de una variable; en el caso de varias variables, en la mayor parte de los casos no se puede demostrar la existencia de ceros, contentándose con la unicidad de éstos a nivel local. Un teorema muy importante, es el de la inversión local que será enunciado sin demostración, para el lector interesado, puede encontrarla en cualquier libro de Análisis, por ejemplo en Rudin.

Teorema IV.2.6.- Sean $\mathcal{D} \subset \mathbb{R}^n$ abierto, $f : \mathcal{D} \longrightarrow \mathbb{R}^n$ continuamente diferenciable. Si:

- a) $f(x^*) = 0$,
- b) f'_{x^*} es inversible,

entonces, existe dos vecindarios, \mathcal{U} de x^* y \mathcal{V} de 0, tales que para todo $v \in \mathcal{V}$ existe un único $x \in \mathcal{U}$, tal que $f(x) = v$ y la aplicación

$$\begin{aligned} g : \mathcal{V} &\longrightarrow \mathcal{U} \\ v &\longrightarrow x(v) \end{aligned}$$

es continuamente diferenciable y además $g'_v = (f'_x)^{-1}$.

Consecuencias de este teorema son: la unicidad local de la solución; si \hat{x} es solución numérica obtenida mediante un metodo cualquiera se tiene la siguiente estimación del error

$$\hat{x} - x^* = (f'_{x^*})^{-1} f(\hat{x}) + \mathcal{O}(\|f(\hat{x})\|^2),$$

de donde si $(f'_{x^*})^{-1}$ es casi singular, $\hat{x} - x^*$ puede ser muy grande, aun si $f(\hat{x})$ es pequeño.

Un método Iterativo simple

Existe otra clase de ecuaciones que pueden ser escritas de la forma

$$f(x) = x, \quad (\text{IV.2.17})$$

donde f es una función de varias variables con las condiciones dadas al inicio de este paragrafo. Cabe remarcar que la ecuación (IV.2.15) puede ser expresada como $x = g(x)$ con $g(x) = x - Af(x)$, donde A es generalmente una matriz. La existencia y unicidad de las ecuaciones de la forma (IV.2.17) son resueltas utilizando el teorema del punto fijo en espacios métricos. Para saber más sobre espacios métricos ver Schawartz. A continuación, se enunciará el teorema del punto fijo.

Teorema IV.2.7.- Sean X un espacio métrico completo, $f : X \longrightarrow X$ una aplicación verificando

$$d(f(x), f(y)) \leq Cd(x, y), \quad C < 1;$$

donde d denota la distancia en X . Entonces la ecuación $f(x) = x$ admite una y una sola solución.

Demostración.- Sea $x_0 \in X$ un punto arbitrario, se define la sucesión $\{x_k\}$ de manera recursiva por

$$x_{k+1} = f(x_k),$$

esta sucesión es de Cauchy, en efecto, si $n \geq m$, se tiene:

$$\begin{aligned} d(x_n, x_m) &\leq d(x_n, x_{n-1}) + \cdots + d(x_{m+1}, x_m), \\ d(x_{m+1}, x_m) &\leq C^m d(x_1, x_0), \end{aligned}$$

por lo tanto

$$d(x_n, x_m) \leq \frac{C^m}{1-C},$$

que tiende a cero, cuando n, m tienden a ∞ . Sea $x^* = \lim_{n \rightarrow \infty} x_n$, se deduce que $f(x^*) = x^*$. Con esto se ha demostrado la existencia de la solución. Para la unicidad se considera x, y son soluciones del problema, por lo tanto

$$d(x, y) = d(f(x), f(y)) \leq C d(x, y),$$

y la única posibilidad que suceda esto es que $x = y$. \square

Se puede dar condiciones menos fuertes sobre el espacio X o sobre f , por ejemplo que f sea localmente una contracción, con solamente esta hipótesis se mantiene la unicidad, pero esta vez localmente, la existencia no está asegurada.

Por lo expuesto, se puede formular el método iterativo dado en la demostración del teorema precedente para el siguiente problema:

Sea $\Phi : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ continua, determinar $x \in \mathbb{R}^n$ tal que $\Phi(x) = x$. El método iterativo está dado por:

$$\begin{cases} x_0, & \text{arbitrario;} \\ x_n = \Phi(x_{n-1}). \end{cases} \quad (\text{IV.2.18})$$

Teorema IV.2.8.- Si $\{x_n\}$ converge hacia x^* , y Φ es continua en x^* , entonces x^* es solución de $x = \Phi(x)$

Demostración.- Se deja al lector. \square

El anterior teorema señala claramente que si la función Φ es continua, y la sucesión definida por (IV.2.18) es convergente, entonces el límite de esta sucesión es la solución del problema $\Phi(x) = x$. Por consiguiente, la pregunta natural que uno puede plantearse es cuando existen condiciones suficientes para tener convergencia. Definiendo el error e_n como

$$e_n = x_n - x^*,$$

donde x^* es la solución exacta, y x_n la n -sima iteración, se obtiene:

$$\begin{aligned} e_{n+1} &= x_{n+1} - x^* = \Phi(x_n) - \Phi(x^*) \\ &= \Phi'(x^*)(x_n - x^*) + \mathcal{O}(\|x_n - x^*\|^2), \\ e_{n+1} &\approx \Phi'(x^*)e_n. \end{aligned} \quad (\text{IV.2.19})$$

Si $\|e_{n+1}\| \leq q \|e_n\|$ para $q < 1$, entonces $e_n \rightarrow 0$ cuando $n \rightarrow \infty$. Por consiguiente la sucesión es convergente, si existe una norma tal que $\|\Phi'(x^*)\| = q < 1$.

Teorema IV.2.9.- a) Sea $\Phi(x) = Ax + b$, con A matriz, entonces el método iterativo (IV.2.18) converge para todo x , si y solamente si $\rho(A) < 1$, donde

$$\rho(A) = \max \{ |\lambda| \mid \lambda \text{ es un valor propio de } A \}.$$

b) Si $\Phi(x)$ es no lineal y dos veces continuamente diferenciable, entonces se tiene convergencia si:

$$\rho(\Phi'(x^*)) < 1,$$

$x_0 - x^*$ es suficientemente pequeño.

Demostración.- Se mostrará solamente el inciso a), dejando al lector el inciso b) con las observaciones hechas antes de enunciar el teorema.

\Rightarrow Se supone que el método converge $\forall x_0 \in \mathbb{R}^n$. Sea λ un valor propio de A , y e_0 un vector propio respecto a este valor propio, entonces

$$e_n = A^n e_0,$$

$$\lambda^n e_0 \longrightarrow 0,$$

posible solamente si $|\lambda| < 1$.

\Leftarrow Se supone que $\rho(A) < 1$, sea $e_0 \in \mathbb{R}^n$ arbitrario, de donde $e_n = A^n e_0$. Por el teorema de Jordan, existe una matriz T no singular tal que

$$T^{-1}AT = \begin{pmatrix} \lambda_1 & \ddots & & & & \\ & \ddots & 1 & & 0 & \\ & & \lambda_1 & & & \\ & & & \lambda_2 & \ddots & \\ 0 & & & & \ddots & 1 \\ & & & & & \lambda_2 & \\ & & & & & & \ddots \end{pmatrix}.$$

Sea $D = \text{diag}(1, \epsilon, \epsilon^2, \dots, \epsilon^{n-1})$, por consiguiente

$$D^{-1}T^{-1}ATD = \begin{pmatrix} \lambda_1 & \ddots & & & & \\ & \ddots & \epsilon & & 0 & \\ & & \lambda_1 & & & \\ & & & \lambda_2 & \ddots & \\ 0 & & & & \ddots & \epsilon \\ & & & & & \lambda_2 & \\ & & & & & & \ddots \end{pmatrix} = J,$$

si ϵ es suficientemente pequeño, se tiene $\|J\|_\infty < 1$, de donde el método es convergente. \square

Para el problema $f(x) = 0$, se vio antes que se puede convertir en problema de punto fijo, planteando

$$\Phi(x) = x - Af(x), \quad (\text{IV.2.20})$$

donde A es una matriz constante A , inversible. Ahora bien, para cumplir las condiciones del teorema precedente $\rho(\Phi'(x^*)) < 1$, motivo por el cual es suficiente $A \approx (f'(x^*))^{-1}$ para que $\Phi'(x^*) \approx 0$.

Ejemplo

Considérese la ecuación siguiente

$$\begin{aligned} x^2 + y^2 - 1 &= 0 \\ \sin x + \sin y &= \frac{1}{2}, \end{aligned}$$

se desea determinar las soluciones de esta ecuación. Convirtiendo en un problema de punto fijo se tiene

$$\Phi(x, y) = \begin{pmatrix} x \\ y \end{pmatrix} - A \begin{pmatrix} x^2 + y^2 - 1 \\ \sin x + \sin y - \frac{1}{2} \end{pmatrix}$$

donde la A es una matriz de 2×2 . La derivada de f en el punto (x, y) , está dada por

$$f'_{(x,y)} = \begin{pmatrix} 2x & 2y \\ \cos x & \cos y \end{pmatrix},$$

la inversa de la derivada de f , esta dada por

$$(f'_{(x,y)})^{-1} = \frac{1}{2(x \cos y - y \cos x)} \begin{pmatrix} \cos y & -2y \\ -\cos x & 2x \end{pmatrix}.$$

Recorriendo a travez de la circunferencia de radio 1 dada por la primera ecuación se escogen los valores de x, y para los cuales remplazando en la segunda ecuación se tiene valores muy cerca a $1/2$. Una vez determinados estos valores se consigue una aproximación de la matriz A para poder aplicar el método iterativo. Las soluciones obtenidas con una precisión del orden de 10^{-10} son:

$$\begin{aligned} x &= 0.3176821764792814, & y &= -0.948197255188055; \\ x &= -0.948197255188055, & y &= 0.3176821764792814. \end{aligned}$$

Ejercicios

1.- Sea A una matriz que es diagonal dominante, es decir

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \quad i = 1, \dots, n;$$

entonces el método de Jacobi, formulado en el capítulo II, para resolver $Ax = b$ es convergente.

2.- Considérese la ecuación integral

$$y(x) = f(x) + \int_a^b K(x, s, y(s)) ds, \quad (*)$$

donde a, b, f, K están dados, se busca $y(x)$. Mostrar, si el nucleo K es degenerado, es decir

$$K(x, s, y) = \sum_{i=1}^n a_i(x) b_i(s, y),$$

entonces la solución de $(*)$ está dada por

$$y(x) = f(x) + \sum_{i=1}^n c_i a_i(x)$$

donde las constantes c_1, \dots, c_n satisfacen

$$c_i = \int_a^b b_i(s, f(s) + \sum_{j=1}^n c_j a_j(s)) ds \quad i = 1, \dots, n.$$

3.- Calcular la solución de

$$y(x) = 1 + \lambda \int_0^\pi (2 \sin x \sin s + \sin 2x \sin 2s) e^{y(s)} ds, \quad \lambda = \frac{1}{10}.$$

Resolver el sistema no lineal para c_1, c_2 con el método iterativo.

4.- Sean $J = [a, b]$ un intervalo de \mathbb{R} en el cual la función dos veces continuamente diferenciable f verifica las relaciones $|f'(x)| \geq m$, $|f''(x)| \leq M$, $f(a)f(b) < 0$. Mostrar que si $\frac{M}{4m}(b-a) = q < 1$ se puede, por n aplicaciones sucesivas del método de la falsa posición, encontrar un intervalo de extremidades a_n, b_n conteniendo la única raíz de $f(x) = 0$ en $[a, b]$, con

$$|b_n - a_n| \leq \frac{4m}{M} q^{2n}.$$

IV.3 Método de Newton

En lo que va este capítulo, se ha visto dos enfoques para resolver sistemas de ecuaciones, el primero mediante una fórmula que de el resultado de manera explícita en función de los datos del problema. El segundo enfoque consiste en utilizar métodos iterativos que permitan aproximar a la solución de una manera arbitraria. En la sección precedente se vio en los diferentes teoremas de convergencia que la velocidad de convergencia es lineal, lo que puede constituir una desventaja, desde el momento en que se requiera resolver grandes y muchos sistemas de ecuaciones. Lo deseable sería cambiar el orden de convergencia, por ejemplo a una reducción cuadrática del error.

Recordando el método de la Falsa Pendiente, en lugar de tomar una secante para determinar una aproximación del cero de la función estudiada, se podría tomar la tangente en un punto de la curva inducida por la función. El método iterativo formulado en la sección precedente estaba basado en el teorema del punto fijo y el problema a resolver era de la forma $\Phi(x) = x$, para el caso de las ecuaciones de la forma $f(x) = 0$, era suficiente considerar la función $\Phi(x) = x - Af(x)$, donde A es una matriz constante, ahora bien suponiendo que $f'(x)$ es inversible en un vecindario de una de las soluciones de la ecuación, en lugar de tomar A constante, se puede plantear $A = (f'(x))^{-1}$.

Las motivaciones están dadas para formular un nuevo método, cuya velocidad de convergencia sea superior a los métodos anteriormente formulados. Para tal efecto, considerése la función $f : \mathcal{U} \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$, \mathcal{U} un abierto de \mathbb{R}^n , se supone además que f es continuamente derivable, y la derivada en todo punto es inversible. Por consiguiente, se tiene el problema

$$f(x) = 0, \quad (\text{IV.3.1})$$

cuyas soluciones, si éstas existen son localmente únicas, por el teorema de las funciones inversas, dado en la sección precedente. Supóngase que este problema tiene al menos una solución, denotandola por x^* . Sea $x \in \mathcal{U}$ bastante próximo de x^* , aplicando la definición de la derivada en el punto x se obtiene

$$f(x^*) = f(x) + f'(x)(x^* - x) + \mathcal{O}(\|x^* - x\|^2),$$

de donde despreciando el término que contiene \mathcal{O} se obtiene la ecuación lineal dada por

$$f'(x)(x^* - x) = -f(x),$$

que por hipótesis tiene solución y es única. De donde, el método de Newton tiene la formulación siguiente, sea x_0 un punto bastante próximo de la

solución buscada x^* , x_m se define recursivamente, por las ecuaciones

$$f'(x_{m-1})(x_m - x_{m-1}) = -f(x_{m-1}). \quad (\text{IV.3.2})$$

Ejemplos

1.- Considérese la ecuación de una sola variable dada por

$$f(x) = 2x - \tan x,$$

para aplicar el método de Newton la derivada está dada por

$$f'(x) = 2 - \frac{1}{\cos^2 x},$$

partiendo del punto inicial $x_0 = 1,2$ se obtiene los siguientes valores mediante el método de Newton.

Tabla IV.3.1. Valores obtenidos por el método de Newton.

k	x_k	$f(x_k)$	e_k^2/e_{k-1}
0	1,2	-0.172152	
1	1.16934	-1.69993×10^{-2}	
2	1.16561	-0.213431×10^{-3}	-3.97664
3	1.16556	$-0.347355 \times 10^{-07}$	-3.44610
4	1.16556	$-1.133227 \times 10^{-14}$	-3.38337

2.- Considérese el sistema de ecuaciones dado por

$$\begin{cases} x^2 + y^2 - 4 = 0 \\ xy - 4 = 0 \end{cases}.$$

Las soluciones de este sistema de ecuaciones están dadas por la intersección de una circunferencia de radio 2 y centro en el origen y una hipérbola inclinada. Realizando un gráfico se puede observar que una de las raíces está próxima al punto (0.5, 2) que será tomado como punto inicial. La matriz derivada es igual a

$$f'(x, y) = \begin{pmatrix} 2x & 2y \\ y & x \end{pmatrix},$$

utilizando el método de eliminación de Gauss para calcular los valores de (x_k, y_k) se obtiene la siguiente tabla con las primeras 23 iteraciones del método de Newton.

Tabla IV.3.2. Valores obtenidos por el método de Newton.

k	x_k	y_k	$\ f(x_k)\ _2$	$\ e_k\ _2^2 / \ e_{k-1}\ _2$
0	0.5	2.	0.25	
1	.516666	1.966666	0.134722	
2	0.526332	1.94921	0.764322×10^{-1}	14.3688
3	0.532042	1.93948	0.446506×10^{-1}	28.3211
22	0.540690	1.92553	0.345530×10^{-5}	446885.
23	0.540690	1.92553	0.210661×10^{-5}	732995.

En las tablas IV.3.1 y IV.3.2, se observa que el método de Newton converge cuadráticamente. Es importante poder confirmar teóricamente este resultado. Analizando el caso de una variable, se supone que $f : I \rightarrow \mathbb{R}$ es dos veces continuamente derivable, x^* una solución del problema $f(x) = 0$, además $f'(x^*) \neq 0$. Sea x_k el valor de la k -ésima iteración obtenida del método de Newton. El desarrollo de Taylor en x_k de la función f y el método de Newton para obtener x_{k+1} están dados por:

$$0 = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2}f''(x_k) \underbrace{(x^* - x_k)^2}_{e_k^2} + \mathcal{O}((x^* - x_k)^3),$$

$$0 = f(x_k) + f'(x_k)(x_{k+1} - x_k),$$

sustrayendo ambas cantidades se obtiene

$$0 = f'(x_k) \underbrace{(x_{k+1} - x^*)}_{e_{k+1}^2} - \frac{1}{2}f''(x_k)e_k^2 + \mathcal{O}((e_k)^3),$$

de donde

$$e_{k+1} = \frac{1}{2} \frac{f''(x_k)}{f'(x_k)} e_k^2 + \mathcal{O}((e_k)^3),$$

mostrando así el teorema siguiente:

Teorema IV.3.1.- Sea $f : I \rightarrow \mathbb{R}$, tres veces continuamente diferenciable, x^* una solución de la ecuación $f(x) = 0$. Si $f'(x) \neq 0$ en un vecindario de x^* , x_0 bastante próximo de x^* , entonces

$$e_{k+1} = \frac{1}{2} \frac{f''(x_k)}{f'(x_k)} e_k^2 + \mathcal{O}((e_k)^3), \quad (\text{IV.3.3})$$

donde $e_k = x^* - x_k$.

Para el caso de una función de varias variables, la situación es bastante similar, en efecto sea

$$f : \mathcal{U} \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n,$$

donde \mathcal{U} es un abierto de \mathbb{R}^n , f es una función tres veces continuamente derivable, cuya derivada es inversible para todo $x \in \mathcal{U}$. La derivada de f en el punto $x \in \mathcal{U}$ es una aplicación lineal, cuya matriz respecto a la base canónica está dada por

$$f'(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix},$$

donde f_1, \dots, f_n son las componentes de la función f ; la segunda derivada de f en el punto $x \in \mathcal{U}$ es una aplicación bilineal simétrica, que respecto a la base canónica, está dada por

$$f''(x)(h, k) = \sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) h_i k_j,$$

donde $h, k \in \mathbb{R}^n$ y los h_i, k_j son las componentes de h y k respecto a las bases naturales. Para más detalle ver Cartan. El desarrollo de Taylor en x está dado por

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)(h, h) + \mathcal{O}(\|h\|^3).$$

Teorema IV.3.2.- Sean $f : \mathcal{U} \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ tres veces diferenciable, con derivada inversible y $x^* \in \mathcal{U}$ con $f(x^*) = 0$. Si $\{x_k\}$ es la sucesión definida por el método de Newton, entonces el error $e_k = x_k - x^*$ satisface

$$e_{k+1} = \frac{1}{2}(f'(x_k))^{-1}f''(x_k)(e_k, e_k) + \mathcal{O}(\|h\|^3). \quad (\text{IV.3.4})$$

Demostración.- Similar a la del caso de una sola variable. □

Cálculo de la Derivada

Al implementar el método de Newton, se debe calcular la derivada de f en cada punto. La primera forma de determinar la matriz $f'(x_k)$ es de manera analítica, construyendo así una subrutina para evaluar los coeficientes de la matriz jacobiana en cada iteración. Pero lastimosamente, no siempre

es posible calcular las derivadas parciales de manera analítica por varias razones: funciones muy complicadas para derivar, cálculos muy largos, etc. Otra manera de evaluar los coeficientes de $f'(x_k)$ consiste en hacerlo numéricamente. Utilizando un esquema de primer orden para evaluar la derivada parcial de f_j respecto a x_i , se tiene

$$\frac{\partial f_j}{\partial x_i}(x) = \lim_{t \rightarrow 0} \frac{f_j(x + te_i) - f_j(x)}{t},$$

donde $t \in \mathbb{R}$, e_i es el i -ésimo vector de la base canónica. Ahora bien, se plantea $g(t) = f_j(x + te_i)$ dejando fijo x , se tiene

$$\frac{\partial f_j}{\partial x_i}(x) = g'(0).$$

De donde, el problema consiste en calcular $g'(0)$; con un esquema de primer orden se obtiene

$$g'(0) \approx \frac{g(t) - g(0)}{t}. \quad (\text{IV.3.5})$$

Cabe recalcar, que se puede aproximar $g'(0)$ con una mejor aproximación, para eso se puede utilizar los polinomios de interpolación. Una pregunta natural surge, ¿cuál t escoger?, para obtener la mejor aproximación de $g'(0)$. El error de aproximación está dado por

$$\frac{g(t) - g(0)}{t} = g'(0) + \frac{1}{2}g''(0)t + \mathcal{O}(t^2),$$

mientras que los errores de redondeo, suponiendo que g es una función bien condicionada, está dado por los siguientes resultados, suponiendo que $g'(t) \approx g'(0)$:

$$\begin{aligned} & \frac{g(t(1 + \epsilon_1))(1 + \epsilon_2) - g(0)(1 + \epsilon_3)}{t} - \frac{g(t) - g(0)}{t} \\ & \approx \frac{1}{t} \left\{ g(t) + g'(0)\epsilon_1 t(1 + \epsilon_2) - g(0)(1 + \epsilon_3) - g(t) + g(0) \right\} \\ & \approx \frac{1}{t} \left\{ \epsilon_2 g(t) + g'(0)\epsilon_1 t - \epsilon_3 g(0) \right\}, \end{aligned}$$

donde $|\epsilon_i| \leq \text{eps}$, por consiguiente

$$\text{error de redondeo} \leq \frac{1}{|t|} \{ 2|g(0)| + |g'(0)t| \} \text{eps}. \quad (\text{IV.3.6})$$

Lo ideal será, por lo tanto escoger un valor de t de manera que los errores de aproximación y de redondeo se aproximen, ver figura IV.3.1.

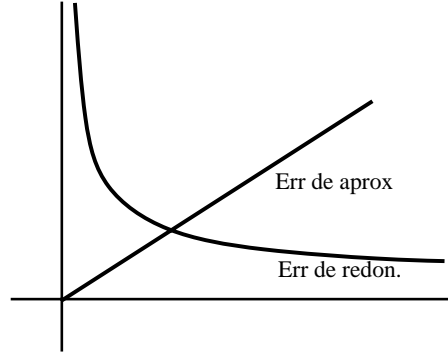


Figura IV.3.1. Error de Aproximación *vs* Error de Redondeo.

Observando la figura, se tiene que el error de aproximación es del orden de t , mientras que el error de redondeo es proporcional a eps/t , de donde equilibrando ambos errores se obtiene

$$t = \sqrt{C \text{ eps}}, \quad (\text{IV.3.7})$$

donde C es una constante elegida de acuerdo a la función f .

El Teorema de Newton-Misovski

Se ha formulado el método de Newton de una manera general, dando inclusive una estimación del error cometido, deduciendo que si el método converge, la convergencia es cuadrática. Pero sin embargo, no se enunció las condiciones suficientes para obtener convergencia de este método. El siguiente teorema dará las condiciones suficientes para asegurar la convergencia del método, cuando la función f cumple ciertas condiciones.

Teorema IV.3.3.- Newton-Misovski. Sean $\mathcal{D} \subset \mathbb{R}^n$ abierto y convexo, $f : \mathcal{D} \rightarrow \mathbb{R}^n$ continuamente diferenciable, $f'(x)$ inversible para todo $x \in \mathcal{D}$, $x_0 \in \mathcal{D}$ satisfaciendo las siguientes condiciones:

- a) $\|\Delta x_0\| \leq \alpha$,
- b) $\left\| (f'(y))^{-1} \left(f'(x + t(x - y)) - f'(x) \right) (y - x) \right\| \leq \omega t \|y - x\|^2, \forall x, y \in \mathcal{D}, t \in [0, 1];$
- c) $\eta := \frac{1}{2} \alpha \omega < 1;$
- d) $\rho := \frac{\alpha}{1 - \eta} < 1;$
- e) $\bar{B}(x_0, \rho) = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq \rho\}.$

Entonces:

- i) La sucesión $\{x_k\}$ definida por el método de Newton se queda dentro $B(x_0, \rho)$ y converge hacia una solución x^* de $f(x) = 0$.
- ii) Se tiene la estimación $\|\Delta x_k\| \leq \frac{\omega}{2} \|\Delta x_{k-1}\|^2$.
- iii) $\|x_k - x^*\| \leq \frac{\omega}{2(1 - \eta^{2^k})} \|\Delta x_{k-1}\|^2$.

Demostración.- Es claro por la definición del método de Newton que si $\{x_k\}$ es convergente, entonces el límite de la sucesión es igual a x^* , donde $f(x^*) = 0$. Se mostrarás la conclusiones por inducción. Se supone que $x_k, x_{k-1} \in \mathcal{D}$, para $k \geq 1$, entonces

$$\|\Delta x_k\| \leq \frac{\omega}{2} \|\Delta x_{k-1}\|^2,$$

en efecto

$$\begin{aligned} \|\Delta x_k\| &= \|(f'(x_k))^{-1} f(x_k)\| \\ &= \|(f'(x_k))^{-1} [f(x_k) - f(x_{k-1}) - f'(x_{k-1})\Delta x_{k-1}]\| \\ &= \left\| f'(x_k) \int_0^1 (f'(x_{k-1} + t\Delta x_{k-1}) - f'(x_{k-1})) \Delta x_{k-1} dt \right\| \\ &\leq \int_0^1 \|f'(x_k) (f'(x_{k-1} + t\Delta x_{k-1}) - f'(x_{k-1})) \Delta x_{k-1}\| dt \\ &\leq \int_0^1 \omega t \|\Delta x_{k-1}\|^2 dt = \frac{\omega}{2} \|\Delta x_{k-1}\|^2. \end{aligned}$$

Se define η_k recursivamente por

$$\eta_k = \eta_{k-1}^2, \quad \eta_0 = \frac{\omega}{2} \alpha = \eta.$$

Si $x_0, x_1, \dots, x_k \in \mathcal{D}$, entonces

$$\frac{\omega}{2} \|\Delta x_k\| \leq \eta_k;$$

es cierto para $k = 0$, supóngase cierto para $k - 1$, por consiguiente

$$\frac{\omega}{2} \|\Delta x_k\| \leq \left(\frac{\omega}{2} \|\Delta x_{k-1}\| \right)^2 \leq \eta_{k-1}^2 = \eta_k.$$

como $\eta_0 = \eta$, se tiene $\eta_k = \eta^{2^k}$, puesto que $\eta < 1$,

$$\lim_{k \rightarrow \infty} \eta_k = 0.$$

El siguiente paso es mostrar que $\{x_k\}$ es una sucesión de Cauchy y que satisface el punto i). Se supone nuevamente, que $x_0, \dots, x_k \in \mathcal{D}$, $0 \leq l \leq k$, obteniendo

$$\begin{aligned} \|x_{k+1} - x_l\| &\leq \underbrace{\|x_{k+1} - x_k\|}_{\frac{2}{\omega}\eta_k} + \dots + \underbrace{\|x_{k+1} - x_l\|}_{\frac{2}{\omega}\eta_l} \\ &\leq \frac{2}{\omega}(1 + \eta_l^2 + \eta_l^4 + \dots) \\ &\leq \frac{2}{\omega} \frac{\eta_l^2}{1 - \eta_l^2}. \end{aligned}$$

Para $l = 0$, se tiene $\|x_{k+1} - x_0\| \leq \frac{2}{\omega} \frac{\eta^2}{1 - \eta^2} < \frac{2}{\omega} \frac{\eta}{1 - \eta} = \frac{\alpha}{1 - \alpha} = \rho$, de donde $x_{k+1} \in B(x_0, \rho)$.

Por otro lado,

$$\|x_{k+1} - x_l\| \leq \frac{2}{\omega} \frac{\eta_l^2}{1 - \eta_l^2} \longrightarrow 0, \quad \text{cuando } k \geq l \rightarrow \infty.$$

de donde la sucesión $\{x_k\}$ es una sucesión de Cauchy, y por lo tanto convergente. Es así que se ha mostrado el punto i) y el punto ii) del teorema quedando pendiente el último punto. Se tiene

$$\|x_{l+1} - x_k\| \leq \|x_{l+1} - x_l\| + \dots + \|x_{k+1} - x_k\|,$$

se plantea:

$$\hat{\eta}_{k-1} = \frac{\omega}{2} \|\Delta x_{k-1}\|, \quad \hat{\eta}_l = \hat{\eta}_{l-1}^2;$$

de donde $\hat{\eta}_{k-1} \leq \eta_{k-1}$ y $\hat{\eta}_k \leq \eta_k - \eta^{2^k}$, además

$$\frac{\omega}{2} \|\Delta x_l\| \leq \hat{\eta}_l \quad \text{para } l \geq l-1.$$

De donde

$$\|x_{l+1} - x_k\| \leq \frac{\omega}{2} \frac{\|\Delta x_{k-1}\|^2}{1 - \eta^{2^k}},$$

haciendo tender l al infinito queda demostrado el punto iii) □

Es necesario remarcar el siguiente hecho concerniente a la hipótesis b) del teorema que se acaba de demostrar. Si f es 3 veces derivable, utilizando la fórmula de Taylor se tiene

$$[f'(x + t(x - y)) - f'(x)](y - x) = f''(x)(t(y - x), y - x) + \mathcal{O}(\|y - x\|^3),$$

por consiguiente

$$(f'(y))^{-1}[f'(x+t(x-y))-f'(x)](y-x) = t(f'(y))^{-1}f''(x)(y-x, y-x) + \mathcal{O}(\|y-x\|^3),$$

si \mathcal{D} es bastante pequeño, entonces se puede despreciar $\mathcal{O}(\|y-x\|^3)$, de donde

$$\|(f'(y))^{-1}[f'(x+t(x-y))-f'(x)](y-x)\| \leq t \|(f'(y))^{-1}f''(x)\| \|y-x\|^2,$$

por consiguiente

$$\omega \approx \sup_{x, y \in \mathcal{D}} \|(f'(y))^{-1}f''(x)\| \quad (\text{IV.3.8})$$

Para comprender el teorema, puede ser útil estudiar en un ejemplo tipo, las hipótesis del teorema y deducir las conclusiones que conducen estas hipótesis

Ejemplo

Sea, $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ definida por

$$f(x) = \begin{pmatrix} x_1^2 + x_2^2 - 1 \\ x_2 - x_1^2 \end{pmatrix},$$

la solución del problema $f(x) = 0$ está dada por la intersección de una circunferencia de centro en el origen y radio 1 y una parábola cuyo vértice se encuentra también en el origen. Este problema puede ser resuelto gráficamente, substituyendo x_2 , pero el propósito del ejemplo es estudiar el teorema de Misovski.

La derivada de f' está dada por la matriz jacobiana

$$f'(x) = \begin{pmatrix} 2x_1 & 2x_2 \\ -2x_1 & 1 \end{pmatrix},$$

cuya inversa es igual a

$$(f'(x))^{-1} = \frac{1}{2x_1(1+2x_2)} \begin{pmatrix} 1 & -2x_2 \\ 2x_1 & 2x_1 \end{pmatrix}.$$

Observando una gráfica del problema se escoge como dominio \mathcal{D} a

$$\mathcal{D} = \left\{ (x_1, x_2) \mid \frac{1}{2} < x_1, x_2 < 2 \right\}.$$

Analizando las condiciones del teorema se tiene:

- a) Se escoge como valor inicial $(1, 1) \in \mathcal{D}$, obteniendo $\Delta x_0 = (-1/6, -1/3)$, de donde

$$\|\Delta x_0\| = \frac{\sqrt{5}}{6} \approx 0,37 = \alpha.$$

- b) El estudio de la segunda condición da:

$$\begin{aligned} & (f'(y))^{-1} \left(f'(x + t(x - y)) - f'(x) \right) (y - x) \\ &= \frac{1}{2y_1(1 + 2y_2)} \begin{pmatrix} 1 & -2y_2 \\ 2y_1 & 2y_1 \end{pmatrix} \begin{pmatrix} 2t(y_1 - x_1) & 2t(y_2 - x_2) \\ -2t(y_1 - x_1) & 0 \end{pmatrix} \begin{pmatrix} y_1 - x_1 \\ y_2 - x_2 \end{pmatrix} \\ &= \frac{t}{2y_1(1 + 2y_2)} \begin{pmatrix} (1 + y_2)(y_1 - x_1)^2 + (y_2 - x_2)^2 \\ 2y_1(y_2 - x_2)^2 \end{pmatrix}. \end{aligned}$$

Se observa inmediatamente que las componentes son positivas, mayorando respecto a y , se obtiene

$$\begin{aligned} & \left| (f'(y))^{-1} \left(f'(x + t(x - y)) - f'(x) \right) (y - x) \right| \\ & \leq \frac{t}{2} \left| \frac{3(y_1 - x_1)^2 + (y_2 - x_2)^2}{4(y_2 - x_2)^2} \right|, \end{aligned}$$

pasando a la norma de la convergencia uniforme, se obtiene

$$\left\| (f'(y))^{-1} \left(f'(x + t(x - y)) - f'(x) \right) (y - x) \right\|_{\infty} \leq 2t \|y - x\| \inf y^2,$$

de donde:

- b) $\omega = 2$.

- c) $\eta = 0,37 < 1$.

- d) $\rho = \frac{0,37}{1 - 0,37} \approx \frac{1}{2}$. Ahora bien, $\bar{B}(x_0, \rho) \not\subset \mathcal{D}$, de donde no se puede garantizar las conclusiones del teorema. Sin embargo, tomando como x_0 el valor de la primera iteración a partir del valor inicial del ejemplo se tiene

$$x_0 = \left(\frac{5}{6}, \frac{2}{3} \right)^t,$$

obteniendo

$$\Delta x_0 = \begin{pmatrix} -19/213 \\ -1/21 \end{pmatrix}, \quad \|\Delta x_0\| = 0,066 = \alpha,$$

de donde

$$\eta = 0,066, \quad \rho = 0,07, \quad \bar{B}(x_0, \rho) \subset \mathcal{D}.$$

Por lo tanto, las tres condiciones del teorema son verificadas, teniendo garantizada de esta manera la convergencia del método de Newton.

El teorema de Newton-Misovski da las ventajas de utilizar el método de Newton para resolver el problema $f(x) = 0$, sin embargo:

Desventajas del método de Newton

- Cada iteración del método es costosa en operaciones. En efecto el cálculo de $f'(x)$ y la resolución del sistema lineal adyacente cuestan mucho si la dimensión del sistema es grande. Resolver el sistema LR cuesta aproximadamente $n^3/3$ operaciones.
- Convergencia local. El método converge solamente si x_0 está suficientemente próximo de la solución x^* .

Método de Newton Simplificado

Una de las mayores desventajas que tiene el método de Newton, consiste en el hecho de calcular en cada iteración la matriz derivada y el sistema lineal asociado. Se puede simplificar el método de Newton de la siguiente manera:

$$\begin{aligned} x_0 & \text{ arbitrario,} \\ \Delta x_k &= -\left(f'(x_0)\right)^{-1} f(x_k), \\ x_{k+1} &= x_k + \Delta x_k. \end{aligned} \tag{IV.3.9}$$

Por consiguiente se debe calcular una sola vez $f'(x_0)$ y calcular una vez la descomposición LR . Por lo tanto, la primera iteración consume aproximadamente $n^3/3$ en el cálculo de la descomposición, y las demás iteraciones necesitan aproximadamente n^2 operaciones. La desventaja de utilizar el método de Newton simplificado reside en el hecho en que se pierde la convergencia cuadrática, obteniendo una convergencia lineal. En efecto, considerando este método como un método iterativo simple, dado en la sección precedente se tiene:

$$\begin{aligned} x_{k+1} &= \Phi(x_k), \quad \text{donde } \beta\Phi(x) = x - \left(f'(x_0)\right)^{-1} f(x); \\ \Phi'(x) &= I - \left(f'(x_0)\right)^{-1} f'(x). \end{aligned}$$

Ahora bien, se tiene convergencia local, si y solamente si

$$\rho\left(I - \left(f'(x_0)\right)^{-1} f'(x)\right) < 1,$$

por el teorema IV.2.9 y la convergencia lineal es consecuencia directa del teorema citado.

Tanto el método de Newton, como su versión simplificada, definen iteraciones de la forma

$$\begin{aligned} \Delta x_k &= -\left(M(x_k)\right)^{-1} f(x_k), \\ x_{k+1} &= x_k + \Delta x_k, \end{aligned} \tag{IV.3.10}$$

donde $M(x)$, es una matriz que depende de x . Para el método de Newton se tiene $M(x) = f'(x)$ y para la versión simplificada $M(x) = f'(x_0)$. En

el primer caso la matriz M es la derivada, en el segundo caso M es la derivada en un punto arbitrario. Existen muchos problemas, en los cuales la matriz derivada tiene ciertas particularidades que constituyen en ventajas y desventajas al mismo tiempo, que puede solucionarse convenientemente si se escoge una matriz M apropiada; por ejemplo $f'(x)$ puede tener una estructura de matriz banda, con algunos elementos no nulos fuera de la banda, es decir

$$f'(x) = \begin{pmatrix} * & * & & * \\ * & * & * & \\ & \ddots & & \ddots \\ * & & & * & * \end{pmatrix},$$

planteando $M(x)$ la matriz cuyos elementos están dados por la banda principal de $f'(x)$ se tiene una buena aproximación de $f'(x)$, disminuyendo de esta manera el costo en operaciones. Las bases teóricas de utilizar la matriz M en lugar de $f'(x)$, están dadas por el siguiente teorema.

Teorema IV.3.4.- *Newton-Kantorovich.* Sean $\mathcal{D} \subset \mathbb{R}^n$ abierto y convexo, $f : \mathcal{D} \rightarrow \mathbb{R}^n$ continuamente diferenciables, $M(x_0)$ inversible y $x_0 \in \mathcal{D}$. Además:

- a) $\|M(x_0)^{-1}f(x_0)\| \leq \alpha$;
- b) $\|M(x_0)^{-1}(f'(y) - f'(x))\| \leq \omega \|y - x\|$, $\forall x, y \in \mathcal{D}$;
- c) $\|M(x_0)^{-1}(f'(x) - M(x))\| \leq \delta_0 + \delta - 1 \|x - x_0\|$,
- d) $\|M(x_0)^{-1}(M(x) - M(x_0))\| \leq \mu \|x - x_0\|$;
- d) $\delta_0 < 1$, $\sigma := \max(\omega, \mu + \delta_1)$ y

$$h = \frac{\alpha\sigma}{(1 - \delta_0)^2} \leq \frac{1}{2};$$

e) $B(x_0, \rho) \subset \mathcal{D}$, con

$$\rho = \frac{1 - \sqrt{1 - 2h}}{h} \frac{\alpha}{1 - \delta_0};$$

entonces:

- i) $M(x)$ es inversible para $x \in B(x_0, \rho)$,
- ii) La sucesión $\{x_k\}$ definida por (IV.3.10) se queda en $B(x_0, \rho)$ y converge hacia una solución x^* de $f(x) = 0$,
- iii) Se define $\bar{h} = \frac{\alpha\omega}{(1 - \delta_0)^2} = \frac{\omega}{\sigma}h$,

$$\rho_{\pm} = \frac{1 \pm \sqrt{1 - 2\bar{h}}}{\bar{h}} \frac{\alpha}{1 - \delta_0},$$

entonces $x^* \in B(x_0, \rho_-)$ y no hay otra solución de $f(x) = 0$ en $B(x_0, \rho_+) \cap \mathcal{D}$.

Antes de demostrar el teorema hay que remarcar que $\bar{h} \leq h$ cuya verificación es inmediata y también $\rho_- \leq \rho$; en efecto considerando los desarrollos en serie de Taylor se tiene:

$$\sqrt{1-x} = \sum_{j \geq 0} \binom{1/2}{j} (-x)^j = 1 - \frac{x}{2} - c_2 x^2 - c_3 x^3 - \dots,$$

con los $c_j \geq 0$,

$$\frac{1 - \sqrt{1-2h}}{h} = 1 + 4c_2 h + 8c_2 h^2 + 16c_4 h^3 + \dots,$$

de donde $\rho_- \leq \rho$. Ver la figura IV.3.2.

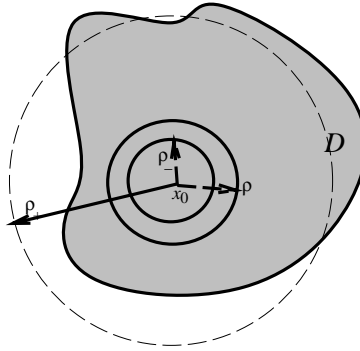


Figura IV.3.2. Resultado del Teorema de Newton-Kantorovich.

Demostración.- Demostrando el punto i), se tiene:

$$M(x) = M(x_0) \left[I + \underbrace{M(x_0)^{-1} (M(x) - M(x_0))}_B \right],$$

$$\|B\| \leq \mu \|x - x_0\| < \mu \rho,$$

$$\mu \rho = \mu \frac{1 - \sqrt{1-2h}}{h\alpha\omega\sigma} \frac{\alpha}{1 - \delta_0} (1 - \delta_0) 62 \leq \frac{\mu}{\sigma} (1 - \delta_0) \leq 1,$$

de donde $\|B\| < 1$, por lo tanto $(I + B)$ es inversible y su inversa está dada por

$$(I + B)^{-1} = \sum_{k=0}^{\infty} (-1)^k B^k,$$

la norma de $(I + B)^{-1}$, está mayorada por

$$\|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|} \leq \frac{1}{1 - \mu \|x - x_0\|},$$

por consiguiente

$$\|M(x)^{-1}M(x_0)\| \leq \frac{1}{1 - \mu \|x - x_0\|} \|x - x_0\| < \rho. \quad (\text{IV.3.11})$$

La demostración del punto ii) se efectúa por inducción. Si $x_{k-1}, x_k \in B(x_0, \rho)$, entonces x_{k+1} existe, se tiene:

$$\begin{aligned} \|x_{k+1} - x_k\| &= \|M(x_k)^{-1}f(x_k)\| \\ &\leq \|M(x_k)^{-1}[f(x_k) - f(x_{k-1}) - f'(x_{k-1})(x_k - x_{k-1})]\| \\ &\quad + \|M(x_k)^{-1}[f'(x_{k-1}) - M(x_{k-1})](x_k - x_{k-1})\| \\ &\leq \|M(x_k)^{-1}[f(x_k) - f(x_{k-1}) - f'(x_{k-1})(x_k - x_{k-1})]\| \\ &\quad + \frac{1}{1 - \mu \|x_k - x_0\|} (\delta_0 + \delta_1 \|x_{k-1} - x_0\|) \|x_k - x_0\| \\ &\leq \left\| M(x_k)^{-1} \int_0^1 [f'(x_{k-1} + t(x_k - x_{k-1})) - f'(x_{k-1})] dt (x_k - x_{k-1}) \right\| \\ &\quad + \frac{1}{1 - \mu \|x_k - x_0\|} (\delta_0 + \delta_1 \|x_{k-1} - x_0\|) \|x_k - x_0\| \\ &\leq \frac{1}{1 - \mu \|x_k - x_0\|} \frac{\omega}{2} \|x_k - x_{k-1}\|^2 \\ &\quad + \frac{1}{1 - \mu \|x_k - x_0\|} (\delta_0 + \delta_1 \|x_{k-1} - x_0\|) \|x_k - x_0\|. \end{aligned}$$

De donde

$$\|x_{k+1} - x_k\| \leq \frac{1}{1 - \mu \|x_k - x_0\|} \left\{ \frac{\sigma}{2} \|x_k - x_{k-1}\|^2 + (\delta_0 + (\sigma - \mu \|x_{k-1} - x_0\|) \|x_k - x_{k-1}\|) \right\}.$$

Para resolver esta desigualdad, se reemplaza $\|x_{k+1} - x_k\|$ por $t_{k+1} - t_k$ y $\|x_k - x_0\|$ por t_k , el símbolo de desigualdad por el de igualdad, definiendo así una sucesión $\{t_k\} \subset \mathbb{R}$, que verifica:

$$\begin{aligned} t_0 &= 0, \\ t_1 &= \alpha, \\ t_{k+1} - t_k &= \frac{1}{1 - \mu t_k} \{ \sigma (t_k - t_{k-1})^2 + (\delta_0 + (\sigma - \mu) t_{k-1}) (t_k - t_{k-1}) \}, \end{aligned}$$

Se demuestra facilmente, por inducción que:

$$\begin{aligned}\|x_{k+1} - x_k\| &\leq t_{k+1} - t_k, \\ \|x_k - x_0\| &\leq t_k;\end{aligned}$$

en efecto, para $k = 0$ se cumple, supóngase cierto para $k - 1$, por consiguiente

$$\begin{aligned}\|x_k - x_0\| &\leq \|x_k - x_{k-1}\| + \cdots + \|x_1 - x_0\| \\ &\leq t_k - t_{k-1} + \cdots + t_0 = t_k; \\ \|x_{k+1} - x_k\| &\leq \frac{1}{1 - \mu t_k} \left\{ \frac{\sigma}{2} (t_k - t_{k-1})^2 + \delta_0 + (\sigma - \mu) t_{k-1} (t_k - t_k - t_{k-1}) \right\} \\ &= t_{k+1} - t_k.\end{aligned}$$

El siguiente paso es estudiar la sucesión $\{t_k\}$, se tiene

$$\begin{aligned}(1 - \mu t_k)(t_{k+1} - t_k) &= \frac{\sigma}{2} (t_k^2 - 2t_k t_{k-1} + t_{k-1}^2) + \delta_0 (t_k - t_{k-1}) \\ &\quad + \sigma t_{k-1} (t_k - t_{k-1}) - \mu t_{k-1} (t_k - t_{k-1}),\end{aligned}$$

efectuando algunos arreglos y simplificaciones se obtiene la siguiente igualdad

$$\begin{aligned}(1 - \mu t_k)(t_{k+1} - t_k) - \frac{\sigma}{2} t_k^2 + (1 - \delta_0) t_k \\ = (1 - \mu t_{k-1})(t_k - t_{k-1}) - \frac{\sigma}{2} t_{k-1}^2 + (1 - \delta_0) t_{k-1} = \alpha,\end{aligned}$$

expresión que es constante por que no depende de k . Expresando

$$t_{k+1} = t_k + \frac{\frac{\sigma}{2} t_k^2 - (1 - \delta_0) t_k + \alpha}{1 - \mu t_k} = t_k + \frac{u(t_k)}{v(t_k)} = \Phi(t_k),$$

donde las funciones u y v están dadas por

$$u(t) = \frac{\sigma}{2} t^2 - (1 - \delta_0) t + \alpha, \quad v(t) = 1 - \mu t.$$

Se debe mostrar por consiguiente que t_k converge y eso sucede si y solamente si Φ tiene un punto fijo, y el radio expectral de Φ' en las proximidades del punto fijo es estrictamente menor a 1.

Se tiene:

$$\Phi(t) = t \iff u(t) = 0 \iff t = \rho_{1,2} = \frac{1 - \delta_0}{\sigma} \pm \sqrt{\left(\frac{1 - \delta_0}{\sigma}\right)^2 - \frac{2\alpha}{\sigma}},$$

donde $\rho_{1,2}$ son las raíces de $u(t)$. Expresando ambas raíces en función de h , se obtiene

$$\rho_{1,2} = \frac{1 \pm \sqrt{1-2h}}{h} \frac{\alpha}{1-\delta_0},$$

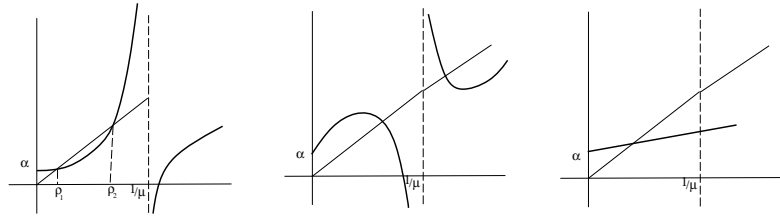
puesto que $h \leq 1/2$, las dos raíces de u son positivas, siendo ρ_1 la raíz más pequeña, se tiene:

$$\rho_1 \leq \rho_2, \quad \rho_1 = \rho.$$

Por otro lado, ambas raíces son acotadas, en efecto

$$\frac{\rho_1 + \rho_2}{2} = \frac{1}{h} \frac{\alpha}{1-\delta_0} = \frac{1-\delta_0}{\sigma} \leq \frac{1}{\mu}.$$

Estudiando la función $\Phi(t)$ se tiene tres casos, que se observan en la figura IV.3.3.



$$\rho_1 \leq \rho_2 \leq \frac{1}{\mu}$$

$$\rho_1 < \frac{1}{\mu} < \rho_2$$

$$\rho_2 = \frac{1}{\mu}$$

Figura IV.3.3. Gráficas de la función Φ .

Para $0 < t < \rho_1$, se tiene que $\Phi'(t) > 0$, en efecto:

$$\begin{aligned} \Phi(t) &= t + \frac{u(t)}{v(t)}, \\ \Phi'(t) &= \frac{v(t) + u'(t)}{v(t)} + u(t) \underbrace{\frac{v'(t)}{v^2(t)}}_{> 0}, \\ v(t) + u'(t) &= 1 - \mu t + \sigma t - (1 - \delta_0) = \delta_0 + (\sigma - \mu)t \geq 0. \end{aligned}$$

Por consiguiente, Φ es creciente sobre el intervalo $(0, \rho_1)$, y además $t_k < \rho_1$; para $k = 0$ es cierto, supóngase cierto para k , entonces

$$t_{k+1} = \Phi(t_k) < \Phi(\rho_1) = \rho_1,$$

La sucesión $\{t_k\}$ es creciente y mayorada, por lo tanto es convergente y converge al primer punto fijo de Φ , el cual es ρ_1 . De aquí, se deduce que

$$x_k \in B(x_0, \rho_1).$$

Ahora se está en condiciones de mostrar que $\{x_k\}$ es una sucesión convergente, para tal motivo se debe demostrar antes que es una sucesión de Cauchy. Se tiene:

$$\begin{aligned} \|x_{k+1} - x_l\| &\leq \|x_{k+1} - x_k\| + \cdots + \|x_{l+1} - x_l\| \\ &\leq t_{k+1} - t_k + \cdots + t_{l+1} - t_l \\ &< \rho_1 - t_l \longrightarrow 0 \end{aligned}$$

cuando $k \geq l \rightarrow \infty$.

De donde $\lim_{k \rightarrow \infty} x_k = x^*$ y como $M(x_k)$ es acotada se deduce inmediatamente

$$f(x^*) = 0,$$

mostrando así el punto ii).

