CURSO DE MÉTODOS NUMÉRICOS

Virginia Muto Foresi

1.360094193
1.367846968
1.363887004
1.365916733
1.364878217
1.365410061
1.365223680
1.365230236

EHU

CURSO DE MÉTODOS NUMÉRICOS

CURSO DE MÉTODOS NUMÉRICOS

impartido por

Virginia Muto Foresi

Departamento de Matemática Aplicada
y Estadística e Investigación Operativa

Facultad de Ciencias / Zientzi Fakultatea
Universidad del País Vasco / Euskal Herriko Unibertsitatea



CIP. Biblioteca Universitaria

Muto Foresi, Virginia

Curso de métodos numéricos [Recurso electrónico] / impartido por Virginia Muto Foresi. – Datos. – [Leioa] : Servicio Editorial de la Universidad del País Vasco = Euskal Herriko Unibertsitateko Argitalpen Zerbitzua, [2026]. – 1 recurso en línea: PDF (345 p.)
Ed. electrónica de la ed. impresa (1998)
Modo de acceso: World Wide Web.
ISBN: 978-84-1319-839-2

1. Análisis numérico.

(0.034)519.6

Debekatuta dago liburu hau osorik edo zatika kopiatzea, bai eta berorri tratamendu informatikoa ematea edota liburua ezein modutan transmititzea, dela bide elektronikoz, mekanikoz, fotokopiaz, erregistroz edo beste edozein eratara, baldin eta *copyrightaren* jabeek ez badute horretarako baimena aurretik eta idatziz eman.

Ninguna parte de esta publicación, incluido el diseño de la cubierta, puede ser reproducida, almacenada o transmitida en manera alguna ni por ningún medio, ya sea eléctrico, químico, mecánico, óptico, de grabación o de fotocopiado, sin permiso previo y por escrito de la entidad editora, sus autores o representantes legales.

© Servicio Editorial de la Universidad del País Vasco
Euskal Herriko Unibertsitateko Argitalpen Zerbitzua

Argitalpen Zerbitzua • Servicio Editorial
editorial@ehu.eus - 94 601 2227
www.ehu.eus/argitalpenak
Edificio Biblioteca, 1.ª planta
B.º Sarriena s/n. Campus de Bizkaia

ISBN: 978-84-1319-839-2

CURSO DE METODOS NUMERICOS — INDICE

PRIMERA PARTE: INTRODUCCION AL ANALISIS NUMERICO Y A LA COMPUTACION

Capítulo I. Introducción al Análisis Numérico.

1. Algoritmos y diagramas de flujo. pg. 3
2. Origen y evolución del Análisis Numérico. pg. 14
3. Objetivos. pg. 15

Ejercicios. pg. 16

Capítulo II. Análisis de los errores.

1. Esquema de resolución numérica de un problema. pg. 17
2. Distintos tipos de errores. pg. 19
3. Convergencia. pg. 21

Ejercicios. pg. 24

Capítulo III. Sistemas de numeración.

1. Representación de la información. pg. 25
2. Introducción a los sistemas numéricos. pg. 25
3. Conversión desde el sistema decimal al sistema numérico en base b. pg. 26
4. Las operaciones aritméticas en base b. pg. 32
5. Conversión desde un sistema numérico en base b_1 a un sistema en base b_2. pg. 35

Ejercicios. pg. 38

Capítulo IV. Aritmética del computador.

1. Representación de los números. pg. 39
2. Introducción a la aritmética de punto flotante. pg. 46
3. Propagación del error. pg. 47

Ejercicios. pg. 58

SEGUNDA PARTE: SOLUCION APROXIMADA DE ECUACIONES DE UNA VARIABLE

Capítulo V. Solución aproximada de ecuaciones de una variable: Preliminares.

1. Separación de raíces. pg. 61
2. Solución gráfica de ecuaciones. pg. 64

Capítulo VI. El algoritmo de bisección.

1. Introducción y método. pg. 67

2. Algoritmo y ejemplos. pg. 68
Ejercicios. pg. 72

Capítulo VII. Iteración del punto fijo.

1. Introducción y método. pg. 73
2. Algoritmo y ejemplos. pg. 76
Ejercicios. pg. 86

Capítulo VIII. El método de la secante.

1. Introducción y método. pg. 87
2. Algoritmo y ejemplos. pg. 91
Ejercicios. pg. 94

Capítulo IX. El método de Newton-Raphson.

1. Introducción y método. pg. 95
2. El algoritmo de Newton-Raphson. pg. 101
3. El algoritmo de la secante modificado. pg. 102
4. El método de Newton modificado. pg. 104
5. El método de combinación. pg. 104
Ejercicios. pg. 108

Capítulo X. Análisis de error y técnicas de aceleración.

1. Análisis de los errores para métodos iterativos. pg. 109
2. Técnicas de aceleración y fórmula de Newton generalizada. pg. 111
3. Convergencia acelerada y el algoritmo Δ^2 de Aitken. pg. 115
4. Convergencia acelerada y el algoritmo de Steffersen. pg. 119
Ejercicios. pg. 122

Capítulo XI. Métodos de interpolación.

1. El método de interpolación de la posición falsa. pg. 123
2. El método de interpolación de Müller. pg. 125
Ejercicios. pg. 128

Capítulo XII. Ceros de polinomios.

1. El método de Horner. pg. 129
2. La técnica de deflación. pg. 135
3. El método de Bairstow. pg. 138
4. El método de Bernoulli. pg. 142
5. El algoritmo del cociente-diferencia. pg. 151
Ejercicios. pg. 160

TERCERA PARTE: **METODOS PARA LA RESOLUCION DE SISTEMAS LINEALES**

Capítulo XIII. Métodos para la resolución de sistemas lineales: Preliminares.

1. Sistemas lineales de ecuaciones. pg. 163
2. Algebra lineal e inversión de una matriz. pg. 166
3. Tipos especiales de matrices. pg. 173
4. Normas de vectores y matrices. pg. 177

Capítulo XIV. Eliminación Gaussiana y sustitución hacia atrás.

1. Introducción y método. pg. 185
2. Algoritmo y ejemplos. pg. 188

Ejercicios. pg. 192

Capítulo XV. Estrategias de pivoteo.

1. Introducción y método. pg. 193
2. Algoritmos de eliminación Gaussiana con pivoteo. pg. 194
3. Ejemplo de algoritmo FORTRAN. pg. 199
4. El algoritmo de Gauss-Jordan. pg. 206

Ejercicios. pg. 213

Capítulo XVI. Factorización directa de matrices.

1. Introducción y método. pg. 215
2. Los algoritmos de Doolittle y de Crout. pg. 217
3. El algoritmo de Cholesky. pg. 226
4. El algoritmo de Crout para sistemas tridiagonales. pg. 228

Ejercicios. pg. 233

Capítulo XVII. Técnicas iterativas para resolver sistemas linales.

1. Introducción y método. pg. 235
2. Los algoritmos de Jacobi y de Gauss-Seidel. pg. 237
3. Convergencia de los procesos iterativos. pg. 240
4. Los métodos de relajación. pg. 246
5. Elección del método para resolver sistemas lineales. pg. 253

Ejercicios. pg. 254

Capítulo XVIII. Estimaciones de error y refinamiento iterativo.

1. Estimaciones de error. pg. 255
2. Refinamiento iterativo. pg. 259

Ejercicios. pg. 262

CUARTA PARTE: **METODOS DE MINIMOS CUADRADOS**

Capítulo XIX. El problema de los mínimos cuadrados.

1. Sistemas lineales de ecuaciones sobredeterminados. pg. 265
2. El vector residual y el problema de los mínimos cuadrados. pg. 266
3. Las ecuaciones normales. pg. 271
4. Aplicaciones. pg. 273

Ejercicios. pg. 276

Capítulo XX. Los métodos de transformación ortogonal.

1. Las transformaciones de Householder. pg. 277
2. La factorización QR con matrices de Householder. pg. 280
3. La factorización QR con matrices de rotación de Givens. pg. 284

Ejercicios. pg. 290

QUINTA PARTE: **SOLUCIONES NUMERICAS A SISTEMAS NO LINEALES**

Capítulo XXI. Puntos fijos para funciones de varias variables.

1. Preliminares. pg. 293
2. Método de iteración y ejemplos. pg. 294
3. Condiciones para la convergencia del proceso de iteración. pg. 297

Ejercicios. pg. 305

Capítulo XXII. Método de Newton.

1. Introducción y método. pg. 307
2. Algoritmo y ejemplos. pg. 313

Ejercicios. pg. 316

Capítulo XXIII. Métodos Cuasi-Newton.

1. El método de Newton modificado. pg. 317
2. El método de Broyden. pg. 317

Ejercicios. pg. 322

Capítulo XXIV. Técnicas de descenso más rápido.

1. Introducción y método. pg. 323
2. Algoritmo y ejemplos. pg. 325

Ejercicios. pg. 328

SEXTA PARTE: **BIBLIOGRAFIA**

Bibliografía básica. pg. 331

Bibliografía complementaria. pg. 331

— PREFACIO —

Este libro es el resultado de la recopilación de una parte de los apuntes del curso de "*Cálculo Numérico*" correspondiente al tercer curso de la licenciatura en Ciencias Matemáticas. Inicialmente estos apuntes fueron pensados para un curso del Plan Antiguo de Estudios (1973-76), anual, por lo que no cabe duda que son perfectamente válidos para la asignatura cuatrimestral "*Métodos Numéricos*" que ahora se puede encontrar en el actual Plan Nuevo de Estudios (1995-96).

Cuando inicié la redacción de estos capítulos, en Enero de 1991, a partir de distintos textos, me propuse como objetivo transmitir con claridad en el concepto y facilidad en la comprensión lo que yo considero que es la herramienta fundamental en el campo de la Matemática Aplicada, al que dedico mis esfuerzos en la investigación científica.

Conceptos como error de cálculo, análisis del error, algoritmo y su coste, implementación de un código en un ordenador usando un lenguaje de programación, ... , van a ser de uso habitual en la actividad laboral que va a poder desarrollar cualquier licenciado en Matemáticas (o de cualquier otra especialidad técnica superior) tras su paso por la Universidad.

La buena acogida que estos apuntes han tenido en el colectivo estudiantil me ha animado a dar un paso más hacia adelante, con el intento de su publicación. Me motiva ahora que más gente que lo que pueda ser el limitado colectivo de alumnos de un curso tenga acceso a esta información, que pueda ahora ser útil a gente con distintos orígenes en cuanto a los problemas que este libro pueda ayudarles a resolver.

Espero conseguirlo, y espero que este libro que tiene el lector en sus manos ayude a ello (¡ y de paso al propio lector a resolver sus problemas de Cálculo Numérico!)

A Ane Miren
y a Mikel.
A Jon, por todo.

CURSO DE MÉTODOS NUMÉRICOS

PRIMERA PARTE

INTRODUCCIÓN AL ANÁLISIS NUMÉRICO Y A LA COMPUTACIÓN

CAPITULO I. INTRODUCCION AL ANALISIS NUMERICO

1. ALGORITMOS Y DIAGRAMAS DE FLUJO

El hecho de que el Análisis Numérico sea tanto una ciencia como un arte es la opinión de los especialistas en este campo pero, frecuentemente es mal entendido por los no especialistas. ¿Se dice que es un arte, y a la vez una ciencia, únicamente como eufemismo, para ocultar el hecho de que el Análisis Numérico no es una disciplina suficientemente precisa para merecer el que se le considere como una ciencia? ¿Es cierto que el nombre de *análisis numérico* se emplea erróneamente, porque el significado clásico del análisis en matemáticas no es aplicable al trabajo numérico? De hecho, la respuesta a ambas preguntas es "no". Más bien la yuxtaposición de ciencia y arte se debe a un principio de incertidumbre que frecuentemente se presenta en la solución de problemas, es decir, el hecho de que para determinar la mejor forma de resolver un problema, puede ser necesaria la solución del problema en sí. En otros casos, la mejor forma de resolver un problema puede depender de un conocimiento de las propiedades de las funciones involucradas, las que no se pueden obtener ni teórica ni prácticamente.

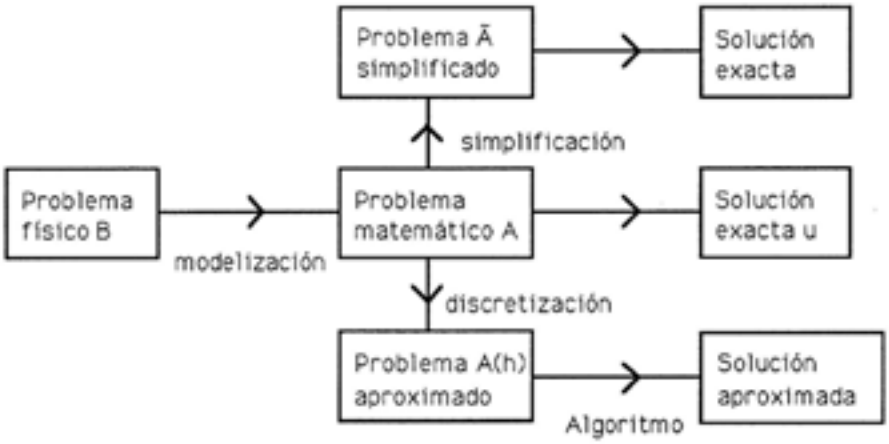
Como una ciencia, el Análisis Numérico está interesado en los procesos por los cuales pueden resolverse los problemas matemáticos, por las operaciones de la aritmética. Algunas veces esto involucrará el desarrollo de algoritmos para resolver un problema que está ya en una forma en la cual pueda encontrarse la solución por medio aritméticos. Frecuentemente involucrará la necesidad de sustituir cantidades que no pueden ser calculadas aritméticamente, por aproximaciones que permiten que sea determinada una solución aproximada. En este caso estaríamos interesados, naturalmente, en los errores cometidos en nuestra aproximación. Pero, en cualquier caso, las herramientas que usaríamos en el desarrollo de los procesos de análisis numérico, serán las herramientas del análisis matemático exacto, tan conocidas clásicamente.

Como un arte, el Análisis Numérico está interesado en la elección del procedimiento, y conveniente aplicación del mismo, "más" adecuado a la solución de un problema particular. Esto implica la necesidad de desarrollar la experiencia y con ello esperar que se desarrolle la intuición del especialista.

Así pues, el Análisis Numérico trata de diseñar métodos para aproximar, de una manera eficiente, las soluciones de problemas expresados matemáticamente. La eficiencia del método depende tanto de la precisión que se requiera como de la facilidad con la que pueda implementarse. En una situación práctica, el problema matemático se deriva de un fenómeno físico sobre el cual se han hecho algunas suposiciones para simplificarlo y para poderlo representar matemáticamente. Generalmente cuando se relajan las suposiciones físicas llegamos a un modelo matemático más apropiado pero, al mismo tiempo, más difícil o imposible de resolver explícitamente.

Ya que normalmente el problema matemático no resuelve el problema físico exactamente, resulta con frecuencia más apropiado encontrar una solución aproximada del modelo matemático más complicado que encontrar una solución exacta del modelo sim-

plificado. Para obtener tal aproximación se idea un método llamado **algoritmo**. El algoritmo consiste en una secuencia de operaciones algebráicas y lógicas que permiten la aproximación al problema matemático y se espera que también al problema físico, con una tolerancia o precisión predeterminada. Los algoritmos determinan los métodos constructivos de resolución de problema matemáticos. Un **método constructivo** es todo proceso que permite obtener la solución a un problema con la precisión que se desee, en un número finito de pasos que se pueden efectuar racionalmente. Obviamente el número de pasos requeridos dependerá de la precisión que se desee en la solución.



Como se ha dicho, los métodos constructivos en matemáticas son métodos que muestran cómo construir soluciones de un problema matemático. Por ejemplo, una demostración constructiva de la existencia de una solución de un problema, no sólo hace ver que la solución existe, si no que describe también cómo se puede determinar esa solución. Una demostración que muestra la existencia de una solución por *reducción al absurdo* no es constructiva.

Ejemplo 1. Demostrar que la ecuación cuadrática

$$x^2 + 2\ b\ x + c = 0 \tag{$I.1$}$$

con coeficientes reales b y c tales que $b^2 > c$ tiene por lo menos dos raíces reales.

Demostración constructiva. Para todo x, b y c

$$\begin{aligned} x^2 + 2\ b\ x + c = x^2 + 2\ b\ x + b^2 + c - b^2 = \\ = (x+b)^2 + c - b^2\ . \end{aligned}$$

Entonces, x es una raíz de $(I.1)$ si y sólo si

$$(x+b)^2 + c - b^2 = 0\ ,$$

es decir,

$$(x+b)^2 = b^2 - c\ .$$

Y tomando la raíz cuadrada, como $b^2 - c > 0$, x será una raíz si y sólo si

$$x + b = \pm\sqrt{b^2 - c}\ ,$$

es decir,

$$x = -b \pm \sqrt{b^2 - c}\ . \tag{I.2}$$

Luego hay dos raíces reales de $(I.1)$ y la fórmula $(I.2)$ muestra como calcularlas.

Demostración no constructiva. Sea

$$q(x) = x^2 + 2\ b\ x + c\ .$$

Supongamos que no existen raíces de $(I.1)$, con lo que no hay ceros de q. Dado que $q(x)$ es una función continua, $q(x)$ será siempre positiva o siempre negativa para cada valor de $x \in \mathcal{R}$. Ahora

$$q(-b) = b^2 - 2b^2 + c = c - b^2 < 0$$

por la condición dada sobre b y c. Entonces q tiene que ser siempre negativa. Sin embargo, para $|x|$ grandes

$$x^2 > |2\ b\ x + c|\ ,$$

con lo que resulta ser $q(x) > 0$. Así que tenemos una contradicción.

Supongamos ahora que q tiene sólo un cero. Dada la continuidad de q se sigue que $q(x) > 0$ para valores x grandes y $q(x) < 0$ para valores $-x$ grandes, o viceversa. Sin embargo, hemos visto que $q(x) > 0$ para valores $|x|$ grandes, cualquiera que sea el signo de x. Tenemos, pues, otra contradicción.

Así, esta demostración dice que existen dos raíces, pero no muestra cómo calcularlas.

Los algoritmos tienen que satisfacer los siguientes requisitos:

a) *generalidad*: un algoritmo se tiene que poder aplicar a cualquier conjunto de datos que pertenezcan a un dominio establecido;
b) *finitud*: un algoritmo tiene que estar constituído por una sucesión de instrucciones que pueden ser ejecutadas por el ordenador un número finito de veces;
c) *no ambigüedad*: un algoritmo no tiene que estar constituído por instrucciones que se contradigan o que lleguen a una paradoja.

Los valores sobre los cuales operan las instrucciones de un lenguaje de programación para producir nuevos valores pueden ser:

a) *numéricos*; ejemplo: 3, 25.4, 0.0004.
b) *lógicos*; ejemplo: Verdadero, Falso (True, False).
c) *alfanuméricos*; ejemplo: "7/1/1991", "1. Primer tema".

Nótese que los valores lógicos dados en el apartado b) son los dos únicos existentes.

Con reglas adecuadas los operadores actúan sobre las variables y las constantes para obtener valores nuevos. Una serie de símbolos usados para indicar los operadores se da

en la tabla 1. Y en la tabla 2 se dan los resultados de los operadores *and*, *or* y *xor* de las variables lógicas.

Tabla 1

Simbolo	Tipo de valor del resultado	Operación
$+$	numérico	suma
$-$	numérico	resta
$*$	numérico	multiplicación
$**$	numérico	exponenciación
$/$	numérico	división
$[\,]$	numérico	parte entera
$\sum$	numérico	suma finita
$=$	lógico	igualdad
$\neq$	lógico	no igualdad
$<$	lógico	menor que
$>$	lógico	mayor que
$\leq$	lógico	menor o igual que
$\geq$	lógico	mayor o igual que
not	lógico	cambio de T (F) en F (T)
and	lógico	(a la vez)
or	lógico	(o bien)
xor	lógico	(o bien exclusivo)

Tabla 2

A	B	not A	A and B	A or B	A xor B
T	T	F	T	T	F
T	F	F	F	T	T
F	T	T	F	T	T
F	F	T	F	F	F

La mayoría de las veces, los métodos de tipo constructivo directos dan lugar a algoritmos finitos, mientras que los métodos iterativos producen algoritmos infinitos (convergentes).

Ejemplo de algoritmo finito.

Un ejemplo clásico de algoritmo finito lo constituye el algoritmo de Euclides para el cálculo del *máximo común divisor* (m.c.d.) de dos números. Sean a, b dos números enteros, $a > b$. Entonces:

$$\begin{aligned}
a &= b * q_1 + r_2, && 0 < r_2 < b \\
b &= r_2 * q_2 + r_3, && 0 < r_3 < r_2 \\
r_2 &= r_3 * q_3 + r_4, && 0 < r_4 < r_3 \\
&\ldots. \\
r_{m-2} &= r_{m-1} * q_{m-1} + r_m, && 0 < r_m < r_{m-1} \\
r_{m-1} &= r_m * q_m\ . &&
\end{aligned}$$

Entonces $r_m = \text{m.c.d.}(a, b)$.
El algoritmo sería:

- hacer $r_0 = \text{máx}\ (a, b)$ y $r_1 = \text{mín}\ (a, b)$
- hallar $r_n =$ resto de dividir r_{n-2} entre r_{n-1}, $n = 2, 3, ...$
- cuando $r_n = 0$, parar: el m.c.d. es r_{n-1}.

Ejemplo a). $a = 37$, $b = 1218$

$1218 = 37 * 32 + 34$
$37 = 34 * 1 + 3$
$34 = 3 * 11 + 1$
$3 = 1 * 3 + 0$
$\text{m.c.d.}(37, 1218) = 1, \Rightarrow a,\ b$ primos.

Ejemplo b). $a = 44$, $b = 420$

$420 = 44 * 9 + 24$
$44 = 24 * 1 + 20$
$24 = 20 * 1 + 4$
$20 = 4 * 5 + 0$
$\text{m.c.d.}(44, 420) = 4.$

Ejemplo de algoritmo iterativo.

Un método para el cálculo de $\sqrt{N}$ es el siguiente:
- hacer $x_0 = \frac{1+N}{2}$
- $x_{n+1} = \frac{1}{2}\left[x_n + \frac{N}{x_n}\right], \qquad n = 0, 1, ...\ ,$

entonces $\lim_{n\to\infty} x_n = \sqrt{N}$.

Es evidente que este método es infinito (la sucesión x_{n+1} no se hace constante porque $\sqrt{N}$ no es racional en la mayor parte de los casos), por lo que debemos indicar un criterio de parada para que el algoritmo sea efectivo. Usualmente el criterio es:

$$|x_{n+1} - x_n| < \varepsilon$$

donde ε es la tolerancia permitida.

Si el algoritmo es visto como una serie temporal de operaciones, una pregunta fundamental es ¿cómo viene controlado el flujo de las operaciones? Cuando el programa en ejecución ha llegado a una instrucción particular ¿cómo determina el ordenador cuál es la próxima instrucción que tiene que ejecutar?

Se ha demostrado que sólo tres principios de control son suficientes para describir cualquier algoritmo.
El primer principio es la **noción de secuencia**; excepto que el ordenador sea instruido distintamente, él ejecuta las instrucciones de un programa secuencialmente.
El segundo principio es la **ejecución condicional** que se indica generalmente en el programa con una instrucción del tipo "*If ... then*" (si ... entonces). En la instrucción *if B then S*, B es una expresión boleana, que puede producir sólo los valores verdadero o falso,

y S es una instrucción cualquiera o grupo de instrucciones. Se evalúa B y se ejecuta S sólo si el resultado es verdadero.
El tercer principio es la **repetición** que puede ser indicado con una instrucción "*While ... do*" (mientras ... ejecuta). *While B do S* examina el valor de B y, si es verdadero, ejecuta S: los dos pasos se repiten hasta que una evaluación de B produce el valor falso. En la mayoría de los casos una evaluación de S determina el cambio del valor de B, de manera que el ciclo no continúe para siempre. Otra manera para indicar la repetición es el bucle *Do ... var = vari, varf, vars S continue*. El *Do ... continue* repite las instrucciones del bloque *S* para los valores de la variable *var* desde *vari* hasta *varf* con paso *vars*.

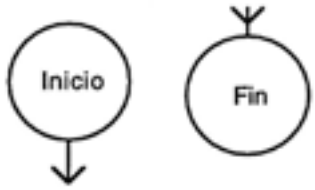
En cada lenguaje de programación, los valores sobre los cuales operan las instrucciones son las constantes y las variables (numéricas, lógicas o alfanuméricas). Además, las instrucciones fundamentales, que se pueden individualizar con un nombre o con un número (llamado *dirección*), son de los tipos siguientes:

a) **instrucciones de asignación**, que permiten asignar el valor de una expresión a una variable;
b) **instrucciones de salto incondicional**, que permiten interrumpir el orden normal de ejecución de las instrucciones de un algoritmo;
c) **instrucciones de condición**, que comparando dos valores, condicionan la ejecución de unas instrucciones en lugar de otras;
d) **instrucciones de transmisión**, que permiten transferir valores entre el mundo externo y el ordenador;
e) **instrucciones de principio de ejecución y de fin de ejecución**, que comandan el inicio o fin de la ejecución de instrucciones del algoritmo.

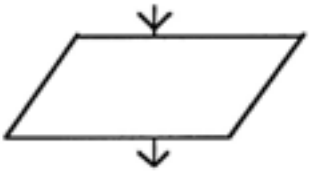
Como se ha dicho antes, excepto las instrucciones de salto incondicional, todas las otras se ejecutan en el orden en el cual están escritas, y la ejecución de una instrucción no empieza hasta que no haya acabado la ejecución de la instrucción previa.

La estructura de un algoritmo se puede representar gráficamente con un diagrama dinámico de líneas que conectan sucesiones de instrucciones del algoritmo. Cada una de esa sucesión de instrucciones es incluida en una figura y las líneas indican la interconexión entre las sucesiones. Conviene dar forma distinta a las figuras dependiendo del tipo de instrucciones que contenga (ver tabla 3). El diagrama dinámico así realizado se llama **diagrama de flujo** (flow chart).

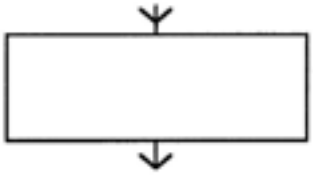
Tabla 3



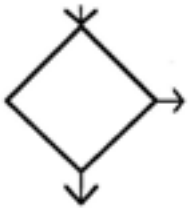
Instrucción de principio
o de fin de ejecución



Instrucción de transmisión



Instrucción de asignación

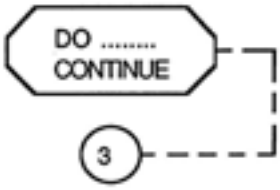


Instrucción de condición

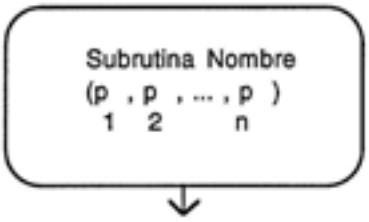


Instrucción de salto incondicional
a la instrucción número n

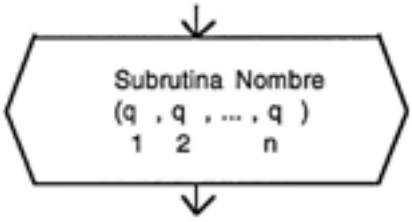
Dirección n de una instrucción



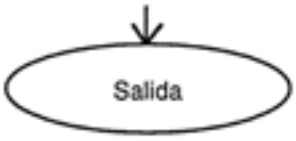
Instrucción para bucle



Instrucción de principio de una
"subroutine" llamada Nombre,
y con parámetros formales p



Instrucción de llamada de un
subrutina llamada Nombre,
y con parámetros actuales q



Instrucción de fin de una
subrutina

Ejemplo 2. Buscar las raíces reales de la ecuación

$$a\ x^2 + b\ x + c = 0\ .$$

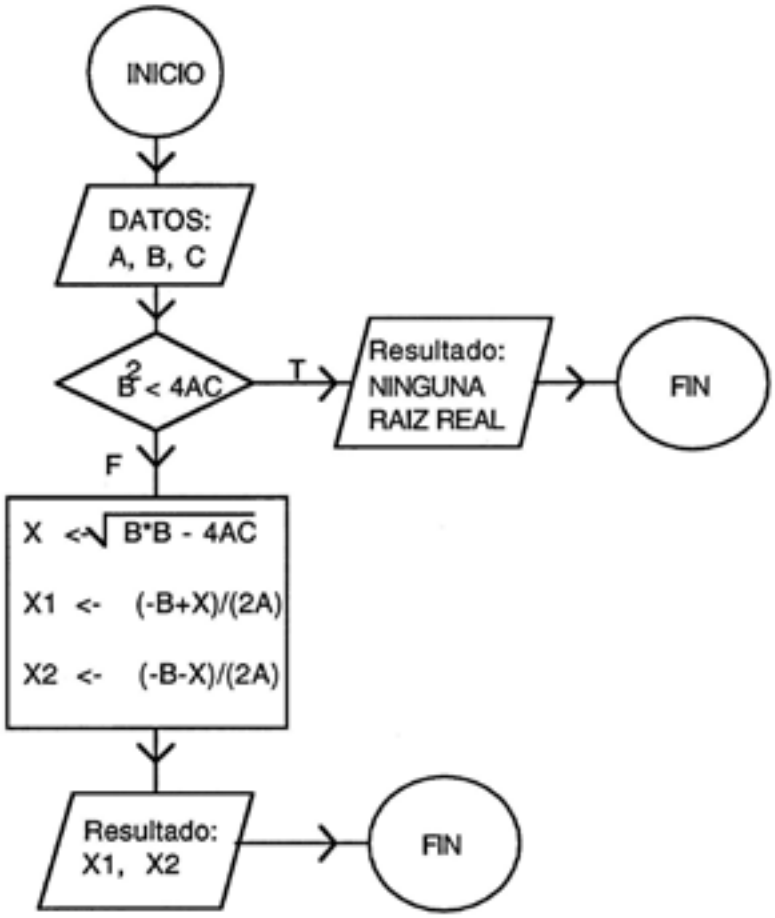
Para $a \neq 0$ las raíces son:

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad\qquad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}\ .$$

Para que x_1 y x_2 sean reales, la expresión $b^2 - 4ac$ no puede ser negativa.

El diagrama de flujo es el siguiente:

Figura 1

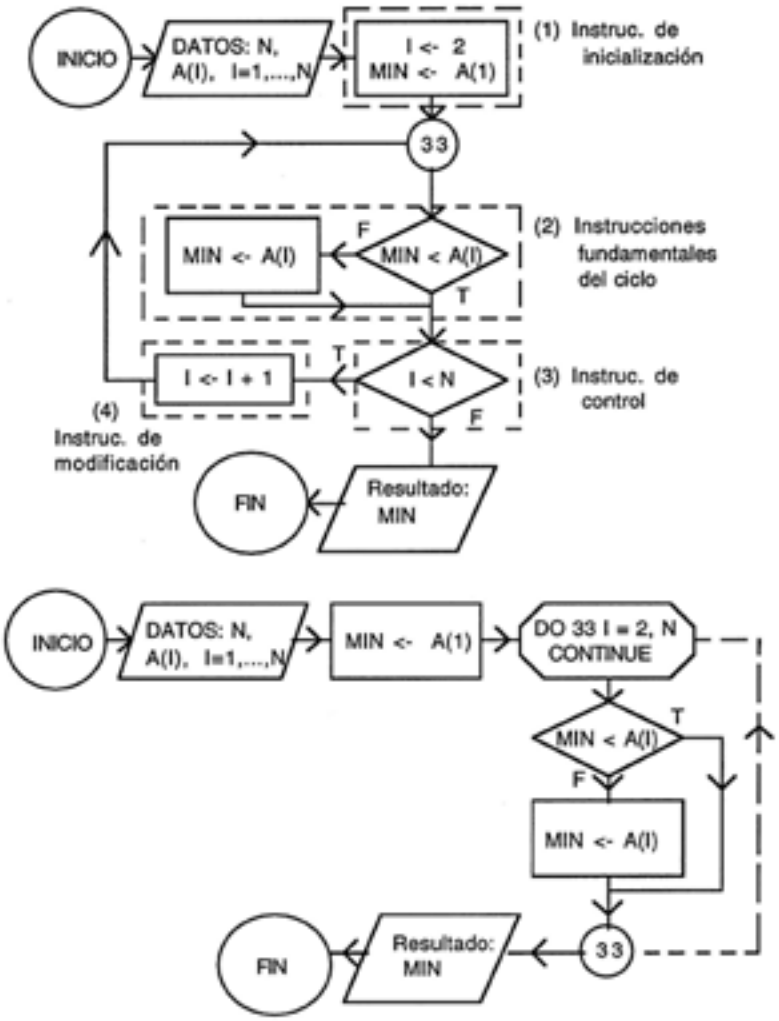


Una parte fundamental de los algoritmos es el **ciclo** (loop). Un ciclo es una sucesión particular de instrucciones que permiten que el ordenador las repita hasta que no sean verificadas unas condiciones que alteren el orden de ejecución. Un ciclo está constituido por cuatro tipos de instrucciones:

(1) instrucciones de inicialización;
(2) instrucciones fundamentales del ciclo;
(3) instrucciones de control;
(4) instrucciones de modificación.

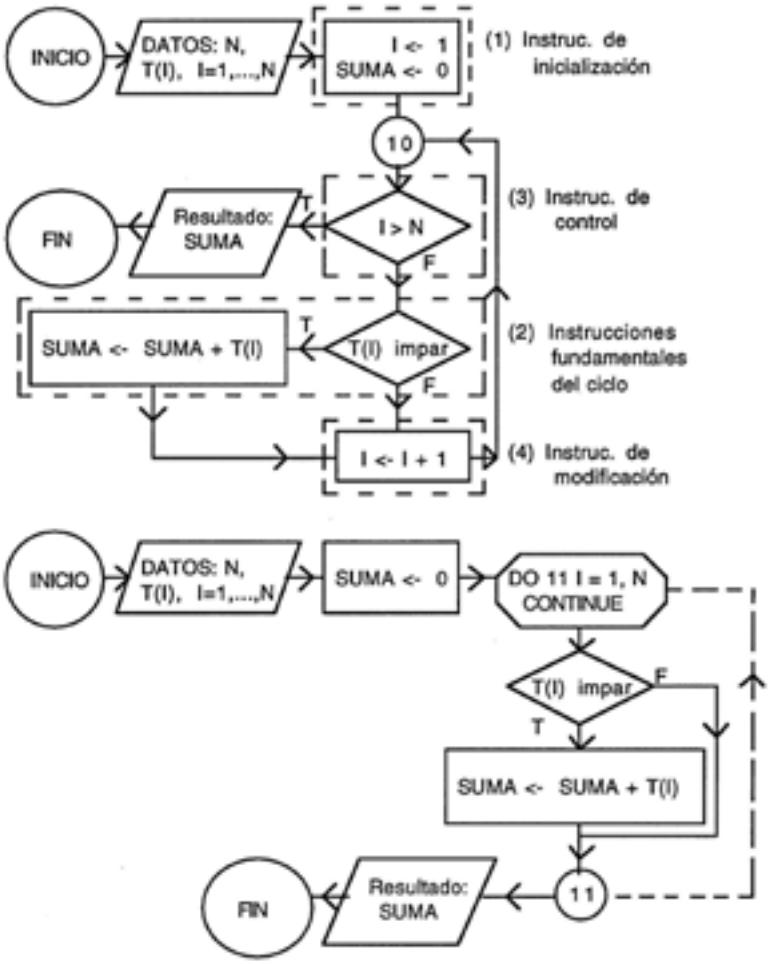
Ejemplo 3. Buscar el minimo de N números reales $a_1, a_2, ..., a_N$.

Figura 2



Ejemplo 4. Buscar la suma de los elementos impares de un vector de N elementos.

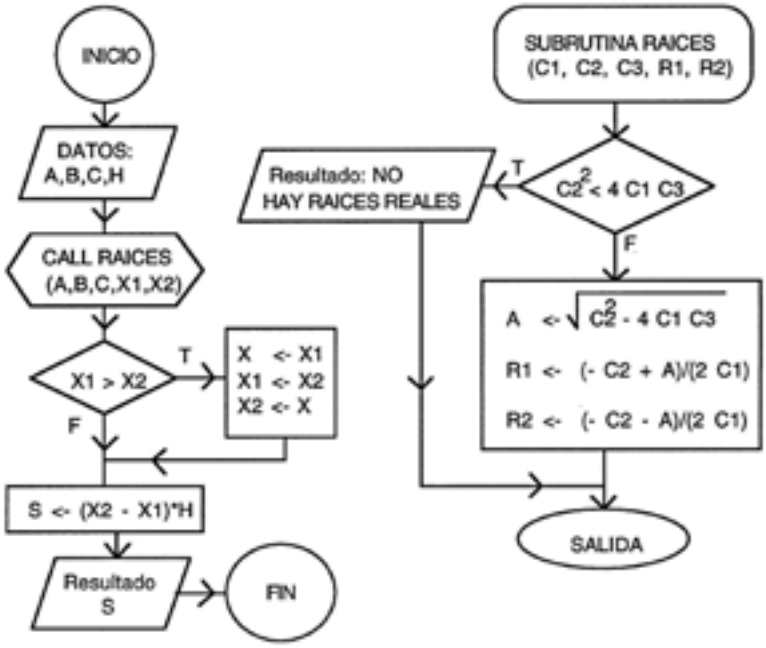
Figura 3



A menudo es conveniente que un problema caracterizado por un algoritmo A sea dividido en un número finito de problemas más sencillos, llamados subrutinas (**subroutines**). Uno de los motivos principales por los cuales es conveniente efectuar esa división en problemas más sencillos es que si se necesita resolver el mismo problema en más de un lugar del algoritmo principal, con diferentes datos, no es muy eficiente repetir las mismas instrucciones que tienen sólo nombres distintos por las variables sobre las cuales operan. Sin embargo, es más conveniente escribir un algoritmo separado que resuelva el problema parcial con **datos formales**, y organizar el problema principal originario de manera que las partes distintas se conectan a la subroutine, transmitiendo los **datos actuales**.

Ejemplo 5. Dada la ecuación $a\ x^2 + b\ x + c = 0$ con raíces x_1, x_2 reales, y $x_2 \geq x_1$, calcular el área del rectángulo de lados h y $x_2 - x_1$.

Figura 4



2. ORIGEN Y EVOLUCION DEL ANALISIS NUMERICO

Debido a la estrecha relación existente entre las diferentes ramas de la Ciencia (y en particular de las Matemáticas), no es fácil determinar dónde acaba una y empieza otra. Por ello la extensión exacta del Análisis Numérico no es conocida. De hecho, el concepto de **Análisis Numérico** no fue creado hasta 1947 en que se fundó el Instituto de Análisis Numérico en la Universidad de California. Sin embargo, el nombre parece estar asociado a aquellos temas que requieran un *procesamientos de datos*. Como la extensión de estos temas es considerable (puede ir, por ejemplo, desde la interpretación de datos médicos hasta la reserva automática de plazas de avión o gestión de una biblioteca), nos limitaremos a ciertos aspectos matemáticos de la idea.

Al principio, la mayor parte del trabajo que se efectuaba en el campo de las **Matemáticas**, inspirado por cuestiones y problemas concretos, se basaba en métodos constructivos para determinar la solución (predicciones sobre eclipses, aparición de un cometa, etc...).

El punto culminante de la utilización de los algoritmos está en Euler (1707–1783), que en los 70 volúmenes que comprenden sus trabajos incluye gran número de algoritmos y fórmulas. Los algoritmos infinitos que presenta, aparecen, normalmente, como desarrollos en serie.

Posteriormente, la perfección de los conocimientos matemáticos y la generalización de los problemas hacen que se sustituyan los razonamientos constructivos por otros de tipo lógico. Así, interesa más determinar si existe la solución a un determinado problema, que calcularlo de forma efectiva. Este proceso sigue hasta aproximadamente el año 1950. La razón del proceso de abstracción era que los algoritmos para el cálculo de las soluciones de los problemas eran, aunque finitos, irrealizables por la gran cantidad de cálculos que exigían. A partir de la segunda mitad del siglo XX, la aparición de las computadoras liberan al algoritmo de la pesadez del cálculo, lo que supone un nuevo auge para los métodos constructivos. Podríamos decir que si desde la antigüedad hasta 1945 la velocidad de cálculo se había multiplicado por 10 mediante rudimentarios artefactos (como el ábaco), desde entonces hasta ahora se ha multiplicado por un millón o más. Esto supone que 1 hora de trabajo de ordenador equivale a 200 años de trabajo de una persona, lo que permite realizar tareas inalcanzables en otros tiempos. Esto no significa que todos los algoritmos puedan ser tratados por un ordenador, pues algunos exigen más de 100 años de trabajo del ordenador actual más potente para poder ser llevados a cabo.

Como la eficiencia de un método depende de su facilidad de implementación, la elección del método apropiado para aproximar la solución de un problema está influenciada significativamente por los cambios tecnológicos en calculadoras y computadoras. El factor limitante en la actualidad es generalmente la capacidad de almacenamiento de la computadora, a pesar de que el costo asociado con los tiempos de cómputo es, desde luego, también un factor importante.

3. OBJETIVOS

El *Análisis Numérico* es *Matemática Aplicada* en el sentido de que toca problemas concretos, reales, de aplicación práctica, pero aprovechando los potentes métodos de la Matemática Pura. Por tanto no son materias opuestas, sino complementarias, lo que hace que la importancia de ambas sea cada vez mayor.

Algunos de los problemas que toca el Análisis Numérico son los siguientes:

a) **Problemas de interpolación**, en los que se sustituye una función poco manejable por otra más sencilla que cumple ciertas condiciones de coincidencia con la primera;
b) Problemas derivados de los anteriores, como pueden ser la **integración aproximada** (cuadratura, cubatura), o **derivación aproximada** de funciones poco manejables;
c) **Problemas de aproximación**, análogos a los anteriores, pero en los que se sustituye una función por otra que sea "próxima", en cierto sentido, a la primera;
d) **Resolución aproximada de ecuaciones diferenciales** tanto ordinarias como en derivadas parciales;
e) Los problemas presentados anteriormente producen, en muchos casos, **sistemas de ecuaciones lineales** con gran número de ecuaciones e incógnitas que por su coste de cálculo son irresolubles por métodos clásicos como la regla de Cramer;
f) **Problemas de tipo matricial**, (hallar valores propios, invertir matrices, etc...) relacionados con los anteriores;
g) **Problemas de optimización**, en los que se maximiza o se minimiza un funcional;
h) **Resolución aproximada de ecuaciones algebráicas y sistemas de ecuaciones no lineales**.

EJERCICIOS.

1. Construir un algoritmo (y dibujar su diagrama de flujo) que tenga como entrada un entero $n \geq 1$, $n+1$ puntos $x_0, x_1, ..., x_n$ y un punto x y que produzca como salida el producto $P = (x - x_0) * (x - x_1) * ... * (x - x_n)$.

2. A) Construir un algoritmo (y dibujar su diagrama de flujo) para buscar la suma de los elementos pares de un vector de n elementos.
B) Ejecutar el algoritmo en el caso del vector: $(2, 3, 4, 7, 8, 9, 12, 27, 38)$.

3. A) Construir un algoritmo (y dibujar su diagrama de flujo) que tomando la ecuación $a\ x^2 + b\ x + c = 0$ con raices reales x_1 y x_2, $x_1 \geq x_2$,
- calcule el perímetro y el área del cuadrado de lado $x_1 - x_2$;
- calcule la circunferencia y el área del círculo de radio $x_1 - x_2$;
B) Ejecutar los algoritmos en el caso que los coeficientes a, b, c de la ecuación sean $a = 3$, $b = -27$ y $c = 42$.

4. Construir un algoritmo (y dibujar su diagrama de flujo) que calcule la media aritmética de n valores dados $a_1, a_2, ..., a_n$.

5. A) Construir un algoritmo (y dibujar su diagrama de flujo) para buscar la suma y el producto de los elementos pares y de los elementos impares de un vector de n elementos.
B) Ejecutar el algoritmo en el caso que el vector esté constituido por los enteros entre 1 y 9 (incluídos).

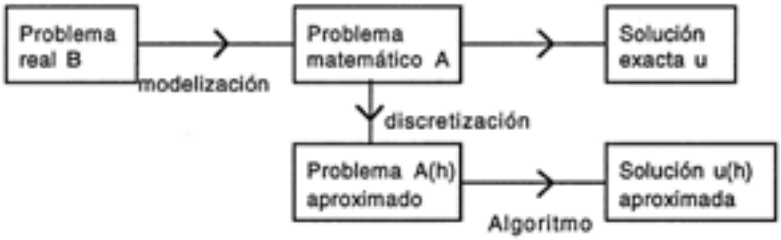
6. A) Construir un algoritmo (y dibujar su diagrama de flujo) para buscar el cuadrado de los elementos pares y el cubo de los elementos impares de un vector de n elementos.
B) Ejecutar el algoritmo en el caso que el vector esté constituido por los enteros entre 1 y 9 (incluídos).

7. Representar los diagramas de flujo de los algoritmos presentados en este capítulo.

CAPITULO II. ANALISIS DE LOS ERRORES

1. ESQUEMA DE RESOLUCION NUMERICA DE UN PROBLEMA

Si se desea resolver un problema físico B, lo primero que se suele hacer es traducirlo al lenguaje matemático para dar un problema matemático A. Se estudia la existencia y unicidad de la solución u de este problema, pero en la mayor parte de los casos y después de probado esto, no se sabe cómo determinar la solución de forma efectiva. Por ello, se sustituye el problema matemático A por un problema proximo a él, A_h, en el que aparecerá algún parámetro h que se va a hacer tender hacia un cierto valor (normalmente 0). Se exige que este problema tenga solución única, u_h, y se espera que al tender h hacia el valor elegido, u_h converja hacia u. Esquemáticamente este tratamiento típico (pero no único), es el siguiente:



De este planteamiento surgen algunos problemas interesantes:

a) ¿Cuál es la **velocidad de convergencia** de u_h hacia u?

b) **Problemas de estabilidad**; es inevitable cometer errores en el cálculo, debido a los redondeos que efectúan los computadores. Interesa que pequeños errores cometidos en los cálculos que conducen a u_h hagan que el resultado no difiera mucho de u; (de eso hablaremos más en el siguiente párrafo).

c) **Coste del proceso**. ¿Cuántas operaciones deben realizarse? ¿Cuánto tiempo se precisará para realizarlas?

Veamos ahora unos ejemplos que muestran la importancia de esas últimas cuestiones:

Ejemplo 1.

Supongamos que se necesita evaluar el polinomio

$$p(x) = 3\,x^4 + 4\,x^3 + 5\,x^2 + 2\,x + 1$$

que es equivalente a:

$$p(x) = ((((3\,x + 4)\,x + 5)\,x + 2)\,x + 1)$$

El número de operaciones para evaluarlo en el primer caso es de:

$4 + 3 + 2 + 1 = 10$ Multiplicaciones

4 Sumas,

mientras que en el segundo se requieren solamente:

4 Multiplicaciones

4 Sumas.

En el caso general de un polinomio de orden n:

$$p(x) = a_0\ x^n + a_1\ x^{n-1} + ... + a_{n-1}\ x + a_n\ ,$$

$$p(x) = ((...((a_0\ x + a_1)\ x + a_2)\ x + ... + a_{n-1})\ x + a_n)\ ,$$

sólo hay diferencia en el número de multiplicaciones (el de sumas es n en ambos casos), que es de

$n + (n-1) + ... + 1 = \frac{(n+1)n}{2} \approx \frac{n^2}{2}$ en el primer caso y

n en el segundo .

Se comprende pues, que es preferible usar el segundo método porque exige menos operaciones (y por lo tanto existen menos posibilidades de que se propaguen los errores de redondeo, lo que dará lugar a una solución más exacta). El algoritmo que lleva a evaluar el polinomio con el segundo método se denomina **algoritmo de Horner** y es:

$b_0 = a_0$

$b_i = a_i + b_{i-1}\ x, i = 1, ..., n$.

Ejemplo 2.

Para resolver sistemas de ecuaciones de orden n con el método de Cramer se precisa un total de $(n+1)!\ (n-1)$ operaciones (multiplicaciones) (cada determinante exige $n!\ (n-1)$ multiplicaciones $a_{p(1)}\ a_{p(2)}...a_{p(n)}$ y hay $n+1$ determinantes a calcular).

El método de Gauss (que explicaremos en un capítulo posterior) exige, sin embargo, sólo $\frac{n^3}{3}$ operaciones. Así, una tabla comparativa de estos métodos sería:

n	5	10	50	100
Cramer	2880	$3.99*10^6$	$7*10^{67}$	10^{160}
Gauss	42	334	42000	334000

Por ejemplo, para $n = 5$, haciendo una operación cada medio minuto (manualmente) se tardarían 24 horas en resolver el sistema por el método de Cramer, mientras que por el de Gauss se tardarían sólo 21 minutos.

Si se intentase utilizar el método de Cramer para resolver un sistema de orden 15 en un ordenador que efectuase 10^6 operaciones por segundo, tardaría más de 9 años en obtener la solución, que además posiblemente no se parecería en nada a la solución verdadera debido a los errores de redondeo que se hubieran producido. ¡Con el método de Gauss, el mismo ordenador tardaría centésimas de segundo!

Este último ejemplo justifica suficientemente la necesidad de buscar algoritmos que sean prácticos.

2. DISTINTOS TIPOS DE ERRORES

Generalmente el resultado de un cálculo numérico es aproximado (sólo en casos excepcionales es exacto), y por eso necesitamos conocer la precisión.

Si p y p^* son dos números reales y p^* se considera como aproximación de p, una medida de la precisión de p^* es

$$E = |p - p^*| \ .$$

De costumbre el conocimiento de E no basta para establecer si p^* es una aproximación buena de p. Por ejemplo:

$p_1 = 5.1346$, $p_1^* = 5.1345$

$E = |p_1 - p_1^*| = 10^{-4}$,

y

$p_2 = 0.0005$, $p_2^* = 0.0004$

$E = |p_2 - p_2^*| = 10^{-4}$.

En los dos casos E es igual a 10^{-4}, pero sólo en el primer caso pensamos que p_1^* es una buena aproximación de p_1. En el segundo caso, p_2 y E son del mismo orden de magnitud, y entonces nos parece mejor considerar su razón.

Damos entonces la siguiente definición: si p^* es una aproximación de p, el **error absoluto** está dado por $E_a = |p - p^*|$, y el **error relativo** está dado por $E_r = \frac{|p-p^*|}{|p|}$, siempre y cuando $p \neq 0$.

En el ejemplo previo, obtenemos:

$E_r = \frac{|p_1-p_1^*|}{|p_1|} = 0.000019476$,

$E_r = \frac{|p_2-p_2^*|}{|p_2|} = 0.2$.

Muchas son las causas que pueden interferir en la precisión de un cálculo, y generar errores. Esos errores se pueden clasificar en:

a) errores iniciales;

b) errores de redondeo;

c) errores de truncamiento;

d) errores de propagación.

Los **errores iniciales** no se pueden evitar si, por ejemplo, son el resultado de medidas de precisión limitada. Supongamos que debemos calcular $f(x)$ en un cierto punto x. Puede ocurrir que estemos obligados a sustituir x por x', con lo cual se calculará $f(x')$ en vez de $f(x)$. Se llama error inicial al valor $f(x') - f(x) = \varepsilon_i$.

Los **errores de redondeo** son debidos a redondeos en los cálculos porque están hechos con un número finito de cifras significativas. Entonces, y continuando con el ejemplo previo, no calcularemos $f(x')$ sino $f_1(x')$. El valor $f_1(x') - f(x') = \varepsilon_r$ se llama error de redondeo.

Los **errores de truncamiento** generalmente corresponden a truncamientos de procedimientos infinitos (desarrollos en serie, etc.). En el ejemplo previo puede ocurrir que

f (y f_1) sea poco manejable y estamos obligados a sustituirla por otra función próxima a ella, f_2. El valor $f_2(x') - f_1(x') = \varepsilon_t$ es llamado error de truncamiento o de discretización.

Aquí es útil, por ejemplo, recordar el **Teorema de Taylor**: supongamos que $f \in C^n[a,b]$ y $f^{(n+1)}$ existe en [a, b). Sea $x_0 \in [a,b]$. Para toda $x \in [a,b]$, existe $\xi(x)$ entre x_0 y x tal que

$$f(x) = P_n(x) + R_n(x)$$

donde

$$P_n(x) = f(x_0) + f'(x_0)\ (x - x_0) + \frac{f''(x_0)}{2!}\ (x - x_0)^2 + ... + \frac{f^{(n)}(x_0)}{n!}\ (x - x_0)^n$$

$$= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}\ (x - x_0)^k$$

y

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\ (x - x_0)^{(n+1)}\ .$$

A $P_n(x)$ se le llama el polinomio de Taylor de grado n para f alrededor de x_0 y a $R_n(x)$ se le llama el residuo (o error de truncamiento) asociado con $P_n(x)$. La serie infinita que se obtiene tomando el límite de $P_n(x)$ cuando $n \to \infty$ se denomina Serie de Taylor para f alrededor de x_0. En el caso de que $x_0 = 0$, el polinomio de Taylor se conoce frecuentemente como polinomio de MacLaurin, y la serie de Taylor se denomina serie de MacLaurin.

Los **errores de propagación** son debidos a la propagación de errores previos en el algoritmo.

Ejemplo 3.

Supongamos que se desea calcular $e^{\sqrt{2}/8}$, y que disponemos de una calculadora de seis dígitos significativos que no dispone de la función exponencial. Es sabido que

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + ...$$

- El primer paso consistirá en calcular $\sqrt{2} \approx 1.41421$
- A continuación el punto en el que se evalúa la función será $x' = 1.41421/8$ en vez del verdadero $\sqrt{2}/8$. El error que estamos cometiendo $f(x') - f(x) = \varepsilon_i$ es el error inicial.
- Pasamos a utilizar una función aproximada a e^x, como puede ser los tres primeros términos de la serie. Se comete así el error de truncamiento $\varepsilon_t = 1 + x' + \frac{x'^2}{2} - e^x$.
- La necesidad de utilizar 6 dígitos significativos hace que las operaciones se hagan sólo con esos 6 dígitos, con lo cual se pierden partes decimales: se está produciendo el error de redondeo.

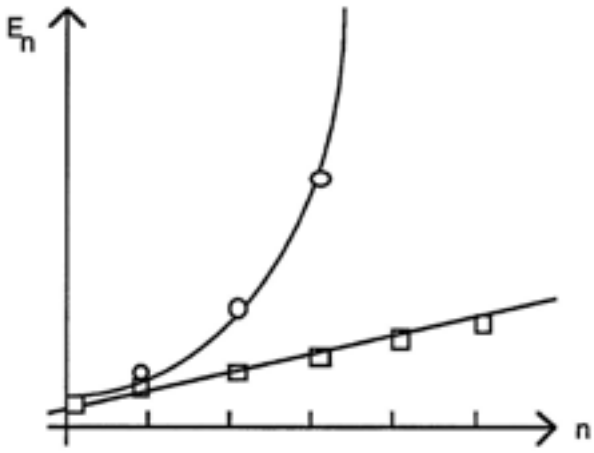
3. CONVERGENCIA

Hemos dicho ya que los cálculos que involucran aproximaciones en la máquina pueden resultar en el crecimiento de los errores de redondeo. Por supuesto, estamos interesados en escoger métodos que produzcan resultados fiables en su precisión. Un criterio que impondremos en un algoritmo, cuando sea posible, es que cambios pequeños en los datos iniciales produzcan correspondientemente cambios pequeños en los resultados finales. Un algoritmo que satisfece esta propriedad se llama **estable**. Es **inestable** cuando este criterio no se cumple. Algunos algoritmos serán estables para ciertos grupos de datos iniciales pero no para todos. Se tratará, siempre que se pueda, de caracterizar las propiedades de estabilidad de los algoritmos.

Para considerar un poco más el tema del crecimiento del error de redondeo y su conexión con la estabilidad de los algoritmos, supongamos que se introduce un error ε en alguna etapa de los cálculos y que el error después de n operaciones subsecuentes se denota por E_n. Los dos casos que se presentan más frecuentemente en la práctica se definen a continuación.

Definición. Supongamos que E_n representa el crecimiento del error después de n operaciones subsecuentes. Si $|E_n| \approx C\ n\ \varepsilon$, donde C es una constante independiente de n, se dice que el crecimiento del error es **lineal**. Si $|E_n| \approx k^n\ \varepsilon$, para algún $k > 1$, el crecimiento del error es **exponencial**.

El crecimiento lineal del error es usualmente inevitable, y cuando C y ε son pequeños los resultados son generalmente aceptables. El crecimiento exponencial del error debe ser evitado, ya que el término k^n será grande aún para valores pequeños de n. Esto lleva a imprecisiones inaceptables, no importando la magnitud de ε. Como consecuencia, un algoritmo que exhibe crecimiento lineal del error es estable, mientras que un algoritmo en el que el crecimiento del error es exponencial es inestable.



Ejemplo 4.

La sucesión $p_n = (\frac{1}{3})^n,\ n > 0$, puede generarse recursivamente tomando $p_0 = 1$ y definiendo $p_n = (\frac{1}{3})\ p_{n-1}$, para $n > 1$. Si obtenemos la sucesión de esta manera, usando aritmética de redondeo a cinco dígitos, los resultados vienen dados en la tabla 1.
El error de redondeo introducido en reemplazar $\frac{1}{3}$ por 0.33333 produce un error de sólo $(0.33333)^n \times 10^{-5}$ en el n-ésimo término de la sucesión. Este método de generar la sucesión es claramente estable.

Tabla 1

n	p_n
0	0.10000×10^1
1	0.33333×10^0
2	0.11111×10^0
3	0.37036×10^{-1}
4	0.12345×10^{-1}

Otra manera de generar la sucesión es definiendo $p_0 = 1$, $p_1 = \frac{1}{3}$, y calculando para cada $n \geq 2$,

$$p_n = (\frac{10}{3})\ p_{n-1} - p_{n-2}\ .$$

La tabla 2 muestra los resultados tanto exactos como redondeados a cinco dígitos usando esta fórmula.

Tabla 2

n	p_n calculado	p_n exacto
0	0.10000×10^1	0.10000×10^1
1	0.33333×10^0	0.33333×10^0
2	0.11111×10^0	0.11111×10^0
3	0.37000×10^{-1}	0.37037×10^{-1}
4	0.12230×10^{-1}	0.12346×10^{-1}
5	0.37660×10^{-2}	0.41152×10^{-2}
6	0.32300×10^{-3}	0.13717×10^{-2}
7	-0.26893×10^{-2}	0.45725×10^{-3}
8	-0.92872×10^{-2}	0.15242×10^{-3}

Este método es obviamente inestable.

Nótese que la fórmula dada, $p_n = (\frac{10}{3})\ p_{n-1} - p_{n-2}$, se satisface si p_n es de la forma

$$p_n = C_1(\frac{1}{3})^n + C_2 3^n$$

para cualquier par de constantes C_1 y C_2. Para verificar esto, notemos que

$$\begin{aligned}\frac{10}{3}\ p_{n-1} - p_{n-2} =& \frac{10}{3}[C_1(\frac{1}{3})^{n-1} + C_2 3^{n-1}] - [C_1(\frac{1}{3})^{n-2} + C_2 3^{n-2}] \\ =& C_1[\frac{10}{3}(\frac{1}{3})^{n-1} - (\frac{1}{3})^{n-2}] + C_2[\frac{10}{3}3^{n-1} - 3^{n-2}] \\ =& C_1(\frac{1}{3})^n + C_2 3^n = p_n\ .\end{aligned}$$

Para tener $p_0 = 1$ y $p_1 = \frac{1}{3}$, las constantes C_1 y C_2 deben escogerse como $C_1 = 1$ y $C_2 = 0$. Sin embargo, en la aproximación de cinco dígitos, los dos primeros términos son $p_0 = 0.10000 \times 10^1$ y $p_1 = 0.33333 \times 10^0$, los cuales requieren una modificación de estas constantes a $C_1 = 0.10000 \times 10^1$ y $C_2 = -0.12500 \times 10-5$. Este pequeño cambio en C_2 da lugar a un error de redondeo de $3^n(-0.12500 \times 10-5)$ al producir p_n. Como consecuencia resulta un crecimiento exponencial del error, lo cual se refleja en la pérdida extrema de exactitud encontrada en la tabla 2.

Para reducir los efectos del error de redondeo, podemos usar una aritmética de un orden grande de dígitos, como las opciones de doble o múltiple precisión, disponibles en la mayoria de las computadoras digitales. Una desventaja del uso de la aritmética de doble precisión es que toma mucho más tiempo de computadora. Por otro lado, no se elimina completamente el crecimiento serio del error de redondeo, sino que sólo se postpone si es que se realizan un gran número de cálculos posteriores. Hay también otros métodos para estimar el error de redondeo (aritmética de intervalo, métodos estadísticos, etc.) que no estudiaremos.

EJERCICIOS.

1. Considerar los siguientes valores de p y p^*. ¿Cuál es el (i) error absoluto, (ii) error relativo al aproximar p^* por p?

a) $p = \pi, \quad p^* = 3.1$;
b) $p = \frac{1}{3}, \quad p^* = 0.333$;
c) $p = \frac{\pi}{1000}, \quad p^* = 0.0031$;
d) $p = \frac{100}{3}, \quad p^* = 33.3$;
e) $p = \pi, \quad p^* = 3.141$;
f) $p = \frac{1}{3}, \quad p^* = 0.33333$;
g) $p = \frac{\pi}{1000}, \quad p^* = 0.00314$;
h) $p = \frac{100}{3}, \quad p^* = 33.33$.

2. Considerar los siguientes valores de p y p^*. ¿Cuál es el (i) error absoluto, (ii) error relativo al aproximar p^* por p?

a) $p = e, \quad p^* = 2.5$;
b) $p = \frac{10}{3}, \quad p^* = 3.33333$;
c) $p = \frac{e}{1000}, \quad p^* = 0.0025$;
d) $p = \frac{10000}{3}, \quad p^* = 3333.333$.

3. Calcular los siguientes valores en el caso que se disponga de una calculadora de seis dígitos significativos que no dispone de la función exponencial (usar el desarrollo en serie de Taylor hasta el cuarto orden):

a) e^2 ;
b) $e^{-1} + e^3$.

CAPITULO III. SISTEMAS DE NUMERACION

1. REPRESENTACION DE LA INFORMACION

El sistema de numeración usado habitualmente es el decimal, de base 10, que no es adecuado para ser manejado por el ordenador, fundamentalmente porque es más sencillo construir un elemento con únicamente dos posibles estados que uno con 10 estados. Y por otra parte como los componentes electrónicos envejecen es más dificil mantener correctamente un dispositivo con 10 estados que uno con dos.

Una información dada al ordenador es, generalmente, representada con una sucesión de caracteres escogidos desde un alfabeto compuesto sólo de dos caracteres, representados, respectivamente, por los simbolos "0" y "1", llamados cifras binarias o **bits** (binary digits).

Cada uno de los caracteres es físicamente representado por uno de los posibles estados de los componentes del ordenador: un núcleo magnético es imanado en una de las dos posibles direcciones de magnetización, un circuito puede ser abierto o cerrado.

El problema de representar los caracteres de un alfabeto hecho con más de dos caracteres se resuelve uniendo más cifras binarias; por ejemplo, los agrupamientos hechos con dos cifras binarias (00, 01, 10, 11) dan la posibilidad de distinguir cuatro caracteres diversos, los obtenidos con tres cifras binarias se pueden usar para distinguir ocho caracteres distintos, y en general los agrupamientos obtenidos con n cifras binarias pueden representar 2^n caracteres distintos.

El usuario de un ordenador no tiene que conocer necesariamente la representación de los datos en la máquina, porque los lenguajes comunes de programación permiten especificar datos e instrucciones con los caracteres y las cifras usadas comunmente por el hombre, y son los compiladores los que convierten al sistema de representación propio del ordenador.

Sin embargo, para la solución de muchos problemas es conveniente que el usuario de un ordenador conozca el sistema binario y los principios fundamentales de la aritmética de un ordenador, aunque nunca contará en la aritmética binaria.

2. INTRODUCCION A LOS SISTEMAS NUMERICOS

La representación común de los números es constituida por sucesiones de los símbolos "$0, 1, 2, 3, 4, 5, 6, 7, 8, 9$". Tales sucesiones pueden ser precedidas por los símbolos "+" y "-" (para indicar un número positivo o negativo), y pueden tener el símbolo "." (**punto raíz**, que separa la parte entera del número, a la izquierda, de su parte decimal, a la derecha). Nuestro sistema decimal es un sistema **posicional**, es decir cada cifra tiene un *peso*. La posición ocupada por cada dígito tiene un significado exacto y determina la contribución de la cifra al valor numérico de la sucesión.

Por ejemplo, 35 y 53 están constituidos por las mismas cifras 3 y 5, pero tienen significados distintos

$$35 = 3 * 10^1 + 5 * 10^0 \ ,$$

$$53 = 5 * 10^1 + 3 * 10^0 \ .$$

Cada cifra de la sucesión es multiplicada por una potencia de 10, con el exponente determinado por la posición de la cifra con respecto al punto raíz. El valor 10 es la **base** de nuestro sistema de numeración, que por esta razón se denomina **sistema posicional en base diez**. Los exponentes de la parte entera son positivos y crecen en unidades a partir de cero, que corresponde al exponente de la potencia de 10 que multiplica la cifra más a la derecha de la parte entera, los exponentes de la parte fraccionaria del número son negativos y disminuyen en unidades a partir de -1, que corresponde al exponente de la potencia de 10 que multiplica la primera cifra de la parte fraccionaria.

Un ejemplo de sistema no posicional lo constituye la numeración romana. En ella, al valor 5 le corresponde el símbolo V, mientras que al valor 50 le corresponde el símbolo L. Para pasar de 5 a 50 no basta con cambiar la posición del símbolo 5 (V), sino que hay que introducir uno nuevo (L).

La descripción hecha del sistema posicional decimal sugiere la posibilidad de usar un sistema de numeración en una base distinta de 10. Con el sistema en base 10 se usan 10 cifras para representar cada número; en general, para representar un número en una base b, se necesitan b símbolos. Entonces, cuando se considera como base un entero $b > 10$, las 10 cifras del sistema decimal no bastan y es necesario usar símbolos nuevos.

Las bases más usadas, además de la 10, son 2, 8 y 16. El sistema en base dos, llamado **sistema binario**, usa las cifras "0, 1". El sistema en base ocho, llamado **sistema octal**, usa las cifras "0, 1, 2, 3, 4, 5, 6, 7". Y el **sistema hexadecimal**, en base 16, usa las cifras "0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F".

3. CONVERSION DESDE EL SISTEMA DECIMAL AL SISTEMA NUMERICO EN BASE b

La existencia de sistemas de numeración en base distintas de 10, hace que nos planteemos el problema de la representación de los números en una nueva base, y la definición de las reglas y de las propiedades formales de las operaciones a ejecutar en la nueva aritmética.

Supongamos que conocemos la representación decimal de un número cualquiera entero positivo, y queremos construir la correspondiente representación en base b. Cada entero n se expresa en la base b en la forma

$$n = d_p * b^p + d_{p-1} * b^{p-1} + ... + d_1 * b^1 + d_0 * b^0 \ , \qquad (III.1)$$

donde p y $d_p, d_{p-1}, ..., d_1, d_0$ son enteros que tenemos que determinar para obtener la representación

$$n_b = d_p \ d_{p-1} \ ... \ d_1 \ d_0 \ .$$

La ecuación (III.1) se puede escribir de la forma

$$n = b * (d_p * b^{p-1} + d_{p-1} * b^{p-2} + ... + d_1) + d_0,$$

por lo cual se deduce que d_0 es el resto de la división de n entre b. Si denotamos con n_1 el cociente de esa división, tenemos

$$n_1 = d_p * b^{p-1} + d_{p-1} * b^{p-2} + ... + d_1$$

que podemos escribir

$$n_1 = b * (d_p * b^{p-2} + d_{p-1} * b^{p-3} + ... + d_2) + d_1.$$

Entonces, se deduce que d_1 es el resto de dividir n_1 entre b. Si denotamos con n_2 el cociente de esa división, tenemos

$$n_2 = d_p * b^{p-2} + d_{p-1} * b^{p-3} + ... + d_2$$

que podemos escribir como

$$n_2 = b * (d_p * b^{p-3} + d_{p-1} * b^{p-4} + ... + d_3) + d_2.$$

Entonces, se deduce que d_2 es el resto de dividir n_2 entre b. Procediendo de manera parecida se llega a determinar d_{p-1} como el resto de la división de n_{p-1} entre b, y donde

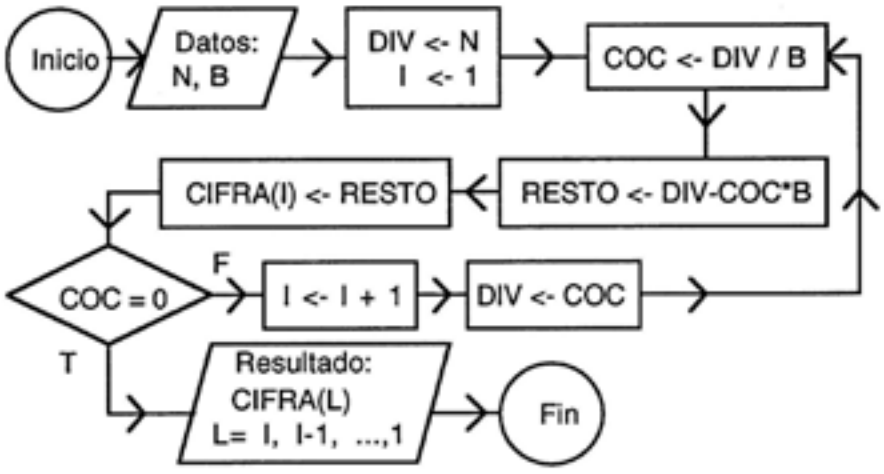
$$n_{p-1} = b * d_p + d_{p-1} \ .$$

Entonces, d_p es el cociente de la última división.

Dado que el valor de p no es conocido, y no se conoce el número de veces que tenemos que dividir entre b, el último cociente d_p se determina de esa manera (conocida como el **método de las divisiones sucesivas**), como el resto de la división por b que da un cociente cero.

En la figura 1 está representado el diagrama de flujo para el método de las divisiones sucesivas. En ese diagrama, N es el número entero para el cual queremos la representación en base B, $CIFRA$ es el vector con las cifras en la representación en base B, COC es el cociente de la división del dividendo DIV entre la base B. $RESTO$ es el resto de la división, e I es un índice con el cual se cuentan las cifras en la representación en base B.

Figura 1



Damos también las potencias de 2:
$2^2 = 4$
$2^3 = 8$
$2^4 = 16$
$2^5 = 32$
$2^6 = 64 = 8^2$
$2^7 = 128$
$2^8 = 256 = 16^2$
$2^9 = 512 = 8^3$
$2^{10} = 1024$
$2^{11} = 2048$
$2^{12} = 4096 = 8^4 = 16^3$

Siguiendo el esquema de la figura 1, es fácil verificar que, por ejemplo

$$1972_{10} = 11110110100_2 = 3664_8 = 7B4_{16}.$$

Construyamos ahora la representación en **base 2**. De hecho es:
$1972/2 = 986$ resto 0
$986/2 = 493$ resto 0
$493/2 = 246$ resto 1
$246/2 = 123$ resto 0
$123/2 = 61$ resto 1
$61/2 = 30$ resto 1
$30/2 = 15$ resto 0
$15/2 = 7$ resto 1
$7/2 = 3$ resto 1
$3/2 = 1$ resto 1
$1/2 = 0$ resto 1.

Luego $1972_{10} = 11110110100_2$, y viceversa:

$$\begin{aligned} 11110110100_2 &= 1*2^{10} + 1*2^9 + 1*2^8 + 1*2^7 + 1*2^5 + 1*2^4 + 1*2^2 \\ &= 1024_{10} + 512_{10} + 256_{10} + 128_{10} + 32_{10} + 16_{10} + 4_{10} = 1972_{10} \end{aligned}$$

Construyamos ahora la representación en **base 8**:
$1972/8 = 246$ resto 4
$246/8 = 30$ resto 6
$30/8 = 3$ resto 6
$3/8 = 0$ resto 3

Entonces $1972_{10} = 3664_8$, y viceversa:

$$\begin{aligned} 3664_8 &= 3*8^3 + 6*8^2 + 6*8^1 + 4*8^0 \\ &= 1536_{10} + 384_{10} + 48_{10} + 4_{10} = 1972_{10} \end{aligned}$$

Y para acabar construyamos la representación en **base 16**:
$1972/16 = 123$ resto 4
$123/16 = 7$ resto B
$7/16 = 0$ resto 7

Entonces $1972_{10} = 7B4_{16}$, y viceversa:

$$7B4_{16} = 7 * 16^2 + B * 16^1 + 4 * 16^0$$
$$= 1792_{10} + 176_{10} + 4_{10} = 1972_{10}.$$

Nótese que el primer resto obtenido en las divisiones es la cifra que tendrá su posición a la inmediata izquierda del punto raíz, así como el último resto obtenido es la cifra que tendrá su posición más a la izquierda del punto raíz.

Consideremos ahora el problema de construir la representación en base b de un número z real y positivo, menor que 1, del cual conocemos la representación decimal. Cada número menor que 1 se expresa en la base b en la forma:

$$z = q_1 * b^{-1} + q_2 * b^{-2} + q_3 * b^{-3} + ... \qquad (III.2)$$

donde $q_1, q_2, q_3, ...$ son enteros que tenemos que determinar para crear la representación

$$z_b = 0.\ q_1\ q_2\ q_3....$$

Supongamos que realizamos la multiplicación de z por b; entonces, de (III.2) obtenemos:

$$z * b = q_1 + q_2 * b^{-1} + q_3 * b^{-2} + ... = q_1 + z_1$$

por lo cual se deduce que q_1 es la parte entera de $z * b$, es decir

$$q_1 = [z * b],$$

y que z_1 es un número menor que 1, obtenido como diferencia del producto $z * b$ menos su parte entera. Del producto $z_1 * b = q_2 + z_2$ se deduce que

$$q_2 = [z_1 * b].$$

De manera análoga se sigue que

$$q_3 = [z_2 * b]$$

$$....$$

$$q_i = [z_{i-1} * b].$$

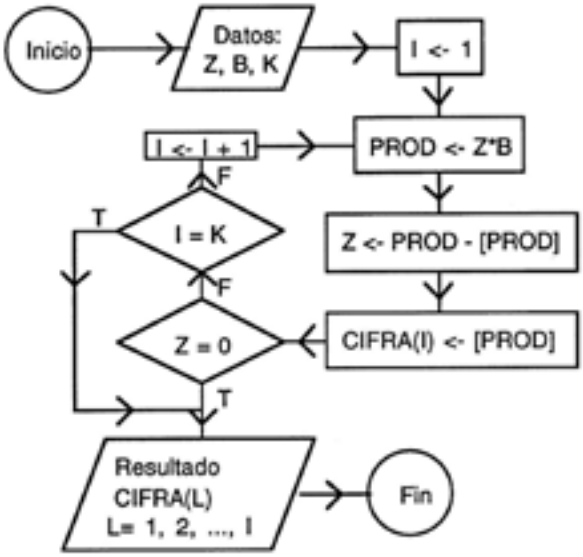
Este método es conocido con el nombre de **método de las multiplicaciones sucesivas**.

En la figura 2 está representado el diagrama de flujo para el método de las multiplicaciones sucesivas. En ese diagrama, Z es el número real positivo menor que 1, para el cual queremos la representación en base B, con no más de K cifras. $CIFRA$ es el vector con las cifras en la representación en base B, y $[PROD]$ es la parte entera del $PROD$.