Para la demostración del punto iii) son necesarias algunas convenciones sobre la escritura de los símbolos utilizados. La sucesión $\{x_k\}$ está definida por

$$x_{k+1} = x_k - M(x_0)^{-1}f(x_k),$$

planteando:

$$\bar{M}(x) = M(x_0), \quad \bar{\alpha} = \alpha, \quad \bar{\omega} = \omega, \quad \bar{\mu} = 0.$$

Por otro lado,

$$\begin{aligned} \|M(x_0)^{-1}(f'(x) - \bar{M}(x))\| &\leq \|M(x_0)^{-1}(f'(x) - M(x))\| \\ &\quad + \|M(x_0)^{-1}(M(x) - \bar{M}(x))\| \\ &\leq \delta_0 + \delta_1 \|x - x_0\| + \mu \|x - x_0\| \\ &= \bar{\delta}_0 + \bar{\delta}_1 \|x - x_0\| \end{aligned}$$

con $\bar{\delta}_0 = \delta_0$, $\bar{\delta}_1 = \delta_1 + \mu$ y $\bar{\sigma} = \sigma$. Definiendo $G(x) = x - M(x_0)^{-1}f(x)$, se tiene:

$$\begin{aligned} G'(x) &= I - M(x_0)^{-1}f'(x), \\ \|G(y) - G(x_{k-1})\| &\leq \|G(y) - G(x_{k-1}) - G'(x_{k-1})(y - x_{k-1})\| \\ &\quad + \|(G'(x_{k-1}) - G'(x_0))(y - x_{k-1})\| \\ &\quad + \|G'(x_0)(y - x_{k-1})\| \\ &\leq \|M(x_0)^{-1}(-f(y) + f(x_{k-1}) + f'(x_{k-1})(x - y_{k-1}))\| \\ &\quad + \|M(x_0)^{-1}(f'(x_{k-1}) - f'(x_0))(y - x_{k-1})\| \\ &\quad + \|M(x_0)^{-1}(M(x_0) - f'(x_0))(y - x_{k-1})\| \end{aligned}$$

utilizando una integral como en el punto ii), se obtiene

$$\begin{aligned} \|G(y) - G(x_{k-1})\| &\leq \frac{\omega}{2} \|y - x_{k-1}\| \\ &\quad + \omega \|x_{k-1} - x_0\| \|y - x_{k-1}\| + \delta_0 \|y - x_{k-1}\|, \end{aligned}$$

Sea $y^* \in \mathcal{D} \cap B(x_0, \rho_+)$, remplazando y por y^* en la última desigualdad, se obtiene

$$\begin{aligned} \|y^* - x_{k-1}\| &\leq \frac{\omega}{2} \|y^* - x_{k-1}\| \\ &\quad + \omega \|x_{k-1} - x_0\| \|y^* - x_{k-1}\| + \delta_0 \|y^* - x_{k-1}\|, \end{aligned}$$

y planteando también $y = x_k$, se obtiene

$$\begin{aligned} \|x_k - x_{k-1}\| &\leq \frac{\omega}{2} \|x_k - x_{k-1}\| \\ &\quad + \omega \|x_{k-1} - x_0\| \|x_k - x_{k-1}\| + \delta_0 \|x_k - x_{k-1}\|, \end{aligned}$$

Como en el punto ii) se remplazan $\|x_k - x_0\|$ por $t_k - t_0$ con $t_0 = 0$ y $\|y^* - x_k\|$ por $s_k - t_k$, el símbolo \leq por la igualdad, obteniendo así:

$$\begin{aligned} t_{k+1} - t_k &= \frac{\omega}{2} (t_k - t_{k-1})^2 + \omega t_{k-1} (t_k - t_{k-1}) + \delta_0 (t_k - t_{k-1}), \\ t_0 &= 0, \quad t_1 = \alpha; \\ s_k - t_k &= \frac{\omega}{2} (s_{k-1} - t_{k-1})^2 + \omega t_{k-1} (s_{k-1} - t_{k-1}) + \delta_0 (s_{k-1} - t_{k-1}), \\ s_0 &= \|y^* - x_0\| < \rho_+. \end{aligned}$$

Se demuestra por inducción que:

$$\begin{aligned} \|x_{k+1} - x_k\| &\leq t_{k+1} - t_k, \\ \|x_k - x_0\| &\leq t_k, \\ \|y^* - x_k\| &\leq s_k - t_k \end{aligned}$$

El siguiente paso en la demostración consiste en estudiar las sucesiones $\{t_k\}$ y s_k . Se tiene:

$$\begin{aligned} t_{k+1} - t_k &= \frac{\omega}{2} (t_k^2 - 2t_k t_{k-1} + t_{k-1}^2) + \omega t_{k-1} (t_k - t_{k-1}) + \delta_0 (t_k - t_{k-1}), \\ t_{k+1} - \frac{\omega}{2} t_k^2 - \delta_0 t_k &= t_k - \frac{\omega}{2} t_{k-1}^2 - \delta_0 t_{k-1} = \alpha, \\ s_k - t_k &= \frac{\omega}{2} (s_{k-1}^2 - 2s_{k-1} t_{k-1} + t_{k-1}^2) + \omega t_{k-1} (s_{k-1} - t_{k-1}) + \delta_0 (s_{k-1} - t_{k-1}) \\ s_k - \frac{\omega}{2} s_{k-1}^2 - \delta_0 s_{k-1} &= t_k - \frac{\omega}{2} t_{k-1}^2 - \delta_0 t_{k-1} = \alpha. \end{aligned}$$

Definiendo la función $\Psi(t)$ por

$$\Psi(t) = \frac{\omega}{2}t^2 + \delta_0 t + \alpha,$$

entonces se obtiene, los resultados siguientes para s_k y t_k :

$$\begin{aligned} t_{k+1} &= \Psi(t_k), & t_0 &= 0; \\ s_k &= \Psi(s_{k-1}), & s_0 &= \|y^* - x_0\| < \rho_+. \end{aligned}$$

Se tiene que, $\Psi'(t) = \omega t + \delta_0 \geq 0$, si $t \geq 0$, para estudiar la convergencia de las sucesiones, se debe determinar los puntos fijos de Ψ , es decir resolver la ecuación

$$\frac{\omega}{2}t^2 + (\delta_0 - 1)t + \alpha = 0,$$

las raíces de ésta, están dadas por:

$$\bar{\rho}_{\pm} = \frac{1 \pm \sqrt{1 - 2\bar{h}}}{\bar{h}},$$

se puede observar, que la sucesión t_k es creciente y tiende hacia ρ_- , por otro lado la sucesión s_k es decreciente y tiende hacia ρ_- . Ver la figura IV.3.4

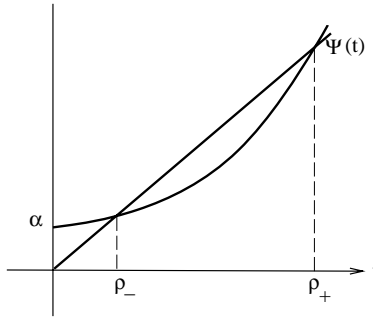


Figura IV.3.4. Gráfica de la función Ψ .

Como en el punto ii) se muestra que la sucesión $\{x_k\}$ es de Cauchy, por consiguiente $x_k \rightarrow x^*$ y $f(x^*) = 0$, además se tiene

$$\begin{aligned} \|x^* - x_0\| &\leq \rho_-, \\ \|y^* - x^*\| &\leq s_k - t_k \rightarrow 0, \\ y^* &= x^*. \end{aligned}$$

□

Para ilustrar la utilización del teorema de Newton-Kantorovich, se tiene el:

Ejemplo

Considérese, la función f dada por

$$f(x) = \begin{pmatrix} x_1^2 + x_2^2 - 1 \\ x_2 - x_1^2 \end{pmatrix},$$

partiendo del punto $(1, 1)$ y tomando como dominio $\mathcal{D} = \mathbb{R}^2$, $M(x) = f'(x)$, es decir el método de Newton en su versión no modificada, se tiene:

$$\begin{aligned} \alpha &= 0,37; & \omega &= 1,2; & \delta_0 &= 0; & \delta_1 &= 0; \\ \mu &= \omega; & \sigma &= 1,2. \end{aligned}$$

Verificando y efectuando cálculos se obtiene:

$$h = 0,444 \leq 1/2, \quad \rho = \rho_- = 0,554; \quad \rho_+ = 0,981.$$

Método de Newton con Relajación

Una de las principales desventajas del método de Newton, tanto en su versión original, como en su versión simplificada, es que es necesario estar cerca de la solución del problema. Lamentablemente, en una gran mayoría de casos eso no es posible. Por lo tanto, es necesario modificar el método de manera a que la convergencia sea una preocupación menor. El método de Newton está dado, por las iteraciones

$$x_{k+1} = x_k - (f'(x_k))^{-1} f(x_k),$$

se tiene convergencia si $\|x_1 - x_0\| \leq \alpha$ suficientemente pequeño, sin embargo esta condición no se cumple en general. Una solución alternativa sería modificar el método de la manera siguiente:

$$\begin{aligned} p_k &= -(f'(x_k))^{-1} f(x_k), \\ x_{k+1} &= x_k + \lambda_k p_k; \end{aligned}$$

con $0 < \lambda \leq 1$. La base teórica de esta modificación está dada por la:

Proposición IV.3.5.- Sean $\mathcal{D} \subset \mathbb{R}^n$, $f : \mathcal{D} \rightarrow \mathbb{R}^n$ continuamente diferenciable, $x_0 \in \mathcal{D}$, $f(x_0) \neq 0$ y $p_0 = -(f'(x_0))^{-1} f(x_0)$. Entonces para toda matriz A invertible, existe $\lambda_0 > 0$ tal que para todo λ , $0 < \lambda < \lambda_0$, se tiene

$$\|Af(x_0 + \lambda p_0)\| < \lambda \|Af(x_0)\|.$$

Demostración.- Se define la función $g : \mathbb{R} \rightarrow \mathbb{R}$ para una matriz A fija e inversible, como

$$\begin{aligned} g(\lambda) &= \|Af(x_0 + \lambda p_0)\|_2^2 \\ &= f(x_0 + \lambda p_0)^t A^t A f(x_0 + \lambda p_0), \end{aligned}$$

calculando la derivada de g , se obtiene:

$$\begin{aligned} g'(\lambda) &= 2f(x_0 + \lambda p_0)^t A^t A f'(x_0 + \lambda p_0)p_0, \\ g'(0) &= -2f(x_0)^t A^t A f'(x_0)(f'(x_0))^{-1}f(x_0) \\ &= -2\|Af(x_0)\|_2^2 < 0, \end{aligned}$$

por consiguiente, existe λ_0 con las conclusiones requeridas por la proposición.

□

Otra de las motivaciones para modificar el método de Newton consiste en el siguiente hecho. Sean $\mathcal{D} \subset \mathbb{R}^n$, $f : \mathcal{D} \rightarrow \mathbb{R}^n$ continuamente diferenciable con derivada inversible para todo $x \in \mathcal{D}$. Llamando $p(x) = -(f'(x))^{-1}f(x)$ la dirección de Newton para la función f , se obtiene la siguiente ecuación diferencial

$$x' = p(x). \quad (\text{IV.3.12})$$

Una manera de resolver numéricamente esta ecuación, ver Capítulo VII, consiste en aproximar la solución mediante segmentos poligonales, es decir, definir una sucesión de vértices, dados por

$$x_{k+1} = x_k - \lambda_k p(x_k), \quad (\text{IV.3.13})$$

este método de resolución de ecuaciones diferenciales es conocido como el método de Euler.

El método modificado se lo conoce como λ -estrategia y su formulación es la siguiente:

$$\begin{aligned} \Delta x_k &= -(f'(x_k))^{-1}f(x_k), \\ x_{k+1} &= x_k + \lambda_k \Delta x_k. \end{aligned} \quad (\text{IV.3.14})$$

Si $\lambda_k = 1$ para todo k , se tiene el método de Newton usual. La otra situación es escoger λ_k , tal que

$$g(\lambda) = \|Af(x_k + \lambda_k \Delta x_k)\| \rightarrow \min, \quad (\text{IV.3.15})$$

presentandose dos interrogantes:

¿Cómo escoger la matriz A ?

¿Cómo calcular el mínimo de $g(\lambda)$?

La elección de A , se la realiza considerando los siguientes hechos. Se desea que

$$\|x_k + \lambda_k \Delta x_k - x^*\| \rightarrow \min,$$

donde x^* es la solución de $f(x) = 0$. No se conoce por el momento x^* , pero se sabe $f(x^*) = 0$. Se tiene, las siguientes aproximaciones:

$$f(x) - f(x^*) \approx f'(x^*)(x - x^*) \approx f'(x_k)(x - x^*)$$

con $x = x_k + \lambda_k \Delta x_k$, de donde

$$x_k + \lambda_k \Delta x_k - x^* = (f'(x_k))^{-1} f(x_k + \lambda \Delta x_k),$$

escogiendo, de esta manera

$$A = (f'(x_k))^{-1}; \quad (\text{IV.3.16})$$

Por lo tanto (IV.3.15), se convierte en $g(\lambda) = \|(f'(x_k))^{-1} f(x_k + \lambda \Delta x_k)\|$ y el problema radica en determinar λ , para que $g(\lambda)$ sea mínimo. A continuación se da un algoritmo simple que permite determinar este λ :

Supóngase que se conoce x_k y una aproximación $\hat{\lambda}_k$ de λ_k . Se tiene dos casos:

a) $g(\hat{\lambda}_k) \geq 0$, es decir $x_k + \hat{\lambda}_k$ es peor que x_k .

Se busca el más pequeño de los $j > 1$, tal que

$$g(\hat{\lambda}_k 2^{-j}) < g(0),$$

se plantea $\lambda_k = \hat{\lambda}_k 2^{-j}$.

b) $g(\hat{\lambda}_k) < 0$. Se busca el $j > 0$ más pequeño tal que

$$g(2^{j+1} \hat{\lambda}_k) > g(2^j \hat{\lambda}_k)$$

y se plantea $\lambda_k = 2^j \hat{\lambda}_k$.

En ambos casos λ_k no debe ser más grande que 1.

Finalmente surge la última interrogante, ¿Cómo determinar $\hat{\lambda}_k$? Utilizando un desarrollo de Taylor se tiene:

$$\begin{aligned} & (f'(x_k))^{-1} f(x_k + \lambda \Delta x_k) \\ &= (f'(x_k))^{-1} \left[f(x_k) + \lambda f'(x_k) \Delta x_k + \frac{\lambda^2}{2} f''(x_k) (\Delta x_k, \Delta x_k) + \mathcal{O}(\lambda^3) \right] \end{aligned}$$

Por otro lado $f(x_k) + \lambda f'(x_k)\Delta x_k = (1 - \lambda)\Delta x_k$ y $(f'(x_k))^{-1}f(x_k) = \Delta x_k$, de donde utilizando el desarrollo de Taylor, la desigualdad del triángulo y las observaciones anteriores, se tiene

$$g(\lambda) \leq \|\Delta x_k\| \left(1 - \lambda + \frac{\lambda^2}{2} h_k + \mathcal{O}(\lambda^3)\right)$$

donde $h_k = \|(f'(x_k))^{-1}f''(x_k)(\Delta x_k, \Delta x_k)\| / \|\Delta x_k\|$. Despreciando el término $\mathcal{O}(\lambda^3)$, se obtiene un polinomio de grado 2, cuyo mínimo se encuentra en

$$\hat{\lambda}_k = \frac{1}{h_k}, \quad (\text{IV.3.17})$$

faltando determinar h_k . Como es prácticamente imposible determinar con exactitud h_k solo se requiere encontrar una buena aproximación de h_k . Para tal efecto, se supone que se ha efectuado $k - 1$ iteraciones, teniendo:

$$\begin{aligned} \Delta x_k &= -(f'(x_k))^{-1}f(x_k), \\ \widehat{\Delta x_k} &= -(f'(x_{k-1}))^{-1}f(x_k), \\ \widehat{\Delta x_k} - \Delta x_k &= (f'(x_{k-1}))^{-1}(f'(x_k) - f'(x_{k-1}))\widehat{\Delta x_k} \\ &\approx (f'(x_{k-1}))^{-1}f''(x_k)(x_k - x_{k-1}, \widehat{\Delta x_k}). \end{aligned}$$

Pasando a las normas se obtiene la siguiente relación:

$$\|\widehat{\Delta x_k} - \Delta x_k\| \approx \frac{h_k}{\|\Delta x_k\|} \lambda_{k-1} \|\Delta x_{k-1}\| \|\widehat{\Delta x_k}\|,$$

de donde

$$\hat{\lambda}_k = \min \left(1, \lambda_{k-1} \frac{\|\Delta x_{k-1}\| \|\Delta x_k\|}{\|\widehat{\Delta x_k} - \Delta x_k\| \|\widehat{\Delta x_k}\|} \right). \quad (\text{IV.3.18})$$

Ejemplo

Considérese, la función f definida por

$$f(x) = \begin{pmatrix} -13x_1 + x_2((5 - x_2) - 2) \\ -29 + x_1 + x_2((1 + x_2)x_2 - 14) \end{pmatrix}$$

Tomando como valor inicial $x_1 = 0,5$ y $x_2 = 2,24$, se resolverá el problema $f(x) = 0$, utilizando la λ -estrategia, el método de Newton en su versión original. A continuación se presentan los resultados obtenidos para ambos:

Tabla IV.3.3. Resultados con λ -estrategia.

Iteración	x_1	x_2	$\ \Delta x\ $	λ
1	-8.5849058	3.94357758	1183.1361	7.8125×10^{-3}
2	4.989738	4.0012650	13.574766	1.0
3	4.9999949	4.0000006	1.03345×10^{-2}	1.0
4	4.9999999	4.000000	5.08213×10^{-6}	1.0
5	5.0	4.0	0.0	1.0

Tabla IV.3.4. Resultados sin λ -estrategia.

Iteración	x_1	x_2	$\ \Delta x\ $
1	-1162.367	220.297	1183.1
2	-47637.0	147.095	46474.7
3	-21035.69	98.296	26601.42
4	-9256.28	65.77062	11779.45
5	-4049.98	44.09533	5206.34
6	-1755.45	29.658	2294.57
7	-748.669	20.056	1006.82
8	-310.0149	13.688	438.7011
9	-121.126	9.5002	188.934
10	-41.5288	6.8016	79.6434
11	-9.51218	5.16002	32.05875
12	1.887	4.3114	11.4307
13	4.7248	4.03167	2.8516
14	4.9968	4.0003	0.27378
15	4.9999	4.0000	3.149×10^{-3}
16	4.9999	4.0	4.566×10^{-7}
17	5.0	4.0	0.0

Aproximación de Broyden

Uno de los principales problemas en la implementación del método de Newton, tanto en su versión con relajación, como en su versión original, es el cálculo de la matriz jacobiana $f'(x_k)$ en cada iteración y por consiguiente mediante el algoritmo de eliminación de Gauss determinar la descomposición LR de esta matriz. Existen muchas alternativas para evitar el cálculo de la matriz jacobiana y su descomposición LR en cada iteración. El siguiente análisis permitirá encontrar algunas de estas alternativas. Utilizando las relaciones dadas por (IV.3.17) y (IV.3.18) respecto a la determinación de los coeficientes de relajación, la experiencia numérica indica que, si $\lambda_k h_k \leq 10^{-1}$ se puede tomar en el lugar de $f'(x_{k+1})$ $f'(x_k)$. En este caso se tiene la versión simplificada del método de Newton. Sin embargo, existe otra alternativa descubierta por Broyden en 1965. Se supone, que se conoce una aproximación J_k de $f'(x_k)$, es decir

$$J_k \approx f'(x_k) \quad (\text{IV.3.19})$$

de donde el método de Newton está dado por

$$\begin{aligned} \Delta x_k &= -J_k^{-1} f(x_k), \\ x_{k+1} &= x_k + \Delta x_k \end{aligned}$$

el objetivo es por consiguiente, encontrar J_{k+1} una aproximación simple de calcular de $f'(x_{k+1})$. Para tal efecto, la matriz J_k debe permanecer igual en las direcciones ortogonales a Δx_k , es decir:

$$\begin{aligned} J_{k+1}p &= J_k p \quad \forall p \perp \Delta x_k; \\ J_{k+1}\Delta x_k &= J_k \Delta x_k + q; \end{aligned} \quad (\text{IV.3.20})$$

por lo tanto

$$q = J_{k+1}\Delta x_k - J_k \Delta x_k = f(x_k) + \underbrace{\underbrace{J_{k+1}\Delta x_k}_{\approx f'(x_{k+1})\Delta x_k}}_{f(x_{k+1}) + \mathcal{O} \|\Delta x\|^2},$$

de donde

$$J_{k+1}\Delta x_k = J_k \Delta x_k + f(x_{k+1}). \quad (\text{IV.3.20})$$

Proposición IV.3.6.- Si J_k es inversible y $\|\bar{\Delta}x_{k+1}\| < \|\Delta x_k\|$, donde

$$\Delta x_k = -J_k^{-1} f(x_k) \quad \bar{\Delta}x_{k+1} = -J_k^{-1} f(x_{k+1}),$$

entonces J_{k+1} es inversible.

Demostración.- Sean, $\alpha \in \mathbb{R}$ y $p \perp \Delta x_k$, se va mostrar que $\ker J_{k+1} = \{0\}$. En efecto:

$$\begin{aligned} 0 &= J_{k+1}(\alpha \Delta x + p) \\ &= \alpha(J_k \Delta x_k + f(x_{k+1})) + J_k p \\ &= J_k(\alpha \Delta x_k + p) + \alpha f(x_{k+1}) \\ &= \alpha \Delta x_k + p - \alpha \bar{\Delta} x_{k+1}, \end{aligned}$$

introduciendo el producto escalar, multiplicando por Δx_k , se obtiene:

$$\alpha \|\Delta x_k\|^2 - \alpha \langle \bar{\Delta} x_{k+1}, \Delta x_k \rangle = 0,$$

por la desigualdad de Cauchy-Schwartz, se tiene

$$|\langle \bar{\Delta} x_{k+1}, \Delta x_k \rangle| \leq \|\bar{\Delta} x_{k+1}\| \|\Delta x_k\| < \|\Delta x_k\|^2,$$

de donde $\alpha = 0$ y por consiguiente $p = 0$. □

La determinación de Δx_{k+1} y $\bar{\Delta} x_k$ se propone en los ejercicios.

Ejercicios

- 1.- Utilizando el teorema de Newton-Kantorovich encontrar $\rho > 0$ tal que la iteración

$$z_{k+1} = z_k - \frac{f(z_k)}{f'(z_k)} \quad f(z) = z^3 - 1$$

converge hacia 1, si $z_0 \in \mathbb{C}$ satisface $|z_0 - 1| < \rho$.

- 2.- Sea $\mathcal{D} \subset \mathbb{R}^n$ y $f : \mathcal{D} \rightarrow \mathbb{R}^n$ continuamente diferenciable. Para $x_0 \in \mathcal{D}$ supóngase que $f(x_0) \neq 0$ y que $f'(x_0)$ sea inversible. Mostrar que

$$p_0 = -f'(x_0)^{-1} f(x_0)$$

es la única dirección que tiene la propiedad siguiente:

para toda matriz inversible A existe $\lambda_0 > 0$ tal que para $0 < \lambda < \lambda_0$

$$\|Af(x_0 + \lambda p_0)\|_2 < \|Af(x_0)\|_2.$$

- 3.- En el artículo

R.S. Dembo, S.C. Eisenstat & T. Steihaug(1982): *Inexact Newton methods*. SIAM J. Numer. Anal., vol. 19,400-408.

los autores consideran la modificación siguiente del método de Newton:

$$\begin{aligned} f'(x_k)\Delta x_k &= -f(x_k) + \tau_k \\ x_{k+1} &= x + l + \Delta x_k. \end{aligned} \quad (\text{IV.3.22})$$

Las perturbaciones τ_k pueden ser interpretadas como la influencia de los errores de redondeo, como el error debido a una aproximación de $f'(x_k), \dots$ Supóngase que $\tau_k = \tau(x_k)$ y que

$$\|\tau(x)\| \leq \eta \|f(x)\| \quad (\text{IV.3.23})$$

con $\eta < 1$.

a) Mostrar el resultado:

Sea $\mathcal{D} \subset \mathbb{R}^n$ abierto, $f : \mathcal{D} \rightarrow \mathbb{R}^n$ y $\tau : \mathcal{D} \rightarrow \mathbb{R}^n$ continuamente diferenciables en \mathcal{D} . Si la condición (IV.3.23) es satisfecha con $\eta < 1$ y si $\|\Delta x_0\|$ es suficientemente pequeño, entonces la iteración (IV.3.21) converge hacia una solución de $f(x) = 0$.

b) El teorema precedente no es cierto, en general, si se reemplaza $\eta < 1$ por $\eta \leq 1$. Mostrarlo.

Indicación. Encontrar una matriz $M(x)$ tal que la iteración (IV.3.21) se vuelva equivalente a

$$M(x_k)\Delta x_k = -f(x_k),$$

y aplicar el teorema de Newton-Kantorovich. Utilizar la norma

$$\|u\|_J = \|f'(x_0)u\|.$$

4.- (U. Ascher & Osborne 1987). Considerar una función $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ que satisfaga

$$\begin{aligned} f(0) &= -\frac{1}{10} \begin{pmatrix} 4\sqrt{3}-3 \\ -4\sqrt{3}-3 \end{pmatrix}, & f'(0) &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ f(u) &= \frac{1}{5} \begin{pmatrix} 4 \\ -3 \end{pmatrix}, & f'(u) &= \begin{pmatrix} 1/\sqrt{3} & -1/\sqrt{3} \\ 1 & 1 \end{pmatrix} \end{aligned}$$

donde $u = -f(0)$. Aplicar el método de Newton con el valor inicial $x_0 = 0$. Para la sucesión $\{x_k\}$ obtenida de esta manera mostrar:

- a) $x_{k+2} = x_k$ para todo $k \geq 0$;
- b) el test de monotonocidad

$$\left\| (f'(x_k))^{-1} f(x_{k+1}) \right\|_2 < \left\| (f'(x_k))^{-1} f(x_k) \right\|_2$$

es satisfecho en cada iteración.

5.- Sea $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ dos veces diferenciable. Mostrar que

$$\|f''(x)(h, k)\|_\infty \leq M \|h\|_\infty \|k\|_\infty$$

$$\text{donde } M = \max_i \sum_j \sum_l \left| \frac{\partial^2 f_i}{\partial x_j \partial x_l}(x) \right|.$$

6.- Sea $f : \mathcal{D} \rightarrow \mathbb{R}^2$ dada por $\mathcal{D} = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mid -1 < x_1, x_2 < 3 \right\}$,

$$f(x) = \begin{pmatrix} x_1 - x_2 \\ (x_1 - 8)x_2 \end{pmatrix}, \quad x_0 = \begin{pmatrix} 0.125 \\ 0.125 \end{pmatrix}.$$

Calcular los valores α y ω del teorema de Newton-Misovski. Mostrar que el metodo de Newton converge hacia una solución de $f(x) = 0$.

7.- Sean $f : \mathcal{D} \rightarrow \mathbb{R}^n$ 3 veces continuamente diferenciable, J_k una aproximación de la matriz $f'(x_k)$ y $p \in \mathbb{R}^n$, $p \neq 0$. Para la aproximación de Broyden

$$J_{k+1} = J_k + \gamma(f(x_k + p) - f(x_k) - J_k p) \cdot \frac{p^t}{p^t p}$$

se tiene:

$$\begin{aligned} \text{a) } (J_{k+1} - f'(x_0 + p))p &= (1 - \gamma)(J_k - f'(x_k))p \\ &\quad + \left(\frac{\gamma}{2} - 1\right)f''(x_k)(p, p) + \mathcal{O}(\|p\|^3). \end{aligned}$$

b) para $\gamma \in [0, 2]$

$$\|J_{k+1} - f'(x_0 + p)\|_2 \leq \|J_k - f'(x_k)\|_2 + \|f''(x_k)(p, \cdot)\|_2 + \mathcal{O}(\|p\|^2).$$

8.- Supóngase que se conoce una aproximación J_0 de $f'(x_0)$, y su descomposición LR . Considere la iteración

$$\begin{aligned} J_k \Delta x_k &= f(x_k) \\ x_{k+1} &= x_k + \Delta x_k \end{aligned}$$

donde J_{k+1} (aproximación de Broyden), está definido por

$$\begin{aligned} J_{k+1} \Delta x_k &= J_k \Delta x_k + f(x_{k+1}); \\ J_{k+1} p &= J_k p, \quad \text{si } p^t \Delta x_k = 0. \end{aligned}$$

Sin calcular explícitamente J_1 y J_2 , encontrar fórmulas para Δx_1 y Δx_2 .

- 9.- Considérese una función $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ donde $m < n$. El conjunto

$$E = \{x | g(x) = 0\}$$

representa una superficie en \mathbb{R}^n .

Dado $\hat{x} \in \mathbb{R}^n$. Proponer un algoritmo para calcular el punto de E el más próximo de \hat{x} . Estudiar las hipótesis para que el algoritmo converga.

- 10.- Considérese la ecuación integrable de Fredholm (cf. Ejercicio 3 sección IV.3)

$$\lambda y(x) = \frac{2}{\pi} \int_0^\pi (3 \sin x \sin t + 2 \sin(2x) \sin(2t))(y(t) + (y(t))^3) dt \quad (\text{IV.3.24})$$

$y(x) = 0$ es solución de (IV.3.24) para todo $\lambda \in \mathbb{R}$. Con las ideas del ejercicio 3 de la anterior sección se puede escribir (IV.3.24) bajo la forma $G(c_1, c_2, \lambda) = 0$.

Calcular los λ , donde $\det \frac{\partial G}{\partial C}(C, 0) = 0$.

Interpretar estos puntos con el teorema de las funciones implícitas. (Soluciones: $\lambda_1 = 2$, $\lambda_2 = 3$)

- 11.- Mostrar numéricamente que para $\lambda_1 = \lambda_1 + \epsilon$ ($\epsilon > 0$) el problema (IV.3.24) posee al menos 3 soluciones, para $\lambda_1 > \lambda_2$ al menos 5 soluciones.

Calcular también todas las soluciones de (IV.3.24) para $\lambda = 10$, (hay 13).

IV.4 Método de Gauss Newton

El estudio de las ecuaciones realizados en las secciones precedentes, consistían en problemas de la forma $f(x) = 0$ o su variante del punto fijo $x = f(x)$, donde $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$. Sin embargo, existen muchos problemas donde intervienen ecuaciones no lineales, que se encuentran en una diversidad de problemas, sobre todo en la determinación de parámetros en el análisis de datos. El problema puede formularse de la siguiente manera. Sea

$$f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}^m \quad \text{con} \quad n \leq m,$$

el problema consiste en encontrar $x \in \mathcal{D}$, tal que

$$\|f(x)\|_2 \longrightarrow \min. \quad (\text{IV.4.1})$$

En el capítulo II.5, se abordó el problema bajo su forma lineal, se formuló el método QR para resolver este problema, y así mismo se introdujo la noción de pseudo-inversa de una matriz. Para el problema no lineal, se sigue el mismo esquema. Pero antes de continuar en esa dirección, existe una alternativa de solución. Se define la función $g : \mathcal{D} \rightarrow \mathbb{R}$ de la manera siguiente

$$g(x) = \|f(x)\|_2^2, \quad (\text{IV.4.2})$$

de donde

$$g'(x) = 2(f'(x))^t f(x), \quad (\text{IV.4.3})$$

se busca, por consiguiente, los $x \in \mathcal{D}$ tales que $(f'(x))^t f(x) = 0$. En principio se podría utilizar el método de Newton para resolver este problema, pero la principal desventaja, consiste en calcular la segunda derivada de f . Este algoritmo da los mínimos locales.

La otra alternativa de solución de este problema, se la dió más arriba, es el método de Gauss-Newton. Al igual que en el método de Newton, la idea principal de este método consiste en linearizar el problema. Sea $x_0 \in \mathcal{D}$ un valor inicial. Utilizando un desarrollo de Taylor en x_0 se obtiene

$$\|f(x)\|_2^2 = \|f(x_0) + f'(x_0)(x - x_0) + \cdots\|_2^2,$$

considerando los términos lineales del segundo miembro de la ecuación se obtiene, al igual que en el capítulo II.5,

$$x - x_0 = -f'(x_0)^+ f(x_0), \quad (\text{IV.4.4})$$

donde $f'(x_0)^+$ es la pseudo inversa de $f'(x)$.

Por lo tanto, se puede formular el método de Gauss-Newton, como:

$$\begin{aligned}\Delta x_k &= -f'(x_k)^+ f(x_k), \\ x_{k+1} &= x_k + \Delta x_k.\end{aligned}\tag{IV.4.5}$$

El cálculo de Δx_k se lo realiza, utilizando la descomposición QR dada en el capítulo II.5. Como se puede observar, la implementación del método de Gauss-Newton es muy simple, y al igual que el método de Newton existen muchas variantes o modificaciones de éste.

Convergencia del Método de Gauss-Newton

Sea, $\{x_k\}$ la sucesión definida por el método de Gauss-Newton, si esta sucesión es convergente cuando $k \rightarrow \infty$, se tiene

$$\Delta x_k \rightarrow 0, \quad f'(x)^+ f(x_k) \rightarrow 0,$$

supóngase que $\lim_{k \rightarrow \infty} x_k = x^*$, que el rango de $f'(x)$ es constante en un vecindario de x^* , por que si no $f'(x)^+$ no sería continua, ver teorema II.5.6, entonces

$$f'(x^*)^+ f(x^*) = 0.\tag{IV.4.6}$$

Por otro lado, si $x \in \mathcal{D}$ es un mínimo local de $g(x) = \|f(x)\|_2^2$, se tiene

$$\frac{\partial}{\partial x_j} \left(\sum_{i=1}^m f_i(x)^2 \right) = 2 \sum_{i=1}^m \frac{\partial f_i}{\partial x_j}(x) \cdot f_i(x) = 0, \quad \forall j;$$

de donde

$$f'(x)^t f(x) = 0.\tag{IV.4.7}$$

Los resultados (IV.4.6) y (IV.4.7) están relacionados por el siguiente:

Proposición IV.4.1.- *Se tiene*

$$f'(x)^+ f(x) = 0 \iff f'(x)^t f(x) = 0.\tag{IV.4.8}$$

Demostración.- Por el teorema II.5.5 existen dos matrices ortogonales U y V , tales que

$$f'(x) = U \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_k & \\ 0 & & & 0 \end{pmatrix} V^t,$$

reemplazando, se tiene

$$f'(x)^t = V \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_k & \\ & 0 & & 0 \end{pmatrix} U^t = 0,$$

si y solamente si las primeras k componentes de $U^t f$ son nulas, si y solamente si

$$f'(x)^+ V \begin{pmatrix} \sigma_1^{-1} & & & \\ & \ddots & & 0 \\ & & \sigma_k^{-1} & \\ & 0 & & 0 \end{pmatrix} U^t = 0.$$

□

A la diferencia del método de Newton, el método de Gauss-Newton ofrece solamente convergencia de tipo lineal, situación que será mostrada a continuación. Como antes, se define $g(x) = (f'(x))^t f(x)$, si x^* es solución del problema de minimización de la norma euclidiana, se tiene $g(x^*) = 0$. El desarrollo en serie de Taylor en punto x de la función g , da el siguiente resultado

$$0 = g(x) + g'(x)(x^* - x) + \mathcal{O}(\|x^* - x\|_2^2),$$

por otra lado, la función g y sus derivadas parciales satisfacen:

$$g_j(x) = \sum_{i=1}^m \frac{\partial f_i(x)}{\partial x_j} f_i(x),$$

$$\frac{\partial g_j}{\partial x_k}(x) = \sum_{i=1}^m \frac{\partial^2 f_i(x)}{\partial x_j \partial x_k} f_i(x) + \sum_{i=1}^m \frac{\partial f_i}{\partial x_j}(x) \frac{\partial f_i}{\partial x_k}(x),$$

de donde

$$g'(x) = B(x)(f(x), \cdot) + f'(x)^t f'(x),$$

con $B(x)$ una forma bilineal simétrica. Por consiguiente el desarrollo de Taylor de g en x_k , el k -ésimo valor obtenido en la iteración del método de Gauss-Newton, es igual a

$$0 = f'(x_k)^t f(x_k) + f'(x_k)^t f'(x_k)(x^* - x_k) + B(x_k)(f(x_k), x^* - x_k) + \mathcal{O}(\|x^* - x_k\|_2^2),$$

y por el método de Gauss Newton, se tiene

$$x_{k+1} - x_k = -f'(x_k)^+ f(x_k),$$

multiplicando esta última expresión por $f'(x_k)^t f'(x_k)$ se obtiene

$$0 = f'(x_k)^t f'(x_k) f'(x_k)^+ f(x_k) + f'(x_k)^t f'(x_k) (x_{k+1} - x_k),$$

y utilizando el teorema II.5.7, el inciso d), se tiene finalmente

$$f'(x_k)^t f(x_k) + f'(x_k)^t f'(x_k) (x_{k+1} - x_k);$$

sustrayendo esta última expresión con aquella concerniente al desarrollo de Taylor, se obtiene

$$0 = f'(x_k)^t f'(x_k) (x_{k+1} - x^*) + B(x_k)(f(x_k), x^* - x_k) + \mathcal{O}(\|x^* - x_k\|_2^2).$$

Ahora bien, supóngase además que $\text{rang } f'(x_k) = n$, es decir que el rango sea maximal. De donde se tiene

$$x_{k+1} - x^* = - \left(f'(x_k)^t f'(x_k) \right)^{-1} B(x_k)(f(x_k), x^* - x_k) + \mathcal{O}(\|x^* - x_k\|_2^2),$$

por consiguiente, el método de Gauss-Newton converge linealmente si

$$\rho \left(\left(f'(x_k)^t f'(x_k) \right)^{-1} B(x_k)(f(x_k),) \right) < 1;$$

y diverge si

$$\rho \left(\left(f'(x_k)^t f'(x_k) \right)^{-1} B(x_k)(f(x_k),) \right) > 1.$$

Debe observarse que si $f(x^*)=0$, el método de Gauss-Newton converge cuadráticamente, en este caso se tiene un problema compatible. Por otro lado, el problema inicial era buscar x tal que $f(x) = 0$, motivo por el cual se debe esperar que la solución encontrada por el método de Gauss-Newton de $f(x^*)$ bastante pequeño, de manera que el radio espectral sea más pequeño que 1.

Al igual que en el método de Newton, el cálculo de $f'(x)$ se lo realiza en muchas ocasiones numéricamente, o en los casos en que se pueda analíticamente. Es indudable que la matriz $f'(x)$ en las iteraciones del método de Gauss-Newton tienen un componente de error, utilizando las relaciones (IV.3.5) y (IV.3.6), este error es del orden de $\sqrt{\epsilon ps}$, donde ϵps es la precisión del computador. Por consiguiente la k -ésima iteración del método de Gauss-Newton, considerando que el error cometido por redondeo se encuentra solamente en el cálculo de $f'(x)$, está dada por

$$\Delta x_k = -(f'(x_k) + E)^+ f(x_k),$$

con $\|E\|_2 = \sqrt{\epsilon ps}$. Efectuando operaciones con la pseudo inversa, dadas en el teorema II.5.7 se obtiene:

$$\begin{aligned}\Delta x_k &= - \left((f'(x_k) + E)^t (f'(x_k) + E) \right)^{-1} (f'(x_k) + E)^t f(x_k), \\ \left((f'(x_k) + E)^t (f'(x_k) + E) \right) \Delta x_k &= -(f'(x_k) + E)^t f(x_k), \\ \left((f'(x_k) + E)^t (f'(x_k) + E) \right) \Delta x_k + (f'(x_k) + E)^t f(x_k) &= 0,\end{aligned}$$

despreciando E en el primer término de la última ecuación, se obtiene

$$\left(f'(x_k)^t f'(x_k) \right) \Delta x_k + f'(x_k)^t f(x_k) + E^t f(x_k) = 0,$$

introduciendo en la serie de Taylor, como se hizo más arriba, se obtiene

$$\begin{aligned}-E^t f(x_k) &= f'(x_k)^t f'(x_k)(x_{k+1} - x^*) + B(x_k)(f(x_k), x^* - x_k) \\ &\quad + \mathcal{O}(\|x^* - x_k\|_2^2),\end{aligned}$$

mostrando así que si $f(x^*) \neq 0$, es inútil buscar una precisión superior a $\sqrt{\epsilon ps}$.

Modificaciones del Método de Gauss-Newton

Similarmente al método de Newton, existen varias modificaciones del método original de Gauss-Newton. Una de las mayores dificultades en la implementación de este método consiste en calcular $f'(x_k)$ en cada iteración. Por consiguiente, una primera modificación consiste en utilizar el método de Gauss-Newton simplificado, el cual está dado por:

$$\begin{cases} x_0 \text{ bastante próximo de la solución,} \\ \Delta x_k = -f'(x_0)^+ f'(x_k), \end{cases}$$

La principal ventaja de utilizar el método de Gauss-Newton simplificado consiste en calcular por una vez $f'(x_0)$ y luego efectuar la descomposición QR . Sin embargo existe una desventaja, que no se puede llegar a la solución x^* , si $f(x^*) \neq 0$, esto se ha observado al finalizar la anterior subsección.

Otra modificación corrientemente utilizada en el método de Gauss-Newton consiste en utilizar coeficientes de relajación o más conocido como la λ -estrategia. Sobre todo, se utiliza la relajación cuando los valores iniciales utilizados para activar el método no están muy cerca de la solución buscada, además como un medio de aceleración de convergencia. Por lo tanto el método de Gauss-Newton con λ -estrategia, está dado por

$$\begin{cases} x_0 \text{ bastante próximo de la solución;} \\ \Delta x_k = -f'(x_0)^+ f'(x_k), \\ x_{k+1} = x_k + \lambda_k \Delta x_k, \quad 0 < \lambda_k \leq 1. \end{cases}$$

El valor λ_k se determina de la misma manera, que para el método de Newton, con la única variación de utilizar la pseudo-inversa en el lugar de la inversa.

Ejemplos

- 1.- Este ejemplo está relacionado con la subsección concerniente a la convergencia del método de Gauss-Newton. Se busca $x \in \mathbb{R}$ tal que $g(t) = e^{xt}$ ajuste lo mejor posible m puntos (t_i, y_i) , $i = 1, \dots, m$; con $m \geq 1$. Por consiguiente, el problema consiste en determinar x tal que

$$\sum_{i=1}^m (y_i - e^{xt_i}) \longrightarrow \min,$$

de donde, dentro las características de la aplicación del método de Gauss-Newton se tiene:

$$f(x) = \begin{pmatrix} e^{xt_1} - y_1 \\ \vdots \\ e^{xt_m} - y_m \end{pmatrix}, \quad f'(x) = \begin{pmatrix} t_1 e^{xt_1} \\ \vdots \\ t_m e^{xt_m} \end{pmatrix},$$

continuando con los cálculos se tiene:

$$f'(x)^t f'(x) = \sum_{i=1}^m (t_i e^{xt_i})^2,$$

$$B(x)(f(x),) = \sum_{i=1}^m t_i^2 e^{xt_i} (e^{xt_i} - y_i).$$

Sea por otro lado ρ tal que

$$|e^{xt_i} - y_i| \leq \rho e^{xt_i} \quad i = 1, \dots, m;$$

obteniendo, la mayoración siguiente

$$\left| \frac{\sum_{i=1}^m t_i^2 e^{xt_i} (e^{xt_i} - y_i)}{\sum_{i=1}^m (t_i e^{xt_i})^2} \right| \leq \rho,$$

se tiene convergencia si $\rho < 1$.

- 2.- Este ejemplo es una ilustración numérica de la implementación del método de Gauss-Newton con relajación. El problema consiste en determinar la circunferencia que pasa lo más cerca posible de los

siguientes puntos: $(2, 3)$, $(2, 1)$, $(1, 2)$, $(3, 2)$ y $(2.5, 2.5)$. Ahora bien la ecuación de una circunferencia de centro (h, k) y radio r está dada por

$$(x - h)^2 + (y - k)^2 - r^2 = 0,$$

por consiguiente el problema consiste en determinar h , k y r de manera que se obtenga la circunferencia más próxima a estos puntos. Para la implementación del método de Gauss-Newton, se tiene que

$$f(h, k, r) = \begin{pmatrix} (h - 2)^2 + (k - 3)^2 - r^2 \\ (h - 2)^2 + (k - 1)^2 - r^2 \\ (h - 1)^2 + (k - 2)^2 - r^2 \\ (h - 3)^2 + (k - 2)^2 - r^2 \\ (h - 2.5)^2 + (k - 2.5)^2 - r^2 \end{pmatrix}$$

Después de un simple gráfico, se puede tomar como valores iniciales $h = 2$, $k = 2$ y $r = 1$, utilizando el método de Gauss-Newton con relajación después de 5 iteraciones, se obtiene los siguientes valores:

$$h = 1.95833333333295,$$

$$k = 1.95833333333295,$$

$$r = 0.959239125904873$$

En la figura IV.4.1, se tiene la gráfica de la circunferencia resultante del método de Gauss-Newton.

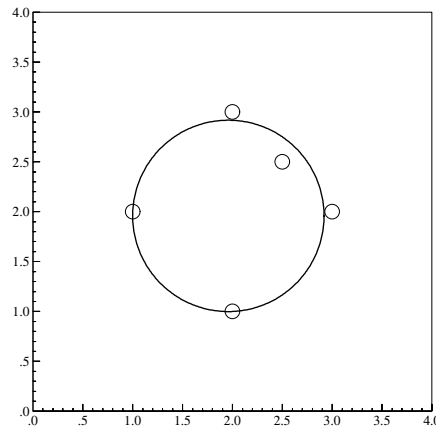


Figura IV.4.1. Circunferencia del método de Gauss-Newton.

El Método de Levenberg-Marquandt

El problema abordado en esta sección, es resolver $\|f(x)\| \rightarrow \min$, donde $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, con $n \leq m$. El método propuesto en el anterior paragrafo fue el método de Gauss-Newton. Como se ha visto, uno de los mayores inconvenientes es que para iniciar las iteraciones se debe partir de un punto inicial bastante próximo de la solución, por otro lado los análisis hechos se basan en el supuesto que $f'(x)$ es de rango igual a n , es decir de rango maximal, además si $f(x)$ no es lo suficientemente pequeño el método de Gauss-Newton diverge. Por las razones expuestas es necesario formular un método alternativo que permita evitar las situaciones enumeradas anteriormente.

El método de Levenberg-Marquandt que constituye un método alternativo para resolver el problema de encontrar x con $\|f(x)\|$ mínima, se basa en las siguientes ideas. Al formular el método de Gauss-Newton, se consideró la serie de Taylor de $f(x)$ en el punto x_k , resultado de la $k - 1$ iteración y se planteó

$$\|f(x)\| = \|f(x_k) + f'(x_k)(x - x_k) + \mathcal{O}(\|x - x_k\|)\| \longrightarrow \min,$$

suponiendo que $x - x_k$ es lo suficientemente pequeño, se obtenía como formulación del método de Gauss-Newton

$$\|f(x_k) + f'(x_k)(x - x_k)\| \longrightarrow \min,$$

y como condición suplementaria deseable

$$\|x - x_k\| \longrightarrow \min,$$

La idea de Levenberg fue de considerar el problema siguiente para determinar el valor de x_{k+1}

$$\|f(x_k) + f'(x_k)(x - x_k)\|_2^2 + p\|x - x_k\|_2^2 \longrightarrow \min, \quad (\text{IV.4.9})$$

donde $p > 0$ es un parámetro fijo. Si se deriva una vez la expresión (IV.4.9), se obtiene:

$$\begin{aligned} 2f'(x_k)^t(f(x_k) + f'(x_k)(x - x_k)) + 2p(x - x_k) &= 0, \\ \left(f'(x_k)^t f'(x_k) + pI\right)(x_{k+1} - x_k) &= -f'(x_k)^t f(x_k). \end{aligned} \quad (\text{IV.4.10})$$

Dependiendo de la elección del parámetro p , el método de Levenberg-Marquandt, tiene los siguientes comportamientos:

Si $p \rightarrow 0$, se tiene el método de Gauss-Newton, para ver realmente lo que sucede ver el capítulo II, sección 5, ejercicio 5.

Si p es muy grande, entonces

$$\Delta x_k \approx -\frac{1}{p} f'(x_k)^t f(x_k),$$

de esta manera, se tiene la dirección de la pendiente más grande, o llamada también del gradiente, pero en este caso la convergencia podría ser muy lenta.

La elección de p puede ser variable, en el sentido siguiente, comenzar con p grande hasta que Δx_k sea bastante pequeño y luego hacer decrecer p hasta obtener el método de Gauss-Newton.

Ejercicios

- 1.- Encontrar una elipse, una hipérbola, que pasa lo mejor posible por los puntos $(x_i, y_i)_{i=1, \dots, m}$. Por ejemplo $m = 10$; $(-1/2, 1)$, $(-1/2, -0.1)$, $(2.5, -1)$, $(1.5, 1.5)$, $(0.3, -1.5)$, $(0.5, -2)$, $(3, 0.1)$, $(3, -0.3)$, $(2.5, 1)$, $(0.5, 3.5)$.

Indicaciones.- Plantear

$$f(x, y) = ax^2 + 2bxy + cy^2 + dx + ey + g$$

y encontrar a, b, c, d, e, g tales que

$$\sum_{i=1}^m (f(x_i, y_i))^2 \longrightarrow \min$$

bajo la condición $ac - b^2 = 1$ para la elipse, y $ac - b^2 = -1$ para la hipérbola.

- 2.- Considérese el problema de localizar un barco en el oceano. Para volver el cálculo más simple, se supondrá que la curvatura terrestre es despreciable y que todos los puntos están situados en el sistema de coordenadas rectangulares (x, y) . Sobre el barco, se mide el ángulo entre el eje x y las varias estaciones emisoras, cuyas coordenadas son conocidas.

i	x_i	y_i	α_i
1	8	6	42
2	-4	5	158
3	1	-3	248

Teóricamente, la posición (x, y) del barco satisface la relación

$$\arctan\left(\frac{y - y_i}{x - x_i}\right) = \alpha_i \quad \forall i.$$

Encontrar la posición probable de éste. Las coordenadas aproximadas del barco son $x_0 = 3$, $y_0 = 2$.

Atención.- Hacer los cálculos en radianes y utilizar siempre la misma rama de la función \arctan .

Capítulo V

Cálculo de Valores Propios

La determinación de valores propios o autovalores de una matriz A es muy corriente en la resolución de múltiples problemas. Sus aplicaciones van desde problemas de minimización y maximización, solución de sistemas de ecuaciones diferenciales, ecuaciones con derivadas parciales, etc. El cálculo de valores propios y la determinación de vectores propios para matriz de un orden inferior o igual a 3 no presenta mayor problema, puesto que su solución depende del polinomio característico cuyo grado en este caso es inferior o igual a 3. Sin embargo, aquéllos problemas, donde se requiere saber los valores y vectores propios, provienen de matrices cuyo orden es mucho mas grande que 3. Es por eso, necesario e imprescindible formular métodos y algoritmos, lo mas eficientes posibles, teniendo en cuenta las particularidades de las matrices que son estudiadas. Ahora bien, la mayor parte de los métodos consevidos sirven para matrices simétricas o matrices normales, pues son éstas, las que se encuentran en la mayor parte de los problemas.

En este capítulo, se estudiarán y formularán tales métodos, pero se comenzará haciendo una introducción teórica del problema de la determinación de autovalores y vectores propios. Como segunda parte de este capítulo, se tendrá la formulación de estos métodos.

V.1 Teoría Clásica y Condición del Problema

En esta sección, se tratará los aspectos teóricos indispensables para la comprensión del problema de la evaluación de valores propios de una matriz A de orden n .

Definición V.1.1.- Sea A una matriz de $n \times n$ a coeficientes complejos o reales. $\lambda \in \mathbb{C}$ es un valor propio, si existe $x \in \mathbb{C}^n$ no nulo, tal que

$$Ax = \lambda x. \quad (\text{V.1.1})$$

Si este es el caso, x es un vector propio asociado al valor propio λ .

Proposición V.1.2.- Sea $A \in M_n(\mathbb{C})$, $\lambda \in \mathbb{C}$ es un valor propio, si y solamente si

$$\ker(A - \lambda I) \neq \{0\}.$$

Demostración.- Resultado inmediato de la definición. \square

Consecuencia de la anterior proposición, se tiene que λ es un valor propio si y solamente si

$$\det(A - \lambda I) = 0, \quad (\text{IV.1.2})$$

de donde, se tiene la:

Definición V.1.3.- El polinomio característico de la matriz $A \in M_n(\mathbb{C})$ está dado por

$$\chi_A(\lambda) = \det(A - \lambda I). \quad (\text{IV.1.3})$$

Por consiguiente, se deduce fácilmente que λ es un valor propio de A si y solamente si λ es una raíz de $\chi_A(\lambda)$. Por otro lado, este polinomio es a coeficientes complejos, de donde existen n valores propios, contando su multiplicidad, de la matriz $A \in M_n(\mathbb{C})$.

Proposición V.1.4.- Valor propio es una propiedad invariante por similitud, es decir si $B = T^{-1}AT$, con T inversible, entonces A y B tienen los mismos valores propios.

Demostración.- En efecto, sea λ valor propio de A , entonces existe un vector propio asociado a λ que se lo denota por v , se tiene

$$BT^{-1}v = T^{-1}ATT^{-1}v = T^{-1}Av = \lambda T^{-1}v,$$

de donde λ es un valor propio de B con $T^{-1}v$ valor propio asociado. \square

Definición V.1.5.- Sea $A \in M_n(\mathbb{C})$, se define la adjunta de A por

$$A_{ij}^* = \bar{a}_{ji}.$$

Definición V.1.6.- Se dice que una matriz U es unitaria si

$$U^*U = I.$$

Vale la pena recalcar, que una matriz ortogonal a coeficientes reales es unitaria, y recíprocamente una matriz unitaria a coeficientes reales es ortogonal, en el caso en que existieran coeficientes complejos no reales la situación cambia. Por otro lado, es fácil ver que el conjunto de las matrices unitarias forman un grupo para la multiplicación de matrices.

Teorema V.1.7.- Schur. Para cada matriz $A \in M_n(\mathbb{C})$, existe una matriz Q unitaria, tal que

$$Q^*AQ = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & & * \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}.$$

Demostración.- Sea A una matriz cualquiera, entonces existe λ_1 que es raíz de $\det(A - \lambda I)$, de donde existe v_1 vector propio asociado a λ_1 , se puede suponer, sin perder generalidad, que $\|v_1\|_2 = 1$. Se plantea

$$Q_1 = (v_1, v_2, \dots, v_n),$$

con v_2, \dots, v_n elegidos de manera que Q_1 sea unitaria. Esta elección es posible con el procedimiento de Gramm-Schmidt en el caso complejo. Se observa inmediatamente, que

$$AQ_1 = Q_1 \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & A^{(1)} & \\ 0 & & & \end{pmatrix}.$$

De la misma manera se encuentra una matriz \bar{Q}_2 de orden $n - 1$, tal que

$$A^{(1)}\bar{Q}_2 = Q_2 \begin{pmatrix} \lambda_2 & * & \cdots & * \\ 0 & & & \\ \vdots & & A^{(2)} & \\ 0 & & & \end{pmatrix},$$

y planteando

$$Q_2 = \begin{pmatrix} 1 & 0 \\ 0 & \bar{Q}_2 \end{pmatrix},$$

se tiene

$$AQ_1Q_2 = Q_2Q_1 \begin{pmatrix} \lambda_1 & * & * & \cdots & * \\ 0 & \lambda_2 & & & \\ & 0 & & & \\ & \vdots & & A^{(2)} & \\ & 0 & & & \end{pmatrix}.$$

repetiendo el procedimiento las veces que sea necesario, se tiene el resultado deseado. Cabe recalcar que este procedimiento da la descomposición requerida en a lo más n pasos. \square

Teorema V.1.8.- Si A es normal, es decir $A^*A = AA^*$, entonces existe una matriz unitaria Q tal que

$$Q^*AQ = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

Demostración.- Si la matriz A es normal y Q es unitaria, entonces Q^*AQ es normal. En efecto

$$\begin{aligned} (Q^*AQ)^* (Q^*AQ) &= Q^*A^*QQ^*AQ \\ &= Q^*AA^*Q \\ &= (Q^*AQ)(Q^*AQ)^*. \end{aligned}$$

queda por demostrar, que si

$$B = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & & * \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}$$

es triangular y normal, entonces

$$B = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Puesto que $BB^* = B^*B$, se tiene

$$\begin{aligned}
& \begin{pmatrix} \bar{\lambda}_1 & 0 & \cdots & \\ * & \bar{\lambda}_2 & & 0 \\ & & \ddots & \\ * & & & \bar{\lambda}_n \end{pmatrix} \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & & * \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} \\
&= \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & & * \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} \begin{pmatrix} \bar{\lambda}_1 & 0 & \cdots & \\ * & \bar{\lambda}_2 & & 0 \\ & & \ddots & \\ * & & & \bar{\lambda}_n \end{pmatrix},
\end{aligned}$$

de donde

$$|\lambda_1|^2 = |\lambda_1|^2 + |*|^2 + \cdots + |*|^2,$$

dando como resultado que la primera fila fuera del coeficiente de la diagonal es nulo. Se continua con el mismo procedimiento hasta mostrar que B es diagonal. \square

La condición del Problema a Valores Propios

Sea, A una matriz cualquiera a coeficientes complejos de orden n . Al igual que en capítulo II, es importante conocer la condición del problema planteado, es decir la determinación de valores propios. Para tal efecto sea \bar{A} la matriz con errores de redondeo de A , repasando la sección II.1, se tiene que los coeficientes de \bar{A} satisfacen

$$\bar{a}_{ij} = a_{ij}(1 + \epsilon_{ij}) \quad |\epsilon_{ij}| \leq eps,$$

donde \bar{a}_{ij} son los coeficientes de \bar{A} , a_{ij} los coeficientes de A y eps la precisión de la computadora. Por consiguiente, el estudio de los valores propios de \bar{A} pueden estudiarse de manera, más general, para $A + \epsilon C$, con $c_{ij} = \frac{\epsilon_{ij}}{eps}$ de donde $|c_{ij}| \leq |a_{ij}|$. Definiendo

$$f(\epsilon, \lambda) = \det(A + \epsilon C - \lambda I),$$

se tiene inmediatamente que

$$f(0, \lambda) = \chi_A(\lambda).$$

Supóngase que λ_1 es una raíz simple de $\chi_A(\lambda)$, entonces se tiene

$$\begin{aligned}
f(0, \lambda_1) &= 0, \\
\frac{\partial f}{\partial \lambda}(0, \lambda_1) &\neq 0.
\end{aligned}$$

Por el teorema de las funciones implícitas, se deduce que existe un vecindario de $(0, \lambda_1)$ en el cual existe una función $\lambda(\epsilon)$, tal que

$$f(\epsilon, \lambda(\epsilon)) = 0,$$

con

$$\lambda(\epsilon) = \lambda_1 + \epsilon \lambda'_1 + \mathcal{O}(\epsilon^2).$$

Por consiguiente, se tiene el siguiente:

Teorema V.1.9.- *Sea λ_1 una raíz simple de $\chi_A(\lambda)$, entonces para $\epsilon \rightarrow 0$, existe un valor propio $\lambda(\epsilon)$ de $A + \epsilon C$, tal que*

$$\lambda(\epsilon) = \lambda_1 + \epsilon \frac{u_1^* C v_1}{u_1^* v_1} + \mathcal{O}(\epsilon^2) \quad (\text{V.1.4})$$

donde $Av_1 = \lambda_1 v_1$ y $u_1^* A = \lambda_1 u_1^*$, u y v no nulos.

Demostración.- Por el teorema, de las funciones implícitas se tiene

$$(A + \epsilon C)v(\epsilon) = \lambda(\epsilon)v(\epsilon),$$

mas precisamente

$$(A + \epsilon C)(v_1 + \epsilon v'_1 + \mathcal{O}(\epsilon^2)) = (\lambda_1 + \epsilon \lambda'_1 + \mathcal{O}(\epsilon^2))v_1,$$

comparando los términos de igual grado en cada miembro de la ecuación anterior, se tiene:

$$\begin{aligned} \epsilon^0 : \quad & Av_1 = \lambda_1 v_1, \\ \epsilon^1 : \quad & Av'_1 + Cv_1 = \lambda_1 v'_1 + \lambda'_1 v_1, \end{aligned}$$

de donde

$$(A - \lambda_1 I)v'_1 = -Cv_1 + \lambda'_1 v_1,$$

multiplicando ésta última por u_1^* , se obtiene

$$0 = -u_1^* C v_1 + \lambda'_1 u_1^* v_1.$$

□

Corolario V.1.10.- *Si A es normal con valores propios diferentes, entonces*

$$\left| \frac{u_1^* C v_1}{u_1^* v_1} \right| \leq \|C\|,$$

es decir, el problema está bien condicionado.

Demostración.- A es normal, por consiguiente existe una matriz Q unitaria tal que

$$Q^* A Q = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix},$$

una verificación inmediata muestra que v_1 es la primera columna de Q , así mismo u_1^* es la primera fila de Q^* , de donde $v_1 = v_1$, por lo tanto, se obtiene

$$\begin{aligned} \left| \frac{u_1^* C v_1}{u_1^* v_1} \right| &\leq \frac{|u_1^* C v_1|}{\|v_1\|^2} \\ &\leq \frac{\|v_1\| \|C v_1\|}{\|v_1\|^2} \\ &\leq \|C\|. \end{aligned}$$

□

Ejemplos

1.- Se considera la matriz A definida por

$$A = \begin{pmatrix} 1 & \alpha \\ 0 & 2 \end{pmatrix}.$$

Es muy simple darse cuenta, que $\lambda = 1$ es un valor propio de A , por consiguiente por simple inspección, se obtiene:

$$\begin{aligned} v_1 &= \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\ u_1 &= \frac{1}{\sqrt{1 + \alpha^2}} \begin{pmatrix} 1 \\ -\alpha \end{pmatrix}, \end{aligned}$$

de donde

$$u_1^* v_1 = \frac{1}{\sqrt{1 + \alpha^2}}.$$

La matriz A es mal condicionada para el cálculo de los valores propios cuando $\alpha \rightarrow \infty$, ya que $u_1^* v_1 \rightarrow 0$.

2.- El teorema de descomposición de Jordan indica que toda matriz A es similar a una matriz de tipo Jordan, es decir existe una matriz T inversible tal que

$$T^{-1} A T = J,$$

donde J es una matriz diagonal por bloques,

$$J = \begin{pmatrix} J(n_1, \lambda_1) & & \\ & \ddots & \\ & & J(n_k, \lambda_k) \end{pmatrix},$$

y

$$J(n_i, \lambda_i) = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}.$$

Ahora bien, supóngase que A es similar a la matriz de tipo Jordan, dada por

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

entonces se tiene que

$$\det(A + \epsilon C - \lambda I) = (1 - \lambda)^4 + \epsilon + \mathcal{O}(\epsilon^2),$$

donde C es una perturbación de A .

Calculando los valores propios de $A + \epsilon C$, ver figura V.1.1, es fácil ver que el problema de determinar valores propios de A es un problema mal condicionado

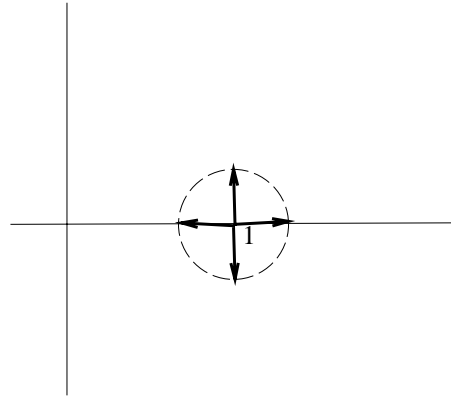


Figura IV.1.1. Determinación de los valores propios de $A + \epsilon C$

Ejercicios

1.- Una matriz A es de tipo Frobenius, si es de la forma

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & & \ddots & \vdots \\ 0 & \cdots & & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-2} & -a_{n-1} \end{pmatrix}.$$

a) Verificar que

$$\det(A - \lambda I) = (-1)^n (\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0).$$

b) Calcular los vectores propios de A .

2.- Calcular los valores propios de la matriz tridiagonal

$$A = \left(\begin{pmatrix} a & c & & 0 \\ b & a & c & \\ & b & a & c \\ 0 & & b & \cdots \\ & & & \ddots \end{pmatrix} \right) \Bigg\} n.$$

Las componentes del vector propio $(v_1, v_2, \dots, v_n)^t$ satisfacen una ecuación de diferencias finitas con $v_0 = v_{n+1} = 0$.

Verificar que $v_j = \text{Const}(\alpha_1^j - \alpha_2^j)$ donde

$$\alpha_1 + \alpha_2 = \frac{a - \lambda}{c}, \quad \alpha_1 \alpha_2 = \frac{b}{c}, \quad \left(\frac{\alpha_1}{\alpha_2} \right)^{n+1} = 1.$$

3.- Mostrar que los valores propios de una matriz A satisfacen

$$\sum_{i=1}^n |\lambda_i|^2 \leq \sum_{i,j=1}^n |a_{ij}|^2.$$

Se tiene igualdad, si y solamente si, A es diagonalizable con una matriz unitaria.

Indicación.- $\sum_{i,j=1}^n |a_{ij}|^2$ es la traza de A^*A que es invariante respecto a la transformación $A \rightarrow Q^*AQ$.

- 4.- Supóngase que los valores propios de A con $a_{ij} \in \mathbb{R}$ son: $\alpha + i\beta$, $\alpha - i\beta$, $\lambda_3, \dots, \lambda_n$ con $\beta \neq 0$, $\lambda_3, \dots, \lambda_n$ diferentes. Mostrar que existe una matriz T inversible con coeficientes reales tal que

$$T^{-1}AT = \begin{pmatrix} \alpha & \beta & & 0 \\ -\beta & \alpha & & \\ & & \lambda_3 & \\ 0 & & & \ddots \\ & & & & \lambda_n \end{pmatrix}.$$

Dar una relación entre las columnas de T y los vectores propios de A .

- 5.- Sea A una matriz simétrica y B una matriz cualquiera. Para cada valor propio λ_B de B existe un valor propio λ_A de A tal que

$$|\lambda_A - \lambda_B| \leq \|A - B\|_2.$$

Indicación.- Mostrar primero para un cierto v

$$v = (A - \lambda_B I)^{-1}(A - B)v,$$

deducir que

$$1 \leq \|(A - \lambda_B I)^{-1}(A - B)\| \leq \|A - B\| \|(A - \lambda_B I)^{-1}\|.$$

- 6.- Sea A una matriz simétrica. Para cada índice i existe un valor propio λ de A tal que

$$|\lambda - a_{ii}| \leq \sqrt{\sum_{j \neq i} |a_{ij}|^2}.$$

Indicación.- Utilizar el ejercicio 5 con una matriz B conveniente.

V.2 Determinación de Valores Propios

En la actualidad, existen muchos métodos numéricos para el cálculo de valores propios de una matriz. Sin embargo, la mayor parte de las aplicaciones en física y otras disciplinas requieren en la mayoría de los casos, la utilización de matrices normales. Motivo por el cual, existen métodos específicos a este tipo de matrices. Se comenzará, esta sección formulando el método de la Potencia.

El Método de la Potencia

Sea, A una matriz de $n \times n$ con coeficientes reales o complejos, el objetivo es determinar el valor propio más grande en valor absoluto. Sea $y_0 \in \mathbb{R}^n$ (\mathbb{C}^n) arbitrario, se define la sucesión $\{y_n\} \subset \mathbb{R}^n$ de manera recursiva, como

$$y_{n+1} = Ay_n, \quad (\text{V.2.1})$$

es evidente que $y_n = A^n y_0$. Esta sucesión tiene propiedades interesantes para la determinación del valor propio más grande, que están dadas en el:

Teorema V.2.1.- *Sea A diagonalizable, es decir que existe una matriz T inversible tal que*

$$T^{-1}AT = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix},$$

supóngase que los valores propios satisfacen

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|$$

y $e_1^ T^{-1} y_0 \neq 0$. Entonces la sucesión $\{y_k\}$ definida por $y_{k+1} = Ay_k$ satisface:*

$$\begin{aligned} \text{a) } y_k &= \lambda_1^k \left[a_1 v_1 + \mathcal{O} \left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right], \\ \text{donde } Av_1 &= \lambda_1 v_1, \quad Av_j = \lambda_j v_j, \\ y_0 &= a_1 v_1 + a_2 v_2 + \dots + a_n v_n, \text{ y} \\ T &= (v_1, \dots, v_n). \end{aligned}$$

b) Se tiene el cociente de Rayleigh, dado por

$$\frac{y_k^* A y_k}{y_k^* y_k} = \lambda_1 + \mathcal{O} \left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right), \quad (\text{V.2.2})$$

si además, la matriz A es normal, entonces

$$\frac{y_k^* A y_k}{y_k^* y_k} = \lambda_1 + \mathcal{O} \left(\left| \frac{\lambda_2}{\lambda_1} \right|^{2k} \right). \quad (\text{V.2.3})$$

Demostración.- Los vectores v_1, \dots, v_n forman una base del espacio \mathbb{C}^n , de donde

$$y_0 = a_1 v_1 + a_2 v_2 + \dots + a_n v_n,$$

deduciendose

$$y_k = a_1 \lambda_1^k v_1 + a_2 \lambda_2^k v_2 + \dots + a_n \lambda_n^k v_n,$$

por consiguiente

$$\frac{y_k}{\lambda_1^k} = a_1 v_1 + a_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k v_2 + \dots + a_n \left(\frac{\lambda_n}{\lambda_1} \right)^k v_n,$$

con lo queda demostrado el punto a). Para la demostración del punto b), se tiene:

$$\begin{aligned} y_k &= \lambda_1^k a_1 v_1 + \lambda_2^k a_2 v_2 + \dots + \lambda_n^k a_n v_n, \\ Ay_k &= \lambda_1^{k+1} a_1 v_1 + \lambda_2^{k+1} a_2 v_2 + \dots + \lambda_n^{k+1} a_n v_n, \\ y_k^* y_k &= \sum_{i,j} \bar{\lambda}_i^k \lambda_j^k v_i^* v_j \bar{a}_i a_j, \\ y_k^* Ay_k &= \sum_{i,j} \bar{\lambda}_i^k \lambda_j^{k+1} v_i^* v_j \bar{a}_i a_j, \end{aligned}$$

obteniendo el cociente de Rayleigh, dado por

$$\frac{y_k^* Ay_k}{y_k^* y_k} = \lambda_1 + \mathcal{O} \left(\left| \frac{\lambda_1}{\lambda_2} \right|^2 \right),$$

ahora bien si A es normal, se tiene que $v_i^* v_j = 0$, si $i \neq j$; obteniendo para

$$\begin{aligned} y_k^* Ay_k &= \sum_i \lambda_i |\lambda_i|^{2k} v_i^* v_i, \\ y_k^* y_k &= \sum_i |\lambda_i|^{2k} v_i^* v_i. \end{aligned}$$

□

Ejemplo

Considérese, la matriz A definida por

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix},$$

utilizando el ejercicio 2 de la sección V.1, se obtiene que el valor propio mas grande está dado por

$$\lambda_1 = 2(1 + \frac{\sqrt{3}}{2}) \approx 3,73205,$$

Tomando $y_0 = (1, 1, 1, 1, 1)^t$, se obtiene los valores de λ_1 dadas en la tabla V.2.1.

Tabla V.2.1. Valores de λ_1 .

Iter.	λ_1	Iter.	λ_1
1	3.6	2	3.696969
3	3.721854	4	3.729110
5	3.731205	6	3.731808
7	3.731981	8	3.732031
9	3.732045	10	3.732049
11	3.732050	12	3.732051
13	3.732051	14	3.732051

Las componentes del valor propio respecto a λ_1 están dados por

$$v = \begin{pmatrix} 1.07735026 \\ 1.86602540 \\ 2.1547005 \\ 1.866025 \\ 1.07735026 \end{pmatrix}.$$

Uno de los inconvenientes de utilizar el método de la potencia es que la convergencia puede ser muy lenta si $|\lambda_1/\lambda_2| \approx 1$. Sin embargo, existe una modificación del método de la potencia para acelerar la convergencia. Este método es más conocido como el:

Método de la Potencia Inversa

La modificación consiste en aplicar el método de la potencia a

$$(\mu I - A)^{-1},$$

donde μ es una aproximación del valor propio λ buscado de la matriz A . La justificación teórica está dada por la siguiente proposición:

Proposición V.2.2.- λ es valor propio de A , si y solamente si,

$$\frac{1}{\mu - \lambda} \quad (V.2.4)$$

es valor propio de $(\mu I - A)^{-1}$.

Demostración.- λ es valor propio de A , si y solamente si existe $v \neq 0$ tal que

$$Av = \lambda v,$$

si y solamente si

$$\begin{aligned} (\mu I - A)v &= (\mu - \lambda)v, \\ (\mu I - A)^{-1}v &= \frac{1}{\mu - \lambda}v. \end{aligned}$$

□

Por otro lado aplicando, el teorema V.2.1 se tiene que la convergencia es del orden

$$\mathcal{O}\left(\left|\frac{\mu - \lambda_1}{\mu - \lambda_2}\right|^k\right).$$

Sea $\bar{\lambda}$ el valor propio más grande de $(\mu I - A)^{-1}$, entonces se tiene que

$$\lambda_1 = \mu - \frac{1}{\bar{\lambda}}. \quad (\text{V.2.5})$$

El método de la potencia da una relación recursiva para los y_k , que está dada por

$$y_{k+1} = (\mu I - A)^{-1}y_k,$$

pero en lugar de calcular la inversa de la matriz, se puede resolver la ecuación

$$(\mu I - A)y_{k+1} = y_k, \quad (\text{V.2.6})$$

utilizando para tal efecto el algoritmo de eliminación de Gauss.

En el anterior ejemplo se tenía como valor de $\lambda_1 = 3.697$ después de 2 iteraciones del método de la potencia inversa. Aplicando el método de la potencia inversa con $\mu = 3.697$, se obtiene después de dos iteraciones:

$$\begin{aligned} \bar{\lambda} &= -232.83, \\ \lambda_1 &= 3.7334052. \end{aligned}$$

Puede suceder que la matriz A sea a coeficientes reales, sin embargo, el valor propio buscado no sea real si no que contenga una parte imaginaria. Supóngase que $\mu = \alpha + \beta i$ sea una aproximación del valor propio buscado. Entonces aplicando el método de la potencia inversa, se tiene

$$((\alpha + \beta i)I - A)y_{k+1} = y_k, \quad (\text{V.2.7})$$

por otro lado los vectores y_k pueden ser descompuestos en un vector real y otro imaginario, de la manera siguiente

$$y_k = u_k + iv_k, \quad u_k, v_k \in \mathbb{R}^n,$$

de donde, se obtiene una nueva formulación para el método de la potencia inversa, dada por

$$\begin{pmatrix} \alpha I - A & -\beta I \\ \beta I & \alpha I - A \end{pmatrix} \begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} u_k \\ v_k \end{pmatrix},$$

y el cociente de Rayleigh está dado por

$$\begin{aligned} \frac{y_k^* y_{k+1}}{y_k^* y_k} &= \frac{(u_k^t - iv_k^t)(u_{k+1} + iv_{k+1})}{u_k^t u_k + v_k^t v_k} \\ &= \frac{(u_k^t u_{k+1} + v_k^t v_{k+1} + i(u_k^t v_{k+1} - v_k^t u_{k+1}))}{u_k^t u_k + v_k^t v_k}. \end{aligned} \quad (\text{V.2.8})$$

Formas Tridiagonales y Matrices de Hessenberg

El método de la potencia y su versión de la potencia inversa son métodos aplicables a matrices, donde el valor propio más grande en valor absoluto lo es estrictamente. Por otro lado es necesario darse un vector inicial cuya proyección sobre el espacio propio, respecto a este valor propio sea no nula, es decir se debe tener una idea clara de la posible ubicación de los vectores propios. Ahora bien, en muchos casos es necesario conocer los diferentes valores propios, situación que no es posible con las dos versiones del método de la potencia estudiados más arriba.

Una gran clase de métodos de cálculo de valores propios están diseñados para operar con matrices tridiagonales, sobre todo si éstas son simétricas. El objetivo de este parágrafo es construir algoritmos que permitan tridiagonalizar una matriz arbitraria, conservando las propiedades originales en los casos en que sea posible.

Sea, A una matriz arbitraria de orden $n \times n$, se busca una transformación, T tal que

$$T^{-1}AT = H = \begin{pmatrix} * & & \cdots & & * \\ * & & & & \vdots \\ & \ddots & & & \vdots \\ & & \ddots & & \vdots \\ & & & * & * \end{pmatrix}.$$

H es conocida bajo el nombre de matriz de Hessenberg. A continuación, se propone como algoritmo de reducción a la forma de Hessenberg, utilizando matrices de tipo L , del algoritmo de eliminación de Gauss.

Algoritmo 1.

- 1.- Sea A una matriz arbitraria,
 se busca $|a_{k1}| = \max_{j=2, \dots, n} |a_{j1}|$,
 se intercambia las filas k y 2 y también las columnas k y 2,
 se define la matriz L_2 por

$$L_2 = \begin{pmatrix} 1 & & & & 0 \\ 0 & 1 & & & \\ 0 & -l_{32} & 1 & & \\ & \vdots & & \ddots & \\ 0 & -l_{n2} & & & 1 \end{pmatrix}, \quad l_{i2} = \frac{a_{i1}}{a_{21}}, \quad i = 2, \dots, n;$$

se obtiene

$$L_2 A L_2^{-1} = \begin{pmatrix} * & * & \cdots & * \\ * & & & \\ 0 & & A^{(1)} & \\ \vdots & & & \\ 0 & & & \end{pmatrix}.$$

- 2.- Se aplica el paso 1 a la matriz $A^{(1)}$, y se continua hasta obtener una matriz de Hessenberg.

La principal desventaja de utilizar este primer algoritmo propuesto es que, si A es una matriz simétrica, H no lo es en general. Recordando el capítulo II.5, existe el algoritmo QR para reducir una matriz A a la forma triangular. Es fácil verificar que $H = Q^t A Q$ es simétrica, si Q es ortogonal y A es simétrica. Por consiguiente, el segundo algoritmo propuesto utiliza matrices de Householder para convertir una matriz A en una matriz de Hessenberg.

Algoritmo 2.

- 1.- Sea A una matriz arbitraria, que puede escribirse, como

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ A'_1 & \cdots & A'_n \end{pmatrix},$$

por consiguiente $A'_k \in \mathbb{R}^{n-1}$.

Se define Q_2 por

$$Q_2 = \begin{pmatrix} 1 & 0 \\ 0 & Q'_2 \end{pmatrix},$$

donde $Q'_2 = I - 2u_2 u_2^t$, u_2 está determinado por la condición, ver Capítulo V.2,

$$Q'_2 A'_1 = \begin{pmatrix} \alpha_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \alpha_2 = \text{signo } a_{21} \|A'_1\|_2;$$

obteniendo

$$Q_2 A Q_2^{-1} = \begin{pmatrix} * & * & \cdots & * \\ * & & & \\ 0 & & A^{(1)} & \\ \vdots & & & \\ 0 & & & \end{pmatrix}.$$

2.- Se procede como en 1, hasta obtener la forma de Hessenberg.

Teorema de Sturm y Método de la Bisección

Al utilizar el algoritmo 2, de la anterior subsección, a una matriz A simétrica se obtiene una matriz tridiagonal simétrica, que se la denota nuevamente por A , para no cargar la notación. Por consiguiente, A es de la forma

$$A = \begin{pmatrix} d_1 & e_2 & & & \\ e_2 & d_2 & e_3 & & \\ & e_3 & \ddots & \ddots & \\ & & \ddots & \ddots & e_n \\ & & & e_n & d_n \end{pmatrix}. \quad (\text{V.2.9})$$

Se define, el polinomio $p_0(x)$ por

$$p_0(x) = 1, \quad (\text{V.2.10})$$

y los polinomios $p_i(x)$ $i = 1, \dots, n$, como

$$p_i(x) = \det(A_i - xI), \quad (\text{V.2.11})$$

donde

$$A = \begin{pmatrix} A_i & 0 \\ 0 & * \end{pmatrix},$$

A_i es la matriz formada por las primeras i filas y columnas de A . Por lo tanto, se tiene

$$P_n(x) = \det(A - xI). \quad (\text{V.2.12})$$

Por otro lado, los determinantes que definen estos polinomios cumplen la siguiente relación recursiva

$$\det(A_i - xI) = (d_i - x) \det(A_{i-1} - xI) - e_i^2 \det(A_{i-2} - xI), \quad i = 2, 3, \dots, n;$$

de donde

$$p_i(x) = (d_i - x)p_{i-1}(x) - e_i^2 p_{i-2}(x), \quad i = 2, 3, \dots, n. \quad (\text{V.2.13})$$

Teorema V.2.3.- Sea A una matriz tridiagonal y simétrica con los coeficientes $e_i \neq 0$ para $i = 2, \dots, n$. Entonces:

- a) Las raíces de $p_n(x) = \chi_A(x)$ son simples,
- b) $p'_n(\lambda_i) \cdot p_{n-1}(\lambda_i) < 0$ si λ_i es una raíz de $p_n(x)$,
- c) $p_j(x^*) = 0$ ($1 \leq j \leq n-1$, $x^* \in \mathbb{R}$) $\Rightarrow p_{j-1}(x^*)p_{j+1}(x^*) < 0$,
- d) $p_0(x) \geq 0$ para todo $x \in \mathbb{R}$.

Demostración.- El punto d) se verifica inmediatamente puesto que $p_0(x) = 1$ por definición.

La demostración del punto c) se la realiza por el absurdo, en efecto, si se tuviera $p_j(x^*) = 0$ y $p_{j+1}(x^*) = 0$, utilizando (V.2.13), se tendría $p_0(x^*) = 0$ lo cual contradice el hecho que $p_0(x^*) = 1$. Ahora bien, utilizando nuevamente (V.2.13) se tiene en el caso en que $p_j(x^*) = 0$

$$p_{j-1}(x^*) = -e_{n-j+1}^2 p_{j+1}(x^*).$$

El punto b) implica el punto a), ya que $p'_n(\lambda_i) \neq 0$ conduce a que λ_i sea raíz simple. Sólo queda el punto b) por demostrar.

Se define, los polinomios:

$$q_i(x) = (-1)^i p_i(x) \frac{1}{e_2 e_3 \cdots e_{i+1}}, \quad i = 1, \dots, n;$$

$$q_0(x) = p_0(x);$$

donde $e_{n+1} = 1$. Efectuando cálculos sobre los polinomios $q_j(x)$, se tiene

$$\begin{aligned} (-1)^i q_i(x) e_2 \cdots e_{i+1} &= (d_i - x)(-1)^{i-1} q_{i-1}(x) e_2 \cdots e_i \\ &\quad - e_i^2 (-1)^{i-2} q_{i-2}(x) e_2 \cdots e_{i-1}, \end{aligned}$$

de donde

$$e_{i+1} q_i(x) + (d_i - x) q_{i-1}(x) + e_i q_{i-2}(x) = 0, \quad i = 2, \dots, n.$$

Esta última relación puede escribirse de manera matricial, como

$$\underbrace{\begin{pmatrix} d_1 - x & e_2 & & & \\ e_2 & d_2 - x & e_3 & & \\ & e_3 & \ddots & \ddots & \\ & & \ddots & \ddots & e_n \\ & & & e_n & d_n - x \end{pmatrix}}_{A - xI} \underbrace{\begin{pmatrix} q_0(x) \\ q_1(x) \\ \vdots \\ q_{n-1}(x) \end{pmatrix}}_{q(x) \neq 0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -q_n(x) \end{pmatrix}.$$

Si λ_i es valor propio de A , entonces $q(\lambda_i)$ es un vector propio, derivando la relación matricial, se obtiene

$$-q(x) + (A - xI)q'(x) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -q'_n(x) \end{pmatrix},$$

multiplicando por la traspuesta de $q(x)$, se tiene

$$-q^t(x)q(x) + \underbrace{q^t(x)(A - xI)q'(x)}_{= 0 \text{ si } x = \lambda_i} = q^t(x) \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -q'_n(x) \end{pmatrix},$$

de donde

$$-q'_n(\lambda_i)q_{n-1}(\lambda_i) = -\|q(\lambda_i)\|^2 < 0,$$

dando por consiguiente

$$p'_n(\lambda_i)p_{n-1}(\lambda_i) < 0.$$

□

Cabe remarcar que los polinomios $p_i(x)$ $i = 0, \dots, n$; forman una sucesión de Sturm, cuyo estudio y cálculo de raíces están dados en la primera sección del capítulo IV concerniente a la resolución de ecuaciones polinomiales. Por consiguiente el número de valores propios de la matriz A en el intervalo $[a, b]$, está dado por

$$w(b) - w(a),$$

donde $w(x)$ indica los cambios de signo de los polinomios $p_i(x)$. La determinación exacta de los valores propios se efectúa mediante el algoritmo de la bisección, ya estudiado en la sección IV.1. Ahora bien, el problema consiste en encontrar un algoritmo simple que permita evaluar $w(x)$.

Algoritmo para calcular $w(x)$

Se define la sucesión $f_i(x)$, $i = 1, \dots, n$; por

$$f_i(x) = \frac{p_i(x)}{p_{i-1}(x)}, \quad (\text{V.2.14})$$

obteniendo de inmediato, la siguiente relación recursiva para $f_i(x)$:

$$\begin{aligned} f_1(x) &= (d_1 - x); \\ f_i(x) &= (d_i - x) - e_i^2 \frac{1}{f_{i-1}(x)}, \quad i = 2, \dots, n. \end{aligned} \quad (\text{V.2.15})$$

Ahora bien, si por azar, existe $x \in \mathbb{R}$, $i = 2, \dots, n$ con $f_i(x) = 0$, (V.2.15) puede ser modificado para evitar la división por 0, de la manera siguiente:

$$f_i(x) = \begin{cases} (d_i - x) - e_i^2 \frac{1}{f_{i-1}(x)}, & \text{si } f_{i-1}(x) \neq 0; \\ (d_i - x) - \frac{|e_i|}{\epsilon}, & \text{si } f_{i-1} = 0. \end{cases} \quad (\text{V.2.16})$$

La justificación de utilizar la sucesión $f_i(x)$ está en la siguiente proposición.

Proposición V.2.4.- $w(x)$ es igual al número de elementos negativos de

$$\{f_1(x), f_2(x), \dots, f_n(x)\}$$

Demostración.- La demostración tiene sus bases teóricas en el Teorema de Sylvester, que indica que si A es una matriz, T una matriz no singular, entonces la matriz $B = T^t A T$ tiene el mismo número de valores propios negativos que la matriz A . Aplicando este teorema a la proposición a demostrar, se tiene

$$A - xI = \begin{pmatrix} 1 & & & & \\ e_2/f_1(x) & 1 & & & \\ & \ddots & & & \\ & & e_n/f_{n-1} & 1 & \end{pmatrix} \begin{pmatrix} f_1(x) & & & & \\ & f_2(x) & & & \\ & & \ddots & & \\ & & & f_n(x) & \end{pmatrix} \begin{pmatrix} 1 & e_2/f_1(x) & & & \\ & \ddots & & & \\ & & 1 & e_n/f_{n-1}(x) & \\ & & & & 1 \end{pmatrix},$$

de donde $A - xI$ tiene el mismo número de valores propios negativos que el conjunto $\{f_1(x), f_2(x), \dots, f_n(x)\}$. Por otro lado este número está dado por $w(x)$, quedando demostrada la proposición. \square

Para poder aplicar el algoritmo de la bisección es necesario conocer los intervalos, donde pueden estar localizados los valores propios, para evitar búsquedas inútiles. El siguiente teorema es muy útil para determinar las regiones donde se encuentran estos.

Teorema V.2.5.- Gerschgorin. Sea λ un valor propio de A matriz, Entonces existe i tal que

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|. \quad (\text{V.2.17})$$

Demostración.- Sea $x \neq 0$, el vector propio asociado a λ , por consiguiente $x = (x_1, x_2, \dots, x_n)^t$. Sea i , tal que

$$|x_i| \geq |x_j| \quad \forall j,$$

puesto que $x \neq 0$, se tiene necesariamente que $x_i \neq 0$. Efectuando cálculos se obtiene:

$$\begin{aligned}\sum_{j=1}^n a_{ij}x_j &= \lambda x_i, \\ \sum_{j \neq i} a_{ij}x_j &= (\lambda - a_{ii})x_i,\end{aligned}$$

pasando a valores absolutos se tiene:

$$\begin{aligned}|\lambda - a_{ii}| |x_i| &\leq \sum_{j \neq i} |a_{ij}| |x_j|, \\ |\lambda - a_{ii}| &\leq \sum_{j \neq i} |a_{ij}|.\end{aligned}$$

□

Generalización del Método de la Potencia

Al formular el método de la potencia, se buscaba el valor propio, cuyo valor absoluto era máximo, de una matriz A . Ahora bien, existen muchos problemas en los cuales se requiere conocer no solamente el valor propio más grande en valor absoluto, sino también los otros valores propios. El propósito de esta subsección es formular un método que permita calcular los dos valores propios más grandes en módulo. Para tal efecto, es necesario suponer que $\lambda_1, \lambda_2, \dots, \lambda_n$ valores propios de A satisfacen

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \dots |\lambda_n|.$$

Recordando el método de la potencia, se tiene la sucesión de los $\{y_k\}$ definida por

$$y_{j+1} = Ay_j,$$

si y_0 cumple ciertas condiciones, se tiene

$$y_j = \lambda_1^j [a_1 v_1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^j\right)],$$

donde v_1 es el vector propio asociado a λ_1 y a_1 la componente de y_0 respecto a v_1 . Para evitar explosiones en los cálculos de los y_j se puede mejorar el algoritmo de la potencia exigiendo que $\|y_j\|_2 = 1$, y además para evitar oscilaciones en los y_j , se plantea por consiguiente

$$y_{j+1} = \frac{Ay_j}{\|Ay_j\|_2} \text{signo} \left(\frac{(Ay_j)_1}{(y_j)_1} \right), \quad (\text{V.2.18})$$

donde $(y_j)_1$ es la primera componente de y_j . Puesto que los y_j son de norma igual a 1, haciendo referencia a vectores ortonormales, se cambia la notación de y_j por q_j , dando por consiguiente la siguiente definición recursiva para los q_j :

$$\begin{aligned}
q_0 & \text{ arbitrario, tal que } \|q_0\|_2 = 1; \\
\|q_j\|_2 & = 1; \\
\lambda_1^{(j+1)} q_{j+1} & = Aq_j.
\end{aligned} \tag{V.2.19}$$

Por el teorema V.2.1, se tiene que:

$$\begin{aligned}
\lim_{j \rightarrow \infty} q_j & = v_1, \\
\lim_{j \rightarrow \infty} \lambda_1^{(j)} & = \lambda_1,
\end{aligned}$$

con v_1 valor propio de norma unitaria respecto a λ_1 . Por otro lado la velocidad de convergencia de (V.2.19), está dada por $\mathcal{O}((\lambda_2/\lambda_1)^j)$.

Por consiguiente, el problema consiste en calcular λ_1 y λ_2 , si es posible al mismo tiempo. Supóngase que conoce de antemano λ_1 y v_1 . Considerando el subespacio vectorial V de \mathbb{C}^n definido por

$$V = \{u \in \mathbb{C}^n \mid u^* v_1 = 0\},$$

espacio de dimensión $n - 1$. La matriz A induce una aplicación lineal que se la denota por la letra A también, la cual está definida por

$$\begin{aligned}
A : \mathbb{C}^n & \longrightarrow \mathbb{C}^n \\
y & \longrightarrow Ay,
\end{aligned}$$

definiendo la aplicación lineal $f : V \longrightarrow V$ como $f = p \circ A|_V$ donde p es la proyección ortogonal de \mathbb{C}^n en V , y $A|_V$ es la restricción de A sobre el espacio V se tiene el:

Teorema V.2.6.- *Los valores propios de f son: $\lambda_2, \dots, \lambda_n$, más precisamente se tiene*

$$f(v) = U \begin{pmatrix} \lambda_1 & 0 & & 0 \\ & \lambda_2 & r_{22} & \cdots & r_{2n} \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_n \end{pmatrix} U^* v,$$

donde $v \in V$ y

$$U^* A U = \begin{pmatrix} \lambda_1 & r_{12} & \cdots & r_{1n} \\ & \ddots & & \\ & & \ddots & \\ & & & \ddots & \\ & & & & \lambda_n \end{pmatrix}$$

es la descomposición de Schur dada en teorema V.1.7.

Demostración.- Sea $U = (u_1, u_2, \dots, u_n)$ matriz unitaria con $u_1 = v_1$, entonces cualquier elemento $v \in V$ se escribe, como

$$v = \sum_{i=2}^n \alpha_i u_i,$$

por consiguiente

$$v = U\alpha, \quad \text{con } \alpha = \begin{pmatrix} 0 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix};$$

de donde $f(v) = AU\alpha$.

Ahora bien, si U es la matriz unitaria de la descomposición de Schur de A , se tiene

$$f(v) = U \begin{pmatrix} \lambda_1 & r_{12} & \cdots & r_{1n} \\ & \ddots & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} U^* U \alpha,$$

como $U\alpha = v$, se tiene después de una simple verificación que

$$f(v) = U \begin{pmatrix} \lambda_1 & 0 & & 0 \\ & \lambda_2 & r_{22} & \cdots & r_{2n} \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_n \end{pmatrix} U^* v.$$

□

El teorema precedente proporciona un medio teórico para calcular λ_2 , el cual es el método de la potencia a f . Por lo tanto, si se tiene determinado λ_1 y v_1 el algoritmo de la potencia, se lo formula de la siguiente manera:

$$\begin{aligned} r_0, \quad & \text{tal que } v_1^* r_0 = 0 \text{ y } \|r_0\|_2 = 1; \\ \alpha_j &= v_1^* A r_j; \\ \|r_j\|_2 &= 1; \\ A r_j - \alpha_j v_1 &= \lambda_2^{(j+1)} r_{j+1}. \end{aligned} \tag{V.2.20}$$

Es facil verificar que

$$\begin{aligned} \lim_{j \rightarrow \infty} \lambda_2^{(j)} &= \lambda_2, \\ \lim_{j \rightarrow \infty} r_j &= u_2; \end{aligned}$$

donde v_2 es el vector propio de norma unitaria respecto a λ_2 .

Sería interesante poder calcular λ_1 y λ_2 al mismo tiempo, sin necesidad de determinar primero λ_1 y luego λ_2 , esto es posible mediante el siguiente algoritmo.

Se da, como vectores iniciales: q_0 y r_0 , tales que:

$$\|q_0\| = 1, \quad \|r_0\| = 1, \quad \text{y } r_0^* q_0 = 0. \quad (\text{V.2.21})$$

Se supone, que se ha calculado

$$q_j, \text{ con } \|q_j\| = 1; \quad r_j, \text{ con } \|r_j\| = 1 \text{ y } r_j^* q_j = 0;$$

entonces se aplica el algoritmo de la potencia de la siguiente manera

$$Aq_j = \lambda_1^{(j+1)} q_{j+1}, \quad \begin{cases} \alpha_{j+1} = q_{j+1}^* A r_j, \\ A r_j - \alpha_{j+1} q_{j+1} = \lambda_2^{j+1} r_{j+1}, \\ \|q_{j+1}\| = 1; \\ \|r_{j+1}\| = 1. \end{cases} \quad (\text{V.2.22})$$

Se puede demostrar que también

$$\lim_{j \rightarrow \infty} \lambda_2^{(j)} = \lambda_2, \\ \lim_{j \rightarrow \infty} r_j = u_2,$$

donde u_2 es la segunda columna de la matriz U de la descomposición de Schur de A . Vale la pena recalcar, el siguiente hecho

$$A \underbrace{(q_j, r_j)}_{U_j} = \underbrace{(q_{j+1}, r_{j+1})}_{U_{j+1}} \underbrace{\begin{pmatrix} \lambda_1^{(j+1)} & \alpha^{(j+1)} \\ 0 & \lambda_2^{(j+1)} \end{pmatrix}}_{R_{j+1}},$$

de donde planteando $U_0 = (q_0, r_0)$ con $U_0^* U_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, el algoritmo de la potencia generalizada, se expresa de manera matricial como

$$A U_j = U_{j+1} R_{j+1},$$

donde el segundo miembro no es nada más que la descomposición QR , pero esta vez utilizando matrices unitarias.

Si la matriz A tiene sus n valores propios diferentes, y además que verifican

$$|\lambda_i| > |\lambda_2| > \cdots > |\lambda_n|,$$

el método de la potencia puede ser aplicado para calcular de manera simultánea los n autovalores. Se toma U_0 una matriz unitaria arbitraria que puede ser por ejemplo $U_0 = I$. Supóngase que se ha calculado R_j y U_j , entonces se tiene

$$AU_j = U_{j+1}R_{j+1}. \quad (\text{V.2.23})$$

Se puede demostrar que:

$$\begin{aligned} R_j &\longrightarrow R = \begin{pmatrix} \lambda_1 & \cdots & * \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \\ U_j &\longrightarrow U, \end{aligned}$$

cuando $j \rightarrow \infty$, donde $AU = UR$ es la descomposición de Schur.

El Método QR

Al finalizar la última subsección, se dio una generalización del método de la potencia. Por razones de presentación, se verá en esta sección un algoritmo equivalente a este último, conocido como el método QR .

La versión simple del algoritmo QR es aplicable a matrices, cuyos valores propios son reales y diferentes en valor absoluto, se lo define recursivamente de la siguiente manera:

$$\begin{aligned} A_1 &= A = Q_1 R_1, \\ A_{j+1} &= R_j Q_j = Q_{j+1} R_{j+1}, \end{aligned} \quad (\text{V.2.24})$$

donde Q_j es una matriz ortogonal, y R es una matriz triangular superior. De donde, este algoritmo es equivalente al método de la potencia generalizada. En efecto, planteando

$$U_j = Q_1 \cdots Q_j,$$

se tiene

$$U_{j+1} R_{j+1} = Q_1 Q_2 \cdots Q_j \underbrace{Q_{j+1} R_{j+1}}_{R_j Q_j},$$

llegando finalmente a

$$U_{j+1} R_{j+1} = \underbrace{Q_1 R_1}_A \underbrace{Q_1 Q_2 \cdots Q_j}_{U_j}.$$

Puesto que el método QR y el algoritmo de la potencia son equivalentes, es fácil ver que

$$\begin{aligned} R_j &\longrightarrow R = \begin{pmatrix} \lambda_1 & \cdots & * \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \\ Q_j &\longrightarrow I. \end{aligned}$$

Por otro lado, las matrices A_{j+1} y A_j tienen los mismos valores propios, pues

$$A_{j+1} = Q_j^* A_j Q_j$$

Indudablemente, el método QR es aplicable a toda matriz cuyos valores propios son diferentes en módulo, sin embargo es conveniente reducir la matriz A a una de tipo Hessenberg, ya que si A es una matriz de Hessenberg, las matrices A_k construidas a partir del algoritmo QR lo son también, ver ejercicio 5, además el costo de operaciones es menor.

El siguiente resultado, enunciado sin demostración, indica la velocidad de convergencia del método QR donde A es una matriz de tipo Hessenberg. Esta relación esta dada por

$$\frac{a_{n,n-1}^{(j+1)}}{a_{n,n-1}^{(j)}} \sim \left(\frac{\lambda_n}{\lambda_{n-1}} \right), \quad (\text{V.2.25})$$

es decir

$$a_{n,n-1}^{(j)} = \mathcal{O} \left(\left(\frac{\lambda_n}{\lambda_{n-1}} \right)^j \right). \quad (\text{V.2.26})$$

Uno de los problemas con que se confronta es que la convergencia puede ser demasiado lenta, sobre todo si $|\lambda_n| \approx |\lambda_{n-1}|$; pero este inconveniente puede superarse aplicando el algoritmo QR , en lugar de la matriz A , a

$$A - pI, \quad (\text{V.2.27})$$

donde $p \approx \lambda_n$. p se lo conoce con el nombre de *shift*. En este caso la velocidad de convergencia está dada por

$$a_{n,n-1}^{(j)} = \mathcal{O} \left(\left(\frac{\lambda_n - p}{\lambda_{n-1} - p} \right)^j \right). \quad (\text{V.2.28})$$

El algoritmo QR con *shift*

Antes de comenzar el algoritmo QR , se supone que la matriz A está expresada bajo la forma de una matriz de Hessenberg. El algoritmo QR con *shift*, se lo define de manera recursiva como:

$$\begin{aligned} A_1 &= A, \\ Q_j R_j &= A_j - p_j I, \\ A_{j+1} &= R_j Q_j + p_j I. \end{aligned} \quad (\text{V.2.29})$$

Las dos maneras más corrientes de elegir el *shift* p_k son:

1.- Se plantea $p_k = a_{n,n}^{(k)}$.

Este procedimiento funciona bien, si todos los valores propios de la matriz A son reales.

2.- Se toma p_k , al valor propio de la matriz

$$\begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{n,n}^{(k)} \end{pmatrix},$$

que es más próximo al coeficiente $a_{n,n}^{(k)}$.

Una interrogante muy importante surge, cuando detener el algoritmo QR . Uno de los criterios más usados es el siguiente. Si

$$\left| a_{n,n-1}^{(k)} \right| \leq eps \left(\left| a_{nn}^{(k)} \right| + \left| a_{n-1,n-1}^{(k)} \right| \right), \quad (\text{V.2.30})$$

se plantea

$$\lambda_n = a_{nn}^{(k)}. \quad (\text{V.2.31})$$

Luego se continua con la matriz $A^{(k)}$ de dimension $n - 1$ resultante, hasta obtener todos los valores propios de la matriz A .

La experiencia numérica muestra que el algoritmo QR converge linealmente, mientras que el algoritmo QR con *shift* tiene convergencia cuadrática.

Ejemplos

1.- Considérese, la matriz A definida por

$$A = \begin{pmatrix} 10 & 7 & 6 \\ 0.1 & 5 & 3 \\ 0 & 0.1 & 1 \end{pmatrix}$$

matriz de tipo Hessenberg. Por lo tanto lista, para ser aplicado el método QR y el método QR con *shift*. A continuación, se mostrará las iteraciones resultantes del método QR sin *shift*

	10.070493	7.0701525	5.8875209
$A_2 =$.49305211 - 001	4.9895287	2.8589253
	.00000000	.19067451 - 001	.93997833
	10.105227	7.0677632	5.8742925
$A_3 =$.24258892 - 001	4.9656988	2.8145741
	.00000000	.35752960 - 002	.92907381
	10.122222	7.0596308	5.8759333
$A_4 =$.11880000 - 001	4.9507267	2.7975584
	.00000000	.66976513 - 003	.92705007

Puede observarse que la convergencia es lineal y además se tiene

$$\frac{a_{21}^{(k+1)}}{a_{21}^{(k)}} \sim \frac{1}{2},$$

$$\frac{a_{32}^{(k+1)}}{a_{32}^{(k)}} \sim \frac{1}{5},$$

por consiguiente se necesitan por lo menos 10 iteraciones para obtener $a_{32}^{(k)} \sim 10^{-8}$ y por lo menos 23 iteraciones para tener $a_{21}^{(k)} \sim 10^{-8}$. Para poder observar la verdadera potencia de aplicar el método QR con *shift*, a continuación se muestran las 6 iteraciones que son necesarias, para obtener los valores propios de la matriz A .

$$p_1 = 1.0$$

	10.078261	7.0950933	5.8540158
$A_2 =$.43358029 - 001	4.9964769	2.8312527
	.00000000	-.19045643 - 002	.92526107

$$p_2 = .92526107$$

	10.112141	7.0679606	5.8707690
$A_3 =$.19226871 - 001	4.9612733	2.8312527
	.00000000	-.62453559 - 006	.92526107

$$p_3 = .92526107$$

	10.126953	7.0571468	5.8766286
$A_4 =$.84142592 - 002	4.9464609	2.7929532
	.00000000	-.67414063 - 013	.92658424

$$p_4 = .84142592 - 002$$

	10.138415	7.0571468	
$A_5 =$	-.18617346 - 004	4.9349986	

$$p_5 = -.18617346 - 004$$

	10.138390	7.0497319	
$A_6 =$	-.90446847 - 010	4.9350238	

obteniendo, así los siguientes valores propios:

$$\lambda_1 = 10.138390,$$

$$\lambda_2 = 4.9350238,$$

$$\lambda_3 = 0.92658424.$$

2.- Puede observarse en el ejemplo anterior que la convergencia del método QR con *shift* es cuadrática, para ilustrar este hecho, considérese

$$A = \begin{pmatrix} 2 & a \\ \epsilon & 1 \end{pmatrix},$$

con ϵ bastante pequeño. Aplicando el método QR con *shift* se tiene $p_1 = 1$ y

$$A - p_1 I = \begin{pmatrix} 1 & a \\ \epsilon & 0 \end{pmatrix},$$

obteniendo

$$Q_1 R_1 = \begin{pmatrix} 1/\sqrt{1+\epsilon^2} & -\epsilon/\sqrt{1+\epsilon^2} \\ \epsilon/\sqrt{1+\epsilon^2} & 1/\sqrt{1+\epsilon^2} \end{pmatrix} \begin{pmatrix} \sqrt{1+\epsilon^2} & a/\sqrt{1+\epsilon^2} \\ 0 & -\epsilon a/\sqrt{1+\epsilon^2} \end{pmatrix},$$

lo que da

$$A_1 = \begin{pmatrix} * & * \\ -\epsilon^2/1+\epsilon^2 & * \end{pmatrix},$$

de donde la convergencia es cuadrática, si a es arbitrario. Si la matriz es simétrica, se puede observar que la convergencia sería cúbica en este ejemplo, tomar $a = \epsilon$.

Ejercicios

1.- Calcular los valores propios y vectores propios de

$$A = \begin{pmatrix} 2 & 1 & & \\ 1 & 2 & 12 & \\ 0 & 1 & 2 & 1 \\ 1 & 2 & & \end{pmatrix},$$

utilizando el método de la potencia inversa de Wielandt.

2.- Sea

$$A = \begin{pmatrix} \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{pmatrix}.$$

Calcular α, β, λ y una matriz T a coeficientes reales, tal que

$$T^{-1} = \begin{pmatrix} \alpha & \beta & 0 \\ -\beta & \alpha & 0 \\ 0 & 0 & \lambda \end{pmatrix}.$$

La matriz A define un método de Runge-Kutta.

3.- Sea A una matriz de orden s que es diagonalizable, y sea J una matriz de orden n . Se define el producto tensorial $A \otimes J$ por

$$A \otimes J = \begin{pmatrix} a_{11}J & \cdots & a_{1s}J \\ \vdots & & \vdots \\ a_{s1}J & \cdots & a_{ss}J \end{pmatrix}.$$

Encontrar un algoritmo eficaz para resolver un sistema lineal con la matriz

$$I - A \otimes J.$$

Indicaciones:

a) Mostrar que

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD).$$

b) Supóngase que

$$T^{-1}AT = \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_s \end{pmatrix}$$

y utilizar la descomposición

$$I - A \otimes J = (T \otimes I)(I - \Lambda \otimes J)(T^{-1} \otimes I) \quad (\text{V2.32})$$

c) estimar el número de multiplicaciones necesarias, si:

— se calcula la descomposición LR de $I - A \otimes J$,

— se utiliza (V.2.32); el trabajo principal es el cálculo de la descomposición LR de $I - \Lambda \otimes J$.

4.- Calcular todos los valores propios de

$$A = \left(\begin{array}{cccc} 1 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 & -1 \\ & & & -1 & 3 \end{array} \right) \Bigg\}^n$$

para $n = 10$ y $n = 20$. Utilizar el método de la bisección.

5.- Mostrar que si la matriz $A = A_1$ es una matriz de Hessenberg, las matrices A_k , $k = 2$, construidas en el método QR son igualmente matrices de Hessenberg.

Capítulo VI

Integración Numérica

En muchos problemas aparecen integrales, ya sean éstas simples, múltiples y de otras clases. En los cursos elementales de Cálculo, la determinación de primitivas es un tópico de bastante importancia. Por consiguiente, existen los elementos teóricos para evaluar primitivas. Sin embargo la mayor parte de las aplicaciones numéricas donde interviene la integración, presenta integrales cuyo cálculo de primitivas es imposible en términos de funciones elementales, o por las características propias de los problemas solo se requiere una cómputo aproximado de éstas.