Damos también las potencias negativas de 2:
$2^{-1} = 0.5$
$2^{-2} = 0.25$

$2^{-3} = 0.125 = 8^{-1}$
$2^{-4} = 0.0625 = 16^{-1}$
$2^{-5} = 0.03125$
$2^{-6} = 0.015625 = 8^{-2}$
$2^{-7} = 0.0078125$
$2^{-8} = 0.00390625 = 16^{-2}$

Figura 2



Siguiendo el esquema de figura 2, es fácil verificar que, por ejemplo:

$$0.828125_{10} = 0.110101_2 = 0.65_8 = 0.D4_{16} \ .$$

Construyamos ahora la representación en **base 2**. De hecho es:
$0.828125 * 2 = 1.65625$ parte entera 1
$0.65625 * 2 = 1.3125$ parte entera 1
$0.3125 * 2 = 0.625$ parte entera 0
$0.625 * 2 = 1.25$ parte entera 1
$0.25 * 2 = 0.5$ parte entera 0
$0.5 * 2 = 1.0$ parte entera 1
$0.0 * 2 = 0.0$

Entonces $0.828125_{10} = 0.110101_2$, y viceversa:

$$0.110101_2 = 1 * 2^{-1} + 1 * 2^{-2} + 1 * 2^{-4} + 1 * 2^{-6}$$
$$= 0.5_{10} + 0.25_{10} + 0.0625_{10} + 0.015625_{10} = 0.828125_{10}$$

Construyamos ahora la representación en **base 8**:
$0.828125 * 8 = 6.625$ parte entera 6
$0.625 * 8 = 5.0$ parte entera 5

$0.0 * 8 = 0.0$

Entonces $0.828125_{10} = 0.65_8$, y viceversa:

$0.65_8 = 6 * 8^{-1} + 5 * 8^{-2}$

$= 6 * 0.125_{10} + 5 * 0.015625_{10} = 0.828125_{10}$

Y para concluir construyamos la representación en **base 16**:

$0.828125 * 16 = 13.25$ parte entera D

$0.25 * 16 = 4.0$ parte entera 4

$0.0 * 16 = 0.0$

Luego $0.828125_{10} = 0.D4_{16}$, y viceversa:

$0.D4_{16} = 13 * 16^{-1} + 4 * 16^{-2}$

$= 13 * 0.0625_{10} + 4 * 0.00390625_{10} = 0.828125_{10}$.

Nótese que la primera parte entera obtenida en las multiplicaciones es la cifra que tendrá su posición más a la inmediata derecha del punto raíz, así como la última parte entera obtenida es la cifra que tendrá su posición más a la derecha del punto raíz.

Aplicamos ahora el método de las multiplicaciones sucesivas para construir la representación en base 2 del número $z_{10} = 0.1_{10}$. Así:

$0.1 * 2 = 0.2$ parte entera 0

$0.2 * 2 = 0.4$ parte entera 0

$0.4 * 2 = 0.8$ parte entera 0

$0.8 * 2 = 1.6$ parte entera 1

$0.6 * 2 = 1.2$ parte entera 1

$0.2 * 2 = 0.4$ parte entera 0

$0.4 * 2 = 0.8$ parte entera 0

$0.8 * 2 = 1.6$ parte entera 1

$0.6 * 2 = 1.2$ parte entera 1

etc.

Por lo tanto tenemos

$$0.1_{10} = 0.000110011...{}_2 = 0.0\overline{0011}_2 \ .$$

Con ese ejemplo, hemos demostrado que un mismo número puede tener, en una base, una representación con un número finito de cifras, mientras que en otra base, puede tener una representación con un número infinito de cifras.

Está claro que ahora sabemos construir la representación en base b de un número real cualquiera. Por ejemplo, el número que en base 10 se reprenta por 1972.828125_{10}, está representado en las bases 2, 8 y 16, respectivamente por:

$$11110110100.110101_2 = 3664.65_8 = 7B4.D4_{16} \ .$$

Buscamos ahora la representación en base 2 del número real -25.375_{10}. Tenemos $25_{10} = 11001_2$, y $0.375_{10} = 0.011_2$. Entonces

$$-25.375_{10} = -11001.011_2 \ .$$

4. LAS OPERACIONES ARITMETICAS EN BASE b

Supongamos que tenemos que calcular la siguiente operación aritmética: $101_2 + 10_2$. Una manera de proceder es convertir los sumandos en base 10, ejecutar la suma, y después convertir el resultado en base 2. Si hacemos así, obtenemos:

$101_2 = 5_{10}, 10_2 = 2_{10}$

$5_{10} + 2_{10} = 7_{10}$

$7_{10} = 111_2$.

Ese método, sin embargo, no siempre es conveniente, sobre todo si pensamos que para las operaciones aritméticas de los números en base b valen las mismas reglas y propiedades formales conocidas en la aritmética en base 10. En el sistema de numeración binario existen sólo dos símbolos: 0_2 y 1_2. Entonces, cuando se efectua la operación $1_2 + 1_2$, no tenemos un único símbolo para representar el resultado, y el resultado es 0 con el reporte de 1, es decir 10_2.

Las reglas para efectuar la suma y la multiplicación de dos cifras binarias están resumidas en la tabla 1.

Tabla 1

+	0	1	*	0	1
0	0	1	0	0	0
1	1	10	1	0	1

Damos también las tablas de la suma y la multiplicación en base 8 (tabla 2) y en base 16 (tabla 3).

Tabla 2

+	0	1	2	3	4	5	6	7
0	0	1	2	3	4	5	6	7
1	1	2	3	4	5	6	7	10
2	2	3	4	5	6	7	10	11
3	3	4	5	6	7	10	11	12
4	4	5	6	7	10	11	12	13
5	5	6	7	10	11	12	13	14
6	6	7	10	11	12	13	14	15
7	7	10	11	12	13	14	15	16

*	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7
2	0	2	4	6	10	12	14	16
3	0	3	6	11	14	17	22	25
4	0	4	10	14	20	24	30	34
5	0	5	12	17	24	31	36	43
6	0	6	14	22	30	36	44	52
7	0	7	16	25	34	43	52	61

Tabla 3

+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
1	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	10
2	2	3	4	5	6	7	8	9	A	B	C	D	E	F	10	11
3	3	4	5	6	7	8	9	A	B	C	D	E	F	10	11	12
4	4	5	6	7	8	9	A	B	C	D	E	F	10	11	12	13
5	5	6	7	8	9	A	B	C	D	E	F	10	11	12	13	14
6	6	7	8	9	A	B	C	D	E	F	10	11	12	13	14	15
7	7	8	9	A	B	C	D	E	F	10	11	12	13	14	15	16
8	8	9	A	B	C	D	E	F	10	11	12	13	14	15	16	17
9	9	A	B	C	D	E	F	10	11	12	13	14	15	16	17	18
A	A	B	C	D	E	F	10	11	12	13	14	15	16	17	18	19
B	B	C	D	E	F	10	11	12	13	14	15	16	17	18	19	1A
C	C	D	E	F	10	11	12	13	14	15	16	17	18	19	1A	1B
D	D	E	F	10	11	12	13	14	15	16	17	18	19	1A	1B	1C
E	E	F	10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D
F	F	10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E

*	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
2	0	2	4	6	8	A	C	E	10	12	14	16	18	1A	1C	1E
3	0	3	6	9	C	F	12	15	18	1B	1E	21	24	27	2A	2D
4	0	4	8	C	10	14	18	1C	20	24	28	2C	30	34	38	3C
5	0	5	A	F	14	19	1E	23	28	2D	32	37	3C	41	46	4B
6	0	6	C	12	18	1E	24	2A	30	36	3C	42	48	4E	54	5A
7	0	7	E	15	1C	23	2A	31	38	3F	46	4D	54	5B	62	69
8	0	8	10	18	20	28	30	38	40	48	50	58	60	68	70	78
9	0	9	12	1B	24	2D	36	3F	48	51	5A	63	6C	75	7E	87
A	0	A	14	1E	28	32	3C	46	50	5A	64	6E	78	82	8C	96
B	0	B	16	21	2C	37	42	4D	58	63	6E	79	84	8F	9A	A5
C	0	C	18	24	30	3C	48	54	60	6C	78	84	90	9C	A8	B4
D	0	D	1A	27	34	41	4E	5B	68	75	82	8F	9C	A9	B6	C3
E	0	E	1C	2A	38	46	54	62	70	7E	8C	9A	A8	B6	C4	D2
F	0	F	1E	2D	3C	4B	5A	69	78	87	96	A5	B4	C3	D2	E1

Veamos ahora unos **ejemplos** de operaciones aritméticas con representaciones en base 2. Para cada operación se usan las tablas dadas, y por control se efectuan también las operaciones en base 10. De manera parecida se procede para las operaciones en bases 8 y 16.

1. Calcular la suma entre los dos números enteros 10001_2 y 11011_2.

$$\begin{array}{r} 10001_2 \\ +\ 11011_2 \\ \cdots\cdots\cdots \\ 101100_2 \end{array}$$

   Entonces, $10001_2 + 11011_2 = 101100_2$, $10001_2 = 17_{10}$, $11011_2 = 27_{10}$, $101100_2 = 44_{10}$.
2. Calcular el producto entre los dos números enteros 1101_2 y 1110_2.

$$\begin{array}{r} 1101_2 \\ *\ 1110_2 \\ \cdots\cdots\cdots\cdots \\ 1101 \\ 1101 \\ 1101 \\ \cdots\cdots\cdots\cdots \\ 10110110_2 \end{array}$$

   Entonces, $1101_2 * 1110_2 = 10110110_2$, $1101_2 = 13_{10}$, $1110_2 = 14_{10}$, $10110110_2 = 182_{10}$.
3. Calcular la diferencia entre los dos números enteros 10100_2 y 1101_2.

$$\begin{array}{r} 10100_2 \\ -\ \ 1101_2 \\ \cdots\cdots\cdots \\ 111_2 \end{array}$$

   Entonces, $10100_2 - 1101_2 = 111_2$, $10100_2 = 20_{10}$, $1101_2 = 13_{10}$, $111_2 = 7_{10}$.
4. Calcular la división entre los dos números enteros 11100_2 y 1001_2.

$$\begin{array}{rcl} 11100 & / & 1001 \\ 1001 & & 11 \\ \cdots\cdots & & \\ 1010 & & \\ 1001 & & \\ \cdots\cdots & & \\ 1 & & \end{array}$$

   Entonces, el cociente y el resto de $11100_2/1001_2$ son, respectivamente, los números 11_2 y 1_2. En la representación decimal, $11100_2 = 28_{10}$, $1001_2 = 9_{10}$, y $28_{10}/9_{10}$ da como resultado 3_{10} y resto 1_{10}.

5. CONVERSION DESDE UN SISTEMA NUMERICO EN BASE b_1 A UN SISTEMA EN BASE b_2

El problema de la conversión de la representación de un mismo número real (no cero) desde una base b_1 a otra base b_2 se puede resolver de distintos modos.

Primer método. Una manera es convertir la representación del número, primero desde la base b_1 a la base 10, y después desde la base 10 a la nueva base b_2. Por ejemplo, para convertir el número X que en base 2 tiene la representación -11101.101_2 desde la base $b_1 = 2$ a la base $b_2 = 8$, se puede proceder de la manera siguiente. Construyamos la representación decimal de X, que se obtiene expresando X en la forma

$$X = -(1 * 2^4 + 1 * 2^3 + 1 * 2^2 + 1 * 2^0 + 1 * 2^{-1} + 1 * 2^{-3}) = -29.625_{10} \ ,$$

y luego se determina la representación de X en la base $b_2 = 8$ en la manera que ya conocemos, obteniendo

$$-11101.101_2 = -29.625_{10} = -35.5_8 \ .$$

Segundo método. Otra manera para obtener la conversión de las representaciones de X desde la base b_1 a la base b_2, no usa la representación decimal intermedia, sino aplica directamente los algoritmos usados para la conversión de un número decimal a la base b, con la condición que las operaciones sean ejecutadas en base b_1.

Por ejemplo, si queremos representar en base $b_2 = 8$ el número X que en base $b_1 = 2$ tiene la representación -11101.101_2, se procede de la forma siguente. Se convierte antes la parte entera del número desde la base 2 a la base 8, recordando que $b_2 = 8$ tiene la representación 1000_2, en base 2. Entonces,

$11101_2/1000_2 = 11_2$ resto 101_2

$11_2/1000_2 = 0$ resto 11_2 .

Dado que $101_2 = 5_8$ y $11_2 = 3_8$, obtenemos $11101_2 = 35_8$. Después se convierte desde base 2 a la base 8 la parte fraccionaria de X, obteniendo

$0.101_2 * 1000_2 = 101.0_2$ parte entera 101_2

$0.000_2 * 1000_2 = 0.0_2$

Dado que $101_2 = 5_8$, obtenemos $0.101_2 = 0.5_8$, y, concluyendo, resulta que

$$-11101.101_2 = -35.5_8 \ .$$

Otro ejemplo: supongamos que se desea construir la representación en base $b_2 = 16$ del número X que en base $b_1 = 8$ está representado por 2777.7661_8. Sabiendo que $b_2 = 16$ se representa en base 8 como 20_8, se convierte la parte entera de X desde la base 8 a la base 16:

$2777_8/20_8 = 137_8$ resto 17_8

$137_8/20_8 = 5_8$ resto 17_8

$5_8/20_8 = 0_8$ resto 5_8 .

Y dado que $17_8 = F_{16}$ y $5_8 = 5_{16}$, obtenemos $2777_8 = 5FF_{16}$. Convertimos ahora desde base 8 a base 16 la parte fraccionaria de X:

$0.7661_8 * 20_8 = 17.542_8$ parte entera 17_8
$0.5420_8 * 20_8 = 13.04_8$ parte entera 13_8
$0.0400_8 * 20_8 = 1.0_8$ parte entera 1_8
$0.0000_8 * 20_8 = 0.0_8$.

Y dado que $17_8 = F_{16}$, $13_8 = B_{16}$ y $1_8 = 1_{16}$, obtenemos $0.7661_8 = 0.FB1_{16}$. Entonces resulta que

$$2777.7661_8 = 5FF.FB1_{16} \ .$$

Tercer método. Otro método para obtener la conversión de un número X desde la base b_1 a la base b_2 es aplicable cuando $b_2 = b_1^k$, siendo k un entero mayor o igual que 2. La conversión se obtiene entonces de la manera siguiente. Sea n un entero positivo del cual tenemos la representación en base b_1. Entonces, descomponemos el alineamiento de los caracteres en agrupamientos, cada uno de k caracteres, partiendo de la derecha hacia la izquierda. Si el agrupamiento más a la izquierda tiene menos de k caracteres se añaden ceros para obtener un agrupamiento de k caracteres. Dado que la tabla:

$$\begin{array}{c} k \\ 000 \ ... \ 000_{b_1} = 0_{b_1^k} \\ 000 \ ... \ 001_{b_1} = 1_{b_1^k} \\ ..... \\ 111 \ ... \ 111_{b_1} = (b_1^k - 1)_{b_1^k} \end{array}$$

da la representación en base b_1 de los enteros $0, 1, ..., (b_1^k - 1)$ representados en base b_1^k, se asocia a cada agrupamiento el correspondiente entero representado en base b_1^k, y el nuevo alineamiento es la representación en base b_1^k de n.

Tabla 4
$8 = 2^3$

$000_2 = 0_8$	$100_2 = 4_8$
$001_2 = 1_8$	$101_2 = 5_8$
$010_2 = 2_8$	$110_2 = 6_8$
$011_2 = 3_8$	$111_2 = 7_8$

Tabla 5
$16 = 2^4$

$0000_2 = 0_{16}$	$0100_2 = 4_{16}$	$1000_2 = 8_{16}$
$0001_2 = 1_{16}$	$0101_2 = 5_{16}$	$1001_2 = 9_{16}$
$0010_2 = 2_{16}$	$0110_2 = 6_{16}$	$1010_2 = A_{16}$
$0011_2 = 3_{16}$	$0111_2 = 7_{16}$	$1011_2 = B_{16}$
		$1100_2 = C_{16}$
		$1101_2 = D_{16}$
		$1110_2 = E_{16}$
		$1111_2 = F_{16}$

En las tablas 4 y 5 se dan las correspondencias entre las bases 2 y 8, y entre las bases 2 y 16. Esas tablas permiten la inmediata conversión de las representaciones de un número entero entre las bases 2, 8 y 16. Por ejemplo:

$$111010_2 = 111\ \ 010_2 = 72_8$$
$$11001101001010_2 = 0011\ \ 0011\ \ 0100\ \ 1010_2 = 334A_{16}$$
$$10101_2 = 010\ \ 101_2 = 25_8$$
$$11001101001010_2 = 011\ \ 001\ \ 101\ \ 001\ \ 010_2 = 31512_8$$
$$111010_2 = 0011\ \ 1010_2 = 3A_{16}$$

Consideramos ahora el caso en que z sea un número positivo menor que 1, del cual tenemos la representación en base b_1. Entonces, se descompone el alineamiento de los caracteres a la derecha del punto raíz, en agrupamientos de k caracteres, partiendo de la izquierda hacia la derecha, y completando el último agrupamiento (más a la derecha) con ceros. Como antes, usando la tabla de correspondencia entre la base b_1 y la base b_1^k, se asocia a cada agrupamiento la cifra correspondiente en base b_1^k. Por ejemplo, usando las tablas 4 y 5, obtenemos:

$$0.1_2 = 0.100_2 = 0.4_8$$
$$0.11111011_2 = 0.111\ \ 110\ \ 110_2 = 0.766_8\ .$$

En el caso de un número real cualquiera, X, del cual tenemos la representación en base b_1, se consigue la representación en base b_1^k aplicando los métodos dados a su parte entera y a su parte fraccionaria. Por ejemplo:

$$11101.101_2 = 011\ \ 101.101_2 = 35.5_8$$
$$11101.101_2 = 0001\ \ 1101.1010_2 = 1D.A_{16}$$

Y viceversa, si tenemos la representación del número X en base $b_2 = b_1^k$, obtenemos su representación en base b_1, sustituyendo cada cifra representada en base b_1^k, con el correspondiente agrupamiento de k caracteres obtenido de la tablas de correspondencia. Por ejemplo:

$$2A_{16} = 0010\ \ 1010_2 = 101010_2$$
$$-102_8 = -001\ \ 000\ \ 010_2 = -1000010_2$$
$$5FF.FB1_{16} = 0101\ \ 1111\ \ 1111.1111\ \ 1011\ \ 0001_2 = 10111111111.11111011_2\ .$$

Para convertir un número desde su representación octal a su representación hexadecimal, se ejecuta un proceso con dos pasos. Antes el número octal se convierte a binario, y después los dígitos binarios se dividen en agrupamientos de cuatro dígitos y convertidos a las cifras hexadecimales. Por ejemplo:

$$14057_8 = 001\ \ 100\ \ 000\ \ 101\ \ 111_2 = 0001\ \ 1000\ \ 0010\ \ 1111_2 = 182F_{16}\ .$$

Un procedimiento analogo se usa si se quiere convertir un número hexadecimal a su representación octal. Antes se escribe el número en su representación hexadecimal y se convierten los dígitos a los correspondientes agrupamientos de cuatro cifras binarias. Después esos dígitos binarios se dividen en agrupamientos de tres dígitos, y esos agrupamientos se escriben en cifras octales. Por ejemplo:

$$1F34B_{16} = 0001\ \ 1111\ \ 0011\ \ 0100\ \ 1011_2 =$$
$$= 011\ \ 111\ \ 001\ \ 101\ \ 001\ \ 011_2 = 371513_8\ .$$

EJERCICIOS.

1. Construir la representación en base $b = 10$ de los siguientes números:
   a) 1532_{16} ; $12F_{16}$;
   b) $82F_{16}$; $A2F_{16}$;
   c) $F2F_{16}$; $1AB_{16}$.

2. Construir la representación en base $b = 10$ de los siguientes números:
   a) 111_8 ; 117_8 ;
   b) 177_8 ; 577_8 ;
   c) 777_8 ; 1000_8 .

3. Dar la representación del número decimal 429_{10} en las bases:
   a) $b = 16$;
   b) $b = 8$;
   c) $b = 2$;
   d) $b = 3$.

4. Convertir el número $AF301_{16}$ a su representación en base $b = 8$.

5. Convertir el número 741731_8 a su representación en base $b = 16$.

6. Convertir el número $7214A_{16}$ a su representación en base $b = 8$.

7. Convertir el número 7214730152_8 a su representación en base $b = 16$.

8. Convertir el número 110111010111000010110111 a sus representaciones en base $b_1 = 8$ y $b_2 = 16$.

9. Convertir el número 101101110.11000010110111 a sus representaciones en base $b_1 = 8$ y $b_2 = 16$.

10. Calcular la suma, la diferencia, el producto, el cociente y el resto de la división entre los dos números:
   a) $n_1 = 11011_2$, $n_2 = 110_2$;
   b) $n_1 = 10101011_2$, $n_2 = 1010_2$.

11. Calcular el cociente y el resto de la división entre los dos números, y ejecutar en base $b = 8$ la prueba de la operación:
   a) $n_1 = 3541.123_8$, $n_2 = 3.52_8$;
   b) $n_1 = 73564.132_8$, $n_2 = 314.5_8$.

12. Calcular la diferencia entre los números enteros $F7A6_{16}$ y $BC43_{16}$.

13. Calcular el producto entre los números reales $40B.D_{16}$ y $2.A8_{16}$.

CAPITULO IV. ARITMETICA DEL COMPUTADOR

1. REPRESENTACION DE LOS NUMEROS

Cuando se usa una calculadora o una computadora digital para realizar cálculos numéricos, se debe considerar un error inevitable, el llamado error de redondeo. Este error se origina porque la aritmética realizada en una máquina involucra números con sólo un número finito de dígitos, con el resultado de que muchos cálculos se realizan con representaciones aproximadas de los números verdaderos.

El ordenador recibe, normalmente, información en decimal, que es transformada a binario por un programa interno. Posteriormente efectúa las operaciones pertinentes, pasa el resultado a decimal e informa al usuario de este resultado.

Así pues, en principio deberíamos hablar de las representación de los números en binario (la forma usual de trabajar del ordenador), pero para facilitar la comprensión, usaremos la representación decimal.

La representación de los números en el sistema decimal no es única (considerar que $0.999... = 1.000...$). Esto también es cierto para otros sistemas de numeración, y en particular para el sistema binario. Para evitar estas ambiguedades, siempre nos referiremos a la representación finita (1 en vez de 0.999...).

En general los ordenadores digitales trabajan con un número fijo (finito) de posiciones, **la longitud de palabra**, cuando representan un número internamente. Esta longitud n depende de la máquina, y además algunas permiten extensiones a múltiplos enteros de n ($2n$, $3n$, ...) que posibilitan una mayor exactitud si se necesita. Una palabra de longitud n se puede utilizar de distintas formas para representar un número:

- la **representación de punto fijo** especifica un número fijo n_1 de lugares enteros, y un número fijo n_2 de decimales, de modo que $n = n_1 + n_2$. En esta representación, la posición del punto decimal está fija y son pocos los dispositivos que la utilizan (ciertas máquinas de calcular, o máquinas de tipo comercial).

Ejemplo 1.

Sea $n = 8$, $n_1 = 3$ y $n_2 = 5$. En el sistema decimal se tienen las siguientes representaciones:

$$81.46 \longrightarrow \underline{| \ 0\ 8\ 1 \ | \ 4\ 6\ 0\ 0\ 0 \ |}$$

$$0.0002 \longrightarrow \underline{| \ 0\ 0\ 0 \ | \ 0\ 0\ 0\ 2\ 0 \ |} \ .$$

- Más importante, sobre todo en el cálculo cientifico, es la **representación en punto flotante**. La posición del punto decimal con respecto al primer dígito se expresa con un número separado, denominado exponente. Así se obtiene la notación científica:

$$x = a * b^t, \qquad \text{con} \qquad |a| < 1, \qquad b \in \mathcal{N}, \qquad t \in \mathcal{Z}$$

donde b es la **base** del sistema de numeración, t es un exponente llamado **característica** y a se llama la **mantisa**. Además si $|a| \geq b^{-1}$, es decir que el primer dígito despues del punto raiz no es cero, se dice que la representación es **normalizada**.

Naturalmente, en un ordenador digital sólo se dispone de un número finito de posiciones para representar un número (n, la longitud de la palabra), por lo que cada ordenador tendrá reservadas m posiciones para la mantisa y c posiciones para la característica ($n = m + c$).

Ejemplo 2.

Sea $m = 5$, $c = 2$ ($\Rightarrow n = 7$), y $b = 10$. El número 5420.0 se representaría:

$$0.54200 * 10^4 \longrightarrow \underline{| \ 5 \ 4 \ 2 \ 0 \ 0 \ | \ 0 \ 4 \ |} \ .$$

Esta notación no es única porque ese número se podría haber expresado también como:

$$0.05420 * 10^5 \longrightarrow \underline{| \ 0 \ 5 \ 4 \ 2 \ 0 \ | \ 0 \ 5 \ |} \ .$$

La primera notación es la representación normalizada.

Son **dígitos significativos** de un número todos los dígitos de la mantisa sin contar los primeros ceros.

Los números m y c, junto con la base b de la representación de un número, determinan un conjunto $\mathcal{A} \subset \mathcal{R}$ de números reales que se pueden representar de forma exacta en una máquina; estos números se denominan **números de la máquina**.

Ejemplo 3.

Imaginemos que una computadora pueda representar exactamente el número decimal 179.015625 , y que el siguiente número de máquina más pequeño sea 179.015609 , mientras que el número de máquina más grande siguiente es 179.015640 . Esto significa que nuestro número de máquina original debe representar no solamente a 179.015625, sino a un número infinito de números reales que estén entre este número y su número de máquina más cercano.

Hay diversos conceptos, relacionados con estas representaciones, que describiremos a continuación: truncamiento, redondeo, underflow y overflow.

Los dos primeros conceptos, **truncamiento** y **redondeo**, son relativos a la forma de representar los números que no pertenecen al conjunto $\mathcal{A}$ definido anteriormente. Supongamos que tenemos una máquina con m dígitos de mantisa y c dígitos de característica. Ya sabemos que el conjunto $\mathcal{A}$ de los números reales que se pueden representar exactamente es un conjunto finito. Sea entonces $x \in \mathcal{R}$, $x \notin \mathcal{A}$ (por ejemplo, por el mayor número de dígitos de mantisa),

$$x = a * b^t, \qquad \text{donde} \qquad b^{-1} \leq |a| < 1$$

$$a = 0.\alpha_1 \ \alpha_2 \ ...\alpha_m \ \alpha_{m+1} \ ..., \qquad 0 \leq \alpha_i \leq b-1, \qquad \alpha_1 \neq 0 \ .$$

Consideremos ahora

$$a' = \begin{cases} \pm 0.\alpha_1 \ \alpha_2 \ ...\alpha_m & 0 \leq \alpha_{m+1} \leq \frac{b}{2} - 1 \\ \pm 0.\alpha_1 \ \alpha_2 \ ...\alpha_m + b^{-m} & \frac{b}{2} \leq \alpha_{m+1} \leq b - 1 \end{cases} \qquad (IV.1a)$$

esto es, se suprimen los dígitos que sigan al último representable si $0 \leq \alpha_{m+1} \leq \frac{b}{2} - 1$, y se aumenta en una unidad α_m si $\alpha_{m+1} \geq \frac{b}{2}$, y después se suprimen los dígitos que sigan al último representable. Entonces, está claro que

$$fl(x) = a' * b^t \in \mathcal{A} \ . \qquad (IV.1b)$$

Por ejemplo, supongamos que se desease utilizar el número $0.34826 * 10^4$ en una máquina en que $m = 4$ y $c = 2$. El truncamiento consiste en suprimir todos los dígitos que existen tras el último representable, sin mirar cuál es el dígito siguiente; por el contrario, el redondeo suprime los dígitos que sigan al último representable si el siguiente es menor o igual que 4 o aumentan en una unidad el último representable, suprimiendo los restantes, si el que sigue es mayor o igual que 5. Así el número anterior se representaría como

$$\underline{|\ 3\ 4\ 8\ 2\ |\ 0\ 4\ |} \quad \text{con truncamiento ,}$$
$$\underline{|\ 3\ 4\ 8\ 3\ |\ 0\ 4\ |} \quad \text{con redondeo .}$$

La forma usual de utilizar los números es la última, porque se usa el número de la máquina que está más próximo al que se necesita.

Entonces, desde un número $x \notin \mathcal{A}$ se puede construir otro número $fl(x) \in \mathcal{A}$, naturalmente haciendo un error de redondeo. Para el error relativo de $fl(x)$ se tiene

$$\Big|\frac{fl(x) - x}{x}\Big| = \Big|\frac{a' * b^t - a * b^t}{a * b^t}\Big| = \Big|\frac{a' - a}{a}\Big| < \frac{\frac{b}{2}\ b^{-(m+1)}}{|a|} \leq \frac{b}{2} b^{-m} \ ,$$

siendo $|a| \geq b^{-1}$. En el caso de $b = 10$

$$\Big|\frac{fl(x) - x}{x}\Big| < \frac{5.0 * 10^{-(m+1)}}{|a|} \leq 5.0 * 10^{-m} \ ,$$

y si ponemos $\nu = 5.0 * 10^{-m}$, se puede poner que

$$fl(x) = x\ (1 + \varepsilon) \ , \qquad \text{donde} \ \ |\varepsilon| \leq \nu \ .$$

El valor ν se llama **precisión de la máquina**.

En ocasiones, el número x no puede ser representado por la máquina al efectuar el redondeo, como indicamos en los cuatro casos siguientes:

$m = 4, c = 2, b = 10$
$fl(0.31794 * 10^{110}) = 0.3179 * 10^{110} \notin \mathcal{A}$
$fl(0.99997 * 10^{99}) = 0.1000 * 10^{100} \notin \mathcal{A}$
$fl(0.012345 * 10^{-99}) = 0.1235 * 10^{-100} \notin \mathcal{A}$
$fl(0.54321 * 10^{-110}) = 0.5432 * 10^{-110} \notin \mathcal{A}$.

En los dos primeros casos, el exponente es demasiado grande para caber en los lugares reservados para él, y se produce un **overflow** (rebasamiento del valor máximo), y en los dos últimos casos el exponente es demasiado pequeño para caber en los lugares reservados

para él, y se produce un **underflow** (rebasamiento del valor mínimo). Los dos últimos casos tienen una posible solución, y es la de prevenirlos definiendo:

$fl(0.012345 * 10^{-99}) = 0.0123 * 10^{-99} \in \mathcal{A}$, (no normalizado)

$fl(0.54321 * 10^{-110}) = 0.0 \in \mathcal{A}$,

pero ahora el redondeo puede no verificar que

$$fl(x) = x\ (1 + \varepsilon)\ , \qquad \text{con}\ \ |\varepsilon| \leq \nu\ .$$

Los ordenadores digitales tratan los fenómenos de overflow y de underflow de forma diferentes, y siempre como irregularidades del cálculo. En cierto casos, al producirse alguno de los rebasamientos, el ordenador continúa los cálculos con el mayor valor permitido (o cero si se trata de un underflow) mostrando o no un mensaje de aviso de lo que ha ocurrido; en otros se muestra un mensaje de error y detiene el programa. Los rebasamientos pueden ser evitados si se hacen escalados adecuados de los datos, y si durante los cálculos se hacen chequeos, efectuando reescalados si fuese preciso.

El uso frecuente de la aritmética de redondeo en computadoras lleva a la siguiente definición: se dice que **el número p^* aproxima a p con m dígitos significativos** (o cifras) si m es el entero más grande no negativo para el cual

$$\left|\frac{p^* - p}{p}\right| \leq 5.0 * 10^{-m}\ .$$

La razón por la cual se usa el error relativo en la definición es que se desea obtener un concepto continuo. Por ejemplo, para que p^* aproxime a 1000 con cuatro cifras significativas, p^* debe satisfacer

$$\left|\frac{p^* - 1000}{1000}\right| \leq 5.0 * 10^{-4}\ , \qquad \text{y eso implica que} \qquad 999.5 \leq p^* \leq 1000.5\ .$$

Pasamos ahora a ver brevemente algunas **representaciones internas** usadas por el ordenador para almacenar los números enteros y los reales.

1. **Representación interna de números enteros en "magnitud-signo"**.
La escritura de un número en el sistema binario es la manera más sencilla de representarlo mediante un patrón de bits. La idea más simple para representar el signo menos es reservar un bit para ello, de forma que si dicho bit vale 0 el número es positivo, y si vale 1 es negativo. Normalmente se suele utilizar el bit situato más a la izquierda.

Ejemplo 4.

Representar los número enteros 17 y -17 en magnitud-signo con 8 bits.

La representación binaria del número 17 es 10001. Entonces, en la representación magnitud-signo con 8 bits del número positivo 17 tendremos que el primer bit, que es el que denota el signo, es cero: 00010001. Para el número negativo -17 se obtiene 10010001.

Ejemplo 5.

¿Qué números decimales representan las series de 8 bits 11100010 y 00111011 codificados en magnitud-signo?

La primera serie 11100010 tiene un uno en el primer bit, indicando que el número representado es un número negativo. Los demás dígitos son la representación binaria del número

$$1100010 = 2 + 32 + 64 = 98 \ .$$

Entonces la primera serie representa al número entero -98.

La segunda serie 00111011 tiene un cero en el primer bit, indicando que el número representado es un número positivo. Los demás dígitos son la representación binaria del número

$$111011 = 1 + 2 + 8 + 16 + 32 = 59 \ .$$

Entonces la segunda serie representa al número entero 59.

2. Representación interna de números enteros en "notación en exceso".
La representación en magnitud-signo es una manera muy natural de codificar números en binario. Sin embargo, hay otras formas de codificación que permiten diseñar circuitos electrónicos más simples para interpretarlas. Una de éstas es la **notación en exceso**. La representación en exceso con p bits de un número decimal entero N consiste en codificar N como el equivalente binario del número $N + 2^{p-1}$, que se denomina **característica** de N con p bits.

Por ejemplo, la tabla siguiente muestra la notación en exceso con 4 bits

$0000 = -8$	$1000 = 0$
$0001 = -7$	$1001 = 1$
$0010 = -6$	$1010 = 2$
$0011 = -5$	$1011 = 3$
$0100 = -4$	$1100 = 4$
$0101 = -3$	$1101 = 5$
$0110 = -2$	$1110 = 6$
$0111 = -1$	$1111 = 7$

El nombre "notación en exceso" se debe a la diferencia que hay entre el número codificado y el número binario directo que representa el patrón de bits. Nótese que a diferencia de la codificación en magnitud-signo, en la notación en exceso los números positivos tienen el primer bit igual a 1, mientras que los números negativos tienen el primer bit igual a 0.

Ejemplo 6.

Representar en exceso con 8 bits los números enteros 23 y -49.

Para dar la representación en exceso del número 23 tenemos que hallar la representación binaria del número $23 + 2^{8-1} = 23 + 2^7 = 23 + 128 = 151$. Tal representación es 10010111 que coincide con la representación en exceso con 8 bits de 23. De manera parecida, para hallar la representación en exceso del número -49 tenemos que hallar la representación binaria del número $-49 + 2^{8-1} = -49 + 2^7 = -49 + 128 = 79$. Tal representación es 1001111, y ahora para tener la representación en exceso se necesitan añadir ceros a la izquierda; entonces, la representación en exceso con 8 bits de -49 es 01001111.

Ejemplo 7.

¿Qué números decimales representan los códigos en exceso 01100111 y 10010011?

Para el primer código: $01100111 = 1+2+4+32+64 = 103$, entonces $N+128 = 103$ lo cual implica $N = -25$.

Para el segundo código: $10010011 = 1+2+16+128 = 147$, entonces $N+128 = 147$ lo cual implica $N = 19$.

3. Representación interna de números enteros en "complemento a dos".

La representación en **complemento a dos** es una manera muy útil de codificar un número debido a que facilita enormemente las operaciones algebraicas. Para obtener el complemento a dos de un número binario hay que considerar en primer lugar el complemento a uno cuya definición es la siguiente: el **complemento a uno** de un número binario es el número que se obtiene al cambiar los ceros por unos y los unos por ceros. Conocido el complemento a uno, el complemento a dos se obtiene fácilmente: el **complemento a dos** de un número binario se obtiene sumando 1 al complemento a uno.

Ahora, la **representación en complemento a dos** de un número consiste en escribir los números positivos como su equivalente en el sistema binario, y los números negativos como el complemento a dos del equivalente en el sistema binario de su valor absoluto.

Para decodificar un número decimal representado en complemento a dos se procede del modo siguiente:

- si el primer bit de la izquierda es 0 el número es positivo. Entonces, el número representado es el equivalente del número binario que forma el resto de los bits.
- si el primer bit de la izquierda es 1 el número es negativo. Entonces el número representado es el opuesto del equivalente decimal del número binario que forma su complemento a dos.

Ejemplo 8.

Representar con 8 bits en complemento a dos los números decimales 17 y -17.

La representación binaria del número 17 es 10001, entonces la representación con 8 bits en complemento a dos de 17 se obtiene añadiendo ceros a la izquierda: 00010001.

Para la representación con 8 bits en complemento a dos de -17 tenemos que realizar el complemento a dos de la representación binaria del su valor absoluto 17. Primero se pasa de 00010001 a su complemento a uno: 11101110. Ahora, se hace el complemento a dos, es decir se le suma 1: $11101110 + 1 = 11101111$. Esta es la representación con 8 bits en complemento a dos de -17.

Ejemplo 9.

¿Qué números decimales representan las series de 8 bits 00101011 y 10101011 codificadas en complemento a dos?

La primera serie 00101011 tiene un cero en el primer bit, indicando que el número representado es un número positivo. Los demás dígitos son la representación binaria del número 43 ($101011 = 1+2+8+32 = 43$). Entonces la primera serie representa al número entero 43.

La segunda serie 10101011 tiene un uno en el primer bit, indicando que el número representado es un número negativo y que entonces tenemos que realizar la operación de complemento a dos. Es decir, primero tenemos que pasar la representación 10101011 a complemento a uno: 01010100 y ahora a complemento a dos añadiendo uno: $01010100 + 1 = 01010101$. Finalmente el número buscado es el opuesto del equivalente decimal: $01010101 = 1 + 4 + 16 + 64 = 85$, es decir -85.

4. Representación interna de números reales en "punto flotante".

Los números fraccionarios y reales se introducen en el ordenador **en punto flotante**. Esta representación consiste en escribirlos en forma exponencial binaria normalizada y codificar tres campos: el signo, el exponente y la matisa. Cada uno de los campos se codifica de la manera siguiente:

1. El bit de signo se pone a 0 cuando el número es positivo y a 1 cuando el número es negativo.
2. El campo exponente se codifica usualmente mediante la notación en exceso.
3. El campo mantisa se codifica como el equivalente binario directo del número decimal dado.

Si se usan 32 bits para la representación pueden dividirse del modo siguiente: 1 bit para el signo, 7 bits para el exponente y 24 bits para la mantisa:

1 bit (signo) | 7 bits (exponente) | 24 bits (mantisa)

Ejemplo 10.

Representar en punto flotante con 32 bits, 1 de signo, 7 de exponente y 24 de mantisa, los números decimales 104.3125 y -13506.96875.

El primer número 104.3125 es positivo, entonces el primer bit será un cero. La representación binaria del número es: 1101000.0101, cuya forma exponencial normalizada es 0.11010000101×2^7. El exponente (7) se codifica en exceso con 7 bits: $7+2^{7-1} = 7+2^6 = 7+64 = 71$ cuya represenatción binaria es 1000111. Finalmente, la mantisa tiene 11 bits (11010000101) y se completa con 13 ceros a la derecha. Entonces, la representación en punto flotante con 32 bits del número 104.3125 es

$$0 \mid 1000111 \mid 11010000101000000000000 \ .$$

De manera parecida, para el segundo número -13506.96875, notamos que es negativo, entonces el primer bit será un uno. La representación binaria del valor absoluto del número es: 11010011000010.11111. Su forma exponencial normalizada es $0.1101001100001011111 \times 2^{14}$. El exponente (14) se codifica en exceso con 7 bits: $14 + 2^{7-1} = 14 + 2^6 = 14 + 64 = 78$ cuya represenatción binaria es 1001110. Finalmente, la mantisa tiene 19 bits (1101001100001011111) y se completa con 5 ceros a la derecha. Entonces, la representación en punto flotante con 32 bits del número 104.3125 es

$$1 \mid 1001110 \mid 11010011000010111111000000 \ .$$

2. INTRODUCCION A LA ARITMETICA DE PUNTO FLOTANTE

Además de dar una representación inexacta de los números, la aritmética realizada en la computadora no es exacta. Sin embargo, usando números con representación en punto flotante con m dígitos de mantisa, las operaciones aritméticas elementales no se pueden siempre ejecutar de manera exacta, y los resultados de las operaciones no necesariamente son números de la máquina aunque los operandos lo sean. Por ello no se puede esperar reproducir de forma exacta las operaciones aritméticas en un ordenador digital. Deberemos contentarnos con sustituirlas por otras $(\oplus, \ominus, \otimes, \oslash)$ llamadas **operaciones de punto flotante**, que las aproximen tanto como sea posible. Esto se puede conseguir, por ejemplo, definiéndolas con la ayuda del redondeo:

suma: $x \oplus y = fl(fl(x) + fl(y))$
resta: $x \ominus y = fl(fl(x) - fl(y))$
multiplicación: $x \otimes y = fl(fl(x) * fl(y))$
división: $x \oslash y = fl(fl(x)/fl(y))$.

Esta aritmética idealizada corresponde a efectuar la aritmética exacta en la representación del punto flotante de x e y, y luego a la conversión del resultado exacto a su representación de punto flotante. Se pueden probar las relaciones

$$x \oplus y = (x + y)\ (1 + \varepsilon_1) \qquad (IV.2a)$$

$$x \ominus y = (x - y)\ (1 + \varepsilon_2) \qquad (IV.2b)$$

$$x \otimes y = (x * y)\ (1 + \varepsilon_3) \qquad (IV.2c)$$

$$x \oslash y = (x/y)\ (1 + \varepsilon_4)\ , \qquad (IV.2d)$$

donde $|\varepsilon_i| \leq \mu_i\ \nu$, y μ_i es un entero $\mu_i \geq 1$, que depende del tipo de máquina usada.

Podemos comprobar que las operaciones en punto flotante no verifican las reglas aritméticas normales:

a) $x \oplus y = x$ **no implica que** $y = 0$. Esta igualdad es cierta para todo y tal que $|y| < \frac{\nu}{b}|x|$ (b es la base del sistema de numeración usado). Según esto, la precisión de la máquina, ν, debería definirse como el menor número positivo g de la máquina, para el cual se cumple $1 \oplus g > 1$, $\nu = \min\{g \in \mathcal{A} \ / \ 1 \oplus g > 1 \ \text{ y } \ g > 0\}$.

Ejemplo 11.
$m = 3, c = 2, b = 10$
$x = 0.123 * 10^0$
$y = 0.000061 = 0.61 * 10^{-4} \leq 0.5 * 10^{-3} * 0.123 = 0.615 * 10^{-4}$
Entonces
$x \oplus y = x$

b) **no asociatividad**: $a \oplus (b \oplus c)$ puede ser diferente de $(a \oplus b) \oplus c$.

Ejemplo 12.
$m = 8, c = 2, b = 10$
$a = 0.23371258 * 10^{-4}$

$b = 0.33678429 * 10^2$
$c = -0.33677811 * 10^2$

Entonces

$a \oplus (b \oplus c) = 0.23371258 * 10^{-4} \oplus 0.61800000 * 10^{-3} = 0.64137126 * 10^{-3}$
$(a \oplus b) \oplus c = 0.33678452 * 10^2 \ominus 0.33677811 * 10^2 = 0.64100000 * 10^{-3}$

y el resultado exacto es

$a + b + c = 0.64137126 * 10^{-3}$.

Esto ha ocurrido porque cuando se restan dos números del mismo signo, se produce un efecto de cancelación si ambos coinciden en uno o más dígitos con respecto al mismo exponente (los dígitos comunes desaparecen). A pesar de que, cuando $x, y \in \mathcal{A}$, su diferencia también es un elemento de $\mathcal{A}$, y por lo tanto no hay errores de redondeo adicionales, veremos que la cancelación es un efecto peligroso cuando se trata de propagación de errores previos (cuando x e y provienen de cálculos que han necesitado redondeo).

c) **no distributividad**: $a \otimes (b \oplus c)$ puede ser diferente de $(a \otimes b) \oplus (a \otimes c)$.

Ejemplo 13.

$m = 2, c = 2, b = 10$
$a = 0.94 * 10^2$
$b = 0.33 * 10^2$
$c = -0.32 * 10^2$

Entonces

$a \otimes (b \oplus c) = 0.94 * 10^2 \otimes 0.1 * 10 = 0.94 * 10^2$
$(a \otimes b) \oplus (a \otimes c) = 0.31 * 10^4 \ominus 0.30 * 10^4 = 0.1 * 10^3$

y el resultado exacto es

$a * (b + c) = 0.94 * 10^2$.

Las operaciones aritméticas $+, -, *, /$, junto a las funciones para las que se hayan especificado sustituciones (por ejemplo raíz cuadrada, funciones trigonometricas, etc...) se llaman **funciones elementales**.

3. PROPAGACION DEL ERROR

Hemos comprobado que la no associatividad de la suma y la no distributividad del producto en un ordenador pueden provocar la obtención de resultados diferentes dependiendo de la técnica que se utilice para efectuar las operaciones. Entonces la propagación del error es un efecto muy importante a tener en cuenta, y se debe evitar en lo posible.

Generalmente, un problema matemático puede ser esquematizado en la manera siguiente: con un número finito de datos iniciales $x_1, x_2, ..., x_n \in \mathcal{R}$ queremos calcular un número finito $y_1, y_2, ..., y_m \in \mathcal{R}$ de resultados. Eso corresponde a asignar una función

$$\phi^{(i)}: D_i \to D_{i+1}, \qquad i = 0, ..., r, \qquad D_j \subseteq \mathcal{R}^{n_j} \tag{IV.3}$$

donde $\phi = \phi^{(r)} \circ \phi^{(r-1)} \circ ... \circ \phi^{(0)}$ y $D_0 = D \subset \mathcal{R}^n$, $D_{r+1} \subseteq \mathcal{R}^{n_r+1} \equiv \mathcal{R}^m$.

Entonces el problema es analizar como un error Δx y los errores de redondeo que se hacen en el cálculo se propagan y cambian el resultado final $y = \phi(x)$. Consideremos

$$\phi\colon D \subset \mathcal{R}^n \to \mathcal{R}^m, \qquad \phi(x) = \begin{pmatrix} \phi_1(x_1, \ldots, x_n) \\ \vdots \\ \phi_m(x_1, \ldots, x_n) \end{pmatrix}$$

con las funciones componentes continuas y con derivadas primeras continuas. Para hacer los calculos más sencillos, analicemos antes sólo la propagación del error sobre el dato inicial Δx, con un procedimiento del primer orden (si ε y η son dos números muy pequeños, entonces el producto $\varepsilon\ \eta$ se puede despreciar con respecto a ε y η).

Si x^* es una aproximación de x, los errores absolutos serán

$$\Delta x_i = x_i^* - x_i, \qquad \Delta x = x^* - x, \qquad \Delta y_i = \phi_i(x^*) - \phi_i(x)\ .$$

Si usamos el desarrollo en serie de Taylor hasta al primer orden

$$\Delta y_i = y_i^* - y_i = \phi_i(x^*) - \phi_i(x) \approx \sum_{j=1}^{n} (x_j^* - x_j)\ \frac{\partial \phi_i(x)}{\partial x_j} = \sum_{j=1}^{n} \Delta x_j\ \frac{\partial \phi_i(x)}{\partial x_j}\ , \qquad (IV.4a)$$

ó en notación matricial

$$\Delta y = \begin{pmatrix} \Delta y_1 \\ \vdots \\ \Delta y_m \end{pmatrix} \approx \begin{pmatrix} \frac{\partial \phi_1(x)}{\partial x_1} & \cdots & \frac{\partial \phi_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial \phi_m(x)}{\partial x_1} & \cdots & \frac{\partial \phi_m(x)}{\partial x_n} \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{pmatrix} = D\phi(x) \cdot \Delta x\ , \qquad (IV.4b)$$

con $D\phi(x)$ la matriz Jacobiana. Aquí el factor de proporcionalidad $\frac{\partial \phi_i(x)}{\partial x_j}$ mide la sensibilidad con la cual y "reacciona" a las variaciones absolutas Δx_j de x_j. La fórmula análoga para la propagación de los errores relativos es:

$$RE_{y_i} \approx \sum_{j=1}^{n} \frac{x_j}{\phi_i(x)}\ \frac{\partial \phi_i(x)}{\partial x_j}\ RE_{x_j} = \sum_{j=1}^{n} \frac{\Delta x_j}{\phi_i(x)}\ \frac{\partial \phi_i(x)}{\partial x_j}\ . \qquad (IV.5)$$

Aquí el factor $\frac{x_i}{\phi(x)}\ \frac{\partial \phi(x)}{\partial x_i}$ (a menudo se le conoce como **índice de condicionamento**) indica cómo el error relativo en x_i repercute en el error relativo de y. Si el índice de condicionamento es de valor absoluto suficientemente grande, errores relativos pequeños en los datos iniciales producen errores relativos muy grandes en los resultados. En ese caso se dice que el problema está **mal planteado**.

La propagación del error relativo en las operaciones elementales viene dada por:

1. $\phi(x, y) = x * y \quad \Rightarrow \quad RE_{x*y} \approx RE_x + RE_y$
2. $\phi(x, y) = x/y \quad \Rightarrow \quad RE_{x/y} \approx RE_x - RE_y$
3. $\phi(x, y) = x \pm y \quad \Rightarrow \quad RE_{x\pm y} \approx \frac{x}{x\pm y}\ RE_x \pm \frac{y}{x\pm y}\ RE_y$
4. $\phi(x) = \sqrt{x} \quad \Rightarrow \quad RE_{\sqrt{x}} \approx \frac{1}{2}\ RE_x\ .$

Consideremos $\phi(x) = \sqrt{x}$. Entonces $\phi'(x) = \frac{1}{2\sqrt{x}}$ y el error relativo es:

$$\frac{|\phi(x) - \phi(x^*)|}{|\phi(x)|} \approx |\phi'(x^*)| \left| \frac{x - x^*}{\phi(x)} \right| = \frac{1}{2} \left| \frac{x - x^*}{\sqrt{x x^*}} \right| \approx \frac{1}{2} \left| \frac{x - x^*}{x} \right| ,$$

por lo que el error relativo en $\phi(x^*)$ es aproximadamente la mitad del error relativo en x^*, y por lo tanto, la operación de calcular la raíz cuadrada es, desde el punto de vista del error relativo, una operación segura.

Es más, también en la multiplicación, división y extracción de raíz, los errores relativos en los datos iniciales no se notan de manera fuerte en el resultado. Eso pasa también en la suma si los operandos x e y tienen el mismo signo: los indices de condicionamento $x/(x+y)$, $y/(x+y)$ tienen un valor entre cero y uno, y su suma es uno, luego

$$|RE_{x+y}| \leq \max(RE_x, RE_y) .$$

Si en la operación de suma los operandos x e y tienen signo contrario, por lo menos uno de los factores $x/(x+y)$, $y/(x+y)$, es mayor que uno, y entonces, por lo menos uno de los errores relativos RE_x, RE_y es mayor. Esa amplificación del error es todavía mayor si $x \approx -y$, porque en ese caso en la expresión de $x + y$ los dos terminos se cancelan.

Ejemplo 14.

Queremos estudiar el error obtenido para hallar la suma

$$\phi(\alpha, \beta, \gamma) = \alpha + \beta + \gamma$$

con $\phi : \mathcal{R}^3 \to \mathcal{R}$.

Para el calculo de ϕ se pueden usar los dos algoritmos:

Algoritmo 1	Algoritmo 2
$\eta = \alpha + \beta$	$\eta = \beta + \gamma$
$y = \phi(\alpha, \beta, \gamma) = \eta + \gamma$	$y = \phi(\alpha, \beta, \gamma) = \alpha + \eta$.

Las decomposiciones ($IV.3$) de ϕ en este caso son:

$$\phi^{(0)} : \mathcal{R}^3 \to \mathcal{R}^2 , \qquad \phi^{(1)} : \mathcal{R}^2 \to \mathcal{R} .$$

Entonces los algoritmos son:

Algoritmo 1	Algoritmo 2
$\phi^{(0)}(\alpha, \beta, \gamma) = \begin{pmatrix} \alpha + \beta \\ \gamma \end{pmatrix} \in \mathcal{R}^2$	$\phi^{(0)}(\alpha, \beta, \gamma) = \begin{pmatrix} \beta + \gamma \\ \alpha \end{pmatrix} \in \mathcal{R}^2$
$\phi^{(1)}(u, v) = u + v \in \mathcal{R}$	$\phi^{(1)}(u, v) = u + v \in \mathcal{R}$.

Usando el cálculo en punto flotante, ($IV.2$), se obtiene para el primer algoritmo:

$\eta = fl(\alpha + \beta) = (\alpha + \beta)\ (1 + \varepsilon_1)$

$\overline{y} = fl(\eta + \gamma) = (\eta + \gamma)\ (1 + \varepsilon_2) = [(\alpha + \beta)\ (1 + \varepsilon_1) + \gamma]\ (1 + \varepsilon_2)$

$= \alpha + \beta + \gamma + (\alpha + \beta)\ \varepsilon_1 + (\alpha + \beta + \gamma)\ \varepsilon_2 + (\alpha + \beta)\ \varepsilon_1\ \varepsilon_2 =$

$= (\alpha+\beta+\gamma)\left[1+\frac{(\alpha+\beta)}{\alpha+\beta+\gamma}\ \varepsilon_1\ (1+\varepsilon_2)+\varepsilon_2\right]$

Y para el error relativo

$$RE_y = \left|\frac{y-\overline{y}}{y}\right| = \left|\frac{\alpha+\beta}{\alpha+\beta+\gamma}\ \varepsilon_1\ (1+\varepsilon_2)+\varepsilon_2\right|$$

Y despreciando los términos de orden superior (procedimento del primer orden):

$$RE_y \approx \left|\frac{\alpha+\beta}{\alpha+\beta+\gamma}\ \varepsilon_1+\varepsilon_2\right| .$$

Si hubiésemos usado el segundo algoritmo, tendríamos:

$$RE_y \approx \left|\frac{\beta+\gamma}{\alpha+\beta+\gamma}\ \varepsilon_1+\varepsilon_2\right| .$$

Los factores de amplificación $\frac{\alpha+\beta}{\alpha+\beta+\gamma}$ y $\frac{\beta+\gamma}{\alpha+\beta+\gamma}$, respectivamente, y 1, indican cómo los errores de redondeo ε_1 y ε_2 influyen sobre el error relativo RE_y del resultado. Dependiendo de cuál de las dos cantitades $(\alpha+\beta)$ o $(\beta+\gamma)$ es menor, se prefiere uno u otro algoritmo. En el caso del ejemplo 12:

$$\frac{\alpha+\beta}{\alpha+\beta+\gamma} \approx 0.5 * 10^5 \qquad \frac{\beta+\gamma}{\alpha+\beta+\gamma} \approx 0.97 .$$

Y eso explica la mayor precisión del segundo algoritmo.

Por lo que concierne a la propagación del error relativo, desde la relación $(IV.5)$, se tiene:

$$RE_y \approx \frac{\alpha}{\alpha+\beta+\gamma}\ RE_\alpha + \frac{\beta}{\alpha+\beta+\gamma}\ RE_\beta + \frac{\gamma}{\alpha+\beta+\gamma}\ RE_\gamma .$$

Y se puede decir que el problema está bien planteado si cada sumando α, β, γ es pequeño con respecto a $(\alpha+\beta+\gamma)$.

Ejemplo 15.

Sabemos que las raices de $a\ x^2 + b\ x + c = 0$, cuando $a \neq 0$, son:

$$x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a} \qquad x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a} .$$

Consideremos la ecuación cuadrática

$$x^2 + 62.10\ x + 1 = 0$$

con raíces aproximadas:

$$x_1 = -0.01610723 \quad \text{y} \quad x_2 = -62.08390 .$$

Para esta ecuación, b^2 es mucho mayor que $4ac$, así que en el cálculo de x_1 y x_2 el numerador involucra la sustracción de números casi iguales. Supongamos que efectuamos los calculos para x_1 usando aritmética de redondeo con cuatro dígitos.

$$\sqrt{b^2-4ac} = \sqrt{(62.10)^2-4.000} = \sqrt{3856.-4.000} = \sqrt{3852.} = 62.06 ,$$

así que

$$fl(x_1) = \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-62.10 + 62.06}{2.000} = \frac{-0.040}{2.000} = -0.020$$

es una representación bastante pobre de $x_1 = -0.01611$ ($RE_{x_1} \approx 0.2415$). Por otro lado, los cálculos para x_2 implican la adición de dos números casi iguales, $-b$ y $-\sqrt{b^2 - 4ac}$, y no presentan ningún problema.

$$fl(x_2) = \frac{-b - \sqrt{b^2 - 4ac}}{2a} = \frac{-62.10 - 62.06}{2.000} = \frac{-124.2}{2.000} = -62.10$$

es una aproximación precisa de $x_2 = -62.08$ ($RE_{x_2} \approx 0.0003222$).

Para obtener una aproximación más exacta de x_1, aún con redondeo de cuatro dígitos, cambiamos la forma de la fórmula cuadrática **racionalizando** el numerador. Entonces:

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}\Big(\frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}}\Big) = \frac{-2c}{b + \sqrt{b^2 - 4ac}}$$

y desde luego

$$fl(x_1) = \frac{-2.000}{62.10 + 62.06} = \frac{-2.000}{124.2} = -0.0161\ .$$

La técnica de racionalización se puede aplicar para obtener una forma alternativa también para x_2

$$x_2 = \frac{-2c}{b - \sqrt{b^2 - 4ac}}\ .$$

Esta sería la expresión a usar si b fuera un número negativo. En nuestro problema, sin embargo, el uso de esta fórmula resulta no sólamente en la sustracción de números casi iguales, sino también en la división entre el resultado pequeño de esta sustracción. La inexactitud que esto produce es dramática:

$$fl(x_2) = \frac{-2c}{b - \sqrt{b^2 - 4ac}} = \frac{-2.000}{62.10 - 62.06} = \frac{-2.000}{0.040} = -50.00\ .$$

Para comprender mejor ese "mal" resultado, hagamos

$$y = \phi(p, q) = p - \sqrt{p^2 + q}, \qquad p > 0$$

y determinemos el error relativo que se propaga en y. Dado que

$$\frac{\partial \phi}{\partial p} = 1 - \frac{p}{\sqrt{p^2 + q}} = \frac{-y}{\sqrt{p^2 + q}} \qquad \text{y} \qquad \frac{\partial \phi}{\partial q} = \frac{-1}{2\sqrt{p^2 + q}}$$

se sigue que

$$RE_y \approx \frac{p}{y}\Big(\frac{-y}{\sqrt{p^2 + q}}\Big)\ RE_p + \frac{q}{y}\Big(\frac{-1}{2\sqrt{p^2 + q}}\Big)\ RE_q =$$

$$= \frac{-p}{\sqrt{p^2 + q}}\ RE_p - \frac{q}{2y\sqrt{p^2 + q}}\ RE_q =$$

$$= \frac{-p}{\sqrt{p^2+q}}\ RE_p + \frac{p+\sqrt{p^2+q}}{2\sqrt{p^2+q}}\ RE_q\ .$$

Dado que, si $q \geq 0$:

$$\left|\frac{p}{\sqrt{p^2+q}}\right| \leq 1 \qquad \text{y} \qquad \left|\frac{p+\sqrt{p^2+q}}{2\sqrt{p^2+q}}\right| \leq 1\ ,$$

entonces ϕ está bien planteada si $q > 0$, y mal planteada si $q \approx -p^2$. Además, si $|q|$ es muy pequeño con respecto a p^2, obtenemos el fenómeno de la cancelación, por el cual los errores de redondeo en el cálculo previo se amplifican notablemente.

Veamos ahora la forma en que se propagan los errores en algunos casos, para poder deducir la forma correcta en que deberían realizarse las operaciones.

a) $S = a_1 + a_2 + a_3 + a_4 + a_5$

$$fl[(((((a_1 + a_2) + a_3) + a_4) + a_5)] = fl(fl(fl(fl(a_1 + a_2) + a_3) + a_4) + a_5) =$$

$$= ((((a_1 + a_2)(1+\varepsilon_2) + a_3)(1+\varepsilon_3) + a_4)(1+\varepsilon_4) + a_5)(1+\varepsilon_5) =$$

$$(a_1 + a_2 + a_3 + a_4 + a_5)(1+\delta)$$

y hay que acotar

$$|\delta| = \left|\frac{S - fl[...]}{S}\right|\ ,$$

sabiendo que $|\varepsilon_i| < \nu$.

$$\begin{aligned} fl[...] = {} & a_1\ (1+\varepsilon_2)\ (1+\varepsilon_3)\ (1+\varepsilon_4)\ (1+\varepsilon_5) + \\ & a_2\ (1+\varepsilon_2)\ (1+\varepsilon_3)\ (1+\varepsilon_4)\ (1+\varepsilon_5) + \\ & a_3\ (1+\varepsilon_3)\ (1+\varepsilon_4)\ (1+\varepsilon_5) + \\ & a_4\ (1+\varepsilon_4)\ (1+\varepsilon_5) + \\ & a_5\ (1+\varepsilon_5) \approx \\ \approx {} & a_1 + a_2 + a_3 + a_4 + a_5 + \\ & a_1\ (\varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5) + a_2\ (\varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5) + \\ & a_3\ (\varepsilon_3 + \varepsilon_4 + \varepsilon_5) + a_4\ (\varepsilon_4 + \varepsilon_5) + a_5\ \varepsilon_5\ , \end{aligned}$$

donde se han desestimado los sumandos $\varepsilon_i\ \varepsilon_j$ que son despreciables respecto a ε_i.

Si ahora consideramos la situación más desfavorable (todos los ε_i iguales en signo y con el mayor valor absoluto ν), tenemos la siguiente acotación:

$$|\delta| \leq \left|\frac{4\ \nu\ a_1}{S}\right| + \left|\frac{4\ \nu\ a_2}{S}\right| + \left|\frac{3\ \nu\ a_3}{S}\right| + \left|\frac{2\ \nu\ a_4}{S}\right| + \left|\frac{\nu\ a_5}{S}\right| =$$

$$= \frac{\nu}{S}\left[|4\ a_1| + |4\ a_2| + |3\ a_3| + |2\ a_4| + |a_5|\right]\ .$$

En general, al sumar progresivamente $a_1 + a_2 + ... + a_n$, el error relativo máximo que se comete es, aproximadamente:

$$\left|\frac{S - fl[...]}{S}\right| \leq \frac{\nu}{S}\left[(n-1)\ (|a_1| + |a_2|) + (n-2)\ |a_3| + ... + 2\ |a_{n-1}| + |a_n|\right]\ .$$

Está claro, pues, que esta acotación es menor si los números $a_1, ..., a_n$ se ordenan de menor a mayor antes de sumarlos. Obtenemos así la siguiente regla práctica:
si se desea hallar $\sum_{i=0}^{n} a_i$, con n grande, y se trata de una serie convergente, entonces $\lim_{n\to\infty} a_i = 0$, y debe efectuarse la suma en orden inverso.

b) Se pretende calcular $\sum_{i=1}^{4} x_i\ y_i$. El resultado que se obtiene es:

$$\begin{aligned} fl\Big[\sum_{i=1}^{4} x_i\ y_i\Big] &= \Big(\Big\{\big[x_1\ y_1\ (1+\delta_1) + x_2\ y_2\ (1+\delta_2)\big]\ (1+\delta_5) + \\ &\qquad x_3\ y_3\ (1+\delta_3)\Big\}\ (1+\delta_6) + x_4\ y_4\ (1+\delta_4)\Big)\ (1+\delta_7) = \\ &= x_1\ y_1\ (1+\delta_1)\ (1+\delta_5)\ (1+\delta_6)\ (1+\delta_7) + \\ &\quad x_2\ y_2\ (1+\delta_2)\ (1+\delta_5)\ (1+\delta_6)\ (1+\delta_7) + \\ &\quad x_3\ y_3\ (1+\delta_3)\ (1+\delta_6)\ (1+\delta_7) + \\ &\quad x_4\ y_4\ (1+\delta_4)\ (1+\delta_7)\ . \end{aligned} \qquad (IV.6)$$

Observamos la falta de simetría en los resultados, debida a la no conmutatividad y no asociatividad de las operaciones de punto flotante.

Para simplificar la expresión anterior, vamos a obtener unas cotas manejables de los productos $(1+\delta_i)$.

Lema. Si $|\delta_i| \le u, i = 1, ..., n$ y $n\ u \le 0.01$, entonces

$$\prod_{i=1}^{n} (1+\delta_i) \le 1 + 1.01\ n\ u\ .$$

Demostración. Antes, consideremos $0 \le x \le 0.01$, y entonces

$$1 + x \le e^x \le 1 + 1.01\ x\ .$$

La primera desigualdad es inmediata, así que sólo veremos la segunda:

$$\begin{aligned} e^x = \sum_{r=0}^{\infty} \tfrac{x^r}{r!} &= 1 + x\ \big(1 + \tfrac{x}{2} + \tfrac{x^2}{3!} + ... + \tfrac{x^n}{(n+1)!} + ...\big) \le \\ &\le 1 + x\ \big(1 + \tfrac{x}{2} + \tfrac{x^2}{4} + ... + \tfrac{x^n}{2^n} + ...\big) \le \\ &\le 1 + x\ \big(1 + x\ (\tfrac{1}{2} + \tfrac{x}{4} + ...)\big) \le \\ &\le 1 + x\ (1+x) \le 1 + 1.01\ x\ . \end{aligned}$$

Entonces, si $n \in \mathcal{N}$ y $0 \le n\ u \le 0.01$

$$(1+u)^n \le (e^u)^n = e^{n\ u} \le 1 + 1.01\ n\ u\ .$$

Ahora está claro que

$$\prod_{i=1}^{n} (1+\delta_i) \le \prod_{i=1}^{n} (1+u) = (1+u)^n \le 1 + 1.01\ n\ u\ .$$

c.q.d.

El resultado de este lema puede expresarse también como:

$$\prod_{i=1}^{n}(1+\delta_i) = 1 + 1.01\ n\ \theta\ u\ , \qquad \text{donde} \qquad |\theta| \leq 1.$$

Entonces, volviendo a (*IV*.6), y suponiendo que $n\ u \leq 0.01$ (lo que se cumple en todas las situaciones reales):

$$fl\Big(\sum_{i=1}^{n} x_i\ y_i\Big) = x_1\ y_1\ (1 + 4.04\ \theta_1\ u) + x_2\ y_2\ (1 + 4.04\ \theta_2\ u) + x_3\ y_3\ (1 + 3.03\ \theta_3\ u) + x_4\ y_4\ (1 + 2.02\ \theta_4\ u)\ , \qquad |\theta_i| \leq 1\ .$$

En general se verifica:

Teorema: Si $n\ u \leq 0.01$ entonces

$$fl\Big(\sum_{i=1}^{n} x_i\ y_i\Big) = x_1\ y_1\ (1 + 1.01\ n\ \theta_1\ u) + \sum_{i=2}^{n} x_i\ y_i\ (1 + 1.01\ (n+2-i)\ \theta_i\ u)\ , \qquad |\theta_i| \leq 1\ .$$

Muchos computadores tienen la ventaja de que en la evaluación de productos escalares $\sum_{i=1}^{n} x_i\ y_i$, pueden ir acumulando los productos parciales $x_1\ y_1$, $x_1\ y_1 + x_2\ y_2$, ..., en doble precisión, de modo que la única vez que se redondea un número con precisión simple es cuando se da el resultado final. Con esta acumulación en doble precisión, el error en el cálculo del producto interno es aproximadamente el de una sola operación.

Queremos ahora usar la fórmula matricial (*IV*.4*b*) para describir la propagación del error de redondeo en un algoritmo. Como hemos ya visto, un algoritmo para calcular una función $\phi\colon D \subset \mathcal{R}^n \to \mathcal{R}^m$, para un dado $x = (x_1, \ldots, x_n)^t \in D$ corresponde a una descomposición de la aplicación ϕ en aplicaciones elementales, diferenciables continuamente, $\phi^{(i)}$, (ver (*IV*.1)), y nos lleva desde x hasta y con resultados intermedios

$$x = x^{(0)} \to \phi^{(0)}(x^{(0)}) = x^{(1)} \to \ldots \to \phi^{(r)}(x^{(r)}) = x^{(r+1)} = y\ .$$

Denotamos con $\psi^{(i)}$ la **aplicación resto**

$$\psi^{(i)} = \phi^{(r)} \circ \phi^{(r-1)} \circ \ldots \circ \phi^{(i)} : D_i \to \mathcal{R}^m\ , \qquad i = 0, 1, 2, \ldots, r\ .$$

Entonces, $\psi^{(0)} \equiv \phi$. $D\phi^{(i)}$ y $D\psi^{(i)}$ son las matrices Jacobianas de las aplicaciones $\phi^{(i)}$ y $\psi^{(i)}$, respectivamente. Dado que las matrices Jacobianas son multiplicativa con respecto de la composición de funciones, tenemos, para $i = 0, 1, 2 \ldots r$

$$D(f \circ g)(x) = Df(g(x)) \cdot Dg(x)\ ,$$

$$D\phi(x) = D\phi^{(r)}(x^{(r)}) \cdot D\phi^{(r-1)}(x^{(r-1)}) \ldots D\phi^{(0)}(x^{(0)})\ ,$$

$$D\psi^{(i)}(x^{(i)}) = D\phi^{(r)}(x^{(r)}) \cdot D\phi^{(r-1)}(x^{(r-1)}) \dots D\phi^{(i)}(x^{(i)}) \ .$$

Con aritmética de punto flotante, los errores iniciales y de redondeo perturberán los resultados intermedios $x^{(i)}$, de manera que se obtendrá el valor aproximado $x^{*(i+1)} = fl(\phi^{(i)}(x^{*(i)}))$. Para los errores absolutos obtenemos

$$\begin{aligned} \Delta x^{(i+1)} =& x^{*(i+1)} - x^{(i+1)} = \\ =& [fl(\phi^{(i)}(x^{*(i)})) - \phi^{(i)}(x^{*(i)})] + [\phi^{(i)}(x^{*(i)}) - \phi^{(i)}(x^{(i)})] \ . \end{aligned} \tag{IV.7}$$

Desde (*IV.4b*) sigue

$$\phi^{(i)}(x^{*(i)}) - \phi^{(i)}(x^{(i)}) \approx D\phi^{(i)}(x^{(i)})\Delta x^{(i)} \tag{IV.8}$$

Nótese que la aplicación $\phi^{(i)}\colon D_i \to D_{i+1} \subseteq \mathcal{R}^{n_i+1}$ es un vector de funciones componentes $\phi_j^{(i)}\colon D_i \to \mathcal{R}$, $j = 1, \dots, n_i + 1$. Entonces, podemos escribir

$$fl(\phi^{(i)}(u)) = (I + E_{i+1}) \cdot \phi^{(i)}(u) \ ,$$

con I la matriz identidad y E_{i+1} la matriz diagonal cuyos elementos son los errores $\varepsilon_j,\ j = 1, \dots, n_i + 1$, $|\varepsilon_j| \leq \nu$. Entonces, para el primer paréntesis de la expresión (*IV.7*) sigue

$$\begin{aligned} fl(\phi^{(i)}(x^{*(i)})) - \phi^{(i)}(x^{*(i)}) =& E_{i+1} \cdot \phi^{(i)}(x^{*(i)}) \\ \approx& E_{i+1} \cdot \phi^{(i)}(x^{(i)}) = E_{i+1} \cdot x^{(i+1)} = \alpha_{i+1} \ . \end{aligned} \tag{IV.9}$$

La cantidad α_{i+1} se puede interpretar como el error absoluto de redondeo creado cuando $\phi^{(i)}$ es evaluada en aritmética de punto flotante, y los elementos diagonales de E_{i+1} se pueden interpretar como los correspondientes errores relativos de redondeo.

Unendo (*IV.7*), (*IV.8*) y (*IV.9*), se puede expresar $\Delta x^{(i+1)}$ como aproximación del primer orden de la maniera siguiente

$$\Delta x^{(i+1)} \approx \alpha_{i+1} + D\phi^{(i)}(x^{(i)}) \cdot \Delta x^{(i)} = E_{i+1} \cdot x^{(i+1)} + D\phi^{(i)}(x^{(i)}) \cdot \Delta x^{(i)} \ .$$

Sigue entonces que

$$\begin{aligned} \Delta x^{(1)} \approx& D\phi^{(0)}(x) \cdot \Delta x + \alpha_1 \ , \\ \Delta x^{(2)} \approx& D\phi^{(1)}(x^{(1)})[D\phi^{(0)}(x) \cdot \Delta x + \alpha_1] + \alpha_2 \ , \\ & \dots\dots\dots \\ \Delta y = \Delta x^{(r+1)} \approx& D\phi^{(r)}(x^{(r)}) \dots D\phi^{(0)}(x) \cdot \Delta x + \\ & + D\phi^{(r)}(x^{(r)}) \dots D\phi^{(1)}(x^{(1)}) \cdot \alpha_1 + \dots + \alpha_{r+1} \ , \end{aligned}$$

que se puede escribir como

$$\begin{aligned} \Delta y \approx& D\phi(x) \cdot \Delta x + D\psi^{(1)}(x^{(1)}) \cdot \alpha_1 + \dots + D\psi^{(r)}(x^{(r)}) \cdot \alpha_r + \alpha_{r+1} \\ \approx& D\phi(x) \cdot \Delta x + D\psi^{(1)}(x^{(1)}) \cdot E_1 x^{(1)} + \dots + D\psi^{(r)}(x^{(r)}) \cdot E_r x^{(r)} + E_{r+1} y \ . \end{aligned} \tag{IV.10}$$

Es entonces la medida de la matriz Jacobiana $D\psi^{(i)}$ de la aplicación resto $\psi^{(i)}$ que es critica para el efecto de los errores de redondeo intermedios α_i ó E_i sobre el resultado final. Está claro que si para hallar el mismo resultados $\phi(x)$ se usan dos algoritmos diferentes, $D\phi(x)$ queda igual mientras que las matrices Jacobianas $D\psi^{(i)}$ que miden la propagación del error de redondeo serán diferentes. Un algoritmo se dirá **numéricamente más fiable** que otro si, por un dado conjunto de datos, el efecto total de redondeo, dado por $D\psi^{(1)}(x^{(1)})\cdot\alpha_1 + \ldots + D\psi^{(r)}(x^{(r)})\cdot\alpha_r + \alpha_{r+1}$ es menor por el primer algoritmo que por el segundo.

Ejemplo 16.

Queremos estudiar el error obtenido para hallar la operacion

$$\phi(a,b) = a^2 - b^2$$

con $\phi : \mathcal{R}^2 \to \mathcal{R}$. Dado que $a^2 - b^2 = (a+b)(a-b)$, se pueden usar para el calculo de ϕ los dos algoritmos

Algoritmo 1	Algoritmo 2
$\eta_1 = a \times a$	$\eta_1 = a + b$
$\eta_2 = b \times b$	$\eta_2 = a - b$
$y = \phi(a,b) = \eta_1 - \eta_2$	$y = \phi(a,b) = \eta_1 \times \eta_2$.

Las correspondientes decomposiciones ($IV.3$) de ϕ en este caso son

Algoritmo 1	Algoritmo 2
$\phi^{(0)}(a,b) = \begin{pmatrix} a^2 \\ b \end{pmatrix} \in \mathcal{R}^2$	$\phi^{(0)}(a,b) = \begin{pmatrix} a+b \\ a-b \end{pmatrix} \in \mathcal{R}^2$
$\phi^{(1)}(u,v) = \begin{pmatrix} u \\ v^2 \end{pmatrix} \in \mathcal{R}^2$	$\phi^{(1)}(u,v) = u \cdot v \in \mathcal{R}$
$\phi^{(2)}(\alpha,\beta) = \alpha - \beta \in \mathcal{R}$.	