En este capítulo se tratará las base teóricas de la integral de Riemann, luego se abordará la noción de fórmula de cuadratura y como consecuencia lógica se definirá el orden de una fórmula de cuadratura. El segundo tema que será estudiado, como parte integrante del Análisis Numérico, está relacionado con la estimación del error cometido por la utilización de métodos numéricos de integración. Se comparará, diferentes fórmulas de cuadratura haciendo hincapié en aquéllas de orden elevado. Las fórmulas de cuadratura de Gauss serán estudiadas como un caso de fórmulas de orden elevado y para comprenderlas mejor se introducirá la noción de polinomios ortogonales. Después, se formulará un método adaptativo para determinar integrales y como corolario se hará un tratamiento de singularidades, es decir se implementará métodos para resolver integrales impropias. Como último tópico, se verá la interpolación trigonométrica, es decir se hará un estudio de las Transformadas de Fourier discretas, como también la Transformada de Fourier Rápida, mas conocida como FTP.

VI.1 Bases Teóricas

En esta sección, se verán las bases teóricas de la teoría de la integración, pero ésta limitada a la integral de Riemann. Por otro lado, también serán abordados las motivaciones para tener la integral como un instrumento matemático de Cálculo.

Uno de los problemas con el cual el hombre ha confortado desde épocas lejanas, ha sido el cálculo de áreas, volúmenes y otros, tanto por la necesidad cotidiana, como también dentro el espíritu de reflexión e investigación que lo ha caracterizado siempre.

La integral más utilizada, desde el punto de vista de cálculo y práctico, es la integral de Riemann, cuya definición permitió que el cálculo integral tuviese bases teóricas cimentadas. En lo que sigue, se hará una presentación teórica de esta importante noción.

Definición VI.1.1.- Sea, $[a, b]$ un intervalo compacto de la recta real. Se llama subdivisión del intervalo $[a, b]$ un conjunto $\mathcal{S} \subset [a, b]$ finito.

Por consiguiente \mathcal{S} puede ser ordenado, como una sucesión finita de puntos de $[a, b]$ expresada de la manera siguiente

$$\mathcal{S} = \{x_0 < x_1 < \cdots < x_n\} \subset [a, b].$$

Definición VI.1.2.- Sea \mathcal{S} una subdivisión de $[a, b]$, se llama paso de la subdivisión \mathcal{S} a

$$\delta(\mathcal{S}) = \max_{i=1, \dots, n} |x_i - x_{i-1}|.$$

Ahora bien, la noción de integral está definida para funciones cuyo dominio son intervalos compactos, además se exige que la función sea acotada. Como no es propósito del libro hacer una exposición sobre la teoría de la integración, se dará una de las definiciones de una función integrable en el sentido de Riemann.

Definición VI.1.3.- Sea $f : [a, b] \rightarrow \mathbb{R}$ una función acotada sobre un intervalo compacto. Se dira que f es Riemann integrable si

$$\lim_{\delta(\mathcal{S}) \rightarrow 0} \sum_{i=1}^n f(\xi_i)(x_i - x_{i-1}) \text{ existe,}$$

con $\xi_i \in [x_{i-1}, x_i]$, independientemente de \mathcal{S} y de los ξ_i . Si éste es el caso, este límite se lo denota por

$$\int_a^b f(x)dx.$$

Se puede demostrar que las funciones monótonas, continuas y en escalera son Riemann integrables. El cálculo de la integral como un límite puede ser bastante tedioso, motivo por el cual, los cursos de Cálculo en los niveles básicos universitarios y últimos cursos de colegio, dan un énfasis al cálculo de primitivas, que para recordar es:

Definición VI.1.4.- Dada una función φ sobre un intervalo I de \mathbb{R} . Se dice que una función $\phi : I \rightarrow \mathbb{R}$ es una primitiva de φ si, en todo punto x de I , la función ϕ es derivable y si $\phi'(x) = \varphi(x)$.

Sea $f : [a, b] \rightarrow \mathbb{R}$ una función integrable en el sentido de Riemann, se define para todo $x \in [a, b]$

$$F(x) = \int_a^x f(t)dt.$$

Proposición VI.1.5.- La función F es continua, además si la función f es continua en un punto x de $[a, b]$, la función F es derivable en x y $F'(x) = f(x)$.

Corolario VI.1.6.- Una función continua sobre un intervalo compacto admite una primitiva.

Las demostraciones de la proposición y el corolario precedentes se las puede encontrar en cualquier libro de Análisis. A continuación, se enuncia uno de los teoremas fundamentales del cálculo.

Teorema VI.1.7.- Sea $f : [a, b] \rightarrow \mathbb{R}$ una función integrable en el sentido de Riemann que, además admite una primitiva G . Para todo $x \in [a, b]$, se tiene:

$$G(x) - G(a) = \int_a^x f(t)dt.$$

Ahora bien, desde el punto de vista numérico, el cálculo de primitivas no tiene mayor utilidad práctica, pues en la mayor parte de los casos la determinación analítica de éstas es costosa en tiempo. El tratamiento numérico que se realiza en el cálculo de integrales está basado en la misma definición de integral. Por consiguiente utilizando la definición, se tiene que $\forall \epsilon > 0$, existe $\mathcal{S} \subset [a, b]$ y $\xi_i \in [x_{i-1}, x_i]$ tal que

$$\left| \sum_{i=1}^n f(\xi_i)(x_i - x_{i-1}) - \int_a^b f(x)dx \right| < \epsilon,$$

de donde el enfoque numérico consiste en aproximar la integral utilizando, sumas finitas de Riemann. Por otro lado, una de las tareas del Análisis Numérico en lo que respecta el cálculo de integrales, está dada en la construcción de fórmulas de cuadratura con un costo razonable en operaciones, y una precisión aceptable en la determinación de la integral.

A continuación, se mostrará las fórmulas de cuadratura o de integración mas rudimentarias.

a) Regla del Punto Medio

Consiste en utilizar la suma de Riemann con $\xi_i = \frac{x_{i-1} + x_i}{2}$, por consiguiente

$$\sum_{i=1}^n f\left(\frac{x_{i-1} + x_i}{2}\right) (x_i - x_{i-1}) \approx \int_a^b f(x) dx.$$

La regla del punto medio, da resultados exactos cuando f es un polinomio de grado menor o igual a 1. Ver figura VI.1.1.

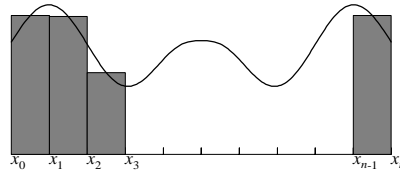


Figura VI.1.1. Regla del Punto Medio.

b) Regla del Trapecio

Consiste en aproximar $f(\xi_i)(x_{i-1} - x_i)$ con el area de un trapecio de alturas $f(x_{i-1})$ y $f(x_i)$. Por consiguiente

$$\sum_{i=1}^n \frac{f(x_{i-1}) + f(x_i)}{2} (x_i - x_{i-1}) \approx \int_a^b f(x) dx,$$

de donde, esta suma puede expresarse de manera más simple, como

$$\int_a^b f(x) dx \approx f(x_0) \frac{x_1 - x_0}{2} + f(x_1) \frac{x_2 - x_1}{2} + \dots \\ \dots + f(x_{n-1}) \frac{x_n - x_{n-2}}{2} + f(x_n) \frac{x_n - x_{n-1}}{2}.$$

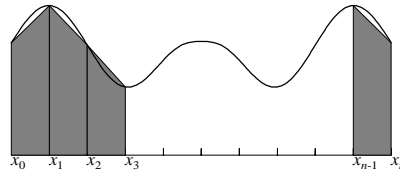


Figura VI.1.2. Regla del Trapecio.

Esta fórmula es exacta para polinomios de grado igual o inferior a 1. Ver figura VI.1.2.

c) Regla de Simpson

Consiste en aproximar $f(\xi_i)(x_{i-1} - x_i)$, con el área de la superficie cuyo lado superior, ver figura VI.1.3, esta dada por la parábola que pasa por los puntos $(x_{i-1}, f(x_{i-1}))$, $(\frac{x_{i-1} + x_i}{2}, f(\frac{x_{i-1} + x_i}{2}))$ y $(x_i, f(x_i))$. Para simplificar los cálculos se toma para x_0 , y x_1 , obteniendo como polinomio de interpolación

$$p(x) = f(x_0) + 2 \frac{f(x_0) + f((x_0 + x_1)/2)}{x_1 - x_0} (x - x_0) + 2 \frac{f(x_0) - 2f((x_0 + x_1)/2) + f(x_1)}{(x_1 - x_0)^2} (x - (x_0 + x_1)/2)(x - x_1).$$

Ahora bien, integrando este polinomio de segundo grado entre x_0 y x_1 , se obtiene

$$\int_{x_0}^{x_1} f(x) dx \approx \left(\frac{1}{6} f(x_0) + \frac{4}{6} f\left(\frac{x_0 + x_1}{2}\right) + \frac{1}{6} f(x_1) \right) h_1,$$

para finalmente tener

$$\int_a^b f(x) dx \approx \sum_{i=1}^n \left(\frac{1}{6} f(x_{i-1}) + \frac{4}{6} f\left(x_{i-1} + \frac{h_i}{2}\right) + \frac{1}{6} f(x_i) \right) h_i.$$

La fórmula de Simpson es exacta para polinomios de grado igual o inferior a 3.

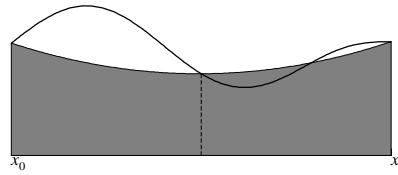


Figura VI.1.3. Regla de Simpson.

d) Regla de Newton

Consiste en aproximar $f(\xi_i)(x_{i-1} - x_i)$, con el área de la superficie cuyo lado superior, ver figura VI.1.4, esta dada por la parábola cúbica que pasa por los puntos $(x_{i-1}, f(x_{i-1}))$, $(\frac{2x_{i-1} + x_i}{3}, f(\frac{2x_{i-1} + x_i}{3}))$,

$(\frac{x_{i-1} + 2x_i}{3}, f(\frac{x_{i-1} + 2x_i}{3}))$ y $(x_i, f(x_i))$. De la misma manera que en la regla de Simpson, se calcula esta parábola cúbica utilizando por ejemplo la fórmula de Newton para determinar polinomios de interpolación, luego se integra para obtener

$$\int_{x_0}^{x_1} f(x)dx \approx \left[\frac{1}{8}f(x_0) + \frac{3}{8}f\left(x_0 + \frac{h_1}{3}\right) + \frac{3}{8}f\left(x_0 + \frac{2h_1}{3}\right) + \frac{1}{8}f(x_1) \right] h_1.$$

La regla de Newton es exacta para polinomios de grado menor o igual a 3.

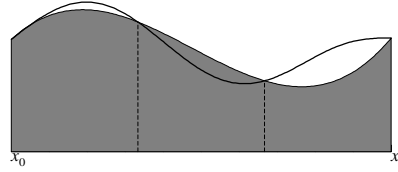


Figura VI.1.4. Regla de Newton.

Fórmulas de Cuadratura

En los cuatro ejemplos precedentes de la anterior subsección, puede observarse claramente que las reglas de integración formuladas tienen la misma estructura, la cual define tácitamente una fórmula de cuadratura. La mayor parte de los métodos numéricos están basados en este principio.

Se desea integrar la función $f : [a, b] \rightarrow \mathbb{R}$, donde f es Riemann integrable. Para tal efecto se considera una subdivisión

$$\mathcal{S} = \{a = x_0 < x_1 < \dots < x_n = b\}.$$

La integral puede ser aproximada mediante la fórmula de cuadratura siguiente

$$\sum_{j=1}^n \left(\sum_{i=1}^s b_i f(x_{j-1} + c_i h_j) \right), \quad (\text{VI.1.1})$$

los c_i se llaman nudos de la fórmula de cuadratura y los b_i son los coeficientes de la fórmula de cuadratura.

Para la regla del Trapecio, se tiene $s = 2$, $b_1 = b_2 = 1/2$ y $c_1 = 0$, $c_2 = 1$. Así mismo, para la regla de Simpson se tiene $s = 3$, $b_1 = b_3 = 1/6$, $b_2 = 4/6$ y $c_1 = 0$, $c_2 = 1/2$, $c_3 = 1$.

Dada una fórmula de cuadratura, uno de los objetivos principales es medir la precisión de ésta. Por razones de simplicidad, es preferible estudiar

la fórmula de cuadratura en una función $f : [0, 1] \rightarrow \mathbb{R}$ y considerar $h_1 = 1$, es decir integrar numéricamente con un paso de integración. Por consiguiente, se analizará el problema

$$\sum_{i=0}^s b_i f(c_i) \approx \int_0^1 f(x) dx.$$

El Orden de una Fórmula de Cuadratura

Definición VI.1.8.- Una fórmula de cuadratura tiene orden p , si y solamente si, p es el entero más grande, tal que

$$\sum_{i=0}^s b_i f(c_i) = \int_0^1 f(x) dx, \quad (\text{VI.1.2})$$

donde f es un polinomio de grado $\leq p - 1$.

Proposición VI.1.9.- Una fórmula de cuadratura es de orden p , si y solamente si

$$\sum_{i=0}^s b_i c_i^{q-1} = \frac{1}{q}, \quad q = 1, \dots, p. \quad (\text{VI.1.3})$$

Demostración. La integración es una operación lineal, por lo tanto es suficiente mostrar que (VI.1.2) se cumple para $f(x) = x^{q-1}$, donde $q = 1, \dots, p - 1$. Ahora bien,

$$\int_0^1 x^{p-1} dx = \frac{1}{p}.$$

□

Por simple verificación, puede comprobarse que la fórmula de cuadratura de la regla del Trapecio es $p = 2$, la de la regla de Simpson es $p = 4$.

Definición VI.1.10.- Una fórmula de cuadratura se dice simétrica, si:

$$\begin{aligned} c_i &= 1 - c_{s+1-i}, & i &= 1, \dots, s. \\ b_i &= b_{s+1-i}, \end{aligned} \quad (\text{IV.1.3})$$

Teorema VI.1.11.- Una fórmula de cuadratura simétrica tiene un orden par.

Demostración.- Supóngase que la fórmula de cuadratura sea exacta para $f(x)$ polinomio de grado $\leq 2k$. Por consiguiente se debe demostrar que la

fórmula de cuadratura es exacta para los polinomios de grado igual a $2k + 1$. Sea $f(x)$ un polinomio de grado igual a $2k + 1$. Ahora bien, $f(x)$ puede expresarse de la siguiente manera

$$f(x) = c \left(x - \frac{1}{2} \right)^{2k+1} + g(x),$$

donde $g(x)$ es un polinomio de grado a lo más $2k$. Por otro lado

$$\int_0^1 c \left(x - \frac{1}{2} \right)^{2k+1} dx = 0.$$

Por consiguiente, es suficiente mostrar que

$$\sum_{i=1}^s b_i \left(c_i - \frac{1}{2} \right)^{2k+1} = 0,$$

esto es cierto debido a la simetría de la fórmula de cuadratura. □

Estimacion del Error

Dada una función $f : [a, b] \rightarrow \mathbb{R}$, se quiere tener una estimación del error cometido por la utilización de una fórmula de cuadratura. Sea c_1, \dots, c_s y b_1, \dots, b_s los coeficientes y los nudos respectivamente, de una fórmula de cuadratura dada. Por consiguiente, se tiene

$$\begin{aligned} \int_a^b f(x) dx &= \sum_{j=1}^n h_j \sum_{i=1}^s b_i f(x_j + c_i h_j) = \\ &= \sum_{j=1}^n h \left[\int_{x_{j-1}}^{x_j} f(x) dx - h_j \sum_{i=1}^s b_i f(x_j + c_i h_j) \right]. \end{aligned}$$

Para poder determinar el error, es decir la diferencia entre la integral y la fórmula de cuadratura que aproxima la integral, se define

$$E(f) = \int_{x_0}^{x_0+h} f(x) dx - h \sum_{i=1}^s b_i f(x_0 + c_i h). \quad (\text{VI.1.4})$$

Habiendo definido $E(f)$, se puede determinar su valor, el cual está dado en el siguiente:

Teorema VI.1.12.- Sea $f \in \mathcal{C}^k[a, b]$, es decir f k veces continuamente derivable. Entonces, se tiene

$$E(f) = \sum_{j=0}^{k-1} \frac{h^{j+1}}{j!} \left[\frac{1}{j+1} - \sum_{i=1}^s b_i c_i^j \right] + h^{k+1} \int_0^1 P_k(t) f^{(k)}(x_0 + th) dt \quad (\text{VI.1.5})$$

donde

$$P_k(t) = E \left(\frac{(x-t)_+^{k-1}}{(k-1)!} \right), \quad \alpha_+^{k-1} = \begin{cases} \alpha^{k-1} & \alpha \geq 0 \\ 0 & \alpha < 0 \end{cases}.$$

Demostración.- Se tiene, efectuando un cambio de variable en la integral, que

$$E(f) = h \int_0^1 f(x_0 + th) dt - h \sum_{i=1}^s b_i f(x_0 + c_i h),$$

por otro lado, el desarrollo en serie de Taylor de $f(x_0 + \tau h)$ con resto en forma de integral está dado por

$$f(x_0 + h) = \sum_{j=0}^{k-1} \frac{h^j}{j!} f^{(j)}(x_0) + h^k \int_0^1 \frac{(1-\tau)^{k-1}}{(k-1)!} f^{(k)}(x_0 + \tau h) d\tau,$$

por lo tanto

$$f(x_0 + th) = \sum_{j=0}^{k-1} \frac{h^j}{j!} f^{(j)}(x_0) t^j + h^k \int_0^t \frac{(t-\tau)^{k-1}}{(k-1)!} f^{(k)}(x_0 + \tau h) d\tau,$$

introduciendo esta serie en $E(f)$, se obtiene

$$\begin{aligned} E(f) &= \sum_{j=0}^{k-1} \frac{h^{j+1}}{j!} f^{(j)}(x_0) \left[\underbrace{\int_0^1 t^j dt}_{\frac{1}{j+1}} - \sum_{i=1}^s b_i c_i^j \right] \\ &\quad + h^{k+1} \left[\int_0^1 \int_0^t \frac{(t-\tau)_+^{k-1}}{(k-1)!} f^{(k)}(x_0 + \tau h) d\tau dt \right. \\ &\quad \left. - \sum_{i=1}^s b_i \int_0^{c_i} \frac{(c_i-\tau)_+^{k-1}}{(k-1)!} f^{(k)}(x_0 + \tau h) d\tau \right] \end{aligned}$$

reemplazando en el límite de integración t por 1, se tiene que el último término del lado derecho de la anterior relación, está dado por

$$h^{k+1} \int_0^1 \left[\int_0^1 \frac{(t-\tau)_+^{k-1}}{(k-1)!} dt - \sum_{i=1}^s b_i \frac{(c_i-\tau)_+^{k-1}}{(k-1)!} \right] f^{(k)}(x_0 + \tau h) d\tau.$$

□

La función $P_k(t)$ definida en el teorema, es conocida por el nombre de Nucleo de Peano de la fórmula de cuadratura $c_1, \dots, c_s; b_1, \dots, b_s$.

Teorema VI.1.13.- *El nucleo de Peano de una fórmula de cuadratura dada, tiene las siguientes propiedades:*

$$a); \quad P_k(\tau) = \frac{(1-\tau)^k}{k!} - \sum_{i=1}^s \frac{(c_i - \tau)_+^{k-1}}{(k-1)!}$$

$$b) \quad P'_k(\tau) = -P_{k-1}(\tau), \quad \text{para } k \geq 2;$$

$$c) \quad P_k(1) = 0, \quad \text{para } k \geq 2, \quad \text{si } c_i \leq 1;$$

$$d) \quad P_k(0) = 0, \quad \text{para } 2 \leq k \leq p, \quad \text{si } c_i \geq 0 \text{ y } p \text{ orden de la f.q.};$$

e) Si la fórmula de cuadratura es $0 \leq c_1 < \dots < c_s \leq 1$ y $\sum_{i=1}^s b_i = 1$, entonces $P_1(\tau)$ es lineal por trozos, con las siguientes propiedades: $P_1(0) = 0$, $P_1|_{(c_{i-1}, c_i)}(x) = -x + d_i$, donde d_i es una constante, además $P_1(c_{i+}) - P_1(c_{i-}) = b_i$, ver figura VI.1.4.

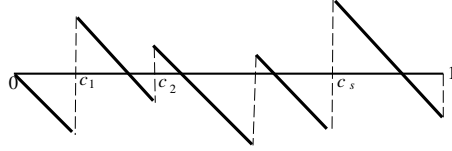


Figura VI.1.5. Gráfica de $P_1(\tau)$

Demostración.- Para el punto a), se tiene que

$$\begin{aligned} P_k(\tau) &= \int_0^1 \frac{(x-\tau)_+^{k-1}}{(k-1)!} dx - \sum_{i=1}^s b_i \frac{(x-\tau)_+^{k-1}}{(k-1)!} \\ &= \int_\tau^1 \frac{(x-\tau)^{k-1}}{(k-1)!} dx - \sum_{i=1}^s b_i \frac{(x-\tau)_+^{k-1}}{(k-1)!} \\ &= \frac{(1-\tau)^k}{k!} - \sum_{i=1}^s b_i \frac{(x-\tau)_+^{k-1}}{(k-1)!}. \end{aligned}$$

El punto b), se obtiene del punto a), derivando para $k \geq 2$. El punto c) es verificación inmediata, reemplazando en $\tau = 1$, siempre que los $c_i \leq 1$.

La demostración del punto d) se basa en que la fórmula de cuadratura, es exacta para los polinomios de grado inferior a p y

$$\begin{aligned} P_k(0) &= \frac{1}{k!} - \sum_{i=1}^s b_i \frac{c_i^{k-1}}{(k-1)!} \\ &= \frac{1}{(k-1)!} \left[\frac{1}{k} - \sum_{i=1}^s b_i c_i^{k-1} \right]. \end{aligned}$$

El punto e) es una verificación sencilla que se deja al lector. \square

Ejemplo

La regla de Simpson es una fórmula de cuadratura, dada por

$$\begin{aligned} c_1 &= 0, & c_2 &= 1/2, & c_3 &= 1; \\ b_1 &= 1, & b_2 &= 4/6, & b_3 &= 1/6. \end{aligned}$$

Utilizando las propiedades dadas en el teorema precedente, se tiene:

$$P_1(\tau) = \begin{cases} \frac{1}{6} - \tau, & 0 \leq \tau \leq \frac{1}{2}; \\ (1 - \tau) - \frac{1}{6}, & \frac{1}{2} \leq \tau \leq 1. \end{cases}$$

$P_2(\tau)$ se obtiene integrando $P_1(\tau)$ por el punto b) del teorema precedente. Por consiguiente

$$P_2(\tau) = \begin{cases} -\frac{\tau}{6} + \frac{\tau^2}{2}, & 0 \leq \tau \leq \frac{1}{2}; \\ \frac{(1 - \tau)^2}{2} - \frac{(1 - \tau)}{6}, & \frac{1}{2} < \tau \leq 1. \end{cases}$$

De la misma manera, se obtiene $P_3(\tau)$ de $P_2(\tau)$, dando como resultado:

$$\begin{aligned} P_3(\tau) &= \begin{cases} \frac{\tau^2}{12} - \frac{\tau^3}{6}, & 0 \leq \tau \leq \frac{1}{2}; \\ -\frac{(1 - \tau)^3}{6} - \frac{(1 - \tau)^2}{12}, & \frac{1}{2} < \tau \leq 1; \end{cases} \\ P_4(\tau) &= \begin{cases} -\frac{\tau^3}{36} + \frac{\tau^4}{24}, & 0 \leq \tau \leq \frac{1}{2}; \\ \frac{(1 - \tau)^4}{24} - \frac{(1 - \tau)^3}{36}, & \frac{1}{2} < \tau \leq 1. \end{cases} \end{aligned}$$

Ver los gráficos en la figura IV.1.6.

Consecuencia inmediata del teorema VI.1.12, se tiene el siguiente:

Teorema VI.1.14.- Sean, $f \in \mathcal{C}^k[a, b]$, c_1, \dots, c_s ; b_1, \dots, b_s una fórmula de cuadratura cuyo orden $p \geq k$, entonces

$$|E(f)| \leq h^{k+1} \int_0^1 |P_k(\tau)| d\tau \max_{x \in [x_0, x_0+h]} |f^{(k)}(x)|. \quad (\text{VI.1.6})$$

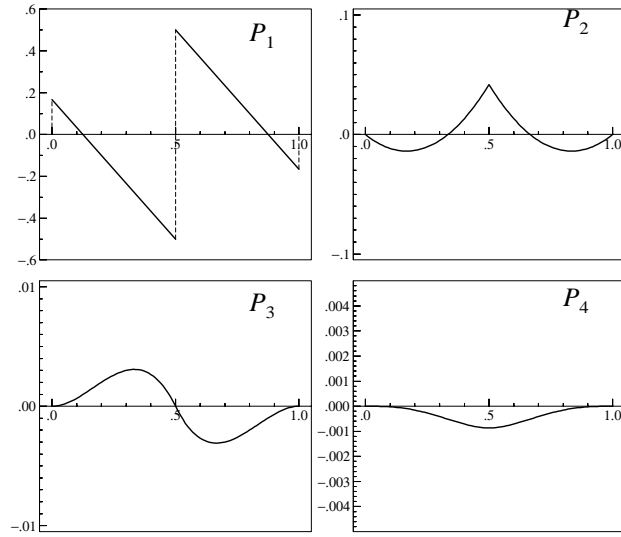


Figura VI.1.6. Nucleos de Peano, para la fórmula de Simpson.

Para el ejemplo precedente, se tiene que la fórmula de Simpson es una fórmula de cuadratura de orden 4, por consiguiente

$$\int_0^1 |P_4(\tau)| d\tau = -2 \int_0^{1/2} P_4(\tau) d\tau = \frac{1}{2880},$$

de donde, si f es cuatro veces continuamente derivable, se tiene que el error verifica

$$|E(f)| \leq \frac{h^5}{2880} \max_{x \in [x_0, x_0+h]} |f^{(4)}(x)|.$$

Teorema VI.1.15.- Si $f \in \mathcal{C}^k[a, b]$, $k \leq p$, donde p es el orden de la fórmula de cuadratura, entonces

$$\left| \int_a^b f(x) dx - \sum_{j=1}^n h_j \sum_{i=1}^s b_i f(x_{j-1} + c_i h_j) \right| \leq h^k (b-a) \int_0^1 |P_k(\tau)| d\tau \max_{x \in [a, b]} |f^{(k)}(x)|, \quad (\text{VI.6.7})$$

donde $h = \max h_i$.

Demostración.- Se tiene

$$\begin{aligned} \int_a^b f(x)dx - \sum_{j=1}^n h_j \left[\sum_{i=1}^s b_i f(x_{j-1} + c_i h_j) \right] &= \\ &= \sum_{j=1}^n \left[\int_{x_{j-1}}^{x_j} f(x)dx - h_j \sum_{i=1}^s b_i f(x_{j-1} + c_i h_j) \right] \\ &\leq h_j^{k+1} \int_0^1 |P_k(\tau)| d\tau \max_{x \in [a, b]} |f^{(k)}(x)|, \end{aligned}$$

por otro lado, se tiene

$$\sum_{i=1}^n h_j^{k+1} = \sum_{i=1}^n h_j h_j^k \leq (b-a)h^k.$$

□

Por lo tanto, la Regla de Simpson da la siguiente estimación para el error global

$$|error| \leq \frac{h^4(b-a)}{2880} \max_{x \in [a, b]} |f^{(4)}(x)|.$$

Los teoremas VI.1.13 a VI.1.15 dan estimaciones teóricas del error cometido, cuando se utiliza una fórmula de cuadratura. Sin embargo, es importante comprobar estas estimaciones teóricas con experimentos numéricos. Suponiendo que la subdivisión del intervalo $[a, b]$ es uniforme, de la fórmula (VI.1.7), suponiendo f suficientemente derivable, se deduce,

$$\int_a^b f(x)dx = \sum_{j=1}^n h \sum_{i=1}^s b_i f(x_{j-1} + c_i h) + Ch^p + \mathcal{O}(h^{p+1}), \quad (\text{VI.1.8})$$

donde C depende solamente de a, b y $f(x)$. Por consiguiente, el error satisface

$$Error \approx Ch^p. \quad (\text{VI.1.9})$$

Introduciendo logaritmos, se tiene

$$\log_{10}(Error) = \log_{10} C + p \log h,$$

denotando \mathbf{fe} , la cantidad de evaluaciones de la función $f(x)$, en el proceso de integración numérica, se tiene

$$\mathbf{fe} = \frac{C'}{h},$$

donde C' es una constante, Por lo tanto, se obtiene

$$-\log_{10}(\text{Error}) = C + p \log_{10} \mathbf{fe}. \quad (\text{VI.1.10})$$

De esta última relación, se deduce que $-\log_{10}(\text{Error})$ y $\log_{10} \mathbf{fe}$ tienen una relación lineal de pendiente p , donde p es el orden de la fórmula de cuadratura. En la figura VI.1.7, se comprueba este hecho, utilizando como integral test

$$\int_0^1 \cos(\pi e^x) e^x dx = \frac{1}{\pi} \sin(\pi e).$$

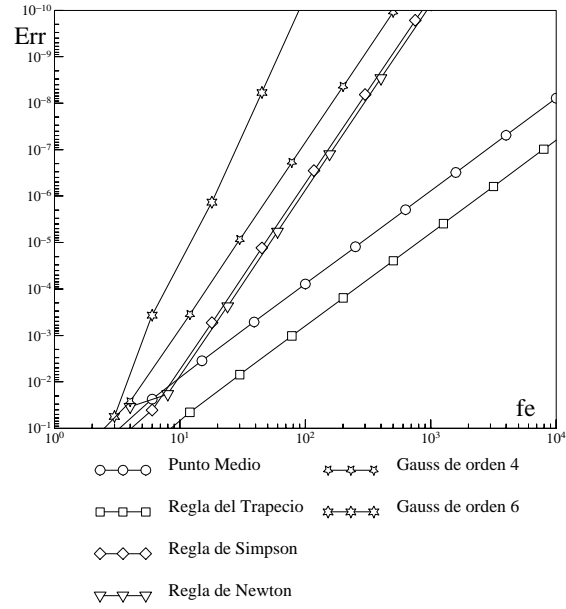


Figura VI.1.7. Gráfica del *Error* vs *fe*.

Ejercicios

1.- Calcular el orden de la regla de Newton:

$$(c_i) = (0, \frac{1}{3}, \frac{2}{3}, 1), \quad (b_i) = (\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}).$$

2.- Sean $0 \leq c_1 < c_2 < \dots < c_s \leq 1$ dados. Mostrar que existe una fórmula de cuadratura única con nudos c_i y con un orden $\geq s$.

3.- Mostrar que si la fórmula de cuadratura satisface

$$c_i = 1 - c_{s+1-i},$$

y si es de orden $\geq s$, entonces se tiene también

$$b_i = b_{s+1-i}.$$

4.- ¿Cómo se debe elegir c_1, c_2 ? para que la fórmula de cuadratura

$$b_1 f(c_1) + b_2 f(c_2)$$

tenga un orden maximal. ¿Cuál es el orden?, ¿La fórmula de cuadratura es simétrica?

5.- Calcular $\int_1^2 \frac{dx}{x} = \ln 2$ mediante la regla del trapecio, la regla de Simpson y la regla de Newton. ¿Cuál fórmula es la mejor? Hacer un gráfico logarítmico para el número de evaluaciones de la función f respecto al error.

6.- Lo mismo que el ejercicio 5, pero para

$$\int_0^{2\pi} \exp(\sin x) dx.$$

7.- Calcular π utilizando

$$\pi = 4 \int_0^1 \frac{dx}{1+x^2} \quad \text{y} \quad \pi = 4 \int_0^1 \sqrt{1-x^2} dx.$$

¿Por qué el segundo resultado es menos bueno?

8.- Mostrar que el resultado obtenido por la regla del trapecio, o por la regla de Simpson, es una suma de Riemann.

9.- Sean $h = (b-a)/n$, $x_i = a + ih$ y

$$T(h) = h \left[\frac{1}{2} f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2} f(x_n) \right].$$

Demostrar que para f suficientemente derivable

$$\int_a^b f(x) dx - T(h) = -\frac{h^2}{12} (f'(b) - f'(a)) + \mathcal{O}(h^3).$$

10.- Calcular los nucleos de Peano para la fórmula del ejercicio 4 y hacer sus gráficas.

11.- Calcular la expresión

$$\frac{\left(\int_a^b f(x) dx - T(h) \right)}{h^2},$$

para $f(x) = 1/x$, $a = 1$, $b = 10$; con varios valores de h . Verificar la fórmula del ejercicio 9.

VI.2 Cuadraturas de Orden Elevado

En esta sección, se darán las herramientas teóricas, para construir fórmulas de cuadratura de orden elevado. El interés que se tiene para utilizar fórmulas de cuadratura del mayor orden posible, reside sustancialmente en el hecho de aumentar la precisión en el cálculo de integrales definidas, como también de disminuir el número de evaluaciones de la función a integrar, ver la figura VI.1.7.

Mediante el ejercicio 2, de la sección precedente, se demuestra que si c_1, \dots, c_s dados, existe una única fórmula de cuadratura que es de orden $\geq s$, es decir

$$\sum_{i=1}^s b_i c_i^{q-1} = \frac{1}{q}, \quad (\text{VI.2.1})$$

para $q = 1, \dots, s$.

Sea m un entero no negativo, la fórmula de cuadratura $c_1, \dots, c_s; b_1, \dots, b_s$; es de orden $\leq s + m$, si y solamente si, la fórmula de cuadratura es exacta para polinomios de grado $\leq s + m - 1$. Ahora bien, se define el polinomio $M(x)$ de grado s , por

$$M(x) = (x - c_1)(x - c_2) \cdots (x - c_s), \quad (\text{VI.2.2})$$

donde los c_i son los nudos de la fórmula de cuadratura estudiada. Sea $f(x)$ un polinomio de grado $\leq s + m - 1$, por la división con resto se tiene que

$$f(x) = q(x)M(x) + r(x),$$

donde $q(x)$ y $r(x)$ son polinomios con $\deg q \leq m - 1$ y $\deg r \leq s - 1$. Por consiguiente

$$\int_0^1 f(x) dx = \int_0^1 q(x)M(x) dx + \int_0^1 r(x) dx.$$

Utilizando la fórmula de cuadratura en la anterior expresión, se obtiene

$$\sum_{i=1}^s b_i f(c_i) = \underbrace{\sum_{i=1}^s b_i q(c_i) M(c_i)}_{=0} + \sum_{i=1}^s b_i r(c_i),$$

suponiendo que el orden de la fórmula de cuadratura sea $\geq s$, se acaba de demostrar el:

Teorema VI.2.1.- Sea (b_i, c_i) , $i = 1, \dots, s$; una fórmula de cuadratura con orden $\geq s$. Entonces

$$\left. \begin{array}{l} \text{el orden de la fórmula de} \\ \text{cuadratura es } \geq s + m \end{array} \right\} \iff \left\{ \begin{array}{l} \int_0^1 q(x)M(x)dx = 0, \\ \text{para todo polinomio } q(x) \\ \text{con } \deg q \leq m - 1. \end{array} \right.$$

Ejemplo

El ejercicio 4 de la sección precedente demanda la construcción de una fórmula de cuadratura del orden más elevado posible. En consecuencia, se define

$$M(x) = (x - c_1)(x - c_2).$$

Se tiene que la fórmula de cuadratura es de orden ≥ 3 si y solamente si

$$\int_0^1 M(x)dx = 0,$$

$$\frac{1}{3} - \frac{1}{2}(c_1 + c_2) + c_1c_2 = 0,$$

la fórmula de cuadratura es de orden más grande o igual a 4, si además

$$\int_1^0 xM(x)dx = 0,$$

$$\frac{1}{4} - \frac{1}{3}(c_1 + c_2) + \frac{1}{2}c_1c_2 = 0,$$

obteniendo así un sistema de dos ecuaciones y dos incógnitas. Una vez determinados los c_i , el siguiente paso es determinar los b_i .

Polinomios Ortogonales

Uno de los mayores inconvenientes en la construcción de fórmulas de cuadratura de orden elevado utilizando el teorema VI.2.1, consiste en el hecho siguiente: se deben resolver sistemas de ecuaciones no lineales para determinar los nudos de la fórmula de cuadratura. Está claro que para fórmulas de cuadratura con una cantidad no muy grande de nudos esto es posible, sin embargo para s ya más grande esto presenta una desventaja. En esta subsección se estudiarán polinomios ortogonales, cuyas raíces son precisamente los nudos.

Se iniciará esta subsección, definiendo un conjunto de funciones particulares. Sea $\omega : (a, b) \rightarrow \mathbb{R}$ una función, llamada función de peso. Sea

$$\mathcal{E} = \left\{ f : (a, b) \rightarrow \mathbb{R} \mid \int_a^b \omega(x) |f(x)|^2 dx < \infty \right\}. \quad (\text{VI.2.3})$$

Puede mostrarse que \mathcal{E} es un espacio vectorial para la adición de funciones y la multiplicación por escalares reales. Para las aplicaciones en general, puede suponerse que f es una función continua.

Suponiendo que $\omega(x) > 0$, puede definirse un producto escalar sobre \mathcal{E} , de la siguiente manera,

$$\langle f, g \rangle = \int_a^b \omega(x) f(x) g(x) dx. \quad (\text{VI.2.4})$$

Hay que remarcar dos hechos importantes; el primero \mathcal{E} es en realidad un conjunto de clases de equivalencia de funciones, donde la relación de equivalencia está definida por

$$f \sim g \iff f(x) \neq g(x) \text{ en un conjunto de medida nula.}$$

La segunda observación es consecuencia de la primera, se puede suponer por lo tanto que \mathcal{E} está constituida por funciones continuas, a lo sumo por funciones continuas por trozos, y con ciertas condiciones impuestas a $\omega(x)$ puede demostrarse que

$$\langle f, f \rangle = 0 \iff f = 0.$$

Definición VI.2.2.- f es ortogonal a g , que se denota por $f \perp g$, si

$$\langle f, g \rangle = 0. \quad (\text{VI.2.5})$$

El objetivo central será, por consiguiente, encontrar $M(x) \perp q(x)$ si $\deg q \leq m - 1$.

Teorema VI.2.3.- Sea $\omega(x)$ dado, entonces existe una sucesión de polinomios $p_0(x), p_1(x), p_2(x), \dots$, tal que:

$$\begin{aligned} \deg p_j &= j; \\ \langle p_k, q \rangle &= 0, \quad \forall q \text{ polinomio de grado } \leq k - 1. \end{aligned}$$

Si se supone que $p_k(x) = x^k + r(x)$ con $\deg r(x) \leq k - 1$, los polinomios son únicos.

Los polinomios satisfacen la siguiente relación recursiva:

$$p_{-1}(x) := 0, \quad p_0(x) = 1, \quad (\text{VI.2.6a})$$

$$p_{k+1} = (x - \delta_{k+1})p_k(x) - \gamma_{k+1}^2 p_{k-1}(x), \quad (\text{VI.2.6b})$$

donde

$$\delta_{k+1} = \frac{\langle xp_k, p_k \rangle}{\langle p_k, p_k \rangle}, \quad \gamma_{k+1}^2 = \frac{\langle p_k, p_k \rangle}{\langle p_{k-1}, p_{k-1} \rangle}. \quad (\text{VI.2.6c})$$

Demostración.- El primer punto del teorema, se obtiene a partir del proceso de ortogonalización de Gramm-Schmidt. Las relaciones entre los diferentes polinomios ortogonales se demuestra por inducción. Por consiguiente, se supone cierto para los polinomios p_0, p_1, \dots, p_k . Ahora bien, se tiene

$$p_{k+1}(x) = xp_k(x) + \sum_{j=0}^k c_j p_j(x),$$

por que el coeficiente dominante de p_k es 1. Aplicando el producto escalar se tiene

$$\begin{aligned} 0 &= \langle p_{k+1}, p_k \rangle \\ &= \langle xp_k, p_k \rangle + \sum_{j=0}^k c_j \langle p_j, p_k \rangle \\ &= \langle xp_k, p_k \rangle + c_k \langle p_k, p_k \rangle, \end{aligned}$$

de donde

$$c_k = -\frac{\langle xp_k, p_k \rangle}{\langle p_k, p_k \rangle}.$$

Por otro lado, aplicando nuevamente el producto escalar se tiene

$$\begin{aligned} 0 &= \langle p_{k+1}, p_{k-1} \rangle \\ &= \langle xp_k, p_{k-1} \rangle + \sum_{j=0}^k c_j \langle p_j, p_{k-1} \rangle \\ &= \langle xp_k, p_{k-1} \rangle + c_{k-1} \langle p_{k-1}, p_{k-1} \rangle \end{aligned}$$

Ahora bien, se verifica facilmente, utilizando la definición del producto escalar definido en \mathcal{E} , que

$$\langle xp_k, p_{k-1} \rangle = \langle p_k, xp_{k-1} \rangle,$$

con la hipótesis de inducción, se verifica inmediatamente que

$$\langle p_k, xp_{k-1} \rangle = \langle p_k, p_k \rangle,$$

de donde

$$c_{k-1} = \frac{\langle p_k, p_k \rangle}{\langle p_{k-1}, p_{k-1} \rangle}.$$

Los restantes c_j , son nulos, utilizando la hipótesis de inducción sobre la ortogonalidad. \square

Consecuencia de este teorema, es que si los nudos de una fórmula de cuadratura son las raíces del polinomio p_s , definido en el teorema precedente, se tiene $M(x) = p_s(x)$ y por el teorema VI.2.1 el orden de la fórmula de cuadratura es igual o mayor a $2s$.

Teorema VI.2.4.- Sean los p_k , como en el teorema VI.2.3, entonces las raíces de $p_k(x)$ son reales, simples y están localizadas en el intervalo (a, b) .

Demostración.- Se denota por τ_1, \dots, τ_T las raíces distintas de p_k donde la función $p_k(x)$ cambia de signo. Se define el polinomio $g(x)$ por

$$g(x) = (x - \tau_1) \cdots (x - \tau_T),$$

por consiguiente, se tiene que

$$g(x)p_k(x)$$

no cambia de signo, de donde

$$\langle g(x), p_k(x) \rangle \neq 0,$$

por lo tanto $\deg g \geq k$, y por la hipótesis inicial se tiene necesariamente que $T = k$. \square

En la tabla VI.2.1, se tienen los diferentes tipos de polinomios ortonormales para diferentes tipos de función de peso $\omega(x)$.

Tabla VI.2.1. Ejemplos de Polinomios Ortonormales

$\omega(x)$	(a, b)	Notación	Nombre
1	$(-1, 1)$	$P_k(x)$	Polinomios de Legendre
$\frac{1}{\sqrt{1-x^2}}$	$(-1, 1)$	$T_k(x)$	Polinomios de Chebichef
$(1-x)^\alpha(1+x)^\beta$	$(-1, 1)$	$P_k^{(\alpha, \beta)}(x)$	Jacobi, $\alpha, \beta > -1$
$x^\alpha e^{-x}$	$(0, \infty)$	$L_k^{(\alpha)}(x)$	Pol. de Laguerre, $\alpha > -1$
e^{-x^2}	$(-\infty, \infty)$	$H_k(x)$	Polinomios de Hermite

Teorema VI.2.5.- *Fórmula de Rodriguez.* Sea $\omega(x)$ una función de peso definida como antes. Entonces

$$p_k(x) = C_k \frac{1}{\omega(x)} \frac{d^k}{dx^k} \{ \omega(x)(x-a)^k(b-x)^k \}. \quad (\text{VI.2.7})$$

Demostración.- Una verificación simple sobre $p_k(x)$, muestra que se tratan efectivamente de polinomios. Para la relación de ortogonalidad con k dado, es suficiente mostrar que si $q(x)$ es un polinomio de grado $\leq k-1$ entonces $q \perp p_k$. En efecto

$$\begin{aligned} \int_a^b \omega(x) p_k(x) q(x) dx &= \int_a^b \frac{d^k}{dx^k} \{ \omega(x)(x-a)^k(b-x)^k \} q(x) dx \\ &= \underbrace{\frac{d^{k-1}}{dx^{k-1}} \{ \omega(x)(x-a)^k(b-x)^k \} q(x) \Big|_a^b}_{=0} \\ &\quad - \int_a^b \frac{d^{k-1}}{dx^{k-1}} \{ \omega(x)(x-a)^k(b-x)^k \} q'(x) dx \\ &\quad \vdots \\ &= \pm \int_a^b \{ \omega(x)(x-a)^k(b-x)^k \} q^{(k)}(x) dx \\ &= 0. \end{aligned}$$

□

La mayor parte de los cálculos de integrales definidas, tienen como función de peso $\omega(x) = 1$. A continuación se estudiará con mayor detalle los polinomios de Legendre.

Los Polinomios de Legendre

Los polinomios de Legendre son los polinomios ortogonales para la función de peso $\omega(x) = 1$ definida en el intervalo $(-1, 1)$, ver la tabla VI.2.1. Por otro lado, utilizando la fórmula de Rodriguez se puede elegir los C_k de manera que $P_k(1) = 1$. Por consiguiente, se tiene

$$1 = P_k(1) = C_k \frac{d^k}{dx^k} \{ (1-x)^k(1+x)^k \} \Big|_{x=1} = C_k (-2)^k k!,$$

de donde

$$C_k = \frac{(-1)^k}{2^k k!},$$

por lo tanto

$$P_k(x) = \frac{(-1)^k}{2^k k!} \frac{d^k}{dx^k} ((1-x^2)^k). \quad (\text{VI.2.8})$$

Los polinomios de Legendre pueden ser calculados mediante la fórmula de Rodriguez, o mediante una relación recursiva, ver ejercicio 1. En la tabla VI.2.2, se da los cuatro primeros polinomios de Legendre.

Tabla VI.2.2. Polinomios de Legendre

k	$P_k(x)$
0	1
1	x
2	$\frac{3}{2}x^2 - \frac{1}{2}$
3	$\frac{5}{2}x^3 - \frac{3}{2}x$

Puede observarse que si k es par, entonces $P_k(x) = Q(x^2)$ donde Q es un polinomio; de la misma manera si k es impar, se tiene que $P_k(x) = xQ(x^2)$.

Las Fórmulas de Cuadratura de Gauss

La verificación del orden de una fórmula de cuadratura, se la realiza en el intervalo $[0, 1]$. Efectuando una transformación afín, se tiene que $M(x)$ del teorema VI.2.1, es igual a

$$M(x) = (x - c_1) \cdots (x - c_s) = P_s(2x - 1) \quad (\text{VI.2.9})$$

con P_s el s -simo polinomio de Legendre, si se desea que orden de la fórmula de cuadratura sea al menos s . El siguiente teorema, tiene una importancia en lo concerniente al orden de una fórmula de cuadratura.

Teorema VI.2.6.- *El orden de una fórmula de cuadratura dada por (b_i, c_i) , $i = 1, \dots, s$; es menor o igual a $2s$.*

Demostración.- Por el absurdo. Supóngase que existe una fórmula de cuadratura de orden superior o igual a $2s + 1$, de donde para todo polinomio $l(x)$ de grado s , se tiene

$$\int_0^1 l(x)M(x)dx,$$

sin embargo, tomando $l(x) = M(x)$ se tiene

$$\int_0^1 M^2(x) dx = 0,$$

lo que conduce a una contradicción \square

El objetivo, será por consiguiente, encontrar una fórmula de cuadratura de orden máximo, es decir $2s$. Por la observación hecha al inicio de esta última subsección, los nudos c_i de la fórmula de cuadratura de orden s son las raíces de $P_s(2x - 1)$ polinomio de Legendre. Como consecuencia de lo anteriormente expuesto se tiene el siguiente teorema formulado por Gauss en 1814.

Teorema VI.2.7.- *Una fórmula de cuadratura (c_i, b_i) , $i = 1, \dots, s$; es de orden $2s$ si y solamente si c_1, \dots, c_s son las raíces de $P_s(2x - 1)$, $P_s(t)$ polinomio de Legendre y los b_i están determinados por*

$$\sum_{i=1}^s b_i c_i^{q-1} = \frac{1}{q}, \quad q = 1, \dots, s. \quad (\text{VI.2.10})$$

Por otro lado, debe observarse que los coeficientes de una fórmula de cuadratura de Gauss son estrictamente positivos, en efecto, se define

$$l_i(x) = \prod_{\substack{j=1 \\ j \neq i}}^s \frac{x - c_j}{c_i - c_j}$$

el i -ésimo polinomio de Lagrange para la subdivisión $c_1 < \dots < c_s$. El grado de este polinomio es igual a $s - 1$ y verifica

$$l_i(c_j) = \begin{cases} 0, & j \neq i; \\ 1, & j = i. \end{cases}$$

Ahora bien, se tiene

$$b_i = \sum_{j=1}^s b_j (l_i(c_j))^2 = \int_0^1 (l_i(x))^2 dx > 0,$$

ya que, $\deg l_i^2 = 2s - 2 < 2s$.

El cálculo de los coeficientes b_i de una fórmula de cuadratura de Gauss, pueden ser resueltos mediante el sistema lineal dado por (VI.2.10). Sin embargo este procedimiento no es el mejor, debido a la acumulación de los

errores de redondeo. Existe un procedimiento que determina los b_i de una manera sencilla, el está dado en el:

Teorema VI.2.8.- Para la formula de cuadratura de Gauss de orden $2s$, se tiene

$$b_i = \frac{1}{(1 - x_i^2)P'_s(x_i)^2}, \quad i = 1, \dots, s; \quad (\text{VI.2.11})$$

donde $x_i = 2c_i - 1$.

Demostración.- Se tiene,

$$b_i = \sum_{j=1}^s b_j l_i(c_j) = \int_0^1 l_i(t) dt,$$

realizando la transformación $x = 2t - 1$, se obtiene

$$b_i = \frac{1}{2} \int_{-1}^1 l_i \left(\frac{x+1}{2} \right) dx,$$

por otro lado, se tiene

$$l_i \left(\frac{x+1}{2} \right) = C \frac{P_s(x)}{x - x_i},$$

de donde pasando al límite, se obtiene

$$\lim_{x \rightarrow x_i} C \frac{P_s(x)}{x - x_i} = C P'_s(x_i) = 1,$$

despejando C , b_i está dado por

$$b_i = \frac{1}{2} \int_{-1}^1 \frac{P_s(x)}{(x - x_i)P'_s(x_i)} dx. \quad (\text{VI.2.12})$$

Además,

$$b_i = \sum_{j=1}^s b_j l_i(c_j) = \frac{1}{2} \int_{-1}^1 \left(\frac{P_s(x)}{(x - x_i)P'_s(x_i)} \right)^2 dx, \quad (\text{VI.2.13})$$

esta integral se la resuelve por partes, obteniendo así

$$b_i = \frac{1}{2(P'_s(x_i))^2} \int_{-1}^1 (P_s(x))^2 \frac{1}{(x - x_i)^2} dx$$

$$\begin{aligned}
&= \frac{1}{2(P'_s(x_i))^2} \left[\frac{-1}{1-x_i} + \frac{1}{-1-x_i} \right] + \int_{-1}^1 \frac{P'_s(x)}{P'_s(x_i)} \underbrace{\left(\frac{P_s(x)}{(x-x_i)P'_s(x_i)} \right)}_{= l_i \left(\frac{x+1}{2} \right)} dx \\
&= \frac{1}{2(P'_s(x_i))^2} \frac{-1}{1-x_i^2} + \sum_{j=1}^s b_j \frac{P'_s(x_j)}{P'_s(x_i)} l_i \left(\frac{x_j+1}{2} \right) \\
&= \frac{1}{2(P'_s(x_i))^2} \frac{-1}{1-x_i^2} + 2b_i
\end{aligned}$$

□

En el ejercicio 1 de esta sección, se mostrará que los polinomios de Legendre verifican la siguiente relación recursiva

$$(1-x^2)P'_s(x) = -s x P_s(x) + s P_{s-1}(x),$$

de donde

$$(1-x_i^2)P'_s(x_i) = s P_{s-1}(x_i),$$

obteniendo

$$b_i = \frac{1-x_i^2}{s^2(P_{s-1}(x_i))^2}. \quad (\text{VI.2.14})$$

En la tabla VI.6.3, se dan las primeras fórmulas de cuadratura de Gauss.

Tabla VI.6.3. Primeras Fórmulas de Cuadratura de Gauss.

s	c_1	c_2	c_3	b_1	b_2	b_3
1	$\frac{1}{2}$			1		
2	$\frac{1}{2} - \frac{\sqrt{3}}{6}$	$\frac{1}{2} + \frac{\sqrt{3}}{6}$		$\frac{1}{2}$	$\frac{1}{2}$	
3	$\frac{1}{2} - \frac{\sqrt{15}}{10}$	$\frac{1}{2}$	$\frac{1}{2} + \frac{\sqrt{15}}{10}$	$\frac{5}{18}$	$\frac{8}{18}$	$\frac{5}{18}$

Ejercicios

1.- Para los polinomios de Legendre, demostrar que:

$$(k+1)P_{k+1}(x) = (2k+1)xP_k(x) - kP_{k-1}(x);$$

$$(1-x^2)P'_k(x) = -kxP_k(x) + kP_{k-1}(x).$$

2.- Los polinomios de Chebychef están definidos por

$$T_k(x) = \cos(k \arccos x).$$

Verificar que:

$$\begin{aligned} T_0(x) &= 1; \quad T_1(x) = x; \\ T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x); \\ \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} T_k(x) T_j(x) dx, &\quad \text{para } i \neq j. \end{aligned}$$

3.- Calcular las raíces de $P_8(x)$ con un método iterativo, como por ejemplo el método de la bisección.