Para el algoritmo 1 obtenemos

$$x = x^{(0)} = \begin{pmatrix} a \\ b \end{pmatrix} \ , \quad x^{(1)} = \begin{pmatrix} a^2 \\ b \end{pmatrix} \ , \quad x^{(2)} = \begin{pmatrix} a^2 \\ b^2 \end{pmatrix} \ , \quad x^{(3)} = y = a^2 - b^2 \ ,$$

$$\psi^{(1)}(u,v) = u - v^2 \ , \quad \psi^{(2)}(u,v) = u - v \ ,$$

$$D\phi(x) = (2a, -2b) \ , \quad D\psi^{(1)}(x^{(1)}) = (1, -2b) \ , \quad D\psi^{(2)}(x^{(2)}) = (1,-1) \ .$$

Además, dado que $fl(\phi^{(0)}(x^{(0)})) - \phi^{(0)}(x^{(0)}) = \begin{pmatrix} a \otimes a \\ b \end{pmatrix} - \begin{pmatrix} a^2 \\ b \end{pmatrix}$, tenemos, con $|\varepsilon_i| < \nu$

$$E_1 = \begin{pmatrix} \varepsilon_1 & 0 \\ 0 & 0 \end{pmatrix} \ , \ \alpha_1 = \begin{pmatrix} \varepsilon_1 \ a^2 \\ 0 \end{pmatrix} \ , \ E_2 = \begin{pmatrix} 0 & 0 \\ 0 & \varepsilon_2 \end{pmatrix} \ , \ \alpha_2 = \begin{pmatrix} 0 \\ \varepsilon_2 \ b^2 \end{pmatrix} \ , \ \alpha_3 = \varepsilon_3(a^2 - b^2) \ .$$

Desde ($IV.10$) con $\Delta x = (\Delta a, \Delta b)^t$ sigue

$$\Delta y \approx 2a\Delta a - 2b\Delta b + a^2\varepsilon_1 - b^2\varepsilon_2 + (a^2 - b^2)\varepsilon_3 \ . \qquad (IV.11)$$

De la misma manera para el algoritmo 2 sigue

$$x = x^{(0)} = \begin{pmatrix} a \\ b \end{pmatrix} \ , \quad x^{(1)} = \begin{pmatrix} a+b \\ a-b \end{pmatrix} \ , \quad x^{(2)} = y = a^2 - b^2 \ ,$$

$$\psi^{(1)}(u,v) = u \cdot v \ , \qquad D\phi(x) = (2a, -2b) \ , \qquad D\psi^{(1)}(x^{(1)}) = (a-b, a+b) \ ,$$

$$E_1 = \begin{pmatrix} \varepsilon_1 & 0 \\ 0 & \varepsilon_2 \end{pmatrix} \ , \quad \alpha_1 = \begin{pmatrix} \varepsilon_1(a+b) \\ \varepsilon_2(a-b) \end{pmatrix} \ , \quad \alpha_2 = \varepsilon_3(a^2 - b^2) \ .$$

Y entonces desde (IV.10) sigue

$$\Delta y \approx 2a\Delta a - 2b\Delta b + (a^2 - b^2)(\varepsilon_1 + \varepsilon_2 + \varepsilon_3) \ . \qquad (IV.12)$$

Desde las ecuaciones (IV.11) y (IV.12) se obtienen los siguientes efectos totales de redondeo:

$$|a^2\varepsilon_1 - b^2\varepsilon_2 + (a^2 - b^2)\varepsilon_3| \leq (a^2 + b^2 + |a^2 - b^2|)\nu \ ,$$

para el algoritmo 1, y

$$|(a^2 - b^2)(\varepsilon_1 + \varepsilon_2 + \varepsilon_3)| \leq 3|a^2 - b^2|\nu \ ,$$

para el algoritmo 2. Entonces, podemos decir que el algoritmo 2 es numéricamente más fiable que el algoritmo 1 cada vez que $\frac{1}{3} < |\frac{a}{b}|^2 < 3$; en los otros casos el algoritmo 1 es más fiable. Esto sigue desde la equivalencia de las dos relaciones $\frac{1}{3} \leq |\frac{a}{b}|^2 \leq 3$ y $3|a^2 - b^2| \leq a^2 + b^2 + |a^2 - b^2|$.

Por ejemplo para $a = 0.3237$ y $b = 0.3134$, con aritmética de cuatro dígitos significativos, se obtienen los siguientes resultados:

Algoritmo 1: $a \otimes a = 0.1048$, $b \otimes b = 0.9822 \times 10^{-1}$

$a \otimes a - b \otimes b = 0.6580 \times 10^{-2}$.

Algoritmo 2: $a \oplus b = 0.6371$, $a \ominus b = 0.1030 \times 10^{-1}$

$a^2 - b^2 = 0.6562 \times 10^{-2}$.

Resultado exacto: $a^2 - b^2 = 0.656213 \times 10^{-2}$.

EJERCICIOS.

1. Encuentre cotas para x, si x es una aproximación con cuatro cifras significativas de

a) π ;
b) e .

2. Suponga que p^* aproxima a p con 3 dígitos significativos. Encuentre el intervalo en el cual p^* debe estar, si p es

a) 150 ; 900 ;
b) 1500 ; 90 .

3. Considere los siguientes valores de p y p^*. ¿Con cuántas cifras significativas aproxima p^* a p?

a) $p = \pi, \quad p^* = 3.1$;
b) $p = \frac{1}{3}, \quad p^* = 0.333$;
c) $p = \frac{\pi}{1000}, \quad p^* = 0.0031$;
d) $p = \frac{100}{3}, \quad p^* = 33.3$;
e) $p = \pi, \quad p^* = 3.141$;
f) $p = \frac{1}{3}, \quad p^* = 0.33333$;
g) $p = \frac{\pi}{1000}, \quad p^* = 0.00314$;
h) $p = \frac{100}{3}, \quad p^* = 33.33$.

4. Efectúe los siguientes cálculos (i) exactamente, (ii) usando aritmética cortando a tres dígitos, (iii) usando aritmética de redondeo a tres dígitos. Determine luego las pérdidas de dígitos significativos suponiendo que los números dados son exactos.

a) $14.1 + 0.0981$;
b) $0.0218 * 179.$;
c) $(164. + 0.913) - (143. + 21.0)$;
d) $(164. - 143.) + (0.913 - 21.0)$.

5. Efectúe los siguientes cálculos (i) exactamente, (ii) usando aritmética cortando a tres dígitos, (iii) usando aritmética de redondeo a tres dígitos.

a) $\frac{4}{5} + \frac{1}{3}$;
b) $\frac{4}{5} * \frac{1}{3}$;
c) $\left(\frac{1}{3} - \frac{3}{11}\right) + \frac{3}{20}$;
d) $\left(\frac{1}{3} + \frac{3}{11}\right) - \frac{3}{20}$.

6. Representar en (a) magnitud-signo, (b) notación en exceso, (c) complemento a dos, todos con 8 bits, los siguientes números: 18, -23, 100, 127, -127, 0, -55 y -99.

7. Representar en punto flotante con 32 bits (1 bit para el signo, 7 bits para el exponente y 24 bits para la mantisa) los números: 1234.5675, -1234.5675 y -234.25.

CURSO DE MÉTODOS NUMÉRICOS

SEGUNDA PARTE

SOLUCIÓN APROXIMADA DE ECUACIONES DE UNA VARIABLE

CAPITULO V. SOLUCION APROXIMADA DE ECUACIONES DE UNA VARIABLE: PRELIMINARES

1. SEPARACION DE RAICES

En esta segunda parte analizaremos uno de los problemas básicos del análisis numérico: el **problema de búsqueda de raíces**.

Si una ecuación algebráica o trascendente es relativamente complicada, no resulta posible por lo general hallar raíces exactas. Es más, en algunos casos las ecuaciones tienen coeficientes conocidos sólo de forma aproximada, y por tanto, carece de sentido tratar de hallar las raíces exactas de la ecuación. Por consiguiente, adquieren particular importancia los procedimientos de cálculo aproximado de raíces de una ecuación así como la estimación de su grado de exactitud.

El problema consiste en encontrar los valores de la variable x que satisfacen la ecuación

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

para una función f dada, que está definida y es continua en un cierto intervalo finito o infinito $a < x < b$. En ciertos casos se necesitará la existencia y continuidad de la primera derivada $f'(x)$ e incluso de la segunda derivada $f''(x)$.

A una solución de este problema, es decir a todo valor p para el cual la función $f(x)$ es cero, se le llama **cero** de la función $f(x)$ o una **raíz** de $f(x) = 0$.

Supondremos que la ecuación (V.1) tiene únicamente raíces separadas, es decir, para cada raíz existe un entorno que no contiene otras raíces de la ecuación.

El cálculo aproximado de las raíces reales separadas de (V.1) se efectúa por lo general en dos etapas:

(a) separación de raíces, es decir, establecer los intervalos más pequeños posibles $[\alpha, \beta]$ que contengan una y solamente una raíz de la ecuación (V.1);

(b) mejorar los valores de las raíces aproximadas, es decir, manipularlos hasta que presenten el grado de exactitud especificado.

Recordemos antes el **Teorema del Valor Intermedio**:

Si $f \in \mathcal{C}[a, b]$ y K es un número cualquiera entre $f(a)$ y $f(b)$, entonces existe c en (a, b) tal que $f(c) = K$.

Y un Corolario de ese Teorema:

Corolario V.1

Si $f \in \mathcal{C}[a, b]$ asume valores de signo opuesto en los extremos de un intervalo $[\alpha, \beta]$, es decir, $f(\alpha) \cdot f(\beta) < 0$, entonces el intervalo contendrá al menos una raíz de la ecuación $f(x) = 0$; en otras palabras, habrá al menos un número $p \in (\alpha, \beta)$ tal que $f(p) = 0$.

La raíz p será única si la derivada $f'(x)$ existe y mantiene el signo dentro del intervalo (α, β); esto es, si $f'(x) > 0$ (ó $f'(x) < 0$) para $\alpha < x < \beta$.

El proceso de separación de raíces comienza estableciendo los signos de la función $f(x)$ en los puntos extremos $x = a$ y $x = b$ de sus dominios de existencia. A continuación se determinan los signos de la función $f(x)$ para un número intermedio de puntos $x = \alpha_1, \alpha_2, ...$, cuya elección depende de la peculiaridades de la función $f(x)$. Si se cumple que $f(\alpha_k) \cdot f(\alpha_{k+1}) < 0$, entonces, en virtud del Corolario V.1, existe una raíz de la ecuación $f(x) = 0$ en el intervalo (α_k, α_{k+1}). Debemos asegurarnos que esta raíz es la única.

En la práctica suele ser suficiente, en el caso de separación de raíces, efectuar el proceso de bisección (que analizaremos en más detalle en el próximo capítulo), dividiendo aproximadamente el intervalo dado (α, β) en dos, cuatro, ocho, ..., partes iguales (hasta un cierto intervalo) y determinar el signo de $f(x)$ en los puntos de división. Conviene recordar que en una ecuación algebráica de grado n,

$$a_0\ x^n + a_1\ x^{n-1} + ... + a_n = 0\ , \qquad a_0 \neq 0$$

tiene a lo sumo n raíces reales. Por consiguiente, si para una ecuación de este tipo se obtienen n cambios de signo (es decir, $n + 1$ intervalos el los cuales la función tiene signo distinto), habrán quedado separadas todas las raíces de la ecuación.

Ejemplo 1.

Queremos separar las raíces de la ecuación:

$$f(x) \equiv x^3 - 6\ x + 2 = 0\ .$$

De la tabla siguiente

x	$-\infty$	-3	-1	0	$+1$	$+3$	$+\infty$
$f(x)$	$-$	$-$	$+$	$+$	$-$	$+$	$+$

se sigue que la ecuación dada tiene tres raíces reales dentro de los intervalos $(-3, -1)$, $(0, 1)$ y $(1, 3)$.

Si existe una derivada continua $f'(x)$ y pueden calcularse fácilmente las raíces de la ecuación $f'(x) = 0$, puede regularizarse el proceso de separación de raíces de la ecuación $(V.1)$. Evidentemente es suficiente contar únicamente los signos de la función $f(x)$ para los ceros de su derivada y en los puntos extremos $x = a$ y $x = b$.

Ejemplo 2.

Se quieren separar las raíces de la ecuación

$$f(x) \equiv x^4 - 4\ x - 1 = 0\ .$$

Dado que $f'(x) = 4(x^3 - 1)$, será $f'(x) = 0$ para $x = 1$. Tenemos: $f(-\infty) > 0$ $(+)$, $f(1) < 0$ $(-)$, $f(+\infty) > 0$ $(+)$. Entonces, la ecuación tiene únicamente dos raíces reales, una de las cuales cae en el intervalo $(-\infty, 1)$, y la otra en el $(1, +\infty)$.

Ejemplo 3.

Se quiere determinar el número de raíces reales de la ecuación

$$f(x) \equiv x + e^x = 0 \ .$$

Como $f'(x) = 1 + e^x > 0$ y $f(-\infty) = -\infty$, $f(+\infty) = +\infty$, se deduce que la ecuación tiene únicamente una raíz real.

Vamos ahora a recordar dos Teoremas que usaremos más adelante:

Teorema del Valor Medio

Si $f \in \mathcal{C}[a,b]$ y f es diferenciable en (a,b), entonces existe un número c, $a < c < b$, tal que

$$f'(c) = \frac{f(b) - f(a)}{b - a} \ .$$

Teorema del Valor Extremo

Si $f \in \mathcal{C}[a,b]$, entonces existen $c_1, c_2 \in [a,b]$ tales que $f(c_1) \leq f(x) \leq f(c_2)$ para todo $x \in [a,b]$. Si además, f es diferenciable en (a,b), entonces los números c_1 y c_2 existirán ya sea en los extremos de $[a,b]$, o donde f' sea cero.

Veamos ahora una estimación del error de una raíz aproximada.

Teorema V.2

Sea p una raíz exacta y $\overline{x}$ una raíz aproximada de la ecuación $f(x) = 0$, situadas ambas en el mismo intervalo $[\alpha, \beta]$, y

$$|f'(x)| \geq m_1 > 0 \ ,$$

para $\alpha \leq x \leq \beta$. Se cumple entonces la siguiente aproximación:

$$|\overline{x} - p| \leq \frac{|f(\overline{x})|}{m_1} \ . \qquad (V.2)$$

Demostración: aplicando el Teorema del valor medio, se tiene

$$f(\overline{x}) - f(p) = (\overline{x} - p)\ f'(c)$$

donde c es un valor intermedio entre $\overline{x}$ y p, es decir, $c \in (\alpha, \beta)$.
De aquí , ya que $f(p) = 0$ y $|f'(c)| \geq m_1$ (puede tomarse para m_1, por ejemplo, el valor más pequeño de $|f'(x)|$ cuando $\alpha \leq x \leq \beta$), tenemos

$$|f(\overline{x}) - f(p)| = |f(\overline{x})| \geq m_1\ |\overline{x} - p|$$

y entonces,

$$|\overline{x} - p| \leq \frac{|f(\overline{x})|}{m_1} \ .$$

c.q.d.

Nótese que la fórmula (*V*.2) puede ofrecer sólo resultados someros y por tanto no es siempre conveniente utilizarla. Por esta razón en la práctica resulta mejor estrechar el intervalo general (α, β) que contiene la raíz p y su valor aproximado $\overline{x}$, y considerar $|\overline{x} - p| \leq \beta - \alpha$.

Ejemplo 4.

Como valor aproximado de la raíz de la ecuación $f(x) \equiv x^4 - x - 1 = 0$ tenemos $\overline{x} = 1.22$. Estímese el error absoluto en esta raíz.

$$f(\overline{x}) = 2.2153 - 1.22 - 1 = -0.0047$$

Como para $\overline{\overline{x}} = 1.23$, tenemos

$$f(\overline{\overline{x}}) = 2.2889 - 1.23 - 1 = 0.0589$$

la raíz exacta p cae en el intervalo $(1.22, 1.23)$. La derivada $f'(x) = 4\ x^3 - 1$ crece en forma monótona y por tanto su valor más pequeño en el intervalo dado es

$$m_1 = 4 * 1.22^3 - 1 = 4 * 1.8158 - 1 = 6.2632$$

de donde, mediante la fórmula (*V*.2), tenemos

$$|\overline{x} - p| \leq \frac{0.0047}{6.2632} \approx 0.000750415\ .$$

Nótese que ocasionalmente, en la práctica, la exactitud de una raíz aproximada x se estima en función de cómo satisfaga la ecuación dada $f(x) = 0$; es decir, si el número $|f(\overline{x})|$ es pequeño, se considera entonces $\overline{x}$ una buena aproximación a la raíz exacta p; pero si $|f(\overline{x})|$ es grande, entonces $\overline{x}$ se toma como aproximación grosera de la raíz exacta p. Pero esa forma de proceder es errónea, porque hay funciones que crecen muy rápidamente y entonces el valor $|f(\overline{x})|$ es grande aunque $\overline{x}$ está cerca de p, y hay funciones que crecen muy lentamente y entonces el valor $|f(\overline{x})|$ es pequeño aunque $\overline{x}$ esté lejano de p.

2. SOLUCION GRAFICA DE ECUACIONES

Las raíces reales de la ecuación

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

pueden determinarse en forma aproximada considerando las abscisas de los puntos de intersección de la gráfica de la función $y = f(x)$ con el eje x.
Resulta aconsejable a veces sustituir la ecuación dada por una ecuación equivalente (dos ecuaciones se denominan equivalentes si tienen exactamente las mismas raíces):

$$\phi(x) = \psi(x)$$

donde las funciones $\phi(x)$ y $\psi(x)$ son más sencillas que $f(x)$. Constrúyanse entonces las gráficas de las funciones $y = \phi(x)$ e $y = \psi(x)$, y las raíces deseadas serán entonces las abscisas de los puntos de intersección de estas gráficas.

Ejemplo 5.

Resuélvase gráficamente la siguiente ecuación

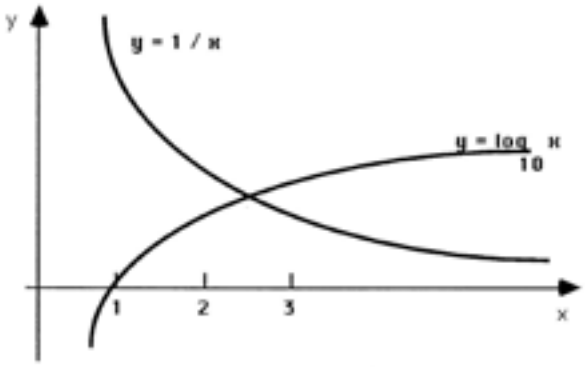
$$x \ \log_{10} x = 1 \ .$$

Para ello, escríbimos la ecuación de la forma

$$\log_{10} x = \frac{1}{x} \ .$$

Las raíces pueden entonces hallarse fácilmente, ya que son las abscisas de los puntos de intersección de la curva logarítmica $y = \log_{10} x$ y la hipérbola $y = \frac{1}{x}$. Construyendo estas curvas, tendremos un valor aproximado $p \approx 2.5$ de la única raíz de la ecuación dada.

Figura 1



Si una de las funciones $\phi(x)$ ó $\psi(x)$ es lineal queda simplificada la operación de hallar las raíces de la ecuación. Las raíces son entonces las abscisas de los puntos de intersección de la curva $y = \phi(x)$ y la línea recta $y = a \ x + b$. Este procedimiento es particularmente ventajoso cuando han de resolverse series de ecuaciones del mismo tipo que difieren únicamente en los coeficientes a y b de una función lineal. La construcción gráfica se reduce entonces a hallar los puntos de intersección de una gráfica dada y varias líneas rectas. Este caso incluye evidentemente las ecuaciones de tres términos

$$x^n + a \ x + b = 0 \ .$$

Ejemplo 6.

Resuélvanse las ecuaciones cúbicas

$$x^3 - 1.75 \ x + 0.75 = 0$$

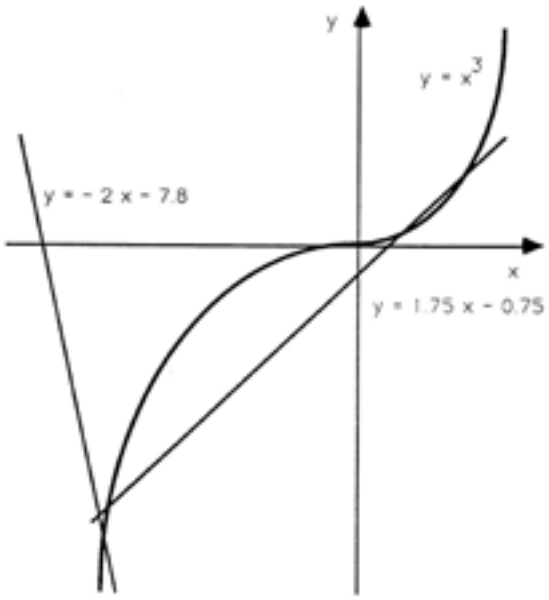
y

$$x^3 + 2\ x + 7.8 = 0\ .$$

Construímos la cúbica $y = x^3$. Las raíces deseadas son las abscisas de los puntos de intersección de esta cúbica con las líneas rectas $y = 1.75\ x - 0.75$ e $y = -2\ x - 7.8$.

Se deduce claramente de la figura que la primera ecuación tiene tres raíces reales: $x_1 = -1.5$, $x_2 = 0.5$ y $x_3 = 1$, y la segunda ecuación tiene únicamente una raíz real, $x_1 = -1.65$.

Figura 2



CAPITULO VI. EL ALGORITMO DE BISECCION

1. INTRODUCCION Y METODO

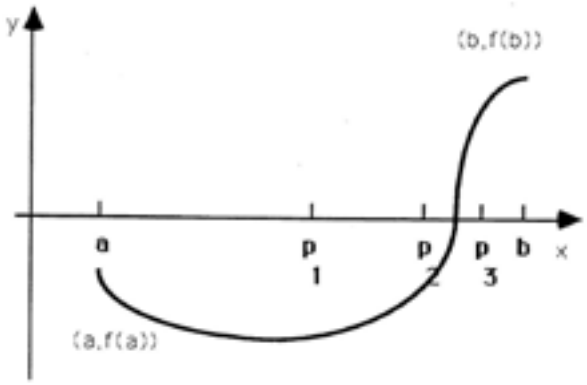
En este capítulo comenzaremos a analizar uno de los problemas más básicos del análisis numérico: el **problema de búsqueda de raíces**. El problema consiste en encontrar los valores de la variable x que satisfacen la ecuación $f(x) = 0$, para una función f dada.

La primera técnica, basada en el Teorema del Valor Intermedio, se llama **algoritmo de bisección** ó **método de búsqueda binaria**, ó también **método de Bolzano**.

Supongamos que tenemos una función continua f definida en el intervalo $[a, b]$, con $f(a)$ y $f(b)$ de signos distintos. Entonces por el corolario V.1 del Teorema del Valor Intermedio, existe p, $a < p < b$, tal que $f(p) = 0$. Aunque el procedimiento sirve para el caso en el que $f(a)$ y $f(b)$ tienen signos opuestos y hay más de una raíz en el intervalo $[a, b]$, por simplicidad se supondrá que la raíz en este intervalo es única.

El método requiere dividir repetidamente a la mitad los subintervalos de $[a, b]$ y, en cada paso, localizar la mitad que contiene a p. Para empezar, tomemos $a_1 = a$ y $b_1 = b$ y p_1 el punto medio de $[a, b]$; o sea $p_1 = \frac{1}{2}(a_1 + b_1)$. Si $f(p_1) = 0$, entonces $p = p_1$; si no, entonces $f(p_1)$ tiene el mismo signo que $f(a_1)$ o $f(b_1)$. Si $f(p_1)$ y $f(a_1)$ tienen el mismo signo, entonces $p \in (p_1, b_1)$, y tomamos $a_2 = p_1$ y $b_2 = b_1$. Si $f(p_1)$ y $f(b_1)$ son del mismo signo, entonces $p \in (a_1, p_1)$, y tomamos $a_2 = a_1$ y $b_2 = p_1$. Ahora re-aplicamos el proceso al intervalo $[a_2, b_2]$. Y así hasta que se encuentra $f(p) = 0$ ó el i-ésimo intervalo $[a_i, b_i]$ es más pequeño que una toleracia TOL prefijada, ó hasta que se cumpla alguna otra condición de paro.

Figura 1



El procedimiento de paro más común es el de dar un número máximo de iteraciones N_0. Cuando usamos un ordenador para generar las aproximaciones, conviene añadir una condición que imponga un máximo al número de iteraciones realizadas. Así se elimina la posibilidad de poner a la máquina en un ciclo infinito, una posibilidad que puede

surgir cuando la sucesión diverge (y también cuando el programa está codificado incorrectamente). Esto se hace fácilmente dando una cota inicial N_0 y requiriendo que el procedimiento termine si se supera esa cota.

Otros procedimientos de paro que se pueden aplicar a cualquier técnica iterativa son los de dar una tolerancia $\varepsilon > 0$ y generar una sucesión $p_1, p_2, ..., p_n$ hasta que una de las siguientes condiciones se satisfaga:

$$|p_n - p_{n-1}| < \varepsilon \tag{$VI.1$}$$

$$\frac{|p_n - p_{n-1}|}{|p_n|} < \varepsilon, \qquad p_n \neq 0 \tag{$VI.2$}$$

$$|f(p_n)| < \varepsilon \ . \tag{$VI.3$}$$

Desafortunadamente, pueden surgir dificultades usando cualquiera de estos criterios de paro. Por ejemplo, existen sucesiones $\{p_n\}$ con la propiedad de que las diferencias $p_n - p_{n-1}$ convergen a cero mientras que la sucesión misma diverge. Es posible también que $f(p_n)$ esté cerca de cero mientras que p_n difiere significativamente de p. Sin conocimiento adicional acerca de f ó p, la desigualdad $(VI.2)$ es el mejor criterio de paro que puede aplicarse porque verifica el error relativo.

Nótese que para empezar el algoritmo de bisección, se debe encontrar un intervalo $[a, b]$ tal que $f(a) \cdot f(b) < 0$. En cada paso del algoritmo de bisección, la longitud del intervalo que contiene el cero de f se reduce por un factor de dos; por lo tanto es ventajoso escoger el intervalo inicial $[a, b]$ tan pequeño como sea posible. Por ejemplo, si $f(x) = 2\ x^3 - x^2 + x - 1$,

$$f(-4) \cdot f(4) < 0 \qquad \text{y} \qquad f(0) \cdot f(1) < 0 \ ,$$

así que el algoritmo de bisección puede usarse con cualquiera de los intervalos $[-4, 4]$ ó $[0, 1]$. Empezando el algoritmo de bisección con $[0, 1]$ en vez de con $[-4, 4]$, reducirá en tres el número de iteraciones requeridas para alcanzar una precisión específica.

2. ALGORITMO Y EJEMPLOS

Algoritmo de bisección.

==========

Para encontrar una solución de $f(x) = 0$ dada la función f en el intervalo $[a, b]$ donde $f(a)$ y $f(b)$ tienen signo opuestos:

Entrada: extremos a y b; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada p ó mensaje de fracaso.

Paso 1: tomar $i = 1$;

Paso 2: mientras que $i \leq N_0$ seguir pasos 3–6;

Paso 3: tomar $p = a + \dfrac{(b-a)}{2}$ *(calcular p_i)*;

Paso 4: si $f(p) = 0$ ó $\dfrac{(b-a)}{2} < TOL$ entonces SALIDA (p);
(procedimiento completado satisfactoriamente) PARAR;

Paso 5: tomar $i = i + 1$

Paso 6: si $f(a) \cdot f(p) > 0$ entonces tomar $a = p$, si no, tomar $b = p$ *(calcular a_i, b_i)*;

Paso 7: SALIDA ($'$*El método fracasó después de N_0 iteraciones, $N_0 =$* $', N_0$); *(procedimiento completado sin éxito)*; PARAR.

==========

Para ilustrar el algoritmo de bisección, considérese el siguiente ejemplo. En este caso se termina la iteración cuando $\dfrac{|p_{n-1} - p_n|}{|p_n|} < 10^{-4}$.

Ejemplo 1.

La función $f(x) = x^3 + 4\,x^2 - 10$ tiene una raíz en $[1, 2]$ ya que $f(1) = -5$ y $f(2) = 14$. Es fácil ver que hay una sóla raíz en $[1, 2]$. El algoritmo de bisección da los valores de la tabla 1.

Después de 13 iteraciones, podemos ver que $p_{13} = 1.365112305$ aproxima a la raíz p con un error de $|p - p_{13}| < |b_{14} - a_{14}| = |1.365234375 - 1.365112305| = 0.000122070$ y como $|a_{14}| < |p|$,

$$\frac{|p - p_{13}|}{|p|} < \frac{|b_{14} - a_{14}|}{|a_{14}|} \leq 9.0 \times 10^{-5}\ ,$$

la aproximación es correcta al menos con cuatro cifras significativas. El valor correcto de p, con nueve cifras decimales, es $p = 1.365230013$.

Es interesante notar que p_9 está más cerca de p que la aproximación final p_{13}, pero no hay manera de determinar esto a menos que se conozca la respuesta correcta.

Tabla 1

n	a_n	b_n	p_n	$f(p_n)$	$\left\lvert \frac{p_n - p_{n-1}}{p_n} \right\rvert$
1	1.0	2.0	1.5	2.375	
2	1.0	1.5	1.25	−1.796875	0.2
3	1.25	1.5	1.375	0.16211	0.090909
4	1.25	1.375	1.3125	−0.84839	0.047619
5	1.3125	1.375	1.34375	−0.35098	0.023256
6	1.34375	1.375	1.359375	−0.09641	0.011494
7	1.359375	1.375	1.3671875	0.03236	0.0057143
8	1.359375	1.3671875	1.36328125	−0.03215	0.0028653
9	1.36328125	1.3671875	1.365234375	0.00007	0.0014306
10	1.36328125	1.365234375	1.364257813	−0.01605	0.00071582
11	1.364257813	1.365234375	1.364746094	−0.00799	0.00035778
12	1.364746094	1.365234375	1.364990234	−0.00396	0.00017886
13	1.364990234	1.365234375	1.365112305	−0.00194	0.000089422

El algoritmo de bisección, aunque conceptualmente claro, tiene inconvenientes importantes. Converge muy lentamente (o sea, N puede ser muy grande antes que $|p - p_N|$ sea suficientemente pequeño) y, más aún, una buena aproximación intermedia puede ser desechada sin que nos demos cuenta. Sin embargo, el método tiene la propiedad importante de que converge siempre a una solución y, por esta razón se usa frecuentemente para "poner en marcha" a los métodos más eficientes que se presentarán más adelante.

Definición

Decimos que $\{\alpha_n\}_{n=1}^{\infty}$ converge a α con *rapidez de convergencia* $O(\beta_n)$, donde $\{\beta_n\}_{n=1}^{\infty}$ es otra sucesión con $\beta_n \neq 0$ para cada n, si

$$\frac{|\alpha_n - \alpha|}{|\beta_n|} \leq K \qquad \text{para } n \text{ suficientemente grande}$$

donde K es una constante independiente de n. Esto se indica por lo general escribiendo $\alpha_n = \alpha + O(\beta_n)$ ó $\alpha_n \to \alpha$ con una rapidez de convergencia $O(\beta_n)$.

Teorema VI.1

Sea $f \in \mathcal{C}[a,b]$ y supongamos que $f(a) \cdot f(b) < 0$. El procedimiento de bisección genera una sucesión $\{p_n\}$ que aproxima a p con la propiedad

$$|p_n - p| \leq \frac{b-a}{2^n}, \qquad n \geq 1 \; . \tag{$VI.4$}$$

Demostración: para cada $n \geq 1$, tenemos

$$b_n - a_n = \frac{1}{2^{n-1}}(b-a) \qquad \text{y} \qquad p \in (a_n, b_n) \; .$$

Ya que $p_n = \frac{1}{2}(a_n + b_n)$, para todo $n \geq 1$, se sigue que

$$|p_n - p| = \left|\frac{1}{2}(a_n + b_n) - p\right| \leq \left|\frac{1}{2}(a_n + b_n) - a_n\right| = \frac{1}{2}\,(b_n - a_n) = 2^{-n}\,(b-a) \; .$$

c.q.d.

De acuerdo con la definición de rapidez de convergencia, la desigualdad $(VI.4)$ implica que $\{p_n\}_{n=1}^{\infty}$ converge a p y está acotada por una sucesión que converge a cero con una rapidez de convergencia $O(2^{-n})$. Es importante hacer notar que Teoremas como éste dan solamente cotas aproximadas para los errores.

Por ejemplo, esta cota aplicada al problema del ejemplo 1 afirma únicamente que

$$|p - p_9| \leq \frac{2-1}{2^9} \approx 2.0 \times 10^{-3} \; ,$$

siendo el error real mucho más pequeño:

$$|p - p_9| = |1.365230013 - 1.365234275| \approx 4.3 \times 10^{-6} \; .$$

Ejemplo 2.

Determinar aproximadamente cuántas iteraciones son necesarias para resolver $f(x) = x^3 + 4\,x^2 - 10 = 0$ con una precisión de $\varepsilon = 10^{-5}$ para $a_1 = 1$ y $b_1 = 2$. Esto requiere encontrar un entero N que satisfaga:

$$|p_N - p| \leq 2^{-N}\,(b-a) = 2^{-N} \leq 10^{-5} \; .$$

Para determinar N usamos logaritmos. Aunque sería suficiente usar logaritmos de cualquier base, usaremos logaritmos de base 10 pues la tolerancia está dada como una potencia de 10. Ya que $2^{-N} \leq 10^{-5}$ implica que $\log_{10}(2^{-N}) \leq \log_{10}(10^{-5}) = -5$,

$$-N \log_{10} 2 \leq -5 \qquad \text{ó} \qquad N \geq \frac{5}{\log_{10} 2} \approx 16.6 \ .$$

Parecería que se requieren 17 iteraciones para obtener una aproximación exacta a 10^{-5}. Con $\varepsilon = 10^{-3}$, se requieren $N \geq 10$ iteraciones y el valor de $p_9 = 1.365234275$ es exacto dentro de 10^{-5}. Es importante notar que estas técnicas dan solamente una cota para el número de iteraciones necesarias, y en muchos casos esta cota es mucho más grande que el número realmente requerido.

Ejemplo 3.

Utilizando el algoritmo de bisección para mejorar una raíz de la ecuación

$$f(x) \equiv x^4 + 2\ x^3 - x - 1 = 0$$

que caiga en el intervalo $[0, 1]$ y tal que $|f(p)| \leq 5.0 * 10^{-2}$, obtenemos los valores de la tabla 2.

Tabla 2

n	a_n	b_n	p_n	$f(p_n)$
1	0.0	1.0	0.5	−1.1875
2	0.5	1.0	0.75	−0.58984375
3	0.75	1.0	0.875	0.05102539
4	0.75	0.875	0.8125	−0.30393982
5	0.8125	0.875	0.84375	−0.13557339
6	0.84375	0.875	0.859375	−0.04461473

Entonces, la aproximación $p_6 = 0.859375$ es la mejora de la raíz buscada. Además,

$$\frac{|p - p_6|}{|p|} < \frac{|b_7 - a_7|}{|a_7|} = \frac{|0.875 - 0.859375|}{0.859375} \approx 0.018182 < 5.0 \times 10^{-2} \ ,$$

es decir, la aproximación es correcta al menos con dos cifras significativas.

EJERCICIOS.

1. Determinar el número de raíces y separarlas, es decir, encontrar los intervalos adecuados para los cuales se da la unicidad de la raíz, y resolver también gráficamente las ecuaciones siguientes:

a) $x - 2^{-x} = 0$ *b)* $e^x - x^2 + 3\,x - 2 = 0$

Aproxime el cero con 10^{-2} de precisión usando el algoritmo de bisección.

2. Determinar el número de raíces y separarlas, es decir, encontrar los intervalos adecuados para los cuales se da la unicidad de la raíz, y resolver también gráficamente las ecuaciones siguientes:

a) $x^2 + 10\,\cos x = 0$ *b)* $2\,\sin \pi x + x = 0$

Aproxime el cero con 10^{-2} de precisión usando el algoritmo de bisección.

3. Demuestre que $f(x) = x^3 - x - 1$ tiene exactamente un cero en el intervalo $[1, 2]$. Aproxime el cero con 10^{-2} de precisión usando el algoritmo de bisección.

4. Use el algoritmo de bisección para encontrar soluciones con una exactitud de 10^{-2} para $x^4 - 2\,x^3 - 4\,x^2 + 4\,x + 4 = 0$ en:

a) $[-2, 0]$ *b)* $[0, 2]$ *c)* $[1, 2]$

5. Use el algoritmo de bisección para encontrar todas las soluciones de $x^3 - 7\,x^2 + 14\,x - 6 = 0$ con una precisión de 10^{-3}.

6. Use el algoritmo de bisección para encontrar soluciones correctas a 10^{-5} para los siguientes problemas:

a) $x - 2^x = 0$ para $0 \le x \le 1$,

b) $e^x + 2^{-x} + 2\,\cos x - 6 = 0$ para $1 \le x \le 2$,

c) $e^x - x^2 + 3\,x - 2 = 0$ para $0 \le x \le 1$.

7. Use el algoritmo de bisección para encontrar una solución, correcta en 10^{-2}, para $x + 0.5 + 2\,\cos \pi x = 0$ en $[0.5, 1.5]$.

8. Cambie el paso 6 en el algoritmo de bisección a:

"Si $f(p) \cdot f(b) > 0$ entonces tomar $b = p$, si no, tomar $a = p$"

Use este nuevo algoritmo para encontrar una solución, correcta a 10^{-2} , para $x + 0.5 + 2\,\cos \pi x = 0$ en $[0.5, 1.5]$. Discuta cualquier discrepancia entre estos resultados y los del problema 5.

9. Encontrar una aproximación de $\sqrt{3}$ correcta en 10^{-4}, usando el algoritmo de bisección.

10. Use el Teorema VI.1 para encontrar una cota al número de iteraciones necesarias para alcanzar una aproximación con exactitud de 10^{-3} en la solución de $x^3 + x - 4 = 0$, que se encuentra en el intervalo $[1, 4]$. Encuentre una aproximación a esta raíz con este grado de exactitud.

CAPITULO VII. ITERACION DEL PUNTO FIJO

1. INTRODUCCION Y METODO

En este capítulo consideraremos un método para determinar la solución de una ecuación que se expresa, para alguna función g, de la forma

$$g(x) = x \ .$$

A una solución de esta ecuación se le llama un **punto fijo** de la función g.

Si para cualquier función g dada se puede encontrar un punto fijo, entonces cada problema de búsqueda de las raíces de $f(x) = 0$ tiene soluciones que corresponden precisamente a los puntos fijos de $g(x) = x$ con $g(x) = x - f(x)$. La primera tarea entonces es decidir cuándo una función tendrá un punto fijo y cómo se pueden determinar (es decir, aproximar con suficiente grado de precisión) dichos puntos fijos.

El siguiente Teorema da las condiciones suficientes para la existencia y unicidad de un punto fijo.

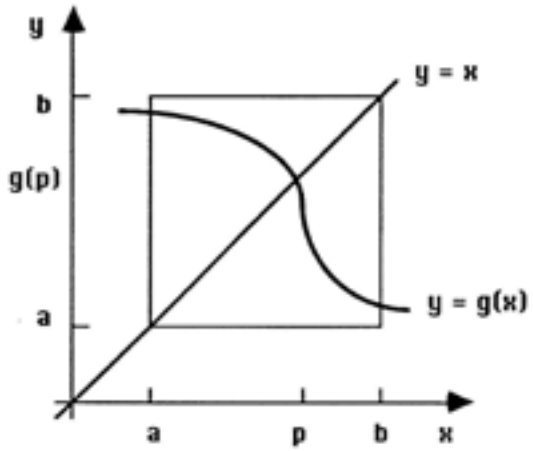
Teorema VII.1

Si $g \in \mathcal{C}[a,b]$ y $g(x) \in [a,b]$ para todo $x \in [a,b]$, entonces g tiene un punto fijo en $[a,b]$. Si además, $g'(x)$ existe en (a,b) y

$$|g'(x)| \leq k < 1 \qquad \text{para todo } x \in (a,b) \ , \tag{$VII.1$}$$

entonces g tiene un punto fijo único p en $[a,b]$.

Figura 1



Demostración: si $g(a) = a$ ó $g(b) = b$, la existencia del punto fijo es obvia. Supongamos que no es así; entonces debe ser cierto que $g(a) > a$ y $g(b) < b$. Definamos $h(x) = g(x) - x$; h es continua en $[a,b]$, y

$$h(a) = g(a) - a > 0 \ , \qquad h(b) = g(b) - b < 0 \ .$$

El corolario V.1 del Teorema del Valor Intermedio implica que existe $p \in (a, b)$ tal que $h(p) = 0$. Por lo tanto $g(p) - p = 0$ y p es un punto fijo de g.

Supongamos además, que la desigualdad $(VII.1)$ se satisface y que p y q son puntos fijos en $[a, b]$ con $p \neq q$. Por el Teorema del Valor Medio, existe un número ξ entre p y q y por lo tanto en $[a, b]$ tal que

$$|p - q| = |g(p) - g(q)| = |g'(\xi)|\ |p - q| \leq k\ |p - q| < |p - q|\ ,$$

lo cual es una contradicción. Esta contradicción debe venir de la única suposición, $p \neq q$. Por lo tanto, $p = q$ y el punto fijo en $[a, b]$ es único. **c.q.d.**

Ejemplo 1.

Sea $g(x) = \dfrac{x^2 - 1}{3}$ en el intervalo $[-1, 1]$. Usando el Teorema del Valor Extremo es fácil demostrar que el mínimo absoluto de g está en $x = 0$ y es $g(0) = -1/3$. Similarmente el máximo absoluto de g ocurre en $x = \pm 1$ y tiene el valor $g(\pm 1) = 0$. Además, g es continua y

$$|g'(x)| = |\frac{2\ x}{3}| \leq \frac{2}{3} \qquad \text{para todo } x \in [-1, 1]\ ,$$

así que g satisface las hipótesis del Teorema VII.1 y tiene un único punto fijo en $[-1, 1]$. En este ejemplo, el único punto fijo p en el intervalo $[-1, 1]$ puede determinarse exactamente. Si

$$p = g(p) = \frac{p^2 - 1}{3}$$

entonces, $p^2 - 3\ p - 1 = 0$, lo cual implica que $p = \frac{3-\sqrt{13}}{2}$. Nótese que g también tiene un punto fijo único $p = \frac{3+\sqrt{13}}{2}$ en el intervalo $[3, 4]$. Sin embargo, $g(4) = 5$ y $g'(4) = 8/3 > 1$; así que g no satisface las hipótesis del Teorema VII.1. Esto muestra que las hipótesis del Teorema VII.1 son suficientes para garantizar un punto fijo único, pero no son necesarias.

Ejemplo 2.

Sea $g(x) = 3^{-x}$. Como $g'(x) = -3^{-x} \ln 3 < 0$ en $[0, 1]$, la función g es decreciente en $[0, 1]$. Entonces, $g(1) = \frac{1}{3} \leq g(x) \leq 1 = g(0)$ para $0 \leq x \leq 1$. Por lo tanto, para $x \in [0, 1]$, $g(x) \in [0, 1]$. Esto implica que g tiene un punto fijo en $[0, 1]$. Como

$$g'(0) = -\ln 3 = -1.098612289\ ,$$

$|g'(x)| \not< 1$ en $[0, 1]$ y el Teorema VII.1 no puede ser usado para determinar la unicidad. Sin embargo, g es decreciente, así que está claro que el punto fijo debe ser único.

Geométricamente, el método de iteración puede explicarse de la siguiente manera: dibújese sobre un plano xy las gráficas de las funciones $y = x$ e $y = g(x)$. Cada raíz real p de la ecuación $x = g(x)$ es la abscisa del punto de intersección M de la curva $y = g(x)$ con la línea recta $y = x$ (ver figuras 2 y 3).

Comenzando a partir de un punto A_0 $(p_0, g(p_0))$, construyamos la línea poligonal $A_0\ B_0\ A_1\ B_1 ...$ (*escalera*), cuyos segmentos son alternativamente paralelos al eje x y al

eje y, los vértices A_0, A_1, A_2, ... caen sobre la curva $y = g(x)$, y los vértices B_0, B_1, B_2, B_3, ... caen sobre la línea recta $y = x$. Las abscisas comunes de los puntos A_1 y B_0, y A_2 y B_1, ..., evidentemente serán las aproximaciones sucesivas p_1, p_2, ... a la raíz p. También es posible tener una línea poligonal diferente A_0 B_0 A_1 B_1 ... (*espiral*).

Figura 2

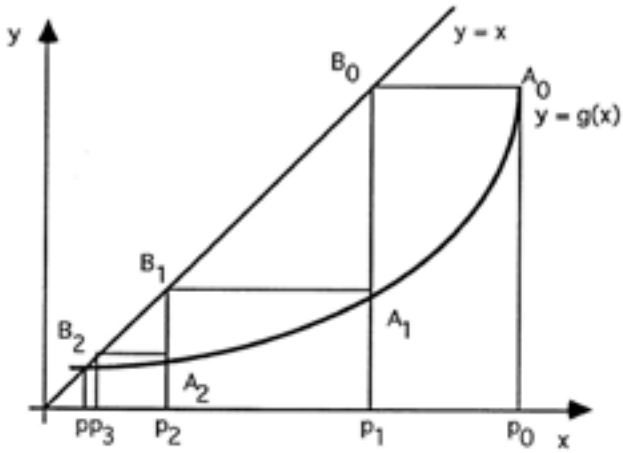
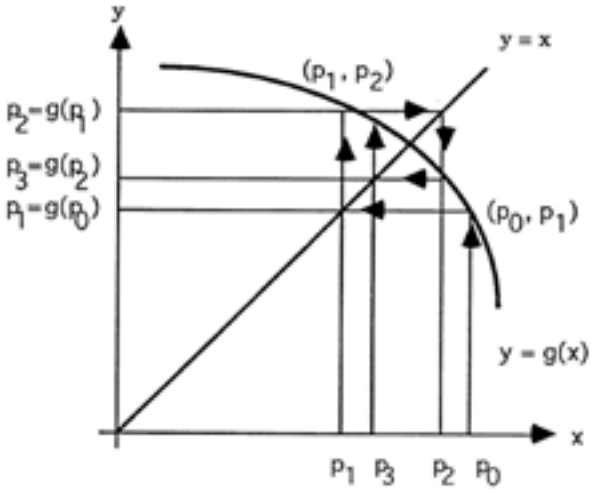


Figura 3

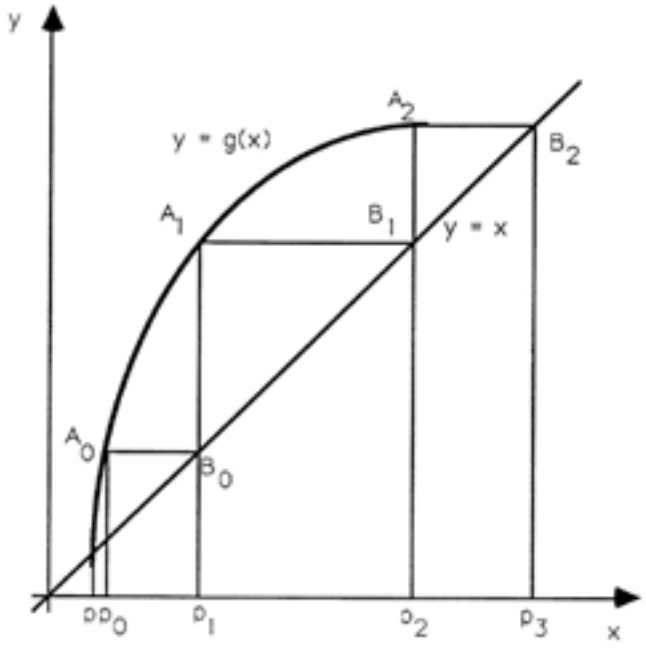


Evidentemente se tiene la solución escalera si la derivada $g'(x)$ es positiva (ver figura 2), y la solución espiral si $g'(x)$ es negativa (ver figura 3).

En las figuras 2 y 3, la curva $y = g(x)$ se "inclina" en la vecindad de la raíz p, es decir, $|g'(x)| < 1$ y el proceso de iteración converge. No obstante, si consideramos el caso en que $|g'(x)| > 1$, entonces el proceso de iteración puede ser divergente (ver figura 4). Por

consiguiente, para aplicar de una manera práctica el método de iteración del punto fijo, hemos de asegurarnos de que se cumplen las condiciones de suficiencia de convergencia del proceso de iteración.

Figura 4



2. ALGORITMO Y EJEMPLOS

Para aproximar el punto fijo de una función g, escogemos una aproximación inicial p_0 y generamos la sucesión $\{p_n\}_{n=0}^{\infty}$ tomando $p_n = g(p_{n-1})$ para cada $n \geq 1$. Si la sucesión converge a p y g es continua, entonces

$$p = \lim_{n\to\infty} p_n = \lim_{n\to\infty} g(p_{n-1}) = g\big(\lim_{n\to\infty} p_{n-1}\big) = g(p) \ ,$$

y se obtiene una solución de $x = g(x)$. Esta técnica se llama **técnica iterativa de punto fijo** ó **iteración funcional**. El procedimiento está detallado en el algoritmo conocido como algoritmo de punto fijo y está descrito en las figuras 2 y 3.

Algoritmo de punto fijo.

==========

Para encontrar una solución de $g(p) = p$ dada una aproximación inicial p_0:

Entrada: aproximación inicial p_0; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada p ó mensaje de fracaso.

Paso 1: tomar $i = 1$;

Paso 2: mientras que $i \leq N_0$ seguir pasos 3–6;

Paso 3: tomar $p = g(p_0)$ *(calcular p_i)*;

Paso 4: si $|p - p_0| < TOL$ entonces SALIDA (p);
(procedimiento completado satisfactoriamente) PARAR;

Paso 5: tomar $i = i + 1$

Paso 6: tomar $p_0 = p$ *(redefinir p_0)*;

Paso 7: SALIDA ($'$*El método fracasó después de N_0 iteraciones, $N_0 =$ $'$*, N_0);
(procedimiento completado sin éxito); PARAR.

==========

Ejemplo 3.

La ecuación $x^3+4\,x^2-10 = 0$ tiene una sola raíz en $[1, 2]$. Existen muchas maneras de cambiar la ecuación a la forma $x = g(x)$, efectuando manipulaciones algebráicas simples. Debe verificarse que el punto fijo de la función $g(x)$ es en realidad una solución de la ecuación original.

(a) $x = g_1(x) = x - x^3 - 4\,x^2 + 10$,

(b) $x = g_2(x) = \left(\frac{10}{x} - 4\,x\right)^{1/2}$,

(c) $x = g_3(x) = \frac{1}{2}\,(10 - x^3)^{1/2}$,

(d) $x = g_4(x) = \left(\frac{10}{4+x}\right)^{1/2}$,

(e) $x = g_5(x) = x - \frac{x^3+4\,x^2-10}{3\,x^2+8\,x}$.

Tabla 1

n	p_n (a)	p_n (b)	p_n (c)	p_n (d)	p_n (e)
0	1.5	1.5	1.5	1.5	1.5
1	-0.875	0.8165	1.286953768	1.348399725	1.373333333
2	6.732	2.9969	1.402540804	1.367376372	1.365262015
3	-469.7	$(-8.65)^{1/2}$	1.345458374	1.364957015	1.365230014
4	1.03×10^8		1.375170253	1.365264748	1.365230013
5			1.360094193	1.365225594	
6			1.367846968	1.365230576	
7			1.363887004	1.365229942	
8			1.365916733	1.365230023	
9			1.364878217	1.365230012	
10			1.365410061	1.365230014	
15			1.365223680	1.365230013	
20			1.365230236		
25			1.365230006		
30			1.365230014		

Con $p_0 = 1.5$, la tabla 1 muestra los resultados del método de iteración de punto fijo para las cinco alternativas para g.

La raíz real es 1.365230013, como se hizo notar en el ejemplo 1 del capítulo VI. Comparando los resultados con el algoritmo de bisección dado en aquel ejemplo, se puede ver que se han obtenido excelentes resultados para los casos (c), (d) y (e), mientras que con la técnica de bisección se necesitan 27 iteraciones para lograr esta precisión. Es

interesante notar que la elección (a) produce divergencia y (b) se vuelve indefinida debido a que lleva a la raíz cuadrada de un número negativo.

Este ejemplo ilustra la necesidad de un procedimiento que garantice que la función g converja a una solución de $x = g(x)$ y que escoja también a g de tal manera que haga la convergencia tan rápida como sea posible. El Teorema siguiente es el primer paso para determinar este procedimiento.

Teorema VII.2

Sea $g \in \mathcal{C}[a,b]$ y supongamos que $g(x) \in [a,b] \ \ \forall\ x \in [a,b]$. Además, supongamos que g' existe en (a,b) con

$$|g'(x)| \leq k < 1 \qquad \text{para toda } \ x \in (a,b)\ . \tag{$VII.1$}$$

Si p_0 es cualquier número en $[a,b]$, entonces la sucesión definida por

$$p_n = g(p_{n-1})\ , \qquad n \geq 1\ ,$$

converge al único punto fijo p en $[a,b]$.

Demostración: por el Teorema VII.1, existe un punto fijo único en $[a,b]$. Como g manda a $[a,b]$ a él mismo, la sucesión $\{p_n\}_{n=0}^{\infty}$ está definida para toda $n \geq 0$ y $p_n \in [a,b]$ para toda n. Usando la desigualdad $(VII.1)$ y el Teorema del valor medio,

$$|p_n - p| = |g(p_{n-1}) - g(p)| = |g'(\xi)|\ |p_{n-1} - p| \leq k\ |p_{n-1} - p|\ , \tag{$VII.2$}$$

donde $\xi \in (a,b)$. Aplicando la desigualdad $(VII.2)$ inductivamente resulta:

$$|p_n - p| \leq k\ |p_{n-1} - p| \leq k^2\ |p_{n-2} - p| \leq\ \ldots\ \leq k^n\ |p_0 - p|\ . \tag{$VII.3$}$$

Como $k < 1$,

$$\lim_{n\to\infty} |p_n - p| \leq \lim_{n\to\infty} k^n\ |p_0 - p| = 0$$

y $\{p_n\}_{n=0}^{\infty}$ converge a p. **c.q.d.**

El Teorema permanece válido si la función $g(x)$ es definida y diferenciable en el intervalo infinito $-\infty < x < +\infty$, y la desigualdad $(VII.1)$ se cumple cuando $x \in (-\infty, +\infty)$.

Nótese que en las condiciones del Teorema VII.2, el método del punto fijo converge para cualquier valor inicial p_0 en $[a,b]$. Por esta razón es **autocorrector**, esto es, un error individual en los cálculos que no vaya por encima de los límites del intervalo $[a,b]$ no afecterá el resultado final, ya que un valor erróneo puede ser considerado como un nuevo valor inicial p_0. Unicamente se habrá trabajado más. La propiedad de autocorrección hace que el método de iteración del punto fijo sea uno de los más fiables. Naturalmente, los errores sistemáticos al aplicar este método pueden hacer que no se obtenga el resultado requerido.

Corolario VII.3

Si g satisface las hipótesis del Teorema VII.2, una cota para el error involucrado al usar p_n para aproximar a p está dada por

$$|p_n - p| \leq k^n \ \max\{p_0 - a, b - p_0\} \qquad \text{para cada } n \geq 1 \ . \tag{$VII.4$}$$

Demostración: de la desigualdad $(VII.3)$,

$$\begin{aligned} |p_n - p| &\leq k^n \ |p_0 - p| \\ &\leq k^n \ \max\{p_0 - a, b - p_0\} \ , \end{aligned}$$

ya que $p \in [a, b]$ y $p_0 \in [a, b]$. **c.q.d.**

Corolario VII.4

Si g satisface las hipótesis del Teorema VII.2, entonces

$$|p_n - p| \leq \frac{k^n}{1-k} \ |p_0 - p_1| \qquad \text{para todo } n \geq 1 \ . \tag{$VII.5$}$$

Demostración: para $n \geq 1$, el procedimiento usado en la demostración del Teorema VII.2 implica que

$$|p_{n+1} - p_n| = |g(p_n) - g(p_{n-1})| \leq k \ |p_n - p_{n-1}| \leq ... \leq k^n \ |p_1 - p_0| \ .$$

Por lo tanto, para $m > n \geq 1$,

$$\begin{aligned} |p_m - p_n| &= |p_m - p_{m-1} + p_{m-1} - ... - p_{n+1} + p_{n+1} - p_n| \\ &\leq |p_m - p_{m-1}| + |p_{m-1} - p_{m-2}| + ... + |p_{n+1} - p_n| \\ &\leq k^{m-1} \ |p_1 - p_0| + k^{m-2} \ |p_1 - p_0| + ... + k^n \ |p_1 - p_0| \\ &= k^n \ (1 + k + k^2 + ... + k^{m-n-1}) \ |p_1 - p_0| \ . \end{aligned}$$

Por el Teorema VII.2, $\lim\limits_{m \to \infty} p_m = p$, así que

$$|p - p_n| = \lim_{m \to \infty} |p_m - p_n| \leq k^n \ |p_1 - p_0| \sum_{i=0}^{\infty} k^i = \frac{k^n}{1-k} \ |p_1 - p_0| \ .$$

c.q.d.

Ambos corolarios relacionan la rapidez de convergencia con la cota k de la primera derivada.

Está claro que la rapidez de convergencia depende del factor $k^n/(1-k)$, y que cuanto más peque˜no se pueda hacer k, más rápida será la convergencia. La convergencia puede ser muy lenta si k es próximo a 1.

Nótese que existe una extendida opinión de que, si al utilizar el método del punto fijo, dos aproximaciones sucesivas p_{n-1} y p_n coinciden dentro de la exactitud especificada

ε (por ejemplo, los primeros m decimales están estabilizados en estas aproximaciones), entonces se cumple $p \approx p_n$ con la misma exactitud (esto es, en particular el número aproximado p_n tiene m cifras exactas). En el caso general esta afirmación es errónea. Es más, resulta fácil demostrar que si $g'(x)$ está próxima a la unidad, entonces la cantidad $|p - p_n|$ puede ser grande, aún cuando $|p_n - p_{n-1}|$ sea extremadamente pequeña.

Los métodos de punto fijo del ejemplo 3 serán reconsiderados tomando en cuenta los resultados descritos en el Teorema VII.2.

Ejemplo 4.

(a) Cuando $g_1(x) = x - x^3 - 4\,x^2 + 10$, $g_1'(x) = 1 - 3\,x^2 - 8\,x$. No hay ningún intervalo $[a, b]$, conteniendo a p, para el cual $|g_1'(x)| < 1$. Aunque el Teorema VII.2 no garantiza que el método debe fracasar para esta elección de g, no hay ninguna razón para sospechar la convergencia.

(b) Con $g_2(x) = [10/x - 4\,x]^{1/2}$, vemos que g_2 no manda al intervalo $[1, 2]$ a $[1, 2]$ y la sucesión $\{p_n\}_{n=0}^{\infty}$ no está definida con $p_0 = 1.5$. Más aún, no hay ningún intervalo que contenga a p tal que $|g_2'(x)| < 1$, ya que $|g_2'(p)| \approx 3.43$.

(c) Para la función $g_3(x) = \frac{1}{2}(10 - x^3)^{1/2}$, $g_3'(x) = -\frac{3}{4}\,x^2\,(10 - x^3)^{-1/2} < 0$ en $[1, 2]$, así que g_3 es estrictamente decreciente en $[1, 2]$. Sin embargo, $|g_3'(2)| \approx 2.12$, así que la desigualdad $(VII.1)$ no se satisface en $[1, 2]$. Examinando más de cerca la sucesión $\{p_n\}_{n=0}^{\infty}$ comenzando con $p_0 = 1.5$ podemos ver que es suficiente considerar el intervalo $[1, 1.5]$ en vez de $[1, 2]$. En este intervalo sigue siendo cierto que $g_3'(x) < 0$ y que g es estrictamente decreciente, pero, además $1 < 1.28 \approx g_3(1.5) \leq g_3(x) \leq g_3(1) = 1.5$ para todo $x \in [1, 1.5]$. Esto demuestra que g_3 manda al intervalo $[1, 1.5]$ a sí mismo. Como también es cierto que $|g_3'(x)| < |g_3'(1.5)| \approx 0.66$ en este intervalo, el Teorema VII.2 confirma la convergencia de la cual ya estábamos enterados.

Las otras partes del ejemplo 3 se pueden manejar de una manera similar.

Teorema VII.5

Sea $g(x)$ una función definida y diferenciable en un intervalo $[a, b]$, y supongamos que la ecuación

$$x = g(x)$$

tenga una raíz p situada en un intervalo más pequeño $[\alpha, \beta]$, donde $\alpha = a + \frac{1}{3}\,(b - a)$ y $\beta = b - \frac{1}{3}\,(b - a)$. En este caso si (a) $|g'(x)| \leq k < 1$ para $a < x < b$ y (b) la aproximación inicial $p_0 \in [\alpha, \beta]$, entonces:

(1) todas las aproximaciones sucesivas caen en el intervalo (a, b): $p_n = g(p_{n-1}) \in (a, b)$, $n = 1, 2, ...$

(2) el proceso de aproximaciones sucesivas es convergente, es decir, existe $\lim_{n\to\infty} p_n = p$ y p es la única raíz en el intervalo $[a, b]$ de la ecuación dada;

(3) el supuesto $(VII.5)$ es válido.

Demostración: (1) en efecto, supongamos

$$p_0 \in [\alpha, \beta]\ .$$

Entonces, la ecuación

$$p_1 = g(p_0)$$

tiene evidentemente significado. Utilizando la ecuación

$$p = g(p)$$

tendremos, en base al Teorema de valor medio,

$$|p_1 - p| = |g(p_0) - g(p)| = |p_0 - p|\ |g'(\overline{p}_0)| \leq k\ (\beta - \alpha) < \frac{b-a}{3}\ .$$

Por consiguiente $p_1 \in (a, b)$. Por lo general, si $p_{n-1} \in (a, b)$ $(n = 1, 2, ...)$ y $|p_{n-1} - p| < \frac{b-a}{3}$, entonces

$$p_n = g(p_{n-1})$$

es significativa y

$$|p_n - p| = |g(p_{n-1}) - g(p)| = |p_{n-1} - p|\ |g'(\overline{p}_{n-1})| \leq k\ (\beta - \alpha) < \frac{b-a}{3}\ .$$

En consecuencia, $p_n \in (a, b)$, donde $n = 1, 2, 3, ...$

Las demostraciones de las afirmaciones (2) y (3) son completamente análogas a las del Teorema VII.1. **c.q.d.**

Teorema VII.6

Sea $g(x)$ una función definida y diferenciable en un intervalo $[a, b]$, y supongamos que la ecuación

$$x = g(x)$$

tenga una raíz p situada en el intervalo $[a, b]$. Además supongamos que la derivada $g'(x)$ conserva el signo y la desigualdad $(VII.1)$ sea válida. Por consiguiente,
(1) si la derivada $g'(x)$ es positiva, las aproximaciones sucesivas

$$p_n = g(p_{n-1})\ , \quad (n = 1, 2, ...)\ , \quad p_0 \in (a, b)$$

convergen monótonamente hacia la raíz p.
(2) Sin embargo, si la derivada $g'(x)$ es negativa, las aproximaciones sucesivas oscilan entonces alrededor de la raíz p.

Demostración: (1) en efecto, hagamos $0 \leq g'(x) \leq k < 1$ y, por ejemplo,

$$p_0 < p\ .$$

En tal caso

$$p_1 - p = g(p_0) - g(p) = (p_0 - p)\ g'(\xi_1) < 0\ ,$$

donde $\xi_1 \in (p_0, p)$, y

$$|p_1 - p| \leq k\ |p_0 - p| < |p_0 - p|\ .$$

En consecuencia,

$$p_0 < p_1 < p \ .$$

Utilizando el método de inducción matemática, obtenemos

$$p_0 < p_1 < p_2 < ... < p \ .$$

Un resultado análogo se obtiene cuando $p_0 > p$.
(2) Hagamos $-1 < -k \leq g'(x) \leq 0$ y, por ejemplo, $p_0 < p$; $p_1 = g(p_0) \in (a, b)$. Tenemos,

$$p_1 - p = g(p_0) - g(p) = (p_0 - p)\ g'(\xi_1) > 0$$

es decir, $p_1 > p$ y $|p_1 - p| < |p_0 - p|$. Repitiendo estos argumentos para las aproximaciones p_1, p_2, ..., tenemos

$$p_0 < p_2 < ... < p < ... < p_3 < p_1 \ .$$

De este modo, las aproximaciones sucesivas oscilarán alrededor de la raíz p. **c.q.d.**

De este modo, si la derivada $g'(x)$ es positiva, solamente es necesario elegir la aproximación inicial p_0 de forma que pertenezca al entorno (a, b) de la raíz p por la que estamos interesados; todas las aproximaciones restantes p_n $(n = 1, 2, ...)$ caerán automáticamente en este entorno y se acercarán monótonamente hacia la raíz p a medida que n aumente.

Por otra parte, en el caso de una derivada negativa $g'(x)$, si dos aproximaciones p_0 y p_1 pertenecen al entorno (a, b) de la raíz p, todas las demás aproximaciones p_n $(n = 1, 2, ...)$ pertenecerán también al mismo intervalo; la secuencia $\{p_n\}$ *estrangula* la raíz p. Además, en este caso

$$|p - p_n| \leq |p_n - p_{n-1}|$$

es decir, los dígitos estabilizados de la aproximación p_n pertenecen definitivamente a la raíz exacta p.

Nótese que dada una ecuación

$$f(x) = 0 \qquad (VII.6)$$

esta puede escribirse de la forma

$$x = g(x) \qquad (VII.7)$$

eligiendo la función $g(x)$ de diferentes maneras.

La notación $(VII.7)$ no deja de tener su importancia; en ciertos casos $|g'(x)|$ demostrará ser pequeña en la vecindad de la raíz p, en otros será grande. Para el método de iteración del punto fijo, la representación de $(VII.7)$ más ventajosa es aquella en la cual la desigualdad $(VII.1)$ es válida, y cuanto menor sea el número k, más rápida será, hablando en términos generales, la convergencia de las aproximaciones sucesivas a la raíz p.

Estudiaremos a continuación una técnica bastante general para reducir la ecuación $(VII.6)$ a la forma $(VII.7)$, y para la cual se asegure la validez de la desigualdad $(VII.1)$. Supongamos que la raíz deseada p de la ecuación cae en el intervalo $[a, b]$, y

$$0 < m_1 \leq f'(x) \leq M_1$$

para $a \leq x \leq b$. [Si la derivada $f'(x)$ es negativa, entonces consideraremos la ecuación $-f(x) = 0$ en lugar de $(VII.6)$]. En particular, podemos tomar para m_1 el valor más pequeño de la derivada $f'(x)$ en el intervalo $[a, b]$, cuyo valor debe ser positivo, y para M_1 el valor más grande de $f'(x)$ en el intervalo $[a, b]$. Reemplácese $(VII.6)$ por la ecuación equivalente

$$x = x - \lambda\ f(x) \qquad (\lambda > 0)\ .$$

Podemos establecer $g(x) = x - \lambda\ f(x)$. Elijamos el parámetro λ de tal manera que en la vecindad dada de la raíz $[a, b]$ sea válida la desigualdad

$$0 \leq g'(x) = 1 - \lambda\ f'(x) \leq k < 1\ ,$$

de donde tenemos

$$0 \leq 1 - \lambda\ M_1 \leq 1 - \lambda\ m_1 \leq k\ .$$

En consecuencia, podemos elegir

$$\lambda = \frac{1}{M_1} \qquad \text{y} \qquad k = 1 - \frac{m_1}{M_1} < 1\ .$$

Ejemplo 5.

Hállese la mayor raíz p positiva de la ecuación

$$x^3 + x = 1000\ ,$$

con un error menor que 10^{-4}.

Un cálculo por encima nos ofrece el valor aproximado de la raíz $p_0 = 10$; es evidente que $p < p_0$.

Podemos reescribir la ecuación de las formas:

$$x = 1000 - x^3\ , \quad \text{ó} \quad x = \frac{1000}{x^2} - \frac{1}{x}\ , \quad \text{ó} \quad x = \sqrt[3]{1000 - x}\ .$$

La versión más adecuada es la tercera, ya que tomando el intervalo (9, 10) como intervalo principal y estableciendo

$$g(x) = \sqrt[3]{1000 - x}\ ,$$

tenemos

$$g'(x) = \frac{-1}{3\ \sqrt[3]{(1000 - x)^2}}\ ,$$

de donde

$$|g'(x)| \leq \frac{1}{3\ \sqrt[3]{990^2}} \approx \frac{1}{300} = k\ .$$

Calculemos las aproximaciones sucesivas p_n utilizando las fórmulas

$$y_n = 1000 - p_n\ , \qquad p_{n+1} = \sqrt[3]{y_n}\ , \qquad n = 0, 1, 2, ...$$

Los valores hallados de este modo se encuentran reflejados en la tabla siguiente:

Tabla 2

n	p_n	y_n
0	10	990
1	9.966555	990.033445
2	9.966667	990.033333
3	9.966667	

Como $1 - k \approx 1$, podemos establecer $p = 9.9667$ con un error menor que 10^{-4}.

Aplicamos ahora la técnica dada previamente:

$$f(x) = x^3 + x - 1000 \qquad \text{y} \qquad f'(x) = 3\ x^2 + 1$$

entonces

$$m_1 = 244 \leq f'(x) \leq 301 = M_1\ .$$

Definimos

$$x = g(x) = x - \lambda(x^3 + x - 1000)\ ,$$

donde

$$\lambda = \frac{1}{M_1} = \frac{1}{301} \approx 0.00332$$

y por la cual $k = 1 - \frac{m_1}{M_1} = 1 - \frac{244}{301} \approx 0.19$.

Los valores hallados de este modo se encuentran reflejados en la tabla siguiente:

Tabla 3

n	p_n
0	10
1	9.9668
2	9.966668
3	9.966667
4	9.966667

Indicaremos ahora otra técnica para acelerar la convergencia del proceso de iteración la cual puede ser útil en ciertos casos. Supóngase que disponemos de una ecuación

$$x = g(x)$$

tal que la desigualdad

$$|g'(x)| \geq q > 1$$

sea cierta dentro del entorno de la raíz deseada p. El proceso de iteración del punto fijo divergirá entonces para esta ecuación. Pero si la ecuación dada es sustituida por la ecuación equivalente

$$x = \phi(x)\ ,$$

donde $\phi(x) = g^{-1}(x)$ es la función inversa, tendremos una ecuación para la cual el proceso de iteración converge, ya que

$$|\phi'(x)| = \Big|\frac{1}{g'(\phi(x))}\Big| \leq \frac{1}{q} = k < 1 \ .$$

Ejemplo 6.

La ecuación

$$f(x) = x^3 - x - 1 = 0$$

tiene una raíz $p \in (1, 2)$, ya que $f(1) = -1 < 0$ y $f(2) = 5 > 0$.

Podemos reescribir la ecuación como

$$x = x^3 - 1$$

en este caso

$$g(x) = x^3 - 1 \qquad \text{y} \qquad g'(x) = 3\ x^2$$

y por tanto $g'(x) \geq 3$ para $1 \leq x \leq 2$ y, en consecuencia, no se cumplen las condiciones de convergencia del proceso de iteración.

Si escribimos la ecuación de la forma

$$x = \sqrt[3]{x+1}$$

tendremos

$$\phi(x) = \sqrt[3]{x+1} \qquad \text{y} \qquad \phi'(x) = \frac{1}{3\ \sqrt[3]{(x+1)^2}} \ .$$

Por consiguiente $0 < \phi'(x) < \frac{1}{3\ \sqrt[3]{4}} < \frac{1}{4}$ para $1 \leq x \leq 2$ y, de aquí que, para la nueva ecuación el proceso de iteración convergerá rápidamente.

Tabla 4

n	p_n	n	p_n
0	1.33	4	1.324725
1	1.325720	5	1.324719
2	1.324908	6	1.324718
3	1.324754	7	1.324718

EJERCICIOS.

1. Use el método de iteración del punto fijo para determinar una solución exacta (en 10^{-2}) de la ecuación $2\sin \pi x + x = 0$ en $[1, 2]$. Use $p_0 = 1$.

2. Usar el Teorema VII.1 para demostrar que $g(x) = 2^{-x}$ tiene un punto fijo único en $[1/3, 1]$. Use la iteración del punto fijo para encontrar una aproximación al punto fijo, con una precisión de 10^{-4}. Use el corolario VII.3 ó el VII.4 para estimar el número de iteraciones requeridas para alcanzar 10^{-4} de precisión, y compare esta estimación teórica con el número realmente necesario.

3. Usar el Teorema VII.1 para demostrar que $g(x) = \pi + 0.5\sin x$ tiene un punto fijo único en $[0, 2\pi]$. Use la iteración del punto fijo para encontrar una aproximación al punto fijo, con una precisión de 10^{-2}. Use el corolario VII.3 ó el VII.4 para estimar el número de iteraciones requeridas para alcanzar 10^{-2} de precisión, y compare esta estimación teórica con el número realmente necesario.

4. Para cada una de las siguientes ecuaciones, determine un intervalo $[a, b]$ en el cual la iteración del punto fijo converja. Estime el número de iteraciones necesarias para obtener aproximaciones exactas en 10^{-5}, y efectúe los cálculos.

a) $x = \frac{2-e^x+x^2}{3}$;
b) $x = \sqrt{\frac{1}{3}e^x}$;
c) $x = 5^{-x}$;
d) $x = 6^{-x}$;
e) $x = 1.75 + \frac{4\ x-7}{x-2}$;
f) $x = \frac{5}{x^2} + 2$.

5. Demuestre que la sucesión definida por

$$x_n = \frac{1}{2}\left(x_{n-1} + \frac{2}{x_{n-1}}\right) , \qquad \text{para } \ n \geq 1 ,$$

converge a $\sqrt{2}$ para cualquier $x_0 > 0$.

CAPITULO VIII. EL METODO DE LA SECANTE

1. INTRODUCCION Y METODO

Utilizando los supuestos de los capítulos anteriores, daremos en este capítulo un procedimiento más rápido para hallar una raíz p de la ecuación $f(x) = 0$ que caiga en un intervalo especificado $[a, b]$ tal que $f(a) \cdot f(b) < 0$. En lugar de dividir por la midad el intervalo $[a, b]$ (método de bisección, Cap. XV), es mejor dividirlo en la relación $-f(a) : f(b)$. Esto ofrece un valor aproximado de la raíz

$$p_1 = a + h_1 = b - \tilde{h}_1 \ , \qquad (VIII.1)$$

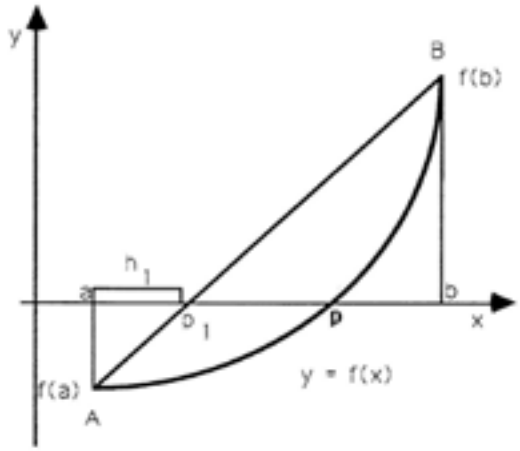
siendo

$$h_1 = -\frac{f(a)}{-f(a) + f(b)} \ (b - a) \ , \qquad \tilde{h}_1 = \frac{f(b)}{-f(a) + f(b)} \ (b - a) \ . \qquad (VIII.2)$$

Aplicando este procedimiento al intervalo $[a, p_1]$ o $[p_1, b]$ en cuyos extremos la función $f(x)$ tenga signos opuestos, tendremos una segunda aproximación p_2 de la raíz, etc. Este método es conocido con el nombre de **método de las partes proporcionales o método de la secante**.

Geométricamente, el método de las partes proporcionales es equivalente a sustituir la curva $y = f(x)$ por una cuerda que pase por los puntos $A\ [a, f(a)]$ y $B\ [b, f(b)]$.

Figura 1



En efecto la ecuación de la secante que pasa por A y B es

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

De aquí , considerando $x = p_1$ e $y = 0$, tenemos

$$p_1 = a - \frac{f(a)}{-f(a) + f(b)} \ (b - a) \ .$$

Para probar la convergencia del proceso, consideremos que la raíz está separada y la segunda derivada $f''(x)$ tiene signo constante en el intervalo $[a, b]$.

Supongamos que $f''(x) > 0$ para $a \leq x \leq b$ (el caso $f''(x) < 0$ se reduce a nuestro caso si escribimos la ecuación de la forma $-f(x) = 0$). La curva $y = f(x)$ será convexa hacia abajo, y por tanto estará localizada por debajo de su secante $A\ B$. Son posibles dos casos:

(1) $f(a) > 0$, (ver figura 2),

(2) $f(a) < 0$, (ver figura 3).

Figura 2

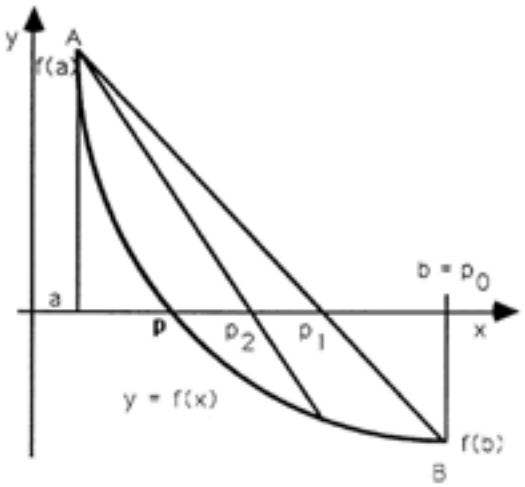
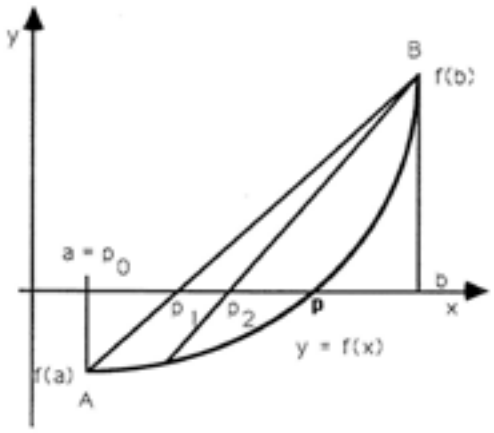


Figura 3



En el primer caso, el extremo a está fijo y las aproximaciones sucesivas:

$$p_0 = b\ , \qquad p_{n+1} = p_n - \frac{f(p_n)}{f(p_n) - f(a)}\ (p_n - a)\ , \qquad n = 0, 1, 2, ... \qquad (VIII.3)$$

forman una secuencia monótona decreciente acotada, y

$$a < p < ... < p_{n+1} < p_n < ... < p_1 < p_0 \ .$$

En el segundo caso, el extremo b está fijo y las aproximaciones sucesivas:

$$p_0 = a \ , \quad p_{n+1} = p_n - \frac{f(p_n)}{f(b) - f(p_n)} \ (b - p_n) \ , \qquad n = 0, 1, 2, ... \qquad (VIII.4)$$

forman una secuencia monótona creciente acotada, y

$$p_0 < p_1 < ... < p_n < p_{n+1} < ... < p < b \ .$$

Resumiendo, sacamos las siguientes conclusiones:

(1) el extremo fijado es aquél para el cual el signo de la función $f(x)$ coincide con el signo de su segunda derivada $f''(x)$;

(2) las aproximaciones sucesivas p_n caen en el lado de la raíz p, donde el signo de la función $f(x)$ es opuesto al signo de su segunda derivada $f''(x)$.

En ambos casos, cada aproximación sucesiva p_{n+1} está más próxima a la raíz p que la precedente, p_n. Supongamos

$$\overline{p} = \lim_{n \to \infty} p_n \qquad (a < \overline{p} < b)$$

(existe límite, ya que la secuencia $\{p_n\}$ está acotada y es monótona). Pasando al límite en $(VIII.3)$, tenemos para el primer caso

$$\overline{p} = \overline{p} - \frac{f(\overline{p})}{f(\overline{p}) - f(a)} \ (\overline{p} - a) \ ,$$

donde $f(\overline{p}) = 0$. Como viene dado que la ecuación $f(x) = 0$ tiene solamente una raíz p en el intervalo (a, b), se deduce que $\overline{p} = p$.

Mediante el mismo procedimiento puede probarse en $(VIII.4)$, que $\overline{p} = p$ para el segundo caso.

Para hallar una estimación de la exactitud de la aproximación, podemos utilizar la fórmula $(V.2)$

$$|p_n - p| \leq \frac{|f(p_n)|}{m_1} \ ,$$

donde $|f'(x)| \geq m_1$ para $a \leq x \leq b$.

Daremos otra fórmula que permita estimar el error absoluto de un valor aproximado p_n conocidos dos valores sucesivos p_{n-1} y p_n.

Teorema VIII.1

Sea $f \in \mathcal{C}^2[a, b]$ y supongamos que $f(a) \cdot f(b) < 0$, y que la derivada $f'(x)$, continua en el intervalo $[a, b]$ que contiene toda las aproximaciones, conserva el signo y sea tal que

$$0 < m_1 \leq |f'(x)| \leq M_1 < +\infty \ .$$

Entonces se estima el error absoluto de un valor aproximado p_n dado por las relaciones iterativas $(VIII.3)$ ó $(VIII.4)$ como

$$|p - p_n| \leq \frac{M_1 - m_1}{m_1} \, |p_n - p_{n-1}| \, . \qquad (VIII.5)$$

Demostración: para mayor claridad, supongamos que las aproximaciones sucesivas p_n a la raíz exacta p están generadas por la fórmula $(VIII.3)$ (análogamente puede considerarse la fórmula $(VIII.4)$):

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f(p_{n-1}) - f(a)} \, (p_{n-1} - a) \, ,$$

con $n = 1, 2, ...$ y donde el extremo a es fijo. Entonces:

$$-f(p_{n-1}) = \frac{f(p_{n-1}) - f(a)}{p_{n-1} - a} \, (p_n - p_{n-1}) \, .$$

Teniendo en cuenta el hecho de que $f(p) = 0$, tenemos

$$f(p) - f(p_{n-1}) = \frac{f(p_{n-1}) - f(a)}{p_{n-1} - a} \, (p_n - p_{n-1}) \, .$$

Utilizando el Teorema de Valor Medio, tendremos

$$(p - p_{n-1}) \; f'(\xi_{n-1}) = (p - p_n + p_n - p_{n-1}) \; f'(\xi_{n-1}) = (p_n - p_{n-1}) \; f'(\overline{p}_{n-1}) \, ,$$

donde $\xi_{n-1} \in (p_{n-1}, p)$ y $\overline{p}_{n-1} \in (a, p_{n-1})$. De aquí que

$$(p - p_n) \; f'(\xi_{n-1}) = [f'(\overline{p}_{n-1}) - f'(\xi_{n-1})] \; [p_n - p_{n-1}] \, ,$$

y entonces:

$$|p - p_n| = \frac{|f'(\overline{p}_{n-1}) - f'(\xi_{n-1})|}{|f'(\xi_{n-1})|} \, |p_n - p_{n-1}| \, .$$

Como $f'(x)$ tiene signo constante en el intervalo $[a, b]$ y $\overline{p}_{n-1} \in [a, b]$ y $\xi_{n-1} \in [a, b]$, tenemos sencillamente

$$|f'(\overline{p}_{n-1}) - f'(\xi_{n-1})| \leq M_1 - m_1 \, .$$

Deducimos, por tanto, que

$$|p - p_n| \leq \frac{M_1 - m_1}{m_1} \, |p_n - p_{n-1}| \, ,$$

donde podemos tomar respectivamente para m_1 y M_1 los valores menor y mayor del módulo de la derivada $f'(x)$ en el intervalo $[a, b]$. **c.q.d.**

Si el intervalo $[a, b]$ es tan estrecho que se cumple la desigualdad

$$M_1 \leq 2 \; m_1$$

obtenemos entonces de la fórmula $(VIII.5)$

$$|p - p_n| \leq |p_n - p_{n-1}| \ .$$

En este caso, cuando

$$|p_n - p_{n-1}| \leq \varepsilon \ ,$$

donde ε es la cota de error absoluto especificada, puede garantizarse que

$$|p - p_n| \leq \varepsilon \ .$$

2. ALGORITMO Y EJEMPLOS

Algoritmo de la secante.

==========

Para encontrar una solución de $f(x) = 0$, dada la función f en el intervalo $[a, b]$ donde $f(a)$ y $f(b)$ tienen signo opuesto:

Entrada: extremos a y b; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada p ó mensaje de fracaso.

Paso 1: tomar $i = 2$, y definir:
$p_0 = b$, $q_0 = f(a)$ y $q_1 = f(b)$, si $f(a) > 0$;
$p_0 = a$, $q_0 = f(b)$ y $q_1 = f(a)$, si $f(a) < 0$;

Paso 2: mientras que $i \leq N_0$ seguir pasos 3–6;

Paso 3: tomar *(calcular p_i)*:
$p = p_0 - \frac{q_1}{q_1 - q_0} \ (p_0 - a)$, si $f(a) > 0$;
$p = p_0 - \frac{q_1}{q_0 - q_1} \ (b - p_0)$, si $f(a) < 0$;

Paso 4: si $|p - p_0| < TOL$ entonces SALIDA (p);
(procedimiento completado satisfactoriamente) PARAR;

Paso 5: tomar $i = i + 1$

Paso 6: tomar $p_0 = p$; $q_1 = f(p)$ *(redefinir p_0, q_1)*;

Paso 7: SALIDA ($'$*El método fracasó después de N_0 iteraciones, $N_0 = \ '$*, N_0); *(procedimiento completado sin éxito)*; PARAR.

==========

Para ilustrar el algoritmo de la secante, considérese los siguientes ejemplos.

Ejemplo 1.

Hállese una raíz positiva de la ecuación

$$f(x) \equiv x^3 + 4 \ x^2 - 10 = 0$$

con una exactitud de 0.0002.

Primeramente separamos la raíz. Ya que

$$f(1.3) = -1.043 < 0 \quad \text{y} \quad f(1.4) = 0.584 > 0$$

la raíz deseada p está en el intervalo $(1.3, 1.4)$. Además, estamos en el caso $f(a) < 0$, y entonces consideramos la fórmula $(VIII.4)$ en la cual el extremo b está fijo:

$$p_0 = a \ , \quad p_{n+1} = p_n - \frac{f(p_n)}{f(b) - f(p_n)} \ (b - p_n) \ , \qquad n = 0, 1, 2, \ldots$$

Entonces, tenemos

$$\begin{aligned} p_0 =& 1.3 \ , \\ p_1 =& 1.3 + \frac{1.043}{0.584 + 1.043} \ (1.4 - 1.3) = 1.364105716 \ , \\ f(p_1) =& - 0.01855573934 \ , \\ p_2 =& 1.364105716 + \frac{0.01855573934}{0.584 + 0.01855573934} \ (1.4 - 1.364105716) \\ =& 1.365211083 \ , \\ f(p_2) =& - 0.00031260885 \ . \end{aligned}$$

Como $f'(x) = 3\ x^2 + 8\ x$ y para $p_2 < x < 1.4$ se tiene

$$m_1 = \min_{x \in [p_2, 1.4]} |f'(x)| = f'(p_2) = 16.513092561 \ ,$$

y

$$M_1 = \max_{x \in [p_2, 1.4]} |f'(x)| = f'(1.4) = 17.08 \ .$$

Luego podemos considerar que

$$0 < p - p_2 < \frac{|f(p_2)|}{m_1} \approx 0.189309694014 \times 10^{-4} < 2.0 \times 10^{-4} \ .$$

Obsérvese que la raíz con diez dígitos exacto de la ecuación es $p = 1.365230013$.

Si consideremos la cota $|p - p_2| \leq \frac{M_1 - m_1}{m_1} \ |p_2 - p_1|$, obtedríamos:

$$|p - p_2| \leq \frac{M_1 - m_1}{m_1} \ |p_2 - p_1| \approx 0.37948 \times 10^{-3} \ ,$$

ilustrandonos que el primer estimado es mucho mejor en este caso, dado que con esta segunda cota tendríamos que iterar una vez más.

Ejemplo 2.

Hállese la raíz negativa de la ecuación

$$f(x) \equiv x^4 + 0.1\ x^3 - 0.56\ x^2 + 0.1\ x - 1.56 = 0$$

con una exactitud de 7.5×10^{-3}.

Primeramente separamos la raíz. Ya que

$$f(-1.5) = 1.755 > 0 \quad \text{y} \quad f(-1.0) = -1.32 < 0$$

la raíz deseada p está en el intervalo $(-1.5, -1.0)$. Además, estamos en el caso $f(a) > 0$, y entonces consideramos la fórmula $(VIII.3)$ en la cual el extremo a está fijo:

$$p_0 = b \ , \quad p_{n+1} = p_n - \frac{f(p_n)}{f(p_n) - f(a)} \ (p_n - a) \ , \qquad n = 0, 1, 2, \ldots$$

Entonces, tenemos

$$\begin{aligned}
p_0 &= -\,1.0 \ , \\
p_1 &= -\,1.0 - \frac{1.32}{1.32 + 1.755} \ (-1.0 + 1.5) = -1.214634146 \ , \\
f(p_1) &= -\,0.510234362 \ , \\
p_2 &= -\,1.214634146 - \frac{0.510234362}{1.755 + 0.510234362} \ (-1.214634146 + 1.5) \\
&= -\,1.278911586 \ , \\
f(p_2) &= -\,0.137780231 \ , \\
p_3 &= -\,1.278911586 - \frac{0.137780231}{1.755 + 0.137780231} \ (-1.278911586 + 1.5) \\
&= -\,1.295005167 \ , \\
f(p_3) &= -\,0.03336161091 \ .
\end{aligned}$$

Como $f'(x) = 4\ x^3 + 0.3\ x^2 - 1.12\ x + 0.1$ y para $-1.5 < x < p_3$ se tiene

$$m_1 = \min_{x \in [-1.5, p_3]} |f'(x)| = |f'(p_3)| = 6.633576181 \ ,$$

y

$$M_1 = \max_{x \in [-1.5, p_3]} |f'(x)| = |f'(-1.5)| = 11.045 \ .$$

Luego podemos considerar que

$$0 < p - p_3 < \frac{|f(p_3)|}{m_1} \approx 0.5029 \times 10^{-2} < 7.5 \times 10^{-3} \ .$$

Obsérvese que la raíz con diez dígitos exacto de la ecuación es $p = 1.365230013$.

Si consideremos la cota $|p - p_3| \leq \frac{M_1 - m_1}{m_1} \ |p_3 - p_2|$, obtedríamos:

$$|p - p_3| \leq \frac{M_1 - m_1}{m_1} \ |p_3 - p_2| \approx 0.010702 \ ,$$

ilustrandonos que el primer estimado es mucho mejor también en este caso (pero no siempre es así), dado que con esta segunda cota tendríamos que iterar una vez más.

EJERCICIOS.

1. Aproxime con 10^{-4} de precisión las raíces de las siguientes ecuaciones en los intervalos dados usando el método de la secante:

- (a) $x^3 - 2\,x^2 - 5 = 0$ $[1, 4]$
- (b) $x - \cos x = 0$ $[0, \pi/2]$
- (c) $x^3 + 3\,x^2 - 1 = 0$ $[-4, 0]$
- (d) $x - 0.8 - 0.2 \sin x = 0$ $[0, \pi/2]$

2. Encuentre una raíz aproximada de $x^3 - x - 1 = 0$ en $[1, 2]$ con precisión de 10^{-5} por el método de la secante.

3. Use el método de la secante para aproximar las soluciones de las ecuaciones siguientes con precisión de 10^{-5}:

- (a) $x = \frac{2 - e^x + x^2}{3}$
- (b) $3\,x^2 - e^x = 0$
- (c) $e^x + 2^{-x} + 2\cos x - 6 = 0$
- (d) $x^2 + 10\cos x = 0$

4. Resolver $4\cos x = e^x$ con exactidud de 10^{-4}, usando el método de la secante con $p_0 = 1$.

5. Calcule una aproximación de $\sqrt{3}$ exacta en 10^{-4}, usando el método de la secante.

CAPITULO IX. EL METODO DE NEWTON-RAPHSON

1. INTRODUCCION Y METODO

El método de Newton-Raphson (o simplemente Newton) es uno de los métodos numéricos más conocidos y poderosos para la resolución del problema de búsqueda de raíces de $f(x) = 0$. Para introducir el método de Newton usaremos un enfoque intuitivo basado en el polinomio de Taylor.

Supóngase que la función f es continuamente diferenciable dos veces en el intervalo $[a, b]$; o sea, $f \in \mathcal{C}^2[a, b]$. Sea $\overline{x} \in [a, b]$ una aproximación a la raíz p tal que $f'(\overline{x}) \neq 0$ y $|\overline{x} - p|$ es *pequeño*. Considérese el polinomio de Taylor de primer grado para $f(x)$ alrededor de $\overline{x}$

$$f(x) = f(\overline{x}) + (x - \overline{x})\ f'(\overline{x}) + \frac{(x - \overline{x})^2}{2}\ f''(\zeta(x))\ , \qquad (IX.1)$$

donde $\zeta(x)$ está entre x y $\overline{x}$. Como $f(p) = 0$, la ecuación $(IX.1)$, con $x = p$, nos da

$$0 = f(\overline{x}) + (p - \overline{x})\ f'(\overline{x}) + \frac{(p - \overline{x})^2}{2}\ f''(\zeta(p))\ . \qquad (IX.2)$$

El método de Newton se deriva suponiendo que el término que contiene a $(p - \overline{x})^2$ es despreciable y que

$$0 \approx f(\overline{x}) + (p - \overline{x})\ f'(\overline{x})\ . \qquad (IX.3)$$

Despejando p de esta ecuación resulta:

$$p \approx \overline{x} - \frac{f(\overline{x})}{f'(\overline{x})}\ , \qquad (IX.4)$$

lo cual debe ser una mejor aproximación a p que $\overline{x}$.

El método de Newton-Raphson implica el generar la sucesión $\{p_n\}$ definida por

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}\ , \qquad n \geq 1\ . \qquad (IX.5)$$

Geométricamente, el método de Newton es equivalente a sustituir un arco pequeño de la curva $y = f(x)$ por una tangente trazada por un punto de la curva. Supongamos, por definición, que $f''(x) > 0$ para $a \leq x \leq b$ y $f(b) > 0$ (ver figura 1).

Tomemos, por ejemplo, $p_0 = b$ para el cual $f(p_0) \cdot f''(p_0) > 0$. Trácese la tangente a la curva $y = f(x)$ en el punto $B(p_0, f(p_0))$. Como primera aproximación p_1 de la raíz p tomemos la abscisa del punto de intersección de esta tangente con el eje x. Trácese nuevamente una tangente por el punto de coordenadas $(p_1, f(p_1))$, cuya abscisa del punto de intersección con el eje x ofrece una segunda aproximación p_2 de la raíz p, y así sucesivamente.

La ecuación de la tangente en el punto de coordenadas $(p_n, f(p_n))$ $(n = 0, 1, ...)$, es

$$y - f(p_n) = f'(p_n)\ (x - p_n)\ .$$

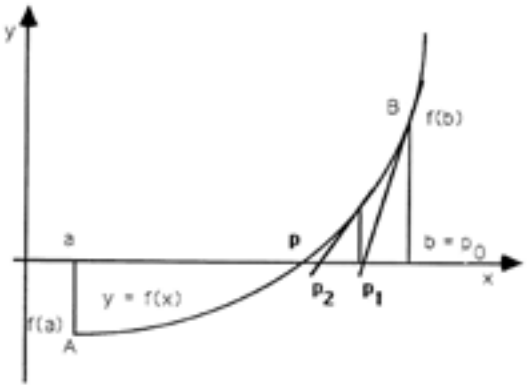
Haciendo $y = 0$ y $x = p_{n+1}$, tendremos la fórmula $(IX.5)$.

Nótese que si en nuestro caso hacemos $p_0 = a$, y por tanto $f(p_0) \cdot f''(p_0) < 0$, y trazamos entonces la tangente a la curva $y = f(x)$ por el punto $A(a, f(a))$, tendremos que el punto p_1' cae fuera del intervalo $[a, b]$; en otras palabras, el procedimiento de Newton no es práctico para este valor inicial. Por tanto, en el caso dado, una *buena* aproximación inicial p_0 es aquella para la cual resulta válida la desigualdad

$$f(p_0) \cdot f''(p_0) > 0 \ .$$

Demostraremos ahora que esta regla es general.

Figura 1



Teorema IX.1

Sea $f \in \mathcal{C}^2[a, b]$. Si $f(a) \cdot f(b) < 0$, y $f'(x)$ y $f''(x)$ son no nulas y conservan el signo para $a \leq x \leq b$, entonces, a partir de la aproximación inicial $p_0 \in [a, b]$ que satisface

$$f(p_0) \cdot f''(p_0) > 0 \ , \qquad (IX.6)$$

es posible, utilizando el método de Newton (fórmula $(IX.3)$), calcular la raíz única p de la ecuación $f(x) = 0$ con cualquier grado de exactitud.

Demostración: supongamos $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$, $f''(x) > 0$ para $a \leq x \leq b$. Por la desigualdad $(IX.6)$ tenemos $f(p_0) > 0$ (podemos, por ejemplo, tomar $p_0 = b$). Por inducción matemática demostraremos que todas las aproximaciones $p_n > p$ $(n = 0, 1, 2, ...)$ y, por consiguiente, $f(p_n) > 0$. En efecto, ante todo, $p_0 > p$.

Establezcamos ahora $p_n > p$. Pongamos

$$p = p_n + (p - p_n) \ .$$

Utilizando la fórmula de Taylor, tendremos

$$0 = f(p) = f(p_n) + f'(p_n)\ (p - p_n) + \frac{1}{2}\ f''(c_n)\ (p - p_n)^2 \ ,$$

donde $p < c_n < p_n$.

Como $f''(x) > 0$, tenemos

$$f(p_n) + f'(p_n)\ (p - p_n) < 0\ ,$$

y, de aquí que

$$p_{n+1} = p_n - \frac{f(p_n)}{f'(p_n)} > p$$

que es lo que se quería demostrar.

Tomando en consideración los signos de $f(p_n)$ y $f'(p_n)$ tenemos, de la fórmula $(IX.5)$, $p_{n+1} < p_n$ $(n = 0, 1, ...)$, es decir, las aproximaciones sucesivas $p_0, p_1, ..., p_n, ...$ forman una secuencia acotada monótona decreciente. Por consiguiente, existe el límite $\overline{p} = \lim_{n\to\infty} p_n$. Pasando al límite en $(IX.5)$, tenemos

$$\overline{p} = \overline{p} - \frac{f(\overline{p})}{f'(\overline{p})}$$

ó $f(\overline{p}) = 0$, de donde $\overline{p} = p$. **c.q.d.**

Por esta razón, al aplicar el método de Newton debe guiarse uno por la **regla** siguiente: para el punto inicial p_0 elíjase el final del intervalo (a, b) asociado con una ordenada del mismo signo que el de $f''(x)$.

Teorema IX.2

Sea $f \in \mathcal{C}(-\infty, +\infty)$, $f(a) \cdot f(b) < 0$, $f'(x) \neq 0$ para $a \leq x \leq b$ y si $f''(x)$ existe en cualquier punto y conserva el signo, entonces puede tomarse cualquier valor $c \in [a, b]$ como aproximación inicial p_0 al utilizarse el método de Newton para hallar una raíz de la ecuación $f(x) = 0$ que caiga en el intervalo (a, b). Se puede, por ejemplo, tomar $p_0 = a$ ó $p_0 = b$.

Demostración: en efecto, supongamos, por ejemplo, $f'(x) > 0$ para $a \leq x \leq b$, $f''(x) > 0$ y $p_0 = c$, donde $a \leq c \leq b$. Si $f(c) = 0$, la raíz $p = c$ y el problema queda resuelto. Si $f(c) > 0$, el razonamiento anterior se cumple y el proceso de Newton con valor inicial c convergerá hacia la raíz $p \in (a, b)$.

Finalmente, si $f(c) < 0$, hallaremos

$$p_1 = p_0 - \frac{f(p_0)}{f'(p_0)} = c - \frac{f(c)}{f'(c)} > c\ .$$

Utilizando la fórmula de Taylor tendremos

$$f(p_1) = f(c) - \frac{f(c)}{f'(c)}\ f'(c) + \frac{1}{2}\ \Big[\frac{f(c)}{f'(c)}\Big]^2\ f''(\overline{c}) = \frac{1}{2}\ \Big[\frac{f(c)}{f'(c)}\Big]^2\ f''(\overline{c}) > 0$$

donde $\overline{c}$ es un cierto valor intermedio entre c y p_1. De este modo

$$f(p_1) \cdot f''(p_1) > 0\ .$$

Además, de la condición $f''(x) > 0$ se deduce que $f'(x)$ es una función creciente y, en consecuencia, $f'(x) > f'(a) > 0$ para $x > a$. Es posible por tanto tomar p_1 como valor inicial del proceso de Newton convergente hacia una cierta raíz $\overline{p}$ de la función $f(x)$ tal que $\overline{p} > c \geq a$. Como la derivada $f'(x)$ es positiva cuando $p > a$, la función $f(x)$ tiene raíz única en el intervalo $(a, +\infty)$, de donde se deduce que

$$\overline{p} = p \in (a, b) \ .$$

Puede establecerse un argumento similar para otras combinaciones de signos de las derivadas $f'(x)$ y $f''(x)$. **c.q.d.**

Nótese que de la fórmula $(IX.5)$ está claro que cuanto mayor sea el valor numérico de la derivada $f'(x)$ en la vecindad de la raíz, tanto menor será la corrección que ha de añadirse a la aproximación $n-$ésima para obtener la aproximación $(n+1)$. El método de Newton es por consiguiente muy conveniente cuando la gráfica de la función tiene una gran pendiente en la vecindad de la raíz dada, pero si el valor numérico de la derivada $f'(x)$ es pequeño cerca de ella, las correcciones serán entonces mayores, y calcular la raíz mediante este procedimiento puede ser un proceso largo o a veces incluso imposible. Resumiendo: no utilice el método de Newton para resolver una ecuación $f(x) = 0$ si la curva $y = f(x)$ es casi horizontal cerca del punto de intersección con el eje x.

El método de Newton es un técnica de iteración funcional $p_n = g(p_{n-1})$, $n \geq 1$ para la cual

$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})} \ , \qquad n \geq 1 \ .$$

Se ve claramente de esta ecuación que el método de Newton no puede continuarse si $f'(p_{n-1}) = 0$ para algún n. Veremos que el método es más eficaz cuando f' está acotada fuera de cero cerca del punto fijo p.

La derivación del método de Newton con serie de Taylor resalta la importancia de una buena aproximación inicial. La suposición crucial al pasar de $(IX.2)$ a $(IX.3)$, es que el término que contiene $(p - \overline{x})^2$ puede ser eliminado. Esta, claramente será una suposición falsa a menos que $\overline{x}$ sea una buena aproximación de p. En particular, si p_0 no está lo suficientemente cerca de la raíz real, el método de Newton puede no converger a la raíz.

El siguiente Teorema de convergencia para el método de Newton ilustra la importancia teórica de la elección de p_0.

Teorema IX.3

Sea $f \in \mathcal{C}^2[a, b]$. Si $p \in [a, b]$ es tal que $f(p) = 0$ y $f'(p) \neq 0$, entonces existe $\delta > 0$ tal que el método de Newton genera una sucesión $\{p_n\}_{n=1}^{\infty}$ que converge a p para cualquier aproximación inicial $p_0 \in [p - \delta, p + \delta]$.

Demostración: la demostración está basada en un análisis del método de Newton como un esquema de iteración funcional $p_n = g(p_{n-1})$, para $n \geq 1$, con

$$g(x) = x - \frac{f(x)}{f'(x)} \ .$$

El objetivo es encontrar, para cualquier valor k en $(0,1)$, un intervalo $[p-\delta, p+\delta]$ tal que g mande al intervalo $[p-\delta, p+\delta]$ a sí mismo y que $|g'(x)| \leq k < 1$ para $x \in [p-\delta, p+\delta]$, donde k es una constante fija en $(0,1)$.

Ya que $f'(p) \neq 0$ y f' es continua, existe $\delta_1 > 0$ tal que $f'(x) \neq 0$ para $x \in [p-\delta_1, p+\delta_1] \subset [a,b]$. Entonces, g está definida y es continua en $[p-\delta_1, p+\delta_1]$. También,

$$g'(x) = 1 - \frac{f'(x)\ f'(x) - f(x)\ f''(x)}{[f'(x)]^2} = \frac{f(x)\ f''(x)}{[f'(x)]^2}$$

para $x \in [p-\delta_1, p+\delta_1]$; y como $f \in \mathcal{C}^2[a,b]$, $g \in \mathcal{C}[p-\delta_1, p+\delta_1]$. De la suposición $f(p) = 0$, se tiene

$$g'(p) = \frac{f(p)\ f''(p)}{[f'(p)]^2} = 0\ .$$

Y como g' es continua, esa ecuación implica que existe un δ con $0 < \delta < \delta_1$, y

$$|g'(x)| \leq k < 1 \qquad \text{para} \quad x \in [p-\delta, p+\delta]\ .$$

Falta todavía demostar que $g : [p-\delta, p+\delta] \to [p-\delta, p+\delta]$. Si $x \in [p-\delta, p+\delta]$, el Teorema del Valor Medio implica que, para algún número ξ entre x y p, $|g(x)-g(p)| = |g'(\xi)|\ |x-p|$. Así que,

$$|g(x)-p| = |g(x)-g(p)| = |g'(\xi)|\ |x-p| \leq k\ |x-p| < |x-p|\ .$$

Como $x \in [p-\delta, p+\delta]$, se sigue que $|x-p| < \delta$ y que $|g(x)-p| < \delta$. Esto implica que $g : [p-\delta, p+\delta] \to [p-\delta, p+\delta]$.

Todas las hipótesis del Teorema VII.2 se satisfacen para $g(x) = x - f(x)/f'(x)$, así que la sucesión $\{p_n\}_{n=1}^{\infty}$ definida por

$$p_n = g(p_{n-1}) \qquad \text{para} \quad n = 1, 2, 3, ...$$

converge a p para cualquier $p_0 \in [p-\delta, p+\delta]$. **c.q.d.**

Para estimar el error de la aproximación p_n de orden n, se puede utilizar la fórmula general (V.2) del capítulo V, $|p-p_n| \leq |f(p_n)|/m_1$, donde m_1 es el valor más pequeño de $|f'(x)|$ en el intervalo $[a,b]$.

Obtendremos ahora otra fórmula para estimar la exactitud de la aproximación p_n. Aplicando la fórmula de Taylor, tenemos

$$f(p_n) = f[p_{n-1} + (p_n - p_{n-1})] =$$

$$= f(p_{n-1}) + f'(p_{n-1})\ (p_n - p_{n-1}) + \frac{1}{2}\ f''(\xi_{n-1})\ (p_n - p_{n-1})^2$$

donde $\xi_{n-1} \in (p_{n-1}, p_n)$. Ya que, en virtud de la definición de la aproximación p_n, tenemos

$$f(p_{n-1}) + f'(p_{n-1})\ (p_n - p_{n-1}) = 0\ ,$$

se deduce que

$$|f(p_n)| \leq \frac{1}{2}\ M_2\ (p_n - p_{n-1})^2$$

donde M_2 es el valor más elevado de $|f''(x)|$ en el entervalo $[a, b]$. En consecuencia, basándose en la fórmula $(V.2)$ tenemos finalmente

$$|p - p_n| \leq \frac{M_2}{2m_1}(p_n - p_{n-1})^2\ . \qquad (IX.7)$$

Si el proceso de Newton converge, entonces $|p_n - p_{n-1}| \to 0$ para $n \to \infty$. Y por tanto para $n \geq N$ tenemos

$$|p - p_n| \leq |p_n - p_{n-1}|\ ,$$

es decir, los decimales iniciales "estabilizados" de las aproximaciones p_{n-1} y p_n son exactos comenzando con una cierta aproximación.

Téngase en cuenta que en el caso general, una coincidencia hasta de ε, de dos aproximaciones sucesivas p_{n-1} y p_n no garantiza que los valores de p_n y la raíz exacta p coincidan con el mismo grado de exactitud.

Obtendremos ahora una fórmula que ligue los errores absolutos de dos aproximaciones sucesivas p_n y p_{n+1}. Utilizando la fórmula de Taylor tendremos

$$0 = f(p) = f(p_n) + f'(p_n)\ (p - p_n) + \frac{1}{2}\ f''(c_n)\ (p - p_n)^2\ ,$$

donde $p < c_n < p_n$, y entonces

$$p = p_n - \frac{f(p_n)}{f'(p_n)} - \frac{1}{2} \cdot \frac{f''(c_n)}{f'(p_n)}\ (p - p_n)^2\ ,$$

y, teniendo en cuenta que $p_{n+1} = p_n - \dfrac{f(p_n)}{f'(p_n)}$, tenemos

$$p - p_{n+1} = -\frac{1}{2} \cdot \frac{f''(c_n)}{f'(p_n)}\ (p - p_n)^2\ ,$$

y consecuentemente,

$$|p - p_{n+1}| \leq \frac{M_2}{2m_1}\ (p - p_n)^2\ . \qquad (IX.8)$$

La fórmula $(IX.8)$ asegura una rápida convergencia del proceso de Newton si la aproximación inicial p_0 es tal que

$$\frac{M_2}{2m_1}\ |p - p_0| \leq k < 1\ .$$

En particular, si

$$\mu = \frac{M_2}{2m_1} \leq 1 \quad \text{y} \quad |p - p_n| < 10^{-m}$$

entonces de $(IX.8)$ tenemos

$$|p - p_{n+1}| < 10^{-2m}\ .$$

Esto es, en este caso, si la aproximación p_n es exacta con m decimales, la siguiente aproximación p_{n+1} lo será como mínimo con $2m$ decimales; en otras palabras, si $\mu \leq 1$, el procedimiento de Newton asegura entonces el doble del número de decimales exactos de la raíz deseada en cada paso.

2. EL ALGORITMO DE NEWTON-RAPHSON

Algoritmo de Newton-Raphson.

==========

Para encontrar una solución de $f(x) = 0$ dada una aproximación inicial p_0:

Entrada: aproximación inicial p_0; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada p ó mensaje de fracaso.

Paso 1: tomar $i = 1$;

Paso 2: mientras que $i \leq N_0$ seguir pasos 3–6;

Paso 3: tomar $p = p_0 - \frac{f(p_0)}{f'(p_0)}$ *(calcular p_i)*;

Paso 4: si $|p - p_0| < TOL$ entonces SALIDA (p);
(procedimiento completado satisfactoriamente) PARAR;

Paso 5: tomar $i = i + 1$

Paso 6: tomar $p_0 = p$. *(redefinir p_0)*;

Paso 7: SALIDA ('*El método fracasó después de N_0 iteraciones, $N_0 =$ '*, N_0);
(procedimiento completado sin éxito); PARAR.

==========

Ejemplo 1.

Supóngase que se necesita una solución a la ecuación $x = \cos x$. Sea $f(x) = \cos x - x$. Entonces

$$f(\frac{\pi}{2}) = -\frac{\pi}{2} < 0 < 1 = f(0) \ ,$$

y, por el Teorema del Valor Intermedio, existe un cero de f en $[0, \pi/2]$. De un estudio gráfico de las ecuaciones $y = x$ e $y = \cos x$ es claro que $f(x) = 0$ tiene una solución única en $[0, \pi/2]$.

Tabla 1

n	p_n
0	0.785398163
1	0.739536134
2	0.739085178
3	0.739085133
4	0.739085133

Como $f'(x) = -\sin x - 1$, el método de Newton tiene la forma

$$p_n = p_{n-1} - \frac{(\cos p_{n-1} - p_{n-1})}{(-\sin p_{n-1} - 1)} \ , \qquad n \geq 1 \ ,$$

donde p_0 tiene todavía que ser seleccionado. Para algunos problemas es suficiente tomar p_0 arbitrariamente, mientras que en otros es importante seleccionar una buena aproximación inicial. Para el problema en consideración, un estudio gráfico sugiere $p_0 = \pi/4$ como una primera aproximación. Con $p_0 = \pi/4$, se generan las aproximaciones de la tabla 1. Con $n = 3$ se obtiene una aproximación excelente.

Ejemplo 2.

Para obtener la solución única de

$$f(x) = x^3 + 4\ x^2 - 10 = 0$$

en el intervalo $[1, 2]$ por el método de Newton generamos la sucesión $\{p_n\}_{n=1}^{\infty}$ dada por

$$p_n = p_{n-1} - \frac{p_{n-1}^3 + 4\ p_{n-1}^2 - 10}{3\ p_{n-1}^2 + 8\ p_{n-1}}\ , \qquad n \geq 1\ .$$

Seleccionando $p_0 = 1.5$ obtenemos los resultados del ejemplo 3 del capítulo XVI en los cuales $p_3 = 1.36523001$ es correcto en ocho decimales.

3. EL ALGORITMO DE LA SECANTE MODIFICADO

El Teorema IX.3 dice que, bajo condiciones razonables, el método de Newton convergerá siempre y cuando se escoja una aproximación inicial lo suficientemente exacta. También implica que la constante k que acota la derivada de g decrece conforme el procedimiento va avanzando y, consecuentemente, indica la rapidez de convergencia del método.

El método de Newton es una técnica extremadamente poderosa, pero tiene una dificultad grande: la necesidad de saber el valor de la derivada de f en cada aproximación. Frecuentemente ocurre que $f'(x)$ es mucho más complicada y necesita más operaciones aritméticas para su cálculo que $f(x)$.

Para evitar el problema de la evaluación de la derivada en el método de Newton, podemos derivar una pequeña variación de éste, relacionada con el método de la secante que hemos visto en el capítulo anterior.

Por definición

$$f'(p_{n-1}) = \lim_{x \to p_{n-1}} \frac{f(x) - f(p_{n-1})}{x - p_{n-1}}\ .$$

Tomando $x = p_{n-2}$

$$f'(p_{n-1}) \approx \frac{f(p_{n-2}) - f(p_{n-1})}{p_{n-2} - p_{n-1}} = \frac{f(p_{n-1}) - f(p_{n-2})}{p_{n-1} - p_{n-2}}\ .$$

Usando esta aproximación para $f'(p_{n-1})$ en la fórmula de Newton da

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f(p_{n-1}) - f(p_{n-2})}\ (p_{n-1} - p_{n-2})\ . \qquad (IX.9)$$

Algoritmo de la secante modificado.

==========

Para encontrar una solución de $f(x) = 0$, dadas las aproximaciones iniciales p_0 y p_1;

Entrada: aproximaciones iniciales p_0 y p_1; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada p ó mensaje de fracaso.

Paso 1: tomar $i = 2$, y definir:
$q_0 = f(p_0)$
$q_1 = f(p_1)$;

Paso 2: mientras que $i \leq N_0$ seguir pasos 3–6;

Paso 3: tomar *(calcular p_i)*:
$p = p_1 - \frac{q_1}{q_1 - q_0}\ (p_1 - p_0)$;

Paso 4: si $|p - p_1| < TOL$ entonces SALIDA (p);
(procedimiento completado satisfactoriamente) PARAR;

Paso 5: tomar $i = i + 1$

Paso 6: tomar *(redefinir p_0, p_1, q_0, q_1)*;
$p_0 = p_1$;
$q_0 = q_1$;
$p_1 = p$;
$q_1 = f(p)$;

Paso 7: SALIDA ($'$*El método fracasó después de N_0 iteraciones, $N_0 =$ $'$*, N_0);
(procedimiento completado sin éxito); PARAR.

==========

Ejemplo 3.

Consideremos el mismo caso del ejemplo 1, donde se usó el método de Newton con $p_0 = \pi/4$ para encontrar un cero de $f(x) = \cos x - x$. Pero ahora queremos usar el método de la secante modificado. Aquí , necesitamos dos aproximaciones iniciales. En la tabla 2 están los cálculos con $p_0 = 0.5$ y $p_1 = \pi/4$, habiendo usado la fórmula

$$p_n = p_{n-1} - \frac{(p_{n-1} - p_{n-2})\ (\cos p_{n-1} - p_{n-1})}{(\cos p_{n-1} - p_{n-1}) - (\cos p_{n-2} - p_{n-2})} \qquad \text{para} \quad n \geq 2\ .$$

Tabla 2

n	p_n
0	0.5
1	0.785398163
2	0.736384139
3	0.739058139
4	0.739085149
5	0.739085133

Comparando estos resultados con los del ejemplo 1, podemos ver que p_5 es exacto hasta la décima cifra decimal. Es interesante hacer notar que la convergencia del método de

la secante es un poco más lenta en este ejemplo que la del método de Newton, el cual obtuvo este grado de precisión con p_3. Este resultado es generalmente cierto.

El método de la secante o el método de Newton se usan frecuentemente para refinar las respuestas obtenidas con otras técnicas, como el método de bisección. Como estos métodos requieren una buena primera aproximación, pero generalmente dan una convergencia rápida, cumplen muy bien con su propósito.

4. EL METODO DE NEWTON MODIFICADO

Si la derivada $f'(x)$ varía, aunque ligeramente, en el intervalo $[a, b]$, en tal caso en la fórmula $(IX.5)$ podemos poner

$$f'(p_n) \approx f'(p_0) \ .$$

De aquí, para la raíz p de la ecuación $f(x) = 0$ tendremos las aproximaciones sucesivas

$$p_{n+1} = p_n - \frac{f(p_n)}{f'(p_0)} \ , \qquad n \geq 0 \ . \tag{IX.10}$$

La fórmula de iteración $(IX.10)$ es conocida también como la **fórmula de Von Mises**.

Geométricamente, este método significa que sustituímos las tangentes en los puntos $B_n[p_n, f(p_n)]$ por líneas rectas paralelas a la tangente a la curva $y = f(x)$ en el punto $B_0[p_0, f(p_0)]$.

La fórmula de Von Mises nos evita la necesidad de calcular los valores de la derivada $f'(p_n)$ cada vez; por lo tanto esta fórmula es muy útil si $f'(p_n)$ es complicada.

Puede demostrarse que supuesta la constancia de los signos de las derivadas $f'(x)$ y $f''(x)$ las aproximaciones sucesivas $(IX.10)$ presentan un proceso convergente.

5. EL METODO DE COMBINACION

Supongamos $f(a) \cdot f(b) < 0$ y que $f'(x)$ y $f''(x)$ conservan los signos en el intervalo $[a, b]$. Combinando el método de la secante modificado y el de Newton, obtenemos un método en el que en cada una de sus etapas encontraremos aproximaciones menores (demasiado pequeñas) y mayores (demasiado grandes) a la raíz exacta p de la ecuación $f(x) = 0$. Este método es también conocido con el nombre de **método de Dandelin**.

Una de sus consecuencias es que los dígitos comunes a p_n y $\overline{p}_n$ deben pertenecer definitivamente a la raíz exacta p. Existen cuatro casos teóricamente posibles:

$$(1) \qquad f'(x) > 0 \quad f''(x) > 0 \ ;$$

$$(2) \qquad f'(x) > 0 \quad f''(x) < 0 \ ;$$

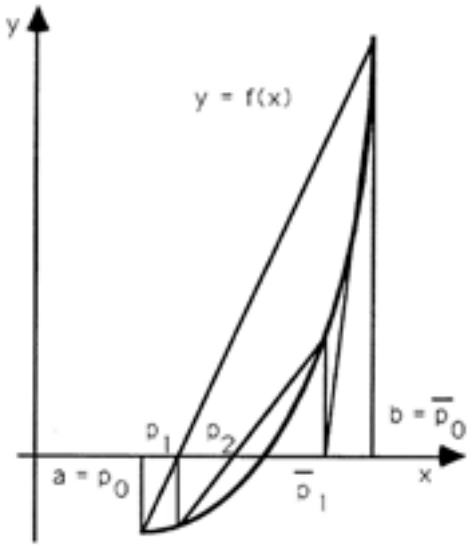
$$(3) \qquad f'(x) < 0 \quad f''(x) > 0 \ ;$$

$$(4) \qquad f'(x) < 0 \quad f''(x) < 0 \ .$$

Limitaremos nuestro análisis al primer caso (ver figura 2). Los casos restantes se estudian de forma análoga y el carácter de los cálculos se comprende fácilmente en base

a las figuras. Conviene tener en cuenta que estos casos pueden reducirse al primero si sustituímos la ecuación $f(x) = 0$ por las ecuaciones equivalentes $-f(x) = 0$ ó $\pm f(-z) = 0$, donde $z = -x$.

Figura 2



De este modo, supongamos $f'(x) > 0$ y $f''(x) > 0$ para $a \leq x \leq b$. Hagamos

$$p_0 = a \qquad \text{y} \qquad \overline{p}_0 = b \ ,$$

y

$$p_{n+1} = p_n - \frac{f(p_n)}{f(\overline{p}_n) - f(p_n)} \ (\overline{p}_n - p_n) \ , \tag{IX.11}$$

$$\overline{p}_{n+1} = \overline{p}_n - \frac{f(\overline{p}_n)}{f'(\overline{p}_n)} \qquad (n = 0, 1, 2, ...) \ . \tag{IX.12}$$

Es decir, en cada paso se aplica el método de la secante para un nuevo intervalo $[p_n, \overline{p}_n]$.

Por lo demostrado anteriormente se deduce que

$$p_n < p < \overline{p}_n$$

y

$$0 < p - p_n < \overline{p}_n - p_n \ .$$

Si el error absoluto permisible en una raíz aproximada p_n se ha especificado de antemano y es igual a ε, el proceso de aproximación termina tan pronto veamos que $\overline{p}_n - p_n < \varepsilon$. Al final del proceso, lo mejor es tomar como valor de la raíz p la media aritmética de los últimos valores obtenidos:

$$\overline{p} = \frac{1}{2} \ (p_n + \overline{p}_n) \ .$$

Ejemplo 4.

Calcúlese con exactitud 0.0005 la única raíz positiva de la ecuación

$$f(x) = x^5 - x - 0.2 = 0 \ .$$

Como $f(1) = -0.2 < 0$ y $f(1.1) = 0.31051 > 0$, la raíz está en el intervalo $(1, 1.1)$. Tenemos

$$f'(x) = 5\ x^4 - 1 \qquad \text{y} \qquad f''(x) = 20\ x^3 \ .$$

En el intervalo elegido, $f'(x) > 0$, $f''(x) > 0$, lo cual quiere decir que se han conservado los signos de las derivadas.

Apliquemos el método de combinación suponiendo que $p_0 = 1$ y $\overline{p}_0 = 1.1$. Ya que

$$f(p_0) = f(1) = -0.2 \ , \qquad f(\overline{p}_0) = f(1.1) = 0.3105 \ , \qquad f'(\overline{p}_0) = f'(1.1) = 6.3205$$

las fórmulas $(IX.11)$ y $(IX.12)$ se convierten en

$$p_1 = 1 + \frac{0.2 \cdot 0.1}{0.5105} \approx 1.03918 \qquad \text{y} \qquad \overline{p}_1 = 1.1 - \frac{0.3105}{6.3205} \approx 1.05087 \ ,$$

con $f(p_1) = -0.0273160$ y $f(\overline{p}_1) = 0.0307078$. Como $\overline{p}_1 - p_1 = 0.01169$, la exactitud no es suficiente. Halleremos el siguiente par de aproximaciones, con $f'(\overline{p}_1) \approx 5.09770$:

$$p_2 = 1.03919 + \frac{0.027316 \cdot 0.01169}{0.0580238} \approx 1.04468$$

y

$$\overline{p}_2 = 1.05087 - \frac{0.0307078}{5.0977} \approx 1.04485 \ ,$$

con $f(p_2) = -0.0000404924$ y $f(\overline{p}_2) = 0.000437805$. En este caso, $\overline{p}_2 - p_2 = 0.00017$, lo cual indica que se ha conseguido el grado de exactitud deseado. Podemos poner

$$\overline{p} = \frac{1}{2}\ (1.04468 + 1.04485) = 1.044765 \approx 1.045$$

con error absoluto menor de $0.0002 < 0.0005$. En efecto:

$$|p - \overline{p}| = |p - \frac{1}{2}(p_2 + \overline{p}_2)| = |p - p_2 + \frac{1}{2}(p_2 - \overline{p}_2)| \le |p - p_2| + \frac{1}{2}|p_2 - \overline{p}_2| \le \frac{1}{2}|p_2 - \overline{p}_2| + \frac{|f(p_2)|}{m_1} \ ,$$

o también

$$|p - \overline{p}| = |p - \overline{p}_2 + \frac{1}{2}(\overline{p}_2 - p_2)| \le \frac{1}{2}|\overline{p}_2 - p_2| + \frac{|f(\overline{p}_2)|}{m_1} \ .$$

Dado que $m_1 = \min\limits_{x \in [1,1.1]} |f'(x)| = 4$, obtenemos que $|p - \overline{p}| \le \frac{1}{2}\ 0.00017 + \frac{|-0.000404924|}{4} = 0.000186231$ en el primer caso y $|p - \overline{p}| \le \frac{1}{2}\ 0.00017 + \frac{|0.000437805|}{4} = 0.000194451$, en el segundo.

Ejemplo 5.

Hallar las primeras dos iteraciones de la raíz negativa de la ecuación

$$f(x) = x^4 + 0.1\ x^3 - 0.56\ x^2 + 0.1\ x - 1.56 = 0\ .$$

Como $f(-1.5) = 1.755 > 0$ y $f(-1.0) = -1.32 < 0$, la raíz está en el intervalo $(-1.5, -1.0)$. Tenemos

$$f'(x) = 4\ x^3 + 0.3\ x^2 - 1.12\ x + 0.1 < 0 \quad \text{en} \quad [-1.5, -1.0]\ ,$$

y

$$f''(x) = 12\ x^2 + 0.6\ x - 1.12 > 0 \quad \text{en} \quad [-1.5, -1.0]\ .$$

En el intervalo elegido no se han conservado los signos de las derivadas. Entonces, tenemos que aplicar el método de combinación de la siguiente manera:

$$p_{n+1} = \overline{p}_n - \frac{f(\overline{p}_n)}{f(p_n) - f(\overline{p}_n)}\ (p_n - \overline{p}_n)\ ,$$

$$\overline{p}_{n+1} = \overline{p}_n - \frac{f(\overline{p}_n)}{f'(\overline{p}_n)} \qquad (n = 0, 1, 2, ...)\ ,$$

y podemos escoger

$$\overline{p}_0 = a = -1.5 \qquad \text{y} \qquad p_0 = b = -1.0\ .$$

Ahora

$$\begin{aligned}
p_0 &= -1.0\ , \qquad \overline{p}_0 = -1.5\ , \\
f(p_0) &= 1.32\ , \qquad f(\overline{p}_0) = 1.755\ , \qquad f'(\overline{p}_0) = -11.045\ , \\
p_1 &= -1.5 + \frac{1.755}{1.32 + 1.755}\ (-1.0 + 1.5) \approx -1.214634146\ , \\
f(p_1) &= -0.510234362\ , \\
\overline{p}_1 &= -1.5 + \frac{1.755}{11.045} \approx -1.341104572\ , \\
f(\overline{p}_1) &= 0.292312590\ , \qquad f'(\overline{p}_1) = -7.506630499\ , \\
p_2 &= -1.341104572 + \frac{0.292312590}{7.506630499 + 0.292312590}\ (-1.214634146 + 1.341104572) \\
&\approx -1.295040105\ , \\
\overline{p}_2 &= -1.341104572 + \frac{0.292312590}{7.506630499} \approx -1.302163986\ , \\
f(p_2) &= -0.033129838\ , \qquad f(\overline{p}_2) = 0.014595893\ .
\end{aligned}$$

Entonces, podemos poner

$$\overline{p} = \frac{1}{2}\ (p_2 + \overline{p}_2) = -1.298602046$$

con error absoluto $|p - \overline{p}| = |1.3 - 1.298602046| = 0.00139795 < 0.0002$.

EJERCICIOS.

1. Aproxime con 10^{-4} de precisión las raíces de las siguientes ecuaciones en los intervalos dados usando el método de Newton-Raphson y el método de la secante modificado:

(a) $x^3 - 2\,x^2 - 5 = 0$ $[1, 4]$ (b) $x - \cos x = 0$ $[0, \pi/2]$

(c) $x^3 + 3\,x^2 - 1 = 0$ $[-4, 0]$ (d) $x - 0.8 - 0.2\sin x = 0$ $[0, \pi/2]$

2. Encuentre una raíz aproximada de $x^3 - x - 1 = 0$ en $[1, 2]$ con precisión de 10^{-5} por el método de Newton y el método de la secante modificado.

3. Usar el método de Newton y el método de la secante modificado para aproximar las soluciones de las ecuaciones siguientes con precisión de 10^{-5}:

(a) $x = \frac{2 - e^x + x^2}{3}$ (b) $3\,x^2 - e^x = 0$

(c) $e^x + 2^{-x} + 2\cos x - 6 = 0$ (d) $x^2 + 10\cos x = 0$

4. Resolver $4\cos x = e^x$ con exactidud de 10^{-4}, usando el método de Newton con $p_0 = 1$ y el método de la secante modificado con $p_0 = \pi/4$ y $p_1 = \pi/2$.

5. Calcular una aproximación de $\sqrt{3}$ exacta en 10^{-4}, usando el método de Newton-Raphson con $p_0 = 2$ y el método de la secante modificado.

6. Use el método de Newton para aproximar, con una exactitud de 10^{-4}, el valor de x que produce el punto en la gráfica de $y = x^2$ más cercano a $(1, 0)$. [Minimizar $[d(x)]^2$, donde $d(x)$ representa la distancia de (x, x^2) a $(1, 0)$].

7. Use el método de Newton-Raphson para aproximar, con una exactitud de 10^{-4}, el valor de x que produce el punto en la gráfica de $y = 1/x$ más cercano a $(1, 0)$.

8. Usar el método de Newton-Raphson para resolver la ecuación

$$f(x) = \big(\sin x - \frac{x}{2}\big)^2 \qquad \text{con} \quad p_0 = \frac{\pi}{2}\ .$$

Iterar hasta que se obtenga una precisión de 10^{-5} para la raíz aproximada. Resuélvase también con $p_0 = 5\pi$ y $p_0 = 10\pi$.

9. La función $f(x) = (4\,x - 7)\cdot(x - 2)$ tiene un cero en $p = 1.75$. Usar el método de Newton con las siguientes aproximaciones iniciales:

(a) $p_0 = 1.625$ (b) $p_0 = 1.875$ (c) $p_0 = 1.5$

(d) $p_0 = 1.95$ (e) $p_0 = 3$ (f) $p_0 = 7$

Explicar los resultados gráficamente.

10. La función descrita por $f(x) = \ln(x^2 + 1) - e^{0.4x}\cos \pi x$ tiene un número infinito de ceros.

(a) Usar el método de Newton para determinar, dentro de 10^{-6}, los únicos ceros negativos;

(b) Usar el método de Newton para determinar, dentro de 10^{-6}, los cuatro ceros positivos más pequeños;

(c) Determinar una aproximación inicial razonable (usar una gráfica aproximada de f) para determinar el enésimo cero positivo más pequeño de f.

CAPITULO X. ANALISIS DE ERROR Y TECNICAS DE ACELERACION

1. ANALISIS DE LOS ERRORES PARA METODOS ITERATIVOS

Este capítulo se dedica a investigar el orden de convergencia de los esquemas de iteración funcional y, con la idea de obtener una convergencia rápida, reutilizar el método de Newton. Consideraremos también maneras de acelerar la convergencia del método de Newton en circunstancias especiales. Pero, antes, tenemos que definir un procedimiento para medir la rapidez de la convergencia.

Definición

Supongamos que $\{p_n\}_{n=0}^{\infty}$ es una sucesión que converge a p y que $e_n = p_n - p$ para cada $n \geq 0$. Si existen dos números positivos λ y α tales que

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^{\alpha}} = \lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^{\alpha}} = \lambda \ ,$$

entonces se dice que $\{p_n\}_{n=0}^{\infty}$ **converge a** p **de orden** α, con una constante de error asintótico λ.

A una técnica iterativa para resolver un problema de la forma $x = g(x)$ se le denomina de orden α si, siempre que el método produce convergencia para una sucesión $\{p_n\}_{n=0}^{\infty}$ donde $p_n = g(p_{n-1})$ para $n \geq 1$, la sucesión converge a la solución de orden α.

En general, una sucesión con un orden de convergencia grande convergerá más rápidamente que una sucesión con un orden más bajo. La constante asintótica afectará la rapidez de convergencia, pero no es tan importante como el orden. Se dará atención especial a dos casos:

i) si $\alpha = 1$, entonces el método se denomina **lineal**;

ii) si $\alpha = 2$, entonces el método se denomina **cuadrático**.

Ejemplo 1.

Supongamos que queremos encontrar una solución aproximada de $g(x) = x$, usando el esquema de iteración de punto fijo $p_n = g(p_{n-1})$ para toda $n \geq 1$. Supongamos también que g manda el intervalo $[a, b]$ a sí mismo y que existe un número positivo k tal que $|g'(x)| \leq k < 1$ para todo $x \in [a, b]$. El Teorema VII.2 implica que g tiene un punto fijo único $p \in [a, b]$ y que si $p_0 \in [a, b]$ entonces la sucesión de punto fijo $\{p_n\}_{n=0}^{\infty}$ converge a p. Se mostrará que la convergencia es lineal, siempre que $g'(p) \neq 0$. Si n es cualquier entero positivo, entonces

$$e_{n+1} = p_{n+1} - p = g(p_n) - g(p) = g'(\xi_n)\ (p_n - p) = g'(\xi_n)\ e_n \ ,$$

donde ξ_n está entre p_n y p. Como $\{p_n\}_{n=0}^{\infty}$ converge a p, $\{\xi_n\}_{n=0}^{\infty}$ también converge a p. Suponiendo que g' es continua en $[a, b]$, tenemos que

$$\lim_{n\to\infty} g'(\xi_n) = g'(p) \ ,$$

y por lo tanto,

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n} = \lim_{n\to\infty} g'(\xi_n) = g'(p) \ , \qquad \text{y} \qquad \lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|} = |g'(p)| \ .$$

Por lo tanto, la iteración del punto fijo exhibe convergencia lineal si $g'(p) \neq 0$. La convergencia de orden mayor puede ocurrir sólo cuando $g'(p) = 0$.

Ejemplo 2.

Supongamos que tenemos dos esquemas iterativos convergentes descritos por

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|} = 0.75 \ , \qquad \text{un método lineal,}$$

y

$$\lim_{n\to\infty} \frac{|\tilde{e}_{n+1}|}{|\tilde{e}_n|^2} = 0.75 \ , \qquad \text{un método cuadrático.}$$

Supongamos también, por simplicidad, que

$$\frac{|e_{n+1}|}{|e_n|} \approx 0.75 \qquad \text{y} \qquad \frac{|\tilde{e}_{n+1}|}{|\tilde{e}_n|^2} \approx 0.75 \ .$$

Para el esquema de convergencia lineal, esto significa que

$$|e_n| \approx 0.75 \ |e_{n-1}| \approx (0.75)^2 \ |e_{n-2}| \approx \ldots \approx (0.75)^n \ |e_0| \ ,$$

mientras que el procedimiento convergente cuadráticamente tiene

$$\begin{aligned} |\tilde{e}_n| &\approx 0.75 \ |\tilde{e}_{n-1}|^2 \approx (0.75) \ [(0.75) \ |\tilde{e}_{n-2}|^2]^2 = (0.75)^3 \ |\tilde{e}_{n-2}|^4 \\ &\approx (0.75)^3 \ [(0.75) \ |\tilde{e}_{n-3}|^2]^4 = (0.75)^7 \ |\tilde{e}_{n-3}|^8 \approx \ldots \approx (0.75)^{2^n-1} \ |\tilde{e}_0|^{2^n} \ . \end{aligned}$$

Para comparar la rapidez de convergencia supondremos que $|e_0| = |\tilde{e}_0| = 0.5$ y usaremos las estimaciones para determinar el valor mínimo de n necesario para obtener un error que no exceda de 10^{-8}. Para el método lineal, esto implica que n debe ser tal que

$$|e_n| = (0.75)^n \ |e_0| \leq 10^{-8} \ ,$$

esto es

$$n \geq \frac{\log_{10} 2 - 8}{\log_{10} 0.75} \approx 62 \ .$$

Para el método de convergencia cuadrática

$$|\tilde{e}_n| = (0.75)^{2^n-1} \ |\tilde{e}_0|^{2^n} = (0.75)^{-1} \ (0.375)^{2^n} \leq 10^{-8} \ ,$$

implica que

$$2^n \log_{10} 0.375 \leq \log_{10} 0.75 - 8 \ ,$$

y por lo tanto,

$$2^n \geq \frac{\log_{10} 0.75 - 8}{\log_{10} 0.375} \approx 19.1 \ ,$$

así que

$$n \geq 5 \ .$$

En estas circunstancias, el método convergente cuadráticamente, requiriendo sólo 5 iteraciones es muy superior al lineal requiriendo 62.

2. TECNICAS DE ACELERACION Y FORMULA DE NEWTON GENERALIZADA

Vamos ahora a determinar y caracterizar esquemas de iteración funcional cuadrática.

Teorema X.1

Sea p una solución de $x = g(x)$. Supongamos que $g'(p) = 0$ y g'' es continua en un intervalo abierto que contiene a p. Entonces existe un $\delta > 0$ tal que, para $p_0 \in [p-\delta, p+\delta]$, la sucesión definida por $p_n = g(p_{n-1})$, para toda $n \geq 1$, es convergente cuadráticamente.

Demostración: escogeremos $\delta > 0$ tal que en el intervalo $[p-\delta, p+\delta]$, $|g'(x)| \leq k < 1$ y g'' sea continua. Como $|g'(x)| \leq k < 1$ se sigue que los términos de la sucesión $\{p_n\}_{n=0}^{\infty}$ están contenidos en $[p-\delta, p+\delta]$. Desarrollando $g(x)$ en un polinomio de Taylor lineal para $x \in [p-\delta, p+\delta]$ resulta

$$g(x) = g(p) + g'(p)(x-p) + \frac{g''(\xi)}{2} \ (x-p)^2 \ ,$$

donde ξ está entre x y p. Usando las hipótesis $g(p) = p$ y $g'(p) = 0$, tenemos que:

$$g(x) = p + \frac{g''(\xi)}{2} \ (x-p)^2 \ .$$

En particular, cuando $x = p_n$ para algún n,

$$p_{n+1} = g(p_n) = p + \frac{g''(\xi_n)}{2} \ (p_n - p)^2$$

con ξ_n entre p_n y p. Por lo tanto,

$$p_{n+1} - p = e_{n+1} = \frac{g''(\xi_n)}{2} \ e_n^2 \ .$$

Como $|g'(x)| \leq k < 1$ en $[p-\delta, p+\delta]$, y g manda $[p-\delta, p+\delta]$ a sí mismo, del Teorema VII.2 tenemos que $\{p_n\}_{n=0}^{\infty}$ converge a p. Como ξ_n está entre p y p_n para cada n, $\{\xi_n\}_{n=0}^{\infty}$ converge también a p, y

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^2} = \frac{|g''(p)|}{2} \ .$$

Esto implica que la sucesión $\{p_n\}_{n=0}^{\infty}$ converge cuadráticamente. **c.q.d.**

Para usar el Teorema X.1 para resolver una ecuación de la forma $f(x) = 0$, supongamos que la ecuación $f(x) = 0$ tiene una solución p tal que $f'(p) \neq 0$. Consideremos el esquema de punto fijo

$$p_n = g(p_{n-1}) \ , \qquad n \geq 1 \ ,$$

con g de la forma

$$g(x) = x - \phi(x)\ f(x)\ ,$$

donde ϕ es una función arbitraria que se escogerá más adelante.

Si $\phi(x)$ está acotada, entonces $g(p) = p$, y, para que el procedimiento iterativo derivado de g sea cuadráticamente convergente, es suficiente que $g'(p) = 0$. Pero

$$g'(x) = 1 - \phi'(x)\ f(x) - \phi(x)\ f'(x) \quad \text{y} \quad g'(p) = 1 - \phi(p)\ f'(p)\ .$$

Consecuentemente, $g'(p) = 0$ si y sólo si $\phi(p) = \dfrac{1}{f'(p)}$.

En particular, se obtendrá convergencia cuadrática para el esquema

$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}\ ,$$

el cual puede reconocerse como el método de Newton.

En la discusión anterior, se impuso la restricción de que $f'(p) \neq 0$, donde p es la solución de $f(x) = 0$. De la definición del método de Newton, es claro que pueden presentarse dificultades si $f'(p_n)$ tiende a cero simultáneamente con $f(p_n)$. En particular, este método y el método de la secante traerán generalmente problemas si $f'(p) = 0$ cuando $f(p) = 0$. Para examinar estas dificultades con más detalle haremos la siguiente definición.

Definición

Se dice que una solución p de $f(x) = 0$ es un **cero de multiplicidad** m de f si $f(x)$ puede escribirse como $f(x) = (x - p)^m\ q(x)$, para $x \neq p$, donde $\lim\limits_{x \to p} q(x) \neq 0$.

Esencialmente $q(x)$ representa la porción de $f(x)$ que no contribuye al cero de f. El siguiente resultado da una manera fácil de identificar a los ceros de las funciones que tienen multiplicidad uno. A estos ceros se les llama **simples**.

Teorema X.2

Una función $f \in \mathcal{C}^1[a, b]$ tiene un cero simple en p en (a, b) si y sólo si $f(p) = 0$, pero $f'(p) \neq 0$.

El resultado del Teorema X.2 implica que existe un intervalo alrededor de p en el cual el método de Newton converge cuadráticamente a p para cualquier aproximación inicial, siempre y cuando p sea una raíz simple. El ejemplo siguiente muestra que no necesariamente hay convergencia cuadrática si la raíz no es simple.

Ejemplo 3.

La función descrita por $f(x) = e^x - x - 1$ tiene una raíz de multiplicidad dos en $p = 0$.

Para demostrar esto, consideremos la función

$$q(x) = \frac{e^x - x - 1}{x^2}\ .$$

Usando la regla de l'Hôpital,

$$\lim_{x\to 0} \frac{e^x - x - 1}{x^2} = \lim_{x\to 0} \frac{e^x - 1}{2\ x} = \lim_{x\to 0} \frac{e^x}{2} = \frac{1}{2} \neq 0\ .$$

Por lo tanto,

$$f(x) = (x-0)^2 \frac{e^x - x - 1}{x^2}\ ,$$

donde $\lim_{x\to 0} \frac{e^x - x - 1}{x^2} \neq 0$, y la multiplicidad de la raíz en $p = 0$ es dos.

Los términos generados por el método de Newton aplicado a f con $p_0 = 1$ se muestran en la tabla siguiente. Está claro que la sucesión no converge cuadráticamente a cero.

Tabla 1

n	p_n	n	p_n
0	1.0	9	2.7750×10^{-3}
1	0.58198	10	1.3881×10^{-3}
2	0.31906	11	6.9424×10^{-4}
3	0.16800	12	3.4716×10^{-4}
4	0.08635	13	1.7358×10^{-4}
5	0.04380	14	8.6773×10^{-5}
6	0.02206	15	4.3329×10^{-5}
7	0.01107	16	2.1635×10^{-5}
8	0.005545		

Una manera de *atacar* el problema de raíces múltiples consiste en definir una función $\mu(x)$ por

$$\mu(x) = \frac{f(x)}{f'(x)}\ .$$

Si p es una raíz de multiplicidad $m \geq 1$, y $f(x) = (x-p)^m\ q(x)$, entonces

$$\mu(x) = \frac{(x-p)^m\ q(x)}{m\ (x-p)^{m-1}\ q(x) + (x-p)^m\ q'(x)} = \frac{(x-p)\ q(x)}{m\ q(x) + (x-p)\ q'(x)}$$

tendrá también una raíz en p, pero de multiplicidad uno. El método de Newton puede entonces aplicarse a la función μ para dar

$$g(x) = x - \frac{\mu(x)}{\mu'(x)} = x - \frac{f(x)/f'(x)}{\{[f'(x)]^2 - f(x)\ f''(x)\}/[f'(x)]^2}$$

o

$$g(x) = x - \frac{f(x)\ f'(x)}{[f'(x)]^2 - f(x)\ f''(x)}\ . \qquad (X.1)$$

La fórmula $(X.1)$ se conoce como **fórmula de Newton generalizada** para raíces multiples.

Si g cumple con las condiciones de continuidad requeridas, la iteración funcional aplicada a g tendrá convergencia cuadrática independientemente de la multiplicidad de la raíz. Teóricamente, las únicas desventajas de este método son los cálculos adicionales de

$f''(x)$ y el hecho de que el procedimiento es más laborioso para calcular las iteraciones. En la práctica, sin embargo, la presencia de una raíz múltiple puede causar serios problemas de redondeo.

Ejemplo 4.

En el ejemplo 3 del capítulo XVI resolvimos $f(x) = x^3 + 4\ x^2 - 10 = 0$ para la raíz $p = 1.365230013$. Para comparar la convergencia del método de Newton y el método de Newton generalizado, ecuación $(X.1)$, para una raíz de multiplicidad uno, sea (i):

$$p_n = p_{n-1} - \frac{p_{n-1}^3 + 4\ p_{n-1}^2 - 10}{3\ p_{n-1}^2 + 8\ p_{n-1}}\ , \qquad \text{del método de Newton}$$

y, de la ecuación $(X.1)$, (ii):

$$p_n = p_{n-1} - \frac{(p_{n-1}^3 + 4\ p_{n-1}^2 - 10)\ (3\ p_{n-1}^2 + 8\ p_{n-1})}{(3\ p_{n-1}^2 + 8\ p_{n-1})^2 - (p_{n-1}^3 + 4\ p_{n-1}^2 - 10)\ (6\ p_{n-1} + 8)}\ .$$

Con $p_0 = 1.5$, las primeras iteraciones para (i) y (ii) son las siguientes.

Tabla 2

	(i)	(ii)
p_1	1.373333333	1.356898976
p_2	1.365262015	1.365195849
p_3	1.365230014	1.365230013
p_4	1.365230013	1.365230013

Ejemplo 5.

Para ilustrar la situación que se presenta en una raíz múltiple, consideremos la ecuación $f(x) = x^4 - 4\ x^2 + 4 = 0$, que tiene una raíz de multiplicidad dos en $x = \sqrt{2} = 1.414213562$. Usar el método de Newton y la versión modificada $(X.1)$ produce, después de algunas simplificaciones, las sucesiones con términos (i):

$$p_n = p_{n-1} - \frac{p_{n-1}^2 - 2}{4\ p_{n-1}}\ , \qquad \text{del método de Newton}$$

y, (ii):

$$p_n = p_{n-1} - \frac{(p_{n-1}^2 - 2)\ p_{n-1}}{(p_{n-1}^2 + 2)}\ , \qquad \text{de la ecuación (X.1).}$$

Con $p_0 = 1.5$, las tres primeras iteraciones para (i) y (ii) nos dan lo siguiente:

Tabla 3

	(i)	(ii)
p_1	1.458333333	1.411764706
p_2	1.436607143	1.414211438
p_3	1.425497619	1.414213562

La solución real correcta en 10^{-9} es la que aparece para p_3 en (ii). Para obtener esta precisión con el método normal de Newton-Raphson se requerirían 20 iteraciones.

Ejemplo 6.

Consideremos la ecuación $f(x) = x^3 - 5\ x^2 + 3\ x + 9 = 0$, que tiene una raíz de multiplicidad dos en $x = 3.0$, dado que $f(x) = x^3 - 5\ x^2 + 3\ x + 9 = (x-3)^2(x+1)$. Para comparar la convergencia del método de Newton y el método de Newton generalizado, ecuación $(X.1)$, después de algunas simplificaciones, obtenemos las sucesiones con términos (i):

$$p_{n+1} = p_n - \frac{(p_n - 3)\ (p_n + 1)}{(3\ p_n - 1)}\ ,$$

para el método de Newton-Raphson, y, (ii):

$$p_{n+1} = p_n - \frac{(p_n - 3)\ (p_n + 1)\ (3\ p_n - 1)}{(3\ p_n^2 + 2\ p_n + 11)}\ ,$$

para el método de Newton generalizado, ecuación $(X.1)$.

Con $p_0 = 1.5$, las iteraciones para (i) y (ii) nos dan los resultados siguientes:

Tabla 4

	(i)	(ii)
p_0	2.5	2.5
p_1	2.769230769	2.959595960
p_2	2.888259109	2.999791764
p_3	2.944944061	2.999999995
p_4	2.972665472	3.000000000
p_5	2.986379918	
p_{10}	2.999575778	
p_{15}	2.999986744	
p_{20}	2.999999586	
p_{25}	2.999999987	
p_{29}	2.999999999	
p_{30}	3.000000000	

3. CONVERGENCIA ACELERADA Y EL ALGORITMO Δ^2 DE AITKEN

En este apartado consideraremos una técnica, llamada **método Δ^2 de Aitken**, que se usa para acelerar la convergencia de cualquier sucesión que converja linealmente, independentemente de su origen.

Supongamos que $\{p_n\}_{n=0}^{\infty}$ es una sucesión linealmente convergente con límite p; o sea que, para $e_n = p_n - p$,

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|} = \lambda \quad \text{y} \quad 0 < \lambda < 1\ .$$

Para investigar la construcción de una sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ que converja más rápidamente a p, supongamos que n es lo suficientemente grande para que el cociente pueda usarse para aproximar el límite. Si suponemos también que todas las e_n tienen el mismo signo, entonces

$$e_{n+1} \approx \lambda\ e_n \quad \text{y} \quad e_{n+2} \approx \lambda\ e_{n+1}\ .$$

Así que

$$p_{n+2} = e_{n+2} + p \approx \lambda\ e_{n+1} + p$$

ó

$$p_{n+2} \approx \lambda\ (p_{n+1} - p) + p \qquad \text{o sea} \qquad \lambda = \frac{p_{n+2} - p}{p_{n+1} - p}\ . \qquad (X.2a, b)\ .$$

Reemplazando $(n+1)$ por n en las ecuaciones $(X.2)$ da

$$p_{n+1} \approx \lambda\ (p_n - p) + p \qquad \text{o sea} \qquad \lambda = \frac{p_{n+1} - p}{p_n - p}\ ; \qquad (X.3a, b)$$

y resolviendo las ecuaciones $(X.2)$ y $(X.3)$ para p mientras se elimina λ nos lleva a que

$$\begin{aligned} p &\approx \frac{p_{n+2}\ p_n - p_{n+1}^2}{p_{n+2} - 2\ p_{n+1} + p_n} \approx \\ &\approx \frac{p_n^2 + p_n\ p_{n+2} + 2\ p_n\ p_{n+1} - 2\ p_n\ p_{n+1} - p_n^2 - p_{n+1}^2}{p_{n+2} - 2\ p_{n+1} + p_n} \approx \\ &\approx \frac{(p_n^2 + p_n\ p_{n+2} - 2\ p_n\ p_{n+1}) - (p_n^2 - 2\ p_n\ p_{n+1} + p_{n+1}^2)}{p_{n+2} - 2\ p_{n+1} + p_n} \approx \\ &\approx p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2\ p_{n+1} + p_n}\ . \end{aligned}$$

El método Δ^2 de Aitken está basado en la suposición de que la sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ definida por

$$\hat{p}_n = p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2\ p_{n+1} + p_n}\ , \qquad (X.4)$$

converge más rápidamente a p que la sucesión original $\{p_n\}_{n=0}^{\infty}$.

Ejemplo 7.

La sucesión $\{p_n\}_{n=1}^{\infty}$, donde $p_n = \cos(\frac{1}{n})$, converge linealmente a $p = 1$. Los primeros términos de la sucesión $\{p_n\}_{n=1}^{\infty}$ y $\{\hat{p}_n\}_{n=1}^{\infty}$ están dados en la siguiente tabla.

Tabla 5

n	p_n	$\hat{p}_n$
1	0.54030	0.96178
2	0.87758	0.98213
3	0.94496	0.98979
4	0.96891	0.99342
5	0.98007	0.99541
6	0.98614	
7	0.98981	

Es evidente que $\{\hat{p}_n\}_{n=0}^{\infty}$ converge más rápidamente a p que $\{p_n\}_{n=0}^{\infty}$.

La notación Δ asociada con esta técnica tiene su origen en la siguiente definición.