4.- Mostrar que el nucleo de Peano $P_k(t)$ de una fórmula de cuadratura satisface

$$\int_0^1 P_k(t) dt = \frac{1}{k+1} - \sum_{i=1}^s b_i c_i^k.$$

5.- Sea p el orden de una fórmula de cuadratura y supóngase que el nucleo de Peano $P_p(t)$ no cambia de signo sobre $[0, 1]$. Mostrar que

$$\int_{x_0}^{x_0+h} f(x) dx - h \sum_{i=1}^s b_i f(x_0 + c_i h) = h^{p+1} \left(\frac{1}{p+1} - h \sum_{i=1}^s b_i c_i^p \right) f^{(p)}(\xi),$$

con $\xi \in (x_0, x_0 + h)$.

Indicación.- Utilizar el teorema VI.1.15 con $k = p$.

6.- Mostrar que para las fórmulas de cuadratura de Gauss de orden $2s$, el nucleo de Peano $P_{2s}(t)$ no cambia de signo.

VI.3 Implementación Numérica

El cálculo numérico de integrales definidas, requiere la implementación de las fórmulas de cuadratura en forma de programas o subrutinas. Dada una función $f : [a, b] \rightarrow \mathbb{R}$, el cálculo de

$$\int_a^b f(x)dx, \quad (\text{VI.3.1})$$

se lo realiza teniendo en cuenta el error que se desea cometer. Generalmente se da una tolerancia que se la denota por **TOL**, por consiguiente se busca una aproximación I , tal que

$$\left| \int_a^b f(x)dx - I \right| \leq \text{TOL} \int_a^b |f(x)| dx. \quad (\text{VI.3.2})$$

Ahora bien, una manera de conseguir (VI.3.2), es subdividir el intervalo $[a, b]$ en subintervalos y aplicar el teorema VI.1.16 de manera de obtener un h óptimo. Sin embargo este procedimiento presenta dos inconvenientes: es necesario conocer de antemano este h óptimo; además de conocer las propiedades de la función f a integrar. La segunda manera de resolver (VI.3.2) es de concebir un algoritmo, donde el cálculo del error se haga de manera automática es decir utilizando un método adaptativo.

Lo primero que se debe tener en la implementación de un programa que permita evaluar la integral de una función f , es una estimación del error. Sea (b_i, c_i) $i = 1, \dots, s$, una fórmula de cuadratura y $a = x_0 < x_1 < \dots < x_n = b$, una subdivisión de $[a, b]$; se define el error en cada subintervalo $[x_i, x_{i+1}]$ a

$$E(f, x_i, x_{i+1}) = \int_{x_i}^{x_{i+1}} f(x)dx - (x_{i+1} - x_i) \sum_{j=1}^s b_j f(x_i + c_j(x_{i+1} - x_i)). \quad (\text{VI.3.3})$$

Por lo tanto, el programa que va calcular numéricamente la integral debe tener en cuenta dos aspectos:

- 1.- Estimación del error.
- 2.- Elección de la subdivisión de $[a, b]$, $x_0 < x_1 < \dots < x_n$; tal que

$$\sum_{j=0}^n |E(f, x_j, x_{j+1})| \leq \text{TOL} \int_a^b |f(x)| dx.$$

Por consiguiente, denotando por:

$$\mathbf{res}_{[a,b]} = I, \quad \mathbf{resabs}_{[a,b]} = \int_a^b |f(x)| dx, \quad \mathbf{err}_{[a,b]} = \left| \int_a^b f(x) dx - I \right|,$$

se tiene el siguiente algoritmo.

Algoritmo

- Calcular: $\mathbf{res}_{[a,b]}$, $\mathbf{resabs}_{[a,b]}$ y $\mathbf{err}_{[a,b]}$.
 $\mathbf{err}_{[a,b]} \leq \text{TOL} \mathbf{resabs}_{[a,b]}$: si es cierto se ha terminado, si no continuar el siguiente paso.
- Plantear $c = c = \frac{a+b}{2}$ y calcular:

$$\begin{array}{lll} \mathbf{res}_{[a,c]}, & \mathbf{err}_{[a,c]}, & \mathbf{resabs}_{[a,c]}, \\ \mathbf{res}_{[c,b]}, & \mathbf{err}_{[c,b]}, & \mathbf{resabs}_{[c,b]}; \end{array}$$

$(\mathbf{err}_{[a,c]} + \mathbf{err}_{[c,b]}) \leq \text{TOL} (\mathbf{resabs}_{[a,c]} + \mathbf{resabs}_{[c,b]})$: si es cierto se ha terminado, si no dividir el subintervalo con error maximal y continuar con el siguiente paso.

- Continuar hasta que

$$\sum \mathbf{err}_{[a_j, c_j]} \leq \text{TOL} \left(\sum \mathbf{resabs}_{[a_j, c_j]} \right).$$

Una vez planteado el algoritmo, el principal problema consiste en estimar el error cometido en el cálculo de la integral, en cada subintervalo $[x_i, x_{i+1}]$. El teorema VI.1.12 da una estimación cuando la fórmula de cuadratura tiene un orden p , la cual está dada por

$$E(f, x_0, x_1) = h^{p+1} \int_0^1 P_p(t) f^{(p)}(x_0 + th) dt,$$

donde $P_p(t)$ es el k -simo nucleo de Peano, para la fórmula de cuadratura. El principal inconveniente de utilizar esta estimación radica en que el cálculo de la derivada de orden p de f puede ser tan complicado, como encontrar una primitiva de f , y por otro lado, determinar el nucleo de Peano no es una tarea nada simple, motivos por los cuales, la estimación del teorema VI.1.12 no es nada práctica desde el punto de vista computacional.

Ahora bien, existen dos métodos numéricos que permiten encontrar una estimación del error cometido en cada subintervalo, sin necesidad de conocer las propiedades del nucleo de Peano y de las derivadas de la función a

integrar. Por consiguiente sea (c_i, b_i) , una fórmula de cuadratura dada de orden p , se desea estimar

$$E(f, x_0, x_1) = \int_{x_0}^{x_1} f(x) dx - (x_1 - x_0) \sum_{i=1}^s b_i (f(x_0 + c_i(x_1 - x_0))).$$

El primer método para estimar E es conocido como **QUADPACK**, algoritmo implementado en algunas bibliotecas de programas. La idea central de este método es tomar una segunda fórmula de cuadratura (\hat{c}_i, \hat{b}_i) con un orden $\hat{p} > p$ y estimar $E(f, x_0, x_1)$, como

$$E(f, x_0, x_1) = (x_1 - x_0) \left[\sum_{i=1}^{\hat{s}} \hat{b}_i f(x_0 + \hat{c}_i(x_1 - x_0)) - \sum_{i=1}^s b_i f(x_0 + c_i(x_1 - x_0)) \right], \quad (\text{VI.3.4})$$

Para evitar demasiadas evaluaciones de la función f , es deseable que

$$\{c_1, \dots, c_s\} \subset \{\hat{c}_1, \dots, \hat{c}_{\hat{s}}\},$$

es decir

$$\{\hat{c}_1, \dots, \hat{c}_{\hat{s}}\} = \{c_1, \dots, c_s\} \cup \{c_{s+1}, \dots, c_{s+m}\},$$

con $s + m = \hat{s}$. Sin embargo m debe ser más grande que s , si la fórmula de cuadratura (c_i, b_i) es una de tipo Gauss; en efecto, si $m \leq s$, se tiene:

$$\begin{aligned} \sum_{i=1}^{s+m} \hat{b}_i c_i^{j-1} &= \frac{1}{j}, \quad j = 1, \dots, s+m; \\ \sum_{i=1}^s b_i c_i^{j-1} &= \frac{1}{j}, \quad j = 1, \dots, 2s; \end{aligned}$$

lo cual conduce a que

$$\hat{b}_i = \begin{cases} b_i, & i = 1, \dots, s; \\ 0, & i = s+1, \dots, s+m; \end{cases}$$

obteniendo así, la misma fórmula de cuadratura. Por consiguiente es necesario elegir $m > s$, por ejemplo $m = s + 1$.

Por otro lado, se pueden elegir los c_i restantes de manera que la fórmula de cuadratura (\hat{c}_i, \hat{b}_i) , $i = 1, \dots, 2s + 1$; tenga un orden igual a $3s + 2$. Esta fórmula de cuadratura lleva el nombre de Konrad, en honor a su descubridor. El método **QUADPACK**, toma como resultado de la integral al resultado numérico proporcionado por la fórmula de cuadratura de Konrad, es decir

$$\mathbf{res} = (x_1 - x_0) \sum_{i=1}^{2s+1} \hat{b}_i f(x_0 + c_i(x_1 - x_0)), \quad (\text{VI.3.5})$$

la estimación del error cometido es de la fórmula de cuadratura de Gauss y no así de la de Konrad. No obstante, se puede obtener una estimación de este error. En efecto los errores de las fórmulas de cuadratura están dados por:

$$\begin{aligned}\mathbf{err}_{\text{Gauss}} &= Ch^{2s+1} + \dots, \\ \mathbf{err}_{\text{Konrad}} &= \hat{C}h^{3s+3} + \dots\end{aligned}$$

Para simplificar los cálculos se puede suponer que tanto C , como \hat{C} son iguales y valen 1, obteniendo así, cuando h tiende a 0

$$(h^{2s+1})^{3/2} \approx h^{3s+3},$$

de donde la estimación del error de la fórmula de cuadratura multiplicada por una constante de seguridad está dada por:

$$\mathbf{err} = \left[(x_1 - x_0) \left(\sum_{i=1}^{2s+1} \hat{b}_i f(x_0 + c_i h) - \sum_{i=1}^s b_i f(x_0 + c_i h) \right) \right]^{3/2} \cdot 100, \quad (\text{VI.3.6})$$

finalmente se tiene

$$\mathbf{resabs} = (x_1 - x_0) \sum_{i=1}^{2s+1} \hat{b}_i |f(x_0 + c_i(x_1 - x_0))|. \quad (\text{VI.3.7})$$

El segundo método es conocido como **GAUINIT, GAUSS**. Al igual que en el método QUADPACK, se considera una formula de cuadratura de tipo Gauss (c_i, b_i) , $i = 1, \dots, s$; pero s impar, de manera que uno de los nudos sea igual a $1/2$. Luego se considera la fórmula de cuadratura de orden al menos $s - 1$ obtenida de la fórmula original, cuyos nudos están dados por

$$\{\hat{c}_1, \dots, \hat{c}_{s-1}\} = \{c_1, \dots, c_s\} \setminus \{1/2\}.$$

Con los mismos argumentos desarrollados para el método QUADPACK, pero esta vez tomando el resultado numérico proporcionado por la fórmula de cuadratura de Gauss, se obtiene:

$$\mathbf{res} = (x_1 - x_0) \sum_{i=1}^s b_i f(x_0 + c_i h), \quad (\text{VI.3.8})$$

$$\mathbf{err} = \left[(x_1 - x_0) \left(\sum_{i=1}^s b_i f(x_0 + c_i h) - \sum_{i=1}^s \hat{b}_i f(x_0 + \hat{c}_i h) \right) \right]^2 \cdot 100, \quad (\text{VI.3.9})$$

$$\mathbf{resabs} = (x_1 - x_0) \sum_{i=1}^s b_i |f(x_0 + c_i h)|. \quad (\text{VI.3.10})$$

Las experiencias numéricas muestran que en la mayoría de los casos las estimaciones del error cometido son demasiado pesimistas, ver en la tabla VI.3.1, las experiencias numéricas han sido realizadas por el método **GAUINIT**. El programa **GAUINIT**, para estas experiencias numéricas, utiliza una fórmula de cuadratura de Gauss de orden 30.

Tabla VI.3.1 Error exacto *vs* Error estimado.

$f(x)$	$[a, b]$	Error exacto	err
$\frac{1}{x^4 + x^2 + 1}$	$[0, 2]$	0.23×10^{-10}	0.90×10^{-10}
$25e^{-25x}$	$[0, 1]$	0.14×10^{-11}	0.23×10^{-5}
\sqrt{x}	$[0, 1/2]$	0.98×10^{-5}	0.14×10^{-8}

Puede observarse que la tercera función a ser integrada es una excepción de la regla anteriormente formulada, eso se debe a que \sqrt{x} no es lo suficientemente derivable, y el algoritmo ha sido concebido para funciones lo suficientemente lisas.

Tratamiento de singularidades

Los métodos desarrollados en la anterior subsección, tal como se puede observar en la tabla precedente, son utilizables para funciones lo suficientemente derivables. Por lo tanto, no son muy eficientes para resolver integrales definidas de funciones no muy lisas, además existen integrales impropias cuyo cálculo es frecuente en diversas aplicaciones, como por ejemplo integrales de los tipos:

$$\int_0^1 \frac{f(x)}{\sqrt{x}} dx, \quad \int_0^1 (\log x) f(x) dx.$$

Ejemplo

Considérese, la función

$$f(x) = -\frac{4x \log x}{x^4 + 100},$$

se desea calcular $\int_0^1 f(x) dx$. Esta integral es impropia, no obstante que una fórmula de cuadratura de tipo Gauss proporciona resultados que se aproximan al valor exacto de esta integral. Esto se debe a que los nudos de la fórmula utilizada son diferentes de 0 y que la función $f(x)$ es singular en $x = 0$. Ahora bien, el error exacto al integrar sobre el intervalo $[0, 1]$ es del orden de 0.18×10^{-4} , el cual está cerca del 5% del

valor exacto, valor muy grande. Un procedimiento para obtener un error que este en el orden de TOL, es definir la sucesión S_k dada por

$$S_k = \sum_{j=0}^k \int_{a_j}^{b_j} f(x) dx,$$

donde $b_j = 2^{j-k}$ y $a_j = b_j/2$, para $j > 0$. Con este procedimiento, solamente se debe calcular la integral en el intervalo más pequeño. Para poder comparar los resultados obtenidos con el método numérico, el valor exacto de la integral, calculada mediante series, es igual a

$$\begin{aligned} \int_0^1 -\frac{4x \log x}{x^4 + 100} dx &= \int_0^1 \frac{1}{100} \frac{-4x \log x}{1 + x^4/100} dx \\ &= \frac{1}{100} \int_0^1 (-4x \log x) \sum_{k=0}^{\infty} (-1)^k \frac{x^{4k}}{100^k} dx \\ &= \frac{-4}{100} \sum_{k=0}^{\infty} \frac{(-1)^k}{100^k} \int_0^1 x^{4k+1} \log x dx \\ &= \frac{1}{100} \sum_{k=0}^{\infty} \frac{(-1)^k}{100^k} \frac{1}{(2k+1)^2}, \end{aligned}$$

lo que es igual con 16 cifras de precisión a

$$\int_0^1 -\frac{4x \log x}{x^4 + 100} dx = 9.9889286860336184 \times 10^{-03}.$$

Aplicando el procedimiento mencionado más arriba, se obtiene la tabla VI.3.2.

Tabla VI.3.2. Cálculo Integral.

S_k	Error Exacto	S_k	Error Exacto
S_0	$1.748733098830973E-07$	S_7	$1.067342048077790E-11$
S_1	$4.371832748595316E-08$	S_8	$2.668355120194476E-12$
S_2	$1.092958187148829E-08$	S_9	$6.670896474103571E-13$
S_3	$2.732395467872073E-09$	S_{10}	$1.667728455334582E-13$
S_4	$6.830988674016991E-10$	S_{11}	$4.169407874510255E-14$
S_5	$1.707747172841056E-10$	S_{12}	$1.042395336714463E-14$
S_6	$4.269368018838815E-11$	S_{13}	$2.605554660917164E-15$

Se puede observar inmediatamente, que la convergencia para calcular la integral es muy lenta, es necesario, efectuar 13 subdivisiones para obtener un error igual o inferior a 2.61×10^{-15} . Cada utilización del programa GAUINT requiere 15 evaluaciones de la función f , por consiguiente para obtener el error mencionado, es necesario por lo menos 14×15 evaluaciones de la función f .

Para evitar tantas evaluaciones de la función f , es necesario construir un algoritmo que permita acelerar la convergencia. Una forma de hacerlo es utilizar procedimientos de extrapolación al límite dado en el capítulo III.3. Ahora bien, el método que será estudiado para acelerar la convergencia en el cálculo de estas integrales será tratado con un procedimiento equivalente, el cual consiste en utilizar diferencias finitas.

A partir de la tabla precedente puede observarse, el siguiente hecho: Denótese por S el valor exacto de la integral, entonces

$$S_{n+1} - S \approx \frac{1}{4}(S_n - S),$$

es decir

$$S_{n+1} - S \approx \rho(S_n - S). \quad (\text{VI.3.11})$$

Supóngase, que se conoce tres valores consecutivos de la sucesión $\{S_k\}$, por decir: S_n, S_{n+1} y S_{n+2} , utilizando la notación de diferencias finitas dada en el capítulo III.1, se tiene el siguiente sistema lineal

$$\begin{cases} S_{n+1} - S = \rho(S_n - S) \\ S_{n+2} - S = \rho(S_{n+1} - S) \end{cases}, \quad (\text{VI.3.12})$$

de donde sustrayendo ambas ecuaciones, se obtiene

$$\Delta S_{n+1} = \rho \Delta S_n,$$

por consiguiente

$$\rho = \frac{\Delta S_{n+1}}{\Delta S_n}. \quad (\text{VI.3.13})$$

Despejando S de la segunda ecuación de (VI.3.12), se tiene

$$\begin{aligned} S &= S_{n+1} - \frac{1}{\rho - 1} \Delta S_{n+1} \\ &= S_{n+1} - \frac{\Delta S_n \Delta S_{n+1}}{\Delta S_{n+1} - \Delta S_n}, \end{aligned}$$

obteniendo así

$$S'_n = S_{n+1} - \frac{\Delta S_n \Delta S_{n+1}}{\Delta^2 S_n}. \quad (\text{VI.3.14})$$

El método que acaba de ser formulado por (VI.3.14), es conocido por el procedimiento Δ^2 de *Aitken*. En la tabla VI.3.3, se dan los valores obtenidos por este procedimiento, para el ejemplo precedente.

Tabla VI.3.3. Procedimiento Δ de Aitken.

S'_k	Valor integral	Error Exacto
S'_0	$9.988928686033609E-03$	$0.35E-14$
S'_1	$9.988928686033618E-03$	$0.26E-14$
S'_2	$9.988928686033618E-03$	$0.26E-14$

Con la finalidad de comparar la eficiencia, del procedimiento Δ^2 de Aitken, para obtener un error del orden de 0.26×10^{-14} solo se necesitan 3 evaluaciones de integrales de f , mientras que, sin el procedimiento de aceleración es necesario 14 evaluaciones de integral.

El siguiente paso en lograr una convergencia más rápida en el cálculo de integrales, consiste en generalizar el procedimiento Δ^2 de Aitken, para tal efecto se supuso que

$$S_{n+1} - S = \rho(S_n - S),$$

por consiguiente

$$S_n - S = C\rho^n.$$

Ahora bien, para ser más precisos se puede suponer que

$$S_n - S = C_1\rho_1 + C_2\rho_2 + \cdots C_k\rho_k, \quad (\text{VI.3.15})$$

con los ρ_i diferentes dos a dos, de donde la diferencia $\mu_n = S_n - S$ satisface una ecuación de diferencias finitas o relación recursiva de la forma

$$\mu_{n+k} + a_1\mu_{n+k-1} + \cdots + a_k\mu_n = 0. \quad (\text{VI.3.16})$$

La teoría de ecuaciones de diferencias finitas, tiene como resultado central, que los ρ_i , $i = 1, \dots, k$; son raíces del polinomio característico de (VI.3.16) dado por

$$\lambda^k + a_1\lambda^{k-1} + \cdots + a_k. \quad (\text{VI.3.17})$$

Se tiene un problema inverso, pues no se conocen los valores de los a_k , pero si los valores de μ_n , los cuales pueden servir para determinar los valores de los a_k a partir del sistema lineal

$$\begin{pmatrix} S_n - S & \cdots & S_{n+k} - S \\ \vdots & & \vdots \\ S_{n+k} - S & \cdots & S_{n+2k} - S \end{pmatrix} \begin{pmatrix} a_k \\ \vdots \\ a_1 \end{pmatrix} = 0. \quad (\text{VI.3.18})$$

Este sistema tiene soluciones no triviales para el sistema lineal homogéneo, por lo tanto el determinante de la matriz es nulo. Efectuando sustracciones sobre las filas de la matriz, se obtiene

$$\det \begin{pmatrix} S_n - S & S_{n+1} - S & \cdots & S_{n+k} - S \\ \Delta S_n & \Delta S_{n+1} & \cdots & \Delta S_{n+k} \\ \vdots & & & \vdots \\ \Delta S_{n+k-1} & \cdots & \cdots & \Delta S_{n-k-1} \end{pmatrix} = 0,$$

luego, se tiene

$$\begin{vmatrix} S_n & S_{n+1} & \cdots & S_{n+k} \\ \Delta S_n & \Delta S_{n+1} & \cdots & \Delta S_{n+k} \\ \vdots & & & \vdots \\ \Delta S_{n+k-1} & \cdots & \cdots & \Delta S_{n-k-1} \end{vmatrix} = S \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \Delta S_n & \Delta S_{n+1} & \cdots & \Delta S_{n+k} \\ \vdots & & & \vdots \\ \Delta S_{n+k-1} & \cdots & \cdots & \Delta S_{n-k-1} \end{vmatrix},$$

efectuando sustracciones sobre la columna de la matriz del lado derecho de la ecuación, se obtiene finalmente

$$S = \frac{\begin{vmatrix} S_n & S_{n+1} & \cdots & S_{n+k} \\ \Delta S_n & \Delta S_{n+1} & \cdots & \Delta S_{n+k} \\ \vdots & & & \vdots \\ \Delta S_{n+k-1} & \cdots & \cdots & \Delta S_{n-k-1} \end{vmatrix}}{\begin{vmatrix} \Delta^2 S_n & \cdots & \Delta^2 S_{n+k-1} \\ \vdots & & \vdots \\ \Delta^2 S_{n+k-1} & \cdots & \Delta^2 S_{n+k-2} \end{vmatrix}} \quad (\text{VI.3.19})$$

Por ultimo la relación (VI.3.19), puede mejorarse si al determinante del numerador se agrega la primera linea a la segunda linea, la segunda linea a la tercera y así sucesivamente, convirtiéndose en

$$S^{(k)} = \frac{\begin{vmatrix} S_n & S_{n+1} & \cdots & S_{n+k} \\ S_{n+1} & S_{n+1} & \cdots & S_{n+k+1} \\ \vdots & & & \vdots \\ S_{n+k} & \cdots & \cdots & S_{n-k} \end{vmatrix}}{\begin{vmatrix} \Delta^2 S_n & \cdots & \Delta^2 S_{n+k-1} \\ \vdots & & \vdots \\ \Delta^2 S_{n+k-1} & \cdots & \Delta^2 S_{n+k-2} \end{vmatrix}}. \quad (\text{VI.3.20})$$

Este resultado constituye una joya desde el punto de vista teórico, pero es una catástrofe, si se quiere implementar numéricamente, las razones son obvias. Por lo tanto es necesario construir un algoritmo que permita determinar $S^{(k)}$, sin necesidad de calcular explícitamente los determinantes encontrados en la última expresión. El método que será explicado constituye el algoritmo *epsilon* o más simplemente ϵ -algoritmo.

Algoritmo Epsilon

El siguiente teorema formulado por *Wynn* en 1956, permite calcular $S^{(k)}$.

Teorema VI.3.1.- *Dados S_0, S_1, S_2, \dots , se define la sucesión $\epsilon_k^{(n)}$ $k = -1, 0, \dots; n = 0, 1, \dots$, de manera recursiva, como*

$$\begin{aligned}\epsilon_{-1}^{(n)} &= 0, \\ \epsilon_0^{(n)} &= S_n, \\ \epsilon_{k+1}^{(n)} &= \epsilon_{k-1}^{(n+1)} + \frac{1}{\epsilon_k^{(n+1)} - \epsilon_k^{(n)}};\end{aligned}\tag{VI.3.21}$$

entonces

$$\epsilon_2^{(n)} = S'_n, \quad \epsilon_4^{(n)} = S''_n, \quad \epsilon_6^{(n)} = S_n^{(3)}, \dots\tag{VI.3.22}$$

Demostración.- Una demostración completa y una explicación detallada puede encontrarse en Brezinski. \square

La sucesión definida por el teorema precedente permite formular el ϵ -algoritmo en forma de un tablero, ver la figura VI.3.1.

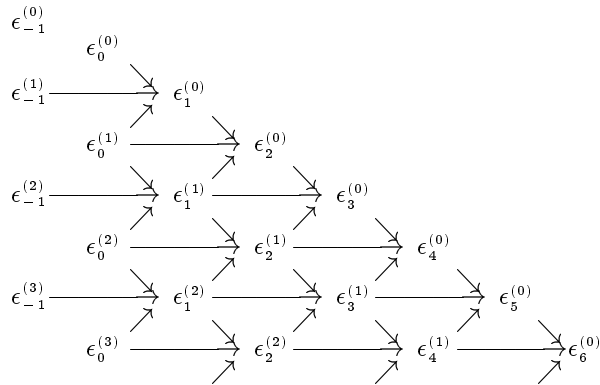


Figura VI.3.1. Esquema ϵ -algoritmo.

En la figura VI.3.4, podrá apreciarse la verdadera potencia del ϵ -algoritmo. La sucesión definida por

$$S_n = 4 \sum_{k=0}^n \frac{(-1)^k}{2k+1}, \quad (\text{VI.3.23})$$

converge hacia π , sin embargo la convergencia de esta sucesión es muy lenta. Para obtener una precisión de 10^{-35} , son necesarias por lo menos 10^{35} evaluaciones de esta sucesión, lo cual es imposible: por el tiempo de cálculo y por el error de redondeo. Aplicando el *epsilon*-algoritmo, se obtiene la precisión requerida, los errores de $\epsilon_k^{(n)}$ son dados en la figura VI.3.4.

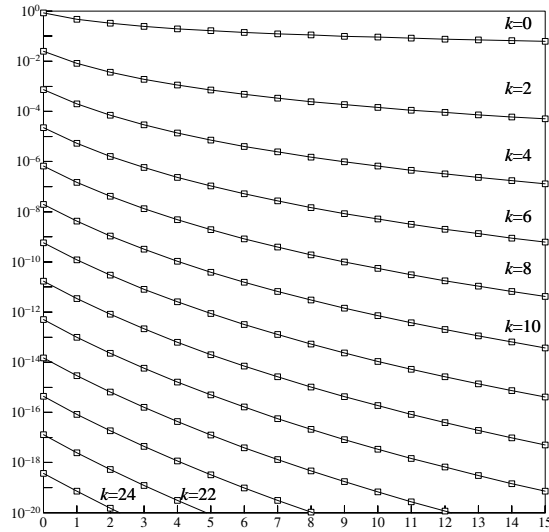


Figura VI.3.4. Error de $\epsilon_k^{(n)}$ en función de n .

Otro medio para acelerar la convergencia en el cálculo de integrales impropias, consiste en efectuar un cambio de variable conveniente. A continuación se presentará algunos ejemplos donde el cálculo de la integral converge con mas rapidez o mayor lentitud dependiendo del cambio de variable elegido.

Ejemplos

a) Considérese, la integral impropia

$$\int_0^1 \frac{\log x}{x^2 + 100} dx.$$

Se observa inmediatamente que $f(x)$ no está acotada en las proximidades del origen. Aplicando la función **GAUINT** con una tolerancia igual a $\text{TOL} = 10^{-14}$. Si se desea obtener el valor exacto de esta integral con un error exacto inferior a 10^{-13} son necesarias 69×15 evaluaciones de la función f .

Efectuando el cambio de variable $x = t^2$, se obtiene la integral

$$\int_0^1 \frac{4t \log t}{t^4 + 100} dt,$$

integral que ha sido ya evaluada, ver la tabla VI.3.2. La función a integrar es acotada, y son necesarias 27×15 evaluaciones de la función integrada. Si nuevamente se realiza otro cambio de variable, como por ejemplo, $t = s^2$, se obtiene la integral

$$\int_0^1 \frac{16s^3 \log s}{s^8 + 100} ds,$$

denotando por $h(s)$ a la función integrada, se puede mostrar que $h(s)$ es dos veces continuamente diferenciable. Resolviendo por la función **GAUINT** son necesarias 7×15 evaluaciones de h .

b) Las integrales de la forma

$$\int_0^1 \frac{f(x)}{\sqrt{x}} dx,$$

pueden ser calculadas con menos evaluaciones, si se hace el cambio de variable $x = t^2$, obteniendo así

$$2 \int_0^1 f(t^2) dt.$$

c) Las integrales impropias del tipo

$$\int_1^\infty f(x) dx,$$

mediante el cambio de variable $x = 1/t$ se convierten en

$$\int_0^1 f(1/t) \frac{1}{t^2} dt.$$

Por lo tanto, para acelerar la convergencia, puede utilizarse de manera combinada el ϵ -algoritmo, con un cambio de variable adecuado. Vale la pena

recaltar que el objetivo principal del cambio de variable es volver la función integrada más lisa, es decir que sea lo suficientemente derivable para poder aplicar la subrutina **GAUINT** en el máximo de su eficiencia. Sin embargo el cambio de variable puede volver más complicada la función a integrar, motivo por el cual la ganancia obtenida en una disminución de evaluaciones de la función integrada puede perderse con las mismas evaluaciones de la función.

Uno de los tipos de integral donde mejor se ajusta los métodos de aceleración propuestos, como el cambio de variable conveniente o el algoritmo epsilon consiste en:

Funciones con Oscilaciones

Escapando un poco a la rutina del libro de presentar las bases teóricas de la solución de un problema, para luego tratar algunos ejemplos, se analizará este tipo de evaluación de integral impropia con un ejemplo.

Considérese la integral de Fresnel, dada por

$$\int_0^{\infty} \sin(x^2) dx = \frac{1}{2} \sqrt{\frac{\pi}{2}}.$$

La función $\sin(x^2)$ se anula en $x^2 = k\pi$ con $k \in \mathbb{N}$, definiendo así una sucesión de números positivos $\{x_k\}$, dada por

$$x_k = \sqrt{k\pi}.$$

Planteando

$$I_k = \int_{x_k}^{x_{k+1}} \sin(x^2) dx, \quad k = 0, 1, 2, \dots;$$

donde I_k pueden ser calculadas por **GAUINT** se define la sucesión $\{S_k\}$, por

$$S_k = \sum_{j=0}^k I_j,$$

teniendo como resultado

$$\int_0^{\infty} \sin(x^2) dx = \lim_{k \rightarrow \infty} S_k.$$

Ahora bien la convergencia de S_k es muy lenta, utilizando ϵ -algoritmo la velocidad de la convergencia hacia la integral, se aumenta de manera ostensible. Ver la tabla VI.3.4.

Tabla VI.3.4. Cálculo de la integral de Fresnel

k	S_k	S'_k	S''_k	$S^{(3)}$	$S^{(4)}$	$S^{(5)}$
0	.89483147	.63252334	.62682808	.62666213	.62665722	.62665721
1	.43040772	.62447449	.62660885	.62665582	.62665582	
2	.78825896	.62773451	.62667509	.62665746	.62665746	
3	.48624702	.62603581	.62664903	.62665692		
4	.75244267	.62705261	.62666112	.62665713		
5	.51172983	.62638728	.62665483			
6	.73311637	.62685063	.62665838			
7	.52703834	.62651269				
8	.72060138	.62676812				
9	.53751806					
10	.71165881					

Ejercicios

1.- Para una sucesión $\{S_n\}_{n \geq 0}$, el ϵ -algoritmo está definido por:

$$\begin{aligned}\epsilon_{-1}^{(n)} &= 0, \\ \epsilon_0^{(n)} &= S_n, \\ \epsilon_{k+1}^{(n)} &= \epsilon_{k-1}^{(n+1)} + \frac{1}{\epsilon_k^{(n+1)} - \epsilon_k^{(n)}}.\end{aligned}$$

Si la aplicación del ϵ -algoritmo a $\{S_n\}_{n \geq 0}$ y a $\{\hat{S}_n\}_{n \geq 0} = \{aS_n + b\}$ proporciona respectivamente las cantidades $\epsilon_k^{(n)}$ y $\hat{\epsilon}_k^{(n)}$. Mostrar que

$$\hat{\epsilon}_{2l}^{(n)} = a\epsilon_{2l}^{(n)} + b, \quad \hat{\epsilon}_{2l+1}^{(n)} = \frac{1}{a}\epsilon_{2l+1}^{(n)}.$$

2.- Utilizar el ϵ -algoritmo para el cálculo de las integrales:

$$\text{a) } \frac{1}{100} \int_0^1 x^{-0.99} dx, \quad \text{b) } \int_0^1 \frac{\log x}{\sqrt{x}} dx.$$

3.- Supóngase que la sucesión $\{S_n\}$ satisface

$$S_{n+1} - S = (\rho + \alpha_n)(S_n - S)$$

con $|\rho| < 1$, $\lim_{n \rightarrow \infty} \alpha_n = 0$ y considérese el procedimiento Δ^2 de Aitken

$$S'_n = S_{n+1} - \frac{\Delta S_n \Delta S_{n+1}}{\Delta^2 S_n}, \quad n = 0, 1, \dots$$

Demostrar que la sucesión $\{S'_n\}$ converge más rápidamente hacia S que la sucesión $\{S_n\}$; es decir

$$\lim_{n \rightarrow \infty} \frac{S'_n - S}{S_n - S} = 0.$$

Indicación.- Verificar que $\Delta S_n = (\rho - 1 + \alpha_n)(S_n - S)$ y encontrar una fórmula similar para $\Delta^2 S_n$

VI.4 Transformación de Fourier

En esta sección será abordado el cálculo numérico de las transformadas de Fourier, es decir los coeficientes de las series de Fourier para una determinada función. Se iniciará un repaso teórico sobre las series de Fourier, luego se introducirá la transformada discreta de Fourier, cuya abreviación usual es *TDF*, para finalmente ver la transformación rápida de Fourier más conocida como *FFT*.

Las motivaciones de la utilización de series de Fourier están dadas por sus diferentes aplicaciones en numerosas áreas de la ciencia, como de la tecnología; para citar algunas de ellas, se tiene el tratamiento de señales, la resolución de ecuaciones diferenciales, la construcción de métodos espectrales en la resolución de ecuaciones a derivadas parciales, etc.

La teoría de la transformación de Fourier está íntimamente ligada a las funciones 2π -periódicas e integrables. En este libro se supondrá que las funciones son integrables en el sentido de Riemann, y no se considerará el caso más general. Recordando la:

Definición VI.4.1.- La serie de Fourier de una función 2π -periódica e integrable, está dada de manera formal por

$$f(x) \sim \sum_{k \in \mathbb{Z}} \hat{f}(k) e^{ikx}, \quad (\text{VI.4.1})$$

donde los coeficientes de Fourier están definidos por

$$\hat{f}(k) = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx. \quad (\text{VI.4.2})$$

Denotando por \mathcal{E} , el espacio de las funciones 2π -periódicas, tales que

$$\int_0^{2\pi} f(x) \overline{f(x)} dx < \infty, \quad \forall f \in \mathcal{E},$$

se tiene que \mathcal{E} es un espacio vectorial provisto del producto sesquilinear, dado por

$$\langle f, g \rangle = \int_0^{2\pi} f(x) \overline{g(x)} dx. \quad (\text{VI.4.3})$$

Una simple verificación muestra que las funciones definidas por

$$\varphi_k(x) = e^{ikx}, \quad (\text{VI.4.4})$$

constituyen una familia ortogonal de funciones, es decir

$$\langle \varphi_k, \varphi_j \rangle = 0, \quad \text{si } j \neq k.$$

Para conocer más respecto a las propiedades de espacios de Hilbert, familias ortonormales y series de Fourier existe una abundante bibliografía, por ejemplo Rudin.

La definición VII.4.1 es una definición formal, es decir la serie de Fourier de una función dada, no necesariamente debe converger hacia la función. Sin embargo existen condiciones suficientes sobre la función f , para que la serie de Fourier converga hacia f o por lo menos en casi todos los puntos. A continuación se enunciará estas condiciones suficientes y el tipo de convergencia que uno puede esperar obtener.

Teorema VI.4.2.- Dirichlet. *Sea $f : \mathbb{R} \rightarrow \mathbb{C}$ una función de clase \mathcal{C}^1 por trozos y periódica de periodo 2π . La serie de Fourier de f es convergente en todo punto de \mathbb{R} . En un punto x donde la función es continua, el límite de la serie es $f(x)$. En un punto x donde f no es continua, la suma de la serie es*

$$\frac{1}{2}(f(x_-) + f(x_+)). \quad (\text{VI.4.5})$$

Además, la convergencia de la serie de Fourier de f es uniforme en todo intervalo compacto que no contiene ningún punto de discontinuidad de f .

Demostración.- Una demostración de este teorema puede encontrarse en Gramain. \square

Con la formulación de este teorema, se conoce la clase de funciones de las cuales la serie de Fourier es igual a la función, en todo caso en los puntos donde la función es continua. Es propósito de esta sección estudiar los métodos numéricos que permitan calcular los coeficientes de Fourier, y por ende la serie de Fourier asociada. Una primera alternativa de cálculo de estos coeficientes consiste en utilizar un método de integración propuesto en las secciones precedentes de este capítulo. Sin embargo existen alternativas menos costosas y más simples que dan excelentes resultados.

Sea $f : [0, 2\pi] \rightarrow \mathbb{C}$, supóngase que la función $f(x)$ es conocida para los x dados por la subdivisión equidistante

$$x_l = \frac{2\pi l}{N}, \quad l = 0, 1, \dots, N. \quad (\text{VI.4.6})$$

Como $f(x_N) = f(x_0)$ por hipótesis, el cálculo de (VI.4.2) puede realizarse mediante la regla del trapecio, obteniendo como aproximación de $\hat{f}(k)$

$$\hat{f}_N(k) = \frac{1}{N} \sum_{l=0}^{N-1} f(x_l) e^{-ikx_l}. \quad (\text{VI.4.7})$$

Ahora bien, (VI.4.7) induce las definiciones siguientes.

Considérese, el espacio de las sucesiones N -periódicas

$$\mathcal{P}_N = \{(y_k)_{k \in \mathbb{Z}} | y_k \in \mathbb{C}, y_{k+N} = y_k\}. \quad (\text{VI.4.8})$$

Definición VI.4.3.- La transformada discreta de Fourier (DFT) de $y \in \mathcal{P}_N$ es la sucesión $(z_k)_{k \in \mathbb{Z}}$, donde

$$z_k = \frac{1}{N} \sum_{l=0}^{N-1} y_l e^{-ikx_l} = \frac{1}{N} \sum_{l=0}^{N-1} y_l \omega^{-kl}, \quad \text{con} \quad \omega = e^{2i\pi/N}.$$

Se la denota $z = \mathcal{F}_N y$.

Proposición VI.4.4.- La transformada discreta de Fourier satisface las siguientes propiedades:

- a) Para $y \in \mathcal{P}_N$, se tiene que $\mathcal{F}_N y \in \mathcal{P}_N$.
- b) La aplicación $\mathcal{F}_N : \mathcal{P}_N \rightarrow \mathcal{P}_N$ es lineal y biyectiva.
- c) La aplicación inversa de \mathcal{F}_N está dada por

$$\mathcal{F}_N^{-1} = N \cdot \bar{\mathcal{F}}_N, \quad (\text{VI.4.9})$$

donde

$$(\bar{\mathcal{F}}_N z)_k := \overline{(\mathcal{F}_N \bar{z})_k} = \frac{1}{N} \sum_{l=0}^{N-1} z_l \omega^{kl}. \quad (\text{VI.4.10})$$

Demostración.- Utilizando el hecho que $\omega^N = e^{2\pi i} = 1$ y $\omega^{-lN} = (\omega^N)^{-l} = 1$, se obtiene

$$z_{k+N} = \frac{1}{N} \sum_{l=0}^{N-1} y_l \omega^{-(k+N)l} = \frac{1}{N} \sum_{l=0}^{N-1} y_l \omega^{-kl} = z_k,$$

mostrando así la periodicidad de z_k . La linealidad de \mathcal{F}_N resulta de una verificación inmediata. Para mostrar la biyectividad y al mismo tiempo la fórmula (VI.4.10), se calcula

$$\begin{aligned} (\bar{\mathcal{F}}_N \mathcal{F}_N y)_j &= \frac{1}{N} \sum_{k=0}^{N-1} (\mathcal{F}_N y)_k \omega^{kj} \\ &= \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} y_l \omega^{-kl} \omega^{kj} \\ &= \frac{1}{N^2} \sum_{l=0}^{N-1} y_l \left(\frac{1}{N} \sum_{k=0}^{N-1} \omega^{k(j-l)} \right) \\ &= \frac{1}{N} y_j. \end{aligned}$$

La última igualdad de este cálculo, es consecuencia de

$$\sum_{k=0}^{N-1} \omega^{km} = \sum_{k=0}^{N-1} (\omega^m)^k = \begin{cases} N & \text{si } m = 0 \bmod N \\ \frac{\omega^{mN}-1}{\omega^m-1} = 0 & \text{si no.} \end{cases}$$

Hay que remarcar que $\omega^m = 1$ si $m = 0 \bmod N$. \square

Estudio del Error

Supóngase que

$$y_l = f(x_l), \quad x_l = \frac{2\pi l}{N}, \quad l = 0, 1, \dots, N;$$

para una función $f : \mathbb{R} \rightarrow \mathbb{C}$ que es 2π -periódica. La fórmula siguiente describe cómo la transformada de Fourier discreta dada por (VI.4.7) aproxima los coeficientes de Fourier dados por (VI.4.2).

Teorema VI.4.5.- *Si la serie $\sum_{k \in \mathbb{Z}} \hat{f}(k)$ es absolutamente convergente, entonces*

$$\hat{f}_N(k) - \hat{f}(k) = \sum_{\substack{j \in \mathbb{Z} \\ j \neq 0}} \hat{f}(k + jN). \quad (\text{VI.4.11})$$

Demostración.- La hipótesis sobre los coeficientes de Fourier implica que se tenga igualdad en la fórmula (VI.4.1), ver Gramain. Por lo tanto, se tiene

$$\begin{aligned} \hat{f}_N(k) &= \frac{1}{N} \sum_{l=0}^{N-1} \left(\sum_{n \in \mathbb{Z}} \hat{f}(n) e^{inx_l} \right) \omega^{-kl} = \sum_{n \in \mathbb{Z}} \hat{f}(n) \underbrace{\left(\frac{1}{N} \sum_{l=0}^{N-1} \omega^{(n-k)l} \right)}_{= \begin{cases} 1 & \text{si } n = k \pmod{N} \\ 0 & \text{si no} \end{cases}} \\ &= \sum_{j \in \mathbb{Z}} \hat{f}(k + jN) \end{aligned}$$

\square

Corolario VI.4.6.- *Sea $f : \mathbb{R} \rightarrow \mathbb{C}$, p veces continuamente derivable ($p \geq 2$) y 2π -periódica. Entonces,*

$$\hat{f}_N(k) - \hat{f}(k) = \mathcal{O}(N^{-p}), \quad \text{para } |k| \leq \frac{N}{2}. \quad (\text{VI.4.12})$$

En particular, con $h = 2\pi/N$, se tiene

$$\frac{h}{2\pi} \sum_{j=0}^{N-1} f(x_j) - \frac{1}{2\pi} \int_0^{2\pi} f(x) dx = \mathcal{O}(h^p), \quad (\text{VI.4.13})$$

lo que significa que, para funciones lisas y periódicas, la fórmula del trapecio es muy precisa.

Demostración.- Se mostrará primero que los coeficientes de Fourier satisfacen

$$\left| \hat{f}(k) \right| \leq C \cdot k^{-p}. \quad (\text{VI.4.14})$$

En efecto, varias integraciones por partes dan

$$\begin{aligned} \hat{f}(k) &= \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx \\ &= \underbrace{f(x) \frac{e^{-ikx}}{-ik} \Big|_0^{2\pi}}_0 + \frac{(i\pi)^{-1}}{2\pi} \int_0^{2\pi} f'(x) e^{-ikx} dx \\ &\vdots \\ &= \frac{(ik)^{-p}}{2\pi} \int_0^{2\pi} f^{(p)}(x) e^{-ikx} dx \end{aligned}$$

teniendo así, (VI.4.14) con $C = \frac{1}{2\pi} \int_0^{2\pi} |f^{(p)}(x)| dx$.

Para $|k| \leq N/2$ y $j \neq 0$ se tiene que $|k + jN| \geq (|j| - 1/2)N$, utilizando (VI.4.11), se obtiene

$$\left| \hat{f}_N(k) - \hat{f}(k) \right| \leq \sum_{j \geq 1} C(j - 1/2)^{-p} N^{-p} = C_1 \cdot N^{-p}.$$

Obsérvese que la serie en esta fórmula converge para $p > 1$. \square

Es muy importante remarcar que $\hat{f}_n(k)$ es una sucesión N -periódica, propiedad de la transformada discreta de Fourier, y que por otro lado $\hat{f}(k)$ converge muy rápidamente hacia 0 por (VI.4.14). Por consiguiente para k grande, por ejemplo $k \approx N$, \hat{f}_N es una mala aproximación de $\hat{f}(k)$; mientras que para $|k| \leq N/2$ la aproximación es en general muy buena.

Interpolación Trigonométrica

Para la división equidistante (VI.4.6) del intervalo $[0, 2\pi]$ y para y_0, y_1, \dots, y_{N-1} dados, se busca un polinomio trigonométrico, es decir una

combinación lineal finita de funciones e^{ikx} , que pase por (x_l, y_l) , para $l = 0, 1, \dots, N-1$. La existencia de tal polinomio trigonométrico está asegurada por el siguiente:

Teorema VI.4.7.- *Sea $y \in \mathcal{P}_N$ y $z = \mathcal{F}_N y$ su transformada discreta de Fourier. Entonces, el polinomio trigonométrico*

$$p_N(x) = \sum_{k=-N/2}^{N/2-1} z_k e^{ikx} := \frac{1}{2} \left(z_{-N/2} e^{-iNx/2} + z_{N/2} e^{iNx/2} \right) + \sum_{|k| < N/2} z_k e^{ikx}, \quad (\text{VI.4.15})$$

satisface $p_N(x_l) = y_l$ para $l = 0, 1, \dots, N-1$.

Hay que remarcar que si los y_k son reales, $\{z_k\}$ es una sucesión hermítica, es decir $z_{-k} = \bar{z}_k$ y por lo tanto el polinomio $p_N(x)$ es un polinomio a coeficientes reales.

Demostración.- Para l fijo, la sucesión $\{z_k e^{ikx_l}\}$ es N -periódica, por consiguiente

$$p_N(x_l) = \sum_{k=0}^{N-1} z_k e^{ikx_l} = N \cdot (\bar{\mathcal{F}}_N z)_l = N (\bar{\mathcal{F}}_N \mathcal{F}_N y)_l = y_l.$$

□

El siguiente teorema a ser enunciado provee una estimación del error de la interpolación trigonométrica con consecuencias importantes, que serán explicadas posteriormente.

Teorema VI.4.8.- *Sea $f : \mathbb{R} \rightarrow \mathbb{C}$ una función 2π -periódica tal que $\sum_{k \in \mathbb{Z}} \hat{f}(k)$ sea absolutamente convergente. Entonces, el polinomio trigonométrico dado por (VI.4.15), para $y_l = f(x_l)$ satisface para todo $x \in \mathbb{R}$*

$$|p_N(x) - f(x)| \leq 2p_N(x) = \sum_{|k| \geq N/2} |\hat{f}(k)|. \quad (\text{VI.4.16})$$

Demostración.- Restando (VI.4.1) de (VI.4.15) se obtiene

$$p_N(x) - f(x) = \sum_{k=-N/2}^{N/2-1} \left(\hat{f}_N(k) - \hat{f}(k) \right) e^{ikx} = \sum_{|k| \geq N/2} \hat{f}(k) e^{ikx}.$$

La aserción es pues consecuencia de (VI.4.11) y de la desigualdad del triángulo □

Este teorema permite una interpretación interesante. Considérese una función 2π -periódica de frecuencia maximal M , es decir $\hat{f}(k) = 0$ para $|k| > M$. Entonces, el polinomio trigonométrico da el resultado exacto $p_N(x) = f(x)$ para todo x , si

$$N > 2M. \quad (\text{VI.4.17})$$

Este resultado, el *Teorema del Muestreo*, da una fórmula para el número de muestras necesarias para obtener una representación exacta de una función.

La evaluación de \mathcal{F}_N requiere N^2 multiplicaciones y adiciones, si se la realiza directamente. Sin embargo existe un procedimiento que permite descender el costo en operaciones a $N \log_2 N$. Este procedimiento será visto en la siguiente subsección.

Transformación Rápida de Fourier (FFT)

El algoritmo que será estudiado, se debe a Cooley & Tukey en 1965, se basa sobre las ideas de Runge 1925. Para poder formular éste, es necesario la siguiente:

Proposición VI.4.9.- Sean $u = (u_0, u_1, \dots, u_{N-1}) \in \mathcal{P}_N$, $v = (v_0, v_1, \dots, v_{N-1}) \in \mathcal{P}_N$ y defínase

$$y = (u_0, v_0, u_1, v_1, \dots, u_{N-1}, v_{N-1}) \in \mathcal{P}_{2N}. \quad (\text{VI.4.18})$$

Entonces, para $k = 0, 1, \dots, N-1$, se tiene $(\omega_{2N} = e^{2i\pi/2N} = e^{i\pi/N})$

$$\begin{aligned} 2N(\mathcal{F}_{2N}y)_k &= N(\mathcal{F}_Nu)_k + \omega_{2N}^{-k}N(\mathcal{F}_Nv)_k, \\ 2N(\mathcal{F}_{2N}y)_{k+N} &= N(\mathcal{F}_Nu)_k - \omega_{2N}^{-k}N(\mathcal{F}_Nv)_k. \end{aligned} \quad (\text{VI.4.18})$$

Demostración.- Utilizando el hecho que $\omega_{2N}^2 = \omega_N$, un cálculo directo da para k arbitrario

$$\begin{aligned} 2N(\mathcal{F}_{2N}y)_k &= \sum_{j=0}^{2N-1} y_j e^{-2\pi i j k / 2N} \\ &= \sum_{j=0}^{2N-1} y_j \omega_{2N}^{-jk} \\ &= \sum_{l=0}^{N-1} \underbrace{y_{2l}}_{u_l} \underbrace{\omega_{2N}^{-2lk}}_{\omega_N^{-lk}} + \sum_{l=0}^{N-1} \underbrace{y_{2l+1}}_{v_l} \underbrace{\omega_{2N}^{-(2l+1)k}}_{\omega_{2N}^{-k} \dots \omega_N^{-lk}} \\ &= N(\mathcal{F}_Nu)_k + \omega_{2N}^{-k}N(\mathcal{F}_Nv)_k. \end{aligned}$$

La segunda fórmula de (VI.4.18) resulta de $\omega_{2N}^{-N} = -1$. \square

La fórmula (VI.4.18) permite calcular, con N multiplicaciones y $2N$ adiciones, la transformada discreta de Fourier de $y \in \mathcal{P}_{2N}$ a partir de $\mathcal{F}_N u$ y $\mathcal{F}_N v$. El mismo procedimiento puede ser aplicado recursivamente a las sucesiones u y v , si éstas tienen una longitud par.

Si se supone que $N = 2^m$, se obtiene el algoritmo presentado en el esquema siguiente (para $N = 8 = 2^3$).

$$\begin{array}{c}
 \mathcal{F}_N \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix} \left\langle \begin{array}{c} F_{N/2} \begin{pmatrix} y_0 \\ y_2 \\ y_4 \\ y_6 \end{pmatrix} \left\langle \begin{array}{c} F_{N/4} \begin{pmatrix} y_0 \\ y_4 \end{pmatrix} \left\langle \begin{array}{l} \mathcal{F}_{N/8} y_0 = y_0 \\ \mathcal{F}_{N/8} y_4 = y_4 \end{array} \right. \\ F_{N/4} \begin{pmatrix} y_2 \\ y_6 \end{pmatrix} \left\langle \begin{array}{l} \mathcal{F}_{N/8} y_2 = y_2 \\ \mathcal{F}_{N/8} y_6 = y_6 \end{array} \right. \\ F_{N/2} \begin{pmatrix} y_1 \\ y_3 \\ y_5 \\ y_7 \end{pmatrix} \left\langle \begin{array}{c} F_{N/4} \begin{pmatrix} y_1 \\ y_5 \end{pmatrix} \left\langle \begin{array}{l} \mathcal{F}_{N/8} y_1 = y_1 \\ \mathcal{F}_{N/8} y_5 = y_5 \end{array} \right. \\ F_{N/4} \begin{pmatrix} y_3 \\ y_7 \end{pmatrix} \left\langle \begin{array}{l} \mathcal{F}_{N/8} y_3 = y_3 \\ \mathcal{F}_{N/8} y_7 = y_7 \end{array} \right. \end{array} \right. \end{array} \right. \end{array} \right. \end{array} \quad (VI.4.19)$$

Figura VI.4.1. Esquema para Cálculo de FFT.