Definición

Dada la sucesión $\{p_n\}_{n=0}^{\infty}$, se define la **diferencia progresiva** Δp_n mediante

$$\Delta p_n = p_{n+1} - p_n \qquad \text{para}\ \ n \geq 0\ .$$

Las potencias mayores $\Delta^k p_n$ se definen recursivamente mediante

$$\Delta^k p_n = \Delta^{k-1}(\Delta p_n) \qquad \text{para} \quad k \geq 2 \ .$$

Debido a la definición,

$$\begin{aligned} \Delta^2 p_n &= \Delta(p_{n+1} - p_n) = \\ &= \Delta p_{n+1} - \Delta p_n = \\ &= (p_{n+2} - p_{n+1}) - (p_{n+1} - p_n) = \\ &= p_{n+2} - 2\ p_{n+1} + p_n \ . \end{aligned}$$

Por lo tanto, la fórmula para $\hat{p}_n$ dada en $(X.4)$ se puede escribir como

$$\hat{p}_n = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n} \qquad \text{para toda} \quad n \geq 0 \ . \tag{X.5}$$

Para ilustrar el método Δ^2 de Aitken de otra manera, supongamos que la sucesión $\{p_n\}_{n=0}^{\infty}$ converge al límite p como una sucesión geométrica decreciente con factor k:

$$p_{n+1} - p = k(p_n - p), \qquad |k| < 1 \qquad n = 0, 1, 2, \ldots \ .$$

Entonces, k y p pueden ser obtenidos a partir de p_n, p_{n+1} y p_{n+2} usando las ecuaciones

$$\begin{aligned} p_{n+1} - p &= k(p_n - p) \ , \\ p_{n+2} - p &= k(p_{n+1} - p) \ . \end{aligned}$$

Haciendo la resta de estas ecuaciones:

$$k = \frac{p_{n+2} - p_{n+1}}{p_{n+1} - p_n} \ ,$$

y sustituyendo en la primera ecuación, dado que $k \neq 1$:

$$p = \frac{k\ p_n - p_{n+1}}{k - 1} = \frac{p_n\ p_{n+2} - p_{n+1}^2}{p_{n+2} - 2\ p_{n+1} + p_n} \ ,$$

que es la misma ecuación $(X.4)$.

Hasta ahora, en nuestra discusión del método Δ^2 de Aitken, hemos dicho que la sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ converge más rápidamente a p que la sucesión original $\{p_n\}_{n=0}^{\infty}$, pero no hemos dicho qué se entiende por *convergencia más rápida*. Los siguientes Teoremas explican esta terminología.

Teorema X.3

Sea $\{p_n\}_{n=0}^{\infty}$ cualquier sucesión que converja linealmente a un límite p con $e_n = p_n - p \neq 0$ para toda $n \geq 0$. Entonces la sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ converge a p más rápidamente que $\{p_n\}_{n=0}^{\infty}$ en el sentido de que

$$\lim_{n \to \infty} \frac{\hat{p}_n - p}{p_n - p} = 0 \ .$$

Demostración: si la sucesión converge linealmente a p con $e_n = p_n - p \neq 0, \quad \forall n \geq 0$, entonces $\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|} = \lambda$. Supongamos que n es lo suficientemente grande para que el cociente pueda usarse para aproximar el límite y que toda las e_n tienen el mismo signo, entonces $e_{n+1} \approx \lambda\ e_n$. Ahora calculamos el cociente:

$$\frac{\hat{p}_n - p}{p_n - p} = \frac{p_n - \frac{(p_{n+1}-p_n)^2}{p_{n+2}-2\ p_{n+1}+p_n} - p}{p_n - p} = \frac{e_n - \frac{(p_{n+1}-p_n)^2}{p_{n+2}-2\ p_{n+1}+p_n}}{e_n} =$$
$$= 1 - \frac{1}{e_n}\ \frac{(p_{n+1} - p_n + p - p)^2}{p_{n+2} - 2\ p_{n+1} + p_n + 2\ p - 2\ p} = 1 - \frac{(e_{n+1}/e_n)^2 - 2\ e_{n+1}/e_n + 1}{e_{n+2}/e_n - 2\ e_{n+1}/e_n + 1}\ .$$

Pasando al límite para $n \to \infty$ obtenemos:

$$\lim_{n\to\infty} \frac{\hat{p}_n - p}{p_n - p} = \lim_{n\to\infty} \left[1 - \frac{\lambda^2 - 2\ \lambda + 1}{\lambda^2 - 2\ \lambda + 1}\right] = 1 - 1 = 0\ .$$

c.q.d.

Teorema X.4

Sea $\{p_n\}_{n=0}^{\infty}$ cualquier sucesión que se comporte asintóticamente como una sucesión geométrica, es decir existe k, $|k| < 1$, tal que

$$p_{n+1} - p = (k + \delta_n)\ (p_n - p), \qquad \lim_{n\to\infty} \delta_n = 0\ .$$

Entonces la sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ converge a p más rápidamente que $\{p_n\}_{n=0}^{\infty}$ en el sentido de que

$$\lim_{n\to\infty} \frac{\hat{p}_n - p}{p_n - p} = 0\ .$$

Demostración: por hipótesis el error $e_n = p_n - p$ satisface $e_{n+1} = (k+\delta_n)\ e_n$. Entonces:

$$\begin{aligned} p_{n+2} - 2\ p_{n+1} + p_n &= e_{n+2} - 2\ e_{n+1} + e_n = \\ &= e_n\ [(k + \delta_{n+1})\ (k + \delta_n) - 2\ (k + \delta_n) + 1] = \\ &= e_n\ ((k-1)^2 + \mu_n) \qquad \text{con}\ \ \mu_n \to 0\ , \end{aligned}$$

y

$$p_{n+1} - p_n = e_{n+1} - e_n = e_n\ [(k-1) + \delta_n]\ .$$

Desde luego, $p_{n+2} - 2\ p_{n+1} + p_n \neq 0$ para grandes valores de n, dado que $e_n \neq 0$, $k \neq 1$ y $\mu_i \to 0$. Entonces la sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ definida en $(X.4)$ está bien definida. Además,

$$\hat{p}_n - p = p_n - p - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2\ p_{n+1} + p_n} = e_n - e_n \frac{[(k-1) + \delta_n]^2}{(k-1)^2 + \mu_n}$$

para valores grandes de n, y entonces (dado que $\delta_n \to 0$ y $\mu_n \to 0$ para $n \to \infty$):

$$\lim_{n\to\infty} \frac{\hat{p}_n - p}{p_n - p} = \lim_{n\to\infty} \left\{1 - \frac{[(k-1) + \delta_n]^2}{(k-1)^2 + \mu_n}\right\} = 0\ .$$

c.q.d.

Algoritmo Δ^2 de Aitken.

==========

Para encontrar una solución de $p = g(p)$, dada una aproximación inicial p_0:

Entrada: aproximación inicial p_0; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada p o mensaje de fracaso.

Paso 1: tomar $i = 1$, y calcular $p_1 = g(p_0)$;

Paso 2: mientras que $i \leq N_0$ seguir pasos 3–6;

Paso 3: tomar:
$p_2 = g(p_1)$; (*calcular* p_{1+i});
$p = p_0 - \frac{(p_1 - p_0)^2}{(p_2 - 2\ p_1 + p_0)}$; (*calcular* $\hat{p}_{i-1}$);

Paso 4: si $|p - p_2| < TOL$ entonces SALIDA (p); *(procedimiento completado satisfactoriamente)* PARAR;

Paso 5: tomar $i = i + 1$

Paso 6: tomar
$p_0 = p_1$; *(redefinir* p_0*)*;
$p_1 = p_2$; *(redefinir* p_1*)*;

Paso 7: SALIDA ('*El método fracasó después de* N_0 *iteraciones,* $N_0 =$ $', N_0$); *(procedimiento completado sin éxito)*; PARAR.

==========

4. CONVERGENCIA ACELERADA Y EL ALGORITMO DE STEFFERSEN

Aplicando el método Δ^2 de Aitken a una sucesión que converge linealmente obtenida de la iteración de punto fijo, podemos acelerar la convergencia a cuadrática. Este procedimiento es conocido como el **método de Steffersen** y difiere un poco de aplicar el método Δ^2 de Aitken directamente a una sucesión de iteración de punto fijo que sea linealmente convergente. El procedimiento directo construiría en orden

$$p_0\ , p_1 = g(p_0)\ , p_2 = g(p_1)\ , \quad \rightarrow \quad \hat{p}_0 = \{\Delta^2\} p_0\ ,$$

$$p_3 = g(p_2)\ , \quad \rightarrow \quad \hat{p}_1 = \{\Delta^2\} p_1\ , \ldots\ ,$$

donde $\{\Delta^2\}$ se usa para indicar que se emplea la técnica Δ^2 de Aitken.

El método de Steffersen construye los mismos primeros cuatro términos p_0, p_1, p_2, $\hat{p}_0$; sin embargo, en el siguiente paso, supone que $\hat{p}_0$ es una mejor aproximación a p que p_2 y aplica iteración de punto fijo a $\hat{p}_0$ en lugar de a p_2. Cada tercer término es generado usando la técnica Δ^2 de Aitken; para los otros, se usa la iteración de punto fijo en el térmmino anterior. La sucesión generada es entonces:

$$p_0\ , p_1 = g(p_0)\ , p_2 = g(p_1)\ , \quad \rightarrow \quad \hat{p}_0 = \{\Delta^2\} p_0\ ,$$

$$p_3 = \hat{p}_0\ , p_4 = g(p_3)\ , p_5 = g(p_4)\ , \quad \rightarrow \quad \hat{p}_1 = \{\Delta^2\} p_3\ , \ldots\ .$$

Es decir, usando una nueva notación útil para el algoritmo empezando con la aproximación inicial $p_0 \equiv p_0^{(0)}$ tendremos

$$p_0^{(0)}\ , p_1^{(0)} = g(p_0^{(0)})\ , p_2^{(0)} = g(p_1^{(0)})\ , \quad \rightarrow \quad p_0^{(1)} = \{\Delta^2\} p_0^{(0)}\ ,$$

$$p_1^{(1)} = g(p_0^{(1)}) \ , p_2^{(1)} = g(p_1^{(1)}) \ , \quad \rightarrow \quad p_0^{(2)} = \{\Delta^2\} p_0^{(1)} \ ,$$
$$p_1^{(2)} = g(p_0^{(2)}) \ , p_2^{(2)} = g(p_1^{(2)}) \ , \quad \rightarrow \quad p_0^{(3)} = \{\Delta^2\} p_0^{(2)} \ ,$$
$$\vdots$$

Algoritmo de Steffersen.

==========

Para encontrar una solución de $p = g(p)$, dada una aproximación inicial p_0:

Entrada: aproximación inicial p_0; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada p o mensaje de fracaso.

Paso 1: tomar $i = 1$;

Paso 2: mientras que $i \leq N_0$ seguir pasos 3–6;

Paso 3: tomar:
$p_1 = g(p_0)$; (*calcular* $p_1^{(i-1)}$);
$p_2 = g(p_1)$; (*calcular* $p_2^{(i-1)}$);
$p = p_0 - \frac{(p_1 - p_0)^2}{(p_2 - 2\ p_1 + p_0)}$; (*calcular* $p_0^{(i)}$);

Paso 4: si $|p - p_2| < TOL$ entonces SALIDA (p); *(procedimiento completado satisfactoriamente)* PARAR;

Paso 5: tomar $i = i + 1$

Paso 6: tomar $p_0 = p$; *(redefinir* p_0*)*;

Paso 7: SALIDA ($'$*El método fracasó después de* N_0 *iteraciones,* $N_0 =$ $'$, N_0); *(procedimiento completado sin éxito)*; PARAR.

==========

Nótese que $\Delta^2 p_n$ puede ser cero. Si esto sucediera, terminaríamos la sucesión y seleccionaríamos $p_2^{(n-1)}$ como la respuesta aproximada, ya que de otra manera esto introduciría un cero en el denominador de la siguiente iteración.

Ejemplo 8.

Para resolver $x^3 + 4\,x^2 - 10 = 0$ usando el método de Steffersen, escribimos $x^3 + 4\,x^2 = 10$ y resolvemos x dividiendo entre $x + 4$. Entonces,

$$g(x) = \sqrt{\frac{10}{x+4}} \ .$$

Usando $p_0 = 1.5$ el procedimiento de Steffersen da los valores de la siguiente tabla.

Tabla 6

k	$p_0^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$
0	1.5	1.348399725	1.367376372
1	1.365265224	1.365225534	1.365230583
2	1.365230013		

El valor de $p_0^{(2)} = 1.365230013$ es exacto en nueve cifras decimales. En este ejemplo, el método de Steffersen da aproximadamente la misma rapidez de convergencia que el método de Newton.

Del ejemplo 8, parece que el método de Steffersen da convergencia cuadrática sin evaluar derivadas. En realidad éste es el caso.

Teorema X.5

Supongamos que $x = g(x)$ tiene la solución p con $g'(p) \neq 1$. Si existe $\delta > 0$ tal que $g \in \mathcal{C}^3[p-\delta, p+\delta]$, entonces el método de Steffersen da convergencia cuadrática para cualquier $p_0 \in [p-\delta, p+\delta]$.

La debilidad del método de Steffersen reside en la necesidad de que $g'(p) \neq 1$, condición que es equivalente a requerir que la multiplicidad del cero p sea uno, para el problema correspondiente de búsqueda de la raíz de $f(x) = 0$. Como consecuencia de esto, no se puede esperar que el método de Steffersen acelere a cuadrática la convergencia lineal que resulta generalmente cuando el método de Newton se usa para aproximar un cero de multiplicidad mayor que uno.

Ejemplo 9.

Queremos resolver $f(x) = x^2 - \cos x = 0$ usando el método de Steffersen, y comparar con el método Δ^2 de Aitken y con el de Newton-Raphson. Entonces, escribimos

$$x = g(x) = \sqrt{\cos x} \ ,$$

para los métodos de Aitken y de Steffersen, y

$$p_{n+1} = p_n - \frac{p_n^2 - \cos p_n}{2\ p_n + \sin p_n} \ ,$$

para las iteraciones de Newton-Raphson. Usando $p_0 = 1.0$, la iteración funcional, el método Δ^2 de Aitken, el algoritmo de Steffersen y el de Newton-Raphson dan los resultados de la tabla siguiente:

Tabla 7

k	Punto fijo	Aitken	Steffersen	Newton
0	1.0	0.820545868	0.820545868	1.0
1	0.735052587	0.823387630	0.824131023	0.838218410
2	0.861275501	0.823989495	0.824132312	0.824241868
3	0.807137107	0.824103654	0.824132312	0.824132319
4	0.831606374	0.824126663		0.824132312
5	0.820785901	0.824131189		0.824132312
6	0.825618791	0.824132090		
7	0.823469674	0.824132268		
8	0.824427236	0.824132304		
9	0.824000957	0.824132311		
10	0.824190798	0.824132312		
11	0.824106268	0.824132312		
15	0.824131288			
20	0.824132330			
25	0.824132312			

EJERCICIOS.

1. Usar el método de Newton para resolver la ecuación

$$0 = \left(\text{sen}(x) - \frac{x}{2}\right)^2$$

con $p_0 = \frac{\pi}{2}$, hasta que se obtenga una precisión de 10^{-5} para la raíz aproximada. Luego usar la ecuación $(X.1)$ e iteración funcional para encontrar una raíz, exacta en 10^{-5} de la misma función $f(x)$, empezando con $p_0 = \frac{\pi}{2}$.

2. Usar el método generalizado de Newton-Raphson descrito en la ecuación $(X.1)$ para encontrar una aproximación a la raíz de

$$f(x) = x^2 + 2\ x\ e^x + e^{2x} = 0$$

empezando con $p_0 = 0$ y efectuando 10 iteraciones.

3. Usar el método generalizado de Newton-Raphson descrito en la ecuación $(X.1)$ para encontrar una aproximación a la raíz de

$$f(x) = e^{6x} + 3\ (\ln 2)^2\ e^{2x} - (\ln 8)\ e^{4x} - (\ln 2)^3 = 0$$

empezando con $p_0 = 0$ y hasta que $|p_{n+1} - p_n| < 0.0002$. Repetir los calculos con las constantes en $f(x)$ reemplazadas por sus aproximaciones a cuatro dígitos $(f(x) = e^{6x} + 1.441\ e^{2x} - 2.079\ e^{4x} - 0.3330 = 0)$.

4. Demostrar que las siguientes sucesiones $\{p_n\}_{n=0}^{\infty}$ convergen linealmente a $p = 0$. Encontrar cuántos términos deben generarse antes de que $|p_n - p| \leq 0.05$.

$$a) \quad p_n = \frac{1}{n}, \quad n \geq 1, \qquad\qquad b) \quad p_n = \frac{1}{n^2}, \quad n \geq 1\ .$$

5. Resolver $x^3 - x - 1 = 0$ para la raíz en $[1, 2]$ con una exactitud de 10^{-4} usando el método de Steffersen.

6. Resolver $x - 2^{-x} = 0$ para la raíz en $[0, 1]$ con una exactitud de 10^{-4} usando el método de Steffersen.

7. Usar el método de Steffersen con $p_0 = 0$ para calcular una aproximación de $\sqrt{3}$ exacta en cuatro dígitos significativos.

8. Aproximar las soluciones de las siguientes ecuaciones con una precisión de 10^{-5}, usando el método de Steffersen.

$$a) \quad x = \frac{2 - e^x + x^2}{3}, \qquad b) \quad 3\ x^2 - e^x = 0, \qquad c) \quad x - \cos x = 0\ .$$

9. Para las siguientes sucesiones $\{p_n\}_{n=0}^{\infty}$, linealmente convergentes, usar el método Δ^2 de Aitken para generar una sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ hasta que $|\hat{p}_n - p| \leq 5 * 10^{-2}$.

$$a) \quad p_n = \frac{1}{n}, \quad n \geq 1, \qquad\qquad b) \quad p_n = \frac{1}{n^2}, \quad n \geq 1\ .$$

CAPITULO XI. METODOS DE INTERPOLACION

1. EL METODO DE INTERPOLACION DE LA POSICION FALSA

Los métodos de interpolación que vamos a discutir en el resto de este capítulo son muy útiles para determinar los ceros de una función real $f(x)$ cualquiera. A diferencia del método de Newton, en los métodos de interpolación que veremos no se necesita calcular la derivada de f, y además convergen más rapidamente.

El método de interpolación más simple es el conocido como **regula falsi** o método de **la falsa posición**. Es muy parecido al método de bisección en el cual dos números p_n y a_n se obtienen en cada paso de forma que $f(p_n) \cdot f(a_n) < 0$. El intervalo $[p_n, a_n]$ contiene entonces al menos un cero de $f(x)$, y los valores p_n vienen determinados en manera que converjan hacia uno de estos ceros.

En el método de interpolación de la posición falsa, para definir los valores p_{n+1}, a_{n+1}, se considera μ_n el cero de la función interpolante lineal:

$$P(x) = f(p_n) + (x - p_n)\ \frac{f(p_n) - f(a_n)}{p_n - a_n}$$

donde $P(p_n) = f(p_n)$ y $P(a_n) = f(a_n)$, es decir

$$\mu_n = p_n - f(p_n)\ \frac{p_n - a_n}{f(p_n) - f(a_n)} = \frac{a_n\ f(p_n) - p_n\ f(a_n)}{f(p_n) - f(a_n)}\ . \qquad (XI.1a)$$

Por el hecho de que $f(p_n) \cdot f(a_n) < 0$, tenemos que $f(p_n) - f(a_n) \neq 0$; entonces μ_n está siempre bien definido y satisface ó $p_n < \mu_n < a_n$ ó $a_n < \mu_n < p_n$. A menos que $f(\mu_n) = 0$, definimos:

$$\begin{cases} p_{n+1} = \mu_n \quad \text{y} \quad a_{n+1} = a_n, \quad \text{si} \quad f(\mu_n) \cdot f(p_n) > 0, \\ p_{n+1} = \mu_n \quad \text{y} \quad a_{n+1} = p_n, \quad \text{si} \quad f(\mu_n) \cdot f(p_n) < 0. \end{cases} \qquad (XI.1b, c)$$

El algoritmo termina si $f(\mu_n) = 0$, es decir si μ_n es el cero.

Para discutir la convergencia del método de la falsa posición, asumiremos por simplicidad que f'' existe y que para algun valor i:

$$p_i < a_i, \qquad (XI.2a)$$

$$f(p_i) < 0, \qquad f(a_i) > 0, \qquad (XI.2b)$$

$$f''(x) \geq 0, \qquad \forall\ x \in [p_i, a_i], \qquad (XI.2c)$$

Con estas hipótesis ó $f(\mu_i) = 0$ ó

$$f(\mu_i) \cdot f(p_i) > 0$$

y entonces $p_i < p_{i+1} = \mu_i < a_{i+1} = a_i$ (ver figuras 1 y 2).

Es ahora fácil ver que las fórmulas $(XI.2)$ son válidas para todo $i \geq i_0$ si son válidas para un i_0. Entonces, $a_i = a$ para $i \geq i_0$, y las p_i forman una secuencia monótona acotada creciente, y el límite $\lim_{i\to\infty} p_i = \overline{p}$ existe. Por el hecho de que f es continua, y por $(XI.2)$ se sigue que

$$f(a) > 0, \qquad f(\overline{p}) \leq 0 \ .$$

Además, pasando al límite en $(XI.1)$

$$\overline{p} = \frac{a\ f(\overline{p}) - \overline{p}\ f(a)}{f(\overline{p}) - f(a)}$$

que implica

$$(\overline{p} - a)\ f(\overline{p}) = 0 \ .$$

Pero $\overline{p} \neq a$, y entonces $f(\overline{p}) = 0$.

Figura 1

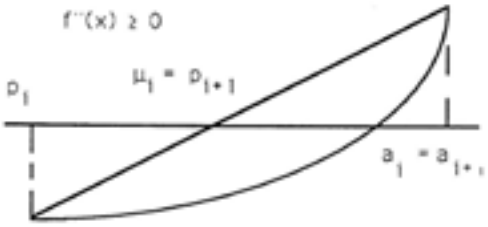
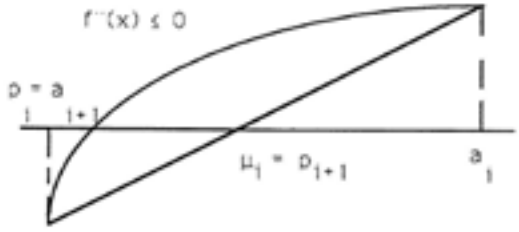


Figura 2



Está claro que bajo las hipótesis $(XI.2)$, la regula falsi, usará sólo las primeras dos fórmulas de recursión $(XI.1b)$. Nótese que este caso se reduce al método de la secante presentado en el Capítulo XVII, en el cual el extremo fijo es aquél para el cual el signo de la función $f(x)$ coincide con el signo de su segunda derivada $f''(x)$.

La variación del método de la falsa posición, basada exclusivamente en la segundas dos fórmulas de recursión $(XI.1c)$

$$p_{n+1} = \frac{p_{n-1}\ f(p_n) - p_n\ f(p_{n-1})}{f(p_n) - f(p_{n-1})}$$

no es nada más que el método de la secante modificado que hemos encontrado en el Capítulo IX.

Ejemplo 1.

Consideramos el polinomio $P(x) = -x^3 + 6\ x^2 + 4\ x - 24$. En el intervalo $[0, 3]$, donde hay un cero, no se puede usar el método de la secante, pero si el de la secante modificada y el de la regula falsi. Los resultados están resumidos en la siguientes tablas, respectivamente.

$p_0 = 3.0$	$f(p_0) = 15.0$	$p_1 = 0.0$	$f(p_1) = -24.0$
i	p_i		$f(p_i)$
2	1.846153846		-2.457897135
3	2.056795132		0.90853891
4	1.99994694		-8.4896×10^{-4}
5	2.0000000011		1.76×10^{-7}
6	2.0		0.0

$p_1 = 0$	$f(p_1) = -24.0$	$a_1 = 3.0$	$f(a_1) = 15.0$
i	μ_i		$f(\mu_i)$
1	1.846153846		-2.457897135
2	2.008603833		0.137660691
3	1.999987967		-1.92528×10^{-4}
4	2.0		0.0

2. EL METODO DE INTERPOLACION DE MÜLLER

Estudiaremos ahora un método presentado por primera vez por D.E. Müller en 1956. Esta técnica puede ser usada en cualquier problema de búsqueda de raíces, pero es particularmente útil para aproximar raíces de polinomios.

El método de Müller es una generalización del método de la secante. El método de la secante modificado empieza con dos aproximaciones iniciales x_0 y x_1 y determina la siguiente aproximación x_2 como la intersección del eje x con la recta que pasa por $(x_0, f(x_0))$ y $(x_1, f(x_1))$. El método de Müller usa tres aproximaciones iniciales x_0, x_1 y x_2 y determina la siguiente aproximación x_3 considerando la intersección del eje x con la parábola que pasa por $(x_0, f(x_0))$, $(x_1, f(x_1))$ y $(x_2, f(x_2))$.
La derivación del procedimiento de Müller comienza considerando el polinomio cuadrático

$$P(x) = a\ (x - x_2)^2 + b\ (x - x_2) + c$$

que pasa por $(x_0, f(x_0))$, $(x_1, f(x_1))$ y $(x_2, f(x_2))$. Las constantes a, b y c pueden determinarse de las condiciones

$$\begin{aligned} f(x_0) = P(x_0) &= a\ (x_0 - x_2)^2 + b\ (x_0 - x_2) + c\ , \\ f(x_1) = P(x_1) &= a\ (x_1 - x_2)^2 + b\ (x_1 - x_2) + c\ , \\ f(x_2) = P(x_2) &= c\ ; \end{aligned}$$

las cuales nos dan

$$\begin{aligned} c &= f(x_2)\ , \\ b &= \frac{(x_0 - x_2)^2\ [f(x_1) - f(x_2)] - (x_1 - x_2)^2\ [f(x_0) - f(x_2)]}{(x_0 - x_2)\ (x_1 - x_2)\ (x_0 - x_1)}\ , \\ a &= \frac{(x_1 - x_2)\ [f(x_0) - f(x_2)] - (x_0 - x_2)\ [f(x_1) - f(x_2)]}{(x_0 - x_2)\ (x_1 - x_2)\ (x_0 - x_1)}\ . \end{aligned} \qquad (XI.3)$$

Para determinar x_3, la raíz de P, aplicamos la fórmula cuadrática a P. Debido a problemas del error de redondeo causados por la subtracción de números casi iguales, se aplica la fórmula

$$x_3 - x_2 = \frac{-2\ c}{b \pm \sqrt{b^2 - 4\ a\ c}}\ .$$

Esto da dos posibilidades para x_3 dependiendo del signo que precede al término bajo radical. En el método de Müller, el signo se elije para que coincida con el de b. Escogido de esta manera, el denominador será el más grande en magnitud y resultará en seleccionar a x_3 como la raíz de P más cercana a x_2. Así,

$$x_3 = x_2 - \frac{2\ c}{b + signo(b)\sqrt{b^2 - 4\ a\ c}}$$

donde a, b y c están dadas en $(XI.3)$. Una vez que se determina x_3, el procedimiento se reinicializa usando x_1, x_2 y x_3 en lugar de x_0, x_1 y x_2 para determinar la siguiente aproximación x_4. El método continúa hasta que se obtiene una conclusión satisfactoria. Ya que el método involucra en cada paso el radical $\sqrt{b^2 - 4\ a\ c}$, el método aproximará raíces complejas cuando sea apropiado.

Algoritmo de Müller.

==========

Para encontrar una solución a $f(x) = 0$ dadas tres aproximaciones x_0, x_1 y x_2:

Entrada: aproximaciones iniciales x_0, x_1 y x_2; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada de p ó mensaje de fracaso.

Paso 1: tomar
$h_1 = x_1 - x_0$; $h_2 = x_2 - x_1$;
$\delta_1 = [f(x_1) - f(x_0)]/h_1$; $\delta_2 = [f(x_2) - f(x_1)]/h_2$;
$a = (\delta_2 - \delta_1)/(h_2 + h_1)$;
$i = 2$;

Paso 2: mientras que $i \leq N_0$ seguir pasos 3–7;

Paso 3: tomar:
$b = \delta_2 + h_2\ a$;
$D = \sqrt{b^2 - 4\ f(x_2)a}$;

Paso 4: si $|b-D| < |b+D|$ entonces tomar $E = b+D$, si no tomar $E = b-D$;

Paso 5: tomar:
$h = -2\ f(x_2)/E$;
$p = x_2 + h$;

Paso 6: si $|h| < TOL$ entonces SALIDA (p); *(procedimiento completado satisactoriamente)* PARAR;

Paso 7: tomar (*preparar para la siguiente iteración*):
$x_0 = x_1$; $x_1 = x_2$; $x_2 = p$;
$h_1 = x_1 - x_0$; $h_2 = x_2 - x_1$;
$\delta_1 = [f(x_1) - f(x_0)]/h_1$; $\delta_2 = [f(x_2) - f(x_1)]/h_2$;
$a = (\delta_2 - \delta_1)/(h_2 + h_1)$;
$i = i + 1$;

Paso 8: SALIDA ($'$*El método fracasó después de* N_0 *iteraciones,* $N_0 =$ $', N_0$); *(procedimiento completado sin éxito)*; PARAR.

==========

Ejemplo 2.

Consideramos el polinomio $P(x) = 16\ x^4 - 40\ x^3 + 5\ x^2 + 20\ x + 6$. Usando el algoritmo de Müller con $TOL = 10^{-5}$ y diferentes valores de x_0, x_1 y x_2, tenemos los resultados que se muestran el la tabla siguente. Los valores reales de las raíces de la ecuación son 1.241677445, 1.970446079 y $-0.356062 \pm 0.162758\ i$, lo que demuestra que las aproximaciones del método de Müller son excelentes.

$x_0 = 0.5$	$x_1 = -0.5$	$x_2 = 0.0$	
i	x_i		$f(x_i)$
3	$-0.555556 + 0.598352\ i$		$-29.4007 - 3.89872\ i$
4	$-0.435450 + 0.102101\ i$		$1.33223 - 1.19309\ i$
5	$-0.390631 + 0.141852\ i$		$0.375057 - 0.670164\ i$
6	$-0.357699 + 0.169926\ i$		$-0.146746 - 0.00744629\ i$
7	$-0.356051 + 0.162856\ i$		$-0.183868 \times 10^{-2} + 0.539780 \times 10^{-3}\ i$
8	$-0.356062 + 0.162758\ i$		$0.286102 \times 10^{-2} + 0.953674 \times 10^{-6}\ i$

$x_0 = 0.5$	$x_1 = 1.0$	$x_2 = 1.5$	
i	x_i		$f(x_i)$
3	1.287855		-1.376275
4	1.237459		1.269422×10^{-1}
5	1.241604		2.194520×10^{-3}
6	1.241677		1.321123×10^{-6}
7	1.241677		1.321123×10^{-6}

$x_0 = 2.5$	$x_1 = 2.0$	$x_2 = 2.25$	
i	x_i		$f(x_i)$
3	1.960592		-6.113129×10^{-1}
4	1.970564		7.456961×10^{-3}
5	1.970447		3.133506×10^{-5}
6	1.970447		2.720395×10^{-6}

El ejemplo 2 ilustra que el método de Müller puede aproximar las raíces del polinomio con una variedad de valores iniciales. De hecho, la importancia del método de Müller reside en que esta técnica generalmente convergerá a la raíz del polinomio para cualquier elección de las aproximaciones iniciales. Se pueden construir problemas en los que no habrá convergencia para ciertas aproximaciones iniciales. Por ejemplo, si x_i, x_{i+1} y x_{i+2} para alguna i tienen la propiedad de que $f(x_i) = f(x_{i+1}) = f(x_{i+2})$, la ecuación cuadrática se reducirá a una función constante no cero y nunca cuzará al eje x; sin embargo, éste no es usualmente el caso.

El método de Müller no es tan eficiente como el método de Newton: su orden de convergencia cerca de una raíz es aproximadamente $\alpha = 1.84$ comparado con el cuadrático, $\alpha = 2$, del método de Newton, pero es mejor que el método de la secante, cuyo orden es aproximadamente $\alpha = 1.62$.

EJERCICIOS.

1. En 1225 Leonardo de Pisa estudió la ecuación

$$f(x) = x^3 + 2\,x^2 + 10\,x - 20 = 0$$

y encontró $x \approx 1.368808107$. Nadie sabe que método usó Leonardo. Usar el método de la regula falsi para reencontrar este resultado.

2. Encontrar aproximaciones exactas en 10^{-4} de todos los ceros reales de las siguientes funciones usando el método de regula falsi.

a) $f(x) = x - e^{-x}$,

b) $P(x) = x^3 - x^2 - x - 1$.

3. Encontrar aproximaciones exactas en 10^{-4} de todos los ceros reales de los siguientes polinomios usando el método de Müller.

a) $P(x) = x^3 - 2\,x^2 - 5$, *b)* $P(x) = x^3 + 3\,x^2 - 1$,

c) $P(x) = x^3 - x - 1$, *d)* $P(x) = x^4 + 2\,x^2 - x - 3$.

4. Encontrar aproximaciones exactas en 10^{-5} de todos los ceros reales de los siguientes polinomios usando el método de Müller.

a) $P(x) = x^4 + 5\,x^3 - 9\,x^2 - 85\,x - 136$,

b) $P(x) = x^4 - 2\,x^3 - 12\,x^2 + 16\,x - 40$,

c) $P(x) = x^4 + x^3 + 3\,x^2 + 2\,x + 2$,

d) $P(x) = x^5 + 11\,x^4 - 21\,x^3 - 10\,x^2 - 21\,x - 5$,

e) $P(x) = 16\,x^4 + 88\,x^3 + 159\,x^2 + 76\,x - 240$,

f) $P(x) = x^4 - 4\,x^2 - 3\,x + 5$,

g) $P(x) = x^4 - 2\,x^3 - 4\,x^2 + 4\,x + 4$,

h) $P(x) = x^3 - 7\,x^2 + 14\,x - 6$,

CAPITULO XII. CEROS DE POLINOMIOS

1. EL METODO DE HORNER

Una función de la forma

$$P(x) = a_0\ x^N + a_1\ x^{N-1} + \ldots + a_{N-1}\ x + a_N\ , \qquad (XII.1)$$

donde las a_i, llamadas los **coeficientes** de P, son constantes y $a_0 \neq 0$, se llama un **polinomio de grado** N. La función cero, $P(x) = 0$ para todos los valores de x, se considera un polinomio pero no se le asigna ningun grado.

Teorema XII.1 (Teorema Fundamental del Algebra)
Si P es un polinomio de grado $N \geq 1$, entonces $P(x) = 0$ tiene al menos una raíz (posiblemente compleja).

Corolario XII.2
Si $P(x) = a_0\ x^N + a_1\ x^{N-1} + \ldots + a_{N-1}\ x + a_N$ es un polinomio de grado $N \geq 1$, entonces existen constantes únicas $x_1,\ x_2,\ \ldots,\ x_k$, posiblemente complejas, y enteros positivos, $m_1,\ m_2,\ \ldots,\ m_k$ tales que $\sum_{i=1}^{k} m_i = N$ y

$$P(x) = a_0\ (x - x_1)^{m_1}\ (x - x_2)^{m_2}\ \ \ldots\ (x - x_k)^{m_k}\ . \qquad (XII.2)$$

El Corolario XII.2 afirma que los ceros de un polinomio son únicos y que si cada cero x_i es contado tantas veces como su multiplicidad m_i, entonces un polinomio de grado N tiene exactamente N ceros.

Corolario XII.3
Sean P y Q polinomios a lo sumo de grado N. Si $x_1,\ x_2,\ \ldots,\ x_k$, $k > N$, son números distintos con $P(x_i) = Q(x_i)$ para $i = 1, 2, \ldots, k$, entonces $P(x) = Q(x)$ para todo valor de x.

Para usar el procedimiento de Newton-Raphson en localizar aproximadamente los ceros de un polinomio P, es necesario evaluar a P y a su derivada en valores específicos. Como P y sus derivadas son polinomios, la eficiencia computacional requerirá que la evalución de estas funciones sea hecha de manera anidada. El método de Horner descrito en el siguiente Teorema incorpora esta técnica y como consecuencia requiere solamente de N multiplicaciones y N sumas para evaluar un polinomio de enésimo grado arbitrario.

Teorema XII.4 (Método de Horner)
Sea

$$P(x) = a_0\ x^N + a_1\ x^{N-1} + \ldots + a_{N-1}\ x + a_N\ .$$

Si

$$d_0 = a_0 \qquad \text{y} \qquad d_k = a_k + d_{k-1}\ x_0\ , \qquad (XII.3)$$

para $k = 1, 2, \ldots, N-1, N$, entonces

$$d_N = P(x_0)\ . \qquad (XII.4)$$

Además, si

$$Q(x) = d_0\ x^{N-1} + d_1\ x^{N-2} + \ldots + d_{N-2}\ x + d_{N-1}\ , \qquad (XII.5)$$

entonces

$$P(x) = (x - x_0)\ Q(x) + d_N\ . \qquad (XII.6)$$

Demostración: la primera parte de la demostración es obvia, debido a la definición de los coeficientes d_k (basta sólo escribir el polinomio en forma *annidada*).

Veamos ahora la segunda parte. Por la definición de $Q(x)$:

$$\begin{aligned}(x - x_0)\ Q(x) + d_N &= (x - x_0)\ (d_0\ x^{N-1} + d_1\ x^{N-2} + \ldots + d_{N-2}\ x + d_{N-1}) + d_N = \\ &= (d_0\ x^N + d_1\ x^{N-1} + \ldots + d_{N-2}\ x^2 + d_{N-1}\ x)\ + \\ &\quad - (d_0\ x_0\ x^{N-1} + d_1\ x_0\ x^{N-2} + \ldots + d_{N-2}\ x_0\ x + d_{N-1}\ x_0) + \\ &\quad + d_N = \\ &= d_0\ x^N + (d_1 - d_0\ x_0)\ x^{N-1} + \ldots + (d_{N-2} - d_{N-3}\ x_0)\ x^2\ + \\ &\quad (d_{N-1} - d_{N-2}\ x_0)\ x + (d_N - d_{N-1}\ x_0)\ .\end{aligned}$$

Ahora, por las hipótesis $d_0 = a_0$ y $d_k - d_{k-1}\ x_0 = a_k$, así que

$$(x - x_0)\ Q(x) + d_N = a_0\ x^N + a_1\ x^{N-1} + \ldots + a_{N-1}\ x + a_N = P(x) \quad \text{y} \quad d_N = P(x_0)\ .$$

c.q.d.

Ejemplo 1.

Evaluar $P(x) = 2\ x^4 - 3\ x^2 + 3\ x - 4$ en $x_0 = -2$ usando el método de Horner.

Usando el Teorema XII.4

$$\begin{aligned} d_0 &= 2, & d_1 &= 2(-2) + 0 = -4, \\ d_2 &= (-4)(-2) - 3 = 5, & d_3 &= 5(-2) + 3 = -7, \end{aligned}$$

y finalmente

$$P(-2) = d_4 = (-7)(-2) - 4 = 10\ .$$

Además, el Teorema XII.4 nos dice que

$$P(x) = (x + 2)(2\ x^3 - 4\ x^2 + 5\ x - 7) + 10\ .$$

Cuando en el método de Horner se hacen los cálculos a mano, se construye primero una tabla, que sugiere el nombre de **división sintética** con frecuencia aplicado a esta técnica. Para el problema del ejemplo anterior, la tabla aparecería como:

	Coef. de x^4	Coef. de x^3	Coef. de x^2	Coef. de x	Término constante
	$a_0 = 2$	$a_1 = 0$	$a_2 = -3$	$a_3 = 3$	$a_4 = -4$
$x_0 = -2$		$d_0x_0 = -4$	$d_1x_0 = 8$	$d_2x_0 = -10$	$d_3x_0 = 14$
	$d_0 = 2$	$d_1 = -4$	$d_2 = 5$	$d_3 = -7$	$d_4 = 10$

Una ventaja adicional al usar el procedimiento de Horner es que, como

$$P(x) = (x - x_0)\ Q(x) + d_N\ ,$$

donde

$$Q(x) = d_0\ x^{N-1} + d_1\ x^{N-2} + \ldots + d_{N-2}\ x + d_{N-1}\ ,$$

diferenciando con respecto a x da

$$P'(x) = Q(x) + (x - x_0)\ Q'(x)$$

y

$$P'(x_0) = Q(x_0)\ . \qquad (XII.7)$$

Así, cuando se use el método de Newton-Raphson para encontrar un cero aproximado de un polinomio P, ambos P y P' pueden ser evaluados de esta manera. El algoritmo siguiente calcula $P(x_0)$ y $P'(x_0)$ usando el método de Horner.

Algoritmo de Horner.

==========

Para evaluar el polinomio

$$P(x) = a_0\ x^N + a_1\ x^{N-1} + \ldots + a_{N-1}\ x + a_N\ ,$$

y su derivada en x_0:

Entrada: grado N; coeficientes a_0, a_1, ..., a_N; punto donde evaluar el polinomio x_0;

Salida: $y = P(x_0)$ y $z = P'(x_0)$.

Paso 1: tomar
$y = a_0$; (*calcular* d_0 *para* P);
$z = a_0$; (*calcular* $\tilde{d}_0$ *para* Q);

Paso 2: para $j = 1, 2, \ldots, N-1$ tomar
$y = x_0\ y + a_j$; (*calcular* d_j *para* P);
$z = x_0\ z + y$; (*calcular* $\tilde{d}_j$ *para* Q);

Paso 3: tomar:
$y = x_0\ y + a_N$; (*calcular* d_N *para* P);

Paso 4: SALIDA (y, z); PARAR.

==========

Un uso interesante del algoritmo de Horner es expresar el desarrollo de Taylor de un polinomio alrededor de cualquier punto. Sea el polinomio P dado por $(XII.1)$, y suponemos que buscamos los coeficientes c_k de la ecuación

$$\begin{aligned} P(x) &= a_0\ x^N + a_1\ x^{N-1} + \ldots + a_{N-1}\ x + a_N\ , \\ &= c_0\ (x - x_0)^N + c_1\ (x - x_0)^{N-1} + \ldots + c_{N-1}\ (x - x_0) + c_N\ . \end{aligned}$$

Es obvio por el teorema de Taylor de que $c_k = \dfrac{1}{(N-k)!}\ P^{(N-k)}(x_0)$, para $k = 0, 1, \ldots, N$, pero es nuestra intención buscar un algoritmo más eficiente. Claramente, $c_N = P(x_0)$, de

modo que este coeficiente se obtiene aplicando el algoritmo de Horner al polinomio P en el punto x_0. El algoritmo también genera el polinomio:

$$Q(x) = \frac{P(x) - P(x_0)}{x - x_0} = d_0\ x^{N-1} + d_1\ x^{N-2} + \ldots + d_{N-2}\ x + d_{N-1} = \\ = c_0\ (x - x_0)^{N-1} + c_1\ (x - x_0)^{N-2} + \ldots + c_{N-1}\ .$$

Esto demuestra que el segundo coeficiente, c_{N-1}, se puede obtener aplicando el algoritmo de Horner al polinomio Q con el punto x_0, ya que $d_{N-1} = c_{N-1} = Q(x_0)$. El proceso se repite hasta que se encuentren todos los coeficientes c_k.

Ejemplo 2.

Encontrar el desarrollo de Taylor alrededor del punto $x_0 = 3$, para el polinomio

$$P(x) = x^4 - 4\ x^3 + 7\ x^2 - 5\ x - 2\ .$$

El trabajo se puede arreglar de la manera siguiente:

	1	-4	7	-5	-2
$x_0 = 3$		3	-3	12	21
	1	-1	4	7	$19 = P(3) = c_4$
$x_0 = 3$		3	6	30	
	1	2	10	$37 = c_3$	
$x_0 = 3$		3	15		
	1	5	$25 = c_2$		
$x_0 = 3$		3			
	$1 = c_0$	$8 = c_1$			

Entonces, el cálculo muestra que

$$P(x) = (x - 3)^4 + 8\ (x - 3)^3 + 25\ (x - 3)^2 + 37\ (x - 3) + 19\ .$$

Ejemplo 3.

Encontrar una aproximación a uno de los ceros de

$$P(x) = 2\ x^4 - 3\ x^2 + 3\ x - 4\ .$$

Hacer los calculos con aritmética de cuatro dígitos significativos y usar el procedimiento de Newton-Raphson y división sintética para evaluar $P(x_n)$ y $P'(x_n)$ para cada iteración.

Usando $x_0 = -2$ como una aproximación inicial, obtenemos $P(-2)$ por:

	2	0	-3	3	-4
$x_0 = -2$		-4	8	-10	14
	2	-4	5	-7	$10 = P(-2)$

Usando el Teorema XII.4 y la ecuación $(XII.7)$, obtenemos

$$Q(x) = 2\,x^3 - 4\,x^2 + 5\,x - 7 \qquad \text{y} \qquad P'(-2) = Q(-2)\ ;$$

así, $P'(-2)$ se puede encontrar de una manera similar, evaluando $Q(-2)$:

	2	-4	5	-7
$x_0 = -2$		-4	16	-42
	2	-8	21	$-49 = Q(-2) = P'(-2)$

y

$$x_1 = x_0 - \frac{P(x_0)}{P'(x_0)} = -2 - \frac{10}{-49} \approx -1.796\ .$$

Repitiendo el procedimiento para encontrar x_2, tenemos que

	2	0	-3	3	-4
-1.796		-3.592	6.451	-6.198	5.744
	2	-3.592	3.451	-3.198	$1.744 = P(x_1)$
-1.796		-3.592	12.90	-29.36	
	2	-7.184	16.35	$-32.56 = Q(x_1) = P'(x_1)$	

Así $P(-1.796) = 1.744$, $P'(-1.796) = -32.56$, y

$$x_2 = x_1 - \frac{P(x_1)}{P'(x_1)} = -1.796 - \frac{1.744}{-32.56} \approx -1.742\ .$$

Un cero real con cinco dígitos decimales significativos es -1.74259.

Nótese que el polinomio denotado por Q depende de la aproximación usada y cambia de iteración a iteración.

Un problema al aplicar el método de Newton a polinomios es el concerniente a la posibilidad de que el polinomio tenga raíces complejas aún cuando todos los coeficientes sean números reales. Si la aproximación inicial al usar el método de Newton es un número real, todas las aproximaciones sucesivas serán también números reales. Una manera de superar esta dificultad es empezar con aproximaciones iniciales no reales y hacer todos los cálculos usando aritmética compleja. Un enfoque diferente se basa en el siguiente Teorema.

Teorema XII.5

Si $z = \beta + \gamma\ i$ es un cero complejo de multiplicidad m del polinomio P, entonces $\bar{z} = \beta - \gamma\ i$ es también un cero de multiplicidad m del polinomio P y $(x^2 - 2\,\beta\,x + \beta^2 + \gamma^2)^m$ es un factor de P.

Consideremos ahora el problema de evaluar un polinomio $P(x)$ en un valor complejo del argumento $x = \beta + i\ \gamma$, donde los coeficientes $a_k = b_k + i\ c_k$ son complejos. Poniendo $d_k = Q_k + i\ R_k$ obtenemos:

$$\begin{cases} Q_n = b_0\ , \quad R_0 = c_n \\ Q_k = Q_{k-1}\ \beta - R_{k-1}\ \gamma + b_k\ , \quad k = 1, 2, \ldots, N\ , \\ R_k = R_{k-1}\ \beta + Q_{k-1}\ \gamma + c_k\ , \quad k = 1, 2, \ldots, N\ , \end{cases}$$

Entonces, la división sintetica compleja funciona de la siguiente manera:

	Coef. de x^N	Coef. de x^{N-1}	... ...	Coef. de x	Término constante
	$b_0\ ,\ \ c_0$	$b_1\ ,\ \ c_1$	...	$b_{N-1}\ ,\ \ c_{N-1}$	$b_N\ ,\ \ c_N$
$\beta + i\ \gamma$		$Q_0\beta - R_0\gamma\ ,$ ... $Q_0\gamma + R_0\beta$ ...		$Q_{N-2}\beta - R_{N-2}\gamma\ ,$ $Q_{N-2}\gamma + R_{N-2}\beta$	$Q_{N-1}\beta - R_{N-1}\gamma\ ,$ $Q_{N-1}\gamma + R_{N-1}\beta$
	$Q_0\ ,\ \ R_0$	$Q_1\ ,\ \ R_1$	...	$Q_{N-1}\ ,\ \ R_{N-1}$	$Q_N\ ,\ \ R_N$

Ejemplo 4.

Evaluar $P(x) = (1 + i)\ x^3 + 2$ en $x_0 = (1 - i)$ usando el método de Horner y la división sintetica compleja.

La tabla de la división sintetica será la siguiente:

	$1\ ,\ \ 1$	$0\ ,\ \ 0$	$0\ ,\ \ 0$	$2\ ,\ \ 0$
$1 - i$		$2\ ,\ \ 0$	$2\ ,-2$	$0\ ,-4$
	$1\ ,\ \ 1$	$2\ ,\ \ 0$	$2\ ,-2$	$2\ ,-4$

Entonces, $P(1 - i) = 2 - 4\ i$, como se puede verificar desde $(1 + i)\ (1 - i)^3 + 2 = 2\ (1 - i)^2 + 2 = -4\ i + 2$. Las derivadas se pueden obtener de la misma manera.

Vamos a dar ahora otro ejemplo de la aplicación del método de Newton-Raphson.

Ejemplo 5.

Encontrar una aproximación a los ceros de

$$P(x) = x^3 - 2 = 0\ ,$$

usando el procedimiento de Newton-Raphson y división sintética para evaluar $P(x_n)$ y $P'(x_n)$ para cada iteración, con aritmética de cuatro dígitos.

Con el valor inicial $x_0 = 1$, obtenemos:

	1	0	0	-2
$x_0 = 1$		1	1	1
	1	1	1	$-1 = P(1)$
$x_0 = 1$		1	2	
	1	2	$3 = P'(1)$	

Entonces,

$$x_1 = x_0 - \frac{P(x_0)}{P'(x_0)} = 1 - \frac{-1}{3} \approx 1.333 \ .$$

Repitiendo el procedimiento para encontrar x_2, tenemos que

	1	0	0	-2
$x_1 = 1.333$		1.333	1.777	2.369
	1	1.333	1.777	$0.369 = P(1.333)$
$x_1 = 1.333$		1.333	3.553	
	1	2.666	$5.330 = P'(1.333)$	

Así $P(1.333) = 0.369$, $P'(1.333) = 5.330$, y

$$x_2 = x_1 - \frac{P(x_1)}{P'(x_1)} = 1.333 - \frac{0.369}{5.330} \approx 1.264 \ .$$

Después de dos iteraciones hemos obtenido un valor aproximado de 1.264, que no está mal comparado con el valor verdadero $p \approx 1.260$ ($p^3 = 2$). Evidentemente el proceso es convergente. Sin embargo, no hay ninguna posibilidad de convergencia a una de las dos raíces complejas $-0.630 \pm 1.091\ i$ si no usamos un valor inicial complejo. Así que ahora repetimos la división sintetica y las iteraciones del método de Newton con la aproximación inicial $x_0 = i$.

	1 , 0	0 , 0	0 , 0	-2 , 0
$0 + 1\ i$		0 , 1	-1 , 0	0 , -1
	1 , 0	0 , 1	-1 , 0	-2 , -1
$0 + 1\ i$		0 , 1	-2 , 0	
	1 , 0	0 , 2	-3 , 0	

Así $P(i) = -2 - i$, $P'(i) = -3$, y

$$x_1 = x_0 - \frac{P(x_0)}{P'(x_0)} = i - \frac{-2 - i}{-3} = -\frac{2}{3} + \frac{2}{3}\ i \ .$$

Entonces, parece que el método converge a la raíz compleja $-0.630 + 1.091\ i$.

2. LA TECNICA DE DEFLACION

Si la $n-$ésima iteración, x_n, en el procedimiento de Newton-Raphson es un cero aproximado del polinomio P de grado N, entonces

$$P(x) = (x - x_n)\ Q(x) + d_N = (x - x_n)\ Q(x) + P(x_n) \approx (x - x_n)\ Q(x) \ ;$$

de lo cual, $x - x_n$ es un factor aproximado de $P(x)$. Tomando $\hat{x}_1 = x_n$ como un cero aproximado de P y $Q_1(x)$ como el factor aproximado,

$$P(x) \approx (x - \hat{x}_1)\ Q_1(x) \ ,$$

podemos encontrar un segundo cero aproximado de P aplicando el procedimiento de Newton-Raphson a $Q_1(x)$. Si P es un polinomio de grado N con N ceros reales, este procedimiento aplicado repetidamente, resultará eventualmente en $(N-2)$ ceros aproximados de P y en un factor cuadrático aproximado $Q_{N-2}(x)$. A este nivel, $Q_{N-2}(x) = 0$ puede resolverse por la fórmula cuadrática para encontrar los dos últimos ceros aproximados de P. Aún cuando este método puede ser usado para encontrar ceros aproximados de muchos polinomios, depende del uso repetido de aproximaciones y en ocasiones puede llevar a aproximaciones muy imprecisas. Este procedimiento se llama **deflación**. La dificultad de precisión de la deflación se debe al hecho de que, cuando obtenemos los ceros aproximados de P, el procedimiento de Newton-Raphson se usa en el polinomio reducido Q_k, o sea, el polinomio con la propiedad de que

$$P(x) \approx (x - \hat{x}_1)\ (x - \hat{x}_2)\ \ldots\ (x - \hat{x}_k)\ Q_k(x)\ .$$

Un cero aproximado $\hat{x}_{k+1}$ de Q_k generalmente no aproximará a una raíz de $P(x) = 0$ tan bien como una raíz de $Q_k(x) = 0$. La imprecisión usualmente es incrementada conforme k crezca. Una manera de eliminar esta dificultad consiste en usar las ecuaciones reducidas, esto es, los factores aproximados del polinomio original P, para encontrar aproximaciones, $\hat{x}_2$, $\hat{x}_3$, ..., $\hat{x}_k$ a los ceros de P y luego mejorar estas aproximaciones aplicando el procedimiento de Newton-Raphson al polinomio original P.

La deflación se usa con el método de Müller una vez que se ha determinado una raíz aproximada. Después de que se ha determinado una aproximación a la raíz de la ecuación deflactada es aconsejable usar, ya sea en el método de Müller o en el método de Newton, el polinomio original con esta aproximación como condición inicial. Esto asegurará que la raíz que se está aproximando sea una solución de la ecuación verdadera y no de la ecuación deflactada.

La siguiente técnica ha sido sugerida por Wilkinson: una vez encontrada una raíz p, entonces se considera la función

$$T(x) = \frac{P(x)}{x - p}\ .$$

El método de Newton se aplica entonces a la función $T(x)$ para dar

$$x_{n+1} = x_n - \frac{T(x_n)}{T'(x_n)} = x_n - [\frac{P'(x_n)}{P(x_n)} - \frac{1}{x_n - p}]^{-1}\ .$$

De esta manera uno puede trabajar con el polinomio original $P(x)$ en vez del polinomio deflactado, reduciendo el error. En general, habiendo encontrado los ceros p_1, p_2, ..., p_s, se puede usar la fórmula general

$$x_{n+1} = x_n - [\frac{P'(x_n)}{P(x_n)} - \sum_{k=1}^{s} \frac{1}{x_n - p_k}]^{-1}\ .$$

Se ha indicado previamente que el éxito del método de Newton depende frecuentemente de la obtención de una buena aproximación inicial. Una aproximación inicial x_0

mal escogida puede originar que la sucesión $\{p_n\}_{n=0}^{\infty}$ diverga también por polinomios. Si el polinomio real $P(x)$ no tiene raíces reales, entonces el método de Newton tiene que diverger para cualquier valor inicial $p_0 \in \mathcal{R}$. No hay reglas generales para escoger valores iniciales en el caso de polinomios genéricos, aunque la idea básica para encontrar ceros aproximados de P es la siguiente: evaluar P en puntos x_i para $i = 1, 2, \ldots, k$. Si $P(x_i)\ P(x_j) < 0$, entonces P tiene un cero entre x_i y x_j. El problema se transforma en escoger las x_i de tal manera que la posibilidad de perder un cambio de signo se minimice, mientras se mantiene el número de las x_i razonablemente pequeño. Sin embargo, existe una regla en el caso en que el polinomio tenga todas las raíces reales.

Teorema XII.6

Sea $P(x)$ un polinomio de grado $N \geq 2$ con coeficientes reales. Si todas las raíces ξ_i de $P(x)$ son reales y $\xi_N \leq \xi_{N-1} \leq \ldots \leq \xi_2 \leq \xi_1$, entonces el método de Newton lleva a una sucesión $\{p_n\}_{n=0}^{\infty}$ convergente y estrictamente decreciente para cualquier valor inicial $p_0 > \xi_1$.

Demostración: sin perder generalidad podemos asumir que $P(p_0) > 0$. Dado que $P(x)$ no cambia de signo para $x > \xi_1$, tenemos que $P(x) = a_0\ x^N + \ldots + a_N > 0$ para $x > \xi_1$, y entonces $a_0 > 0$. La derivada P' tiene $N-1$ ceros reales α_i con (para el Teorema de Rolle)

$$\xi_N \leq \alpha_{N-1} \leq \xi_{N-1} \leq \ldots \leq \alpha_2 \leq \xi_2 \leq \alpha_1 \leq \xi_1\ .$$

Dado que $P'(x)$ es de grado $N - 1 \geq 1$, éstas son todas las raíces, y además $P'(x) > 0$ para $x > \alpha_1$, dado que $a_0 > 0$. Usando otra vez el Teorema de Rolle, y recordando que $N \geq 2$, obtenemos:

$$P''(x) > 0 \quad \text{y} \quad P'''(x) \geq 0 \qquad \text{para} \quad x \geq \alpha_1\ .$$

Entonces, P y P' son funciones convexas para $x \geq \alpha_1$. Ahora bien, el hecho de que $p_n \geq \xi_1$ implica que

$$p_{n+1} = p_n - \frac{P(p_n)}{P'(p_n)} < p_n$$

dado que $P'(p_n) > 0$ y $P(p_n) > 0$.

Nos queda por demostrar que $p_{n+1} > \xi_1$. Por el Teorema de Taylor tenemos:

$$\begin{aligned} 0 = P(\xi_1) &= P(p_n) + (\xi_1 - p_n)\ P'(p_n) + \frac{(\xi_1 - p_n)^2}{2}\ P''(\delta) \\ &> P(p_n) + (\xi_1 - p_n)\ P'(p_n) \end{aligned}$$

dado que $\alpha_1 \leq \xi_1 < \delta < p_n$ implica que $P''(\delta) > 0$. De la definición de p_{n+1} se tiene que $P(p_n) = P'(p_n)\ (p_n - p_{n+1})$. Entonces,

$$0 > P'(p_n)\ (p_n - p_{n+1} + \xi_1 - p_n) = P'(p_n)\ (\xi_1 - p_{n+1})$$

que implica $\xi_1 - p_{n+1} < 0$ dado que $P'(p_n) > 0$, es decir, $p_{n+1} > \xi_1$. **c.q.d.**

3. EL METODO DE BAIRSTOW

Basandose sobre el Teorema XII.5, se puede diseñar una división sintética que involucre polinomios cuadráticos para factorizar aproximadamente el polinomio, de tal manera que uno de los términos sea un polinomio cuadrático cuyas raíces complejas sean aproximaciones a las raíces del polinomio original. Para introducir la **división sintetica cuadrática**, consideremos el polinomio $P(x)$ de grado N, de la forma $(XII.1)$, $P(x) = a_0\ x^N + a_1\ x^{N-1} + \ldots + a_{N-1}\ x + a_N$ y sea $x^2 - r\ x - s$ un término cuadrático fijo. Entonces, podemos escribir $P(x)$ de la forma

$$P(x) = (x^2 - r\ x - s)\ Q(x) + u\ (x - r) + v\ , \qquad (XII.8)$$

donde los términos $u\ (x - r) + v$ costituyen el resto cuando el polinomio $P(x)$ se divide entre $x^2 - r\ x - s$. Así, $Q(x)$ es un polinomio de grado $(N - 2)$ y se puede representar como

$$Q(x) = b_0\ x^{N-2} + b_1\ x^{N-3} + \ldots + b_{N-3}\ x + b_{N-2}\ . \qquad (XII.9)$$

Si ponemos $b_{N-1} = u$ y $b_N = v$, entonces la ecuación $(XII.8)$ se puede reescribir como

$$P(x) = (x^2 - r\ x - s)\ (b_0\ x^{N-2} + b_1\ x^{N-3} + \ldots + b_{N-3}\ x + b_{N-2}) + b_{N-1}\ (x - r) + b_N\ ,$$

que representado en potencias de x tiene la forma

$$\begin{aligned} P(x) = {} & b_0\ x^N + (b_1 - r\ b_0)\ x^{N-1} + (b_2 - r\ b_1 - s\ b_0)\ x^{N-2} + \\ & + \ldots + (b_k - r\ b_{k-1} - s\ b_{k-2})\ x^k + \ldots + \\ & + (b_{N-1} - r\ b_{N-2} - s\ b_{N-3})\ x + b_N - r\ b_{N-1} - s\ b_{N-2}\ . \end{aligned} \qquad (XII.10)$$

Comparando los coeficientes de las potencias x^k de la ecuación $(XII.10)$ con los de la $(XII.1)$, obtenemos los números b_k. Las fórmulas recursivas son las siguientes:

$$\begin{cases} b_0 = a_0 \\ b_1 = a_1 + r\ b_0 \\ b_k = a_k + r\ b_{k-1} + s\ b_{k-2} \quad \text{para } k = 2, 3, \ldots, N\ . \end{cases} \qquad (XII.11)$$

Cuando se hacen los cálculos a mano, se construye una nuova tabla para esta división sintetica cuadrática que tiene la siguiente forma.

	a_0	a_1	a_2	a_3	$\ldots$	a_k	$\ldots$	a_{N-2}	a_{N-1}	a_N
s			$s\ b_0$	$s\ b_1$	$\ldots$	$s\ b_{k-2}$	$\ldots$	$s\ b_{N-4}$	$s\ b_{N-3}$	$s\ b_{N-2}$
r		$r\ b_0$	$r\ b_1$	$r\ b_2$	$\ldots$	$r\ b_{k-1}$	$\ldots$	$r\ b_{N-3}$	$r\ b_{N-2}$	$r\ b_{N-1}$
	b_0	b_1	b_2	b_3	$\ldots$	b_k	$\ldots$	b_{N-2}	b_{N-1}	b_N

Ejemplo 6.

Usar la división sintetica cuadrática para dividir entre $(x^2 + 2\ x - 3)$ el polinomio $P(x) = x^5 + 6\ x^4 - 20\ x^2 + 22\ x + 8$.

En este caso nesesitamos usar $r = -2$ y $s = 3$; entonces se construye la tabla siguiente:

	1	6	0	-20	22	8
$s = 3$			3	12	-15	6
$r = -2$		-2	-8	10	-4	-6
	1	4	-5	2	$b_1 = 3$	$b_0 = 8$

Entonces, el polinomio $P(x)$ se puede factorizar como

$$P(x) = (x^2 + 2\ x - 3)\ (x^3 + 4\ x^2 - 5\ x + 2) + 3\ (x + 2) + 8\ .$$

Ahora usaremos la división sintetica cuadrática para introducir una técnica, conocida como el **método de Bairstow**, que es usada para encontrar un factor cuadrático, del tipo $(x^2 - r\ x - s)$, del polinomio $P(x)$.

Suponemos que empezamos con un factor cuadrático inicial

$$x^2 - r_0\ x - s_0 \tag{XII.12}$$

y que el polinomio $P(x)$ se pueda expresar como

$$P(x) = (x^2 - r_0\ x - s_0)\ Q(x) + u\ (x - r_0) + v\ . \tag{XII.13}$$

Cuando u y v son pequeños, el polinomio cuadrádico (XII.12) está cerca del factor del polinomio $P(x)$. Queremos encontrar nuevos valores r_1 y s_1 de manera que

$$x^2 - r_1\ x - s_1 \tag{XII.14}$$

sea más cerca al factor de $P(x)$ que el polinomio cuadrático inicial (XII.12). Nótese que u y v en (XII.13) son funciones de r y s, así que $u = u(r, s)$ y $v = v(r, s)$. Los nuevos valores r_1 y s_1 satisfacen las relaciones

$$r_1 = r_0 + \Delta r \qquad \text{y} \qquad s_1 = s_0 + \Delta s\ . \tag{XII.15}$$

Los diferenciales de las funciones u y v dan las aproximaciones

$$\begin{aligned} v(r_1, s_1) &\approx v(r_0, s_0) + v_r(r_0, s_0)\Delta r + v_s(r_0, s_0)\Delta s\ , \\ u(r_1, s_1) &\approx u(r_0, s_0) + u_r(r_0, s_0)\Delta r + u_s(r_0, s_0)\Delta s\ . \end{aligned} \tag{XII.16}$$

Si el polinomio cuadrático (XII.14) es un factor del polinomio $P(x)$, entonces los nuevos valores r_1 y s_1 tienen que satisfacer

$$u(r_1, s_1) = 0 \qquad \text{y} \qquad v(r_1, s_1) = 0\ . \tag{XII.17}$$

Cuando las cuantidades Δr y Δs son pequeñas, las aproximaciones $(XII.16)$ se pueden usar de manera que Δr y Δs sean la solución del sistema lineal

$$\begin{aligned} 0 &= v(r_0, s_0) + v_r(r_0, s_0)\Delta r + v_s(r_0, s_0)\Delta s \ , \\ 0 &= u(r_0, s_0) + u_r(r_0, s_0)\Delta r + u_s(r_0, s_0)\Delta s \ . \end{aligned} \qquad (XII.18)$$

Si se conocen los valores de las derivadas parciales que aparecen en el sistema $(XII.18)$, entonces Δr y Δs se pueden calcular usando las fórmulas de Cramer, y los nuevos valores de r_1 y s_1 se obtienen desde las ecuaciones $(XII.15)$. Deduciremos más adelante las expresiones de las derivadas parciales; por el momento decimos que estas están dadas por

$$v_r = c_{N-1}, \qquad v_s = c_{N-2}, \qquad u_r = c_{N-2}, \qquad u_s = c_{N-3} \ , \qquad (XII.19)$$

donde los coeficientes c_k están dados por las fórmulas recursivas

$$\begin{cases} c_0 = b_0 \\ c_1 = b_1 + r\ c_0 \\ c_k = b_k + r\ c_{k-1} + s\ c_{k-2} \quad \text{para } k = 2, 3, \ldots, N \ . \end{cases} \qquad (XII.20)$$

Las fórmulas $(XII.20)$ usan los coeficientes b_k que se habían calculado en las fórmulas recursivas $(XII.11)$. Dado que $u(r_0, s_0) = b_{N-1}$ y $v(r_0, s_0) = b_N$, el sistema lineal $(XII.18)$ se puede reescribir como

$$\begin{aligned} c_{N-1}\Delta r + c_{N-2}\Delta s &= -b_N \ , \\ c_{N-2}\Delta r + c_{N-3}\Delta s &= -b_{N-1} \ . \end{aligned} \qquad (XII.21)$$

Usamos ahora las fórmulas de Cramer para resolver el sistema $(XII.21)$. Los determinantes que se necesitan son

$$D = \det \begin{pmatrix} c_{N-1} & c_{N-2} \\ c_{N-2} & c_{N-3} \end{pmatrix}, \quad D_1 = \det \begin{pmatrix} -b_N & c_{N-2} \\ -b_{N-1} & c_{N-3} \end{pmatrix}, \quad D_2 = \det \begin{pmatrix} c_{N-1} & -b_N \\ c_{N-2} & -b_{N-1} \end{pmatrix},$$

y los nuevos valores r_1 y s_1 se calculan como

$$r_1 = r_0 + \frac{D_1}{D} \qquad \text{y} \qquad s_1 = s_0 + \frac{D_2}{D} \ . \qquad (XII.22)$$

El proceso iterativo continua hasta que se encontren buenas aproximaciones de r y s. Si las aproximaciones iniciales r_0 y s_0 se escogen pequeñas, la iteración en general converge. Cuando $x \approx 0$, las potencias grandes de x en el polinomio $P(x)$, $(XII.1)$, se pueden trascurar, y tenemos la aproximación $0 \approx P(x) \approx a_{N-2}\ x^2 + a_{N-1}\ x + a_N$. Entonces, las aproximaciones iniciales podrían ser

$$r_0 = -\frac{a_{N-1}}{a_{N-2}} \qquad \text{y} \qquad s_0 = -\frac{a_N}{a_{N-2}} \ , \qquad (XII.23)$$

siempre que $a_{N-2} \neq 0$.

Vamos ahora a derivar las fórmulas $(XII.20)$. La idea es diferenciar las ecuaciones $(XII.11)$ con respecto a r y s. Para empezar, nótese que $b_0 = a_0$ es una constante, así que sus derivadas parciales son cero. Continuando en la lista obtenemos

$$\begin{aligned}
&&\frac{\partial}{\partial s}b_0 &= 0\ ,\\
\frac{\partial}{\partial r}b_0 &= 0\ , &\frac{\partial}{\partial s}b_1 &= 0\ ,\\
\frac{\partial}{\partial r}b_1 &= b_0\ , &\frac{\partial}{\partial s}b_2 &= b_0 + r\ \frac{\partial}{\partial s}b_1 = b_0\ ,\\
\frac{\partial}{\partial r}b_2 &= b_1 + r\ \frac{\partial}{\partial r}b_1 &\frac{\partial}{\partial s}b_3 &= b_1 + r\ \frac{\partial}{\partial s}b_2 + s\ \frac{\partial}{\partial s}b_1\\
&= b_1 + r\ b_0\ , &&= b_1 + r\ b_0\ .
\end{aligned} \qquad (XII.24)$$

Diferenciando el término general en $(XII.11)$ con respecto a r y s, obtenemos

$$\frac{\partial}{\partial r}b_k = 0 + b_{k-1} + r\ \frac{\partial}{\partial r}b_{k-1} + 0 + s\ \frac{\partial}{\partial r}b_{k-2}\ , \qquad (XII.25)$$

y

$$\frac{\partial}{\partial s}b_{k+1} = 0 + 0 + r\ \frac{\partial}{\partial s}b_k + b_{k-1} + s\ \frac{\partial}{\partial s}b_{k-1}\ . \qquad (XII.26)$$

Entonces, empezando con las ecuaciones $(XII.24)$ y usando las $(XII.25)$ y $(XII.26)$, sigue que

$$\frac{\partial}{\partial r}b_k = \frac{\partial}{\partial s}b_{k+1}\ , \qquad \text{para}\ \ k = 0, 1, \ldots, N\ . \qquad (XII.27)$$

Y si definimos c_{k-1} el término común en $(XII.27)$, entonces, $(XII.25)$ se puede usar para mostrar que

$$\begin{aligned}
c_{k-1} = \frac{\partial}{\partial r}b_k &= b_{k-1} + r\ \frac{\partial}{\partial r}b_{k-1} + s\ \frac{\partial}{\partial r}b_{k-2} =\\
&= b_{k-1} + r\ c_{k-2} + s\ c_{k-3}\ .
\end{aligned} \qquad (XII.28)$$

Un método compacto para el cálculo es poner

$$b_{-1} = b_{-2} = c_{-1} = c_{-2} = 0\ , \qquad (XII.29)$$

$$b_k = a_k + r\ b_{k-1} + s\ b_{k-2} \quad \text{y} \quad c_k = b_k + r\ c_{k-1} + s\ c_{k-2}\ , \qquad (XII.30)$$

para $k = 0, 1, \ldots, N$, como dicho en las fórmulas $(XII.20)$.

Cuando se hacen los cálculos a mano del método de Bairstow, se construye una extensión de la tabla para la división sintetica cuadrática que tiene la siguiente forma.

	a_0	a_1	a_2	a_3	$\ldots$	a_{N-3}	a_{N-2}	a_{N-1}	a_N
s			$s\ b_0$	$s\ b_1$	$\ldots$	$s\ b_{N-5}$	$s\ b_{N-4}$	$s\ b_{N-3}$	$s\ b_{N-2}$
r		$r\ b_0$	$r\ b_1$	$r\ b_2$	$\ldots$	$r\ b_{N-4}$	$r\ b_{N-3}$	$r\ b_{N-2}$	$r\ b_{N-1}$
	b_0	b_1	b_2	b_3	$\ldots$	b_{N-3}	b_{N-2}	b_{N-1}	b_N
s			$s\ c_0$	$s\ c_1$	$\ldots$	$s\ c_{N-5}$	$s\ c_{N-4}$	$s\ c_{N-3}$	
r		$r\ c_0$	$r\ c_1$	$r\ c_2$	$\ldots$	$r\ c_{N-4}$	$r\ c_{N-3}$	$r\ c_{N-2}$	
	c_0	c_1	c_2	c_3	$\ldots$	c_{N-3}	c_{N-2}	c_{N-1}	

Ejemplo 7.

Dado el polinomio $P(x) = x^4 + x^3 + 3\ x^2 + 4\ x + 6$, usar el método de Bairstow, empezando con $r_0 = -2.1$ y $s_0 = -1.9$, para encontrar r_1, s_1, r_2, s_2, ..., los factores cuadráticos y las raíces de $P(x)$.

La tabla para calcular r_1 y s_1 es

	1.0000	1.0000	3.0000	4.0000	6.0000
$s = -1.9$			-1.9000	2.0900	-6.4790
$r = -2.1$		-2.1000	2.3100	-7.1610	2.2491
	1.0000	-1.1000	3.4100	$-1.0710 = b_3$	$1.7701 = b_4$
$s = -1.9$			-1.9000	6.0800	
$r = -2.1$		-2.1000	6.7200	-17.283	
	1.0000	$-3.2000 = c_1$	$8.2300 = c_2$	$-12.274 = c_3$	

El sistema lineal para Δr y Δs resultante es entonces

$$\begin{aligned} -12.274\ \Delta r + 8.2300\ \Delta s &= -1.7701\ , \\ 8.2300\ \Delta r - 3.2000\ \Delta s &= \ \ 1.0710\ . \end{aligned}$$

Usamos ahora las fórmulas de Cramer para resolver este sistema. Los determinantes son

$$D = -28.4561\ , \qquad D_1 = -3.15001\ , \qquad D_2 = 1.422469\ .$$

Entonces, los nuevos valores r_1 y s_1 son

$$r_1 = -2.1 + \frac{-3.15001}{-28.4561} = -1.98930282 \quad \text{y} \quad s_1 = -1.9 + \frac{1.422469}{-28.4561} = -1.94998819\ .$$

Otra iteración nos daría $r_2 = -1.99999277$ y $s_2 = -2.00015098$. Las sucesiones convergen a los valores $r = -2$ y $s = -2$, y $P(x)$ tiene la siguiente factorización

$$P(x) = (x^2 + 2\ x + 2)\ (x^2 - x + 3)\ .$$

Finalmente, sigue que las cuatro raíces complejas son

$$x_{1,2} = 1 \pm i \qquad \text{y} \qquad x_{3,4} = 0.5 \pm i\ 1.65831239\ .$$

4. EL METODO DE BERNOULLI

Los métodos discutidos en los capítulos anteriores son muy efectivos, pero solamente si se conoce una razonable primera aproximación a la solución deseada. Como obtener una tal primera aproximación es un problema que, para ecuaciones sin propiedades especiales, es de tal generalidad que no puede resolverse por reglas o algoritmos aplicables en todos los casos. Para polinomios, sin embargo, existen algoritmos que nos proporcionan la deseada primera aproximación sin que se tenga que usar otra información que el conocimiento de

los coeficientes del polinomio. El primero de los dos algoritmos que vamos a estudiar es el método clásico de Bernoulli.

Para comenzar con el caso más censillo, supongamos que el polinomio de grado N

$$P(x) = a_0\ x^N + a_1\ x^{N-1} + \ldots + a_{N-1}\ x + a_N\ , \qquad (XII.1)$$

cuyos coeficientes pueden ser complejos, tiene N ceros distintos, z_1, z_2, …, z_N. Si miramos a la ecuación de diferencias

$$a_0\ x_n + a_1\ x_{n-1} + \ldots + a_N\ x_{n-N} = 0\ , \qquad (XII.31)$$

la solución $X = \{x_n\}$ (cualesquiera que sean sus valores iniciales) debe poderse representar en la forma

$$x_n = c_1\ z_1^n + c_2\ z_2^n + \ldots + c_N\ z_N^n\ , \qquad (XII.32)$$

en donde c_1, c_2, …, c_N son constantes adecuadas.