La programación de este algoritmo se la realiza en dos etapas. La primera, se ordena los y_i en el orden exigido por (VI.4.19), es decir es necesario invertir los bits en la representación binaria de los índices:

$$\begin{array}{ll}
 0=(0,0,0) & 0=(0,0,0) \\
 1=(0,0,1) & 4=(1,0,0) \\
 2=(0,1,0) & 2=(0,1,0) \\
 3=(0,1,1) & \longleftrightarrow 6=(1,1,0) \\
 4=(1,0,0) & 1=(0,0,1) \\
 5=(1,0,1) & 5=(1,0,1) \\
 6=(1,1,0) & 3=(0,1,1) \\
 7=(1,1,1) & 7=(1,1,1)
 \end{array}$$

Después, se efectúa las operaciones de (VI.4.18) de la manera como indica el esquema (VI.4.19).

Para pasar de una columna a otra en el esquema (VI.4.19) son necesarias $N/2$ multiplicaciones complejas y de otras N adiciones o sustracciones. Como $m = \log_2 N$ pasajes son necesarios, entonces se tiene el:

Teorema VI.4.10.- Para $N = 2^m$, el cálculo de $\mathcal{F}_N y$ puede ser realizado con:

$$\begin{array}{ll} \frac{N}{2} \log_2 N & \text{multiplicaciones complejas y} \\ N \log_2 N & \text{adiciones complejas.} \end{array}$$

Para ilustrar mejor la importancia de este algoritmo, ver la tabla VI.4.1 para comparar el cálculo de $\mathcal{F}_N y$ con o sin FFT.

Tabla VI.4.1. Comparación de FFT con DFT.

N	N^2	$N \log_2 N$	cociente
$2^5 = 32$	$\approx 10^3$	160	≈ 6.4
$2^{10} \approx 10^3$	$\approx 10^6$	$\approx 10^4$	100
$2^{20} \approx 10^6$	$\approx 10^{12}$	$\approx 2 \cdot 10^7$	$5 \cdot 10^4$

Aplicaciones de la FFT

La transformada rápida de Fourier, tiene una gran cantidad de aplicaciones, desde el cálculo de espectrogramas, resolución de ecuaciones diferenciales ordinarias o a derivadas parciales, hasta la solución de sistemas lineales.

Definiendo el producto de convolución de dos sucesiones N -periódicas $y \in \mathcal{P}_N$ y $z \in \mathcal{P}_n$, por

$$(y * z)_k = \sum_{l=0}^{N-1} y_{k-l} z_l. \quad (\text{VI.4.20})$$

Se tiene la siguiente propiedad, ver ejercicio 1,

$$\mathcal{F}_N(y * z) = N \cdot \mathcal{F}_N y \cdot \mathcal{F}_N z, \quad (\text{VI.4.21})$$

de donde (VI.4.20) puede ser calculado mediante $\mathcal{O}(N \log_2 N)$ operaciones.

La resolución de un sistema lineal con una matriz de Toeplitz circular puede ser resuelto utilizando FFT. En efecto, un tal sistema es de la forma

$$\begin{pmatrix} a_0 & a_{N-1} & a_{N-2} & \cdots & a_1 \\ a_1 & a_0 & a_{N-1} & \cdots & a_2 \\ a_2 & a_1 & a_0 & \cdots & a_3 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{N-1} & a_{N-2} & a_{N-3} & \cdots & a_0 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{N-1} \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{N-1} \end{pmatrix}. \quad (\text{VI.4.22})$$

Evidentemente, el sistema lineal (VI.4.22) es equivalente a $a * x = b$, si se considera (a_i) , (x_i) y (b_i) como sucesiones de \mathcal{P}_N .

La multiplicación de una matriz de Toeplitz arbitraria con un vector

$$\begin{pmatrix} a_0 & a_{-1} & a_{-2} & \cdots & a_{-N+1} \\ a_1 & a_0 & a_{-1} & \cdots & a_{-N+2} \\ a_2 & a_1 & a_0 & \cdots & a_{-N+3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{N-1} & a_{N-2} & a_{N-3} & \cdots & a_0 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{N-1} \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{N-1} \end{pmatrix}, \quad (\text{VI.4.23})$$

puede ser resuelta utilizando FFT, considerando las sucesiones en \mathcal{P}_{2N}

$$\begin{aligned} a &= (a_0, a_1, \dots, a_{N-1}, 0, a_{-N+1}, a_{-N+2}, \dots, a_{-1}), \\ x &= (x_0, x_1, \dots, x_{N-1}, 0, 0, \dots, 0). \end{aligned}$$

Se puede verificar facilmente que el resultado del producto (VI.4.23) es la primera mitad del producto de convolución $a * x$. Por consiguiente, el cálculo con FFT da un algoritmo rápido para efectuar el producto (VI.4.23).

Ejercicios

1.- Mostrar que

$$\mathcal{F}_n(y * z) = N \cdot \mathcal{F}_N y \cdot \mathcal{F}_N z,$$

para el producto de convolución

$$(y * z)_k = \sum_{l=0}^{N-1} y_{k-l} z_l,$$

de dos sucesiones N -periódicas. Deducir que

$$y * z = N \cdot \mathcal{F}_N^{-1}(\mathcal{F}_N y \cdot \mathcal{F}_N z).$$

2.- Resolver la ecuación diferencial con valores en la frontera

$$u''(x) = -1 \quad u(0) = u(2\pi) = 0,$$

graficar la solución.

Capítulo VII

Ecuaciones Diferenciales

El estudio de una gran cantidad de fenómenos de las más diversas características se traducen en ecuaciones diferenciales. La descripción de un fenómeno mediante ecuaciones diferenciales tiene un propósito primordial que es la predecibilidad. Por otro lado, permite obtener conclusiones de carácter local, que serán extrapoladas para tener informaciones globales del modelo estudiado. El énfasis que se hace a la resolución analítica de las ecuaciones diferenciales en los cursos de Ecuaciones Diferenciales que se dictan en los primeros niveles de las universidades, tienen el objetivo de encontrar soluciones generales a los diversos problemas diferenciales que se encuentran en el transcurso de los estudios universitarios, como también en el ejercicio profesional. Este hecho se debe fundamentalmente que hasta hace no mucho, no se contaban con los medios tecnológicos que permitan resolver ecuaciones diferenciales con la precisión que se requería. Por consiguiente, el objetivo de estos cursos eran esencialmente obtener las soluciones en forma de fórmulas, perdiéndose así el carácter esencial de las ecuaciones diferenciales que es el estudio local de los fenómenos. Además cuestiones como existencia y unicidad no son abordadas por falta de tiempo.

Este capítulo tiene como objetivo principal la formulación de métodos numéricos de resolución de problemas diferenciales a valores iniciales o problemas de Cauchy. La primera parte tratará sobre cuestiones de existencia y unicidad de las ecuaciones diferenciales. Luego se abordará los métodos a un paso y como expresión de estos; los métodos de Runge-Kutta, desde la construcción de estos, estimaciones de error y como corolario los métodos encajonados del tipo Dormand & Prince. La tercera parte de este capítulo tratará los métodos numéricos a paso múltiple, se verá la construcción de estos, cuestiones de estabilidad y convergencia.

VII.1 Generalidades

En este capítulo I , I_0 designan intervalos abiertos de \mathbb{R} no reducidos a un punto y t_0 un punto fijo de I_0 ; se da una función f definida y continua sobre $I_0 \times \mathbb{R}^m$ con valores en \mathbb{R}^m , un elemento $y_0 \in \mathbb{R}^m$, y se desea encontrar una función y continua y derivable sobre el intervalo I_0 , con valores en \mathbb{R}^m , tal que:

$$y'(t) = f(t, y(t)), \quad \forall t \in I_0; \quad (\text{VII.1.1})$$

$$y(t_0) = y_0. \quad (\text{VII.1.2})$$

Este problema se lo conoce con el nombre de problema de Cauchy para el sistema diferencial (VII.1.1); la condición (VII.1.2) se llama una condición de Cauchy. Una función y que satisface el sistema (VII.1.1) es llamada una integral del sistema (VII.1.1). En numerosos ejemplos físicos, la variable t representa el tiempo, el instante t_0 , es por consiguiente, llamado instante inicial y la condición (VII.1.2) llamada condición inicial.

Se puede remarcar que si se denota por y_1, y_2, \dots, y_m las componentes de y , por $f_1(t, y_1, \dots, y_m), \dots, f_m(t, y_1, \dots, y_m)$ las componentes de $f(t, y)$ la ecuación (VII.1.1) es equivalente al sistema

$$\begin{cases} y'_1(t) = f_1(t, y_1(t), \dots, y_m(t)) \\ y'_2(t) = f_2(t, y_1(t), \dots, y_m(t)) \\ \vdots \\ y'_m(t) = f_m(t, y_1(t), \dots, y_m(t)) \end{cases}. \quad (\text{VII.1.3})$$

Las ecuaciones del sistema VII.1.3 son de primer orden, pues en éstas, el orden de derivación más alto que aparece es el primero. Considérese ahora un problema diferencial de orden p , de la forma

$$y^{(p)}(t) = f(t, y(t), y'(t), \dots, y^{(p-1)}(t)), \quad (\text{VII.1.4})$$

el cual puede convertirse en problema de la forma (VII.1.1), planteando

$$z_1(t) = y(t), \quad z_2(t) = y'(t), \dots, z_p(t) = y^{(p-1)}(t);$$

el problema diferencial (VII.1.4), por consiguiente es equivalente al sistema

$$\begin{cases} z'_1(t) = z_2(t) \\ \vdots \\ z'_{p-1}(t) = z_p(t) \\ z'_p(t) = f(t, z_1(t), \dots, z_p(t)) \end{cases}.$$

de donde planteando

$$z = (z_1, z_2, \dots, z_p)^t \quad \text{y} \quad F(t, z) = (z_2, \dots, z_p, f(t, z_1, \dots, z_p))^t,$$

se tiene

$$z'(t) = F(t, z(t)). \quad (\text{VII.1.5})$$

La condición de Cauchy para el problema (VII.1.5) está dada por $y(t_0), y'(t_0), \dots, y^{(p-1)}(t_0)$.

Ahora bien, en este capítulo no será tratado el problema diferencial general de orden p , dado por

$$F(t, y(t), y'(t), \dots, y^{(n)}(t)) = 0, \quad \forall t \in I_0. \quad (\text{VII.1.6})$$

Cuando se puede aplicar el teorema de las funciones implícitas, (VII.1.6) es localmente equivalente a la ecuación de la forma (VII.1.4) y la teoría que será desarrollada en este capítulo podrá ser aplicada a este tipo de problema sin inconvenientes. Si el teorema de las funciones implícitas no es aplicable, serias dificultades matemáticas y numéricas pueden aparecer, en este caso se habla de ecuaciones diferenciales algebraicas, para saber más sobre este tipo de ecuaciones referirse a Hairer & Wanner.

Finalmente es necesario remarcar que, si bien I_0 es un intervalo abierto, el estudio de las soluciones de los problemas diferenciales permite considerar los intervalos de la forma $[t_0, t_0 + T)$ y $(t_0 - T, t_0]$, obteniendo el intervalo abierto por recolamiento de ambos subintervalos semiabiertos.

Teoremas de Existencia y Unicidad

En esta subsección se supondrá que la terminología básica es conocida por el lector. No obstante, se enunciará dos teoremas muy importantes en lo que concierne el análisis numérico de ecuaciones diferenciales. El primer teorema a enunciarse da condiciones suficientes para asegurar la existencia y unicidad de las soluciones de los problemas a valores iniciales. El segundo teorema está relacionado con la condición misma del problema, pues cuando se trata numéricamente un problema es de suponer que se trabaja con una solución aproximada del problema.

Teorema VII.1.1.- *Cauchy-Lipschitz.* Supóngase que f es continua sobre $I_0 \times \mathbb{R}^m$ y que satisface una condición de Lipschitz, es decir que existe un real L tal que

$$\|f(t, z) - f(t, y)\| \leq L \|z - y\| \quad \forall (t, y) \text{ y } (t, z) \in I_0 \times \mathbb{R}^m; \quad (\text{VII.1.7})$$

entonces el problema (VII.1.1,2) admite una solución y una sola.

Demostración.- Se dará una demostración directa que tiene la ventaja de ser válida cuando se reemplaza \mathbb{R}^n por un espacio de Banach.

Para fijar las ideas, supóngase que $I_0 = [t_0, t + t_0]$ y considérese la aplicación Φ que a $y \in \mathcal{C}^0([t_0, t + t_0])$ asocia $\Phi(y) \in \mathcal{C}^0([t_0, t + t_0])$ definida por

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

Introduciendo la norma

$$\|y\|_L = \max_{s \in I_0} (e^{-2L(s-t_0)} \|y(s)\|)$$

que dota $\mathcal{C}^0(I_0)$ de una estructura de espacio de Banach. Se tiene

$$\begin{aligned} \|(\Phi(y) - \Phi(y^*))(t)\| &\leq \int_{t_0}^t \|f(s, y(s)) - f(s, y^*(s))\| ds \\ &\leq \int_{t_0}^t L e^{2L(s-t_0)} ds \|y - y^*\|_L \\ &\leq \frac{1}{2} e^{2L(t-t_0)} \|y - y^*\|_L, \end{aligned}$$

deduciéndose

$$\|\Phi(y) - \Phi(y^*)\|_L \leq \frac{1}{2} \|y - y^*\|_L.$$

El teorema del punto fijo implica que Φ tiene un solo punto fijo en $\mathcal{C}^0(I_0)$, de donde se tiene el resultado. \square

Teorema VII.1.2.- Sea $f : \mathcal{V} \rightarrow \mathbb{R}^n$ continua, donde $\mathcal{V} \subset \mathbb{R}^{n+1}$ abierto. Supóngase que: $y(x)$ es una solución de $y' = f(x, y)$ sobre $[x_0, \bar{x}]$, tal que $y(x_0) = y_0$; $v(x)$ una solución aproximada de la ecuación diferencial sobre $[x_0, \bar{x}_0]$ tal que

$$\|v'(x) - f(x, v(x))\| \leq \delta \quad (\text{VII.1.8})$$

y f satisface una condición de Lipschitz sobre un conjunto que contenga $\{(x, y(x)), (x, z(x))\}$, es decir

$$\|f(x, y) - f(x, z)\| \leq L \|y - z\|. \quad (\text{VII.1.9})$$

Entonces

$$\|y(x) - v(x)\| \leq \|y_0 - v(x_0)\| e^{L(x-x_0)} + \frac{\delta}{L} (e^{L(x-x_0)} - 1). \quad (\text{VII.1.10})$$

Demostración.- Se tiene:

$$\begin{aligned} y(x) - v(x) &= y(x_0) - v(x_0) + \int_{x_0}^x (y'(s) - v'(s)) ds \\ &= y(x_0) - v(x_0) + \int_{x_0}^x (f(s, y(s)) - f(s, v(s)) + f(s, v(s)) - v'(s)) ds, \end{aligned}$$

pasando a las normas y aplicando las hipótesis (VII.1.8) y (VII.1.9), se obtiene

$$\|y(x) - v(x)\| \leq \|y_0 - v(x_0)\| + \int_{x_0}^x (L \|y(s) - v(s)\| + \delta) ds.$$

Planteando

$$u(x) = \|y_0 - v(x_0)\| + \int_{x_0}^x (L \|y(s) - v(s)\| + \delta) ds,$$

se deduce

$$u'(x) = L \|y(x) - v(x)\| + \delta \leq Lu(x) + \delta,$$

de esta manera se obtiene la desigualdad diferencial

$$\begin{aligned} u'(x) &\leq Lu(x) + \delta \\ u(x_0) &= \|y_0 - v(x_0)\|. \end{aligned} \tag{VII.1.11}$$

Para resolver esta desigualdad, se considera la familia de ecuaciones:

$$w'_n(x) = Lw_n(x) + \delta, \quad w_n(x_0) = u(x_0) + \frac{1}{n}, \tag{VII.1.12}$$

cuyas soluciones están dadas por:

$$w_n(x) = w_n(x_0)e^{L(x-x_0)} + \frac{\delta}{L} \left(e^{L(x-x_0)} - 1 \right).$$

El siguiente paso en la demostración es mostrar que

$$u(x) \leq w_n(x). \tag{VII.1.13}$$

Supóngase lo contrario, es decir que existe un n y $s > x_0$, tales que $u(s) > w_n(s)$. Considérese el conjunto

$$\mathcal{A} = \{x > x_0 | u(x) > w_n(x)\}$$

y sea $x_1 = \inf \mathcal{A}$. Por continuidad se tiene $u(x_1) = w_n(x_1)$. De donde:

$$\begin{aligned} w_n(x_1) - w_n(x_0) &= \int_{x_0}^{x_1} w'_n(s) ds = \int_{x_0}^{x_1} (Lw_n(s) + \delta) ds \\ &\geq \int_{x_0}^{x_1} (Lu(s) + \delta) ds \geq \int_{x_0}^{x_1} u'(s) ds \\ &= u(x_1) - u(x_0), \end{aligned}$$

por lo tanto

$$w(x_0) \leq u(x_0),$$

llegando a una contradicción con $-1/n \geq 0$.

□

Problemas con Valores en la Frontera

La teoría de existencia y unicidad de ecuaciones diferenciales son generalmente formuladas para problemas de Cauchy o problemas a valores iniciales, sin embargo existe una gran variedad de problemas diferenciales de otras características que serán tratados en esta subsección.

Toda ecuación diferencial puede expresarse como un sistema de ecuaciones diferenciales de primer orden de la manera siguiente

$$y' = f(x, y), \quad (\text{VII.1.14})$$

donde $f : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$.

Un problema diferencial con valores en la frontera es darse una ecuación diferencial del tipo (VII.1.14) con n condiciones para $y(a)$ y $y(b)$, donde $a, b \in \mathbb{R}$. Existe una gama de problemas diferenciales con valores en la frontera. A continuación se mostrarán los ejemplos más representativos.

Ejemplos

a) Problemas a valores iniciales. Son de la forma

$$\begin{aligned} y' &= f(x, y), \\ y(x_0) &= y_0. \end{aligned}$$

Este tipo de problema tiene solución única si f es continua y verifica las condiciones de Lipschitz.

b) Considérese el problema

$$\begin{aligned} y'' &= y; & y(a) &= A, \\ & & y(b) &= B. \end{aligned}$$

La ecuación diferencial de segundo orden puede reducirse al siguiente sistema de primer orden

$$\begin{aligned} y_1' &= y_2, \\ y_2' &= y_1; \end{aligned}$$

con condiciones de borde dadas por $y_1(a) = A$ y $y_1(b) = B$. Este problema siempre tiene solución única cuando $a \neq b$.

- c) Encontrar una solución T periódica de una ecuación diferencial, por ejemplo

$$y' = y + \cos x.$$

Es muy fácil deducir que $T = 2\pi k$, con k entero. El problema es encontrar una solución de la ecuación diferencial que satisfaga

$$y(x) = y(x + T).$$

- d) Determinar el parámetro $\lambda \in \mathbb{R}^p$ de la ecuación

$$y' = f(x, y, \lambda),$$

donde la solución buscada verifica $y(a) = y_a$ y $g(y(a))$ con $g : \mathbb{R}^n \rightarrow \mathbb{R}^p$. Ahora bien, este problema puede expresarse como el problema diferencial equivalente

$$\begin{aligned} y' &= f(x, y, \lambda), \\ \lambda' &= 0, \end{aligned}$$

con condiciones de frontera

$$\begin{aligned} y(a) &= y_a, \\ g(y(b)) &= 0. \end{aligned}$$

- e) Problemas a frontera libre, son de la forma

$$\begin{aligned} y'' &= f(x, y, y'); & y(0) &= A, \\ & & y(l) &= B, \\ & & y'(l) &= 0; \end{aligned}$$

con l desconocido. Este problema mediante una transformación afín de x , puede expresarse de la siguiente forma

$$\begin{aligned} z'' &= l^2 f\left(lt, z, \frac{z'}{l}\right), \\ l' &= 0, \end{aligned}$$

con condiciones de borde dadas por

$$z(0) = A, \quad z(1) = B, \quad z'(1) = 0.$$

En base a los ejemplos expuestos más arriba, el problema diferencial con valores en la frontera puede expresarse como

$$\begin{aligned} y' &= f(x, y), \\ r(y(a), y(b)), \end{aligned} \tag{VII.1.15}$$

donde $f : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ y $r : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$.

Por consiguiente la función r en los diferentes ejemplos, será igual a:

$$\begin{aligned} r(y(a), y(b)) &= y(a) - y_a && \text{ejemplo a),} \\ r\left(\begin{pmatrix} y_1(a) \\ y_2(a) \end{pmatrix}, \begin{pmatrix} y_1(b) \\ y_2(b) \end{pmatrix}\right) &= \begin{pmatrix} y_1(a) - A \\ y_1(b) - B \end{pmatrix} && \text{ejemplo b),} \\ r(y(x_0), y(x_0 + T)) &= y(x_0) - y(x_0 + T) && \text{ejemplo c),} \end{aligned}$$

Introduciendo la notación siguiente

$$y(x, a, y_a) \quad (\text{VII.1.16})$$

para expresar la solución $y(x)$ que satisface $y(a) = y_a$, el problema diferencial con condiciones en la frontera consiste en encontrar y_a , tal que

$$r(y_a, y(b, a, y_a)) = 0. \quad (\text{VIII.1.7})$$

Definiendo la función $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ por

$$F(y_a) = r(y_a, y(b, a, y_a)), \quad (\text{VII.1.18})$$

resumiendose el problema diferencial con valores en la frontera a encontrar y_a tal que

$$\begin{aligned} y' &= f(x, y), \\ F(y_a) &= 0. \end{aligned} \quad (\text{VII.1.19})$$

Supóngase que $y^*(x)$ sea una solución de (VII.1.19), es decir $y_a^* = y^*(a)$, además que $F'(y_a^*)$ sea inversible, por el teorema de la inversión local la solución y_a^* es localmente única y por consiguiente $y^*(x)$ lo es también.

La ecuación $F(y_a)$ se resuelve generalmente por un método iterativo, si F no es lineal, se utiliza por ejemplo el método de Newton. Para poder aplicar el método de Newton es necesario conocer $F'(y_a)$. Ahora bien, se tiene

$$F'(y_a) = \frac{\partial r}{\partial y_a}(y_a, y(b, a, y_a)) + \frac{\partial r}{\partial y_b}(y_a, y(b, a, y_a)) \frac{\partial y}{\partial y_a}(b, a, y_a). \quad (\text{VII.1.20})$$

Diferenciabilidad respecto a los Valores Iniciales

En la expresión (VII.1.20) puede observarse que existe una expresión que es derivada respecto al valor inicial y_a . Retomando la notación de la sección precedente se tiene $y(x, x_0, y_0)$ es la solución que pasa por (x_0, y_0) de la

ecuación diferencial $y' = f(x, y)$, donde $f : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$. El problema consiste en determinar

$$\frac{\partial y}{\partial y_0}(x, x_0, y_0). \quad (\text{VII.1.21})$$

En el caso lineal, se tiene que la ecuación diferencial es de la forma

$$y' = A(x)y \quad (\text{VII.1.22})$$

donde $A(x)$ es una matriz de coeficientes $(a_{ij}(x))$ continuos. La solución general de (VII.1.22) está dada por

$$y(x, x_0, y_0) = \sum_{i=1}^n y(x, x_0, e_i) y_{0i} = R(x, x_0) y_0 \quad (\text{VII.1.23})$$

donde los e_i son los vectores de la base canónica de \mathbb{R}^n . La matriz $R(x, x_0)$ se llama el núcleo resolvente o la resolvente de la ecuación (VII.1.22). Es fácil de verificar que

$$\frac{\partial y}{\partial y_0}(x, x_0, y_0) = R(x, x_0) \quad (\text{VII.1.24})$$

Para el caso no lineal la situación es un poco más complicada. Se tiene

$$\frac{\partial y}{\partial x}(x, x_0, y_0) = f(x, y(x, x_0, y_0)) \quad (\text{VII.1.25})$$

suponiendo que $\partial y / \partial y_0$ existe y el orden de derivación conmuta, se obtiene

$$\frac{\partial}{\partial x} \left(\frac{\partial y}{\partial y_0}(x, x_0, y_0) \right) = \frac{\partial f}{\partial y}(x, y(x, x_0, y_0)) \frac{\partial y}{\partial y_0}(x, x_0, y_0)$$

con

$$\frac{\partial y}{\partial y_0}(x_0, x_0, y_0) = I,$$

de donde $\frac{\partial y}{\partial y_0}(x, x_0, y_0)$ es la resolvente de la ecuación diferencial lineal

$$\Psi' = \frac{\partial f}{\partial y}(x, y(x, x_0, y_0)) \Psi. \quad (\text{VII.1.26})$$

Shooting Simple

Retomando el problema con valores en la frontera formulado bajo la forma de las ecuaciones (VII.1.18) y (VII.1.19) puede ser resuelto utilizando el método conocido como *shooting* simple, que en síntesis es utilizar un método

número para resolver (VII.1.19). Este método puede ser Newton u otro método numérico adaptado al problema. Sin pretender dar una teoría que justifique tal método, para tal efecto referirse a Stoer, se mostrará este método implementado en ejemplos.

Ejemplos

a) Considérese el problema con valores en la frontera dado por:

$$\begin{aligned} y'' &= -e^y, \\ y(0) &= 1, \\ y(1) &= \frac{1}{2}. \end{aligned} \tag{VII.1.27}$$

Utilizando la notación de la subsección precedente, se plantea

$$y(x, \alpha),$$

la solución del problema diferencial a valores iniciales

$$\begin{aligned} y'' &= -e^y, \\ y(0) &= 1, \\ y'(0) &= \alpha. \end{aligned} \tag{VII.1.27b}$$

El problema (VII.1.27) se traduce en encontrar α , tal que

$$y(1, \alpha) = \frac{1}{2}.$$

En la figura VII.1.1, puede observarse en la gráfica los valores de $y(1)$ en función de α .

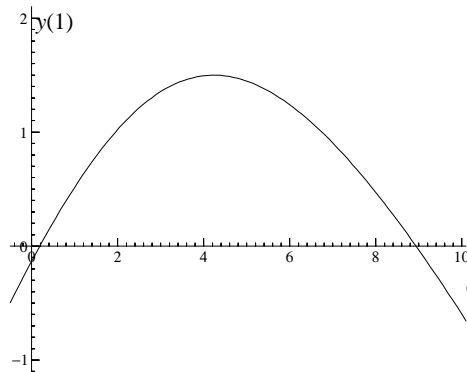


Figura VII.1.1. Valores de $y(1)$ en función de α .

Observando la gráfica, se puede deducir que el problema (VII.1.27) tiene dos soluciones. El siguiente paso es aplicar el método del *shooting* simple, obteniendo de esta manera dos soluciones. La primera con

$$\alpha = 0.9708369956661049,$$

la segunda solución con

$$\alpha = 7.93719815816973.$$

Puede apreciarse las graficas de las dos soluciones, en la figura (VII.1.2)

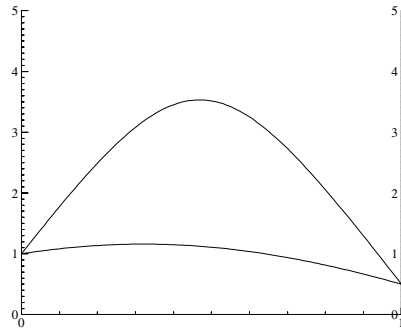


Figura VII.1.2. Soluciones del Problema (VII.1.26).

- b) En este ejemplo se analizará un problema de soluciones periódicas. Considérese la ecuación diferencial de Van der Pol, dada por

$$y'' = (1 - y^2)y' - y. \quad (\text{VII.1.28})$$

El problema consiste en determinar si existe una solución periódica y sobre todo cómo es ésta. Ahora bien, (VII.1.28) puede escribirse como el sistema de ecuaciones diferenciales

$$\begin{aligned} y_1' &= y_2, \\ y_2' &= (1 - y_1^2)y_2 - y_1. \end{aligned} \quad (\text{VII.1.29})$$

Planteando $y(x) = (y_1(x), y_2(x))$, el problema se resume en encontrar $T > 0$, tal que $y(T) = y(0)$, que puede formularse de la siguiente manera

$$F(T, y_0) = y(T, 0, y_0) - y_0 = 0. \quad (\text{VII.1.30})$$

Supóngase que T, y_0 sean una aproximación de la solución del problema (VII.1.30), de donde

$$F(T + \Delta T, y_0 + \Delta y_0) = 0,$$

con ΔT , Δy_0 escogidos convenientemente. Desarrollando en serie de Taylor, se deduce:

$$F(T, y_0) + \frac{\partial F}{\partial T}(T, y_0)\Delta T + \frac{\partial F}{\partial y_0}(T, y_0)\Delta y_0 = 0,$$

$$y(T, 0, y_0) - y_0 + f(y(T, 0, y_0))\Delta T + \left(\frac{\partial y}{\partial y_0}(T, 0, y_0) - I \right) \Delta y_0 = 0.$$

Por consiguiente, se tiene n ecuaciones lineales, con $n + 1$ incognitas, agregando la condición suplementaria

$$\Delta y_0 \perp f(y(T, 0, y_0)), \quad (\text{VII.1.31})$$

se obtiene

$$\begin{pmatrix} \frac{\partial y}{\partial y_0}(T, 0, y_0) - I & f(y(T, 0, y_0)) \\ f^t(y(T, 0, y_0)) & 0 \end{pmatrix} \begin{pmatrix} \Delta y_0 \\ \Delta T \end{pmatrix} = - \begin{pmatrix} y(T, 0, y_0) - y_0 \\ 0 \end{pmatrix}. \quad (\text{VII.1.32})$$

Partiendo de los valores iniciales

$$y_1(0) = 1.,$$

$$y_2(0) = 2.,$$

$$T = 7.;$$

luego de 14 iteraciones se obtiene con una precisión del orden de 10^{-13} , los valores iniciales y el periodo

$$y_1(0) = 2.00861986087296,$$

$$y_2(0) = 0.,$$

$$T = 6.66328685933633.$$

La solución periódica de la ecuación Van der Pol puede apreciarse en la figura VII.1.3

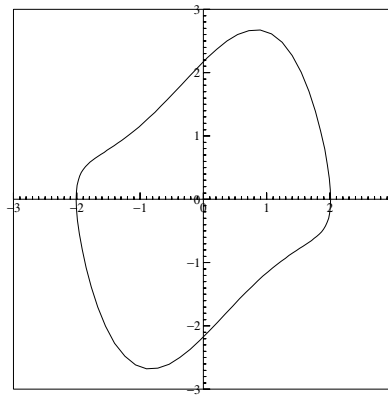


Figura VII.1.3. Soluciones Periodica de Van der Pol.

Shooting Múltiple

La solución del problema con valores en la frontera

$$y' = f(x, y), \quad r(y(a), y(b)) = 0,$$

requiere que para todo punto x de la región donde está definida la solución y se tenga una aproximación numérica de $y(x)$. En el método *shooting* descrito en la anterior subsección, solamente el valor inicial $y(a) = \hat{y}_a$ es determinado. En muchos problemas es suficiente conocer este valor, sin embargo en una gran variedad de problemas con valores en la frontera conocer este valor no es nada fiable. Para ilustrar esto, considérese la ecuación diferencial.

$$y'' - y' + 110y = 0, \quad (\text{VII.1.33})$$

cuya solución general está dada por

$$y(x) = C_1 e^{-10x} + C_2 e^{11x}, \quad (\text{VII.1.34})$$

donde C_1 y C_2 son constantes reales arbitrarias. El problema a valor inicial $y(0) = y_0$ y $y'(0) = y'_0$ tiene como solución

$$y(x) = \frac{11y_0 - y'_0}{21} e^{-10x} + \frac{y'_0 + 10y_0}{21} e^{11x}. \quad (\text{VII.1.35})$$

Ahora bien, considérese el problema con valores en la frontera dado por

$$y(0) = 1, \quad y(10) = 1;$$

la solución de este problema consiste en determinar y'_0 de la fórmula (VII.1.35). Por consiguiente, resolviendo la ecuación lineal

$$\frac{11 - y'_0}{21} e^{-100} + \frac{10 + y'_0}{21} e^{110} = 1;$$

se obtiene

$$y'_0 = \frac{21 - 10e^{110} - 11e^{-100}}{e^{110} - e^{100}}.$$

Debido a la aritmética de punto flotante que las computadoras poseen, en lugar de y'_0 , se manipula la cantidad

$$\bar{y}'_0 = -10(1 + \epsilon), \quad \text{con } |\epsilon| \leq \epsilon ps. \quad (\text{VII.1.36})$$

Suponiendo que los cálculos se hacen en doble precisión, se puede tomar por ejemplo $\epsilon = 10^{-16}$. Remplazando y'_0 en (VII.1.35), se obtiene

$$y(100) \approx \frac{10^{-15}}{21} e^{110} \approx 2.8 \times 10^{31}.$$

Otra de las dificultades mayores en la implementación del método *shooting* en su versión simple, es la necesidad de contar con una buena aproximación de y_a , lo que en general no sucede. Por otra lado, los valores que pueden tomar los y_a , en muchas situaciones, están confinados a regiones demasiado pequeñas; por ejemplo considerese el problema

$$\begin{aligned} y'' &= y^3, \\ y(0) &= 1, \\ y(100) &= 2; \end{aligned} \tag{VII.1.37}$$

los valores que puede tomar $y'(0)$, para que $y(x)$ esté definida en el intervalo $[0, 100]$, están restringidos a un intervalo de longitud no mayor a 10^{-6} . Por este motivo puede deducirse la dificultad de implementar el método *shooting* simple.

El remedio a estas dificultades enumeradas más arriba está en la implementación del método *shooting* múltiple, que consiste en subdividir el intervalo $[a, b]$ en subintervalos de extremidades x_i , es decir tomar una subdivisión $a = x_0 < x_1 < \dots < x_n = b$. Luego, se denota por $y_i = y(x_i)$. De esta manera se obtiene el sistema de ecuaciones

$$\begin{cases} y(x_1, x_0, y_0) - y_1 = 0 \\ y(x_2, x_1, y_1) - y_2 = 0 \\ \vdots \\ y(x_n, x_{n-1}, y_{n-1}) = 0 \\ r(y_0, y_n) = 0. \end{cases} \tag{VII.1.38}$$

La solución de (VII.1.38) se la encuentra utilizando un método iterativo, que puede ser Newton si el problema no es lineal. Como ilustración de este Método se tiene los siguientes dos ejemplos.

Ejemplos

- a) Considérese el problema (VII.1.37). Para su resolución se ha subdividido equidistantemente en 100 subintervalos. La solución del sistema (VII.1.38) se la hecho utilizando el método de Newton. Las iteraciones y las gráficas pueden apreciarse en la figura VII.1.4.

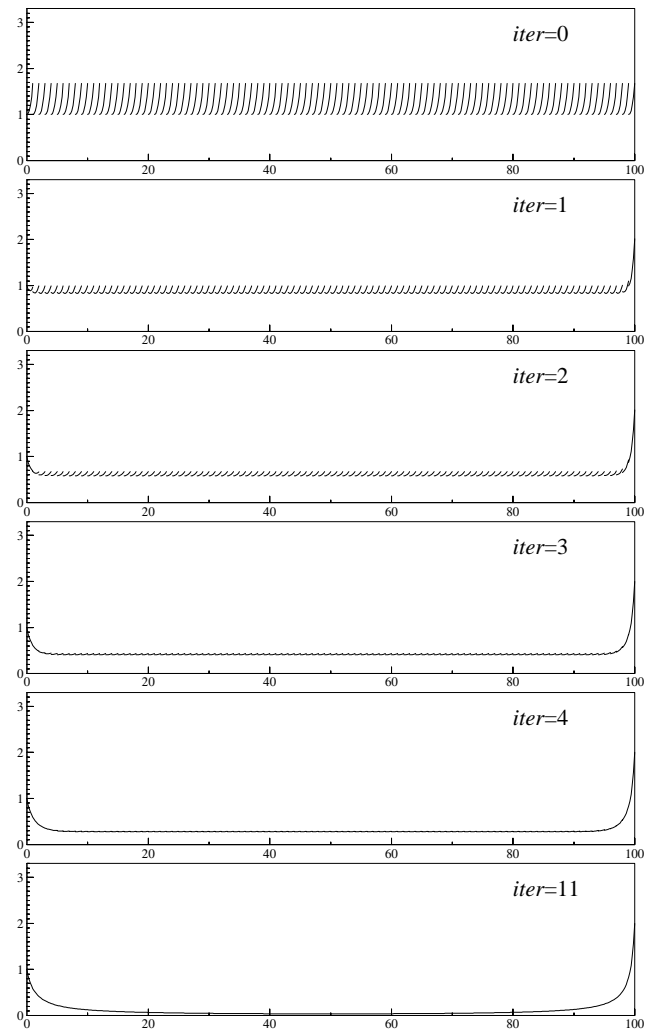


Figura VII.1.4. Implementación Múltiple *Shooting*.

- b) Este ejemplo está relacionado con un problema ingenieril. Un granjero desea un silo cuya base sea un círculo de radio 5 m , de altura 10 m y el techo sea un círculo de radio 2.5 m , la capacidad del silo debe ser exactamente de 550 m^3 . Suponiendo que el costo es proporcional al área del silo. ¿Cuál es la forma de éste?

Matemáticamente el problema puede ser formulado como:

$$\begin{aligned} \text{area lateral} &\longrightarrow \min, \\ \text{volumen} &= 550. \end{aligned}$$

Por la simetría del problema, se puede deducir que el silo es una superficie de revolución, por lo tanto el problema puede expresarse de la manera siguiente: Encontrar una función $y(x)$, tal que

$$\begin{aligned} \int_0^{10} y \sqrt{1 + y'^2} dx &\longrightarrow \min, \\ \pi \int_0^{10} y^2 dx &= 550, \\ y(0) &= 5, \\ y(10) &= 2.5. \end{aligned}$$

Planteando

$$L(\lambda, y, y') = y \sqrt{1 + y'^2} - \lambda \left(y^2 - \frac{55}{\pi} \right),$$

el problema es equivalente, por los Multiplicadores de Lagrange a

$$\int_0^{10} L(\lambda, y, y') dx \rightarrow \min.$$

Este problema es de tipo variacional. Aplicando las ecuaciones de Euler-Lagrange se convierte en el problema diferencial siguiente

$$\begin{aligned} \frac{d}{dx} \left(\frac{\partial}{\partial y'} \left(y \sqrt{1 + y'^2} - \lambda \left(y^2 - \frac{55}{\pi} \right) \right) \right) - \\ \frac{\partial}{\partial y} \left(y \sqrt{1 + y'^2} - \lambda \left(y^2 - \frac{55}{\pi} \right) \right) = 0. \end{aligned}$$

La solución de este problema ha sido efectuada utilizando el método de *shooting* múltiple. Como solución inicial se ha utilizado una parábola que satisfaga las condiciones de contorno y la condición de volumen. El intervalo $[0, 10]$ ha sido subdividido en 10 subintervalos de igual longitud. Después de 5 iteraciones se llega a una precisión de 10^{-10} . Las iteraciones

del método pueden apreciarse en la figura VII.1.5.

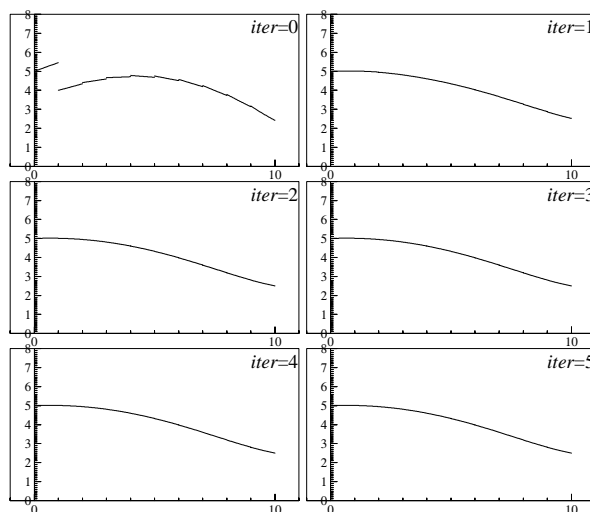


Figura VII.1.4. Determinación del Silo óptimo.

A continuación se presenta una serie de ejercicios, cuya finalidad es recordar las nociones básicas sobre las ecuaciones diferenciales.

Ejercicios

1.- Resolver:

$$y' = -x \operatorname{signo}(y) \sqrt{|x|},$$

$$y' = \exp(y) \sin(x),$$

$$y' = (x - y + 3)^2.$$

Dibujar los campos de vectores.

2.- Dibujar el campo de vectores de la ecuación

$$xy' = \sqrt{x^2 - y^2} + y.$$

Encontrar una fórmula explícita para las soluciones.

3.- La ecuación de un cuerpo en caída libre está dada por

$$r'' = -\frac{\gamma M}{r^2}, \quad r(0) = R, \quad r'(0) = v_0. \quad (\text{VII.1.39})$$

Plantear $r' = p(r)$, deducir una ecuación diferencial para p y resolverla. Encontrar una condición para v_0 , tal que $r(t) \rightarrow \infty$ si $t \rightarrow \infty$. Calcular las soluciones de (VII.1.39) que tienen la forma $a(b \pm x)^\alpha$.

4.- Transformar la ecuación de Bernoulli

$$y' + \frac{y}{1+x} + (1+x)y^4 = 0,$$

en una ecuación lineal. Plantear $y(x) = z(x)^q$ con q conveniente.

5.- Resolver la ecuación de Riccati

$$y' = y^2 + 1 - x^2. \quad (\text{VII.1.40})$$

Dibujar el campo de vectores considerando las curvas donde $y' = \text{cons}$, las isoclinas. Deducir una solución particular $\phi(x)$ de (VII.1.40). Calcular las otras soluciones mediante la transformación $z(x) = y(x) - \phi(x)$.

6.- Encontrar la solución de

$$y'' - 3y' - 4y = g(x), \quad g(x) = \begin{cases} \cos x, & 0 \leq x \leq \pi/2; \\ 0, & \pi/2 \leq x; \end{cases}$$

que satisface $y(0) = y'(0) = 0$.

7.- Resolver las ecuaciones lineales

$$y' = \begin{pmatrix} 3 & 6 \\ -2 & -3 \end{pmatrix} y, \quad y' = \begin{pmatrix} 1 & -1 \\ 4 & -3 \end{pmatrix}.$$

VII.2 Método de Euler

En esta sección será estudiado el método más sencillo de resolución de ecuaciones diferenciales con valores iniciales.

Se considera el problema a valor inicial o problema de Cauchy dado por

$$\begin{cases} y' = f(x, y), \\ y(t_0) = y_0, \end{cases} \quad (\text{VII.2.1})$$

donde $f : \mathcal{V} \rightarrow \mathbb{R}^n$ continua, con $\mathcal{V} \subset \mathbb{R} \times \mathbb{R}^n$. Se desea determinar $y(t_0 + T)$, para eso se considera la subdivisión del intervalo $[t_0, t_0 + T]$ dada por

$$t_0 < t_1 < t_2 < \cdots < t_n = t_0 + T,$$

definiendo

$$h_i = x_{i+1} - x_i, \quad i = 0, \dots, n-1. \quad (\text{VII.2.2})$$

La idea fundamental del método de Euler consiste en aproximar la solución exacta en x_1 , utilizando una tangente que pase por (x_0, y_0) , es decir

$$y_1 = y_0 + h_0 f(x_0, y_0). \quad (\text{VII.2.3})$$

De manera general

$$\begin{aligned} x_{i+1} &= x_i + h_i, \\ y_{i+1} &= y_i + h_i f(x_i, y_i). \end{aligned} \quad (\text{VII.2.4})$$

La solución numérica proporcionada por el método de Euler es, por consiguiente una función poligonal, ver figura VI.2.1, denotada por $y_h(x)$, llamada polígono de Euler; esta función esta definida por:

$$\begin{cases} h = (h_0, \dots, h_n), \\ x \in [x_k, x_{k+1}], \\ y_h(x) = y_k + (x - x_k) f(x_k, y_k). \end{cases} \quad (\text{VII.2.5})$$

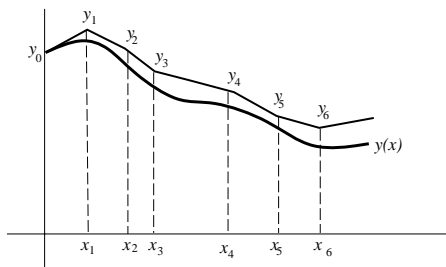


Figura VII.2.1. El Polígono de Euler

Una pregunta muy natural, que uno puede plantearse, consiste en determinar bajo que condiciones el método de Euler converge hacia la solución exacta del problema diferencial de Cauchy, cuando $n \rightarrow \infty$.

Considérese el problema

$$\begin{aligned} y' &= f(x, y), \\ y(x_0) &= y_0, \end{aligned} \tag{VII.2.6}$$

donde $f : \mathcal{U} \rightarrow \mathbb{R}^n$, con $\mathcal{U} \subset \mathbb{R} \times \mathbb{R}^n$ abierto. Se define

$$|h| = \max_{i=0, \dots, n-1} h_i. \tag{VII.2.7}$$

Proposición VII.2.1.- Sea $\mathcal{D} = \{(x, y) | x \in [x_0, x_0 + a], \|y - y_0\| \leq b\} \subset \mathcal{U}$. Supóngase que $f|_{\mathcal{D}}$ continua, $A = \max_{(x, y) \in \mathcal{D}} \|f(x, y)\|$, $\alpha = \min(a, b/A)$.

Entonces para cada división h del intervalo $[x_0, x_0 + \alpha]$ se tiene:

- a) $\|y_h(x) - y_h(\bar{x})\| \leq A \|x - \bar{x}\|$; $x, \bar{x} \in [x_0, x_0 + \alpha]$.
- b) si $\|f(x, y) - f(x_0, y_0)\| \leq \epsilon$ sobre \mathcal{D} , entonces $\|y_h(x) - (y_0 + (x - x_0)f(x_0, y_0))\| \leq \epsilon |x - x_0|$.

Demostración.-El punto a) es trivial, ya que es consecuencia de la definición de y_h , A y la desigualdad del triángulo aplicada una cantidad finita de veces.

La demostración del punto b) es la siguiente. Sea $x \in [x_0, x_0 + \alpha]$, por consiguiente $x \in [x_{k-1}, x_k]$, de donde

$$\begin{aligned} y_h(x) &= y_k + (x - x_{k-1})f(x_{k-1}, y_{k-1}) \\ &= y_0 + h_0 f(x_0, y_0) + h_1 f(x_1, y_1) + \dots + h_{k-2} f(x_{k-2}, y_{k-2}) \\ &\quad + (x - x_{k-1})f(x_{k-1}, y_{k-1}). \end{aligned}$$

Ahora bien,

$$\begin{aligned} y_0 + (x - x_0)f(x_0, y_0) &= y_0 + h_0 f(x_0, y_0) + \dots + h_{k-2} f(x_{k-2}, y_{k-2}) \\ &\quad + (x - x_{k-1})f(x_{k-1}, y_{k-1}), \end{aligned}$$

obteniendo finalmente b). □

Proposición VII.2.2.- Sea h una división del intervalo I . Sean $y_h(x)$ el polígono de Euler para (x_0, y_0) y $z_k(x)$ el polígono de Euler para (x_0, z_0) . Si

$$\|f(x, y) - f(x, z)\| \leq L \|y - z\|, \quad \forall (x, y), (x, z) \in \mathcal{D}; \tag{VII.2.8}$$

entonces

$$\|y_h(x) - z_h(x)\| \leq \|x_0 - z_0\| e^{L(x-x_0)}. \tag{VII.2.9}$$

Antes de demostrar esta proposición vale la pena recalcar que si f satisface una condición de Lipschitz, es decir (VII.2.8), el polígono de Euler es único paa h dado.

Demostración.- Se tiene:

$$\begin{aligned} y_1 &= y_0 + h_0 f(x_0, y_0), \\ z_1 &= z_0 + h_0 f(x_0, z_0), \end{aligned}$$

obteniendo como desigualdad

$$\begin{aligned} \|y_1 - z_1\| &\leq \|y_0 - z_0\| + h_0 L \|y_0 - z_0\| \\ &\leq (1 + h_0 L) \|y_0 - z_0\| \\ &\leq e^{h_0 L}, \end{aligned}$$

procediendo de manera recursiva, se obtiene

$$\begin{aligned} \|y_k - z_k\| &\leq e^{L h_{k-1}} \|y_{k-1} - z_{k-1}\| \\ &\leq e^{L(x-x_0)} \|y_0 - z_0\|. \end{aligned}$$

□

Teorema VII.2.3.- Cauchy. Sea $\mathcal{D} = \{(x, y) | x_0 \leq x \leq x_0 + a, \|y - y_0\|\} \subset \mathcal{U}$; supóngase:

i) $f|_{\mathcal{D}}$ continua, $A = \max_{(x,y) \in \mathcal{D}} \|f(x, y)\|$, $\alpha = \min(a, b/A)$,

ii) $\|f(x, y) - f(x, z)\| \leq L \|y - z\|$ si $(x, y), (x, z) \in \mathcal{D}$;

entonces:

a) Si $|h| \rightarrow 0$, entonces los polígonos de Euler $y_h(x)$ convergen uniformemente sobre $[x_0, x_0 + \alpha]$ hacia una función $\varphi(x)$.

b) $\varphi(x)$ es solución de $y' = f(x, y)$, $y(x_0) = y_0$.

c) La solución es única sobre $[x_0, x_0 + \alpha]$.

Demostración.- Inicialmente se mostrará, que si h_k es una sucesión de subdivisiones de $[x_0, x_0 + \alpha]$ con $|h_k| \rightarrow 0$, entonces la sucesión de poligonos de Euler asociada, es una sucesión de Cauchy para la norma de la convergencia uniforme. Para tal efecto, se mostrará que

$$\begin{aligned} \forall \epsilon > 0 \exists \delta > 0 \text{ tal que } |h| < \delta, |\hat{h}| < \delta \\ \implies \forall x \in [x_0, x_0 + \alpha] \text{ se tiene } \|y_h(x) - y_{\hat{h}}(x)\| < \epsilon. \end{aligned}$$

donde h es una división con $|h| < \delta$ y \hat{h} una división más fina que h .

Sea $\epsilon > 0$ dado, entonces existe $\delta > 0$ tal que si $|x - \bar{x}| \leq \delta$ y $\|y - \bar{y}\| \leq A\delta$ implica que $\|f(x, y) - f(\bar{x}, \bar{y})\| \leq \epsilon$. Esto es cierto por que f es uniformemente continua sobre \mathcal{D} .

Por la proposición VII.2.1, se tiene

$$\|y_h(x) - \{y_0 + (x - x_0)f(x_0, y_0)\}\| \leq \epsilon |x - x_0|,$$

de donde

$$\begin{aligned} \|y_h(x) - y_{\hat{h}}(x)\| &\leq \epsilon \left[(x_1 - x_0)e^{L(x-x_1)} + \dots + (x - x_k)e^{L(x-x_k)} \right] \\ &\leq \int_{x_0}^x e^{L(x-s)} ds = \epsilon \frac{e^{L(x-x_0)} - 1}{L} \\ &\leq \epsilon \frac{e^{L\alpha} - 1}{L}, \end{aligned}$$

por consiguiente, $\{y_h(x)\}$ es una sucesión de Cauchy, con δ que no depende de x . Como consecuencia inmediata, se tiene que la sucesión converge hacia una función $\varphi(x)$ que además es continua.

La demostración del punto b) se basa en los siguientes hechos: $y_h(x_0) = y_0$ implica que $\varphi(x_0) = y_0$, se considera el módulo de continuidad de f , definido por

$$\epsilon(\delta) = \sup\{\|f(x, y) - f(\bar{x}, \bar{y})\| \mid |x - \bar{x}| \leq \delta, \|y - \bar{y}\| \leq A\delta\},$$

se observa inmediatamente, que $\epsilon(\delta) \rightarrow 0$ si $\delta \rightarrow 0$. Utilizando la proposición VII.2.2, se obtiene

$$\|y_h(x + \delta) - y_h(x) - \delta f(x, y_h(x))\| \leq \delta \epsilon(\delta),$$

de donde, se tiene

$$\|\varphi(x + \delta) - \varphi(x) - \delta f(x, \varphi(x))\| \leq \delta \epsilon(\delta),$$

efectuando el pasaje al límite, se obtiene

$$\varphi'(x) = f(x, \varphi(x)).$$

La unicidad del punto c) ha sido ya demostrada en el corolario VII.1.12 \square

Corolario VII.2.4.- $f : \mathcal{U} \rightarrow \mathbb{R}^n$ ($\mathcal{U} \subset \times \mathbb{R}^n$) continuamente diferenciable.

$$\mathcal{D} = \{(x, y) \mid x_0 \leq x \leq x_0 + \alpha, \|y - y_0\| \leq b\} \subset \mathcal{U}.$$

Sobre \mathcal{D} se tiene $\|f(x, y)\| \leq A$, $\left\|\frac{\partial f}{\partial y}(x, y)\right\| \leq L$, $\left\|\frac{\partial f}{\partial x}(x, y)\right\| \leq M$.