Para proseguir, formulemos las dos hipótesis siguientes:

i) El polinomio P tiene un solo cero dominante, es decir, uno de los ceros (por ejemplo, z_1) tiene un módulo mayor que el de todos los restantes:

$$|z_1| > |z_k|\ , \qquad k = 2, 3, \ldots, N\ . \qquad (XII.33)$$

ii) Los valores iniciales son tales que el cero dominante está representado en la solución ($XII.32$), es decir, se tiene que

$$c_1 \neq 0\ . \qquad (XII.34)$$

Consideramos ahora la razón de dos valores consecutivos de la sucesión solución X. Usando ($XII.32$) encontramos

$$\frac{x_{n+1}}{x_n} = \frac{c_1\ z_1^{n+1} + c_2\ z_2^{n+1} + \ldots + c_N\ z_N^{n+1}}{c_1\ z_1^n + c_2\ z_2^n + \ldots + c_N\ z_N^n}\ .$$

En virtud de ($XII.34$) esto puede escribirse

$$\frac{x_{n+1}}{x_n} = z_1\ \frac{1 + \dfrac{c_2}{c_1}\ \left(\dfrac{z_2}{z_1}\right)^{n+1} + \ldots + \dfrac{c_N}{c_1}\ \left(\dfrac{z_N}{z_1}\right)^{n+1}}{1 + \dfrac{c_2}{c_1}\ \left(\dfrac{z_2}{z_1}\right)^{n} + \ldots + \dfrac{c_N}{c_1}\ \left(\dfrac{z_N}{z_1}\right)^{n}}\ . \qquad (XII.35)$$

Por ($XII.33$), $|z_k/z_1| < 1$ para $k = 2, 3, \ldots, N$. Se sigue que

$$\left(\frac{z_k}{z_1}\right)^n \to 0 \qquad \text{cuando} \quad n \to \infty \qquad (XII.36)$$

para $k = 2, 3, \ldots, N$. La fracción que multiplica a z_1 tiende pues a 1 cuando $n \to \infty$, y tenemos por tanto

$$\lim_{n\to\infty}\ \frac{x_{n+1}}{x_n} = z_1\ .$$

Tenemos pues la siguiente formulación tentativa del **método de Bernoulli**:

- Escójanse valores arbitrarios x_0, $x_{-1}, \ldots$, x_{1-N}, y determínase la sucesión $X = \{x_n\}$ por la relación de recurrencia

$$x_n = -\frac{a_1\ x_{n-1} + a_2\ x_{n-2} + \ldots + a_N\ x_{n-N}}{a_0}\ , \qquad n = 1, 2, \ldots . \tag{XII.37}$$

- Fórmese luego la sucesión de cocientes

$$q_n = \frac{x_{n+1}}{x_n}\ . \tag{XII.38}$$

Hemos entonces probado que

Teorema XII.7

Si el cero dominante del polinomio P dado por (XII.1) es único y si se satisface la condición (XII.34), entonces los cocientes q_n están "*al final*" definidos y convergen al cero dominante del polinomio P.

Ejemplo 8.

Sea

$$P(x) = 70\ x^4 - 140\ x^3 + 90\ x^2 - 20\ x + 1\ .$$

La ecuación de diferencia (XII.31) (resuelta para x_n), o que es lo mismo la ecuación (XII.37), toma aquí la forma

$$x_n = \frac{140\ x_{n-1} - 90\ x_{n-2} + 20\ x_{n-3} - x_{n-4}}{70}\ .$$

En la siguiente tabla se dan los valores de n, de la sucesión x_n (iniciada con $x_0 = 1$, $x_n = 0$ para $n < 0$) y de la sucesión q_n.

Tabla 1

n	x_n	q_n
0	1.0	2.0
1	2.0	1.3571429
2	2.7142857	1.1578947
3	3.1428571	1.0668831
4	3.3530611	1.0176506
5	3.4122447	0.9883800
6	3.3725945	0.9699170
7	3.2711367	0.9578036
⋮	⋮	⋮
14	2.1627466	0.9330692
15	2.0179923	0.9323640
16	1.8815034	0.9318586
17	1.8815034	
⋮	⋮	⋮
⋮	⋮	
⋮	⋮	0.93057

Lo precedente es un mero bosquejo del método de Bernoulli en el caso más sencillo posible. Todavía queda por hacer frente a cierto número de complicaciones como, por

ejemplo, las siguentes: convergencia lenta, ceros de multiplicidad mayor que uno, elección desafortunada de valores iniciales y la presencia de varios ceros dominantes.

Incluso si las condiciones de convergencia del método de Bernoulli se satisfacen, la velocidad de convergencia puede ser baja. Entendemos por esto que el error de la aproximación q_n al cero z_1, $e_n = q_n - z_1 = \frac{x_{n+1}}{x_n} - z_1$ tiende a cero lentamente. Como en el caso de la iteración, puede ser posible acelerar la convergencia haciendo un uso juicioso de la información sobre la forma en que e_n tiende a cero. Para descrubir esta forma de convergencia, analizaremos los errores e_n con más atención. Continuaremos con nuestras hipótesis (i) y (ii), y aparte, suponemos que

$$|z_1| > |z_2| > |z_k| \ , \qquad k = 3, 4, \ldots, N \ , \tag{XII.39}$$

(el cero que sigue al dominante es el único cero de su módulo), y que

$$c_2 \neq 0 \ , \tag{XII.40}$$

(el cero que sigue al dominante está representado en la solución $\{x_n\}$).

Bajo estas hipótesis, el error

$$e_n = \frac{c_1\ z_1^{n+1} + c_2\ z_2^{n+1} + \ldots + c_N\ z_N^{n+1} - z_1(c_1\ z_1^n + c_2\ z_2^n + \ldots + c_N\ z_N^n)}{c_1\ z_1^n + c_2\ z_2^n + \ldots + c_N\ z_N^n} =$$
$$= \frac{c_2\ (z_2 - z_1)\ z_2^n + \ldots + c_n\ (z_n - z_1)\ z_N^n}{c_1\ z_1^n + c_2\ z_2^n + \ldots + c_N\ z_N^n} \ ,$$

puede escribirse en la forma

$$e_n = A\ t^n\ (1 + \varepsilon_n) \ , \tag{XII.41}$$

donde

$$A = \frac{c_2(z_2 - z_1)}{c_1} \ , \qquad t = \frac{z_2}{z_1}$$

y

$$1 + \varepsilon_n = \frac{1 + \dfrac{c_3(z_3 - z_1)}{c_2(z_2 - z_1)} \left(\dfrac{z_3}{z_2}\right)^n + \ldots + \dfrac{c_N(z_N - z_1)}{c_2(z_2 - z_1)} \left(\dfrac{z_N}{z_2}\right)^n}{1 + \dfrac{c_2}{c_1} \left(\dfrac{z_2}{z_1}\right)^n + \ldots + \dfrac{c_N}{c_1} \left(\dfrac{z_N}{z_1}\right)^n} \ .$$

En virtud de la condición (XII.39), tenemos, aparte de (XII.36), $\left(\dfrac{z_k}{z_2}\right)^n \to 0$ cuando $n \to \infty$, para $k = 3, 4, \ldots, N$, y la razón que aparece en el segundo miembro tiene el límite 1 cuando $n \to \infty$. Se sigue que

$$\lim_{n \to \infty} \varepsilon_n = 0 \ .$$

Como consecuencia,

$$\frac{e_{n+1}}{e_n} = t\ \frac{1 + \varepsilon_{n+1}}{1 + \varepsilon_n} = t\ (1 + \delta_n)$$

donde $\delta_n = \frac{\varepsilon_{n+1}-\varepsilon_n}{1+\varepsilon_n} \to 0$ cuando $n \to \infty$. Los errores e_n satisfacen pues las condiciones del Teorema X.4 que hace efectivo el proceso Δ^2 de Aitken. Tenemos pues como consecuencia inmediata el siguiente

Teorema XII.8

Bajo las hipótesis formuladas al comienzo de la presente sección, la sucesión $\{q'_n\}_{n=0}^{\infty}$ derivada de la sucesión $\{q_n\}$ por medio de la fórmula Δ^2 de Aitken

$$q'_n = q_n - \frac{(\Delta q_n)^2}{\Delta^2 q_n}$$

converge más rápidamente al cero dominante z_1 que la sucesión $\{q_n\}_{n=0}^{\infty}$ en el sentido de que

$$\lim_{n\to\infty} \frac{q'_n - z_1}{q_n - z_1} = 0 \ .$$

Ejemplo 9.

Sea el mismo polinomio del ejemplo 8

$$P(x) = 70\ x^4 - 140\ x^3 + 90\ x^2 - 20\ x + 1 \ .$$

La siguiente tabla es una extensión de la tabla 1: se dan los valores de n, de la sucesión x_n (iniciada con $x_0 = 1$, $x_n = 0$ para $n < 0$), de la sucesión q_n y en la última columna los valores q'_n. También están dados los valores intermedios requeridos para calcular q'_n. La convergencia más rápida de la sucesión $\{q'_n\}$ al cero exacto $z_1 = 0.93057$ es evidente.

Tabla 2

n	x_n	q_n	Δq_n	$\Delta^2 q_n$	$-\frac{(\Delta q_n)^2}{\Delta^2 q_n}$	q'_n
0	1.0	2.0				
1	2.0	1.3571429				
2	2.7142857	1.1578947	−0.0910116	0.0417791	−0.1982597	0.9596350
3	3.1428571	1.0668831	−0.0492325	0.0199619	−0.1214233	0.9454595
4	3.3530611	1.0176506	−0.0292706	0.0108076	−0.0792746	0.9383760
5	3.4122447	0.9883800	−0.0184630	0.0063496	−0.0536856	0.9346444
6	3.3725945	0.9699170	−0.0121134	0.0039481	−0.0371658	0.9327512
7	3.2711367	0.9578036	−0.0081653	0.0025560		
⋮	⋮	⋮	⋮	⋮	⋮	⋮
13	2.3154382	0.9340550	−0.0009858	0.0002806	−0.0034633	0.9305917
14	2.1627466	0.9330692	−0.0007052	0.0001998	−0.0024890	0.9305802
15	2.0179923	0.9323640	−0.0005054			
16	1.8815034	0.9318586				
17	1.8815034					
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	0.93057	⋮	⋮	⋮	0.93057

Si el polinomio P tiene ceros no dominantes repetidos $z_2, \ldots, z_N$, entonces la fórmula $(XII.32)$ para la solución general de la ecuación de diferencias $(XII.31)$ contiene también

términos de la forma $n^k\ z_2^n$. Por tanto, la expresión $(XII.35)$ contiene términos como $n^k\ \left(\dfrac{z_2}{z_1}\right)^n$ además de $\left(\dfrac{z_2}{z_1}\right)^n$. Sin embargo, estos términos no pueden alterar la convergencia del método, ya que si $|q|<1$, entonces tenemos no solamente que $q_n \to 0$ cuando $n \to \infty$, sino que también $n^k\ q_n \to 0$ para cualquier valor fijo de k.

La situación es diferente si el cero *dominante* tiene una multiplicidad >1. (Todavía estamos suponiendo que hay solo un único cero dominante). Para fijar las ideas, supongamos que la multiplicidad es 2. La relación $(XII.32)$ toma ahora la forma (suponiendo aún que $c_1 \neq 0$)

$$x_n = c_1\ n\ z_1^n + c_2\ z_1^n + c_3\ z_3^n + \dots\ ,$$

donde $|z_3| < |z_1|$. La razón $(XII.35)$ toma ahora la forma (suponiendo aún que $c_1 \neq 0$)

$$\begin{aligned}\frac{x_{n+1}}{x_n} &= \frac{c_1\ (n+1)\ z_1^{n+1} + c_2\ z_1^{n+1} + c_3\ z_3^{n+1} + \dots}{c_1\ n\ z_1^n + c_2\ z_1^n + c_3\ z_3^n + \dots} = \\ &= z_1\ \frac{(n+1)\ c_1 + c_2}{n\ c_1 + c_2}\ \frac{1 + \dfrac{c_3}{(n+1)\ c_1 + c_2}\ \left(\dfrac{z_3}{z_1}\right)^{n+1} + \dots}{1 + \dfrac{c_3}{n\ c_1 + c_2}\ \left(\dfrac{z_3}{z_1}\right)^{n} + \dots}\ .\end{aligned} \qquad (XII.42)$$

La convergencia aún tiene lugar, pero, debido al factor $\frac{(n+1)\ c_1+c_2}{n\ c_1+c_2} = 1 + \frac{1}{n+\frac{c_2}{c_1}}$ con una velocidad mucho menor; el error después de n pasos es ahora del orden de $1/n$ y no del de $(z_1/z_2)^n$ como sería el caso si el cero dominante tuviera multiplicidad igual a uno.

Una de las condiciones para la convergencia del algoritmo expresado en las fórmulas $(XII.37)$ y $(XII.38)$, era que $c_1 \neq 0$. Se puede mostrar por medio de la teoría de funciones de variables compleja que esta condición se satisface siempre si los valores iniciales se escogen como sigue:

$$x_{-N+1} = x_{-N+2} = \dots = x_{-1} = 0\ , \quad x_0 = 1\ . \qquad (XII.43)$$

Una elección de valores iniciales diferente y más sofisticada queda definida por el siguiente **algoritmo de Bernoulli**:

- Si los coeficientes $a_0, a_1, \dots, a_N$ han sido dados, calcúnse x_0, $x_1, \dots$, x_{N-1} por las fórmulas

$$\begin{aligned} x_0 &= -\frac{a_1}{a_0}\ , \\ x_1 &= -\frac{1}{a_0}\ (2\ a_2 + a_1\ x_0)\ , \\ x_2 &= -\frac{1}{a_0}\ (3\ a_3 + a_2\ x_0 + a_1\ x_1)\ , \\ x_k &= -\frac{1}{a_0}\ ((k+1)\ a_{k+1} + a_k\ x_0 + a_{k-1}\ x_1 + \dots + a_1\ x_{k-1})\ , \end{aligned} \qquad (XII.44)$$

en general, para $k = 1, 2, \dots N-1$. Luego, como antes, determínase la sucesión $X = \{x_n\}$ por la relación de recurrencia

$$x_n = -\frac{a_1\ x_{n-1} + a_2\ x_{n-2} + \dots + a_N\ x_{n-N}}{a_0}\ , \quad n = N, N+1, N+2, \dots . \qquad (XII.45)$$

- Fórmese luego la sucesión de cocientes

$$q_n = \frac{x_{n+1}}{x_n} \ . \qquad (XII.38)$$

Se puede demostrar por medio de la teoría de las funciones complejas el siguiente teorema del cual omitimos la demostración.

Teorema XII.9

Si cada uno de los ceros distintos z_i tienen multiplicidad m_i, con $i = 1, 2, \ldots, M$, $(M \leq N)$, y si los valores iniciales para el método de Bernoulli están determinados por $(XII.44)$, entonces que la relación $(XII.32)$ toma la forma

$$x_n = m_1 \ z_1^{n+1} + m_2 \ z_2^{n+1} + \ldots + m_M \ z_M^{n+1} \ , \qquad (XII.46)$$

La relación $(XII.46)$ es notable por el hecho de que no aparecen potencias de n, no obstante la posible presencia de ceros de multiplicidad más alta que uno. La dificultad de los ceros con multiplicidad > 1 puede pues así siempre ser evitada por una elección adecuada de los valores iniciales. La razón x_{n+1}/x_n converge entonces a una velocidad determinada solamente por la magnitud de los ceros mayores.

Ejemplo 10.

Vamos a comparar las sucesiones $\{q_n\}$ para el polinomio

$$p(x) = (x-3)^2(x+1)^2 = x^4 - 4 \ x^3 - 2 \ x^2 + 12 \ x + 9$$

generadas iniciando el método de Bernoulli por la $(XII.43)$ y por la $(XII.44)$. La relación de recurrencia es, tanto en un caso como en otro,

$$x_n = 4 \ x_{n-1} + 2 \ x_{n-2} - 12 \ x_{n-3} - 9 \ x_{n-4}$$

Tabla 3

n	x_n	q_n	x_n	q_n
-3	0			
-2	0			
-1	0			
0	1	4.0	4	5.0
1	4	4.5	20	2.6
2	18	3.78	52	3.15385
3	68	3.69	164	2.95122
4	251	3.54	484	3.01653
5	888	3.46	1460	2.99452
6	3076	3.40	4372	3.00183
7	10456	3.35	13124	2.99939
8	35061	3.32	39364	3.00020
9	116252	3.29	118100	2.99993
10	381974	3.26086	354292	3.00002
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		3.00000		3.00000

La teoría que hasta el momento hemos venido desarrollando es aplicable perfectamente tanto si los coeficientes del polinomio P son reales como si son complejos. Sin embargo, siempre hemos supuesto que z_1 es el único cero dominante de P. Consideremos ahora el caso en que P es un polinomio con coeficientes reales que tiene un par de complejos conjugados como ceros dominantes, z_1 y $z_2 = \overline{z}_1$, ambos de multiplicidad 1. Los restantes ceros satisfarán la disegualdad

$$|z_k| < |z_1| \ , \qquad k = 3, 4, \ldots, N \ . \tag{XII.47}$$

Si los valores iniciales para la sucesión $\{x_n\}$ son reales, la ecuación (XII.32) toma la forma

$$x_n = c_1 \ z_1^n + \overline{c}_1 \ \overline{z}_1^n + c_3 \ z_3^n + \ldots + c_N \ z_N^n \ .$$

Representando los número complejos c_1 y z_1 en forma polar, escribimos $z_1 = r \ e^{i \ \varphi}$, $c_1 = a \ e^{i \ \delta}$, donde $r > 0$ y $a > 0$. Podemos suponer, además, que z_1 es el cero situado en el semiplano superior y, por consiguiente, que $0 < \varphi < \pi$. La expresión para x_n toma ahora la forma

$$x_n = 2 \ a \ r^n \ \cos(n \ \varphi + \delta) + c_3 \ z_3^n + \ldots + c_N \ z_N^n \ .$$

Esto puede escribirse

$$x_n = 2 \ a \ r^n \ [\cos(n \ \varphi + \delta) + \theta_n] \ , \tag{XII.48}$$

con $\theta_n = \dfrac{c_3}{2 \ a} \ \left(\dfrac{z_3}{r}\right)^n + \ldots + \dfrac{c_N}{2 \ a} \ \left(\dfrac{z_N}{r}\right)^n$, de donde, para alguna C constante adecuada $|\theta_n| \leq C \ t^n$, en donde t denota la mayor de las razones $|z_k/r|$, $k = 3, \ldots N$. Por (XII.47), el número t es menor que 1, y por tanto

$$\left| \lim_{n \to \infty} \theta_n \right| = 0 \ . \tag{XII.49}$$

Nuestro problema es el de recobrar las cantidades r y φ de la sucesión $\{x_n\}$. Para encontrar su solución, comencemos por suponer que $\theta_n = 0$ y tendremos en cuenta que los elementos $x_n = 2 \ a \ r^n \ \cos(n \ \varphi + \delta)$ son solución de la ecuación de diferencias

$$x_n + A \ x_{n-1} + B \ x_{n-2} = 0 \ , \tag{XII.50}$$

donde $B = r^2$ y $A = -2 \ r \ \cos\varphi$. Para determinar los coeficientes A y B partiendo de la solución conocida $\{x_n\}$, observamos que la ecuación (XII.50) junto con la ecuación correspondiente con n aumentado de una unidad, $x_{n+1} + A \ x_n + B \ x_{n-1} = 0$, representa un sistema de dos ecuaciones lineales para las dos incognitas A y B. El determinante

$$D_n = \begin{vmatrix} x_{n-1} & x_{n-2} \\ x_n & x_{n-1} \end{vmatrix} \tag{XII.51}$$

es igual a $4 \ a^2 \ r^{2n-2}\{\cos^2[(n-1)\varphi+\delta] - \cos(n\varphi+\delta) \ \cos[(n-2)\varphi+\delta]\} = 4 \ a^2 \ r^{2n-2} \sin^2\varphi$, por una identidad trigonométrica, y es, por tanto, diferente de cero, pues $0 < \varphi < \pi$. Podemos por tanto resolverlo para A y B, encontrando que

$$A = -\frac{E_n}{D_n} \ , \qquad B = \frac{D_{n+1}}{D_n} \ ,$$

donde

$$E_n = \begin{vmatrix} x_n & x_{n-2} \\ x_{n+1} & x_{n-1} \end{vmatrix} . \qquad (XII.52)$$

Las cantidades buscadas r y φ pueden ahora encontrarse por las igualdades

$$r = \sqrt{B} = \sqrt{\frac{D_{n+1}}{D_n}} , \quad \cos\varphi = -\frac{A}{2r} = \frac{E_n}{2\sqrt{D_n\ D_{n+1}}} . \qquad (XII.53)$$

Las relaciones $(XII.53)$ resuelven nuestro problema si $\{x_n\}$ es una solución de la ecuación de diferencias $(XII.31)$ tal que $\theta_n = 0$ en $(XII.48)$. Si $\{x_n\}$ es una solución cualquiera de $(XII.31)$, $\theta_n \neq 0$ en general; sin embargo, usando $(XII.49)$ no es difícil probar que

$$\frac{D_{n+1}}{D_n} \to r^2 \quad \text{y} \quad \frac{E_n}{2\ D_n} \to r\ \cos\varphi , \qquad (XII.54)$$

cuando $n \to \infty$. Tenemos, por lo tanto, el siguiente algoritmo para encontrar una pareja de ceros dominantes complejos conjugados:

- Partiendo de la solución $\{x_n\}$ de la ecuación de diferencias $(XII.31)$, calcúlense los determinantes D_n y E_n definidos por las fórmulas $(XII.51)$ y $(XII.52)$.
- Fórmese con estos la sucesión de razones $\frac{D_{n+1}}{D_n}$ y $\frac{E_n}{2\ D_n}$. Si la solución de la ecuación de diferencias $(XII.31)$ es tal que $c_1 \neq 0$ en $(XII.32)$, entonces las relaciones límites $(XII.54)$ se verifican.

Nótese que este algoritmo tiene la desventaja de no ser muy exacto cuando φ, el argumento del cero dominante, es pequeño.

Teorema XII.10

Si el polinomio $(XII.1)$ tiene exactamente dos ceros dominantes $z_{1,2} = r\ e^{\pm i\varphi}$, ambos de multiplicidad uno, con $0 < \varphi < \pi$. Si la solución de la ecuación de diferencias $(XII.31)$ es tal que $c_1 \neq 0$ in $(XII.32)$, entonces las relaciones límites $(XII.54)$ se verifican.

Una pregunta frecuente es: ¿cómo se pueden detectar un par de ceros complejos conjugados? La fórmula $(XII.48)$ muestra que en presencia de un par de ceros dominantes que sean complejos conjugados los elementos de la sucesión $\{x_n\}$ se comportan como $x_n = 2\ a\ r^n\ \cos(n\ \varphi + \delta)$. Esta ecuación implica, en particular, que los signos de las x_n oscilan. Por otra parte, viendo la frecuencia con la que estas oscilaciones se presentan, podemos esperar ser capaces de presentar una proposición acerca del ángulo φ. Es claro que cuando φ es pequeño, las "*ondas de signo*" son largas, y que cuando φ es próximo a π, las ondas son cortas.

Ejemplo 11.

Para el polinomio

$$P(x) = 81\ x^4 - 188\ x^3 + 24\ x + 20 ,$$

con ceros dominantes exactos $z_{1,2} = 1 \pm \frac{1}{3}i$, el cálculo se efectúa como en la tabla siguiente. Es evidentemente innecesario calcular los determinantes D_n y E_n desde el principio.

Tabla 4

x_n	D_n	$\frac{D_{n+1}}{D_n}$	E_n	$\frac{E_n}{2\ D_n}$
1				
1.333333				
1.777778				
2.074074				
2.123457				
1.975309				
1.333333				
1.580247				
0.965707				
0.178022				
−0.718589				
−1.634438				
−2.470444				
−3.124966				
−3.504914	1.229324	1.111187	2.458740	1.000037
−3.537670	1.366009	1.111125	2.732037	1.000006
−3.180991	1.517807	1.111095	3.035585	0.999990
−2.431232	1.686428	1.111115	3.372868	1.000003
−1.328033	1.873816	1.111111	3.747630	0.999999
0.045304	2.082018	1.111113	4.164035	1.000000
1.566200	2.313358		4.626705	
⋮	⋮	⋮	⋮	⋮
		1.111111		1.000000

5. EL ALGORITMO DEL COCIENTE-DIFERENCIA

El método de Bernoulli tiene la desventaja de proporcionar solamente los ceros dominantes de un polinomio. Si se desea calcular un cero no dominante por el método de Bernoulli, es necesario primero calcular todos los ceros mayores, y luego sacarlos del polinomio por la técnica de deflación. Solamente en raras ocasiones se conocen exactamente estos ceros. Las sucesivas deflaciones tenderán pues a falsificar los ceros restantes. Discutiremos ahora una moderna extensión del método de Bernoulli, debida a Rutishauser, que tiene la ventaja de proveer aproximaciones simultáneas a *todos* los ceros. Como en el apartado anterior no se presenterá la demostración de los teoremas debido a la necesidad de usar la teoría de las funciones complejas, que no es prerrequisito de este curso.

El **algoritmo del cociente-diferencia (Q-D)** puede considerarse como una generalización del método de Bernoulli. Como en este último método, se nos da un polinomio

$$P(x) = a_0\ x^N + a_1\ x^{N-1} + \ldots + a_{N-1}\ x + a_N\ , \qquad (XII.1)$$

y formamos una solución de la ecuación de diferencias asociada

$$a_0\ x_n + a_1\ x_{n-1} + \ldots + a_N\ x_{n-N} = 0\ . \qquad (XII.31)$$

La sucesión $\{x_n\}$ puede por ejemplo iniciarse haciendo

$$x_{-N+1} = x_{-N+2} = \ldots = x_{-1} = 0\ , \quad x_0 = 1\ . \qquad (XII.43)$$

En el método de Bernoulli, ahora formamos los cocientes

$$q_n = \frac{x_{n+1}}{x_n} = q_n^{(1)} \ . \qquad (XII.55)$$

Si el polinomio P tiene un solo cero dominante, entonces la sucesión $\{q_n\}$ converge a él.

Los elementos de esta sucesión $\{q_n\}$ se denotan ahora por $q_n^{(1)}$, y forman la primera columna del esquema bidimensional llamado el esquema del **cociente-diferencia (QD)**. Los elementos de las restantes columnas se denotan convencionalmente por $e_n^{(1)}$, $q_n^{(2)}$, $e_n^{(2)}$, $q_n^{(3)}$, ..., $e_n^{(N-1)}$, $q_n^{(N)}$ y se generan por la formación alternada de diferencias y cocientes, como sigue:

$$e_n^{(k)} = [q_{n+1}^{(k)} - q_n^{(k)}] + e_{n+1}^{(k-1)} \ , \qquad (XII.56)$$

$$q_n^{(k+1)} = \frac{e_{n+1}^{(k)}}{e_n^{(k)}} \ q_{n+1}^{(k)} \ , \qquad (XII.57)$$

en donde $k = 1, 2, \ldots, N-1$, $n = 0, 1, 2, \ldots$. En $(XII.56)$ hacemos $e_n^{(0)} = 0$ para $k = 1$. El número de columnas de q formadas es igual al grado del polinomio dado.

Ejemplo 12.

Para el grado $N = 4$, el esquema general QD tiene el siguiente aspecto:

	$q_0^{(1)}$						
0		$e_0^{(1)}$					
	$q_1^{(1)}$		$q_0^{(2)}$				
0		$e_1^{(1)}$		$e_0^{(2)}$			
	$q_2^{(1)}$		$q_1^{(2)}$		$q_0^{(3)}$		
0		$e_2^{(1)}$		$e_1^{(2)}$		$e_0^{(3)}$	
	$q_3^{(1)}$		$q_2^{(2)}$		$q_1^{(3)}$		$q_0^{(4)}$
0		$e_3^{(1)}$		$e_2^{(2)}$		$e_1^{(3)}$	
	$q_4^{(1)}$		$q_3^{(2)}$		$q_2^{(3)}$		$q_1^{(4)}$
0		$e_4^{(1)}$		$e_3^{(2)}$		$e_2^{(3)}$	
$\vdots$	$q_5^{(1)}$	$\vdots$	$q_4^{(2)}$	$\vdots$	$q_3^{(3)}$	$\vdots$	$q_2^{(4)}$
	$\vdots$		$\vdots$		$\vdots$		$\vdots$

En cada columna del esquema los superíndices son constantes, y en cada diagonal son constantes los subíndices. Las reglas $(XII.56)$ y $(XII.57)$ pueden memorizarse observando que en cada una de las configuraciones análogas a rombos que se muestran en el esquema, o las sumas o los productos de los pares de elementos SW o NE, son iguales. Si un rombo está centrado en una columna q, las sumas son iguales; si está centrado en una columna e, son los productos los que son iguales. En vista de esta interpretación, frecuentemente se hace referencia a las fórmulas $(XII.56)$ y $(XII.57)$ llamándolas *reglas del rombo*.

El esquema QD puede describirse también de otra forma si introducimos, aparte del operador de diferencias progresiva Δ (definido por $\Delta\, x_n = x_{n+1} - x_n$), el operador cociente

Q definido por $Q\ x_n = \frac{x_{n+1}}{x_n}$. Las relaciones $(XII.56)$ y $(XII.57)$ pueden entonces escribirse de manera más compacta del siguiente modo:

$$e_n^{(k)} = e_{n+1}^{(k-1)} + \Delta\ q_n^{(k)}\ , \qquad q_n^{(k+1)} = q_{n+1}^{(k)}\ Q\ e_n^{(k)}\ . \tag{$XII.58$}$$

Ejemplo 13.

Veamos el esquema QD numérico del polinomio

$$P(x) = x^2 - x - 1\ .$$

La sucesión $\{x_n\}$ es la de Fibonacci.

Tabla 5

x_n	$e_n^{(0)}$	$q_n^{(1)}$	$e_n^{(1)}$	$q_n^{(2)}$	$e_n^{(2)}$
1	0				
		1.000000			
1	0		1.000000		
		2.000000		−1.000000	
2	0		−0.500000		−0.000001
		1.500000		−0.500001	
3	0		0.166667		−0.000001
		1.666667		−0.666669	
5	0		−0.066667		0.000002
		1.600000		−0.600000	
8	0		0.025000		0.000025
		1.625000		−0.624975	
13	0		−0.009615		−0.000049
		1.615385		−0.615409	
21	0		0.003663		−0.000171
		1.619048		−0.619243	
34	0		−0.001401		⋮
		1.617647		⋮	
55	0		⋮		
		⋮			
⋮	⋮	↓		↓	
		$\frac{1+\sqrt{5}}{2}$		$\frac{1-\sqrt{5}}{2}$	

Evidentemente el esquema QD no existe si un coeficiente $e_n^{(k)}$, con $1 \le k \le N-1$ se hace cero. Otro caso trivial de no existencia del esquema se presenta cuando x_n llega a hacerse cero. Parece difícil enunciar condiciones explícitas necesarias y suficientes para la existencia del esquema en términos del polinomio P. En términos de la sucesión $\{x_n\}$, una condición necesaria y suficiente es que los determinantes

$$H_n^{(k)} = \begin{vmatrix} x_n & x_{n+1} & \dots & x_{n+k-1} \\ x_{n+1} & x_{n+2} & \dots & x_{n+k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n+k-1} & x_{n+k} & \dots & x_{n+2k-2} \end{vmatrix}\ , \tag{$XII.59$}$$

deben ser diferentes de cero para $k = 1, 2, \ldots, N$ y para $n = 0, 1, 2, \ldots$. Es posible enunciar condiciones sencillas *suficientes* (pero no necesarias) para que éste sea el caso. Entre ellas tenemos las siguientes:

(i) Los ceros $z_1, z_2, \ldots, z_N$ de P son positivos, y la sucesión $\{x_n\}$ se inicia con el algoritmo $(XII.44)$.

(ii) Los ceros $z_1, z_2, \ldots, z_N$ de P son simples (pero no necesariamente reales) y tienen valores absolutos distintos, $|z_1| > |z_2| > \ldots > |z_N| > 0$.

Hay una gran cantidad de evidencia numérica de que el esquema QD existe en muchos casos, incluso si no se satisfacen ningun de las anteriores condiciones suficientes, por ejemplo, si P es un polinomio con coeficientes reales que tienen como cero pares de complejos conjugados.

Si el esquema QD existe, son posibles varios enunciados notables acerca de los límites de sus elementos cuando $n \to \infty$. La situación más simple surge si los ceros del polinomio P satisfacen la condición

$$|z_1| > |z_2| > \ldots > |z_N| > 0 \ . \qquad (XII.60)$$

Teorema XII.11

Si los ceros del polinomio P satisfacen la condición $(XII.60)$, tenemos entonces que

$$\lim_{n\to\infty} q_n^{(k)} = z_k \ , \quad k = 1, 2, \ldots, N \ , \qquad (XII.61)$$

es decir, la $k-$ésima q columna del esquema QD converge al $k-$ésimo cero del polinomio.

De $(XII.61)$ se sigue, en virtud de $(XII.56)$, que

$$\lim_{n\to\infty} e_n^{(1)} = 0 \ , \qquad (XII.62)$$

y de esto se deduce por inducción que

$$\lim_{n\to\infty} e_n^{(k)} = 0 \ , \quad k = 1, 2, \ldots, N-1 \ . \qquad (XII.63)$$

Así pues, bajo la condición $(XII.60)$, todas las columnas e del esquema QD tienden a cero.

Lo que no está a este punto claro es como tratar con varios ceros que tienen el mismo valor absoluto; la mayor parte de las veces esta situación se presenta en conexión con ceros complejos conjugados de polinomios reales. Tales ceros pueden también obtenerse de la tabla del QD.

Teorema XII.12

En el caso más común de ceros complejos conjugados se obtienen que los límites

$$\lim_{n\to\infty} [q_{n+1}^{(k+1)} + q_n^{(k+2)}] = A_k \ , \qquad (XII.64)$$

$$\lim_{n\to\infty} q_{n+1}^{(k+1)} \cdot q_n^{(k+2)} = B_k \ , \qquad (XII.65)$$

existen, y que el polinomio $x^2 - A_k\ x + B_k$ tiene los ceros z_{k+1} y z_{k+2}.

Mencionamos para concluir el siguiente hecho, que juega el papel de control calculatorio:

Teorema XII.13
Las cantidades $e_n^{(N)}$ calculadas según $(XII.56)$ con $k = N$, son idénticamente nulas.

Como ya descrito, el esquema QD se construye procediendo de izquierda a derecha. La sucesión $\{x_n\}$ determina la primera columna q, $\{q_n^{(1)}\}$; de ella obtenemos sucesivamente las columnas $\{e_n^{(1)}\}$, $\{q_n^{(2)}\}, \ldots$, por medio de las relaciones $(XII.56)$ y $(XII.57)$. Sin embargo, este método de construcción del esquema QD no es factible en la práctica, porque sufre de **inestabilidad numérica** a causa de la grave pérdida de *dígitos significativos*. En análisis matemático, los números reales (o complejos) se conciben siempre como determinados con precisión absoluta. Numéricamente, los números reales pueden ser pensados como fracciones decimales infinitas. En el cálculo, como ya sabemos, un número real x nunca, salvo en raras excepciones, está representado exactamente, sino que tan solo aparece un número racional x^* que es su aproximación. Ya hemos estudiado las siguientes reglas sencillas, concernientes a la propagación de los errores de redondeo:

(i) Una suma o diferencia de dos números redondeados a^* y b^* tiene un error de redondeo absoluto del orde de la suma de los errores absolutos de a^* y b^*.

(i) Un producto o cociente de dos números redondeados a^* y b^* tiene un error de redondeo relativo del orde de la suma de los errores relativos de a^* y b^*.

El esquema QD ofrece una ilustración interesante de estas reglas. Para simplificar la situación, supongamos que la sucesión $\{x_n\}$ está generada sin errores de redondeo y que se tiene la situación cubierta por las relaciones $(XII.60)$ y $(XII.61)$. Si nos quedamos con t decimales, los números $q_n^{(1)}$ tendrán entonces errores de redondeo del orden de 10^{-t}.. Los elementos $e_n^{(1)}$, formados obteniendo las diferencias de los $q_n^{(1)}$, tendrán de acuerdo con la regla (i) errores absolutos de la misma magnitud. Si embargo, como por la $(XII.63)$ los $e_n^{(1)}$ tienden a cero, sus errores relativos se hacen cada vez más grandes. Por tanto, por la regla (ii), las razones $e_{n+1}^{(1)}/e_n^{(1)}$ son cada vez calculadas con menor y menor exactitud y, de nuevo por la regla (ii), el error relativo de la cantidad $q_n^{(2)}$, según se determina por $(XII.57)$, aumenta sin límite. Como $q_n^{(2)}$ tiende a un límite distinto de cero cuando $n \to \infty$, lo mismo es cierto respecto del valor absoluto de $q_n^{(2)}$. Es claro que esta inestabilidad numérica se hace más y más acentuada a medida que k aumenta.

Ejemplo 14.

La pérdida de dígitos significativos está ilustrada ya en el ejemplo 13. La última columna $e_n^{(2)}$ debe teóricamente estar constituida por ceros. El hecho de que estos elementos no sean ceros, e incluso aumenten cuando n aumenta, muestra la influenci creciente de los errores de redondeo. Si se hacen los cálculos con más decimales, esto retrasará, pero al final no impedirá, el fenómeno de inestabilidad numérica.

El hecho de que el método de generar el esquema QD es inestable, no hace, desde luego, que las propiedades enunciadas anteriormente no sean ciertas, dados que ellas se

refieren al esquema QD matemáticamente exacto, no redondeado.

El esquema QD puede generarse de una forma estable si se construye fila por fila en lugar de columna por columna. Para ver esto, resolvemos cada una de las relaciones de recurrencia ($XII.56$) y ($XII.57$) para el elemento sur del rombo en cuestión:

$$q_{n+1}^{(k)} = [e_n^{(k)} - e_{n+1}^{(k-1)}] + q_n^{(k)} \ , \qquad (XII.66)$$

$$e_{n+1}^{(k)} = \frac{q_n^{(k+1)}}{q_{n+1}^{(k)}} \ e_n^{(k)} \ . \qquad (XII.67)$$

Supongamos que conocemos una fila q y una fila e. La nueva fila q, calculada de acuerdo con ($XII.66$), tendrá entonces, por la regla (i), errores absolutos de igual magnitud. El hecho de que los errores relativos en las e son grandes (debido a la pequeñez de las e) no tiene ahora importancia. Además, los errores relativos en la nueva fila e determinada según ($XII.67$) son del mismo orden que en la anterior fila e, debido a la regla (ii). Aunque debe esperarse un aumento normal de propagación de error en el presente modo de generar el esquema, este es mucho menos serio que cuando el esquema está generado columna por columna.

Si el esquema está generado fila por fila, debe obtenerse de algún modo un primer par de filas. El siguiente algoritmo muestra cómo se consigue esto:

- Sean $a_0, a_1, \ldots, a_N$ constantes, todas distintas de cero. Hagamos:

$$q_0^{(1)} = -\frac{a_1}{a_0} \ , \quad q_{1-k}^{(k)} = 0 \ , \quad k = 2, 3, \ldots, N \ ; \qquad (XII.68)$$

$$e_{1-k}^{(k)} = \frac{a_{k+1}}{a_k} \ , \quad k = 1, 2, 3, \ldots, N-1 \ . \qquad (XII.69)$$

- Considérense los elementos así generados como las primeras dos filas de un esquema QD, y genérense las filas siguientes por medio de ($XII.66$) y ($XII.67$), usando las condiciones suplementarias

$$e_n^{(0)} = e_n^{(N)} = 0 \ , \quad n = 1, 2, \ldots \ . \qquad (XII.70)$$

Como antes, hay la posibilidad teórica de ruptura del esquema por el hecho de que un denominador sea cero.

Ejemplo 15.

Para el grado $N = 4$, las filas superiores del esquema general QD tienen el aspecto de la tabla 6.

Ejemplo 16.

Para el polinomio

$$P(x) = 128 \ x^4 - 256 \ x^3 + 160 \ x^2 - 32 \ x + 1 \ ,$$

obtenemos el esquema QD numérico (hecho por filas) de la tabla 7.

Todas las columnas e tienden a cero, luego el polinomio tiene cuatro ceros cuyos valores absolutos son distintos. Las columnas q convergen a los ceros, cuyos valores exactos son: $z_1 = 0.96194$, $z_2 = 0.69134$, $z_3 = 0.30866$ y $z_4 = 0.038060$.

Tabla 6

	$-\frac{a_1}{a_0}$		0		0		0
0		$\frac{a_2}{a_1}$		$\frac{a_3}{a_2}$		$\frac{a_4}{a_3}$	
	$q_1^{(1)}$		$q_0^{(2)}$		$q_{-1}^{(3)}$		$q_{-2}^{(4)}$
0		$e_1^{(1)}$		$e_0^{(2)}$		$e_{-1}^{(3)}$	
	$q_2^{(1)}$		$q_1^{(2)}$		$q_0^{(3)}$		$q_{-1}^{(4)}$
0		$e_2^{(1)}$		$e_1^{(2)}$		$e_0^{(3)}$	
	$q_3^{(1)}$		$q_2^{(2)}$		$q_1^{(3)}$		$q_0^{(4)}$
0		$e_3^{(1)}$		$e_2^{(2)}$		$e_1^{(3)}$	
	$q_4^{(1)}$		$q_3^{(2)}$		$q_2^{(3)}$		$q_1^{(4)}$
0		$e_4^{(1)}$		$e_3^{(2)}$		$e_2^{(3)}$	
$\vdots$	$q_5^{(1)}$	$\vdots$	$q_4^{(2)}$	$\vdots$	$q_3^{(3)}$	$\vdots$	$q_2^{(4)}$
	$\vdots$		$\vdots$		$\vdots$		$\vdots$

Tabla 7

$q_n^{(1)}$	$e_n^{(1)}$	$q_n^{(2)}$	$e_n^{(2)}$	$q_n^{(3)}$	$e_n^{(3)}$	$q_n^{(4)}$
2.000000		0.000000		0.000000		0.000000
	−0.625000		−0.200000		−0.031250	
1.375000		0.425000		0.168750		0.031250
	−0.193182		−0.079412		−0.005787	
1.181818		0.538770		0.242375		0.037037
	−0.088068		−0.035725		−0.000884	
1.093750		0.591114		0.277215		0.037921
	−0.047596		−0.016754		−0.000121	
1.046154		0.621956		0.293848		0.038042
	−0.028297		−0.007915		−0.000016	
1.017857		0.642337		0.301748		0.038058
	−0.017857		−0.003718		−0.000002	
1.000000		0.656476		0.305464		0.038060
	−0.011723		−0.001730		−0.000000	
0.988277		0.666468		0.307194		0.038060
	−0.007906		−0.000789		−0.000000	
0.980372		0.673576		0.307992		0.038060
	−0.005432		−0.000365		−0.000000	
0.974940		0.678643		0.308356		0.038060
	−0.003781		−0.000166		−0.000000	

Ejemplo 17.

Para el polinomio

$$P(x) = x^4 - 8\,x^3 + 39\,x^2 - 62\,x + 50\ ,$$

obtenemos el esquema QD numérico (hecho por filas) de la tabla 8.

Tenemos ahora que $e_n^{(2)} \to 0$, pero $e_n^{(1)}$ y $e_n^{(3)}$ no tienden a cero. Esto es una indicación de que hay dos pares de ceros complejo conjugados. Formemos entonces las cantidades

$$A_n^{(k)} = q_{n+1}^{(k+1)} + q_n^{(k+2)} \ , \qquad B_k = q_{n+1}^{(k+1)} \cdot q_n^{(k+2)}$$

para $k = 0$ y $k = 2$, obteniendo los valores de la tabla 9. Los límites son 6, 25, 2 y 2 respectivamente, indicando que los polinomios

$$x^2 - 6\ x + 25 \quad \text{y} \quad x^2 - 2\ x + 2$$

son factores cuadráticos del polinomio dado. En realidad,

$$x^4 - 8\ x^3 + 39\ x^2 - 62\ x + 50 = (x^2 - 6\ x + 25)\ (x^2 - 2\ x + 2) \ .$$

Tabla 8

$q_n^{(1)}$	$e_n^{(1)}$	$q_n^{(2)}$	$e_n^{(2)}$	$q_n^{(3)}$	$e_n^{(3)}$	$q_n^{(4)}$
8.000000		0.000000		0.000000		0.000000
	−4.875000		−1.589744		−0.806452	
3.125000		3.285256		0.783292		0.806452
	−5.125000		−0.379037		−0.830296	
−2.000000		8.031220		0.332033		1.636748
	20.580000		−0.015670		−4.092923	
18.580000		−12.564451		−3.745220		5.729671
	−13.916921		−0.004671		6.261609	
4.663079		1.347799		2.521060		−0.531938
	−4.022497		−0.008737		−1.321186	
0.640581		5.361559		1.208612		0.789248
	−33.667611		−0.001970		−0.862761	
−33.027030		39.027201		0.347820		1.652009

Tabla 9

$A_n^{(0)}$	$B_n^{(0)}$	$A_n^{(2)}$	$B_n^{(2)}$
6.410256	26.282051	1.589744	0.000000
6.031220	25.097561	1.968780	1.282051
6.015549	25.128902	1.984451	1.902439
6.010878	25.042114	1.989122	1.992225
6.002141	25.001372	1.997859	1.989741
6.000171	25.000110	1.999829	1.996637

Finalmente, consideramos la complicación que surge si algunos de los coeficientes $a_0, a_1, \ldots, a_N$ es igual a cero. En este caso, el esquema QD extendido definido por el último algoritmo dado es claro que no existe, ya que algunas de las relaciones $(XII.68)$ y $(XII.69)$ dejan de tener sentido. Un remedio posible consiste en introducir una nueva variable $\tilde{x} = x - a$ y considerar el polinomio

$$\tilde{P}(\tilde{x}) = P(a + \tilde{x}) = p(a) + \frac{1}{1!}\ P'(a)\ \tilde{x} + \frac{1}{2!}\ P''(a)\ \tilde{x}^2 + \ldots + \frac{1}{N!}\ P^{(N)}(a)\ \tilde{x}^N \ .$$

Aquí a denota un parámetro adecuadamente escogido. Se puede provar que si P tiene algunos coeficientes iguales a cero, entonces todos los coeficientes de $\tilde{P}$ son distintos de cero para valores suficientemente pequeños de $a \neq 0$. Si los ceros $\tilde{z}_k$ de $\tilde{P}$ han sido calculados, los de P vienen dados por la fórmula $z_k = \tilde{z}_k + a$.

Ejemplo 18.

Sea el polinomio

$$P(x) = 81\ x^4 - 108\ x^3 + 24\ x + 20\ .$$

Aquí $a_2 = 0$, y no podemos usar las fórmulas $(XII.68)$ y $(XII.69)$. Formemos, entonces, $\tilde{P}$ con $a = 1$. El nuevo polinomio será

$$\tilde{P}(\tilde{x}) = 81\ \tilde{x}^4 + 216\ \tilde{x}^3 + 162\ \tilde{x}^2 + 24\ \tilde{x} + 17\ ,$$

y este tiene todos los coeficientes distintos de ceros.

Las primeras dos filas de su esquema QD son las siguientes:

$$\begin{array}{ccccccccc} & & -\dfrac{216}{81} & & 0 & & 0 & & 0 \\ 0 & & & \dfrac{162}{216} & & \dfrac{24}{162} & & \dfrac{17}{24} & \end{array}$$

EJERCICIOS.

1. Encontrar aproximaciones exactas en 10^{-5} de todos los ceros reales de los siguientes polinomios , usando el método de Newton y deflación. Encontrar primero los ceros reales y luego reducir a polinomios de grado menor para determinar los eventuales ceros complejos.

a) $P(x) = x^3 - 2\,x^2 - 5$; *b)* $P(x) = x^3 + 3\,x^2 - 1$;
c) $P(x) = x^3 - x - 1$; *d)* $P(x) = x^4 + 2\,x^2 - x - 3$;
e) $P(x) = x^4 + 5\,x^3 - 9\,x^2 - 85\,x - 136$;
f) $P(x) = x^4 - 2\,x^3 - 12\,x^2 + 16\,x - 40$;
g) $P(x) = x^4 + x^3 + 3\,x^2 + 2\,x + 2$;
h) $P(x) = x^5 + 11\,x^4 - 21\,x^3 - 10\,x^2 - 21\,x - 5$;
i) $P(x) = 16\,x^4 + 88\,x^3 + 159\,x^2 + 76\,x - 240$;
l) $P(x) = x^4 - 4\,x^2 - 3\,x + 5$.

2. Para los siguientes polinomio de orden cuatro, (*i*) usar el método de Horner para mostrar que z_0 es una raíz de $P(x)$, y encontrar el polinomio $Q_0(x)$ tal que $P(x) = (x - z_0)\,Q_0(x)$; (*ii*) mostrar que $Q_0(z_1) = 0$ y encontrar $Q_1(x)$ tal que $Q_0(x) = (x - z_1)\,Q_1(x)$; (*iii*) dado que $P(x) = (x - z_0)\,(x - z_1)\,Q_1(x)$, usar la fórmula cuadrática para encontrar las otras dos raíces; (*iv*) empezar con p_0 y usar el método de Newton para encontrar p_1 y p_2.

a) $P(x) = x^4 + x^3 - 13\,x^2 - x + 12,\ z_0 = 3,\ z_1 = -1,\ p_0 = 1.1$;
b) $P(x) = x^4 + x^3 - 21\,x^2 - x + 20,\ z_0 = 4,\ z_1 = -1,\ p_0 = 1.1$;
c) $P(x) = x^4 - 2\,x^3 - 7\,x^2 + 8\,x + 12,\ z_0 = 3,\ z_1 = 2,\ p_0 = 2.1$;
d) $P(x) = x^4 - 4\,x^3 - 3\,x^2 + 14\,x - 8,\ z_0 = -2,\ z_1 = 4,\ p_0 = 1.1$;
e) $P(x) = x^4 - 6\,x^3 + 9\,x^2 + 4\,x - 12,\ z_0 = 2,\ z_1 = -1,\ p_0 = 2.1$;
f) $P(x) = x^4 - 5\,x^3 + 6\,x^2 + 4\,x - 8,\ z_0 = -1,\ z_1 = 2,\ p_0 = 2.1$.

3. Para los polinomio del problema anterior, usar el método de Bairstow para encontrar el factor cuadrático $(x^2 - r\,x - s)$ de $P(x)$, con r_0 y s_0 dados por:

a) $r_0 = 1.5,\ \ s_0 = 2.5$; *b)* $r_0 = 2.5,\ \ s_0 = 3.5$;
c) $r_0 = 4.5,\ \ s_0 = -5.5$; *d)* $r_0 = 1.5,\ \ s_0 = 7.5$;
e) $r_0 = 0.5,\ \ s_0 = 1.5$; *f)* $r_0 = 0.5,\ \ s_0 = 1.5$.

4. Para los siguientes polinomio, encontrar por el método de Bernoulli el cero dominante con tres cifras significativas. Intentar acelerar la convergencia con el proceso Δ^2 de Aitken. Luego mejorar la aproximación con el método de Newton y usando el algoritmo de Horner.

a) $P(x) = 32\,x^3 - 48\,x^2 + 18\,x - 1$; *b)* $P(x) = x^3 - x^2 + 2$;
c) $P(x) = x^4 - 7\,x^3 + 13\,x^2 - 8\,x + 12$.

5. Genérese los esquemas QD construidos por columnas y por filas para los siguientes polinomios.

a) $P(x) = x^3 + 5\,x^2 + 9\,x - 5$; *b)* $P(x) = 32\,x^3 - 48\,x^2 + 18\,x - 1$;
c) $P(x) = 81\,x^4 - 108\,x^3 + 24\,x + 20$;
d) $P(x) = 70\,x^4 - 140\,x^3 + 90\,x^2 - 20\,x + 1$;
e) $P(x) = x^5 - 3\,x^4 - 20\,x^3 + 60\,x^2 - x - 78$.

CURSO DE MÉTODOS NUMÉRICOS

TERCERA PARTE

MÉTODOS PARA LA RESOLUCIÓN DE SISTEMAS LINEALES

CAPITULO XIII. METODOS PARA LA RESOLUCION DE SISTEMAS LINEALES: PRELIMINARES

1. SISTEMAS LINEALES DE ECUACIONES

En esta tercera parte se consideran técnicas para resolver el sistema de ecuaciones lineales:

$$\begin{aligned} E_1: &\quad a_{11}\,x_1 + a_{12}\,x_2 + \ldots + a_{1n}\,x_n = b_1\ , \\ E_2: &\quad a_{21}\,x_1 + a_{22}\,x_2 + \ldots + a_{2n}\,x_n = b_2\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\,x_1 + a_{n2}\,x_2 + \ldots + a_{nn}\,x_n = b_n\ , \end{aligned} \tag{XIII.1}$$

para $x_1, \ldots, x_n$, dadas las a_{ij} para cada $i, j = 1, 2, \ldots, n$, y las b_i, para cada $i = 1, 2, \ldots, n$.

Los procedimientos de resolución de sistemas de ecuaciones lineales se dividen fundamentalmente en dos grupos:

(1) **procedimientos exactos o técnicas directas**, que son algoritmos finitos para cálculo de las raíces de un sistema (tales como la regla de Cramer, el método de Gauss, etc.);

(2) **procedimientos iterativos**, los cuales permiten obtener las raíces de un sistema con una exactitud dada mediante procesos infinitos convergentes (éstos incluyen el método de iteración, el de Seidel, el de relajación, etc.).

Debido al inevitable redondeo, incluso los resultados de procedimientos exactos son *aproximados*, viéndose comprometida, en el caso general, la estimación del error de las raíces. En el caso de procesos iterativos ha de añadirse el error del método.

Para resolver un sistema lineal como el de (XIII.1) están permitidas tres operaciones en las ecuaciones:

(1) la ecuación E_i puede multiplicarse por cualquier constante λ diferente de cero y se puede usar la ecuación resultante en lugar de E_i. Esta operación se denotará por $(\lambda E_i) \to (E_i)$;

(2) la ecuación E_j puede multiplicarse por cualquier constante λ diferente de cero, sumarla a la ecuación E_i, y usar la ecuación resultante en lugar de E_i. Esta operación se denotará por $(E_i + \lambda E_j) \to (E_i)$;

(3) las ecuaciones E_i y E_j se pueden intercambiar. Esta operación se denotará por $(E_i) \leftrightarrow (E_j)$.

Por medio de una secuencia de las operaciones anteriores, un sistema lineal se puede transformar a un sistema lineal más fácil de resolver y teniendo el mismo conjunto de soluciones. La secuencia de operaciones se ilustrará en el ejemplo siguiente.

Ejemplo 1.

Resolver las cuatro ecuaciones:

$$\begin{aligned} E_1: &\quad x_1 + x_2 + 3\,x_4 = 4\ , \\ E_2: &\quad 2\,x_1 + x_2 - x_3 + x_4 = 1\ , \\ E_3: &\quad 3\,x_1 - x_2 - x_3 + 2\,x_4 = -3\ , \\ E_4: &\quad -x_1 + 2\,x_2 + 3\,x_3 - x_4 = 4\ , \end{aligned} \tag{XIII.2}$$

para las incógnitas x_1, x_2, x_3, x_4. Un primer paso puede ser usar la ecuación E_1 para eliminar la incógnita x_1 de E_2, E_3 y E_4 efectuando $(E_2 - 2E_1) \to (E_2)$, $(E_3 - 3E_1) \to (E_3)$, y $(E_4 + E_1) \to (E_4)$. El sistema resultante es:

$$\begin{array}{lrcrcrcrcr} E_1: & x_1 & + & x_2 & & & + & 3\,x_4 & = & 4\,, \\ E_2: & & - & x_2 & - & x_3 & - & 5\,x_4 & = & -7\,, \\ E_3: & & - & 4\,x_2 & - & x_3 & - & 7\,x_4 & = & -15\,, \\ E_4: & & & 3\,x_2 & + & 3\,x_3 & + & 2\,x_4 & = & 8\,. \end{array} \qquad (XIII.3)$$

En el nuevo sistema, se usa E_2 para eliminar x_2 de E_3 y E_4 por medio de las operaciones $(E_3 - 4E_2) \to (E_3)$ y $(E_4 + 3E_2) \to (E_4)$, resultando el sistema:

$$\begin{array}{lrcrcrcrcr} E_1: & x_1 & + & x_2 & & & + & 3\,x_4 & = & 4\,, \\ E_2: & & - & x_2 & - & x_3 & - & 5\,x_4 & = & -7\,, \\ E_3: & & & & + & 3\,x_3 & + & 13\,x_4 & = & 13\,, \\ E_4: & & & & & & - & 13\,x_4 & = & -13\,. \end{array} \qquad (XIII.4)$$

Este último sistema está ahora en **forma triangular** o reducida y puede resolverse fácilmente para encontrar las incógnitas por un proceso de **sustitución hacia atrás**. Notando que E_4 implica que $x_4 = 1$, E_3 puede resolverse para x_3:

$$x_3 = \frac{1}{3}\,(13 - 13\,x_4) = \frac{1}{3}\,(13 - 13) = 0\,.$$

Continuando, x_2 resulta ser:

$$x_2 = -(-7 + 5\,x_4 + x_3) = -(-7 + 5 + 0) = 2\,;$$

y x_1 es:

$$x_1 = 4 - 3\,x_4 - x_2 = 4 - 3 - 2 = -1\,.$$

Por lo tanto la solución a $(XIII.4)$ es $x_1 = -1$, $x_2 = 2$, $x_3 = 0$ y $x_4 = 1$. Se puede verificar fácilmente que estos valores son también solución de las ecuaciones $(XIII.2)$.

Cuando realizamos los cálculos del ejemplo 1, no necesitamos escribir las ecuaciones completas en cada paso, ya que la única variación de sistema a sistema ocurre en los coeficientes de las incógnitas y en los términos independientes de las ecuaciones. Por esta razón, un sistema lineal se reemplaza frecuentemente por una matriz, que contiene toda la información del sistema que es necesaria para determinar su solución, pero en forma compacta.

La notación para una matriz $n \times m$ será una letra mayúscula como A para la matriz y letras minúsculas con subíndices dobles como a_{ij}, para referirse a la componente en la intersección de la $i-$ésima fila y la $j-$ésima columna:

$$A = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}\,.$$

Para representar al sistema lineal $(XIII.1)$ puede usarse una matriz $n \times (n+1)$, construyendo primero

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{pmatrix}$$

y luego combinando estas matrices para formar la **matriz ampliada**

$$A_a = [A, \mathbf{b}] = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{array}\right),$$

donde se usa la barra para separar los coeficientes de las incógnitas de los términos independientes de las ecuaciones.

Ejemplo 2.

Repetiremos el ejemplo 1 en notación matricial. La matriz ampliada asociada con el sistema $(XIII.2)$ será:

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 2 & 1 & -1 & 1 & 1 \\ 3 & -1 & -1 & 2 & -3 \\ -1 & 2 & 3 & -1 & 4 \end{array}\right). \qquad (XIII.5)$$

Las operaciones asociadas con $(E_2 - 2E_1) \to (E_2)$, $(E_3 - 3E_1) \to (E_3)$, y $(E_4 + E_1) \to (E_4)$ en el sistema $(XIII.2)$ se llevan a cabo manipulando las filas respectivas de la matriz ampliada $(XIII.5)$, la cual se transforma en la matriz correspondiente al sistema $(XIII.3)$:

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & -4 & -1 & -7 & -15 \\ 0 & 3 & 3 & 2 & 8 \end{array}\right). \qquad (XIII.6)$$

Realizando las manipulaciones finales, $(E_3 - 4E_2) \to (E_3)$ y $(E_4 + 3E_2) \to (E_4)$, se obtiene la matriz ampliada correspondiente al sistema $(XIII.4)$:

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right). \qquad (XIII.7)$$

Esta matriz puede transformarse ahora en su correspondiente sistema lineal $(XIII.4)$ y así obtener las soluciones x_1, x_2, x_3 y x_4.

El procedimiento descrito en este proceso se llama **eliminación Gaussiana con sustitución hacia atrás**. En un próximo capítulo consideraremos las condiciones bajo las cuales el método puede usarse con éxito para resolver el sistema lineal.

2. ALGEBRA LINEAL E INVERSION DE UNA MATRIZ

Esta sección se refiere al álgebra asociada con las matrices y la manera en que éstas pueden usarse para resolver problemas que involucran sistemas lineales.

Definición. Se dice que dos matrices A y B son **iguales** si son del mismo tamaño, digamos $m \times n$ y si $a_{ij} = b_{ij}$ para cada $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, n$.

Definición. Si A y B son matrices ambas $m \times n$, entonces la **suma** de A y B, denotada por $A + B$, es la matriz $m \times n$ cuyos elementos son $a_{ij} + b_{ij}$, para cada $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, n$.

Definición. Si A es una matriz $m \times n$ y λ es un número real, entonces el **producto escalar** de λ y A, denotado λA, es la matriz $m \times n$ cuyos elementos son λa_{ij}, para cada $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, n$.

Denotando la matriz que tiene todos sus elementos iguales a cero simplemente como O y como $-A$ la matriz cuyos elementos son $-a_{ij}$, podemos enumerar las siguientes propiedades generales de la adición y de la multiplicación escalar matricial. Estas propiedades son suficientes para clasificar el conjunto de todas las matrices $m \times n$ con elementos reales como un **espacio vectorial** sobre el campo de los números reales.

Teorema XIII.1

Sean A, B y C matrices $m \times n$ y λ y μ números reales. Se satisfacen las siguientes propiedades de la adición y multiplicación escalar:

a) $A + B = B + A$,
b) $(A + B) + C = A + (B + C)$,
c) $A + O = O + A = A$,
d) $A + (-A) = -A + A = O$,
e) $\lambda(A + B) = \lambda A + \lambda B$,
f) $(\lambda + \mu)\ A = \lambda A + \mu A$,
g) $\lambda(\mu A) = (\lambda\mu)A$,
h) $1A = A$.

Definición. Sean A una matriz $m \times n$ y B una matriz $n \times p$. El **producto matricial** de A y B, denotado por $A\ B$, es una matriz $m \times p$, cuyos elementos c_{ij} están dados por

$$c_{ij} = \sum_{k=1}^{n} a_{ik}\ b_{kj} = a_{i1}\ b_{1j} + a_{i2}\ b_{2j} + \ldots + a_{in}\ b_{nj}$$

para cada $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, p$.

Definición. Una matriz **diagonal** de orden n es una matriz $D = (d_{ij})$, $n \times n$, con la propiedad de que $d_{ij} = 0$ siempre que $i \neq j$. La **matriz identidad de orden n**, $I_n = (\delta_{ij})$, es la matriz diagonal con elementos

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} ;$$

Normalmente esta matriz se escribe simplemente como I.

Es bien conocido que la matriz identidad conmuta con una matriz A de orden n, es decir, el orden de la multiplicación no importa. Por otra parte, la propiedad conmutativa, $A\ B = B\ A$, no es generalmente cierta para la multiplicación matricial. Algunas de las propiedades relacionadas con la multiplicación de matrices, que sí se satisfacen, se presentan en el Teorema siguiente:

Teorema XIII.2

Sean A una matriz $n \times m$, B una matriz $m \times k$, C una matriz $k \times p$, D una matriz $m \times k$ y λ un número real. Se satisfacen las siguientes propiedades:

a) $A(B\ C) = (A\ B)C$,

b) $A(B + D) = A\ B + A\ D$,

c) $I_m B = B$, $B\ I_k = B$,

d) $\lambda(A\ B) = (\lambda A)B = A(\lambda B)$.

Un concepto fundamental del álgebra lineal que es muy útil para determinar la existencia y unicidad de soluciones de sistemas lineales es el **determinante** de una matriz $n \times n$. El único enfoque que se dará aquí para calcular el determinante será la definición recursiva. El determinante de una matriz A de denotará por "det A". Una **submatriz** de una matriz A es una matriz "extraída" de A suprimiendo algunas filas y/o columnas de A.

Definición.

a) Si $A = (a)$ es una matriz 1×1, entonces det $A = a$.

b) El **menor** M_{ij} es el determinante de la submatriz $(n-1) \times (n-1)$ de una matriz $n \times n$ de A obtenido suprimiendo la $i-$ésima fila y la $j-$ésima columna.

c) El **cofactor** A_{ij} asociado con M_{ij} se define como $A_{ij} = (-1)^{i+j}\ M_{ij}$.

d) El **determinante** de una matriz A, $n \times n$, donde $n > 1$ está dado ya sea por

$$\det A = \sum_{j=1}^{n} a_{ij} A_{ij} \qquad \text{para cualquier } i = 1, 2, \ldots, n \ , \tag{XIII.8}$$

o

$$\det A = \sum_{i=1}^{n} a_{ij} A_{ij} \qquad \text{para cualquier } j = 1, 2, \ldots, n \ . \tag{XIII.9}$$

Usando inducción matemática, se puede demostrar que, si $n > 1$, el uso de las definiciones dadas para calcular el determinante de una matriz, en general $n \times n$, requiere $n!$ multiplicaciones / divisiones y de $(n! - 1)$ sumas / restas. Incluso para valores relativamente pequeños de n, el número de cálculos puede llegar a ser inmanejable.

Teorema XIII.3

Sea A una matriz $n \times n$:

a) Si cualquier fila o columna de A tiene sólo componentes cero, entonces det $A = 0$.

b) Si $\tilde{A}$ se obtiene de A por medio de la operación $(E_i) \leftrightarrow (E_j)$, con $i \neq j$, entonces det $\tilde{A} = -$det A.

c) Si A tiene dos filas iguales, entonces $\det A = 0$.

d) Si $\tilde{A}$ se obtiene de A por medio de la operación $\lambda(E_i) \to (E_i)$, entonces $\det \tilde{A} = \lambda \det A$.

e) Si $\tilde{A}$ se obtiene de A por medio de la operación $(E_i + \lambda E_j) \to (E_j)$, con $i \neq j$, entonces $\det \tilde{A} = \det A$.

f) Si B es también una matriz $n \times n$ entonces $\det A\ B = \det A \det B$.

Definición. Se dice que una matriz A $n \times n$ es **no singular** si existe una matriz A^{-1}, $n \times n$, tal que $A\ A^{-1} = A^{-1}\ A = I$. La matriz A^{-1} se llama la **inversa** de A. Una matriz que no tiene inversa se llama **singular**.

Para encontrar un método para calcular A^{-1}, suponiendo su existencia, consideramos nuevamente la multiplicación matricial. Sea B_j la $j-$ésima columna de la matriz B $n \times n$. Realizaremos el producto

$$A\ B_j = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} b_{1j} \\ b_{2j} \\ \dots \\ b_{nj} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} a_{1k}\ b_{kj} \\ \sum_{k=1}^{n} a_{2k}\ b_{kj} \\ \dots \\ \sum_{k=1}^{n} a_{nk}\ b_{kj} \end{pmatrix} .$$

Si $A\ B = C$, entonces la $j-$ésima columna de C está dada por

$$C_j = \begin{pmatrix} c_{1j} \\ c_{2j} \\ \dots \\ c_{nj} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} a_{1k}\ b_{kj} \\ \sum_{k=1}^{n} a_{2k}\ b_{kj} \\ \dots \\ \sum_{k=1}^{n} a_{nk}\ b_{kj} \end{pmatrix} .$$

Por lo tanto, la $j-$ésima columna del producto $A\ B$ es el producto de A con la $j-$ésima columna de B. Supongamos que A^{-1} existe y que $A^{-1} = B = (b_{ij})$; entonces $A\ B = I$ y

$$A\ B_j = \begin{pmatrix} 0 \\ \dots \\ 0 \\ 1 \\ 0 \\ \dots \\ 0 \end{pmatrix} ,$$

donde el valor 1 aparece en la $j-$ésima fila. Para encontrar B debemos resolver n sistemas lineales en los cuales la $j-$ésima columna de la matriz inversa es la solución del sistema lineal con término independiente igual a la $j-$ésima columna de I.

Otra manera de calcular A^{-1} es relacionarla con el determinante de la matriz y con su adjunto.

Definición. Se define el **adjunto** de una matriz A, $n \times n$, como la matriz

$$A^{+} = \begin{pmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \dots & \dots & \dots & \dots \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{pmatrix} ,$$

donde A_{ij} son los cofactores (menores con signo) de los elementos correspondientes a_{ij} $(i, j = 1, 2, \dots, n)$. [Nótese que los adjuntos de los elementos de las filas de una matriz caen en las columnas correspondientes al adjunto, es decir, se verifica la operación de transposición].

Para encontrar la inversa de la matriz A, se dividen todos los elementos de la matriz adjunta A^{+} por el valor del determinante de A:

$$A^{-1} = \frac{1}{\det A} \, A^{+} .$$

Presentaremos ahora el resultado clave que relaciona a la no-singularidad, la eliminación Gaussiana, los sistemas lineales y los determinantes.

Teorema XIII.4

Para una matriz A $n \times n$ las siguientes afirmaciones son equivalentes:

a) La ecuación $A\ \mathbf{x} = \mathbf{0}$ tiene la única solución $\mathbf{x} = \mathbf{0}$.
b) El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ tiene una solución única para cualquier vector columna $\mathbf{b}$ $n-$dimensional.
c) La matriz A es no singular, es decir, A^{-1} existe.
d) $\det A \neq 0$.
e) El algoritmo de la eliminación Gaussiana con intercambio de filas (que veremos más adelante) se puede aplicar al sistema lineal $A\ \mathbf{x} = \mathbf{b}$ para cualquier vector columna $\mathbf{b}$ $n-$dimensional.

Por medio de la definición de la multiplicación de matrices se puede discutir la relación entre los sistemas lineales y el álgebra lineal. El sistema lineal

$$\begin{aligned} E_1: &\quad a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = b_1 \ , \\ E_2: &\quad a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = b_2 \ , \\ &\quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \\ E_n: &\quad a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = b_n \ , \end{aligned} \qquad (XIII.1)$$

puede verse como la ecuación matricial

$$A\ \mathbf{x} = \mathbf{b} \ , \qquad (XIII.10)$$

donde

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} , \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \qquad \text{y} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{pmatrix} .$$

El concepto de la **matriz inversa** de una matriz está también relacionado con los sistemas lineales, dado que multiplicando a la izquierda ambos miembros de $(XIII.10)$ por la matriz inversa A^{-1}, obtenemos

$$A^{-1}\ A\ \mathbf{x} = A^{-1}\ \mathbf{b}\ ,$$

o

$$\mathbf{x} = A^{-1}\ \mathbf{b}\ , \qquad (XIII.11)$$

que nos da la solución única del sistema $(XIII.1)$. Ese método es conocido como **regla de Cramer**. Dado que

$$A^{-1} = \frac{A^{+}}{\det\ A}\ ,$$

donde A^{+} es el adjunto de A, se tiene que

$$\mathbf{x} = \frac{A^{+}}{\det\ A}\ \mathbf{b}\ ,$$

o

$$\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} = \frac{1}{\det\ A}\ \begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \dots \\ \Delta_n \end{pmatrix}\ , \qquad (XIII.12)$$

donde

$$\Delta_i = \sum_{j=1}^{n} A_{ji}\ b_j = \det \begin{pmatrix} a_{11} & \dots & a_{1,i-1} & b_1 & a_{1,i+1} & \dots & a_{1n} \\ a_{21} & \dots & a_{2,i-1} & b_2 & a_{2,i+1} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{n,i-1} & b_n & a_{n,i+1} & \dots & a_{nn} \end{pmatrix}$$

son los determinantes obtenidos del determinante $\det\ A$ sustituyendo su $i-$ésima columna por la columna de términos constantes del sistema $(XIII.1)$. De la ecuación $(XIII.12)$ tenemos las **fórmulas de Cramer**:

$$x_1 = \frac{\Delta_1}{\det\ A}\ , \quad x_2 = \frac{\Delta_2}{\det\ A}\ , \quad \dots, \quad x_n = \frac{\Delta_n}{\det\ A}\ . \qquad (XIII.13)$$

De este modo, si el determinante del sistema $(XIII.1)$ es distinto de cero, entonces el sistema tiene una solución única $\mathbf{x}$ definida por la fórmula matricial $(XIII.11)$ o por las fórmulas escalares $(XIII.13)$ equivalentes.

Ejemplo 3.

La utilidad de la matriz inversa se puede mostrar considerando el sistema lineal

$$\begin{array}{rcrcrcl} x_1 & + & 2\ x_2 & - & x_3 & = & 2\ , \\ 2\ x_1 & + & x_2 & & & = & 3\ , \\ -\ x_1 & + & x_2 & + & 2\ x_3 & = & 4\ . \end{array}$$

Primero, convertimos el sistema a la ecuación matricial

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} .$$

Ahora, buscamos la inversa de la matriz A. En primer lugar calculamos el determinante de la matriz A:

$$\det A = (2 + 0 - 2) - (1 + 0 + 8) = -9 \neq 0 ,$$

es decir, la matriz A no es singular.

Luego, usamos la definición de la matriz adjunta de A:

$$A^+ = \begin{pmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{pmatrix} ,$$

donde A_{ij} son los cofactores (menores con signo) de los elementos correspondientes a_{ij} $(i, j = 1, 2, 3)$ [es decir, $A_{ij} = (-1)^{i+j}\ M_{ij}$, donde el menor M_{ij} es el determinante de la submatriz de 2×2 de la matriz A obtenido suprimiendo la $i-$ésima fila y la $j-$ésima columna]. Entonces:

$$A_{11} = (-1)^{1+1}\ M_{11} = (-1)^2 \det \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix} = 2 ,$$

$$A_{12} = (-1)^{1+2}\ M_{12} = (-1)^3 \det \begin{pmatrix} 2 & 0 \\ -1 & 2 \end{pmatrix} = -4 ,$$

$$A_{13} = (-1)^{1+3}\ M_{13} = (-1)^4 \det \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix} = 3 ,$$

$$A_{21} = (-1)^{2+1}\ M_{21} = (-1)^3 \det \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix} = -5 ,$$

$$A_{22} = (-1)^{2+2}\ M_{22} = (-1)^4 \det \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} = 1 ,$$

$$A_{23} = (-1)^{2+3}\ M_{23} = (-1)^5 \det \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix} = -3 ,$$

$$A_{31} = (-1)^{3+1}\ M_{31} = (-1)^4 \det \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix} = 1 ,$$

$$A_{32} = (-1)^{3+2}\ M_{32} = (-1)^5 \det \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix} = -2 ,$$

$$A_{33} = (-1)^{3+3}\ M_{33} = (-1)^6 \det \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = -3 ;$$

y la matriz inversa será:

$$A^{-1} = \frac{1}{\det A}\ A^+ = \frac{1}{-9} \begin{pmatrix} 2 & -5 & 1 \\ -4 & 1 & -2 \\ 3 & -3 & -3 \end{pmatrix} .$$

Finalmente, multiplicamos ambos lados de la ecuación matricial por la inversa:

$$\frac{1}{9}\begin{pmatrix}-2 & 5 & -1\\ 4 & -1 & 2\\ -3 & 3 & 3\end{pmatrix}\begin{pmatrix}1 & 2 & -1\\ 2 & 1 & 0\\ -1 & 1 & 2\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix}=\frac{1}{9}\begin{pmatrix}-2 & 5 & -1\\ 4 & -1 & 2\\ -3 & 3 & 3\end{pmatrix}\begin{pmatrix}2\\ 3\\ 4\end{pmatrix}\ ;$$

con lo que

$$I_3\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix}=\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix}=\begin{pmatrix}\frac{7}{9}\\ \frac{13}{9}\\ \frac{15}{9}\end{pmatrix}\ ,$$

lo cual da la solución

$$x_1=\frac{7}{9}\ ,\quad x_2=\frac{13}{9}\quad\text{y}\quad x_3=\frac{15}{9}\ .$$

Ejemplo 4.

Resuélvase el sistema de ecuaciones lineales

$$\begin{array}{rcrcrcrcr} 2\,x_1 & + & x_2 & - & 5\,x_3 & + & x_4 & = & 8\ ,\\ x_1 & - & 3\,x_2 & & & - & 6\,x_4 & = & 9\ ,\\ & & 2\,x_2 & - & x_3 & + & 2\,x_4 & = & -5\ ,\\ x_1 & + & 4\,x_2 & - & 7\,x_3 & + & 6\,x_4 & = & 0\ . \end{array}$$

El determinante de este sistema es:

$$\det\begin{pmatrix}2 & 1 & -5 & 1\\ 1 & -3 & 0 & -6\\ 0 & 2 & -1 & 2\\ 1 & 4 & -7 & 6\end{pmatrix}=27\neq 0\ .$$

Calculando los determinantes suplementarios, tenemos:

$$\Delta_1=\det\begin{pmatrix}8 & 1 & -5 & 1\\ 9 & -3 & 0 & -6\\ -5 & 2 & -1 & 2\\ 0 & 4 & -7 & 6\end{pmatrix}=81\ ,$$

$$\Delta_2=\det\begin{pmatrix}2 & 8 & -5 & 1\\ 1 & 9 & 0 & -6\\ 0 & -5 & -1 & 2\\ 1 & 0 & -7 & 6\end{pmatrix}=-108\ ,$$

$$\Delta_3=\det\begin{pmatrix}2 & 1 & 8 & 1\\ 1 & -3 & 9 & -6\\ 0 & 2 & -5 & 2\\ 1 & 4 & 0 & 6\end{pmatrix}=-27\ ,$$

$$\Delta_4=\det\begin{pmatrix}2 & 1 & -5 & 8\\ 1 & -3 & 0 & 9\\ 0 & 2 & -1 & -5\\ 1 & 4 & -7 & 0\end{pmatrix}=27\ ,$$

por consiguiente

$$\begin{aligned} x_1 &= \frac{\Delta_1}{\det A} = \frac{81}{27} = 3 \ , \\ x_2 &= \frac{\Delta_2}{\det A} = -\frac{108}{27} = -4 \ , \\ x_3 &= \frac{\Delta_3}{\det A} = -\frac{27}{27} = -1 \ , \\ x_4 &= \frac{\Delta_4}{\det A} = \frac{27}{27} = 1 \ . \end{aligned}$$

De este modo, la solución de un sistema lineal como $(XIII.1)$ con n incógnitas se reduce a evaluar al $(n+1)-$ésimo determinante de orden n. Si n es grande, el cálculo de los determinantes es laborioso. Por esta razón, se han elaborado técnicas directas para hallar las raíces de un sistema lineal de ecuaciones.

3. TIPOS ESPECIALES DE MATRICES

Presentamos ahora material adicional sobre matrices. El primer tipo de matrices que consideraremos es el producido cuando se aplica eliminación Gaussiana a un sistema lineal.

Definición. Una matriz **triangular superior** U $n \times n$ tiene para cada j, los elementos $u_{ij} = 0$ para cada $i = j+1, j+2, \ldots, n$; una matriz **triangular inferior** L $n \times n$ tiene para cada j, los elementos $l_{ij} = 0$ para cada $i = 1, 2, \ldots, j-1$. (Una matriz **diagonal** es a la vez triangular superior e inferior). Es decir,

$$L = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix} \ , \quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} \ ,$$

y

$$D = \begin{pmatrix} d_{11} & 0 & 0 & 0 \\ 0 & d_{22} & 0 & 0 \\ 0 & 0 & d_{33} & 0 \\ 0 & 0 & 0 & d_{44} \end{pmatrix} \ .$$

El cálculo del determinante de una matriz arbitraria puede requerir un gran número de manipulaciones. Sin embargo, una matriz en forma triangular tiene un determinante fácil de calcular.

Teorema XIII.5

Si $A = (a_{ij})$ es una matriz $n \times n$ triangular superior (o triangular inferior o diagonal), entonces $\det A = \prod_{i=1}^{n} a_{ii}$.

Ejemplo 5.

Reconsidereremos los ejemplos 1 y 2, en los que el sistema lineal

$$\begin{array}{lrcrcrcrcr} E_1: & x_1 & + & x_2 & & & + & 3\,x_4 & = & 4 \ , \\ E_2: & 2\,x_1 & + & x_2 & - & x_3 & + & x_4 & = & 1 \ , \\ E_3: & 3\,x_1 & - & x_2 & - & x_3 & + & 2\,x_4 & = & -3 \ , \\ E_4: & -\,x_1 & + & 2\,x_2 & + & 3\,x_3 & - & x_4 & = & 4 \ , \end{array}$$

fue reducido al sistema equivalente

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right) .$$

Sea U la matriz triangular superior de 4×4

$$U = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix} ,$$

la cual es el resultado de efectuar la eliminación Gaussiana a A. Para $i = 1, 2, 3$, definimos m_{ji} para cada $j = i + 1, i + 2, \ldots, 4$ como el número usado en el paso de eliminación $(E_j - m_{ji}E_i) \to E_j$; es decir $m_{21} = 2$, $m_{31} = 3$, $m_{41} = -1$, $m_{32} = 4$, $m_{42} = -3$ y $m_{43} = 0$. Si L se define como la matriz triangular inferior de 4×4 con elementos l_{ji} dados por

$$l_{ji} = \begin{cases} 0 , & \text{cuando } i = 1, 2, \ldots, j - 1 , \\ 1 , & \text{cuando } i = j , \\ m_{ji} , & \text{cuando } i = j + 1, j + 2, \ldots, n , \end{cases}$$

entonces

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} ,$$

y es fácil verificar que

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix} = A .$$

Los resultados de este ejemplo son ciertos en general y están dados en el Teorema siguiente.

Teorema XIII.6

Si el procedimiento de eliminación Gaussiana puede aplicarse al sistema $A\mathbf{x} = \mathbf{b}$ sin intercambio de fila, entonces la matriz A puede factorizarse como el producto de una matriz triangular inferior L con una matriz triangular superior U:

$$A = L\, U ,$$

donde $U = (u_{ij})$ y $L = (l_{ij})$ están definidas para cada j por:

$$u_{ij} = \begin{cases} a_{ij}^{(i)} , & \text{cuando } i = 1, 2, \ldots, j , \\ 0 , & \text{cuando } i = j+1, j+2, \ldots, n , \end{cases}$$

y

$$l_{ij} = \begin{cases} 0 , & \text{cuando } i = 1, 2, \ldots, j-1 , \\ 1 , & \text{cuando } i = j , \\ m_{ij} , & \text{cuando } i = j+1, j+2, \ldots, n , \end{cases}$$

donde $a_{ij}^{(i)}$ es el elemento i, j de la matriz final obtenida por el método de eliminación Gaussiana y m_{ij} es el multiplicador.

Si se tienen que efectuar intercambios de filas para que el procedimiento funcione, entonces A se puede factorizar como $L\,U$, donde U es la misma que en el Teorema XIII.6, pero en general, L no será triangular inferior.

El problema de calcular el determinante de una matriz se puede simplificar reduciendo primero la matriz a forma triangular y después usando el Teorema XIII.5 para encontrar el determinante de una matriz triangular.

Ejemplo 6.

Sea

$$A = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ -1 & 2 & 3 & -1 \\ 3 & -1 & -1 & 2 \end{pmatrix} .$$

Con los tres tipos de operaciones permitidas, la matriz A será deducida a una matriz triangular superior. Primero efectuamos $(E_2 - 2E_1) \to (E_2)$, $(E_3 + E_1) \to (E_3)$, $(E_4 - 3E_1) \to (E_4)$ para obtener:

$$A^{(1)} = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 3 & 3 & 2 \\ 0 & -4 & -1 & -7 \end{pmatrix} .$$

Por el Teorema XIII.3(e) [si $\tilde{A}$ se obtiene de A por medio de la operación $(E_i + \lambda E_j) \to (E_j)$, con $i \neq j$, entonces $\det \tilde{A} = \det A$], $\det A = \det A^{(1)}$. Formando $A^{(2)}$ de $A^{(1)}$ con las operaciones $(E_3 + 3E_2) \to (E_3)$, $(E_4 - 4E_2) \to (E_4)$

$$A^{(2)} = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 0 & -13 \\ 0 & 0 & 3 & 13 \end{pmatrix} ,$$

y otra vez, $\det A = \det A^{(1)} = \det A^{(2)}$. Si $A^{(3)}$ se forma de $A^{(2)}$ con $(E_3) \leftrightarrow (E_4)$, entonces:

$$A^{(3)} = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix} .$$

Por el Teorema XIII.5, $\det A^{(3)} = (1)(-1)(3)(-13) = 39$ y como $A^{(3)}$ se formó de $A^{(2)}$ mediante intercambio de filas, por el Teorema XIII.3(b) [si $\tilde{A}$ se obtiene de A por medio de la operación $(E_i) \leftrightarrow (E_j)$, con $i \neq j$, entonces $\det \tilde{A} = -\det A$],

$$\det A = \det A^{(2)} = -\det A^{(3)} = -39 \ .$$

Definición. La **traspuesta** de una matriz A $m \times n$, denotada por A^t, es una matriz $n \times m$ cuyos elementos son $(A^t)_{ij} = (A)_{ji}$. Una matriz cuya traspuesta es ella misma se llama **simétrica**.

Teorema XIII.7

Las siguientes operaciones que involucran a la traspuesta de una matriz se satisfacen siempre que la operación sea posible:

1. $(A^t)^t = A$,
2. $(A + B)^t = A^t + B^t$,
3. $(A\ B)^t = B^t\ A^t$,
4. si A^{-1} existe, $(A^{-1})^t = (A^t)^{-1}$,
5. $\det A^t = \det A$.

Definición. Una matriz $n \times n$ se llama una **matriz banda** si existen enteros p y q, $1 < p, q < n$, con la propiedad de que $a_{ij} = 0$ siempre que $i + p \leq j$ ó $j + q \leq i$. El **ancho de banda** para una matriz de este tipo se define como $w = p + q - 1$.

La definición de matriz de banda fuerza a estas matrices a concentrar todos sus elementos no cero alrededor de la diagonal. Dos casos especiales de matrices de banda que ocurren frecuentemente en la práctica son $p = q = 2$ y $p = q = 4$. Las matrices con ancho de banda 3 (que se presenta cuando $p = q = 2$) se llaman generalmente **tridiagonales** ya que tienen la forma

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & a_{32} & a_{33} & a_{34} & 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & a_{i,i-1} & a_{ii} & a_{i,i+1} & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & a_{n-2,n-1} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & a_{n,n-1} & a_{nn} \end{pmatrix}$$

Definición. Se dice que la matriz A de orden n es **estrictamente dominante diagonalmente** en el caso de que satisfaga

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \ ,$$

para cada $i = 1, 2, \dots, n$.

Teorema XIII.8

Si A es una matriz $n \times n$ estrictamente dominante diagonalmente, entonces A es no singular. Además, se puede efectuar eliminación Gaussiana en cualquier sistema lineal de la forma $A\ \mathbf{x} = \mathbf{b}$ para obtener su solución única sin intercambios de filas o columnas, y los cálculos son estables con respecto al crecimiento de los errores de redondeo.

La última clase especial de matrices que se discutirá en esta sección se llama positiva definida.

Definición. Una matriz simétrica A $n \times n$ se llama **positiva definida** si $\mathbf{x}^t\ A\ \mathbf{x} > 0$ para todo vector columna $n-$dimensional $\mathbf{x} \neq \mathbf{0}$,

$$\mathbf{x}^t\ A\ \mathbf{x} = (x_1, x_2, \ldots, x_n) \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix} =$$

$$= (x_1, x_2, \ldots, x_n) \begin{pmatrix} \sum_{j=1}^{n} a_{1j}\ x_j \\ \sum_{j=1}^{n} a_{2j}\ x_j \\ \ldots \\ \sum_{j=1}^{n} a_{nj}\ x_j \end{pmatrix} = (\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}\ x_i\ x_j)\ .$$

Teorema XIII.9

Si A es una matriz $n \times n$ positiva definida, entonces A es no singular. Además, se puede efectuar eliminación Gaussiana en cualquier sistema lineal de la forma $A\ \mathbf{x} = \mathbf{b}$ para obtener su solución única sin intercambios de filas o columnas, y los cálculos son estables con respecto al crecimiento de los errores de redondeo.

4. NORMAS DE VECTORES Y MATRICES

Sea $\mathcal{R}^n$ el conjunto de todos los vectores columna con componentes reales. Para definir una distancia en $\mathcal{R}^n$, usaremos la idea de la **norma** de un vector.

Definición. Una **norma vectorial** en $\mathcal{R}^n$ es una función $||\ \cdot\ ||$, de $\mathcal{R}^n$ en $\mathcal{R}$ con las siguientes propiedades:

- $||\mathbf{x}|| \geq 0$ para todo $\mathbf{x} \in \mathcal{R}^n$;
- $||\mathbf{x}|| = 0$ si y sólo si $\mathbf{x} = (0, 0, \ldots, 0)^t \equiv \mathbf{0}$;
- $||\alpha\ \mathbf{x}|| = |\alpha|\ ||\mathbf{x}||$ para todo $\alpha \in \mathcal{R}$ y $\mathbf{x} \in \mathcal{R}^n$;
- $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ para todo $\mathbf{x}, \mathbf{y} \in \mathcal{R}^n$.

Para nuestros propósitos sólo necesitaremos tres normas específicas en $\mathcal{R}^n$.

Definición. Las normas l_1, l_2 y l_∞ para el vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ se definen como

$$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|\ , \qquad ||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \quad \text{y} \quad ||\mathbf{x}||_\infty = \max_{1 \leq i \leq n} |x_i|\ .$$

La norma l_2 se denomina frecuentemente **norma Euclideana** del vector $\mathbf{x}$ ya que representa la noción usual de distancia al origen en el caso en el que $\mathbf{x}$ esté en $\mathcal{R}$, $\mathcal{R}^2$ o $\mathcal{R}^3$.

Ya que la norma de un vector da una medida de la distancia entre el vector y el origen, la distancia entre dos vectores se puede definir como la norma de la diferencia de los dos vectores.

Definición. Si $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ e $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$ son vectores en $\mathcal{R}^n$, las distancias l_1, l_2 y l_∞ entre $\mathbf{x}$ e $\mathbf{y}$ se definen como:

$$||\mathbf{x} - \mathbf{y}||_1 = \sum_{i=1}^{n} |x_i - y_i| \ ,$$

$$||\mathbf{x} - \mathbf{y}||_2 = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \quad \text{y} \quad ||\mathbf{x} - \mathbf{y}||_\infty = \max_{1 \le i \le n} |x_i - y_i| \ .$$

Ejemplo 7.

El sistema lineal

$$\begin{array}{lrcrcrcr} E_1: & 3.3330\, x_1 & + & 15920\, x_2 & + & 10.333\, x_3 & = & 15913\ , \\ E_2: & 2.2220\, x_1 & + & 16.710\, x_2 & - & 9.6120\, x_3 & = & 28.544\ , \\ E_3: & 1.5611\, x_1 & + & 5.1791\, x_2 & - & 1.6852\, x_3 & = & 8.4254\ , \end{array}$$

tiene por solución a $\mathbf{x} = (x_1, x_2, x_3)^t = (1.0000, 1.0000, -1.0000)^t$. Si se efectúa eliminación Gaussiana con aritmética de cinco dígitos usando pivoteo máximo de columna, la solución obtenida es $\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)^t = (1.2001, 0.99991, -0.92538)^t$ (ver ejemplo 3 del Cap. XVIII). Las medidas de $\mathbf{x} - \tilde{\mathbf{x}}$ están dadas por

$$\begin{aligned} ||\mathbf{x} - \tilde{\mathbf{x}}||_1 =& |1.0000 - 1.2001| + |1.0000 - 0.99991| + |1.0000 - 0.92538| = \\ =& 0.2001 + 0.00009 + 0.07462 = 0.27481 \ , \end{aligned}$$

$$\begin{aligned} ||\mathbf{x} - \tilde{\mathbf{x}}||_2 =& [|1.0000 - 1.2001|^2 + |1.0000 - 0.99991|^2 + |1.0000 - 0.92538|^2]^{1/2} = \\ =& \sqrt{(0.2001)^2 + (0.00009)^2 + (0.07462)^2} = 0.21356 \ , \end{aligned}$$

y

$$\begin{aligned} ||\mathbf{x} - \tilde{\mathbf{x}}||_\infty =& \max\{|1.0000 - 1.2001|, |1.0000 - 0.99991|, |1.0000 - 0.92538|\} = \\ =& \max\{0.2001, 0.00009, 0.07462\} = 0.2001 \ . \end{aligned}$$

Aún cuando las componentes $\tilde{x}_2$ y $\tilde{x}_3$ son buenas aproximaciones a x_2 y x_3 la componente $\tilde{x}_1$ es una aproximación pobre de x_1 y $|x_1 - \tilde{x}_1|$ domina a las normas.

El concepto de distancia en $\mathcal{R}^n$ puede usarse también para definir el límite de una sucesión de vectores en este espacio.

Definición. Se dice que una sucesión $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ de vectores en $\mathcal{R}^n$ **converge** a $\mathbf{x}$ con respecto a la norma $||\ \cdot\ ||$ si, dado cualquier $\varepsilon > 0$, existe un entero $N(\varepsilon)$ tal que

$$||\mathbf{x}^{(k)} - \mathbf{x}|| < \varepsilon \qquad \text{para toda} \quad k \ge N(\varepsilon) \ .$$

Teorema XIII.10

La sucesión de vectores $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ converge a $\mathbf{x}$ en $\mathcal{R}^n$ con respecto a $||\ \cdot\ ||_\infty$ si y sólo si $\lim_{k\to\infty} x_i^{(k)} = x_i$ para cada $i = 1, 2, \ldots, n$.

Teorema XIII.11

Para cada $\mathbf{x} \in \mathcal{R}^n$,

$$||\mathbf{x}||_\infty \leq ||\mathbf{x}||_2 \leq \sqrt{n}\ ||\mathbf{x}||_\infty\ .$$

Demostración: sea x_j una coordenada de $\mathbf{x}$ tal que $||\mathbf{x}||_\infty = \max_{1\leq i\leq n} |x_i| = |x_j|$. Entonces

$$||\mathbf{x}||_\infty^2 = |x_j|^2 = x_j^2 \leq \sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} x_j^2 = nx_j^2 = n||\mathbf{x}||_\infty^2\ .$$

Por lo tanto

$$||\mathbf{x}||_\infty \leq \Big[\sum_{i=1}^{n} x_i^2\Big]^{1/2} = ||\mathbf{x}||_2 \leq \sqrt{n}||\mathbf{x}||_\infty\ .$$

c.q.d.

Se puede demostrar que todas las normas en $\mathcal{R}^n$ son equivalentes con respecto a la convergencia; es decir, si $||\ \cdot\ ||$ y $||\ \cdot\ ||'$ son dos normas cualesquiera en $\mathcal{R}^n$ y $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ tiene el límite $\mathbf{x}$ con respecto a $||\ \cdot\ ||$, entonces $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ tiene el límite $\mathbf{x}$ con respecto a $||\ \cdot\ ||'$

Es necesario también tener un método para medir distancias entre dos matrices $n\times n$, lo cual nuevamente requiere el uso del concepto de norma.

Definición. Una **norma matricial** en el conjunto de todas las matrices reales $n \times n$ es una función de valores reales $||\ \cdot\ ||$, definida en este conjunto que satisface, para todas las matrices A y B $n \times n$ y todo número real α:

- $||A|| \geq 0$;
- $||A|| = 0$ si y sólo si $A = O$;
- $||\alpha\ A|| = |\alpha|\ ||A||$;
- $||A + B|| \leq ||A|| + ||B||$;
- $||A \cdot B|| \leq ||A|| \cdot ||B||$.

Una **distancia entre las matrices** A y B $n \times n$ se puede definir de la manera usual como $||A - B||$. Aún cuando las normas de las matrices pueden obtenerse de varias maneras, las únicas normas que consideraremos son aquellas que son una consecuencia natural de las normas vectoriales l_1, l_2 y l_∞.

Teorema XIII.12

Si $||\ \cdot\ ||$ es cualquier norma vectorial en $\mathcal{R}^n$, entonces

$$||A|| = \max_{||\mathbf{x}||=1} ||A\mathbf{x}||$$

define una norma matricial en el conjunto de las matrices reales $n \times n$, que se llama la **norma natural**.

Consecuentemente, las normas matriciales que consideraremos tienen las formas

$$||A||_1 = \max_{||\mathbf{x}||_1=1} ||A\mathbf{x}||_1 \ , \quad \text{norma } \ l_1 \ ,$$

$$||A||_2 = \max_{||\mathbf{x}||_2=1} ||A\mathbf{x}||_2 \ , \quad \text{norma } \ l_2 \ ,$$

y

$$||A||_\infty = \max_{||\mathbf{x}||_\infty=1} ||A\mathbf{x}||_\infty \ , \quad \text{norma } \ l_\infty \ .$$

Teorema XIII.13
Si $A = (a_{ij})$ es una matriz $n \times n$, entonces

$$a) \qquad ||A||_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ ,$$

$$b) \qquad ||A||_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}| \ .$$

Demostración: a) sea $\mathbf{x}$ un vector columna $n-$dimensional tal que su norma l_∞ sea uno; es decir, $||\mathbf{x}||_\infty = \max\limits_{1\le i\le n} |x_i| = 1$. Como $A\mathbf{x}$ es también un vector columna $n-$dimensional,

$$||A\mathbf{x}||_\infty = \max_{1\le i\le n} |(A\ x)_i| = \max_{1\le i\le n} |\sum_{j=1}^{n} a_{ij}\ x_j| \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ \max_{1\le j\le n} |x_j|$$

$$= \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ ||\mathbf{x}||_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ .$$

Así que $||A\mathbf{x}||_\infty \le \max\limits_{1\le i\le n} \sum\limits_{j=1}^{n} |a_{ij}|$ para toda $\mathbf{x}$ con $||\mathbf{x}||_\infty = 1$. Consecuentemente,

$$||A||_\infty = \max_{||\mathbf{x}||_\infty=1} ||A\mathbf{x}||_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ .$$

Por otro lado, si p es el entero $1 \le p \le n$, con

$$\sum_{j=1}^{n} |a_{pj}| = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ ,$$

y $\mathbf{x}$ se escoge de tal manera que

$$x_j = \begin{cases} 1 \ , & \text{si } a_{pj} \ge 0, \\ -1 \ , & \text{si } a_{pj} < 0, \end{cases}$$

entonces $||\mathbf{x}||_\infty = 1$ y $|a_{pj}\ x_j| = |a_{pj}|$ para toda $j = 1, 2, \ldots, n$. Además,

$$||A\mathbf{x}||_\infty = \max_{1\le i\le n} |\sum_{j=1}^{n} a_{ij}\ x_j| \ge |\sum_{j=1}^{n} a_{pj}\ x_j| = \sum_{j=1}^{n} |a_{pj}| = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ .$$

Esto implica que

$$||A||_\infty = \max_{||\mathbf{x}||_\infty=1} ||A\mathbf{x}||_\infty \geq \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \ .$$

Entonces,

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

Demostremos ahora la parte b); sea $\mathbf{x}$ un vector columna $n-$dimensional tal que su norma l_1 sea uno; es decir, $||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i| = 1$. Como $A\mathbf{x}$ es también un vector columna $n-$dimensional,

$$||A\mathbf{x}||_1 = \sum_{i=1}^{n} |(A\ x)_i| = \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{ij}\ x_j| = \sum_{j=1}^{n} |(\sum_{i=1}^{n} a_{ij})\ x_j| \leq$$
$$\leq \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| = \sum_{i=1}^{n} |a_{ij}|\ ||\mathbf{x}||_1 = \sum_{i=1}^{n} |a_{ij}| \ .$$

Así que $||A\mathbf{x}||_1 \leq \sum_{i=1}^{n} |a_{ij}|$ para toda $\mathbf{x}$ con $||\mathbf{x}||_1 = 1$. Consecuentemente,

$$||A||_1 = \max_{||\mathbf{x}||_1=1} ||A\mathbf{x}||_1 \leq \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \ .$$

Por otro lado, si p es el entero $1 \leq p \leq n$, con

$$\sum_{i=1}^{n} |a_{ip}| = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \ ,$$

y $\mathbf{x}$ se escoge de tal manera que

$$x_j = \begin{cases} 1 \ , & \text{si } j = p, \\ 0 \ , & \text{en el resto de los casos,} \end{cases}$$

entonces $||\mathbf{x}||_1 = \sum_{j=1}^{n} |x_j| = 1$. Además,

$$||A\mathbf{x}||_1 = \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{ij}\ x_j| \geq \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{ip}\ x_j| = \sum_{i=1}^{n} |a_{ip}| \sum_{j=1}^{n} |x_j| = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \ .$$

Esto implica que

$$||A\mathbf{x}||_1 = \max_{||\mathbf{x}||_1=1} ||A\mathbf{x}||_1 \geq \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \ .$$

Entonces,

$$||A\mathbf{x}||_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \ .$$

c.q.d.

Para investigar la norma l_2, es necesario discutir algunos conceptos adicionales del álgebra lineal.

Definición. Si A es una matriz real $n \times n$, el polinomio definido por

$$p(\lambda) = \det(A - \lambda\, I)$$

se llama **polinomio característico** de A.

Es fácil demostrar que $p(\lambda)$ es un polinomio de grado n con coeficientes reales y consecuentemente, tiene a lo más n ceros distintos, algunos de los cuales pueden ser complejos. Si λ es un cero de $p(\lambda)$, entonces debido a que $\det(A - \lambda\, I) = 0$, el Teorema XIII.4 implica que el sistema lineal definido por $(A - \lambda\, I)\ \mathbf{x} = \mathbf{0}$ tiene una solución diferente de la solución idénticamente cero (ó solución trivial). Deseamos estudiar los ceros de $p(\lambda)$ y las soluciones no triviales correspondientes de estos sistemas.

Definición. Si $p(\lambda)$ es el polinomio característico de la matriz A los ceros de $p(\lambda)$ se llaman **autovalores** (también llamados valores propios o valores característicos) de la matriz A. Si λ es un valor característico de A y $\mathbf{x} \neq \mathbf{0}$ tiene la propiedad de que $(A - \lambda\, I)\ \mathbf{x} = \mathbf{0}$, entonces $\mathbf{x}$ es el **autovector** (también llamado vector propio o vector característico) de A correspondiente al autovalor λ.

Ejemplo 8.

Sea

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

Para calcular los autovalores de A consideremos

$$\begin{aligned} p(\lambda) = \det(A - \lambda\, I) &= \det \begin{pmatrix} 1-\lambda & 0 & 1 \\ 2 & 2-\lambda & 1 \\ -1 & 0 & -\lambda \end{pmatrix} = \\ &= (1-\lambda)\, \det \begin{pmatrix} 2-\lambda & 1 \\ 0 & -\lambda \end{pmatrix} + 1\, \det \begin{pmatrix} 2 & 2-\lambda \\ -1 & 0 \end{pmatrix} = \\ &= (1-\lambda)(2-\lambda)(-\lambda) + (2-\lambda) = (2-\lambda)(\lambda^2 - \lambda + 1) \ . \end{aligned}$$

Los autovalores de A son las soluciones de $p(\lambda) = 0$, es decir,

$$\lambda_1 = 2 \ , \quad \lambda_2 = \frac{1}{2} + \frac{\sqrt{3}}{2}\ i \ , \quad \lambda_3 = \frac{1}{2} - \frac{\sqrt{3}}{2}\ i \ .$$

Un autovector $\mathbf{x}$ de A asociado con λ_1 es una solución del sistema $(A - \lambda\, I)\ \mathbf{x} = \mathbf{0}$:

$$\begin{pmatrix} -1 & 0 & 1 \\ 2 & 0 & 1 \\ -1 & 0 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \ .$$

Por lo tanto,

$$\begin{aligned} -\ x_1 \ &+ \ \ x_3 \ &= \ \ 0 \ , \\ 2\, x_1 \ &+ \ \ x_3 \ &= \ \ 0 \ , \\ -\ x_1 \ &- \ 2\, x_3 \ &= \ \ 0 \ , \end{aligned}$$

sistema que tiene por solución a $x_1 = x_3 = 0$ y x_2 arbitrario. En particular, $\mathbf{x} = (0, 1, 0)^t$ es un autovector de A correspondiente al autovalor $\lambda_1 = 2$.

Definición. El **radio espectral** $\rho(A)$ de una matriz A se define como

$$\rho(A) = \max |\lambda|$$

donde λ es un valor característico de A.

El radio espectral está relacionado con la norma de una matriz, como muestra el siguiente Teorema.

Teorema XIII.14

Si $A = (a_{ij})$ es una matriz real $n \times n$, entonces

i) $[\rho(A^t\ A)]^{1/2} = ||A||_2$;

ii) $\rho(A) \leq ||A||$ para cualquier norma natural $||\ \cdot\ ||$.

Un resultado útil e interesante es que para cualquier matriz A y cualquier $\varepsilon > 0$, existe una norma $||\ \cdot\ ||$ con la propiedad de que $||A|| < \rho(A) + \varepsilon$. Consecuentemente, $\rho(A)$ es la máxima cota inferior para las normas de A.

Ejemplo 9.

Sea

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

entonces, como A es simétrica, $A^t = A$ y

$$A^t\ A = A^2 = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 3 & 1 \\ 3 & 3 & 3 \\ 1 & 3 & 5 \end{pmatrix} \ .$$

Para calcular $\rho(A^t\ A)$, necesitamos los valores característicos de $A^t\ A$. Si

$$\begin{aligned} 0 = \det \begin{pmatrix} 5-\lambda & 3 & 1 \\ 3 & 3-\lambda & 3 \\ 1 & 3 & 5-\lambda \end{pmatrix} = \\ = (5-\lambda)^2(3-\lambda) + 9 + 9 - (3-\lambda) - 9\ (5-\lambda) - 9\ (5-\lambda) = \\ = \ -\lambda^3 + 13\lambda^2 - 36\lambda = -\lambda(\lambda-4)(\lambda-9)\ . \end{aligned}$$

Entonces, λ es 0, 4 ó 9. Por lo tanto

$$||A||_2 = \sqrt{\rho(A^t\ A)} = \sqrt{\max\{0, 4, 9\}} = 3\ .$$

En el estudio de las técnicas iterativas de matrices, es de particular importancia saber cuándo las potencias de una matriz se hacen pequeñas, es decir, cuándo todas las componentes tienden a cero. Las matrices de este tipo se denominan **convergentes**.

Definición. Llamamos a A $n \times n$ una matriz **convergente** si

$$\lim_{k\to\infty} (A^k)_{ij} = 0$$

para cada $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, n$.

Ejemplo 10.

Sea

$$A = \begin{pmatrix} \frac{1}{2} & 0 \\ \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix} .$$

Calculando las potencias de A, obtenemos:

$$A^2 = \begin{pmatrix} \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix} , \quad A^3 = \begin{pmatrix} \frac{1}{8} & 0 \\ \\ \frac{3}{16} & \frac{1}{8} \end{pmatrix} , \quad A^4 = \begin{pmatrix} \frac{1}{16} & 0 \\ \\ \frac{1}{8} & \frac{1}{16} \end{pmatrix} ,$$

y en general

$$A^k = \begin{pmatrix} \frac{1}{2}^k & 0 \\ \\ \frac{k}{2^{k+1}} & \frac{1}{2}^k \end{pmatrix} .$$

Como

$$\lim_{k\to\infty} (\frac{1}{2})^k = 0 \quad \text{y} \quad \lim_{k\to\infty} \frac{k}{2^{k+1}} = 0 ,$$

A es una matriz convergente. Nótese que $\rho(A) = \frac{1}{2}$, ya que $\frac{1}{2}$ es el único valor característico de A.

Existe una conexión importante entre el radio espectral de una matriz y su convergencia.

Teorema XIII.15

Las siguientes afirmaciones son equivalentes:

1. A es una matriz convergente;
2. $\lim_{n\to\infty} ||A^n|| = 0$, para alguna norma natural $||\ \cdot\ ||$;
3. $\rho(A) \leq 1$;
4. $||A|| \leq 1$;
5. $\lim_{n\to\infty} A^n\ \mathbf{x} = \mathbf{0}$, para toda $\mathbf{x}$.

CAPITULO XIV. ELIMINACION GAUSSIANA Y SUSTITUCION HACIA ATRAS

1. INTRODUCCION Y METODO

El procedimiento general de eliminación Gaussiana aplicado al sistema

$$\begin{aligned} E_1: &\quad a_{11}\,x_1 + a_{12}\,x_2 + \ldots + a_{1n}\,x_n = b_1\ , \\ E_2: &\quad a_{21}\,x_1 + a_{22}\,x_2 + \ldots + a_{2n}\,x_n = b_2\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\,x_1 + a_{n2}\,x_2 + \ldots + a_{nn}\,x_n = b_n\ , \end{aligned} \tag{$XIV.1$}$$

se maneja de una manera similar al procedimiento seguido en el ejemplo 2 del Capítulo XIII. Formamos la matriz ampliada A_a:

$$A_a = [A, \mathbf{b}] = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & a_{1,n+1} \\ a_{21} & a_{22} & \ldots & a_{2n} & a_{2,n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & a_{n,n+1} \end{array}\right)\ , \tag{$XIV.2$}$$

donde A denota la matriz formada por los coeficientes y los elementos en la $(n+1)$-ésima columna son los valores de $\mathbf{b}$, es decir, $a_{i,n+1} = b_i$ para cada $i = 1, 2, \ldots, n$. Siempre y cuando $a_{11} \neq 0$, se efectúan las operaciones correspondientes a $(E_j - (a_{j1}/a_{11})E_1) \to (E_j)$ para cada $j = 2, 3, \ldots, n$ para eliminar el coeficiente de x_1 en cada una de estas filas. Aún cuando se espera que los elementos de las filas $2, 3, \ldots, n$ cambien, para facilitar la notación, denotaremos nuevamente el elemento en la $i-$ésima fila y en la $j-$ésima columna por a_{ij}. Teniendo en cuenta esto, seguiremos un procedimiento secuencial para $i = 2, 3, \ldots, n-1$ y realizamos la operación $(E_j - (a_{ji}/a_{ii})E_i) \to (E_j)$ para cada $j = i+1, i+2, \ldots, n$, siempre que $a_{ii} \neq 0$. Esto eliminará x_i en cada fila debajo de la $i-$ésima para todos los valores de $i = 1, 2, \ldots, n-1$. La matriz resultante tendrá la forma:

$$A_a^{(f)} = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & a_{1,n+1} \\ 0 & a_{22} & \ldots & a_{2n} & a_{2,n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & a_{nn} & a_{n,n+1} \end{array}\right)\ .$$

Esta matriz representa un sistema lineal con el mismo conjunto de soluciones que el sistema ($XIV.1$). Como el sistema lineal equivalente es triangular:

$$\begin{array}{ccccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \ldots & \ldots & + & a_{1n}\,x_n & = & a_{1,n+1}\ , \\ & & a_{22}\,x_2 & + & \ldots & \ldots & + & a_{2n}\,x_n & = & a_{2,n+1}\ , \\ & & & & \ldots & \ldots & \ldots & \ldots & = & \ldots \\ & & & & & a_{n-1,n-1}\,x_{n-1} & + & a_{n-1,n}\,x_n & = & a_{n-1,n+1}\ , \\ & & & & & & & a_{nn}\,x_n & = & a_{n,n+1}\ , \end{array}$$

se puede realizar la sustitución hacia atrás. Resolviendo la $n-$ésima ecuación para x_n se obtiene:

$$x_n = \frac{a_{n,n+1}}{a_{nn}}\ .$$

Resolviendo la ecuación $(n-1)$-ésima para x_{n-1} y usando x_n obtenemos:

$$x_{n-1} = \frac{(a_{n-1,n+1} - a_{n-1,n}\ x_n)}{a_{n-1,n-1}}\ .$$

Y continuando con este proceso, llegamos a que

$$x_i = \frac{(a_{i,n+1} - a_{in}\ x_n - a_{i,n-1}\ x_{n-1} - \ldots - a_{i,i+1}\ x_{i+1})}{a_{ii}} =$$

$$= \frac{(a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}\ x_j)}{a_{ii}}\ ,$$

para cada $i = n-1, n-2, \ldots, 2, 1$.

El procedimiento de eliminación Gaussiana se puede mostrar más detalladamente, aunque de forma más complicada, formando una secuencia de matrices ampliadas $A_a^{(1)}$, $A_a^{(2)}$, ..., $A_a^{(n)}$, donde $A_a^{(1)}$ es la matriz A_a dada en la ecuación $(XIV.2)$ y $A_a^{(k)}$ con $k = 2, 3, \ldots, n$ tiene los elementos $a_{ij}^{(k)}$ de la forma:

$$a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)} & \text{cuando } i = 1, 2, \ldots, k-1 \\ & \text{y } j = 1, 2, \ldots, n+1\ , \\ 0 & \text{cuando } i = k, k+1, \ldots, n \\ & \text{y } j = 1, 2, \ldots, k-1\ , \\ a_{ij}^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}}\ a_{k-1,j}^{(k-1)} & \text{cuando } i = k, k+1, \ldots, n \\ & \text{y } j = k, k+1, \ldots, n+1\ . \end{cases}$$

$$A_a^{(k)} = \left(\begin{array}{ccccccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & \ldots & a_{1n}^{(1)} & | & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & \ldots & a_{2n}^{(2)} & | & a_{2,n+1}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \ldots & a_{3,k-1}^{(3)} & a_{3k}^{(3)} & \ldots & a_{3n}^{(3)} & | & a_{3,n+1}^{(3)} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & | & \ldots \\ 0 & \ldots & \ldots & 0 & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \ldots & a_{k-1,n}^{(k-1)} & | & a_{k-1,n+1}^{(k-1)} \\ 0 & \ldots & \ldots & \ldots & 0 & a_{kk}^{(k)} & \ldots & a_{kn}^{(k)} & | & a_{k,n+1}^{(k)} \\ 0 & \ldots & \ldots & \ldots & 0 & \ldots & \ldots & \ldots & | & \ldots \\ 0 & \ldots & \ldots & \ldots & 0 & a_{nk}^{(k)} & \ldots & a_{nn}^{(k)} & | & a_{n,n+1}^{(k)} \end{array}\right)\ ,$$

es la matriz que representa el sistema lineal equivalente para el cual la variable x_{k-1} acaba de ser eliminada de las ecuaciones $E_k, E_{k+1}, \ldots, E_n$.

El procedimiento no funcionará si alguno de los elementos $a_{11}^{(1)}$, $a_{22}^{(2)}$, ..., $a_{n-1,n-1}^{(n-1)}$, $a_{nn}^{(n)}$ es cero, ya que en este caso el paso $(E_i - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\ E_k) \to E_i$ no se puede realizar (esto ocurre si una de las $a_{11}^{(1)}, a_{22}^{(2)}, a_{33}^{(3)}, \ldots, a_{n-1,n-1}^{(n-1)}$ es cero), o la sustitución hacia atrás no se puede llevar a cabo (en el caso $a_{nn}^{(n)}$). Esto no significa que el sistema lineal no sea resoluble, sino que la técnica de resolución debe alterarse. Cuando $a_{kk}^{(k)} = 0$ para algún $k = 1, 2, \ldots, n-1$, se busca en la $k-$ésima columna de $A_a^{(k-1)}$ desde la fila k hasta la n para encontrar el primer elemento diferente de cero. Si $a_{pk}^{(k)} \neq 0$ para algún p,

$k+1 \leq p \leq n$, entonces se efectúa la operación $(E_k) \leftrightarrow (E_p)$ para obtener $A_a^{(k-1)}$. El procedimiento puede continuar entonces para formar $A_a^{(k)}$, y así proseguir. Si $a_{pk}^{(k)} = 0$ para $p = k, k+1, \ldots, n$, se puede demostrar (Teorema XIII.4) que el sistema lineal no tiene una solución única y el procedimiento se para. Finalmente, si $a_{nn}^{(n)} = 0$ el sistema lineal no tiene una solución única y el procedimiento se para.

El ejemplo siguiente ilustra el funcionamiento de este método:

Ejemplo 1.

Resolver el sistema de ecuaciones:

$$\begin{array}{lrcrcrcrcr} E_1: & x_1 & - & x_2 & + & 2\,x_3 & - & x_4 & = & -8\,, \\ E_2: & 2\,x_1 & - & 2\,x_2 & + & 3\,x_3 & - & 3\,x_4 & = & -20\,, \\ E_3: & x_1 & + & x_2 & + & x_3 & & & = & -2\,, \\ E_4: & x_1 & - & x_2 & + & 4\,x_3 & + & 3\,x_4 & = & 4\,. \end{array}$$

La matriz ampliada es

$$A_a = A_a^{(1)} = \left(\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 2 & -2 & 3 & -3 & -20 \\ 1 & 1 & 1 & 0 & -2 \\ 1 & -1 & 4 & 3 & 4 \end{array}\right)\,,$$

y efectuando las operaciones $(E_2 - 2E_1) \to (E_2)$, $(E_3 - E_1) \to (E_3)$ y $(E_4 - E_1) \to (E_4)$ llegamos a:

$$A_a^{(2)} = \left(\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right)\,.$$

Como el elemento $a_{22}^{(2)}$, llamado **elemento de pivote**, es cero, el procedimiento no puede continuar de la misma forma, pero la operación $(E_i) \leftrightarrow (E_j)$ está permitida, así que se hace una búsqueda de los elementos $a_{32}^{(2)}$ y $a_{42}^{(2)}$ para encontrar el primer elemento no cero. Ya que $a_{32}^{(2)} \neq 0$, se efectúa la operación $(E_2) \leftrightarrow (E_3)$ para obtener una nueva matriz

$$A_a^{(2)'} = \left(\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right)\,.$$

Como x_2 está ya eliminada de E_3 y E_4, $A_a^{(3)}$ será $A_a^{(2)'}$ y los cálculos pueden continuar con la operación $(E_4 + 2E_3) \to (E_4)$, dando

$$A_a^{(4)} = \left(\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 0 & 2 & 4 \end{array}\right)\,.$$

Finalmente, se puede aplicar la sustitución hacia atrás:

$$x_4 = \frac{4}{2} = 2\,, \qquad x_3 = \frac{[-4 - (-1)\,x_4]}{-1} = 2\,,$$
$$x_2 = \frac{[6 - x_4 - (-1)\,x_3]}{2} = 3\,, \qquad x_1 = \frac{[-8 - (-1)\,x_4 - 2\,x_3 - (-1)\,x_2]}{1} = -7\,.$$

2. ALGORITMO Y EJEMPLOS

Para resumir el método de eliminación Gaussiana completo con sustitución hacia atrás, se presenta el siguiente algoritmo.

Algoritmo de eliminación Gaussiana con sustitución hacia atrás.

==========

Para resolver el sistema lineal de $n \times n$:

$$\begin{aligned} E_1: &\quad a_{11}\, x_1 + a_{12}\, x_2 + \ldots + a_{1n}\, x_n = a_{1,n+1} \\ E_2: &\quad a_{21}\, x_1 + a_{22}\, x_2 + \ldots + a_{2n}\, x_n = a_{2,n+1} \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\, x_1 + a_{n2}\, x_2 + \ldots + a_{nn}\, x_n = a_{n,n+1} \end{aligned}$$

Entrada: número de incógnitas y de ecuaciones n; matriz ampliada $A_a = (a_{ij})$ donde $1 \le i \le n$ y $1 \le j \le n+1$.

Salida: solución $x_1, x_2, \ldots, x_n$ ó mensaje de que el sistema lineal no tiene solución única.

Paso 1: Para $i = 1, 2, \ldots, n-1$ seguir los pasos 2–4 *(proceso de eliminación).*

Paso 2: Sea p el menor entero con $i \le p \le n$ y $a_{pi} \neq 0$. Si p no puede encontrarse entonces SALIDA; *(no existe solución única)* PARAR.

Paso 3: Si $p \neq i$ entonces efectuar $(E_p) \leftrightarrow (E_i)$.

Paso 4: Para $j = i+1, i+2, \ldots, n$ seguir los pasos 5 y 6.

Paso 5: Tomar $m_{ji} = \frac{a_{ji}}{a_{ii}}$.

Paso 6: Efectuar $(E_j - m_{ji}\, E_i) \to (E_j)$.

Paso 7: Si $a_{nn} = 0$ entonces SALIDA; *(no existe solución única)* PARAR.

Paso 8: *(Empieza la sustitución hacia atrás)*; tomar

$$x_n = \frac{a_{n,n+1}}{a_{nn}}\ .$$

Paso 9: Para $i = n-1, n-2, \ldots, 1$ tomar

$$x_i = \frac{a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}\, x_j}{a_{ii}}\ .$$

Paso 10: SALIDA $(x_1, x_2, \ldots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

==========

Ejemplo 2.

Resolver los dos sistemas lineales:

$$\begin{array}{lrcrcrcrcr} E_{1,(1)}: & x_1 & + & x_2 & + & x_3 & + & x_4 & = & 7\,, \\ E_{2,(1)}: & x_1 & + & x_2 & & & + & 2\,x_4 & = & 8\,, \\ E_{3,(1)}: & 2\,x_1 & + & 2\,x_2 & + & 3\,x_3 & & & = & 10\,, \\ E_{4,(1)}: & -\,x_1 & - & x_2 & - & 2\,x_3 & + & 2\,x_4 & = & 0\,, \end{array}$$

y

$$\begin{array}{lrcrcrcrcr} E_{1,(2)}: & x_1 & + & x_2 & + & x_3 & + & x_4 & = & 7\,, \\ E_{2,(2)}: & x_1 & + & x_2 & & & + & 2\,x_4 & = & 5\,, \\ E_{3,(2)}: & 2\,x_1 & + & 2\,x_2 & + & 3\,x_3 & & & = & 10\,, \\ E_{4,(2)}: & -\,x_1 & - & x_2 & - & 2\,x_3 & + & 2\,x_4 & = & 0\,. \end{array}$$

Estos sistemas dan lugar a las matrices

$$A^{(1)}_{a(1)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 1 & 1 & 0 & 2 & 8 \\ 2 & 2 & 3 & 0 & 10 \\ -1 & -1 & -2 & 2 & 0 \end{array}\right) \quad \text{y} \quad A^{(1)}_{a(2)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 1 & 1 & 0 & 2 & 5 \\ 2 & 2 & 3 & 0 & 10 \\ -1 & -1 & -2 & 2 & 0 \end{array}\right) .$$

Ya que $a_{11} = 1 \neq 0$, los pasos para eliminar x_1 de E_2, E_3 y E_4 dan, para $i = 1$

$$m_{ji} = m_{j1} = \frac{a_{j1}}{a_{11}} = \frac{a_{j1}}{1} = a_{j1}\ .$$

Entonces:

$$j = 2,\ \ m_{21} = 1\ ; \qquad j = 3,\ \ m_{31} = 2\ ; \qquad j = 4,\ \ m_{41} = -1\ ;$$

y las operaciones a efectuar son:

$$(E_2 - E_1) \to (E_2)\ ; \qquad (E_3 - 2E_1) \to (E_3)\ ; \qquad (E_4 + E_1) \to (E_4)\ .$$

Las matrices se trasforman en:

$$A^{(2)}_{a(1)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 1 & -2 & -4 \\ 0 & 0 & -1 & 3 & 7 \end{array}\right) \quad \text{y} \quad A^{(2)}_{a(2)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 0 & 0 & -1 & 1 & -2 \\ 0 & 0 & 1 & -2 & -4 \\ 0 & 0 & -1 & 3 & 7 \end{array}\right) .$$

Aquí $a_{22} = a_{32} = a_{42} = 0$ y el algoritmo requiere que el procedimiento se detenga y no se obtiene una solución para ninguno de los sistemas.

Para examinar más de cerca la razón de la dificultad, efectuamos $(E_4 + E_3) \to (E_4)$ para obtener $A^{(3)}_{a(1)} = A^{(4)}_{a(1)}$ y $A^{(3)}_{a(2)} = A^{(4)}_{a(2)}$

$$A^{(4)}_{a(1)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 1 & -2 & -4 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right) \quad \text{y} \quad A^{(4)}_{a(2)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 0 & 0 & -1 & 1 & -2 \\ 0 & 0 & 1 & -2 & -4 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right) .$$

Escribiendo las ecuaciones para cada sistema se obtiene:

$$\begin{array}{lcrcrcrcr} x_1 & + & x_2 & + & x_3 & + & x_4 & = & 7\,, \\ & & - & & x_3 & + & x_4 & = & 1\,, \\ & & & & x_3 & - & 2\,x_4 & = & -4\,, \\ & & & & & & x_4 & = & 3\,, \end{array}$$

y

$$\begin{array}{rcrcrcrcl} x_1 & + & x_2 & + & x_3 & + & x_4 & = & 7\,, \\ & & & - & x_3 & + & x_4 & = & -2\,, \\ & & & & x_3 & - & 2\,x_4 & = & -4\,, \\ & & & & & & x_4 & = & 3\,. \end{array}$$

Efectuando sustitución hacia atrás en cada sistema nos lleva a:

$$x_4 = 3 \quad \text{y} \quad x_3 = -4 + 2\,x_4 = 2\,,$$

en ambos sistemas. Si se continúa la sustitución hacia atrás hasta la segunda ecuación en cada caso, la diferencia entre los dos sistemas se hace aparente ya que en el primer sistema

$$-x_3 + x_4 = 1 \quad \text{implica que} \quad 1 = 1\,,$$

mientras que en el segundo sistema

$$-x_3 + x_4 = -2 \quad \text{implica que} \quad 1 = -2\,.$$

El primer sistema lineal tiene un número infinito de soluciones $x_4 = 3$, $x_3 = 2$, x_2 arbitraria y $x_1 = 2 - x_2$, mientras que el segundo nos lleva a una contadicción y no existe solución. En ambos casos, sin embargo, no hay una solución única como concluímos a partir del algoritmo de eliminación Gaussiana con sustitución hacia atrás.

Cuando se comparan las técnicas para resolver sistemas lineales, se necesita considerar otros conceptos además de la cantidad de lugar requerido para almacenamiento. Uno de éstos conceptos es el efecto del error de redondeo y otro es la cantidad de tiempo requerido para completar los cálculos. Ambos dependen del número de operaciones aritméticas que se necesitan efectuar para resolver un problema. En general, el tiempo requerido para realizar una multiplicación o división es considerablemente mayor que el requerido para realizar una suma o una resta. Para mostrar el procedimiento que se emplea para contar las operaciones en un método dado, contaremos las operaciones necesarias para resolver un sistema lineal típico de n ecuaciones con n incógnitas usando el algoritmo de la eliminación Gaussiana con sustitución hacia atrás.

Hasta los pasos 5 y 6 del algoritmo no se efectúan operaciones aritméticas. El paso 5 requiere que se realicen $(n-i)$ divisiones. El reemplazar la ecuación E_j por $(E_j - m_{ji}E_i)$ en el paso 6 requiere que m_{ji} se multiplique por cada término en E_i resultando un total de $(n-i)(n-i+2)$ multiplicaciones. Después de completar esto, cada término de la ecuación resultante se resta del término correspondiente en E_j. Esto requiere $(n-i)(n-i+2)$ restas. Para cada $i = 1, 2, \ldots, n-1$, las operaciones requeridas en los pasos 5 y 6 son

Multiplicaciones/Divisiones

$$(n-i) + (n-i)\,(n-i+2) = (n-i)\,(n-i+3)\,,$$

Sumas/Restas

$$(n-i)\,(n-i+2)\,.$$

El número total de operaciones requeridas en estos pasos se obtiene sumando las cuentas de las operaciones para cada i. Recordando que

$$\sum_{j=1}^{m} 1 = m \ , \quad \sum_{j=1}^{m} j = \frac{m(m+1)}{2} \ , \quad \sum_{j=1}^{m} j^2 = \frac{m(m+1)(2m+1)}{6} \ ,$$

obtenemos

Multiplicaciones/Divisiones

$$\sum_{i=1}^{n-1}(n-i)(n-i+3) = (n^2+3n)\sum_{i=1}^{n-1} 1 - (2n+3)\sum_{i=1}^{n-1} i + \sum_{i=1}^{n-1} i^2 =$$

$$= (n^2+3n)(n-1) - (2n+3)\frac{(n-1)n}{2} + \frac{(n-1)n(2n-1)}{6} = \frac{n^3+3n^2-4n}{3} \ ,$$

Sumas/Restas

$$\sum_{i=1}^{n-1}(n-i)(n-i+2) = (n^2+2n)\sum_{i=1}^{n-1} 1 - 2(n+1)\sum_{i=1}^{n-1} i + \sum_{i=1}^{n-1} i^2 =$$

$$= (n^2+2n)(n-1) - 2(n+1)\frac{(n-1)n}{2} + \frac{(n-1)n(2n-1)}{6} = \frac{2n^3+3n^2-5n}{6} \ .$$

Los otros pasos del algoritmo de la eliminación Gaussiana con sustitución hacia atrás que requieren de operaciones aritméticas son los pasos 8 y 9. El n^0 8 requiere de una división. El n^0 9 requiere de $(n-i)$ multiplicaciones y $(n-i-1)$ sumas para cada término con sumatorio y luego una resta y una división. El número total de operaciones en los pasos 8 y 9 es

Multiplicaciones/Divisiones

$$1 + \sum_{i=1}^{n-1}[(n-i)+1] = \frac{n^2+n}{2} \ ,$$

Sumas/Restas

$$\sum_{i=1}^{n-1}[(n-i-1)+1] = \frac{n^2-n}{2} \ .$$

El total de operaciones aritméticas en el algoritmo de la eliminación Gaussiana con sustitución hacia atrás es por lo tanto

Multiplicaciones/Divisiones

$$\frac{n^3+3n^2-4n}{3} + \frac{n^2+n}{2} = \frac{2n^3+9n^2-5n}{6} \ ,$$

Sumas/Restas

$$\frac{2n^3+3n^2-5n}{6} + \frac{n^2-n}{2} = \frac{n^3+3n^2-4n}{3} \ .$$

Como el número total de multiplicaciones y de divisiones es aproxidamente $n^3/3$, y similar para sumas y restas, la cantidad de cómputo y el tiempo requerido se incrementarán con n proporcionalmente a n^3.

EJERCICIOS.

1. Resolver los siguientes sistemas lineales usando eliminación Gaussiana con sustitución hacia atrás y aritmética de redondeo a dos dígitos, sin reordenar las ecuaciones. (la solución exacta de cada sistema es $x_1 = 1$, $x_2 = -1$, $x_3 = 3$.)

a)
$$\begin{aligned} 4\,x_1 &+ x_2 - x_3 &= 0\,,\\ 2\,x_1 &+ 5\,x_2 + 2\,x_3 &= 3\,,\\ x_1 &+ 2\,x_2 + 4\,x_3 &= 11\,. \end{aligned}$$

b)
$$\begin{aligned} x_1 &+ 2\,x_2 + 4\,x_3 &= 11\,,\\ 4\,x_1 &+ x_2 - x_3 &= 0\,,\\ 2\,x_1 &+ 5\,x_2 + 2\,x_3 &= 3\,. \end{aligned}$$

c)
$$\begin{aligned} 4\,x_1 &+ x_2 + 2\,x_3 &= 9\,,\\ 2\,x_1 &+ 4\,x_2 - x_3 &= -5\,,\\ x_1 &+ x_2 - 3\,x_3 &= -9\,. \end{aligned}$$

d)
$$\begin{aligned} 2\,x_1 &+ 4\,x_2 - x_3 &= -5\,,\\ x_1 &+ x_2 - 3\,x_3 &= -9\,,\\ 4\,x_1 &+ x_2 + 2\,x_3 &= 9\,. \end{aligned}$$

2. Usar el algoritmo de eliminación Gaussiana con sustitución hacia atrás y aritmética exacta para resolver, si es posible, los sistemas lineales siguientes y determinar si son necesarios cambios de filas.

a)
$$\begin{aligned} x_1 &- x_2 + 3\,x_3 &= 2\,,\\ 3\,x_1 &- 3\,x_2 + x_3 &= -1\,,\\ x_1 &+ x_2 &= 3\,. \end{aligned}$$

b)
$$\begin{aligned} x_2 &+ 4\,x_3 &= 0\,,\\ x_1 &- x_2 - x_3 &= 0.375\,,\\ x_1 &- x_2 + 2\,x_3 &= 0\,. \end{aligned}$$

c)
$$\begin{aligned} 2\,x_1 &- 1.5\,x_2 + 3\,x_3 &= 1\,,\\ -\,x_1 &+ 2\,x_3 &= 3\,,\\ 4\,x_1 &- 4.5\,x_2 + 5\,x_3 &= 1\,. \end{aligned}$$

d)
$$\begin{aligned} 2\,x_1 &- x_2 + x_3 &= -1\,,\\ 3\,x_1 &+ 3\,x_2 + 9\,x_3 &= 0\,,\\ 3\,x_1 &+ 3\,x_2 + 5\,x_3 &= 4\,. \end{aligned}$$

CAPITULO XV. ESTRATEGIAS DE PIVOTEO

1. INTRODUCCION Y METODO

Durante la derivación del algoritmo de la eliminación Gaussiana con sustitución hacia atrás, se encontró que para obtener un cero para el elemento pivote $a_{kk}^{(k)}$ era necesario un intercambio de filas de la forma $(E_k) \leftrightarrow (E_p)$ donde $k+1 \leq p \leq n$ era el entero más pequeño con $a_{pk}^{(k)} \neq 0$. En la práctica frecuentemente es deseable realizar intercambios de las filas que contienen a los elementos pivote, aun cuando éstos no sean cero. Cuando los cálculos se realizan usando aritmética de dígitos finitos, como sería el caso de las soluciones generadas con calculadora u ordenador, un elemento pivote que sea pequeño comparado con los elementos de debajo de él en la misma columna puede llevar a un error de redondeo sustancial. En el ejemplo siguiente se da una ilustración de esta dificultad.

Ejemplo 1.

El sistema lineal

$$\begin{aligned} E_1: \quad 0.003\, x_1 \;+\; 59.14\, x_2 &= 59.17\,, \\ E_2: \quad 5.291\, x_1 \;-\; 6.130\, x_2 &= 46.78\,, \end{aligned}$$

tiene la solución exacta $x_1 = 10.00$ y $x_2 = 1.000$.

Para ilustrar las dificultades del error de redondeo, se aplicará eliminación Gaussiana a este sistema usando aritmética de cuatro dígitos con redondeo.

El primer elemento pivote es $a_{11}^{(1)} = 0.003$ y su multiplicador asociado es

$$m_{21} = \frac{5.291}{0.003} = 1763.\overline{6}\,,$$

el cual se redondea a 1764. Efectuando la operación $(E_2 - m_{21}E_1) \rightarrow (E_2)$ y el redondeo apropiado ($1764 \cdot 59.14 = 104322 = 104300$ y $1764 \cdot 59.17 = 104375 = 104400$),

$$\begin{aligned} 0.003\, x_1 \;-\; 59.14\, x_2 &= 59.17\,, \\ -\; 104300\, x_2 &= -104400\,. \end{aligned}$$

La sustitución hacia atrás implica que

$$x_2 = 1.001\,, \qquad x_1 = \frac{59.17 - 59.14 \cdot 1.001}{0.003} = \frac{59.17 - 59.20}{0.003} = -\frac{0.030}{0.003} = -10.00\,.$$

El error absoluto tan grande en la solución numérica de x_1 resulta del error pequeño de 0.001 al resolver para x_2. Este error absoluto fue amplificado por un factor de 20000 en la solución de x_1 debido al orden en el que fueron realizados los cálculos.

El ejemplo 1 ilustra las dificultades que pueden surgir en algunos casos cuando el elemento pivote $a_{kk}^{(k)}$ es pequeño en relación a los elementos $a_{ij}^{(k)}$ para $k \leq i \leq n$ y $k \leq j \leq n$. Las estrategias de pivoteo se llevan a cabo en general seleccionando un nuevo elemento como pivote $a_{pq}^{(k)}$ intercambiando las filas k y p, e intercambiando las columnas k y q, si es necesario. La estrategia más simple consiste en seleccionar el elemento en la

misma columna que está debajo de la diagonal y que tiene el mayor valor absoluto; es decir, se determina p tal que

$$|a_{pk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}| \ ,$$

y se efectúa $(E_k) \leftrightarrow (E_p)$. En este caso no se considera un intercambio de columnas.

Ejemplo 2.

Reconsideremos el sistema lineal del ejemplo 1:

$$\begin{array}{lrcrcl} E_1: & 0.003\ x_1 & + & 59.14\ x_2 & = & 59.17\ , \\ E_2: & 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78\ . \end{array}$$

Usando el procedimiento de pivoteo descrito arriba resulta que primero se encuentra

$$\max\{|a_{11}^{(1)}|, |a_{21}^{(1)}|\} = \max\{|0.003|, |5.291|\} = |5.291| = |a_{21}^{(1)}| \ .$$

Así, se realiza la operación $(E_2) \leftrightarrow (E_1)$ la cual da el sistema

$$\begin{array}{lrcrcl} E_1: & 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78\ , \\ E_2: & 0.003\ x_1 & + & 59.14\ x_2 & = & 59.17\ . \end{array}$$

El multiplicador para este sistema es

$$m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{0.003}{5.291} = 0.000567\ ,$$

y la operación $(E_2 - m_{21}E_1) \to (E_2)$ con el redondeo apropiado ($0.000567 \cdot 6.13 = 0.003476$ y $0.000567 \cdot 46.78 = 0.02652$) reduce el sistema a

$$\begin{array}{rcrcl} 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78\ , \\ & & 59.14\ x_2 & = & 59.14\ . \end{array}$$

Las respuestas con cuatro dígitos que resultan de la sustitución hacia atrás son los valores correctos $x_1 = 10.00$ y $x_2 = 1.000$.

Esta técnica se conoce como **pivoteo máximo de columna** o **pivoteo parcial**.

2. ALGORITMOS DE ELIMINACION GAUSSIANA CON PIVOTEO

A continuación se presenta el algoritmo de eliminación Gaussiana con pivoteo parcial (pivoteo máximo de columna). Los procedimientos detallados en este algoritmo son suficientes para garantizar que cada multiplicador m_{ij} tiene una magnitud que no excede a uno.

Algoritmo de eliminación Gaussiana con pivoteo máximo de columna.

==========

Para resolver el sistema lineal de $n \times n$:

$$\begin{array}{ll} E_1: & a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = a_{1,n+1} \\ E_2: & a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = a_{2,n+1} \\ & \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: & a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = a_{n,n+1} \end{array}$$

Entrada: número de incógnitas y de ecuaciones n; matriz ampliada $A_a = (a_{ij}) = (a(i,j))$ donde $1 \leq i \leq n$ y $1 \leq j \leq n+1$.

Salida: solución $x_1, x_2, \ldots, x_n$ ó mensaje de que el sistema lineal no tiene solución única.

Paso 1: Para $i = 1, 2, \ldots, n$ tomar $F(i) = i$;
(inicializar el indicador de la fila).

Paso 2: Para $i = 1, 2, \ldots, n-1$ seguir los pasos 3–6 *(proceso de eliminación).*

Paso 3: Sea p el menor entero con $i \leq p \leq n$ y

$$|a(F(p), i)| = \max_{i \leq j \leq n} |a(F(j), i)| \ .$$

Paso 4: Si $a(F(p), i) = 0$ entonces SALIDA;
(no existe solución única) PARAR.

Paso 5: Si $F(i) \neq F(p)$ entonces tomar $AUX = F(i)$, $F(i) = F(p)$, $F(p) = AUX$; *(intercambio de filas simulado).*

Paso 6: Para $j = i+1, i+2, \ldots, n$ seguir los pasos 7 y 8.

Paso 7: Tomar $m(F(j), i) = \frac{a(F(j),i)}{a(F(i),i)}$.

Paso 8: Efectuar $(E_{F(j)} - m(F(j), i)\ E_{F(i)}) \rightarrow (E_{F(j)})$.

Paso 9: Si $a(F(n), n) = 0$ entonces SALIDA; *(no existe solución única)* PARAR.

Paso 10: *(Empieza la sustitución hacia atrás)*; tomar

$$x_n = \frac{a(F(n), n+1)}{a(F(n), n)} \ .$$

Paso 11: Para $i = n-1, n-2, \ldots, 1$ tomar

$$x_i = \frac{a(F(i), n+1) - \sum_{j=i+1}^{n} a(F(i), j)\ x_j}{a(F(i), i)} \ .$$

Paso 12: SALIDA $(x_1, x_2, \ldots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

==========

Aún cuando la estrategia del pivoteo máximo de columna es suficiente para la mayoría de los sistemas lineales, se presentan a veces situaciones en las que esta estrategia resulta inadecuada.

Ejemplo 3.

El sistema lineal:

$$\begin{array}{lrcrcr} E_1: & 30.00\ x_1 & + & 591400\ x_2 & = & 591700\ , \\ E_2: & 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78\ , \end{array}$$

es el mismo sistema que el presentado en los ejemplos previos excepto que todos los coeficientes en la primera ecuación están multiplicados por 10^4. El procedimiento descrito

en el algoritmo de eliminación Gaussiana con pivoteo máximo de columna con aritmética de 4 dígitos lleva a los mismos resultados que se obtuvieron en el primer ejemplo.

El máximo valor en la primera columna es 30.00 y el multiplicador

$$m_{21} = \frac{5.291}{30.00} = 0.1764$$

y la operación $(E_2 - m_{21}E_1) \rightarrow (E_2)$ con el redondeo apropiado ($0.1764 \cdot 591400 = 104322 = 104300$ y $0.1764 \cdot 591700 = 104375 = 104400$) transformaría el sistema en

$$\begin{array}{rcrcr} 30.00\, x_1 & + & 591400\, x_2 & = & 591700\ , \\ & - & 104300\, x_2 & = & -104400\ , \end{array}$$

el cual tiene soluciones $x_2 = 1.001$ y $x_1 = -10.00$.

Para el sistema del ejemplo 3 es apropiada una técnica conocida como **pivoteo escalado de columna**. El primer paso en este procedimiento consiste en definir un factor de escala s_l para cada fila $l = 1, \ldots, n$

$$s_l = \max_{1 \le j \le n} |a_{lj}|\ .$$

Si $s_l = 0$ para algún l, los Teoremas XIII.3 y XIII.4 implican que no existe solución única y el procedimiento se detiene. El intercambio apropiado de filas para luego obtener ceros en la primera columna queda determinado escogiendo el primer entero $1 \le k \le n$ con

$$\frac{|a_{k1}|}{s_k} = \max_{1 \le j \le n} \frac{|a_{j1}|}{s_j}\ ,$$

y realizando $(E_1) \leftrightarrow (E_k)$. Igualmente, al paso generico i, el intercambio apropiado para llevar el elemento pivote a_{ii} en su posición, queda determinado escogiendo el menor entero k, $i \le k \le n$, con

$$\frac{|a_{ki}|}{s_k} = \max_{i \le j \le n} \frac{|a_{ji}|}{s_j}\ ,$$

y realizando $(E_i) \leftrightarrow (E_k)$. Si al efectuar este intercambio no se variían los factores de escala, diremos que estamos aplicando una estrategia de **pivoteo escalado de columna con factores de escalas fijos**. Por otra parte, otra estategia es efectuar también el intercambio $(s_i) \leftrightarrow (s_k)$ si se está haciendo el intercambio de filas $(E_i) \leftrightarrow (E_k)$ $(1 \le i \le n$, $i \le k \le n)$. En este caso diremos que se aplica la estrategia de pivoteo escalado de columna con intercambio completo o simplemente **pivoteo escalado de columna**.

Una modificación de esta técnica de pivoteo escalado de columna, que llameremos **pivoteo escalado de columna modificado**, consiste en redefinir los factores de escala a cada paso, es decir, al paso i-ésimo de nuestro algoritmo $(1 \le i \le n)$ se definen los factores de escala s_l para cada fila $l = i, \ldots, n$

$$s_l = \max_{i \le j \le n} |a_{lj}|\ .$$

Entonces, el intercambio apropiado de filas para llevar el elemento pivote a_{ii} en su posición queda determinado escogiendo el primer entero k, $i \leq k \leq n$, con

$$\frac{|a_{ki}|}{s_k} = \max_{i \leq j \leq n} \frac{|a_{j1}|}{s_j} \ ,$$

y realizando luego $(E_i) \leftrightarrow (E_k)$.

El efecto de escalar consiste en asegurar que el elemento mayor de cada fila tenga una magnitud relativa de uno antes de que se empiece la comparación para el intercambio de filas. El escalamiento se hace solamente con propósitos de comparación, así que la división entre los factores de escala no produce un error de redondeo en el sistema.

Aplicando la técnica de pivoteo escalado de columna al ejemplo 3 da

$$s_1 = \max\{|30.00|, |591400|\} = 591400 \ ,$$

$$s_2 = \max\{|5.291|, |-6.130|\} = 6.130 \ .$$

Consecuentemente,

$$\frac{|a_{11}|}{s_1} = \frac{30.00}{591400} = 0.5073 \times 10^{-4} \quad \text{y} \quad \frac{|a_{21}|}{s_2} = \frac{5.291}{6.130} = 0.8631 \ ,$$

y por lo cual se hace el intercambio $(E_1) \leftrightarrow (E_2)$.

Aplicando eliminación Gaussiana, el nuevo sistema

$$\begin{array}{rcrcr} 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78 \ , \\ 30.00\ x_1 & + & 591400\ x_2 & = & 591700 \ , \end{array}$$

producirá los resultados correctos $x_1 = 10.00$ y $x_2 = 1.000$. De hecho, el multiplicador es

$$m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{30.00}{5.291} = 5.67 \ ,$$

y la operación $(E_2 - m_{21}E_1) \rightarrow (E_2)$ (con $5.67 \cdot 6.13 = 34.76$ y $5.67 \cdot 46.78 = 256.2$) reduce el sistema a

$$\begin{array}{rcrcr} 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78 \ , \\ & & 591400\ x_2 & = & 591400 \ . \end{array}$$

Las respuestas con cuatro dígitos que resultan de la sustitución hacia atrás son los valores correctos $x_1 = 10.00$ y $x_2 = 1.000$.

Algoritmo de eliminación Gaussiana con pivoteo escalado de columna.

==========

Para resolver el sistema lineal de $n \times n$:

$$\begin{array}{rl} E_1: & a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = a_{1,n+1} \\ E_2: & a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = a_{2,n+1} \\ & \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: & a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = a_{n,n+1} \end{array}$$

Entrada: número de incógnitas y de ecuaciones n; matriz ampliada $A_a = (a_{ij}) = (a(i,j))$ donde $1 \le i \le n$ y $1 \le j \le n+1$.

Salida: solución $x_1, x_2, \ldots, x_n$ ó mensaje de que el sistema lineal no tiene solución única.

Paso 1: Para $i = 1, 2, \ldots, n$ tomar

$$s_i = s(i) = \max_{1 \le j \le n} |a(i,j)| \ ;$$

si $s_i = 0$ entonces SALIDA; *(no existe solución única)* PARAR. Tomar $F(i) = i$; *(inicializar el indicador de la fila).*

Paso 2: Para $i = 1, 2, \ldots, n-1$ seguir los pasos 3–6 *(proceso de eliminación).*

Paso 3: Sea p el menor entero con $i \le p \le n$ y

$$\frac{|a(F(p),i)|}{s(F(p))} = \max_{i \le j \le n} \frac{|a(F(j),i)|}{s(F(j))} \ .$$

Paso 4: Si $a(F(p),i) = 0$ entonces SALIDA;
(no existe solución única) PARAR.

Paso 5: Si $F(i) \neq F(p)$ entonces tomar $AUX = F(i)$, $F(i) = F(p)$, $F(p) = AUX$; *(intercambio de filas simulado).*

Paso 6: Para $j = i+1, i+2, \ldots, n$ seguir los pasos 7 y 8.

Paso 7: Tomar $m(F(j),i) = \frac{a(F(j),i)}{a(F(i),i)}$.

Paso 8: Efectuar $(E_{F(j)} - m(F(j),i)\ E_{F(i)}) \to (E_{F(j)})$.

Paso 9: Si $a(F(n),n) = 0$ entonces SALIDA; *(no existe solución única)* PARAR.

Paso 10: *(Empieza la sustitución hacia atrás)*; tomar

$$x_n = \frac{a(F(n),n+1)}{a(F(n),n)} \ .$$

Paso 11: Para $i = n-1, n-2, \ldots, 1$ tomar

$$x_i = \frac{a(F(i),n+1) - \sum_{j=i+1}^{n} a(F(i),j)\ x_j}{a(F(i),i)} \ .$$

Paso 12: SALIDA $(x_1, x_2, \ldots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

==========

Los cálculos adicionales requeridos para el pivoteo escalado de columna resultan primero de la determinación de los factores de escala, es decir $(n-1)$ comparaciones para cada uno de las n filas, que da un total de

$$n(n-1) \quad \text{comparaciones} \ .$$

Para determinar el primer intercambio correcto, se realizan n divisiones y se hacen $(n-1)$ comparaciones. La determinación del primer intercambio entonces, añade un total de

$$\textit{comparaciones} \quad n(n-1) + (n-1) \quad \text{y} \quad \textit{divisiones} \quad n \ .$$

Como los factores de escala se calculan sólo una vez, el segundo paso requiere solamente

$$\textit{comparaciones} \quad (n-2) \quad \text{y} \quad \textit{divisiones} \quad (n-1) \ .$$

Procediendo de manera similar, el procedimiento de pivoteo escalado de columna agrega un total de

$$\textit{comparaciones} \quad (n-1) + \sum_{k=2}^{n}(k-1) = \frac{3}{2}n(n-1)$$

y

$$\textit{divisiones} \quad \sum_{k=2}^{n} k = \frac{n(n+1)}{2} - 1 \ ,$$

al procedimiento de eliminación Gaussiana. El tiempo requerido para realizar una comparación es comparable, aunque un poco mayor, al de suma/resta. Entonces la técnica de escalamiento no incrementa significativamente el tiempo de cómputo requerido para resolver un sistema para valores grandes de n.

Si un sistema garantiza el tipo de pivoteo que da un pivoteo escalado de columna modificado, entonces se debe usar **pivoteo máximo** o **total**. Es decir, este pivoteo máximo en el $k-$ésimo paso busca todos los elementos

$$a_{ij} \quad \text{para} \quad i = k, k+1, \ldots, n, \quad \text{y} \quad j = k, k+1, \ldots, n,$$

para encontrar el elemento que tiene la magnitud más grande. Se realizan intercambios de filas y de columnas para traer este elemento a la posición pivote.

El primer paso de pivoteo total requiere que se realicen $(n^2 - 1)$ comparaciones, el segundo paso requiere $[(n-1)^2 - 1]$ comparaciones, y así sucesivamente. El tiempo total adicional requerido para incorporar el pivoteo total en la eliminación Gaussiana es consecuentemente

$$\textit{comparaciones} \quad \sum_{k=2}^{n}(k^2 - 1) = \frac{n(n-1)(2n+5)}{6} \ .$$

Este número es comparable con el número requerido por una técnica de pivoteo de columna modificada, pero no es necesaria ninguna división. El pivoteo total es consecuentemente la estrategia recomendada para la mayoría de los sistemas complicados para los cuales se puede justificar la cantidad de tiempo de ejecución tan intensa.