Entonces

$$\|y(x) - y_h(x)\| \leq \frac{M + AL}{L} \left(e^{L(x-x_0)} - 1 \right) |h|, \quad (\text{VII.2.10})$$

donde $y(x)$ es solución de $y' = f(x, y)$, $y(x_0) = y_0$.

Demostración.- Remontando la demostración del teorema precedente, se tiene

$$\|y_h(x) - y(x)\| \leq \epsilon \frac{e^{L(x-x_0)} - 1}{L}.$$

Puesto que f es diferenciable, se tiene

$$\begin{aligned} \|f(x, y) - f(\bar{x}, \bar{y})\| &\leq L \|y - \bar{y}\| + M |x - \bar{x}| \\ &\leq A |h| + |h|, \end{aligned}$$

de donde planteando $\epsilon = (LA + M) |h|$ se tiene (VII.2.10). \square

Efectos de los Errores de Redondeo

Generalmente no se puede calcular la solución exacta del esquema (VII.2.4); se calcula solamente la solución del esquema perturbado dado por

$$y_{n+1}^* = y_n^* + h_n f(t_n, y_n^*) + h_n \mu_n + \varrho_n \quad (\text{VII.2.11})$$

donde μ_n designa el error con el cual es calculado la función f y ϱ_n los errores de redondeo cometidos por la computadora. Se supondrá que $|\mu_n| \leq \mu$ y $|\varrho_n| \leq \varrho$.

Con la hipótesis suplementaria que f es continuamente diferenciable, como en el corolario VII.2.4, se tiene planteando $e_n = y_n^* - y_n$ y sustrayendo (VII.2.11) con (VII.2.4),

$$e_{n+1}^* = e_n^* + h_n [f(t_n, y_n^*) - f(t_n, y_n)] + h_n \mu_n + \varrho_n. \quad (\text{VII.2.12})$$

Puesto que f es diferenciable, se obtiene de (VII.2.12) el siguiente esquema

$$e_{n+1}^* = e_n^* + h_n \left[\frac{\partial f}{\partial y}(t_n, y_n) e_n^* \right] + h_n \mu_n + \varrho_n + \mathcal{O}(\|e_n^*\|^2). \quad (\text{VII.2.13})$$

Planteando $z_n = h_n e_n^*$, despreciando $\mathcal{O}(\|e_n^*\|^2)$ y suponiendo que $h_n = h$ constante, se obtiene el siguiente esquema para z_n , con

$$z_{n+1} = z_n + h \left[\frac{\partial f}{\partial y}(t_n, y_n) z_n \right] + h(h \mu_n + \varrho_n), \quad (\text{VII.2.14})$$

de donde z_n es la solución numérica exacta obtenida a partir del método de Euler de la ecuación

$$z'(t) = \left[\frac{\partial f}{\partial y}(t, y(t)) \right] z(t) + (h\mu(t) + \varrho(t)). \quad (\text{VII.2.15})$$

En lugar de estudiar la solución numérica de (VII.2.15), se estudiará la solución de este problema cuando h tiende a 0. Cualquier solución de la ecuación diferencial (VII.2.15) no puede ser idénticamente nula por la existencia del término no homogéneo no nulo, para h suficientemente pequeño este término no homogéneo es no nulo. Sea $C = \max \|z(t)\|$ cuando $h = 0$, por otro lado denotando $z_h(t)$ la solución de (VII.2.15) para un h fijo, se tiene que $z_h(t)$ converge uniformemente hacia $z_0(t)$ cuando h tiende a 0. Por lo tanto, existe un intervalo cerrado $J \subset [t_0, t_0 + T]$ y h_0 para los cuales

$$\|z_h(t)\| \geq \frac{C}{2} \quad \forall t \in J, \forall h \leq h_0.$$

Puesto que $e_n^* \approx z(t_n)/h$, se tiene que

$$\lim_{h \rightarrow 0} e_n^* = \infty.$$

Acaba de observarse que cuando la longitud de paso h_n tiende a 0, el error debido al redondeo toma un lugar preponderante, distorsionando completamente cualquier resultado numérico. En la figura VII.2.2 puede verse un comportamiento aproximado del error de redondeo en función de h .

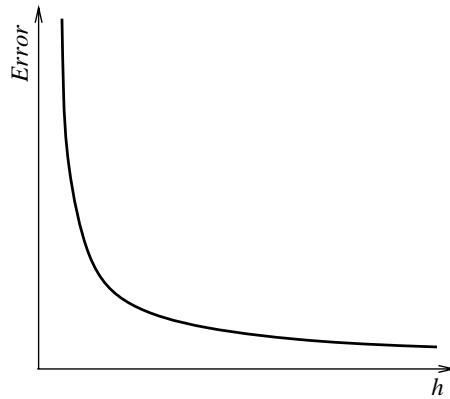


Figura VII.2.2. Error de Redondeo en el Método de Euler.

La pregunta natural que surge es: ¿Cuál es el h mínimo que se puede tomar sin que el error de redondeo sea preponderante? La respuesta a esta

pregunta tiene dos fuentes, la primera que radica en la práctica numérica, y la segunda en un análisis sobre el origen de los otros errores que ocurren durante la implementación del método. Ambos estudios llegan a la conclusión de

$$h_{\min} \geq C\sqrt{\textit{eps}}, \quad (\text{VII.2.16})$$

donde \textit{eps} es la precisión de la computadora.

Las mayoraciones que se acaban de dar son por lo general demasiado pesimistas, por ser rigurosos matemáticamente, uno se situa en la situación más desfavorable posible. Por otro lado si se toma h muy pequeño, inferior a \textit{eps} , el algoritmo se mantiene estacionario.

Estabilidad del Método de Euler

En la anterior subsección se pudo observar la incidencia directa del error de redondeo. En esta parte se analizará la propagación del error de redondeo en la solución numérica. Considérese el problema siguiente

$$y' = \lambda y, \quad y(0) = y_0. \quad (\text{VII.2.17})$$

El método de Euler con paso constante, da el siguiente esquema

$$y_{n+1} = y_n + h\lambda y_n, \quad (\text{VII.2.18})$$

supóngase que en lugar de y_0 , se introduce una aproximación \tilde{y}_0 , planteando $e_n = \tilde{y}_n - y_n$ donde \tilde{y}_n es la solución numérica de la ecuación (VII.2.17) con valor inicial \tilde{y}_0 . Los e_n verifican la siguiente relación:

$$e_{n+1} = e_n + h\lambda e_n, \quad e_0 = (\tilde{y}_0 - y_0),$$

de donde

$$e_n = (\lambda h + 1)^n e_0. \quad (\text{VII.2.19})$$

Por consiguiente, el esquema (VII.2.17) será estable siempre y cuando

$$\lim_{n \rightarrow \infty} e_n = 0,$$

situación que sucede cuando $|\lambda h + 1| < 1$. Por consiguiente, para obtener un método estable es necesario que

$$h_{\max} = \frac{2}{|\lambda|}. \quad (\text{VII.2.20})$$

Por lo expuesto en la anterior subsección y en ésta, se tiene necesariamente que la longitud de paso está acotada inferiormente e superiormente. La cota inferior impide que el error de redondeo distorsione completamente

la solución numérica del problema, mientras que la cota superior en los problemas lineales impide que el método diverga. La idea de estabilidad puede generalizarse a problemas diferenciales lineales de mayor dimensión. Por ejemplo, considérese el problema

$$y' = Ay, \quad y(0) = y_0.$$

Con el procedimiento utilizado anteriormente e introduciendo normas en los lugares apropiados es fácil mostrar que el método de Euler es estable si

$$\rho(A + I)h < 1, \quad (\text{VII.2.21})$$

donde $\rho(A + I)$ es el radio espectral de la matriz $A + I$. Planteando $z = \lambda h$, se tiene estabilidad en el método, si y solamente si $|z + 1| < 1$. De donde, se tiene definida una región de estabilidad, dada en la figura VII.2.3, la parte achurada del círculo corresponde a la región donde el método es estable.

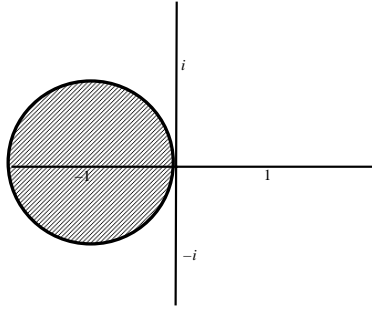


Figura VII.2.3. Región de Estabilidad del Método de Euler

El siguiente ejemplo ilustra, que el método de Euler no es un método muy apropiado para resolver ciertos problemas diferenciales a valor inicial. Considérese, el problema

$$\begin{aligned} y'(x) &= -100y(x) + \cos x, \\ y(0) &= 0. \end{aligned} \quad (\text{VII.2.22})$$

La solución de este problema, está dada por

$$y(x) = \frac{100}{10001} \cos x + \frac{1}{10001} \sin x + Ce^{-100x},$$

con C una constante determinada por la condición inicial. Puede observarse que para $x = \pi$, se tiene

$$y(\pi) \approx -\frac{100}{10001} \approx -0,01. \quad (\text{VII.2.23})$$

Aplicando el método de Euler con paso constante se tiene:

$$\begin{aligned} h_1 &= \frac{\pi}{165} \rightarrow y_h(\pi) = -9.999 \times 10^{-3}, \\ h_2 &= \frac{\pi}{155} \rightarrow y_h(\pi) = -60.101. \end{aligned}$$

No obstante que $h_2 - h_1 \approx 1.2 \times 10^{-3}$, la diferencia de los resultados es abismal. La explicación de este fenómeno de inestabilidad ha sido explicado más arriba.

El problema dado por (VII.2.22) está dentro una categoría de problemas diferenciales conocidos con el nombre de ecuaciones diferenciales rígidas, ver Hairer & Wanner. Para evitar aberraciones en los resultados numéricos en la resolución numérica de esta clase de problemas, el método de Euler puede modificarse, obteniendo así el:

Método de Euler Implícito

En lugar de utilizar el esquema dado por (VII.2.4); el método de Euler implícito, está dado por el esquema

$$y_{k+1} = y_k + h_k f(x_{k+1}, y_{k+1}). \quad (\text{VII.2.24})$$

Puede observarse inmediatamente, que para determinar y_{k+1} se debe resolver una ecuación, que en lo general no es lineal. Comparando con la versión explícita del método de Euler, esto constituye una dificultad adicional, pues para evaluar y_{k+1} debe utilizarse un método de resolución de ecuaciones, como ser el método de Newton. Puede mostrarse, ver Hairer & Wanner, que si h_k es lo suficientemente pequeño, la convergencia del método de Newton, o de otro método de resolución está asegurada. Se ha expuesto la desventaja principal del método de Euler implícito, respecto al método de Euler explícito. A continuación se expondrá la principal motivación de utilizar la versión implícita en determinadas situaciones. La principal desventaja del método de Euler explícito en radica en la falta de estabilidad cuando h no es lo suficientemente pequeño, ver el ejemplo en el cual para 2 pasos muy próximos, las soluciones numéricas del problema (VII.2.22) difieren de manera significativa. El análisis de estabilidad del método de Euler implícito es muy similar a la del método de Euler explícito, en efecto considerando el problema (VII.2.18), se tiene que el método implícito satisface la siguiente relación recursiva para el error

$$e_{n+1} = e_n + \lambda h e_{n+1}, \quad (\text{VII.2.25})$$

de donde, se obtiene de manera explícita

$$e_{n+1} = \frac{1}{1 - \lambda h} e_n, \quad (\text{VII.2.26})$$

teniendo de esta manera estabilidad en la propagación de los errores de redondeo, si y solamente si

$$|1 - \lambda h| > 1. \quad (\text{VII.2.27})$$

Para los problemas de la forma $y' = Ay$, (VII.2.27) y remplazando $z = \lambda h$, se obtiene el dominio de estabilidad dado por

$$|z - 1| > 1, \quad (\text{VII.2.28})$$

el cual está dado en la figura VII.2.3.

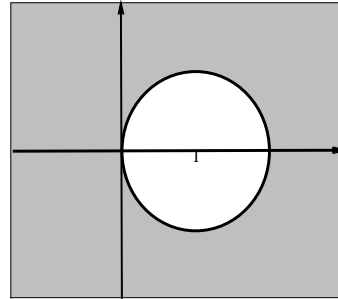


Figura VII.2.3. Región de Estabilidad del Método de Euler Implícito.

Ejercicios

- 1.- Aplicar el método de Euler al problema

$$y' = y^2, \quad y(0) = 1.$$

Evaluar $y(1/4)$.

Utilizar pasos constantes, (por ejemplo $h=1/6$). Estimar el error con el corolario VII.2.4. Comparar esta estimación con el error exacto.

- 2.- Demostrar el resultado siguiente: Si el problema

$$y' = f(x, y), \quad y(x_0) = y_0,$$

con $f : \mathcal{U} \rightarrow \mathbb{R}^n$ continua, $\mathcal{U} \subset \times \mathbb{R}^n$ abierto y $(x_0, y_0) \in \mathcal{U}$ posee una y una sola solución sobre el intervalo I , entonces los polígonos de Euler $y_h(x)$ convergen uniformemente sobre I hacia esta solución.

- 3.- Aplicar el método de Euler al sistema

$$\begin{aligned} y_1' &= -y_2, & y_1(0) &= 1, \\ y_2' &= y_1, & y_2(0) &= 0. \end{aligned}$$

Utilizar pasos constantes para encontrar una aproximación de la solución en el punto $x = 0.4$. Estimar el error como en el ejercicio 1 y compararla con el error exacto. La solución exacta es $y_1(x) = \cos x$, $y_2(x) = -\sin x$.

VII.3 Métodos de Runge-Kutta

En la anterior sección se formuló el método de Euler en su versión explícita como en su versión implícita. Uno de los grandes problemas con el cual se debe confrontar en la utilización de este método, consiste en la falta de precisión de éste, ver el corolario VII.2.4, motivo por el cual, los pasos de integración deben ser demasiado pequeños, induciendo de esta manera grandes perturbaciones debido al error de redondeo.

En esta sección se pretende construir métodos cuya precisión sea más elevada, permitiendo así menos evaluaciones y una menor incidencia del error de redondeo en los cálculos. Se estudiará los métodos conocidos como métodos a un paso, en contraposición a los métodos multipasos que serán tratados en la siguiente sección.

Se pretende calcular una aproximación de la solución de

$$y' = f(x, y), \quad y(x_0) = y_0, \quad (\text{VII.3.1})$$

sobre el intervalo $[x_0, x_e]$. Se procede como sigue: De la misma manera que en el método de Euler, se particiona $[x_0, x_e]$ en subintervalos $x_0 < x_1 < \dots < x_N = x_e$, se denota $h_n = x_{n+1} - x_n$ y se calcula $y_n \approx y(x_n)$ por una fórmula de la forma

$$y_{n+1} = y_n + h_n \Phi(h_n, x_n, y_n). \quad (\text{VII.3.2})$$

Una tal fórmula se llama método a un paso, por que el cálculo de y_{n+1} utiliza los valores de h_n, x_n, y_n de un paso solamente.

Puede observarse fácilmente que el método de Euler corresponde bien a esta categoría de métodos a un paso. Vale la pena recalcar que los métodos de la forma (VII.3.2) no solamente son métodos a un paso, si no que también explícitos. Para saber más referirse a Hairer & Wanner. Para simplificar la notación, solamente se considerará el primer paso, es decir cuando $n = 0$ en (VII.3.2) y escribir h en lugar de h_0 .

Para obtener otros métodos numéricos se integra (VII.3.1) de x_0 a $x_0 + h$. La expresión que se obtiene, está dada por

$$y(x_0 + h) = y_0 + \int_{x_0}^{x_0+h} f(t, y(t)) dt. \quad (\text{VII.3.3})$$

La idea central consiste en remplazar la integral de (VII.3.3) por una expresión que la aproxime. Por ejemplo, si se utiliza $h(f(x_0, y_0))$ como aproximación se tiene el método de Euler Explícito. Por consiguiente, la idea será utilizar fórmulas de cuadratura que tienen un orden más elevado.

Como ejemplo de la utilización de este ingenioso argumento, se tiene el método de Runge. Se toma la fórmula del punto medio que es una fórmula de cuadratura de orden 2. Obteniendo así

$$y(x_0 + h) \approx y_0 + hf \left(x_0 + \frac{h}{2}, y(x_0 + \frac{h}{2}) \right), \quad (\text{VII.3.4})$$

obteniendo posteriormente el valor desconocido $y(x_0 + h/2)$ mediante el método de Euler. Esto da

$$y_1 = y_0 + hf \left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2} f(x_0, y_0) \right). \quad (\text{VII.3.5})$$

La interpretación geométrica del método de Runge, ver figura VII.3.1, consiste en hacer pasar una recta tangente por el punto medio de las curvas integrales de la ecuación diferencial, obteniendo este punto medio mediante una tangente que sale de (x_0, y_0) . Ver la figura VII.3.1.

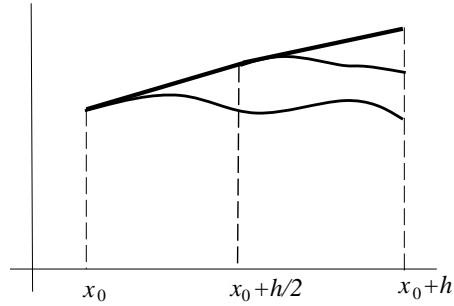


Figura VII.3.1 Método de Runge

Kutta en 1901, generalizó la idea de Runge, conduciendo a la definición siguiente de los métodos que llevan sus nombres.

Definición VII.3.1.- Un método de Runge-Kutta a s pisos, está dada por:

$$\begin{aligned} k_1 &= f(x_0, y_0), \\ k_2 &= f(x_0 + c_2 h, y_0 + h a_{21} k_1), \\ k_3 &= f(x_0 + c_3 h, y_0 + h(a_{31} k_1 + a_{32} k_2)), \\ &\vdots \\ k_s &= f(x_0 + c_s h, y_0 + h(a_{s1} k_1 + \cdots + a_{s,s-1} k_{s-1})); \\ y_1 &= y_0 + h(b_1 k_1 + \cdots + b_s k_s); \end{aligned} \quad (\text{VII.3.6})$$

donde los c_i, a_{ij}, b_j son coeficientes. El método se representa por el esquema

$$\begin{array}{c|ccccc}
c_1 = 0 & & & & & \\
c_2 & a_{21} & & & & \\
c_3 & a_{31} & a_{32} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & \\
\hline
& b_1 & b_2 & \cdots & b_{s-1} & b_s
\end{array} \tag{VII.3.7}$$

El método de Runge está basado sobre la fórmula del punto medio que es una fórmula de cuadratura de orden 2:

$$y(x_0 + h) = y_0 + hf(x_0 + h/2, y(x_0 + h/2)) + \mathcal{O}(h^3).$$

Remplazando $y(x_0 + h/2)$ por el valor $y_0 + (h/2)f(x_0, y_0)$ del método de Euler, se agrega un término de tamaño $\mathcal{O}(h^3)$. Entonces, este método tiene orden $p = 2$.

El método de Heun, ver tabla VII.3.1, se obtiene a partir de la fórmula de cuadratura

$$y(x_0 + h) = y_0 + \frac{h}{4} \left(f(x_0, y_0) + 3f\left(x_0 + \frac{2h}{3}, y(x_0 + \frac{2h}{3})\right) \right) + \mathcal{O}(h^4),$$

si se remplace $y(x_0 + 2h/3)$ por la aproximación del método de Runge. De donde el método de Heun tiene el orden $p = 3$.

Utilizando la suposición (VII.3.8) que es una condición simplificadora, se tiene la siguiente proposición.

Proposición VII.3.3.- *La condición (VII.3.8) implica que es suficiente considerar problemas autónomos de la forma $y' = f(y)$ para verificar la condición de orden.*

Demostración.- Se supone que (VII.3.9) es satisfecha para los problemas autónomos de la forma $y' = F(y)$, $y(x_0) = y_0$.

Considérese un problema de la forma $y' = f(x, y)$, $y(x_0) = y_0$, el cual es equivalente al problema

$$\begin{aligned} \dot{y} &= f(x, y), & y(0) &= y_0, \\ \dot{x} &= 1, & x(0) &= 0, \end{aligned}$$

obteniendo así

$$Y' = F(Y), \quad \text{donde} \quad Y = \begin{pmatrix} x \\ y \end{pmatrix}, \quad F(Y) = \begin{pmatrix} 1 \\ f(x, y) \end{pmatrix} \quad (\text{VII.3.10})$$

con valor inicial $Y_0 = (x_0, y_0)^t$. La aplicación del método de Runge-Kutta al problema (VII.3.10) da

$$K_i = F\left(Y_0 + h \sum_{j=1}^{i-1} a_{ij} K_j\right) = \begin{pmatrix} 1 \\ k_i \end{pmatrix}, \quad Y_1 = Y_0 + h \sum_{i=1}^s b_i K_i = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix},$$

lo que es equivalente a (VII.3.6), por que

$$Y_0 + h \sum_{j=1}^{i-1} a_{ij} K_j = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + h \sum_{j=1}^{i-1} a_{ij} \begin{pmatrix} 1 \\ k_j \end{pmatrix} = \begin{pmatrix} x_0 + c_i h \\ y_0 + h \sum_{j=1}^{i-1} a_{ij} k_j \end{pmatrix}.$$

□

El siguiente paso, es de estudiar un procedimiento que permita determinar métodos de Runge-Kutta de ordenes más elevados. Un buen ejercicio será construir el método de Kutta que es de orden 4.

Construcción de un método de orden 4

La idea principal de la construcción del método de Runge, es considerar los desarrollos en serie de Taylor de la solución exacta $y(x_0 + h)$, como también de la solución numérica obtenida a partir del método, que se la denota por $y_1(h)$. Luego comparar las series, para obtener condiciones sobre los coeficientes del método.

Serie de Taylor de la solución exacta

Considérese el problema de Cauchy

$$y' = f(y), \quad y(x_0) = y_0. \quad (\text{VII.3.11})$$

derivando la ecuación (VII.3.11), se obtiene para la solución exacta

$$\begin{aligned} y'' &= f'_y y' = f'_y f(y) \\ y''' &= f''_y(y', f(y)) + f'_y f'_y y' = f''_y(f(y), f(y)) + f'_y f'_y f(y) \\ y^{(4)} &= f'''_y(f(y), f(y), f(y)) + 3f''_y(f'_y f(y), f(y)) + f'_y f''_y(f(y), f(y)) \\ &\quad + f'_y f'_y f'_y f(y), \end{aligned}$$

de donde la serie de Taylor de la solución exacta está dada por

$$\begin{aligned} y(x_0 + h) &= y_0 + hf(y_0) + \frac{h^2}{2!} f'_y f(y_0) \\ &\quad + \frac{h^3}{3!} (f''_y(f(y_0), f(y_0)) + f'_y f'_y f(y_0)) \\ &\quad + \left(f'''_y(f(y), f(y), f(y)) + 3f''_y(f'_y f(y), f(y)) \right. \\ &\quad \left. + f'_y f''_y(f(y), f(y)) + f'_y f'_y f'_y f(y) \right) + \mathcal{O}(h^5). \end{aligned} \quad (\text{VII.3.12})$$

Serie de Taylor de la solución numérica

Para calcular la serie de Taylor de y_1 , es suficiente determinar las series de los

$$k_i = f(x_0 + h \sum_{j=1}^{i-1} a_{ij} k_j). \quad (\text{VII.3.13})$$

Ahora bien, se puede considerar (VII.3.13) como una ecuación a punto fijo, pudiendo así aproximar los k_i por el método de las aproximaciones sucesivas. Comenzando la primera aproximación de k_i , se tiene

$$k_i = f(y_0) + \mathcal{O}(h),$$

utilizando (VII.3.8), se obtiene

$$k_i = f(y_0) + c_i h f_y f(y_0) + \mathcal{O}(h^2).$$

La segunda iteración da como resultado

$$\begin{aligned} k_i &= f(y_0 + h c_i f(y_0)) + h^2 \sum_j a_{ij} c_j f_y f(y_0) + \mathcal{O}(h^3) \\ &= f(y_0) + c_i h f_y f(y_0) + h^2 \sum_j a_{ij} c_j f_y f_y f(y_0) + \frac{h^2}{2} c_i^2 f_y''(f(y_0), f(y_0)) \\ &\quad + \mathcal{O}(h^3). \end{aligned}$$

Efectuando todavía una vez más una iteración, e introduciendo los valores obtenidos en la definición (VII.3.5) de y_i , se obtiene

$$\begin{aligned} y_1 &= y_0 + h \left(\sum_i b_i \right) f_0 + \frac{h^2}{2} \left(2 \sum_i b_i c_i \right) (f' f)_0 \\ &\quad + \frac{h^3}{3!} \left(\left(3 \sum_i b_i c_i^2 \right) (f''(f, f))_0 + \left(6 \sum_{ij} b_i a_{ij} c_j \right) (f' f' f)_0 \right) \\ &\quad + \frac{h^4}{4!} \left(\left(\sum_i b_i c_i^3 \right) (f'''(f, f, f))_0 + \left(8 \sum_{ij} b_i c_i a_{ij} c_j \right) (f''(f' f' f))_0 \right. \\ &\quad \left. + \left(24 \sum_{ijk} b_i a_{ij} a_{jk} c_k \right) (f' f' f' f)_0 \right) + \mathcal{O}(h^5), \end{aligned} \tag{VII.3.14}$$

donde el subíndice 0 indica que las evaluaciones deben ser hechas en el punto y_0 .

Comparando los coeficientes de (VII.3.12) y (VII.3.14), se encuentran las condiciones que deben satisfacer los coeficientes c_i , b_i y a_{ij} para contar con un método de orden 4. Estas condiciones se las enuncia en el:

Teorema VII.3.4.- Condiciones de Orden. El método de Runge-Kutta (VII.3.6) tiene el orden 4 si los coeficientes satisfacen (VII.3.8) y

$$\sum_i b_i = 1, \tag{VII.3.15a}$$

$$\sum_i b_i c_i = \frac{1}{2}, \tag{VII.3.15b}$$

$$\sum_i b_i c_i^2 = \frac{1}{3}, \tag{VII.3.15c}$$

$$\sum_{ij} b_i a_{ij} = \frac{1}{6}, \quad (\text{VII.3.15d})$$

$$\sum_i b_i c_i^3 = \frac{1}{4}, \quad (\text{VII.3.15e})$$

$$\sum_{ij} b_i c_i a_{ij} c_j = \frac{1}{8}, \quad (\text{VII.3.15f})$$

$$\sum_{ij} b_i a_{ij} c_j^2 = \frac{1}{12}, \quad (\text{VII.3.15g})$$

$$\sum_{ijk} b_i a_{ij} a_{jk} b_k = \frac{1}{24}. \quad (\text{VII.3.15h})$$

Es necesario remarcar que si el método satisface solamente (VII.3.15a), (VII.3.15a,b) o (VII.3.15,a,b,c) tiene el orden 1, 2 o 3 respectivamente.

El siguiente paso es la resolución del sistema (VII.3.15), para $s = 4$. Este sistema consiste de 8 ecuaciones no lineales para 10 parámetros b_i , a_{ij} ; los c_i están determinados por la condición (VII.3.8). Las condiciones (VII.3.15a,b,c,e) traducen el hecho que (b_i, c_i) es una fórmula de cuadratura de orden 4. Por consiguiente los b_i pueden ser determinados si los c_i son conocidos. En particular, se obtiene

$$b_3 c_3 (c_3 - c_2) (c_4 - c_3) = -\frac{c_2 c_4}{2} + \frac{c_2 + c_3}{3} - \frac{1}{4}. \quad (\text{VII.3.16})$$

Por otro lado (VII.3.15g)–(VII.3.15d) da

$$b_4 a_{43} c_3 (c_3 - c_2) = \frac{1}{12} - \frac{c_2}{6}, \quad (\text{VII.3.17})$$

y $c_4 \cdot (\text{VII.3.15d}) - (\text{VII.3.15f})$ implica que

$$b_3 (c_4 - c_3) a_{32} c_2 = \frac{c_4}{6} - \frac{1}{8}. \quad (\text{VII.3.18})$$

Si se multiplica (VII.3.17) con (VII.3.18), luego (VII.3.16) con (VII.3.15h), se obtiene para la misma expresión, las dos fórmulas:

$$b_4 a_{43} a_{32} c_2 \cdot b_3 c_3 (c_3 - c_2) (c_4 - c_3) = \left(\frac{1}{12} - \frac{c_2}{6} \right) \left(\frac{c_4}{6} - \frac{1}{8} \right),$$

$$b_4 a_{43} a_{32} c_2 \cdot b_3 c_3 (c_3 - c_2) (c_4 - c_3) = \left(-\frac{c_2 c_4}{2} + \frac{c_2 + c_3}{3} - \frac{1}{4} \right) / 24,$$

lo que es equivalente a

$$c_2 (1 - c_4) = 0.$$

Puesto que $c_2 \neq 0$, consecuencia de la condición (VII.3.15h), se tiene necesariamente que $c_4 = 1$. La solución general del sistema (VII.3.16), se la obtiene de la manera siguiente.

Método de Resolución.- Plantear $c_1 = 0$, $c_4 = 1$; c_2 y c_3 son parámetros libres; calcular b_1, b_2, b_3, b_4 tales que la fórmula de cuadratura sea de orden 4; calcular a_{32} utilizando (VII.3.18), a_{43} utilizando (VII.3.19) y a_{42} proviene de (VII.3.15d); finalmente calcular a_{21}, a_{31}, a_{41} de (VII.3.8) para $i = 2, 3, 4$.

Entre los métodos de orden 4, los más conocidos están dados en la tabla VII.3.2.

Tabla VII.3.2. Métodos de Kutta

0					0				
1/2	1/2				1/3	1/3			
1/2	0	1/2			2/3	-1/3	1		
1	0	0	1		1	1	-1	1	
	1/6	2/6	2/6	1/6		1/8	3/8	3/8	1/8
Método de Runge-Kutta					Regla 3/8				

Métodos Encajonados

Para resolver un problema realista, un cálculo a paso constante es en general ineficiente. Existen diversas causas para esta ineficiencia: la primera es que se debe conocer a priori una estimación del error local cometido, lo cual es complicado en general ocasionando una pérdida de generalidad en los programas a implementarse. La otra razón fundamental que existen intervalos donde los pasos de integración pueden ser de longitud mayor. El principal problema reside en como escoger los pasos de integración. La idea que puede resolver satisfactoriamente estos problemas consiste en escoger pasos de integración de manera que el error local sea en todo caso inferior o igual a Tol dado por el utilizador. Motivo por el cual es necesario conocer una estimación del error local. Inspirados en el programa GAUINT descrito en VI.3, se construye un método de Runge-Kutta con \hat{y}_1 como aproximación numérica, y se utiliza la diferencia $\hat{y}_1 - y_1$ como estimación del error local del método menos preciso.

Se da un método de orden p a s pisos, con coeficientes c_i, a_{ij}, b_j . Se busca un método de orden $\hat{p} < p$ que utiliza las mismas evaluaciones de f , es decir

$$\hat{y}_1 = y_0 + h(\hat{b}_1 k_1 + \cdots + \hat{b}_s k_s), \quad (\text{VII.3.19})$$

donde los k_i están dados por (VII.3.3). Para tener más libertad, se agrega a menudo un término que contenga $f(x_1, y_1)$ a la fórmula (VII.3.19), de todas maneras se debe calcular $f(x_1, y_1)$ para el paso siguiente, determinando así \hat{y}_1 , como

$$\hat{y}_1 = y_0 + h(\hat{b}_1 k_1 + \cdots + \hat{b}_s k_s + \hat{b}_{s+1} f(x_1, y_1)). \quad (\text{VII.3.20})$$

Un tal método encajonado puede ser representado en forma de un esquema, ver la tabla VII.3.3.

Tabla VII.3.3. Métodos encajonados

$c_1 = 0$					
c_2	a_{21}				
c_3	a_{31}	a_{32}			
\vdots	\vdots	\vdots	\ddots		
c_s	a_{s1}	a_{s2}	\cdots	$a_{s,s-1}$	
(1)	b_1	b_2	\cdots	b_{s-1}	b_s
	\hat{b}_1	\hat{b}_2	\cdots	\hat{b}_{s-1}	$\hat{b}_s \quad (\hat{b}_{s+1})$

El siguiente paso es la determinación del h optimal. Si se aplica el método con un valor h , la estimación del error satisface

$$y_1 - \hat{y}_1 = (y_1 - y(x_0 + h)) + (y(x_0 + h) - \hat{y}_1) = \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{\hat{p}+1}) \approx Ch^{\hat{p}+1}. \quad (\text{VII.3.21})$$

El h optimal, denotado por h_{opt} , es áquel donde esta estimación es próxima de Tol , es decir

$$Tol \approx Ch_{\text{opt}}^{\hat{p}+1}. \quad (\text{VII.3.22})$$

Eliminando C de las dos últimas fórmulas, se obtiene

$$h_{\text{opt}} = 0.9 \cdot h \cdot \sqrt[p+1]{\frac{Tol}{\|y_1 - \hat{y}_1\|}}. \quad (\text{VII.3.23})$$

El factor 0.9 se lo agrega para volver el programa más seguro.

Algoritmo para la selección automática de paso.

Al iniciar los cálculos, el utilizador provee una subrutina que calcule el valor de $f(x, y)$, los valores iniciales x_0, y_0 y una primera elección de h .

A) Con h , calcular y_1 , $err = \|y_1 - \hat{y}_1\|$ y h_{opt} dado por (VII.3.23).

B) Si $err \leq Tol$, el paso es aceptado, entonces

$$x_0 := x_0 + h, \quad y_0 := y_1, \quad h := \min(h_{\text{opt}}, x_{\text{end}} - x_0);$$

si no el paso es rechazado

$$h := h_{\text{opt}}.$$

C) Si $x_0 = x_{\text{end}}$ se ha terminado, si no se recommienza por (A) y se calcula el paso siguiente.

La práctica numérica muestra que es aconsejable remplazar (VII.3.23) por

$$h_{\text{opt}} = h \cdot \min \left(5, \max(0.2, 0.9(Tol/err)^{1/\hat{p}+1}) \right).$$

Para la norma dada en la relación (VII.3.23), se utiliza en general

$$\|y_1 - \hat{y}_1\| = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{y_{i1} - \hat{y}_{i1}}{sc_i} \right)^2}, \quad \text{donde} \quad sc_i = 1 + \max(|y_{i0}|, |y_{i1}|). \quad (\text{VII.3.24})$$

En la literatura especializada, estos métodos encajonados con control de paso automático, se los denota por $RK_{p\hat{p}}$ donde RK son las iniciales de Runge-Kutta y p, \hat{p} son los ordenes de los métodos encajonados. En la actualidad la utilización de tales métodos es muy común. En la tabla VII.3.4 se da el esquema del método de Dormand-Prince RK_{54} ; este método es analizado minuciosamente en Hairer & Wanner.

Tabla VII.3.4. Método Dormand-Prince 5(4)

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3187}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
y_1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
\hat{y}_1	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

En el ejemplo siguiente se detallará la construcción de un método encajonado con selección de paso automático. Por razones de simplicidad, se considerará un método RK_{32} .

Ejemplo

Tómese un método de orden $p = 3$ a $s = 3$ pisos, que en este caso será el método de Heun, ver tabla VII.1.1 y se buscará un método encajonado de orden $\hat{p} = 2$ de la forma (VII.3.20), es decir se aumenta s de 1, agregando un $(s + 1)$ -emo piso con los coeficientes $a_{s+1,i}$, para $i = 1, \dots, s$. De esta manera se obtiene el siguiente sistema de ecuaciones para los coeficientes \hat{b}_i , el cual está dado por:

$$\begin{aligned}\hat{b}_1 + \hat{b}_2 + \hat{b}_3 + \hat{b}_4 &= 1 \\ \frac{1}{3}\hat{b}_2 + \frac{2}{3}\hat{b}_3 + \hat{b}_4 &= \frac{1}{2}.\end{aligned}$$

Puedo observarse que para que el método encajonado tenga un orden igual a 2, se requieren 2 condiciones, quedando así dos grados de libertad. Al igual que en el método de Heun, puede plantearse $\hat{b}_2 = 0$, \hat{b}_4 puede escogerse libremente, por ejemplo $\hat{b}_4 = 1/2$ de donde por la segunda ecuación se deduce fácilmente que $\hat{b}_3 = 0$ y por la primera ecuación $\hat{b}_1 = 1/2$. Por lo tanto se obtiene el esquema dado en la tabla VII.3.5.

Tabla VII.3.5. Ejemplo de Método Encajonado

0				
1/3	1/3			
2/3	0	2/3		
1	1/4	0	3/4	
	1/2	0	0	1/2

Soluciones Continuas

Uno de los inconvenientes mayores de los métodos formulados en esta sección consiste en el hecho que estos dan las soluciones en un conjunto discreto de puntos. Una primera alternativa es unir estas soluciones con polígonos, si los puntos donde el método ha dado las soluciones. El resultado no es satisfactorio desde varios puntos de vista. Si el método es de orden elevado la interpolación lineal, vista en el Capítulo III, tiene como efecto la pérdida de precisión. Desde el punto de vista gráfico las soluciones dejan de ser

lisas. Debido a estas dos razones expuestas es necesario complementar estos métodos con alternativas que permitan dar soluciones densas.

Existen varias alternativas validas que permiten subsanar esta falencia de los métodos de Runge-Kutta. La primera consiste en la construcción de los métodos de Runge-Kutta Continuos. Para tal efecto, se considera un método de Runge-Kutta dado de antemano. Se desea evaluar utilizando este esquema en el punto $x^* = x_0 + \theta h$ con $0 < \theta \leq 1$. Los coeficientes de este método se los denota por $c_i(\theta), a_{ij}(\theta), b_j(\theta)$. Ahora bien, el método será interesante desde el punto de vista numérico, si los coeficientes $c_i(\theta), a_{ij}(\theta)$ no dependen de θ coincidiendo de esta manera con los del método dado de antemano. Sin embargo no es nada raro que el método obtenido a partir de los $b_i(\theta)$ tenga un orden inferior. Esto se debe a que los a_{ij}, c_i ya están prescritos. Para ilustrar esta construcción considérese el ejemplo siguiente.

Ejemplo

Se considera nuevamente el método de Heun dado en la tabla VII.3.1. Utilizando el teorema VII.3.4, remplazando $y(x_0 + h)$ por $y(x_0 + \theta h)$ se obtienen las siguientes condiciones de orden para los coeficientes $b_i(\theta)$:

$$\begin{aligned} b_1 + b_2 + b_3 &= \theta, \\ \frac{1}{3}b_2 + \frac{2}{3}b_3 &= \frac{\theta^2}{2}, \\ \frac{1}{9}b_2 + \frac{4}{9}b_3 &= \frac{\theta^3}{3}, \\ \frac{4}{9}b_3 &= \frac{\theta^3}{6}. \end{aligned}$$

Como puede observarse es imposible que los coeficientes b_i puedan satisfacer las condiciones para obtener un método de orden 3, por lo tanto, uno debe contentarse con obtener un método de orden 2 para θ diferente de 1 y para $\theta = 1$ un método de orden 3. Tomando las dos primeras ecuaciones y planteando $b_2 = 0$, como en el método de Heun, se tiene para $b_3 = \frac{3}{4}\theta^2$ y para $b_1 = \theta(\theta - \frac{3}{4}\theta)$. obteniendo así, el esquema dado en la tabla VII.5.6.

Tabla VII.5.6. Ejemplo de Método Continuo.

0			
1/3	1/3		
2/3	0	2/3	
	$\theta(\theta - \frac{3}{4}\theta)$	0	$\frac{3}{4}\theta^2$

La segunda estrategia para construir soluciones continuas, consiste en utilizar la interpolación de Hermite con polinomios cúbicos, tomando como extremidades y_0, y_1 y como derivadas $f(x_0, y_0)$ y $f(x_1, y_1)$. Es fácil ver, que a diferencia de la primera alternativa es necesario evaluar y_1 . Viendo el capítulo III, la interpolación de Hermite tiene un error del orden $\mathcal{O}(h^4)$, equivalente a un método de tercer orden.

Convergencia de los Métodos de Runge-Kutta

En las subsecciones precedentes, se ha hecho un análisis del error local, intimamente ligado al orden del método. Pero no se ha analizado que sucede con el error global de la solución numérica. Para analizar este tipo de error es necesario introducir alguna notación adicional.

Considérese un método a un paso

$$y_{n+1} = y_n + h_n \Phi(x_n, y_n, h_n), \quad (\text{VII.3.25})$$

aplicado a una ecuación diferencial $y' = f(x, y)$, con valor inicial $y(x_0) = y_0$. Es natural plantearse sobre la magnitud del tamaño del error global $y(x_n) - y_n$. A continuación se enunciará un teorema general sobre esta clase de error.

Teorema VII.3.5.- Sea $y(x)$ la solución de $y' = f(x, y)$, $y(x_0) = y_0$ sobre el intervalo $[x_0, x_e]$. Supóngase que:

a) el error local satisface para $x \in [x_0, x_e]$ y $h \leq h_{\max}$

$$\|y(x+h) - y(x) - h\Phi(x, y(x), h)\| \leq C \cdot h^{p+1}, \quad (\text{VII.3.26})$$

b) la función $\Phi(x, y, z)$ satisface una condición de Lipschitz

$$\|\Phi(x, y, h) - \Phi(x, z, h)\| \leq \Lambda \|y - z\|, \quad (\text{VII.3.27})$$

para $h \leq h_{\max}$ y $(x, y), (x, z)$ en un vecindario de la solución.

Entonces, el error global admite para $x_n \leq x_e$ la estimación

$$\|y(x_n) - y_n\| \leq h^p \frac{C}{\Lambda} \left(e^{\Lambda(x_n - x_0)} - 1 \right), \quad (\text{VII.3.28})$$

donde $h = \max_i h_i$, bajo la condición que h sea lo suficientemente pequeño.

Demostración.- La idea central de la demostración es estudiar la influencia del error local, cometido en el paso i -ésimo a la propagación de y_n . Enseguida, adicionar los errores acumulados. Ver figura VII.3.2.

Sean $\{y_n\}$ y $\{z_n\}$ dos soluciones numéricas obtenidas por (VII.3.25) con valores iniciales y_0 y z_0 respectivamente. Utilizando la condición de Lipschitz dada por (VII.3.27), su diferencia puede ser estimada como

$$\begin{aligned} \|y_{n+1} - z_{n+1}\| &\leq \|y_n - z_n\| + h_n \Lambda \|y_n - z_n\| \\ &= (1 + \Lambda h_n) \|y_n - z_n\| \leq e^{h_n \Lambda} \|y_n - z_n\|. \end{aligned} \quad (\text{VII.3.29})$$

Recursivamente, se obtiene

$$\|y_n - z_n\| \leq e^{h_{n-1}\Lambda} e^{h_{n-2}\Lambda} \dots e^{h_i\Lambda} \|y_i - z_i\|.$$

y el error propagado E_i , ver figura VII.3.2, satisface

$$\|E_i\| \leq e^{\Lambda(x_n - x_i)} \|e_i\| \leq Ch_{i-1}^{p+1} e^{\Lambda(x_n - x_i)}, \quad (\text{VII.3.30})$$

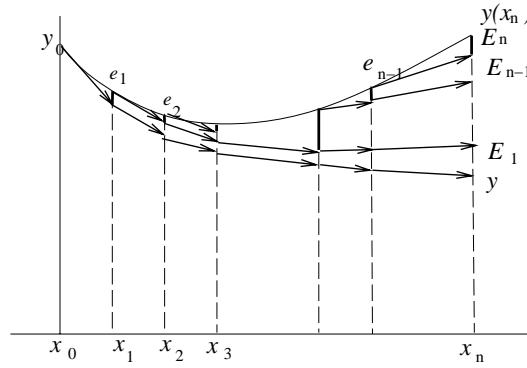


Figura VII.3.2. Estimación del error global.

La desigualdad del triángulo, así como (VII.3.30) da, ver la figura VII.3.3, para la estimación de la suma

$$\begin{aligned} \|y(x_n) - y_n\| &\leq \sum_{i=1}^n \|E_i\| \leq \sum_{i=1}^n h_{i-1}^{p+1} e^{\Lambda(x_n - x_i)} \\ &\leq Ch^p (h_0 e^{\Lambda(x_n - x_1)} + \dots + h_{n-2} e^{\Lambda(x_n - x_{n-1})} + h_{n-1}) \\ &\leq Ch^p \int_{x_0}^{x_n} e^{\Lambda(x_n - t)} dt = \frac{Ch^p}{\Lambda} (e^{\Lambda(x_n - x_0)} - 1). \end{aligned}$$

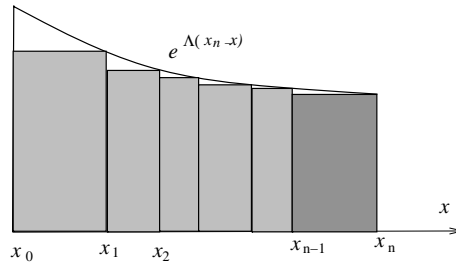


Figura VII.3.3. Estimación de la suma de Riemann.

Solo queda justificar la implicación de (VII.3.25) en (VII.3.27), ya que la estimación (VII.3.25) es válida solamente en un vecindario

$\mathcal{U} = \{(x, y) \mid \|y - y(x)\| \leq b\}$ de la solución exacta. Si se supone que h es lo suficientemente pequeño, más precisamente, si h es tal que

$$\frac{Ch^p}{\Lambda} \left(e^{\Lambda(x_e - x_0)} - 1 \right) \leq b,$$

se estará seguro que todas las soluciones numéricas de la figura VII.3.2 se quedarán en \mathcal{U} . \square

Supóngase que (VII.3.25) representa un método de Runge-Kutta, se verificará las hipótesis del teorema precedente. La condición (VII.3.26) es satisfecha para un método de orden p . Solo queda por verificar la condición de Lipschitz (VII.3.27), para la función

$$\Phi(x, y, h) = \sum_{i=1}^s b_i k_i(y), \quad (\text{VII.3.31})$$

donde

$$k_i(y) = f \left(x + c_i h, y + h \sum_{j=1}^{i-1} a_{ij} k_j(y) \right). \quad (\text{VII.3.32})$$

Proposición VII.3.6.- Si $f(x, y)$ satisface una condición de Lipschitz $\|f(x, y) - f(x, z)\| \leq L \|y - z\|$ en un vecindario de la solución de $y' = f(x, y)$, la expresión $\Phi(x, y, h)$ de (VII.3.31) verifica la condición (VII.3.27) con

$$\Lambda = L \left(\sum_i |b_i| + (h_{\max} L \sum_{ij} |b_i a_{ij}| + (h_{\max} L)^2 \sum_{ijk} |b_i a_{ij} a_{jk}| + \dots) \right). \quad (\text{VII.3.33})$$

Demostración.- La condición de Lipschitz para $f(x, y)$ aplicado a (VII.3.32) da

$$\begin{aligned} \|k_1(y) - k_1(z)\| &= \|f(x, y) - f(x, z)\| \leq L \|y - z\| \\ \|k_2(y) - k_2(z)\| &\leq L \|y - z + h a_{21} (k_1(y) - k_1(z))\| \\ &\leq L(1 + h_{\max} L |a_{21}|) \|y - z\|, \end{aligned} \quad (\text{VII.3.34})$$

etc. Las estimaciones (VII.3.34) introducidas en

$$\|\Phi(x, y, h) - \Phi(x, z, h)\| \leq \sum_{i=1}^s |b_i| \|k_i(y) - k_i(z)\|,$$

implican (VII.3.27) con Λ dado por (VII.3.33). \square

Experiencias Numéricas

En esta subsección se mostrará varios ejemplos donde los métodos de Runge-Kutta actúan. Se comparará varios métodos.

1.- Se considera la ecuación diferencial, con valor inicial,

$$y' = -y + \sin x, \quad y(0) = 1; \quad (\text{VII.3.35})$$

la solución de este problema, está dada por

$$y(x) = \frac{3}{2}e^{-x} - \frac{1}{2}\cos x + \frac{1}{2}\sin x. \quad (\text{VII.3.36})$$

En la figura VII.3.4, se muestra las gráficas de diferentes métodos de Runge-Kutta. El intervalo de integración es $[0, 10]$. Puede comprobarse, que al igual que en los métodos de integración, existe una relación lineal entre $-\log_{10}(\text{Error Global})$ y el número de evaluaciones de la función, fe . Esta relación lineal tiene una pendiente $1/p$, donde p es el orden del método.

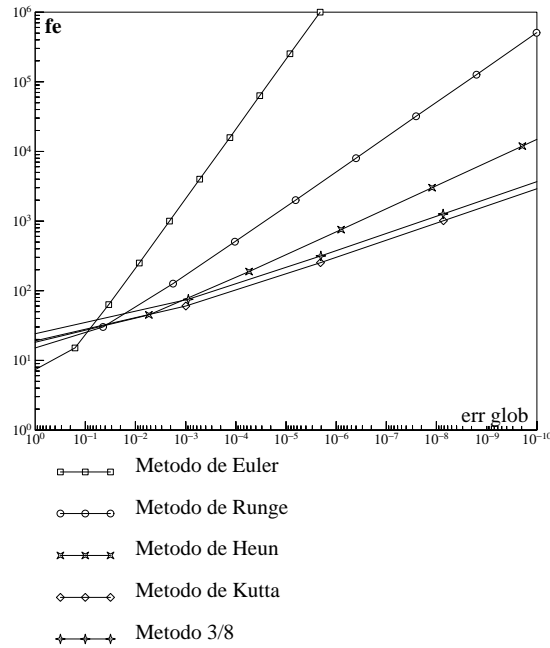


Figura VII.3.4. Relación del error global vs fe .

2.- En esta experiencia numérica, se implementa un método encajonado, para este efecto, se toma el método dado en uno de los ejemplos precedentes. La ecuación diferencial sobre la cual es examinada tal método, es una ecuación diferencial de una reacción química, conocida como Brusselator. Por consiguiente, se debe resolver:

$$\begin{aligned} y_1' &= 1 + y_1^2 y_2 - 4y_1, & y_1(0) &= 1.5, \\ y_2' &= 3y_1 - y_1^2 y_2, & y_2(0) &= 3, \end{aligned} \quad (\text{VII.3.37})$$

sobre el intervalo $[0, 10]$. Los resultados obtenidos con $Tol = 10^{-5}$ son presentados en la figura VII.3.5. En la gráfica superior, las dos componentes de la solución utilizando la solución continua del método de Heun, ver métodos continuos. En la gráfica del medio, la longitud de los pasos, los pasos aceptados están ligados, mientras que los pasos rechazados indicados por \times . En la gráfica inferior, la estimación del error local, como también los valores exactos del error local y global.

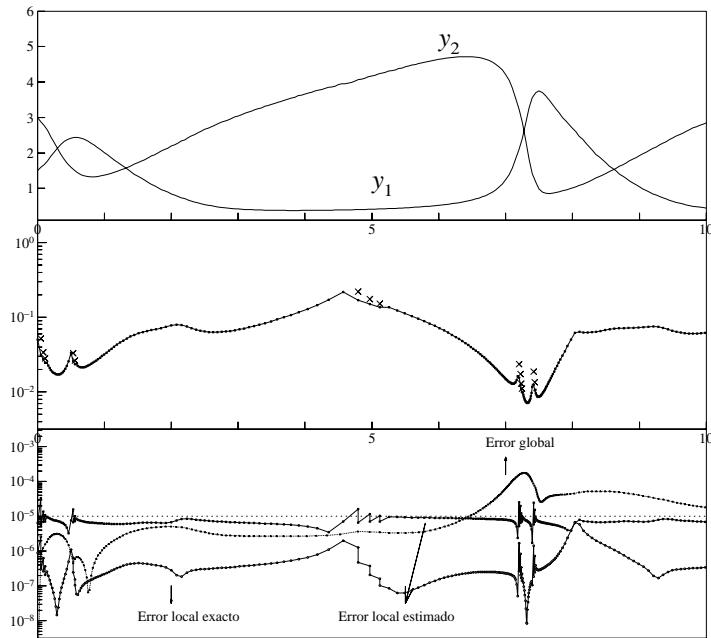


Figura VII.3.5. Experiencia numérica de un método encajonado.

Ejercicios

- 1.- Aplicar el método de Runge al sistema

$$\begin{aligned}y_1' &= -y_2, & y_1(0) &= 1, \\y_2' &= y_1, & y_2(0) &= 0,\end{aligned}$$

con $h = 1/5$. Comparar el trabajo necesario y la precisión del resultado con el método de Euler.

- 2.- Considérese un método de Runge-Kutta a s pisos y de orden p . Mostrar que

$$\sum_{i=1}^s b_i c_i^{q-1} = \frac{1}{q} \quad \text{para } q = 1, \dots, p.$$

es decir, la fórmula de cuadratura asociada tiene al menos el orden p .

- 3.- Aplicar un método de Runge-Kutta de orden $p = s$, s es el número de pisos, al problema $y' = \lambda y$, donde λ es una constante compleja. Mostrar que la solución numérica está dada por

$$y_1 = \left(\sum_{j=0}^s \frac{z^j}{j!} \right) y_0, \quad z = \lambda h.$$

- 4.- Construir todos los métodos de Runge-Kutta de orden 3 con $s = 3$ pisos.
5.- Considérese el problema

$$y' = \frac{\lambda}{x} y + g(x), \quad y(0) = 0$$

con $g(x)$ suficientemente derivable y $\lambda \leq 0$. Mostrar que este problema admite una y una sola solución. Las derivadas de la solución en el punto $x = 0$ son

$$y^{(j)}(0) = \left(1 - \frac{\lambda}{j} \right)^{-1} g^{(j-1)}(0).$$

- 6.- Aplicar un método de Runge-Kutta al problema del ejercicio anterior, utilizar la definición $f(x, y) = \left(1 - \frac{\lambda}{j} \right)^{-1} g(0)$, para $x = 0$.
a) Mostrar que el error del primer paso satisface

$$y(h) - y_1 = C_2 h^2 g'(0) + \mathcal{O}(h^3)$$

donde C_2 depende solamente de los coeficientes del método.

- c) Calcular C_2 para uno de los métodos de Kutta.

VII.4 Métodos Multipasos

Los métodos a un paso calculan un valor aproximado y_{n+1} , en el instante t_{n+1} , utilizando únicamente el valor aproximado y_n . La utilización de varios valores aproximados y_n, y_{n-1}, \dots , permite obtener a precisión igual métodos de costo menos elevado; estos métodos son comúnmente llamados multipasos, más precisamente cuando los cálculos utilizan r valores precedentes: $y_n, y_{n-1}, \dots, y_{n-r+1}$, es decir concernientes los r pasos: $h_n, h_{n-1}, \dots, h_{n-r+1}$, se habla de métodos a r pasos. Entre estos métodos, los métodos de Adams son aquéllos que parecen ser a la hora actual los más eficaces cuando el problema diferencial es bien condicionado, serán estudiados en esta subsección. Los algoritmos modernos utilizan estos métodos con un número variable de pasos y un tamaño de paso variables. La variación del número de pasos y la longitud de los pasos permite adaptar el método a la regularidad de la solución, permite también de levantar las dificultades de arranque que presentan los métodos a número de pasos fijo.

Métodos de Adams Explícitos

Sea una división $x_0 < x_1 < \dots < x_n = x_e$ del intervalo sobre el cual se busca resolver la ecuación diferencial $y' = f(x, y)$ y supóngase que se conoce aproximaciones $y_n, y_{n-1}, \dots, y_{n-k+1}$ de la solución para k pasos consecutivos ($y_i \approx y(x_i)$). De la misma forma que para la derivación de los métodos de Runge-Kutta, se integra la ecuación diferencial, para obtener

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(t, y(t)) dt. \quad (\text{VII.4.1})$$

Pero en lugar de aplicar una fórmula de cuadratura estándar a la integral (VII.4.1), se reemplaza $f(t, y(t))$ por el polinomio de grado $k-1$, que satisfaga

$$p(x_j) = f_j, \quad \text{para } j = n, n-1, \dots, n-k+1; \quad (\text{VII.4.2})$$

donde $f_j = f(x_j, y_j)$. Ver la figura VII.4.1.

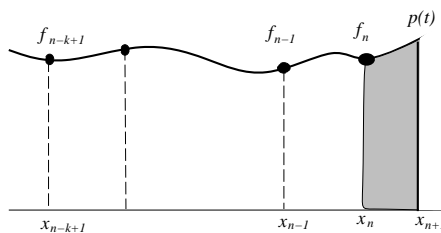


Figura VII.4.1. Método de Adams.

Por consiguiente, la aproximación de $y(x_{n+1})$ está definida por

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p(t) dt. \quad (\text{VII.4.3})$$

Si se representa el polinomio $p(t)$ por la fórmula de Newton, ver el teorema III.1.9,

$$p(t) = \sum_{j=0}^{k-1} \left(\prod_{i=0}^{j-1} (t - x_{n-i}) \right) f[x_n, x_{n-1}, \dots, x_{n-j}], \quad (\text{VII.4.4})$$

el método (VII.4.3) se convierte en

$$y_{n+1} = y_n + \sum_{j=0}^{k-1} \left(\int_{x_n}^{x_{n+1}} \prod_{i=0}^{j-1} (t - x_{n-i}) \right) f[x_n, x_{n-1}, \dots, x_{n-j}]. \quad (\text{VII.4.5})$$

El caso más simple de estas fórmulas de Adams explícitas consiste cuando la división es equidistante. En esta situación se tiene $x_j = x_0 + jh$, las diferencias divididas pueden ser expresadas bajo la forma

$$f[x_n, x_{n-1}, \dots, x_{n-j}] = \frac{\nabla^j f_n}{j! h^j}, \quad (\text{VII.4.6})$$

donde $\nabla^0 f_n = f_n$, $\nabla f_n = f_n - f_{n-1}$, $\nabla^2 f_n = \nabla(\nabla f_n)$, ... son las diferencias finitas regresivas.

La fórmula (VII.4.5), planteando $t = x_n + sh$, se convierte en

$$y_{n+1} = y_n + h \sum_{j=0}^{k-1} \gamma_j \nabla^j f_n, \quad (\text{VII.4.7})$$

donde

$$\gamma_j = \frac{1}{j!} \int_0^1 \prod_{i=0}^{j-1} (i + s) ds = \int_0^1 \binom{s + j - 1}{j} ds.$$

Los primeros coeficientes γ_j están dados en la tabla VII.4.1.

Tabla VII.4.1. Coeficientes para los métodos de Adams Explícitos.

j	0	1	2	3	4	5	6	7	8
γ_j	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19087}{60480}$	$\frac{5257}{17280}$	$\frac{1070017}{3628800}$

Casos particulares de los métodos de Adams explícitos son:

$$k = 1 : \quad y_{n+1} = y_n + hf_n, \quad (\text{método de Euler});$$

$$k = 2 : \quad y_{n+1} = y_n + h \left(\frac{2}{3}f_n - \frac{1}{2}f_{n-1} \right);$$

$$k = 3 : \quad y_{n+1} = y_n + h \left(\frac{23}{12}f_n - \frac{16}{12}f_{n-1} + \frac{5}{12}f_{n-2} \right);$$

$$k = 4 : \quad y_{n+1} = y_n + h \left(\frac{55}{24}f_n - \frac{59}{24}f_{n-1} + \frac{37}{24}f_{n-2} - \frac{9}{24}f_{n-3} \right).$$

Si se quiere aplicar este método, por ejemplo para $k=4$, a la resolución de $y' = f(x, y)$, $y(x_0) = y_0$, es necesario conocer y_0, y_1, y_2 y y_3 . Después se puede utilizar la fórmula de recurrencia para determinar y_4, y_5, \dots . Al formular Adams sus métodos, el utilizaba la serie de Taylor de la solución en el punto inicial, sin embargo es más cómodo empezar con un método a un paso del tipo de Runge-Kutta y luego utilizar el método multipaso correspondiente.

En la construcción del método de Adams explícito dado por (VII.4.5), se ha utilizado el polinomio de interpolación $p(t)$ fuera del intervalo $[x_{n-k+1}, x_n]$. Esta situación puede provocar grandes errores, ver nuevamente III.1. Por lo tanto, este inconveniente puede modificarse considerando los:

Métodos de Adams Implícitos

La idea central de estos métodos consiste en considerar el polinomio $p^*(t)$ de grado k , que satisface

$$p^*(t) = f_j \quad \text{para,} \quad j = n+1, n, n-1, \dots, n-k+1. \quad (\text{VII.4.9})$$

A diferencia de la versión explícita $f_{n+1} = f(x_{n+1}, y_{n+1})$ es todavía desconocido. El siguiente paso es definir la aproximación numérica por

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p^*(t) dt. \quad (\text{VII.4.10})$$

De la misma manera que en el caso explícito, la fórmula de Newton da

$$p^*(t) = \sum_{j=0}^k \left(\prod_{i=0}^{j-1} (t - x_{n-i+1}) \right) f[x_{n+1}, x_n, \dots, x_{n-j+1}], \quad (\text{VII.4.11})$$

de donde el método está dado por

$$y_{n+1} = y_n + \sum_{j=0}^k \left(\int_{x_n}^{x_{n+1}} \prod_{i=0}^{j-1} (t - x_{n-i+1}) \right) f[x_{n+1}, x_n, \dots, x_{n-j+1}]. \quad (\text{VII.4.12})$$

El caso más simple de estas fórmulas de Adams implícitas consiste cuando la división es equidistante, el método está dado por

$$y_{n+1} = y_n + h \sum_{j=0}^k \gamma_j^* \nabla^j f_{n+1}, \quad (\text{VII.4.13})$$

donde

$$\gamma_j^* = \frac{1}{j!} \int_0^1 \prod_{i=0}^{j-1} (i + s - 1) ds = \int_0^1 \binom{s + j - 2}{j} ds. \quad (\text{VII.4.14})$$

Los primeros coeficientes γ_j^* están dados en la tabla VII.4.2.

Tabla VII.4.2. Coeficientes para los métodos de Adams Implícitos.

j	0	1	2	3	4	5	6	7	8
γ_j^*	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$	$-\frac{863}{60480}$	$-\frac{375}{24192}$	$-\frac{33953}{3628800}$

Casos particulares de los métodos de Adams implícitos son:

$$k = 0: \quad y_{n+1} = y_n + h f_{n+1} = y_n + h f(x_{n+1}, y_{n+1});$$

$$k = 1: \quad y_{n+1} = y_n + h \left(\frac{1}{2} f_{n+1} + \frac{1}{2} f_n \right);$$

$$k = 2: \quad y_{n+1} = y_n + h \left(\frac{5}{12} f_{n+1} + \frac{8}{12} f_n - \frac{1}{12} f_{n-1} \right);$$

$$k = 2: \quad y_{n+1} = y_n + h \left(\frac{9}{24} f_{n+1} + \frac{19}{24} f_n - \frac{5}{24} f_{n-1} + \frac{1}{24} f_{n-2} \right).$$

Cada una de estas fórmulas representa una ecuación no lineal para y_{n+1} , que es de la forma

$$y_{n+1} = \eta_n + h \beta f(x_{n+1}, y_{n+1}). \quad (\text{VII.4.15})$$

que puede ser resuelta por el método de Newton o simplemente por el método iterativo simple.

Métodos Predictor-Corrector

Los métodos de Adams implícitos tienen la gran ventaja sobre la versión explícita, por que las soluciones provistas son más realistas. Sin embargo la gran dificultad de estos métodos es la resolución de (VII.4.15). En algunos casos particulares, como en el caso de los sistemas lineales, esta resolución

podrá hacerse directamente, pero en el caso general la utilización de un método iterativo para resolver el sistema no lineal es necesaria. Por otro lado, es necesario recalcar que y_{n+1} es una aproximación de la solución exacta, motivo por el cual es natural pensar que un cálculo muy preciso de y_{n+1} tal vez no sea necesario. Es por eso, que una mejora de esta situación puede darse de la manera siguiente. Primero calcular una primera aproximación por un método explícito, luego corregir este valor, una o varias veces, utilizando la fórmula (VII.4.15). Con este algoritmo, un paso de este método toma la forma siguiente:

P: se calcula el predictor $\hat{y}_{n+1} = \hat{y}_n + h \sum_{j=0}^{k-1} \gamma_j \nabla^j f_n$ por el método de Adams explícito; \hat{y}_{n+1} es ya una aproximación de $y(x_{n+1})$.

E: evaluación de la función: se calcula $\hat{f}_{n+1} = f(x_{n+1}, \hat{y}_{n+1})$.

C: la aproximación corregida está dada por $y_{n+1} = \eta_n + h\beta\hat{f}_{n+1}$.

E: calcular $f_{n+1} = f(x_{n+1}, y_{n+1})$.

Este procedimiento, que se lo denota PECE, es el más utilizado. Otras posibilidades son: efectuar varias iteraciones, por ejemplo PECECE, o de omitir la última evaluación de f , es decir PEC, y de tomar \hat{f}_{n+1} en el lugar de f_{n+1} para el paso siguiente.

Metodos BDF

Se ha visto que los métodos de Adams, tanto en su versión explícita, como en su versión implícita consideran polinomios de interpolación que pasan por los f_j . Esta manera de abordar el problema de la aproximación de las soluciones de un problema diferencial es eficaz cuando el problema diferencial es bien condicionado, pero éstos pueden volverse muy costosos desde el punto de vista computacional para los problemas diferenciales rígidos. Por consiguiente, puede ser interesante de utilizar el método de las diferencias retrógradas que será descrito en esta subsección.

La idea central de los métodos BDF consiste en considerar el polinomio $q(t)$ de grado k , ver figura VII.4.2, definido por

$$q(x_j) = y_j, \quad \text{para } j = n+1, n, \dots, n-k+1; \quad (\text{VII.4.16})$$

y se determina y_{n+1} , de manera que

$$q'(x_{n+1}) = f(x_{n+1}, q(x_{n+1})). \quad (\text{VII.4.17})$$

Como en la fórmula (VII.4.11), la fórmula de Newton da

$$q(t) = \sum_{j=0}^k \left(\prod_{i=0}^{j-1} (t - x_{n-i+1}) \right) y[x_{n+1}, x_n, \dots, x_{n-j+1}]. \quad (\text{VII.4.18})$$

Cada término de esta suma contiene el factor $(t - x_{n+1})$, de donde es muy fácil calcular $q'(x_{n+1})$, obteniendo así

$$\sum_{j=0}^k \left(\prod_{i=1}^{j-1} (x_{n+1} - x_{n-i+1}) \right) y[x_{n+1}, x_n, \dots, x_{n-j+1}] = f(x_{n+1}, y_{n+1}). \quad (\text{VII.4.19})$$

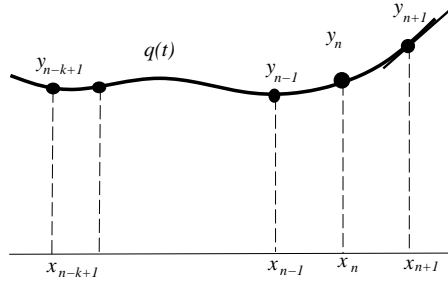


Figura VII.4.2. Método de tipo BDF

Para el caso equidistante, (VII.4.19) utilizando (VII.4.6) se convierte en

$$\sum_{j=1}^k \frac{1}{j} \nabla^j y_{n+1} = h f_{n+1}. \quad (\text{VII.4.20})$$

Obteniendo de esta manera, los siguientes casos particulares:

$$\begin{aligned} k = 1 : & \quad y_{n+1} - y_n = h f_{n+1}; \\ k = 2 : & \quad \frac{3}{2} y_{n+1} - 2 y_n + \frac{1}{2} y_{n-1} = h f_{n+1}; \\ k = 3 : & \quad \frac{11}{6} y_{n+1} - 3 y_n + \frac{3}{2} y_{n-1} - \frac{1}{3} y_{n-2} = h f_{n+1}; \\ k = 4 : & \quad \frac{25}{12} y_{n+1} - 4 y_n + 3 y_{n-1} - \frac{4}{3} y_{n-2} + \frac{1}{4} y_{n-3} = h f_{n+1}. \end{aligned}$$

De nuevo, cada fórmula define implícitamente la aproximación numérica y_{n+1} .

Estudio del Error Local

Al igual de los métodos a un paso, como los métodos de Runge-Kutta, existe la noción de orden para los métodos multipasos, el cual está íntimamente ligado al estudio del error local.

Los métodos multipasos pueden expresarse bajo la forma siguiente

$$\sum_{i=0}^k \alpha_i y_{n+i} = h \sum_{i=0}^k \beta_i f_{n+i}, \quad (\text{VII.4.21})$$

donde $\alpha_k \neq 0$ y $|\alpha_0| + |\beta_0| > 0$. El método es explícito si $\beta_k = 0$, si no el método es implícito.

Definición VII.4.1.- Sea $y(x)$ una solución de $y' = f(x, y(x))$ y sea y_{n+k} el valor obtenido por el método (VII.4.21) utilizando $y_i = y(x_i)$ para $i = n, \dots, n+k-1$, ver figurar VII.4.3. Entonces

$$\text{error local} = y(x_{n+k}) - y_{n+k}. \quad (\text{VII.4.22})$$

Se dice que el método (VII.4.21) tiene el orden p si el error local es $\mathcal{O}(h^{p+1})$.

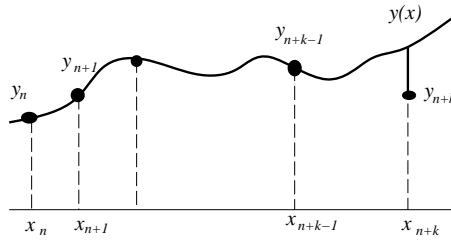


Figura VII.4.3. Definición del error local.

Para estudiar el error local de un método multipaso, se utiliza el operador L definido por:

$$L(y, x, h) = \sum_{i=0}^k (\alpha_i y(x + ih) - h\beta_i y'(x + ih)). \quad (\text{VII.4.23})$$

Como $y_i = y(x_i)$, para $i = n, \dots, n+k-1$ en la definición dada más arriba, se tiene que $f_i = f(x_i, y(x_i)) = y'(x_i)$ para $i = n, \dots, n+k-1$ y la fórmula (VII.4.21) puede ser expresada bajo la forma

$$\begin{aligned} \sum_{i=0}^k \alpha_i y(x_n + ih) + \alpha_k (y_{n+k} - y(x_{n+k})) &= h \sum_{i=0}^k \beta_i y'(x_n + ih) \\ &\quad + h\beta(f_{n+k} - f(x_{n+k}, y(x_{n+k}))), \end{aligned}$$

lo que es equivalente a

$$L(y, x_n, h) = \left(\alpha_k I - h\beta_k \frac{\partial f}{\partial y}(\dots) \right) (y(x_{n+k}) - y_{n+k}), \quad (\text{VII.4.24})$$

el argumento de $\partial f / \partial y$ puede ser diferente para cada línea de esta matriz, ver el teorema del valor medio o incrementos finitos. La fórmula muestra que el error local del método (VII.4.21) es $\mathcal{O}(h^{p+1})$ si y solamente si $L(y, x, h) = \mathcal{O}(h^{p+1})$ para toda función $y(x)$ que es suficientemente derivable.

Teorema VII.4.2.- *Un método multipaso tiene el orden p , si y solamente si sus coeficientes satisfacen*

$$\sum_{i=0}^k \alpha_i = 0 \quad \text{y} \quad \sum_{i=0}^k \alpha_i i^q = q \sum_{i=0}^k \beta_i i^{q-1} \quad \text{para } q = 1, \dots, p. \quad (\text{VII.4.25})$$

Demostración.- En la fórmula (VII.4.23), se desarrolla las expresiones $y(x + ih)$ y $y'(x + ih)$ en serie de Taylor y obteniendo

$$\begin{aligned} L(y, x, h) &= \sum_{i=0}^k \alpha_i \sum_{g \geq 0} y^{(g)}(x) \frac{(ih)^g}{g!} - h \sum_{i=0}^k \beta_i \sum_{r \geq 0} y^{(r+1)}(x) \frac{(ih)^r}{r!} \\ &= y(x) \left(\sum_{i=0}^k \alpha_i \right) + \sum_{g \geq 1} y^{(g)}(x) \frac{h^g}{g!} \left(\sum_{i=0}^k \alpha_i i^g - g \sum_{i=0}^k \beta_i i^{g-1} \right). \end{aligned}$$

La condición $L(y, x, h) = \mathcal{O}(h^{p+1})$ da la condición (VII.4.25). \square

Ejemplo

Considérese el método de Adams explícito a k pasos. Para $q \leq k$, se considera la ecuación diferencial $y' = qx^{q-1}$ cuya solución es $y(x) = x^q$. En esta situación, el polinomio $p(t)$ de (VII.4.2) es igual a $f(t, y(t))$ y el método de Adams explícito da el resultado exacto. Por consiguiente, se tiene $L(y, x, h) = 0$, ver la fórmula (VII.4.24) lo que implica

$$\sum_{i=0}^k (\alpha_i (x + ih)^q - q \beta_i (x + ih)^{q-1} h) = 0$$

y por lo tanto (VII.4.25), planteando $x = 0$. Entonces, el orden de este método es $\geq k$. Se puede mostrar que efectivamente el orden es igual a k .

De la misma manera, se muestra que el método de Adams implícito tiene el orden $p = k + 1$ y el método BDF el orden $p = k$.

Estabilidad

A simple vista la relación (VII.4.25) permite formular métodos multipaso con un número dado de pasos con un orden optimal. Sin embargo la experiencia

numérica mostró que los resultados obtenidos no siempre eran válidos. Es por eso necesario introducir una nueva noción en el estudio de estos métodos. Esta noción está íntimamente ligada a la estabilidad del método. Para comprender mejor el problema que se plantea con los métodos multipaso, es necesario referirse al siguiente ejemplo enunciado por Dahlquist en 1956.

Se plantea $k = 2$ y se construye un método explícito, $\beta_2 = 0$, $\alpha_2 = 1$ con un orden maximal. Utilizando las condiciones dadas por (VII.4.25) con $p = 3$, se llega al siguiente esquema:

$$y_{n+2} + 4y_{n+1} - 5y_n = h(4f_{n+1} + 2f_n). \quad (\text{VII.4.26})$$

Una aplicación a la ecuación diferencial $y' = y$ con $y(0) = 1$ da la fórmula de recurrencia

$$y_{n+2} + 4(1-h)y_{n+1} + (5+2h)y_n = 0. \quad (\text{VII.4.27})$$

Para resolver la ecuación precedente, se calcula el polinomio característico de ésta, obteniendo

$$\zeta^2 + 4(1-h)\zeta - (5+2h) = 0, \quad (\text{VII.4.28})$$

ecuación de segundo grado cuyas soluciones son

$$\zeta_1 = 1 + h + \mathcal{O}(h^2), \quad \zeta_2 = -5 + \mathcal{O}(h).$$

La solución de (VII.4.27) tiene la forma

$$y_n = C_1 \zeta_1^n + C_2 \zeta_2^n \quad (\text{VII.4.29})$$

donde las constantes C_1 y C_2 están determinadas por $y_0 = 1$ y $y_1 = e^h$. Para n grande, el término $C_2 \zeta_2^n \approx C_2 (-5)^n$ es dominante y será muy difícil que la solución numérica converga hacia la solución exacta. Ver figura VII.4.4.

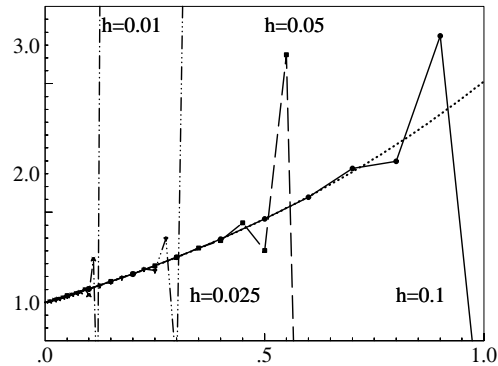


Figura VII.4.4. Inestabilidad del método VII.4.26.

La razón de la divergencia de la solución numérica en el ejemplo precedente, es que el polinomio

$$\varrho(\zeta) = \sum_{i=0}^k \alpha_i \zeta^i, \quad (\text{VII.4.30})$$

tiene una raíz que es más grande que 1 en valor absoluto.

Para encontrar una condición necesaria para la convergencia, se considerará el problema $y' = 0$ con valores iniciales y_0, y_1, \dots, y_{k-1} perturbados. La solución numérica y_n satisface:

$$\alpha_k y_{n+k} + \dots + \alpha_0 y_n = 0, \quad (\text{VII.4.31})$$

y está dada por una combinación lineal de:

$$\begin{aligned} \zeta^n & \dots \dots \dots \text{ si } \zeta \text{ es una raíz simple de } \varrho(\zeta) = 0, \\ \zeta^n, n\zeta^n & \dots \dots \dots \text{ si } \zeta \text{ es una raíz doble de } \varrho(\zeta) = 0, \\ \zeta^n, n\zeta^n, \dots, n^l \zeta^n & \dots \dots \text{ si } \zeta \text{ es una raíz de multiplicidad } l. \end{aligned}$$

Para que la solución numérica quede acotada, es necesario que las condiciones de la definición siguiente sean satisfechas.

Definición VII.4.3.- Un método multipaso es estable, si las raíces del polinomio $\varrho(\zeta)$ satisfacen

- i) si $\varrho(\hat{\zeta}) = 0$ entonces $|\hat{\zeta}| \leq 1$,
- ii) si $\varrho(\hat{\zeta}) = 0$ y $|\hat{\zeta}| = 1$ entonces $\hat{\zeta}$ es una raíz simple de $\varrho(\zeta)$.

Se puede mostrar fácilmente que los métodos de Adams tienen como polinomio característico

$$\varrho(\zeta) = \zeta^{k-1}(\zeta - 1).$$

Por consiguiente los métodos de Adams son estables para la definición que acaba de ser enunciada. Por otro lado se muestra igualmente que los métodos BDF son estables solamente para $k \leq 6$.

Existe un resultado muy importante en la teoría de la estabilidad de los métodos multipaso, cuya demostración escapa a los objetivos del libro. Este resultado es conocido como la primera barrera de Dahlsquist.

Teorema VII.4.4.- *Primera barrera de Dahlsquist. Para un método multipaso estable, el orden k satisface $p \leq k + 2$ si k es par, $p \leq k + 1$ si k es impar y $p \leq k$ si el método es explícito.*

Convergencia de los Métodos Multipaso

Debido a la complejidad de la demostración, en esta subsección se tratará, solamente el caso equidistante.

Teorema VII.4.5.- Supóngase que los k valores de partida satisfagan $\|y(x_i) - y_i\| \leq C_0 h^p$, para $i = 0, 1, \dots, k-1$. Si el método multipaso (VII.4.21) es de orden p y estable, entonces el método es convergente de orden p , es decir el error global satisface

$$\|y(x_n) - y_n\| \leq Ch^p, \quad \text{para } x_n - x_0 = nh \leq \text{Const.} \quad (\text{VII.4.34})$$

Demostración.- El punto esencial de la demostración es el siguiente: Se escribe formalmente el método multipaso (VII.4.21) bajo la forma de un método a un paso. El método multipaso es de la forma, suponiendo $\alpha_k = 1$

$$y_{n+k} = - \sum_{i=0}^{k-1} \alpha_i y_{n+i} + h \Psi(x_n, y_n, \dots, y_{n+k-1}, h). \quad (\text{VII.4.35})$$

Para un método explícito $\beta_k = 0$, Ψ está dada por

$$\Psi(x_n, y_n, \dots, y_{n+k-1}, h) = \sum_{i=0}^{k-1} \beta_i f(x_{n+i}, y_{n+i}),$$

y para un método general, Ψ está definida implícitamente por

$$\begin{aligned} \Psi(x_n, y_n, \dots, y_{n+k-1}, h) = & \sum_{i=0}^{k-1} \beta_i f(x_{n+i}, y_{n+i}) \\ & + \beta_k f \left(x_{n+k}, h \Psi(x_n, y_n, \dots, y_{n+k-1}, h) - \sum_{i=0}^{k-1} \alpha_i y_{n+i} \right). \end{aligned} \quad (\text{VII.4.36})$$

Luego se considera los supervectores

$$Y_n = (y_{n+k-1}, \dots, y_{n+1}, y_n)^t,$$

pudiendo de esta manera escribir el método dado por (VII.4.35), bajo la forma

$$Y_{n+1} = AY_n + h\Phi(x_n, Y_n, h), \quad (\text{VII.4.37})$$

donde

$$A = \begin{pmatrix} -\alpha_{k-1}\mathbb{1} & -\alpha_{k-2}\mathbb{1} & \cdots & -\alpha_1\mathbb{1} & -\alpha_0\mathbb{1} \\ \mathbb{1} & 0 & \cdots & 0 & 0 \\ & \mathbb{1} & \ddots & & O \\ & & \ddots & \ddots & \vdots \\ & & & \mathbb{1} & 0 \end{pmatrix}, \quad \Phi(x, Y, h) = \begin{pmatrix} \Psi(x, Y, h) \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (\text{VII.4.37})$$

El siguiente paso en la demostración es introducir el error local. Se considera los valores y_n, \dots, y_{n+k-1} sobre la solución exacta, denotando

$$Y(x_n) = (y(x_{n+k-1}), \dots, y(x_{n+1}), y(x_n))^t, \quad (\text{VII.4.38})$$

y aplicando una vez el método multipaso. Esto da

$$\hat{Y}_{n+1} = AY(x_n) + h\Phi(x_n, Y(x_n), h).$$

La primera componente de $\hat{Y}_{n+1} - Y(x_{n+1})$ es exactamente el error local dado por (VII.4.22), mientras que las otras componentes son iguales a cero. Como el método es de orden p , por hipótesis, se tiene

$$\|\hat{Y}_{n+1} - Y(x_{n+1})\| \leq C_1 h^{p+1}, \quad \text{para } x_{n+1} - x_0 = (n+1)h \leq \text{Const.} \quad (\text{VII.4.39})$$

A continuación se debe analizar la propagación del error, es decir introducir la estabilidad del método. Considérese una segunda solución numérica, definida por

$$Z_{n+1} = AZ_n + h\Phi(x_n, Z_n, h)$$

y estimar la diferencia $Y_{n+1} - Z_{n+1}$. Como ilustración del siguiente paso en la demostración se considerará solamente los métodos de Adams, pues el caso general requiere de la utilización de otra norma, cuyo estudio escapa el alcance de este libro. Por consiguiente

$$\begin{pmatrix} y_{n+k} - z_{n+k} \\ y_{n+k-1} - z_{n+k-1} \\ \vdots \\ y_{n+1} - z_{n+1} \end{pmatrix} = \begin{pmatrix} y_{n+k-1} - z_{n+k-1} \\ \vdots \\ y_{n+1} - z_{n+1} \\ y_n - z_n \end{pmatrix} + h \begin{pmatrix} \Psi(x_n, Y_n, h) - \Psi(x_n, Z_n, h) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Utilizando la norma infinita y condición de Lipschitz para Ψ que es consecuencia de la de $f(x, y)$, se obtiene

$$\|Y_{n+1} - Z_{n+1}\| \leq (1 + h\Lambda) \|Y_n - Z_n\|. \quad (\text{VII.4.40})$$

Ahora se verá la acumulación de los errores propagados. Esta parte de la demostración es exactamente igual que para los métodos a un paso, ver el

paragrafo VII.3 y la figura VII.4.5. □

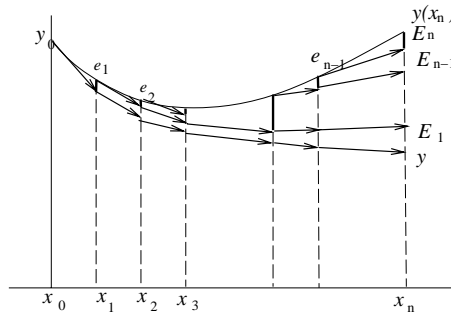


Figura VII.3.2. Estimación del error global para métodos multipaso.

Ejercicios

- 1.- Para los coeficientes del método de Adams explícito, demostrar la fórmula de recurrencia siguiente:

$$\gamma_m + \frac{1}{2}\gamma_{m-1} + \frac{1}{3}\gamma_{m-2} + \cdots + \frac{1}{m+1}\gamma_0 = 1.$$

Indicación.-Introducir la serie

$$G(t) = \sum_{j=0}^{\infty} \gamma_j t^j, \quad \gamma_j = (-1)^j \int_0^1 \binom{-s}{j} ds$$

y demostrar que

$$G(t) = \int_0^1 (1-t)^{-s} ds \quad \text{y} \quad -\frac{\log(1-t)}{t} G(t) = \frac{1}{1-t},$$

enseguida, desarrollar la segunda fórmula en serie de Taylor y comparar los coeficientes de t^m .

- 2.- Considérese la identidad

$$y(x_{n+1}) = y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(t, y(t)) dt$$

para la solución de la ecuación diferencial $y' = f(x, y)$.

a) Remplazar en la identidad la función desconocida $f(t, y(t))$ por el polinomio $p(t)$, como se ha hecho para construir el método de Adams explícito. Deducir la fórmula de Nyström

$$y_{n+1} = y_{n-1} + h \sum_{j=0}^{k-1} \kappa_j \nabla^j f_n.$$

b) Calcular los primeros κ_j .

c) Verificar la identidad $\kappa_j = 2\gamma_j - \gamma_{j-1}$, donde γ_j son los coeficientes del método de Adams explícito.

3.- Mostrar que el método BDF a k pasos tiene orden k .

4.- Calcular el orden para la fórmula de Nyström, ver ejercicio 2.

5.- Un método multipaso se dice simétrico, si

$$\alpha_j = -\alpha_{k-j}, \quad \beta_j = \beta_{k-j}.$$

Mostrar que un método simétrico siempre tiene un orden par.

6.- Dado un predictor de orden $p-1$

$$\sum_{i=0}^k \alpha_i^P y_{n+1} = h \sum_{i=0}^{k-1} \beta_i^P f_{n+1}, \quad (P)$$

y un corrector de orden p

$$\sum_{i=0}^k \alpha_i y_{n+1} = h \sum_{i=0}^k \beta_i f_{n+1}. \quad (C)$$

a) Escribir el método $P(EC)^M E$ bajo la forma

$$Y_{n+1} = AY_n + h\Phi(x_n, Y_n, h).$$

b) Mostrar que este método es convergente de orden p , si el corrector es estable, no es necesario que el predictor sea estable.

Bibliografía

Al final de cada libro o artículo, se indica entre corchetes y caracteres itálicos el capítulo y/o sección al que hace referencia.

J.H Ahlberg, E.N. Nilson & J.L. Walsh (1967): *The Theory of Splines and Their Applications*. Academic Press, New York. [III.2]

G. Arfken (1985): *Mathematical Methods for Physicists*. Academic Press, London. [I.3], [VI.2].

V.I. Arnol'd (1992): *Ordinary Differential Equations*. Springer-Textbook. [VII.1].

K.E. Atkinson (1978): *An Introduction to Numerical Analysis*. John Wiley & Sons. [I], [III], [VI].

O.A Axelsson (1976): *A class of iterative methods for finite element equations*. Comp. Math. in Appl. Mech. and Eng. 9. [II.4].

N. Bakhbalov (1976): *Méthodes Numériques*. Editions Mir, Moscou. [I], [III], [VI].

C. de Boor (1978): *A Practical Guide to Splines*. Springer-Verlag, Berlin. [III.2].

C. Brézinski (1977): *Accélération de la Convergence en Analyse Numérique*. Lectures Notes in Mathematics, Nr. 584, Springer-Verlag. [III.3], [VI.3].

J.C Butcher (1987): *The Numerical Analysis of Ordinary Differential Equations*. John Wiley & Sons. [VII].

H. Cartan (1977): *Cours de Calcul Différentiel*. Hermann, Paris. [IV.2], [IV.3].

P.G Ciarlet (1982): *Introduction à l'analyse numérique matricielle et à l'optimisation*. Mason, Paris. [II].

M. Crouzeix & A.L. Mignot (1984): *Analyse Numérique des Equations Differentielles*. Mason, Paris. [III], [VI], [VII].

G. Dahlquist & A. Björck (1974): *Numerical Methods*. Prentice-Hall. [I], [III], [VI].

R. Dautray & J.L. Lions (1988): *Mathematical Analysis and Numerical Methods for Science and Technology, volume 2, Functional and Variational Methods*. Springer-Verlag. [VI.4].

P.J. Davis & P. Rabinowitz (1975): *Methods of Numerical Integration*. Academic Press, New York. [VI].

- J. Dieudonné (1980): *Calcul Infinitesimal*. Hermann, Paris. [IV.1], [IV.2], [VI.4].
- J.J. Dongarra, C.B. Moler, J.R. Bunch & G.W. Stewart (1979): *LINPACK Users' Guide*. SIAM, Philadelphia. [II].
- A. Gramain (1988): *Integration*. Hermann, Paris. [VI.1], [VI.4].
- R.P. Grimaldi (1989): *Discrete and Combinatorial Mathematics*. Addison-Wesley Publishing Company. [II.3].
- E. Hairer, S.P. Nørsett & G. Wanner (1987): *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer Series in Comput. Math., vol. 8. it [VII].
- E. Hairer & G. Wanner. *Solving Ordinary Differential Equations. Stiff Problems*. Springer Series in Comput. Math., vol. 11. [VII].
- P. Henrici (1962): *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley & Sons. [VII].
- R.A. Horn & C.R. Johnson (1985): *Matrix Analysis*. Cambridge University Press. [II.1], [V.1].
- G.H. Golub & C.F. Van Loan (1989): *Matrix Computations*. The Johns Hopkins University Press, Baltimore and London. [II].
- K.L. Kantorovitch & G. Akilov (1981): *Analyse Fonctionnelle, Volume 2*. Editions Mir, Moscou. [IV.2], [IV.3].
- V.I. Krylov (1962): *Approximate calculation of integrals*. Macmillan, New York. [VI].
- S. Lang (1987): *Linear Algebra*. Springer-Verlag. [II], [V].
- M. Marden (1966): *Geometry of Polynomials*. American Mathematical Society, Providence, Rhode Island, 2nd Edition. [IV.1].
- J.M. Ortega & W.C. Rheinboldt (1970): *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York. [IV].
- A.M. Ostrowski (1966): *Solution of Equations and Systems of Equations*. Academic Press, New York, 2nd edition. [IV].
- B.N. Parlett (1980): *The Symmetric Eigenvalue Problem*. Prentice-Hall, Englewood Cliffs, New Jersey. [V].
- R. Piessens, E. de Doncker-Kapenga, C.W. Überhuber & D.K. Kahaner (1983): *QUAD-PACK. A Subroutine Package for Automatic Integration*. Springer Series in Comput. Math., vol. 1. [VI].
- W.H. Press, B.R. Flannery, S.A. Teukolsky & W.T. Vetterling (1989): *Numerical Recipes. The Art of Scientific Computing* (Versión FORTRAN). Cambridge University Press, Cambridge. [VI.4].
- C.R. Rao: (1973): *Linear Statistical Inference and Its Applications*. John Wiley & Sons. [II.5].

W. Rudin (1976): *Principles of Mathematical Analysis*. MacGraw Hill, third edition. [IV.2], [VI.1].

A.A. Samarski & E.S. Nikolaev (1982): Métodos de solución de las ecuaciones reticulares, Tomo I y II. Editorial Mir, Moscou. [III.3], [III.4]

B.T. Smith, J.M. Boyle, Y. Ikebe, V.C. Klema & C.B. Moler (1970): *Matrix Eigensystem Routines: EISPACK Guide*. 2nd ed., Springer-Verlag, New York. [V]

J. Stoer & R. Bulirsch (1980): *Introduction to Numerical Analysis*. Springer-Verlag, New York. [I], [III], [VI].

G.H. Stewart (1973): *Introduction to Matrix Computations*. Academic Press, New York. [II].

L. Schwartz (1970): *Analyse, Topologie générale et analyse fonctionnelle*. Hermann, Paris. [IV.2].

R.A. DeVore & G.G. Lorentz (1991): *Constructive Approximation*. Springer-Verlag. [III.2].

J.H. Wilkinson (1963): *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford. [V].

J.H. Wilkinson (1965): *Rounding Errors in Algebraic Process*. Prentice-Hall, New York. [I].

J.H. Wilkinson (1969): *Rundungsfehler*. Springer-Verlag, Berlin. [I], [II].

J.H. Wilkinson & C. Reinsch (1971): *Handbook for Automatic Computation, Volume II, Linear Algebra*. Springer-Verlag, New York. [II], [V].

FORTRAN/9000 Reference (1991): Hewlett-Packard Company. [I.2].

Indice de Símbolos

Los símbolos matemáticos que aparecen en el libro, están enunciados en este glosario de la manera siguiente. En la primera columna el símbolo mismo, al medio el nombre de éste y a la derecha la pagina donde está definido.

$\mathcal{P}(x)$	problema, 2.
$arr(x)$	redondeo de x , 8.
eps	precisión de la computadora, 8.
$\ \ $	norma de vector, 25.
$\ \ $	norma de matriz, 26.
$\mathbb{1}$	vector 1, 29.
$cond(A)$	condición de la matriz A, 29.
LR	descomposición de Gauss, 37.
$diag(r_1, r_2, \dots, r_n)$	matriz diagonal, 45.
$D^{1/2}$	matriz raíz cuadrada, 45.
ω	coeficiente de relajación, 56.
\otimes	producto tensorial de matrices, 60.
$\ x\ _A$	norma natural, 72.
Q	matriz ortogonal, 87.
A^+	pseudo inversa de una matriz, 93.
$y[x_{i_0}, \dots, x_{i_k}]$	diferencia dividida de orden k, 107.
$\nabla^k y_i$	diferencia finita progresiva, 109.
$\bar{\nabla}^k y_i$	diferencia finita regresiva, 109.
$T_n(x)$	polinomio de Chebichef, 113.
$l_i(x)$	polinomio de Lagrange, 116.
Λ_n	constante de Lebesgue, 116.
f'_x	aplicación derivada, 168.
$\rho(A)$	radio espectral, 171.
\ker	nucleo de una aplicación lineal, 214.
$\chi_A(\lambda)$	polinomio característico, 214.
A^*	matriz adjunta, 215.
U	matriz unitaria, 215.
$P_k(\tau)$	nucleo de Peano de orden k, 251.
$P_k(x)$	polinomio de Legendre, 262.
\mathcal{F}_N	transformada discreta de Fourier, 286.
$y * k$	producto de convolución, 292.

Indice Alfabético

- acelerador de convergencia, 147, 276
- Adams
 - método explícito, 341
 - método implícito, 343
- Aitken, 276
- algoritmo, 2
 - bisección, 161, 164
 - Cholesky, 43
 - epsilon, 278
 - Euclides, 157
 - Gauss, 37
 - Hörner, 3, 7
- aplicación de *spline*, 142
- aproximación de Broyden, 198
- armónica, sucesión, 148
- axiomas de Moore-Penrose, 96

- BDF, métodos, 345
- Broyden, 198
- Brusselator, 339
- Bulirsch, sucesión, 148
- busqueda
 - de Fibonacci, 129
 - de pivote parcial, 41

- cálculo derivada, 177
- cálculo variacional, 310
- Cardano, 154
- Cauchy, 297
- cero
 - localización, 155
 - multiplicidad, 105
- Chebichef
 - polinomio, 73, 113
 - puntos, 116
 - teorema, 74
- cociente de Rayleigh, 223, 227
- constante
 - de Lebesgue, 116

- condición
 - del problema a valores propios, 217
 - del problema lineal, 30, 33
 - de un problema, 10
 - de una matriz, 29, 30
 - Lipschitz, 297, 298
- construcción
 - polinomio de interpolación, 106
 - método de orden 4, 327
 - spline*, 133
- convergencia
 - acelerador, 147
 - interpolación, 119
 - método de Gauss-Newton, 204
 - método de Gauss-Seidel, 49
 - método de Jacobi, 49
 - métodos de Runge-Kutta, 335
 - métodos multipaso, 350
 - spline*, 133
- costo, 3
 - descomposición *LR*, 39
 - descomposición *QR*, 90
 - transformada rápida de Fourier, 291
- Cramer, regla de, 35

- Dahlsquist, primera barrera, 350
- descomposición
 - Cholesky, 44, 45
 - Jordan, 219
 - LR*, 37, 38, 39
 - QR*, 87
 - Schur, 215, 234
 - valores singulares, 94
- diferencias
 - divididas, 107
 - finitas, 109
- dirección
 - conjugada, 69, 70
 - Newton, 194
- dispositivo
 - ideal de cálculo, 2
 - material de cálculo, 8
- Dormand & Prince, 332

- ecuaciones
 - con derivadas parciales, 59
 - cuadráticas, 152
 - cúbicas, 153
 - de grado cuarto, 154
 - diferenciales, 296
 - Euler-Lagrange, 310
 - no resolubles por radicales, 155
 - resolubles por radicales, 152
- eficiencia, 3
- ϵ -algoritmo, 278
- error
 - de aproximación, 5
 - de la aproximación *spline*, 136
 - de interpolación, 111
 - del método, 5
 - de truncación, 4
 - fórmula de cuadratura, 250
 - extrapolación, 149
 - método gradiente, 68
 - método gradiente condicionado, 72
 - método Gauss-Newton, 205,
 - método Multipaso, 346
 - método Newton, 176
 - transformada discreta de Fourier, 287
- errores de redondeo, 115
- estabilidad
 - algoritmo de Gauss, 41,42,43
 - backward analysis*, 13
 - forward analysis*, 12
 - método de Euler, 319
 - método de Euler explícito, 322
 - método multipaso, 348
 - numéricamente estable, 12
- Euler
 - efecto de los errores de redondeo, 317
 - estabilidad, 319
 - método, 313
 - método implícito, 321
 - polígono, 313
- extrapolación
 - polinomial, 145
 - tablero, 147
 - error, 149
- factorización,
 - incompleta de Cholesky, 77
- fenómeno de Runge, 124
- FFT, 290
- Fibonacci, búsqueda, 129
- formas tridiagonales, 227
- fórmula
 - de Cardano, 154
 - de Newton, 107
 - de Nyström, 354
 - de Rodríguez, 262
- fórmula de cuadratura, 248
 - error, 250
 - Gauss, 264–266
 - orden, 249
 - orden elevado, 259
 - simétrica, 249
- Fourier
 - error, 287
 - serie de, 284
 - transformada, 284
 - transformada discreta, 286
 - transformada rápida, 290
- funcion
 - integrable, 244
 - modelo, 83
 - ortogonal, 260
 - peso, 259
 - funciones con oscilaciones, 281
- Galois, teorema, 155
- GAUINT, 272
- Gauss,
 - algoritmo de eliminación, 36, 37
 - fórmulas de cuadratura, 264–266
 - convergencia Gauss-Newton, 204
 - método Gauss-Newton, 203
- Gerschgorin, teorema, 156, 232
- grado de aproximación, 120
- grafo dirigido, 49
- Hermite, 104, 138
- Heun, 325

- Interpolación
 - convergencia, 119
 - de Lagrange, 104–108
 - de Hermite, 104
 - trigonométrica, 289
- Kantorovich, ver teorema
- Konrad, 271
- Lagrange, 104
- λ -estrategia, 194
- Lebesgue, 116
- Levenberg-Marquandt, 210
- Lipschitz, 297
- mantisa, 8
- matriz
 - adjunta, 215
 - definida positiva, 43
 - Frobenius, 155, 221
 - Hessenberg, 227
 - Hilbert, 31, 87
 - Householder, 88, 228
 - irreducible, 49, 50
 - no-negativa, 50
 - normal, 216
 - ortogonal, 30
 - pseudo-inversa de una, 92
 - simétrica, 43
 - tridiagonal, 227
 - Toeplitz, 292
 - unitaria, 215
 - Vandermonde, 31
- método,
 - Adamas explícitos, 341
 - Adamas implícitos, 343
 - a un paso, 323
 - BDF, 345
 - de la bisección, 229
 - de la potencia, 223
 - de la potencia generalizada, 233
 - de la potencia inversa, 225
 - Dormand & Prince, 332
 - encajonados, 330
 - Euler, 313, 325
 - falsa posición, 165
 - Gauss Newton, 203–205
 - Gauss Seidel, 48–56
 - Gradiente, 67
 - Gradiente Conjugado, 69
 - Gradiente Conjugado
 - Precondicionado, 75
 - Heun, 325
 - iterativos, 163–172
 - iterativo simple, 169
 - Jacobi, 48–56
 - Levenberg-Marquandt, 210
 - Maehly, 158
 - Multipaso, 341
 - Newton, 157, 174–199
 - Newton con Relajación, 193
 - Newton Simplificado, 184
 - predictor-corrector, 344
 - QR , 237
 - QR con *shift*, 238
 - Sobrerelajación SOR, 56, 58
 - Runge-Kutta, ver Runge-Kutta
 - SSOR precondicionado, 78
- mínimos cuadrados, 83
 - interpretación estadística, 84
 - interpretación geométrica, 85
- Misovski,
 - ver teorema Newton-Misovski
- modificaciones de Gauss-Newton, 207
- módulo de continuidad, 121
- Newton
 - dirección, 194
 - Kantorovich, 185
 - método, 157, 174–199
 - método con Relajación, 193
 - método simplificado, 184
 - Misovski, 179
 - regla de, 247
- norma,
 - absoluta, 26
 - ciudad-bloque, 25
 - de la convergencia uniforme, 25
 - de una matriz, 26
 - de un vector, 25
 - euclidiana, 25
 - Frobenius, 33
 - monótona, 26
 - natural, 72
- Núcleo de Peano, 251–254
- núcleo resolvente, 303
- operación elemental, 2
- orden
 - fórmula de cuadratura, 249
 - método de Runge-Kutta, 325
 - construcción de
 - un método de orden 4, 327

- Peano, 251–254
- Pearson, 99
- Perron-Frobenius, 52
- pivote, 40
- polígono de Euler, 313
- polinomio
 - característico, 214
 - Chebichef, 73, 113, 262
 - Hermite, 104
 - interpolación de Hermite, 104, 138, 262
 - interpolación de Lagrange, 104
 - Jacobi, 262
 - ortonormales, 260
 - Lagrange, 104, 105
 - Laguerre, 262
 - Legendre, 262, 263
- precisión de la computadora, 8
- primitiva, 245
- problema, 2
 - a valor inicial, 296
 - bien condicionado, 10
 - con valores en la frontera, 300
 - de Cauchy, 296
 - de minimización, 66
 - mal condicionado, 10
- potencia inversa de Wielandt, 225
- procedimiento Δ^2 de Aitken, 276
- producto
 - de Kronecker, 65
 - escalar de funciones, 260
 - sesquilineal, 284
 - tensorial de matrices, 60
- Property A* de Young, 55
- punto
 - flotante, 8
 - de Chebichef, 117
 - punto fijo, 169
- pseudo-inversa de una matriz, 92
- QUADPACK, 271
- Rayleigh, 223, 227
- redondeado, 8
- región de estabilidad, 320
- regla
 - del punto medio, 246
 - del trapecio, 246
 - de Newton, 247
 - de Simpson, 247
- Cramer, 35
- resolvente, 303
- Rodriguez, 262
- Romberg, 148
- Runge-Kutta
 - condiciones de orden, 327, 328
 - convergencia, 335
 - error global, 335
 - error local, 325
 - Dormand-Prince, 332
 - esquema, 325
 - Euler, 325
 - Heun, 325
 - Kutta, 330
 - métodos, 324
 - métodos encajonados, 330
 - orden, 325
 - regla 3/8, 330
 - Runge, 325
 - soluciones continuas, 333
- Schur, 215, 234
- serie de Fourier, 284
- Shooting*
 - múltiple, 307
 - simple, 303
- soluciones continuas, 333
- SOR, 56
- Spline*
 - aplicación, 142
 - construcción, 133
 - cúbico, 131, 132
 - error, 136
 - fijo en los bordes, 132
 - natural, 132
- SSOR preconditionado, 78
- radio espectral, 58

- ω optimal, 58
- Sturm, 159
- sucesiones
 - armónica, 148
 - Bulirsch, 148
 - Romberg, 148
 - Sturm, 159
- Sylvester, 232
- tablero de extrapolación, 147
- teorema
 - Cauchy-Lipschitz, 297, 315
 - Chebichef, 74
 - Dirichlet, 285
 - del Muestreo, 289
 - Fundamental del Algebra, 152
 - Galois, 155
 - Gerschgorin, 156, 232
 - Jordan, 219
 - Newton-Kantorovich, 185
 - Newton-Misovski, 179
 - Pearson, 99
 - Perron-Frobenius, 52
 - punto fijo, 169
 - Schur, 215
 - Sylvester, 232
 - Weistrass, 124
 - Wilkinson, 42
 - Wynn, 278
- transformada
 - discreta de Fourier, 286
 - rápida de Fourier, 290
- valor absoluto de un vector, 26
- valor propio, 215
- valores singulares, 94
- Van der Pol, 305
- vector propio, 215
- Weistrass, 124
- Wilkinson, 42
- Wynn, 278

Una Introducción al Análisis Numérico es un libro que pretende dar un enfoque moderno de los tópicos introductorios de esta disciplina. Las nociones de estabilidad y error son analizadas minuciosamente en cada tema. La formulación de métodos y algoritmos es tratada de una manera construccionista, evitando de esta manera las recetas y trucos que aparecen en otros libros.

El libro cuenta con siete capítulos que dan una idea de lo que constituye actualmente el Análisis Numérico. Estos son: Preliminares, Sistemas Lineales, Interpolación, Ecuaciones No Lineales, Cálculo de Valores Propios, Integración Numérica y Ecuaciones Diferenciales.

Por sus características, este libro puede utilizarse como texto base, o bien como un complemento bibliográfico. Está destinado a alumnos o profesionales interesados en el Análisis Numérico. Como prerequisite para una buena utilización de este libro, se requiere tener los conocimientos básicos de análisis y álgebra lineal