3. EJEMPLO DE ALGORITMO FORTRAN

En esta sección vamos a presentar una versión FORTRAN muy sencilla del algoritmo de eliminación Gaussiana con pivoteo máximo de columna. En el esquema de la programación estructurada FORTRAN, el problema de la búsqueda de solución de un sistema de ecuaciones lineales será desarrollado dividiéndolo en un programa principal y en varios subprogramas, donde cada uno de ellos resuelve una tarea particular. En nuestro caso, el

problema será resuelto usando un programa principal que llama a la subrutina MATRIZA, para la lectura de los elemento de la matriz ampliada A_a, correspondiente al sistema dado $A\,\mathbf{x} = \mathbf{b}$ y a las subrutinas GAUSELI, GAUSMAX o GAUSESC, dependiendo de qué método se quiere usar, para el desarrollo del algoritmo de eliminación Gaussiana sin pivoteo, la primera, con pivoteo máximo de columna, la segunda, y con pivoteo escalado de columna, la tercera. Aquí se dará solamente la versión FORTRAN de la subrutina GAUSMAX (las otras se pueden obtener de ésta con sencillas modificaciones).

```
C       PROGRAMA PRINCIPAL
        PROGRAM SISLIN
        PARAMETER (M = 20, MM = 21)
        REAL XX(M), AA(M, MM)
        INTEGER N, I, J, INDEX
        EXTERNAL MATRIZA, GAUSELI, GAUSMAX, GAUSESC
C
        PRINT*, 'NUMERO DE DIMENSION MAXIMA', M
        PRINT*, 'DAR LA DIMENSION DEL PROBLEMA'
        READ*, N
        PRINT*, 'ESCOGER EL METODO A USAR'
        PRINT*, 'INDEX = 0, ELIMINACION GAUSSIANA CON'
        PRINT*, '  SUSTITUCION HACIA ATRAS SIN PIVOTEO'
        PRINT*, 'INDEX = 1, ELIMINACION GAUSSIANA CON'
        PRINT*, '  PIVOTEO MAXIMO DE COLUMNA'
        PRINT*, 'INDEX = 2, ELIMINACION GAUSSIANA CON'
        PRINT*, '  PIVOTEO ESCALADO DE COLUMNA'
        READ*, INDEX
        IF(INDEX.EQ.0) PRINT*, ' NO USARE PIVOTEO'
        IF(INDEX.EQ.1) PRINT*, ' USARE PIVOTEO MAXIMO'
        IF(INDEX.EQ.2) PRINT*, ' USARE PIVOTEO ESCALADO'
C
        CALL MATRIZA(N, AA, M)
        IF(INDEX.EQ.0) CALL GAUSELI (N, AA, M, XX)
        IF(INDEX.EQ.1) CALL GAUSMAX (N, AA, M, XX)
        IF(INDEX.EQ.2) CALL GAUSESC (N, AA, M, XX)
C
        PRINT*, ' LA APROXIMACION A LA SOLUCION ES'
        DO 10 I = 1, N
10          PRINT*, XX(I)
        STOP
        END
    ************************************************************
        SUBROUTINE GAUSMAX (N, A, M, XX)
        PARAMETER (MM = 20)
        INTEGER I, J, K, N, IFIL(MM)
        REAL AA(M,*), XX(M), CHECK, CHECK1, MUL(MM,MM)
C
        DO 10 K = 1, N
10          IFIL(K) = K
C
        DO 99 I = 1, N - 1
          PRINT*, '     *** PASO NUMERO ***  I : ', I
```

```
          CHECK = ABS(A(IFIL(I), I))
          IP = I
        DO 20 J = I + 1, N
          CHECK1 = ABS(A(IFIL(J), I))
          IF(CHECK1.GT.CHECK) THEN
            CHECK = CHECK1
            IP = J
            PRINT*, ' HAY INTERCAMBIO DE I: ', I
            PRINT*, '              CON IP: ', IP
          ENDIF
20      CONTINUE
          IF(A(IFIL(IP), I).EQ.0.0) THEN
            PRINT*, ' NO EXISTE SOLUCION UNICA'
            GOTO 999
          ENDIF
          IF(IFIL(I).NE.IFIL(IP)) THEN
            AUX = IFIL(I)
            IFIL(I) = IFIL(IP)
            IFIL(IP) = AUX
          ENDIF
        DO 77 J = I + 1, N
          MUL(IFIL(J), I) = A(IFIL(J), I)/A(IFIL(I), I)
          PRINT*, ' MULTIPLICADOR'
          PRINT*, I, J, MUL(IFIL(J), I)
        DO 88 K = 1, N + 1
88        A(IFIL(J), K) = A(IFIL(J), K) - MUL(IFIL(J), I) * A(IFIL(I), K)
77      CONTINUE
        PRINT*, ((A(K, J) , J = 1, N + 1) , K = 1, N)
99    CONTINUE
C
      IF(A(IFIL(N), N).EQ.0.0) THEN
        PRINT*, ' NO EXISTE SOLUCION UNICA'
        GOTO 999
      ENDIF
C
        XX(N) = A(IFIL(N), N + 1)/A(IFIL(N), N)
      DO 55 I = N - 1, 1, -1
        SUMA = 0.0
        DO 44 J = I + 1, N
44        SUMA = SUMA + A(IFIL(I), J) * XX(J)
          XX(I) = (A(IFIL(I), N + 1) - SUMA)/A(IFIL(I), I)
55    CONTINUE
      PRINT*, ' EL PROCEDIMIENTO HA SIDO'
      PRINT*, ' COMPLETADO SATISFACTORIAMENTE'
999   CONTINUE
      RETURN
      END
*************************************************************
      SUBROUTINE MATRIZA (N, AA, M)
      INTEGER N, I, J, M
      REAL AA(M, *)
      OPEN (UNIT = 13, FILE =' IN.DAT')
C
```

```
      DO  10  I = 1,  N
        DO  10  J = 1,  N + 1
10          READ(13, *) AA(I,  J)
      CLOSE(13)
      RETURN
      END
```

Vamos ahora a presentar un nuevo ejemplo, cuya solución se obtendrá de dos maneras, por supuesto, equivalentes. Primero, seguiremos el algoritmo como si fuese un ordenador el que hace los cálculos, y después resolvemos el sistema con el algoritmo de eliminación Gaussiana con pivoteo máximo de columna, en la manera habitual como cuando se hacen *los cálculos a mano*. En ambos casos, los cálculos serán hechos en aritmética de cinco dígitos (los resultados obtenidos con un ordenador computando con representación numérica de precisión simple son ligeramente distintos).

Ejemplo 4.

El sistema lineal

$$\begin{array}{lrcrcrcl} E_1: & 1.5611\,x_1 & + & 5.1791\,x_2 & - & 1.6852\,x_3 & = & 8.4254\,, \\ E_2: & 3.3330\,x_1 & + & 15920\,x_2 & + & 10.333\,x_3 & = & 15913\,, \\ E_3: & 2.2220\,x_1 & + & 16.710\,x_2 & - & 9.6120\,x_3 & = & 28.544\,, \end{array}$$

tiene la solución exacta $x_1 = 1.0$, $x_2 = 1.0$ y $x_3 = -1.0$; resolvámoslo usando el método de eliminación Gaussiana con pivoteo máximo de columna y aritmética de cinco dígitos.

(a) En primer lugar resolvemos el sistema lineal siguiendo los pasos del algoritmo, como si fuese un ordenador el que efectúa los cálculos.

La matriz ampliada almacenada por el ordenador será:

$$A_a = \left(\begin{array}{ccc|c} 1.5611 & 5.1791 & -1.6852 & 8.4254 \\ 3.3330 & 15920 & 10.333 & 15913 \\ 2.2220 & 16.710 & -9.6120 & 28.544 \end{array}\right) .$$

Paso 1: *(Inicializar el indicador de la fila).* Para $i = 1, 2, \ldots, n$ tomar $F(i) = i$; es decir

$$F(1) = 1\,, \qquad F(2) = 2\,, \qquad F(3) = 3\,.$$

Paso 2: Para $i = 1, 2, \ldots, n-1$ ($i = 1, 2$) seguir los pasos 3–6 *(proceso de eliminación).* Empezamos con $i = 1$.

Paso 3: Sea p el menor entero con $i \le p \le n$ y

$$|a(F(p), i)| = \max_{i \le j \le n} |a(F(j), i)|\,.$$

Es decir, buscar el menor entero $1 \le p \le 3$ tal que

$$|a(F(p), 1)| = \max_{1 \le j \le 3} |a(F(j), 1)|\,.$$

$j = 1, \qquad |a(F(1), 1)| = |a_{11}| = 1.5611\,,$

$j = 2, \qquad |a(F(2),1)| = |a_{21}| = 3.3330 \ ,$
$j = 3, \qquad |a(F(3),1)| = |a_{31}| = 2.2220 \ .$
El valor máximo es el segundo; luego $p = 2$.

Paso 4: Si $a(F(p),i) = 0$ entonces SALIDA; *(no existe solución única)* PARAR.

Paso 5: Si $F(i) \neq F(p)$ entonces tomar $AUX = F(i)$, $F(i) = F(p)$, $F(p) = AUX$; *(intercambio de filas simulado).* En nuestro caso,

$$F(1) = 2 \ , \qquad F(2) = 1 \ , \qquad F(3) = 3 \ .$$

Paso 6: Para $j = i+1, i+2, \ldots, n$ $(j = 2, 3)$ seguir los pasos 7 y 8.

Paso 7: Tomar $m(F(j),i) = \frac{a(F(j),i)}{a(F(i),i)}$.
$j = 2, i = 1,$
$m(F(2),1) = m(1,1) = \frac{a(F(2),1)}{a(F(1),1)} = \frac{a_{11}}{a_{21}} = \frac{1.5611}{3.3330} = 0.46838 \ ,$
$j = 3, i = 1,$
$m(F(3),1) = m(3,1) = \frac{a(F(3),1)}{a(F(1),1)} = \frac{a_{31}}{a_{21}} = \frac{2.2220}{3.3330} = 0.66667 \ ,$

Paso 8: Efectuar $(E_{F(j)} - m(F(j),i)\ E_{F(i)}) \to (E_{F(j)})$.
$j = 2, i = 1, \qquad (E_{F(2)} - m(F(2),1)\ E_{F(1)}) \to (E_{F(2)}) \ ,$
$(E_1 - m(1,1)\ E_2) \to (E_1) \ ,$
$j = 3, i = 1, \qquad (E_{F(3)} - m(F(3),1)\ E_{F(1)}) \to (E_{F(3)}) \ ,$
$(E_3 - m(3,1)\ E_2) \to (E_3) \ ,$
entonces la matriz ampliada se transforma en:

$$A_a = \left(\begin{array}{ccc|c} 0.0000 & 5.1791 - 7456.6 & -1.6852 - 4.8398 & 8.4254 - 7453.3 \\ 3.3330 & 15920 & 10.333 & 15913 \\ 0.0000 & 16.710 - 10613 & -9.6120 - 6.8887 & 28.544 - 10609 \end{array}\right) =$$

$$= \left(\begin{array}{ccc|c} 0.0000 & -7451.4 & -6.525 & -7444.9 \\ 3.3330 & 15920 & 10.333 & 15913 \\ 0.0000 & -10596 & -16.501 & -10580 \end{array}\right) \ .$$

Paso 2: Continuamos con $i = 2$.

Paso 3: Sea p el menor entero con $i \leq p \leq n$ y

$$|a(F(p),i)| = \max_{i \leq j \leq n} |a(F(j),i)| \ .$$

Es decir, buscar el menor entero $2 \leq p \leq 3$ tal que

$$|a(F(p),2)| = \max_{2 \leq j \leq 3} |a(F(j),2)| \ .$$

$j = 2, \qquad |a(F(2),2)| = |a_{12}| = 7451.4 \ ,$
$j = 3, \qquad |a(F(3),2)| = |a_{32}| = 10596 \ .$
El valor máximo es el segundo; luego $p = 3$.

Paso 4: Si $a(F(p),i) = 0$ entonces SALIDA; *(no existe solución única)* PARAR.

Paso 5: Si $F(i) \neq F(p)$ entonces tomar $AUX = F(i)$, $F(i) = F(p)$, $F(p) = AUX$; *(intercambio de filas simulado).* En nuestro caso,

$$F(1) = 2 \ , \qquad F(2) = 3 \ , \qquad F(3) = 1 \ .$$

Paso 6: Para $j = i+1, i+2, \ldots, n$ $(j = 3)$ seguir los pasos 7 y 8.

Paso 7: Tomar $m(F(j), i) = \frac{a(F(j),i)}{a(F(i),i)}$.

$j = 3, i = 2,$

$$m(F(3), 2) = m(1, 2) = \frac{a(F(3),2)}{a(F(2),2)} = \frac{a_{12}}{a_{32}} = \frac{-7451.4}{-10596} = 0.70323 \ ,$$

Paso 8: Efectuar $(E_{F(j)} - m(F(j), i)\ E_{F(i)}) \to (E_{F(j)})$.

$j = 3, i = 2,$ $\qquad (E_{F(3)} - m(F(3), 2)\ E_{F(2)}) \to (E_{F(3)}) \ ,$

$$(E_1 - m(1, 2)\ E_3) \to (E_1) \ ,$$

entonces la matriz ampliada se transforma en:

$$A_a = \begin{pmatrix} 0.0000 & 0.0000 & -6.525 + 11.604 & | & -7444.9 + 7440.2 \\ 3.3330 & 15920 & 10.333 & | & 15913 \\ 0.0000 & -10596 & -16.501 & | & -10580 \end{pmatrix} =$$

$$= \begin{pmatrix} 0.0000 & 0.0000 & 5.0790 & | & -4.7000 \\ 3.3330 & 15920 & 10.333 & | & 15913 \\ 0.0000 & -10596 & -16.501 & | & -10580 \end{pmatrix} \ .$$

Paso 9: Si $a(F(n), n) = 0$ entonces SALIDA; *(no existe solución única)* PARAR.

Paso 10: *(Empieza la sustitución hacia atrás)*; tomar

$$x_n = \frac{a(F(n), n+1)}{a(F(n), n)} \ .$$

Entonces,

$$x_3 = \frac{a(F(3), 4)}{a(F(3), 3)} = \frac{a_{14}}{a_{13}} = \frac{-4.7000}{5.0790} = -0.92538 \ .$$

Paso 11: Para $i = n-1, n-2, \ldots, 1$ $(i = 2, 1)$ tomar

$$x_i = \frac{a(F(i), n+1) - \sum_{j=i+1}^{n} a(F(i), j)\ x_j}{a(F(i), i)} \ .$$

Entonces,

$$x_2 = \frac{a(F(2), 4) - a(F(2), 3)\ x_3}{a(F(2), 2)} = \frac{a_{34} - a_{33}\ x_3}{a_{32}} =$$

$$= \frac{-10580 - (-16.501)(-0.92538)}{-10596} =$$

$$= \frac{-10580 - 15.270}{-10596} = \frac{-10595}{-10596} = 0.99991 \ ,$$

$$x_1 = \frac{a(F(1), 4) - \sum_{j=2}^{3} a(F(1), j)\ x_j}{a(F(i), i)} =$$

$$= \frac{a(F(1), 4) - a(F(1), 2)\ x_2 - a(F(1), 3)\ x_3}{a(F(1), 1)} =$$

$$= \frac{a_{24} - a_{22}\ x_2 - a_{23}\ x_3}{a_{21}} =$$

$$= \frac{15913 - (15920)(0.99991) - (10.333)(-0.92538)}{3.3330} =$$

$$= \frac{15913 - 15919 + 9.5620}{3.3330} = \frac{3.5620}{3.3330} = 1.0687 \ .$$

Paso 12: SALIDA ($x_1 = 1.0687, x_2 = 0.99991, x_3 = -0.92538$);
(procedimiento completado satisfactoriamente) PARAR.

(b) Ahora resolvemos el sistema lineal de la manera más normal mediante el cómputo manual.

La matriz ampliada será:

$$A_a = \left(\begin{array}{ccc|c} 1.5611 & 5.1791 & -1.6852 & 8.4254 \\ 3.3330 & 15920 & 10.333 & 15913 \\ 2.2220 & 16.710 & -9.6120 & 28.544 \end{array}\right) .$$

En primer lugar, tenemos que buscar el menor entero $1 \leq p \leq 3$ tal que

$$|a_{p1}| = \max_{1 \leq j \leq 3} |a_{j1}| .$$

$j = 1, \qquad |a_{11}| = 1.5611 ,$
$j = 2, \qquad |a_{21}| = 3.3330 ,$
$j = 3, \qquad |a_{31}| = 2.2220 .$
El valor máximo es el segundo, entonces, $p = 2$, y tenemos que intercambiar la fila (E_1) con (E_2). La nueva matriz ampliada es

$$A_a = \left(\begin{array}{ccc|c} 3.3330 & 15920 & 10.333 & 15913 \\ 1.5611 & 5.1791 & -1.6852 & 8.4254 \\ 2.2220 & 16.710 & -9.6120 & 28.544 \end{array}\right) .$$

Para $j = 2, 3$ tomar $m_{ji} = \frac{a_{ji}}{a_{ii}}$.
$j = 2, i = 1, \qquad m_{21} = \frac{a_{21}}{a_{11}} = \frac{1.5611}{3.3330} = 0.46838 ,$
$j = 3, i = 1, \qquad m_{31} = \frac{a_{31}}{a_{11}} = \frac{2.2220}{3.3330} = 0.66667 ;$
y efectuar $(E_j - m_{ji}\, E_i) \to (E_j)$.
$j = 2, i = 1, \qquad (E_2 - m_{21}\, E_1) \to (E_2) ,$
$j = 3, i = 1, \qquad (E_3 - m_{31}\, E_1) \to (E_3) ,$
entonces la matriz ampliada se transforma en:

$$A_a = \left(\begin{array}{ccc|c} 3.3330 & 15920 & 10.333 & 15913 \\ 0.0000 & 5.1791 - 7456.6 & -1.6852 - 4.8398 & 8.4254 - 7453.3 \\ 0.0000 & 16.710 - 10613 & -9.6120 - 6.8887 & 28.544 - 10609 \end{array}\right) =$$

$$= \left(\begin{array}{ccc|c} 3.3330 & 15920 & 10.333 & 15913 \\ 0.0000 & -7451.4 & -6.525 & -7444.9 \\ 0.0000 & -10596 & -16.501 & -10580 \end{array}\right) .$$

Ahora tenemos que buscar el menor entero $2 \leq p \leq 3$ tal que

$$|a_{p2}| = \max_{2 \leq j \leq 3} |a_{j2}| .$$

$j = 2, \qquad |a_{22}| = 7451.4 ,$
$j = 3, \qquad |a_{32}| = 10596 .$

El valor máximo es el segundo, entonces, $p = 3$, y tenemos que intercambiar la fila (E_2) con (E_3). La nueva matriz ampliada es

$$A_a = \left(\begin{array}{ccc|c} 3.3330 & 15920 & 10.333 & 15913 \\ 0.0000 & -10596 & -16.501 & -10580 \\ 0.0000 & -7451.4 & -6.525 & -7444.9 \end{array}\right) .$$

Tomar $m_{ji} = \frac{a_{ji}}{a_{ii}}$.

$j = 3, i = 2, \qquad m_{32} = \frac{a_{32}}{a_{22}} = \frac{-7451.4}{-10596} = 0.70323$,

y efectuar $(E_j - m_{ji}\ E_i) \to (E_j)$.

$j = 3, i = 2, \qquad (E_3 - m_{32}\ E_2) \to (E_3)$,

entonces la matriz ampliada se transforma en:

$$A_a = \left(\begin{array}{ccc|c} 3.3330 & 15920 & 10.333 & 15913 \\ 0.0000 & -10596 & -16.501 & -10580 \\ 0.0000 & 0.0000 & -6.525 + 11.604 & -7444.9 + 7440.2 \end{array}\right) =$$

$$= \left(\begin{array}{ccc|c} 3.3330 & 15920 & 10.333 & 15913 \\ 0.0000 & -10596 & -16.501 & -10580 \\ 0.0000 & 0.0000 & 5.0790 & -4.7000 \end{array}\right) .$$

Finalmente,

$$x_3 = \frac{a_{34}}{a_{33}} = \frac{-4.7000}{5.0790} = -0.92538 ,$$

$$x_2 = \frac{a_{24} - a_{23}\ x_3}{a_{22}} = \frac{-10580 - (-16.501)(-0.92538)}{-10596} =$$

$$= \frac{-10580 - 15.270}{-10596} = \frac{-10595}{-10596} = 0.99991 ,$$

$$x_1 = \frac{a_{14} - a_{12}\ x_2 - a_{13}\ x_3}{a_{11}} =$$

$$= \frac{15913 - (15920)(0.99991) - (10.333)(-0.92538)}{3.3330} =$$

$$= \frac{15913 - 15919 + 9.5620}{3.3330} = \frac{3.5620}{3.3330} = 1.0687 .$$

4. EL ALGORITMO DE GAUSS-JORDAN

Como hemos visto en el método conocido como regla de Cramer (ver Capítulo XIII), para resolver el sistema lineal $A\ \mathbf{x} = \mathbf{b}$ se puede necesitar la matriz inversa A^{-1} para obtener $\mathbf{x} = A^{-1}\ \mathbf{b}$ como única solución del sistema. Sin embargo la inversa A^{-1} de una matriz $n \times n$ no singular A no se necesita a menudo, dado que existen otros métodos para resolver los sistemas lineales. De cualquier manera, el **algoritmo de Gauss-Jordan** nos da un método para invertir la aplicación $\mathbf{x} \longrightarrow A\ \mathbf{x} = \mathbf{y}$, $\mathbf{x} \in \mathcal{R}^n$, $\mathbf{y} \in \mathcal{R}^n$, de una manera sistemática.

Consideremos el sistema $A\ \mathbf{x} = \mathbf{y}$:

$$\begin{aligned} E_1 :&\quad a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = y_1 , \\ E_2 :&\quad a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = y_2 , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n :&\quad a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = y_n . \end{aligned} \qquad (XV.1)$$

En el primer paso del método de Gauss-Jordan, la variable x_1 se cambia por una de las variables y_r. Para hacer esto, se busca un coeficiente $a_{r1} \neq 0$, por ejemplo con el pivoteo máximo de columna:

$$|a_{r1}| = \max_{1 \leq i \leq n} |a_{i1}|$$

y las ecuaciones E_1 y E_r vienen intercambiadas, es decir, se hace un intercambio de filas $(E_1) \leftrightarrow (E_r)$. De esta manera se obtiene un sistema:

$$\begin{aligned}
E_1: &\quad \overline{a}_{11}\, x_1 + \overline{a}_{12}\, x_2 + \ldots + \overline{a}_{1n}\, x_n = \overline{y}_1\ , \\
E_2: &\quad \overline{a}_{21}\, x_1 + \overline{a}_{22}\, x_2 + \ldots + \overline{a}_{2n}\, x_n = \overline{y}_2\ , \\
&\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\
E_n: &\quad \overline{a}_{n1}\, x_1 + \overline{a}_{n2}\, x_2 + \ldots + \overline{a}_{nn}\, x_n = \overline{y}_n\ ,
\end{aligned} \tag{$XV.2$}$$

en el cual las variables $\overline{y}_1, \ldots, \overline{y}_n$ son permutaciones de $y_1, \ldots, y_n$, y además $\overline{a}_{11} = a_{r1}$, $\overline{y}_1 = y_r$. Ahora $\overline{a}_{11} \neq 0$, porque si no fuese así,tendríamos $a_{i1} = 0$ para todo i, con lo que A sería singular. Resolvamos la primera ecuación de ($XV.2$) para x_1, y sustituyamos el resultado en todas las demás ecuaciones del sistema. Entonces se obtiene el sistema:

$$\begin{aligned}
E_1: &\quad a'_{11}\, \overline{y}_1 + a'_{12}\, x_2 + \ldots + a'_{1n}\, x_n = x_1\ , \\
E_2: &\quad a'_{21}\, \overline{y}_1 + a'_{22}\, x_2 + \ldots + a'_{2n}\, x_n = \overline{y}_2\ , \\
&\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\
E_n: &\quad a'_{n1}\, \overline{y}_1 + a'_{n2}\, x_2 + \ldots + a'_{nn}\, x_n = \overline{y}_n\ ,
\end{aligned} \tag{$XV.3$}$$

donde, para todo $i, k = 2, 3, \ldots, n$,

$$\begin{aligned}
a'_{11} &= \frac{1}{\overline{a}_{11}}, & a'_{1k} &= -\frac{\overline{a}_{1k}}{\overline{a}_{11}}, \\
a'_{i1} &= \frac{\overline{a}_{i1}}{\overline{a}_{11}}, & a'_{ik} &= \overline{a}_{ik} - \overline{a}_{i1}\, \frac{\overline{a}_{1k}}{\overline{a}_{11}}.
\end{aligned}$$

En el paso siguiente, la variable x_2 se cambia con una de las variables $\overline{y}_2, \ldots, \overline{y}_n$; es decir, se busca $a_{r2} \neq 0$, tal que $|a_{r2}| = \max_{2 \leq i \leq n} |a_{i2}|$ y se hace un intercambio de filas $(E_2) \leftrightarrow (E_r)$; luego se resuelve la segunda ecuación para x_2, y se sustituye en todas las demás ecuaciones del sistema. Después se repite para las variables x_3 y para todas las demás. Si representamos los sistemas con sus matrices, partiendo de $A = A^{(0)}$, se obtiene una sucesión $A^{(0)} \to A^{(1)} \to \ldots \to A^{(n)}$. La matriz genérica $A^{(j)} = a^{(j)}_{ik}$ representa el *sistema mixto de ecuaciones* de la forma

$$\begin{aligned}
E_1: &\quad a^{(j)}_{11}\, \tilde{y}_1 + \ldots + a^{(j)}_{1j}\, \tilde{y}_j + a^{(j)}_{1,j+1}\, x_{j+1} + \ldots + a^{(j)}_{1n}\, x_n = x_1\ , \\
&\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\
E_j: &\quad a^{(j)}_{21}\, \tilde{y}_1 + \ldots + a^{(j)}_{jj}\, \tilde{y}_j + a^{(j)}_{j,j+1}\, x_{j+1} + \ldots + a^{(j)}_{jn}\, x_n = x_j\ , \\
E_{j+1}: &\quad a^{(j)}_{21}\, \tilde{y}_1 + \ldots + a^{(j)}_{j+1,j}\, \tilde{y}_j + a^{(j)}_{j+1,j+1}\, x_{j+1} + \ldots + a^{(j)}_{2n}\, x_n = \tilde{y}_{j+1}\ , \\
&\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\
E_n: &\quad a^{(j)}_{n1}\, \tilde{y}_1 + \ldots + a^{(j)}_{nj}\, \tilde{y}_j + a^{(j)}_{n,j+1}\, x_{j+1} + \ldots + a^{(j)}_{nn}\, x_n = \tilde{y}_n\ .
\end{aligned} \tag{$XV.4$}$$

En este sistema $(\tilde{y}_1, \ldots, \tilde{y}_n)$ indica una permutación de las variables originarias $(y_1, \ldots, y_n)$. En el paso $A^{(j-1)} \to A^{(j)}$ la variable x_j se intercambia por $\tilde{y}_j$. Entonces, se obtiene $A^{(j)}$ de $A^{(j-1)}$ según las reglas dadas abajo. Por simplicidad, los elementos de $A^{(j-1)}$ se indican con a_{ik}, y los elementos de $A^{(j)}$ con a'_{ik}.

Reglas para el algoritmo de Gauss-Jordan con pivoteo máximo de columna.

a) Determinar r como el menor entero $j \le r \le n$ tal que

$$|a_{rj}| = \max_{j \le i \le n} |a_{ij}| \ .$$

Si $a_{rj} = 0$, la matriz es singular y no hay solución.

b) Intercambiar las filas r y j de la matriz $A^{(j-1)}$ y llamar al resultado $\overline{A} = \overline{a}_{ik}$.

c) Calcular $A^{(j)} = a'_{ik}$, para $i, k \ne j$, según las fórmulas

$$a'_{jj} = \frac{1}{\overline{a}_{jj}}, \qquad a'_{jk} = -\frac{\overline{a}_{jk}}{\overline{a}_{jj}},$$
$$a'_{ij} = \frac{\overline{a}_{ij}}{\overline{a}_{jj}}, \qquad a'_{ik} = \overline{a}_{ik} - \overline{a}_{ij}\ \frac{\overline{a}_{jk}}{\overline{a}_{jj}}.$$

El sistema $(XV.4)$ implica que

$$A^{(n)}\ \hat{\mathbf{y}} = \mathbf{x}\ , \qquad \hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^t$$

donde $(\hat{y}_1, \ldots, \hat{y}_n)$ es una permutación de las variables originales $(y_1, \ldots, y_n)$; es decir, $\hat{\mathbf{y}} = P\ \mathbf{y}$ que corresponde a los intercambios de filas hechos en el paso b) del algoritmo de Gauss-Jordan, y puede ser fácilmente determinado. Entonces, $A^{(n)}\ \hat{\mathbf{y}} = A^{(n)}\ P\ \mathbf{y} = \mathbf{x}$ además de $A\ \mathbf{x} = \mathbf{y}$, lo que implica

$$A^{-1} = A^{(n)}\ P\ .$$

Ejemplo 5.

Encontrar la matriz inversa A^{-1} de la matriz

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 6 \end{pmatrix}\ .$$

Para hallar la matriz inversa, aplicamos el método de Gauss-Jordan sin ningún pivoteo (sin intercambios de filas). Entonces la matriz $A^{(0)}$ es la misma matriz A:

$$A^{(0)} = A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 6 \end{pmatrix}\ .$$

Dado que no hay pivoteo, la matriz $\overline{A}$ es la misma matriz $A^{(0)}$; entonces construímos la matriz $A^{(1)}$ de la siguiente manera:

$$A^{(1)} = \begin{pmatrix} \dfrac{1}{\overline{a}_{11}} & -\dfrac{\overline{a}_{12}}{\overline{a}_{11}} & -\dfrac{\overline{a}_{13}}{\overline{a}_{11}} \\ \dfrac{\overline{a}_{21}}{\overline{a}_{11}} & \overline{a}_{22} - \overline{a}_{21}\ \dfrac{\overline{a}_{12}}{\overline{a}_{11}} & \overline{a}_{23} - \overline{a}_{21}\ \dfrac{\overline{a}_{13}}{\overline{a}_{11}} \\ \dfrac{\overline{a}_{31}}{\overline{a}_{11}} & \overline{a}_{32} - \overline{a}_{31}\ \dfrac{\overline{a}_{12}}{\overline{a}_{11}} & \overline{a}_{33} - \overline{a}_{31}\ \dfrac{\overline{a}_{13}}{\overline{a}_{11}} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 \\ 1 & 1 & 2 \\ 1 & 2 & 5 \end{pmatrix}\ .$$

Como antes, dado que no hay pivoteo, la matriz $\overline{A}$ es la misma matriz $A^{(1)}$; entonces construímos la matriz $A^{(2)}$ de la siguiente manera:

$$A^{(2)} = \begin{pmatrix} \overline{a}_{11} - \overline{a}_{12}\,\dfrac{\overline{a}_{21}}{\overline{a}_{22}} & \dfrac{\overline{a}_{12}}{\overline{a}_{22}} & \overline{a}_{13} - \overline{a}_{12}\,\dfrac{\overline{a}_{23}}{\overline{a}_{22}} \\ -\dfrac{\overline{a}_{21}}{\overline{a}_{22}} & \dfrac{1}{\overline{a}_{22}} & -\dfrac{\overline{a}_{23}}{\overline{a}_{22}} \\ \overline{a}_{31} - \overline{a}_{32}\,\dfrac{\overline{a}_{21}}{\overline{a}_{22}} & \dfrac{\overline{a}_{32}}{\overline{a}_{22}} & \overline{a}_{33} - \overline{a}_{32}\,\dfrac{\overline{a}_{23}}{\overline{a}_{22}} \end{pmatrix} = \begin{pmatrix} 2 & -1 & 1 \\ -1 & 1 & -2 \\ -1 & 2 & 1 \end{pmatrix} .$$

De nuovo, la matriz $\overline{A}$ es la misma matriz $A^{(2)}$; entonces construímos la matriz $A^{(3)}$ de la siguiente manera:

$$A^{(3)} = \begin{pmatrix} \overline{a}_{11} - \overline{a}_{13}\,\dfrac{\overline{a}_{31}}{\overline{a}_{33}} & \overline{a}_{12} - \overline{a}_{13}\,\dfrac{\overline{a}_{32}}{\overline{a}_{33}} & \dfrac{\overline{a}_{13}}{\overline{a}_{33}} \\ \overline{a}_{21} - \overline{a}_{23}\,\dfrac{\overline{a}_{31}}{\overline{a}_{33}} & \overline{a}_{22} - \overline{a}_{23}\,\dfrac{\overline{a}_{32}}{\overline{a}_{33}} & \dfrac{\overline{a}_{23}}{\overline{a}_{33}} \\ -\dfrac{\overline{a}_{31}}{\overline{a}_{33}} & -\dfrac{\overline{a}_{32}}{\overline{a}_{33}} & \dfrac{1}{\overline{a}_{33}} \end{pmatrix} = \begin{pmatrix} 3 & -3 & 1 \\ -3 & 5 & -2 \\ 1 & -2 & 1 \end{pmatrix} = A^{-1} .$$

Ejemplo 6.

Volvemos al ejemplo 4 y resolvemos el sistema lineal

$$\begin{array}{lrcrcrcl} E_1: & 1.5611\,x_1 & + & 5.1791\,x_2 & - & 1.6852\,x_3 & = & 8.4254\,, \\ E_2: & 3.3330\,x_1 & + & 15920\,x_2 & + & 10.333\,x_3 & = & 15913\,, \\ E_3: & 2.2220\,x_1 & + & 16.710\,x_2 & - & 9.6120\,x_3 & = & 28.544\,, \end{array}$$

con el método de Gauss-Jordan con pivoteo máximo de columna y aritmética de cinco dígitos.

La primera matriz $A^{(0)}$ será:

$$A^{(0)} = \begin{pmatrix} 1.5611 & 5.1791 & -1.6852 \\ 3.3330 & 15920 & 10.333 \\ 2.2220 & 16.710 & -9.6120 \end{pmatrix} ;$$

para determinar la matriz $\overline{A}$, tenemos que buscar el menor entero $1 \le p \le 3$ tal que

$$|a_{p1}| = \max_{1 \le i \le 3} |a_{i1}| ,$$

que en este caso es $p = 2$, y tenemos que intercambiar la fila (E_1) con (E_2). Entonces

$$\overline{A} = \begin{pmatrix} 3.3330 & 15920 & 10.333 \\ 1.5611 & 5.1791 & -1.6852 \\ 2.2220 & 16.710 & -9.6120 \end{pmatrix} \qquad \text{y} \qquad \hat{\mathbf{y}} = (15913, 8.4254, 28.544)^t .$$

Ahora calculamos la matriz $A^{(1)} = (a'_{ik})$ según:

$$\begin{aligned} A^{(1)} &= \begin{pmatrix} \dfrac{1}{\overline{a}_{11}} & -\dfrac{\overline{a}_{12}}{\overline{a}_{11}} & -\dfrac{\overline{a}_{13}}{\overline{a}_{11}} \\ \dfrac{\overline{a}_{21}}{\overline{a}_{11}} & \overline{a}_{22} - \overline{a}_{21}\,\dfrac{\overline{a}_{12}}{\overline{a}_{11}} & \overline{a}_{23} - \overline{a}_{21}\,\dfrac{\overline{a}_{13}}{\overline{a}_{11}} \\ \dfrac{\overline{a}_{31}}{\overline{a}_{11}} & \overline{a}_{32} - \overline{a}_{31}\,\dfrac{\overline{a}_{12}}{\overline{a}_{11}} & \overline{a}_{33} - \overline{a}_{31}\,\dfrac{\overline{a}_{13}}{\overline{a}_{11}} \end{pmatrix} \\ &= \begin{pmatrix} 0.30003 & -4776.5 & -3.1002 \\ 0.46838 & -7451.4 & -6.5250 \\ 0.66667 & -10596 & -16.501 \end{pmatrix} . \end{aligned}$$

Ahora tenemos que buscar el menor entero $2 \leq p \leq 3$ tal que

$$|a_{p2}| = \max_{2 \leq i \leq 3} |a_{i2}| \ ;$$

y en nuestro caso $p = 3$, y tenemos que intercambiar la fila (E_2) con (E_3). Entonces

$$\overline{A} = \begin{pmatrix} 0.30003 & -4776.5 & -3.1002 \\ 0.66667 & -10596 & -16.501 \\ 0.46838 & -7451.4 & -6.5250 \end{pmatrix} \qquad \text{y} \qquad \hat{\mathbf{y}} = (15913, 28.544, 8.4254)^t \ .$$

Ahora calculamos la matriz $A^{(2)} = (a'_{ik})$ según:

$$\begin{aligned} A^{(2)} &= \begin{pmatrix} \overline{a}_{11} - \overline{a}_{12} \dfrac{\overline{a}_{21}}{\overline{a}_{22}} & \dfrac{\overline{a}_{12}}{\overline{a}_{22}} & \overline{a}_{13} - \overline{a}_{12} \dfrac{\overline{a}_{23}}{\overline{a}_{22}} \\ -\dfrac{\overline{a}_{21}}{\overline{a}_{22}} & \dfrac{1}{\overline{a}_{22}} & -\dfrac{\overline{a}_{23}}{\overline{a}_{22}} \\ \overline{a}_{31} - \overline{a}_{32} \dfrac{\overline{a}_{21}}{\overline{a}_{22}} & \dfrac{\overline{a}_{32}}{\overline{a}_{22}} & \overline{a}_{33} - \overline{a}_{32} \dfrac{\overline{a}_{23}}{\overline{a}_{22}} \end{pmatrix} \\ &= \begin{pmatrix} -0.49 * 10^{-3} & 0.45078 & 4.3382 \\ 0.62917 * 10^{-4} & -0.94375 * 10^{-4} & -0.15573 * 10^{-2} \\ -0.44 * 10^{-3} & 0.70323 & 5.0790 \end{pmatrix} \ . \end{aligned}$$

Ahora la matriz $\overline{A}$ es la misma matriz $A^{(2)}$; entonces construímos la matriz $A^{(3)}$ de la siguiente manera:

$$\begin{aligned} A^{(3)} &= \begin{pmatrix} \overline{a}_{11} - \overline{a}_{13} \dfrac{\overline{a}_{31}}{\overline{a}_{33}} & \overline{a}_{12} - \overline{a}_{13} \dfrac{\overline{a}_{32}}{\overline{a}_{33}} & \dfrac{\overline{a}_{13}}{\overline{a}_{33}} \\ \overline{a}_{21} - \overline{a}_{23} \dfrac{\overline{a}_{31}}{\overline{a}_{33}} & \overline{a}_{22} - \overline{a}_{23} \dfrac{\overline{a}_{32}}{\overline{a}_{33}} & \dfrac{\overline{a}_{23}}{\overline{a}_{33}} \\ -\dfrac{\overline{a}_{31}}{\overline{a}_{33}} & -\dfrac{\overline{a}_{32}}{\overline{a}_{33}} & \dfrac{1}{\overline{a}_{33}} \end{pmatrix} \\ &= \begin{pmatrix} -0.11438 * 10^{-3} & -0.14954 & 0.85414 \\ 0.62895 * 10^{-4} & -0.58724 * 10^{-4} & -0.30667 * 10^{-4} \\ 0.86631 * 10^{-4} & -0.13846 & 0.19689 \end{pmatrix} \ . \end{aligned}$$

Así tenemos la solución $\mathbf{x} = A^{(3)}\hat{\mathbf{y}}$, con $\hat{\mathbf{y}} = (15913, 28.544, 8.4254)$; es decir

$$\begin{aligned} x_1 &= -0.11438 * 10^{-3} * 15913 - 0.14954 * 28.544 + 0.85414 * 8.4254 = \\ &= -1.8201 - 4.2685 + 7.1965 = 1.1079 \ , \\ x_2 &= 0.62895 * 10^{-4} * 15913 - 0.58724 * 10^{-4} * 28.544 - 0.30667 * 10^{-4} * 8.4254 = \\ &= 1.0008 - 0.16762 * 10^{-2} - 0.25838 * 10^{-3} = 0.99886 \ , \\ x_3 &= 0.86631 * 10^{-4} * 15913 - 0.13846 * 28.544 + 0.19689 * 8.4254 = \\ &= 1.3786 - 3.9522 + 1.6589 = -0.91470 \ . \end{aligned}$$

Entonces,

$$\mathbf{x} = (1.1079, 0.99886, -0.91470)$$

es la solución aproximada de la solución exacta $(1.0, 1.0, -1.0)$ obtenida con el método de Gauss-Jordan.

En la práctica, cuando se hacen los cálculos para resolver *a mano* un sistema de ecuaciones lineal, no se construyen las matrices $A^{(k)}$, si no que se trabaja directamente sobre el sistema. Mostraremos esta manera de proceder en el ejemplo siguiente.

Ejemplo 7.

Reconsideremos el ejemplo 3 del Capítulo XIII y resolvemos el sistema lineal

$$\begin{array}{lrcrcrcl} E_1: & x_1 & + & 2\,x_2 & - & x_3 & = & 2\,, \\ E_2: & 2\,x_1 & + & x_2 & & & = & 3\,, \\ E_3: & -\,x_1 & + & x_2 & + & 2\,x_3 & = & 4\,, \end{array}$$

con el método de Gauss-Jordan con pivoteo escalado de columna y aritmética de tres dígitos.

En el Capítulo XIII hemos visto que la solución de este sistema es

$$x_1 = \frac{7}{9}\,, \qquad x_2 = \frac{13}{9}\,, \qquad x_3 = \frac{15}{9}\,.$$

En el primer paso del método de Gauss-Jordan, la variable x_1 se cambia por una de las variables y_r. Para hacer esto, se busca un coeficiente $a_{r1} \neq 0$, con el pivoteo escalado de columna:

$$s_i = \max_{1\leq j\leq 3} |a_{ij}|\,, \qquad \frac{|a_{r1}|}{s_r} = \max_{1\leq i\leq 3} \frac{|a_{i1}|}{s_i}\,,$$

y las ecuaciones E_1 y E_r vienen intercambiadas, $(E_1) \leftrightarrow (E_r)$. En nuestro caso

$$s_1 = s_2 = s_3 = 2\,,$$

$$\frac{|a_{11}|}{s_1} = \frac{1.0}{2.0} = 0.5\,, \qquad \frac{|a_{21}|}{s_2} = \frac{2.0}{2.0} = 1.0\,, \qquad \frac{|a_{31}|}{s_3} = \frac{1.0}{2.0} = 0.5\,,$$

y así tenemos que intercambiar las primera y la segunda ecuación, y también tenemos que intercambiar los factores de escala, aunque en este caso quedan iguales $s_1 = s_2 = s_3 = 2$. De esta manera se obtiene el sistema:

$$\begin{array}{rcrcrcl} 2\,x_1 & + & x_2 & & & = & 3\,, \\ x_1 & + & 2\,x_2 & - & x_3 & = & 2\,, \\ -\,x_1 & + & x_2 & + & 2\,x_3 & = & 4\,. \end{array}$$

Ahora, resolvemos la primera ecuación por x_1, y sustituyemos el resultado en todas las demás ecuaciones:

$$\begin{array}{rcrcrcrcl} 1.5 & - & 0.5\,x_2 & & & & & = & x_1\,, \\ (1.5 & - & 0.5\,x_2) & + & 2\,x_2 & - & x_3 & = & 2\,, \\ (-1.5 & + & 0.5\,x_2) & + & x_2 & + & 2\,x_3 & = & 4\,. \end{array}$$

Entonces,

$$\begin{array}{rcrcrcl} 1.5 & - & 0.5\,x_2 & & & = & x_1\,, \\ 1.5 & + & 1.5\,x_2 & - & x_3 & = & 2\,, \\ -1.5 & + & 1.5\,x_2 & + & 2\,x_3 & = & 4\,. \end{array}$$

Ahora aplicamos otra vez el pivoteo escalado de columna:

$$\frac{|a_{22}|}{s_2} = \frac{1.5}{2.0} = 0.75 \ , \qquad \frac{|a_{32}|}{s_3} = \frac{1.5}{2.0} = 0.75 \ ,$$

y así no hay que intercambiar ecuaciones. Entonces, podemos resolver la segunda ecuación por x_2, y sustituir el resultado en las demás:

$$\begin{array}{rcrcrcl} 1.5 & - & 0.5\,(0.333 + 0.667\,x_3) & & & = & x_1 \ , \\ 0.333 & & & + & 0.667\,x_3 & = & x_2 \ , \\ -1.5 & + & 1.5\,(0.333 + 0.667\,x_3) & + & 2\,x_3 & = & 4 \ , \end{array}$$

que nos da

$$\begin{array}{rcrcl} 1.33 & - & 0.334\,x_3 & = & x_1 \ , \\ 0.333 & + & 0.667\,x_3 & = & x_2 \ , \\ -1.00 & + & 3\,x_3 & = & 4 \ . \end{array}$$

Finalmente, resolvemos la tercera ecuación por la variable x_3, y sustituyamos el resultado en las demás,

$$\begin{array}{rcrcll} 3\,x_3 & = & 5 \ , & \Rightarrow & x_3 = \frac{5}{3} = 1.67 \\ 1.33 & - & 0.334\,(1.67) & = & x_1 \ , \\ 0.333 & + & 0.667\,(1.67) & = & x_2 \ , \end{array}$$

para obtener la solución

$$x_1 = 0.772 \ , \qquad x_2 = 1.44 \ , \qquad x_3 = 1.67 \ ,$$

que es una buena aproximación de la solución exacta

$$x_1 = \frac{7}{9} \ , \qquad x_2 = \frac{13}{9} \ , \qquad x_3 = \frac{15}{9} \ .$$

EJERCICIOS.

1. Resolver los siguientes sistemas lineales con aritmética de redondeo de dos dígitos y i) eliminación Gaussiana con sustitución hacia atrás, ii) eliminación Gaussiana con pivoteo máximo de columna, iii) eliminación Gaussiana con pivoteo escalado de columna, iv) el algoritmo de Gauss-Jordan. Luego resolverlos con aritmética exacta y determinar cuál de i), ii), iii) o iv) es el más exacto.

a)
$$\begin{aligned} x_1 + 2\,x_2 + 3\,x_3 &= 1\,, \\ 2\,x_1 + 3\,x_2 + 4\,x_3 &= -1\,, \\ 3\,x_1 + 4\,x_2 + 6\,x_3 &= 2\,. \end{aligned}$$

b)
$$\begin{aligned} 2\,x_1 + 4\,x_2 - x_3 &= 11\,, \\ x_1 + x_2 - 3\,x_3 &= -5\,, \\ 4\,x_1 + x_2 + 2\,x_3 &= -9\,. \end{aligned}$$

c)
$$\begin{aligned} 0.1\,x_1 + 0.2\,x_2 + 0.4\,x_3 &= 1.1\,, \\ 4\,x_1 + x_2 - x_3 &= 6\,, \\ 2\,x_1 + 5\,x_2 + 2\,x_3 &= 3\,. \end{aligned}$$

d)
$$\begin{aligned} 0.04\,x_1 + 0.01\,x_2 - 0.01\,x_3 &= 0.06\,, \\ 0.2\,x_1 + 0.5\,x_2 - 0.2\,x_3 &= 0.3\,, \\ x_1 + 2\,x_2 + 4\,x_3 &= 11\,. \end{aligned}$$

e)
$$\begin{aligned} 2\,x_1 - x_2 + x_3 &= -1\,, \\ 3\,x_1 + 3\,x_2 + 9\,x_3 &= 0\,, \\ 3\,x_1 + 3\,x_2 + 5\,x_3 &= 4\,. \end{aligned}$$

2. Resolver los siguientes sistemas lineales con aritmética de redondeo de tres dígitos y i) eliminación Gaussiana con sustitución hacia atrás, ii) eliminación Gaussiana con pivoteo máximo de columna, iii) eliminación Gaussiana con pivoteo escalado de columna, iv) el algoritmo de Gauss-Jordan. Luego resolverlos con aritmética exacta y determinar cuál de i), ii), iii) o iv) es el más exacto.

a)
$$\begin{aligned} 0.832\,x_1 + 0.448\,x_2 + 0.193\,x_3 &= 1.00\,, \\ 0.784\,x_1 + 0.421\,x_2 - 0.207\,x_3 &= 0.00\,, \\ 0.784\,x_1 - 0.421\,x_2 + 0.279\,x_3 &= 0.00\,. \end{aligned}$$

b)
$$\begin{aligned} 0.03\,x_1 + 58.9\,x_2 &= 59.2\,, \\ 5.31\,x_1 - 6.10\,x_2 &= 47.0\,. \end{aligned}$$

c)
$$\begin{aligned} 58.9\,x_1 + 0.03\,x_2 &= 59.2\,, \\ -6.10\,x_1 + 5.31\,x_2 &= 47.0\,. \end{aligned}$$

d)
$$\begin{aligned} \pi\,x_1 - e\,x_2 + \sqrt{2}\,x_3 - \sqrt{3}\,x_4 &= 1\,, \\ \pi^2\,x_1 + e\,x_2 - e^2\,x_3 + \tfrac{3}{7}\,x_4 &= -1\,, \\ \sqrt{5}\,x_1 - \sqrt{6}\,x_2 + x_3 - 1.1\,x_4 &= 0\,, \\ \pi^3\,x_1 + e^2\,x_2 - \sqrt{7}\,x_3 + x_4 &= \sqrt{2}\,. \end{aligned}$$

3. Resolver los siguientes sistemas lineales con aritmética de redondeo de cuatro dígitos y i) eliminación Gaussiana con sustitución hacia atrás, ii) eliminación Gaussiana con pivoteo máximo de columna, iii) eliminación Gaussiana con pivoteo escalado de columna, iv) el algoritmo de Gauss-Jordan. Luego resolverlos con aritmética exacta y determinar cuál de i), ii), iii) o iv) es el más exacto.

$$
\begin{array}{lrcrcr}
a) & 58.09\, x_1 & + & 1.003\, x_2 & = & 68.12\ , \\
 & 321.8\, x_1 & + & 5.550\, x_2 & = & 377.3\ .
\end{array}
$$

$$
\begin{array}{lrcrcr}
b) & 1.003\, x_1 & + & 58.09\, x_2 & = & 68.12\ , \\
 & 321.8\, x_1 & + & 5.550\, x_2 & = & 377.3\ .
\end{array}
$$

4. Dadas las siguientes matrices A y vectores $\mathbf{b}$, determinar si los sistemas de ecuaciones lineales $A\,\mathbf{x} = \mathbf{b}$ tienen solución.

$$
a) \quad A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 1 \\ -2 \\ 0 \\ 4 \end{pmatrix} \ ;
$$

$$
b) \quad A = \begin{pmatrix} 2 & 1 & 3 \\ -1 & 0 & 2 \\ 6 & 2 & 2 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \ ;
$$

$$
c) \quad A = \begin{pmatrix} 3 & -2 & 1 \\ 1 & -3 & 1 \\ 1 & 1 & -1 \\ 3 & 0 & 1 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 2 \\ 5 \\ -5 \\ 0 \end{pmatrix} \ ;
$$

$$
d) \quad A = \begin{pmatrix} 2 & 1 & 3 \\ -1 & 0 & 2 \\ 6 & 2 & 2 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \ .
$$

CAPITULO XVI. FACTORIZACION DIRECTA DE MATRICES

1. INTRODUCCION Y METODO

La discusión centrada alrededor del Teorema XIII.6 se refirió a la factorización de una matriz A en términos de una matriz triangular inferior L y de una matriz triangular superior U. Esta factorización existe cuando se puede resolver de manera única el sistema lineal $A\,\mathbf{x} = \mathbf{b}$ por eliminación Gaussiana sin intercambios de filas o columnas. El sistema $L\,U\,\mathbf{x} = A\,\mathbf{x} = \mathbf{b}$ puede transformarse entonces en el sistema $U\,\mathbf{x} = L^{-1}\,\mathbf{b}$ y como U es triangular superior, se puede aplicar una sustitución hacia atrás. Aún cuando las formas específicas de L y U se pueden obtener del proceso de eliminación Gaussiana, es deseable encontrar un método más directo para su determinación, para que, si fuera necesaria la solución de varios sistemas usando A, sólo se necesitaría realizar una sustitución hacia adelante y otra hacia atrás. Para ilustrar un procedimiento para calcular los elementos de estas matrices, consideremos un ejemplo.

Ejemplo 1.

Considere la matriz estrictamente dominante diagonalmente de 4×4:

$$A = \begin{pmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{pmatrix} .$$

Los Teoremas XIII.6 y XIII.8 garantizan que A se puede factorizar en la forma $A = L\,U$, donde:

$$L = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} .$$

Los 16 elementos conocidos de A se pueden usar para determinar parcialmente los diez elementos desconocidos de L y el mismo número de U. Sin embargo si el procedimiento nos debe llevar a una solución única, se necesitan cuatro condiciones adicionales para los elementos de L y de U. El método a usar en este ejemplo consiste en requerir arbitrariamente que $l_{11} = l_{22} = l_{33} = l_{44} = 1$, y se conoce como el **método de Doolittle**. Más adelante en este capítulo, se considerarán métodos que requieren que todos los elementos de la diagonal de U sean uno (**método de Crout**) y que $l_{ii} = u_{ii}$ para cada valor de i (**método de Choleski**).

La parte de la multiplicación de L con U,

$$L\,U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} =$$

$$= \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = A\ ,$$

que determina la primera fila de A, da lugar a las cuatro ecuaciones

$$u_{11} = 6 \ , \quad u_{12} = 2 \ , \quad u_{13} = 1 \ , \quad u_{14} = -1 \ .$$

La parte de la multiplicación de L con U que determina los elementos restantes de la primera columna de A da las ecuaciones

$$l_{21}\ u_{11} = 2 \ , \quad l_{31}\ u_{11} = 1 \ , \quad l_{41}\ u_{11} = -1 \ ,$$

y entonces

$$l_{21} = \frac{1}{3} \ , \quad l_{31} = \frac{1}{6} \ , \quad l_{41} = -\frac{1}{6} \ .$$

Hasta aquí las matrices L y U asumen la forma:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/3 & 1 & 0 & 0 \\ 1/6 & l_{32} & 1 & 0 \\ -1/6 & l_{42} & l_{43} & 1 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 6 & 2 & 1 & -1 \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} \ .$$

La parte de la multiplicación que determina los elementos restantes en la segunda fila de A lleva a las ecuaciones

$$\begin{aligned} l_{21}\ u_{12} + u_{22} &= \frac{2}{3} + u_{22} = 4 \ , \\ l_{21}\ u_{13} + u_{23} &= \frac{1}{3} + u_{23} = 1 \ , \\ l_{21}\ u_{14} + u_{24} &= -\frac{1}{3} + u_{24} = 0 \ , \end{aligned}$$

así que

$$u_{22} = \frac{10}{3} \ , \quad u_{23} = \frac{2}{3} \ , \quad u_{24} = \frac{1}{3} \ ;$$

y la que determina los elementos restantes de la segunda columna de A da

$$\begin{aligned} l_{31}\ u_{12} + l_{32}\ u_{22} &= \frac{2}{6} + \frac{10}{3}\ l_{32} = 1 \ , \\ l_{41}\ u_{12} + l_{32}\ u_{22} &= -\frac{2}{6} + \frac{10}{3}\ l_{42} = 0 \ , \end{aligned}$$

así que

$$l_{32} = \frac{1}{5} \ , \quad l_{42} = \frac{1}{10} \ .$$

Ahora las matrices L y U tienen la forma:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/3 & 1 & 0 & 0 \\ 1/6 & 1/5 & 1 & 0 \\ -1/6 & 1/10 & l_{43} & 1 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 6 & 2 & 1 & -1 \\ 0 & 10/3 & 2/3 & 1/3 \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} \ .$$

La parte de la multiplicación que determina los elementos restantes en la tercera fila de A lleva a las ecuaciones

$$l_{31}\ u_{13} + l_{32}\ u_{23} + u_{33} = \frac{1}{6} + \frac{1}{5}\cdot\frac{2}{3} + u_{33} = 4\ ,$$
$$l_{31}\ u_{14} + l_{32}\ u_{24} + u_{34} = -\frac{1}{6} + \frac{1}{5}\cdot\frac{1}{3} + u_{34} = -1\ ,$$

así que

$$u_{33} = \frac{37}{10} \quad \text{y} \quad u_{34} = -\frac{9}{10}\ ;$$

y la que determina los elementos restantes de la tercera columna de A da

$$l_{41}\ u_{13} + l_{42}\ u_{23} + l_{43}\ u_{33} = -\frac{1}{6} + \frac{1}{10}\ \frac{2}{3} + \frac{37}{10}\ l_{43} = -1\ ,$$

así que

$$l_{43} = -\frac{9}{37}\ .$$

Y finalmente, la última ecuación es:

$$l_{41}\ u_{14} + l_{42}\ u_{24} + l_{43}\ u_{34} + u_{44} = -\frac{1}{6}(-1) + \frac{1}{10}\cdot\frac{1}{3} - \frac{9}{37}\ (-\frac{9}{10}) + u_{44} = 3\ ,$$

así que

$$u_{4u} = \frac{191}{74}\ ;$$

para obtener finalmente:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{3} & 1 & 0 & 0 \\ \frac{1}{6} & \frac{1}{5} & 1 & 0 \\ -\frac{1}{6} & \frac{1}{10} & -\frac{9}{37} & 1 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 6 & 2 & 1 & -1 \\ 0 & \frac{10}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{37}{10} & -\frac{9}{10} \\ 0 & 0 & 0 & \frac{191}{74} \end{pmatrix}\ .$$

2. LOS ALGORITMOS DE DOOLITTLE Y DE CROUT

En el siguiente algoritmo de factorización directa está contenido un procedimiento general para **factorizar matrices** en un producto de **matrices triangulares**. Aunque se construyen nuevas matrices L y U, los valores generados pueden reemplazar a los elementos correspondientes de A que no son ya necesarios. Por lo tanto, la nueva matriz tiene elementos $a_{ij} = l_{ij}$ para cada $i = 2, 3, \ldots, n$ y $j = 1, 2, 3, \ldots, i-1$; y $a_{ij} = u_{ij}$ para cada $i = 1, 2, 3, \ldots, n$ y $j = i, i+1, \ldots, n$.

Algoritmo de factorización directa de Doolittle o de Crout.

==========

Para factorizar una matriz $A = (a_{ij})$ de $n \times n$ en el producto de la matriz triangular inferior $L = (l_{ij})$ con la matriz triangular superior $U = (u_{ij})$; esto es, $A = L\ U$, donde está dada la diagonal principal de L ó U.

Entrada: dimensión n; los elementos $a_{ij}, 1 \leq i, j \leq n$ de A; la diagonal $l_{11}, l_{22}, \ldots, l_{nn}$ de L (método de Doolittle) ó $u_{11}, u_{22}, \ldots, u_{nn}$ de U (método de Crout).

Salida: los elementos l_{ij}, $1 \leq j \leq i$, $1 \leq i \leq n$ de L y los elementos u_{ij}, $1 \leq i \leq n$, $i \leq j \leq n$, de U.

Paso 1: Seleccionar l_{11} y u_{11} satisfaciendo $l_{11}\ u_{11} = a_{11}$.
Si $l_{11}\ u_{11} = 0$ entonces SALIDA; *(factorización imposible)* PARAR.

Paso 2: Para $j = 2, 3, \ldots, n$ tomar
$u_{1j} = \frac{a_{1j}}{l_{11}}$; *(primera fila de U)*;
$l_{j1} = \frac{a_{j1}}{u_{11}}$; *(primera columna de L)*.

Paso 3: Para $i = 2, 3, \ldots, n-1$ seguir los pasos 4 y 5.

Paso 4: Seleccionar l_{ii} y u_{ii} satisfaciendo $l_{ii}\ u_{ii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik}\ u_{ki}$.
Si $l_{ii}\ u_{ii} = 0$ entonces SALIDA; *(factorización imposible)*
PARAR.

Paso 5: Para $j = i+1, i+2, \ldots, n$ tomar
$u_{ij} = \frac{1}{l_{ii}}\ [a_{ij} - \sum_{k=1}^{i-1} l_{ik}\ u_{kj}]$; *(i−ésima fila de U)*;
$l_{ji} = \frac{1}{u_{ii}}\ [a_{ji} - \sum_{k=1}^{i-1} l_{jk}\ u_{ki}]$; *(i−ésima columna de L)*.

Paso 6: Seleccionar l_{nn} y u_{nn} satisfaciendo $l_{nn}\ u_{nn} = a_{nn} - \sum_{k=1}^{n-1} l_{nk}\ u_{kn}$.
Si $l_{nn}\ u_{nn} = 0$ entonces $A = L\ U$ pero A es singular.

Paso 7: SALIDA (l_{ij} y u_{ij} para $j = 1, \ldots, n$ e $i = 1, \ldots, n$); PARAR.

==========

Una dificultad que puede surgir cuando se usa este algoritmo para obtener la factorización de la matriz de coeficientes de un sistema lineal de ecuaciones es la causada por el hecho de que no se usa pivoteo para reducir el efecto del error de redondeo. Se ha visto en cálculos anteriores que el error de redondeo puede ser muy significativo cuando se usa aritmética de dígitos finitos y que cualquier algoritmo eficiente debe de tomar esto en consideración.

Aún cuando el intercambio de columnas es difícil de incorporar en el algoritmo de factorización, el algoritmo puede alterarse fácilmente para incluir una técnica de intercambio de filas equivalente al procedimiento de pivoteo máximo de columna descrito en el capítulo XV. Este intercambio resulta suficiente en la mayoría de los casos.

El siguiente algoritmo incorpora el procedimiento de factorización del algoritmo de factorización directa junto con el pivoteo máximo de columna y la sustitución hacia adelante y hacia atrás para obtener una solución a un sistema lineal de ecuaciones. El proceso requiere que el sistema lineal $A\ \mathbf{x} = \mathbf{b}$ se escriba como $L\ U\ \mathbf{x} = \mathbf{b}$. La sustitución hacia adelante resuelve el sistema $L\ \mathbf{z} = \mathbf{b}$ y la sustitución hacia atrás resuelve al sistema $U\ \mathbf{x} = L^{-1}\ \mathbf{b} = \mathbf{z}$. Se debe hacer notar que los elementos diferentes de cero de L y U se pueden guardar en los elementos correspondientes de A excepto los de la diagonal de L ó U, la cual debe darse en entrada.

Algoritmo de factorización directa con pivoteo máximo de columna.

==========

Para resolver el sistema lineal $n \times n$ $A\,\mathbf{x} = \mathbf{b}$ en la forma:

$$\begin{aligned} E_1: &\quad a_{11}\,x_1 + a_{12}\,x_2 + \ldots + a_{1n}\,x_n = a_{1,n+1} \\ E_2: &\quad a_{21}\,x_1 + a_{22}\,x_2 + \ldots + a_{2n}\,x_n = a_{2,n+1} \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\,x_1 + a_{n2}\,x_2 + \ldots + a_{nn}\,x_n = a_{n,n+1} \end{aligned}$$

factorizando A en $L\,U$ y resolviendo $L\,\mathbf{z} = \mathbf{b}$ y $U\,\mathbf{x} = \mathbf{z}$ donde se da la diagonal principal de L o U.

Entrada: dimensión n; los elementos a_{ij}, $1 \le i \le n$, $1 \le j \le n+1$ de la matriz ampliada de A; la diagonal $l_{11}, l_{22}, \ldots, l_{nn}$ de L (método de Doolittle) o la diagonal $u_{11}, u_{22}, \ldots, u_{nn}$ de U (método de Crout).

Salida: solución $x_1, x_2, \ldots, x_n$ ó mensaje de que el sistema lineal no tiene solución única.

Paso 1: Sea p el menor entero tal que $1 \le p \le n$ y $|a_{p1}| = \max\limits_{1 \le j \le n} |a_{j1}|$; *(encontrar el primer elemento pivote).*
Si $|a_{p1}| = 0$ SALIDA; *(no existe solución única)* PARAR.

Paso 2: Si $p \neq 1$ entonces intercambiar las filas p y 1 en A.

Paso 3: Seleccionar l_{11} y u_{11} satisfaciendo $l_{11}\,u_{11} = a_{11}$.

Paso 4: Para $j = 2, 3, \ldots, n$ tomar
$u_{1j} = \frac{a_{1j}}{l_{11}}$; *(primera fila de U)*;
$l_{j1} = \frac{a_{j1}}{u_{11}}$; *(primera columna de L).*

Paso 5: Para $i = 2, 3, \ldots, n-1$ seguir los pasos 6–9.

Paso 6: Sea p el menor entero tal que $i \le p \le n$ y

$$\left| a_{pi} - \sum_{k=1}^{i-1} l_{pk}\,u_{ki} \right| = \max_{i \le j \le n} \left| a_{ji} - \sum_{k=1}^{i-1} l_{jk}\,u_{ki} \right| ;$$

(encontrar el $i-$ésimo elemento pivote).
Si el máximo es cero entonces SALIDA;
(no existe solución única) PARAR.

Paso 7: Si $p \neq i$ entonces intercambiar las filas p e i en la matriz A e intercambiar los elementos de las filas p e i de las primeras $(i-1)$ columnas de L.

Paso 8: Seleccionar l_{ii} y u_{ii} satisfaciendo $l_{ii}\,u_{ii} = a_{ii} - \sum\limits_{k=1}^{i-1} l_{ik}\,u_{ki}$.

Paso 9: Para $j = i+1, i+2, \ldots, n$ tomar
$u_{ij} = \frac{1}{l_{ii}}\,[a_{ij} - \sum\limits_{k=1}^{i-1} l_{ik}\,u_{kj}]$; *($i-$ésima fila de U)*;
$l_{ji} = \frac{1}{u_{ii}}\,[a_{ji} - \sum\limits_{k=1}^{i-1} l_{jk}\,u_{ki}]$; *($i-$ésima columna de L).*

Paso 10: Tomar $AUX = a_{nn} - \sum\limits_{k=1}^{n-1} l_{nk}\,u_{kn}$.
Si $AUX = 0$ entonces SALIDA; *(no existe solución única)* PARAR.
Seleccionar l_{nn} y u_{nn} que satisfagan $l_{nn}\,u_{nn} = a_{nn} - \sum\limits_{k=1}^{n-1} l_{nk}\,u_{kn}$.

(Los pasos 11 y 12 resuelven el sistema triangular inferior $L\ \mathbf{z} = \mathbf{b}$.)

Paso 11: Tomar $z_1 = \frac{a_{1,n+1}}{l_{11}}$.

Paso 12: Para $i = 2, 3, \ldots, n$ tomar
$z_i = \frac{1}{l_{ii}}\ [a_{i,n+1} - \sum_{j=1}^{i-1} l_{ij}\ z_j]$.

(Los pasos 13 y 14 resuelven el sistema triangular superior $U\ \mathbf{x} = \mathbf{z}$.)

Paso 13: Tomar $x_n = \frac{z_n}{u_{nn}}$.

Paso 14: Para $i = n-1, n-2, \ldots, 1$ tomar
$x_i = \frac{1}{u_{ii}}\ [z_i - \sum_{j=i+1}^{n} u_{ij}\ x_j]$.

Paso 15: SALIDA $(x_1, x_2, \ldots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

==========

Ejemplo 2.

Para ilustrar el procedimiento seguido en el algoritmo de factorización directa con pivoteo máximo de columna, consideremos el sistema lineal

$$\begin{array}{rcrcrcrcl} 1.00\,x_1 & + & 0.333\,x_2 & + & 1.50\,x_3 & - & 0.333\,x_4 & = & 3.00\ , \\ -2.01\,x_1 & + & 1.45\,x_2 & + & 0.50\,x_3 & + & 2.95\,x_4 & = & 5.40\ , \\ 4.32\,x_1 & - & 1.95\,x_2 & & & + & 2.08\,x_4 & = & 0.13\ , \\ 5.11\,x_1 & - & 4.00\,x_2 & + & 3.33\,x_3 & - & 1.11\,x_4 & = & 3.77\ . \end{array}$$

Seguiremos los pasos del algoritmo de factorización directa con pivoteo máximo de columna con $l_{11} = l_{22} = l_{33} = l_{44} = 1$, usando aritmética de redondeo a tres dígitos. En primer lugar escribimos la matriz ampliada:

$$A_a = [A, \mathbf{b}] = \left(\begin{array}{rrrr|r} 1.00 & 0.333 & 1.50 & -0.333 & 3.00 \\ -2.01 & 1.45 & 0.50 & 2.95 & 5.40 \\ 4.32 & -1.95 & 0.00 & 2.08 & 0.13 \\ 5.11 & -4.00 & 3.33 & -1.11 & 3.77 \end{array}\right)\ .$$

Además, las matrices triangular inferior L y triangular superior U son:

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ l_{21} & 1.00 & 0 & 0 \\ l_{31} & l_{32} & 1.00 & 0 \\ l_{41} & l_{42} & l_{43} & 1.00 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix}\ .$$

Paso 1: Tenemos que encontrar el primer elemento pivote, es decir, el menor entero p tal que $1 \le p \le n$ y $|a_{p1}| = \max_{1 \le j \le n} |a_{j1}|$. En nuestro caso

$$p = 4\ .$$

Paso 2: Dado que $p \neq 1$, entonces tenemos que intercambiar las filas $p = 4$ y 1 en A. La matriz ampliada se transforma en

$$[A, \mathbf{b}] = \left(\begin{array}{rrrr|r} 5.11 & -4.00 & 3.33 & -1.11 & 3.77 \\ -2.01 & 1.45 & 0.50 & 2.95 & 5.40 \\ 4.32 & -1.95 & 0.00 & 2.08 & 0.13 \\ 1.00 & 0.333 & 1.50 & -0.333 & 3.00 \end{array}\right)\ .$$

Paso 3: Se necesita seleccionar l_{11} y u_{11} satisfaciendo $l_{11}\ u_{11} = a_{11} = 5.11$. Y como $l_{11} = 1.00$,

$$u_{11} = 5.11$$

Paso 4: Para $j = 2, 3, 4$ debemos tomar $u_{1j} = \frac{a_{1j}}{l_{11}}$ y $l_{j1} = \frac{a_{j1}}{u_{11}}$.
Es decir,

$$u_{12} = \frac{a_{12}}{l_{11}} = -4.00\ , \quad u_{13} = \frac{a_{13}}{l_{11}} = 3.33\ , \quad u_{14} = \frac{a_{14}}{l_{11}} = -1.11\ ,$$

y

$$l_{21} = \frac{a_{21}}{u_{11}} = \frac{-2.01}{5.11} = -0.393\ ,$$
$$l_{31} = \frac{a_{31}}{u_{11}} = \frac{4.32}{5.11} = 0.845\ ,$$
$$l_{41} = \frac{a_{41}}{u_{11}} = \frac{1.00}{5.11} = 0.196\ .$$

Entonces, las matrices L y U asumen la forma

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ -0.393 & 1.00 & 0 & 0 \\ 0.845 & l_{32} & 1.00 & 0 \\ 0.196 & l_{42} & l_{43} & 1.00 \end{pmatrix} \text{ y } U = \begin{pmatrix} 5.11 & -4.00 & 3.33 & -1.11 \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix}\ .$$

Paso 5: Para $i = 2$ seguir los pasos 6–9.

Paso 6: Ahora tenemos que encontrar el segundo elemento pivote, es decir, encontrar el menor entero p tal que $2 \le p \le 4$ y

$$|a_{p2} - l_{p1}\ u_{12}| = \max_{2 \le j \le 4} |a_{j2} - l_{j1}\ u_{12}|\ .$$

En nuestro caso,

$$|a_{22} - l_{21}\ u_{12}| = |1.45 - (-0.393)(-4.00)| = |-0.12| = 0.12\ ,$$
$$|a_{32} - l_{31}\ u_{12}| = |-1.95 - (0.845)(-4.00)| = |1.43| = 1.43\ ,$$
$$|a_{42} - l_{41}\ u_{12}| = |0.333 - (0.196)(-4.00)| = |1.12| = 1.12\ .$$

Así, $p = 3$.

Paso 7: Dado que $p = 3 \neq 2 = i$, tenemos que intercambiar las filas $p = 3$ e $i = 2$ en la matriz A e intercambiar los elementos de las filas $p = 3$ e $i = 2$ de la primera columna de L. Entonces,

$$[A, \mathbf{b}] = \left(\begin{array}{cccc|c} 5.11 & -4.00 & 3.33 & -1.11 & 3.77 \\ 4.32 & -1.95 & 0.00 & 2.08 & 0.13 \\ -2.01 & 1.45 & 0.50 & 2.95 & 5.40 \\ 1.00 & 0.333 & 1.50 & -0.333 & 3.00 \end{array}\right)\ ,$$

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.845 & 1.00 & 0 & 0 \\ -0.393 & l_{32} & 1.00 & 0 \\ 0.196 & l_{42} & l_{43} & 1.00 \end{pmatrix}\ .$$

Paso 8: Tenemos que seleccionar l_{22} y u_{22} satisfaciendo

$$l_{22}\ u_{22} = a_{22} - l_{21}\ u_{12}\ .$$

Dado que $l_{22} = 1.00$, entonces

$$u_{22} = a_{22} - l_{21}\ u_{12} = -1.95 - (0.845)(-4.00) = 1.43\ .$$

Paso 9: Para $j = 3, 4$ tenemos que tomar

$$u_{2j} = \frac{1}{l_{22}}\ [a_{2j} - l_{2k}\ u_{kj}]$$

$$l_{j2} = \frac{1}{u_{22}}\ [a_{j2} - l_{jk}\ u_{k2}]$$

En nuestro caso,

$$\begin{aligned}
u_{23} &= \frac{1}{l_{22}}\ [a_{23} - l_{21}\ u_{13}] = [0.00 - (0.845)(3.33)] = -2.81\ ,\\
u_{24} &= \frac{1}{l_{22}}\ [a_{24} - l_{21}\ u_{14}] = [2.08 - (0.845)(-1.11)] = 3.01\ ,\\
l_{32} &= \frac{1}{u_{22}}\ [a_{32} - l_{31}\ u_{12}] = \frac{1}{1.43}\ [1.45 - (-0.393)(-4.00)] = -0.0839\ ,\\
l_{42} &= \frac{1}{u_{22}}\ [a_{42} - l_{41}\ u_{12}] = \frac{1}{1.43}\ [0.333 - (0.196)(-4.00)] = 0.783\ .
\end{aligned}$$

Entonces, las matrices L y U asumen la forma

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.845 & 1.00 & 0 & 0 \\ -0.393 & -0.0839 & 1.00 & 0 \\ 0.196 & 0.783 & l_{43} & 1.00 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 5.11 & -4.00 & 3.33 & -1.11 \\ 0 & 1.43 & -2.81 & 3.02 \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix}\ .$$

Paso 5: Para $i = 3$ seguir los pasos 6–9.

Paso 6: Ahora tenemos que encontrar el tercer elemento pivote, es decir, encontrar el menor entero p tal que $3 \le p \le 4$ y

$$|a_{p3} - \sum_{k=1}^{2} l_{pk}\ u_{k3}| = \max_{3 \le j \le 4} |a_{j3} - \sum_{k=1}^{2} l_{jk}\ u_{k3}|\ .$$

En nuestro caso,

$$\begin{aligned}
|a_{33} - (l_{31}\ u_{13} + l_{32}\ u_{23})| &= |0.5 - ((-0.393)(3.33) + (-0.0839)(-2.81))| = 1.57\ ,\\
|a_{43} - (l_{41}\ u_{13} + l_{42}\ u_{23})| &= |1.5 - ((0.196)(3.33) + (0.783)(-2.81)))| = 3.05\ .
\end{aligned}$$

Así, $p = 4$.

Paso 7: Dado que $p = 4 \neq 3 = i$, tenemos que intercambiar las filas $p = 5$ y $i = 3$ en la matriz A e intercambiar los elementos de las filas $p = 5$ e $i = 3$ de la primera y segunda columnas de L. Entonces,

$$[A, \mathbf{b}] = \left(\begin{array}{cccc|c} 5.11 & -4.00 & 3.33 & -1.11 & 3.77 \\ 4.32 & -1.95 & 0.00 & 2.08 & 0.13 \\ 1.00 & 0.333 & 1.50 & -0.333 & 3.00 \\ -2.01 & 1.45 & 0.50 & 2.95 & 5.40 \end{array} \right) ,$$

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.845 & 1.00 & 0 & 0 \\ 0.196 & 0.783 & 1.00 & 0 \\ -0.393 & -0.0839 & l_{43} & 1.00 \end{pmatrix} .$$

Paso 8: Tenemos que seleccionar l_{33} y u_{33} satisfaciendo

$$l_{33}\ u_{33} = a_{33} - (l_{31}\ u_{13} + l_{32}\ u_{23}) .$$

Dado que $l_{33} = 1.00$, entonces

$$u_{33} = a_{33} - (l_{31}\ u_{13} + l_{32}\ u_{23}) = 1.50 - (0.196)(3.33) + (-0.0839)(-2.81) = 3.05 .$$

Paso 9: Para $j = 4$ tenemos que tomar

$$u_{3j} = \frac{1}{l_{33}} \left[a_{3j} - (l_{31}\ u_{1j} + l_{32}\ u_{2j}) \right]$$

$$l_{j3} = \frac{1}{u_{33}} \left[a_{j3} - (l_{j1}\ u_{13} + l_{j2}\ u_{23}) \right] .$$

En nuestro caso,

$$\begin{aligned} u_{34} &= \frac{1}{l_{33}} \left[a_{34} - (l_{31}\ u_{14} + l_{32}\ u_{24}) \right] \\ &= [-0.333 - ((0.196)(-1.11) + (0.783)(3.02))] = -2.47 , \\ l_{43} &= \frac{1}{u_{33}} \left[a_{43} - (l_{41}\ u_{13} + l_{42}\ u_{23}) \right] \\ &= \frac{1}{3.05} \left[0.5 - ((-0.393)(3.33) + (-0.0839)(-2.81)) \right] = 0.515 . \end{aligned}$$

Entonces, las matrices L y U asumen la forma

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.845 & 1.00 & 0 & 0 \\ 0.196 & 0.783 & 1.00 & 0 \\ -0.393 & -0.0839 & 0.515 & 1.00 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 5.11 & -4.00 & 3.33 & -1.11 \\ 0 & 1.43 & -2.81 & 3.02 \\ 0 & 0 & 3.05 & -2.47 \\ 0 & 0 & 0 & u_{44} \end{pmatrix} .$$

Paso 10: Finalmente, tenemos que seleccionar l_{44} y u_{44} que satisfagan

$$l_{44}\ u_{44} = a_{44} - \sum_{k=1}^{3} l_{4k}\ u_{k4} .$$

Dado que $l_{44} = 1.00$, entonces

$$\begin{aligned} u_{44} &= a_{44} - (l_{41}\ u_{14} + l_{42}\ u_{24} + l_{43}\ u_{23}) \\ &= 2.95 - ((-0.393)(-1.11) + (-0.0839)(3.02) + (0.515)(-2.47)) = 4.04\ . \end{aligned}$$

La factorización está completa:

$$A = \begin{pmatrix} 5.11 & -4.00 & 3.33 & -1.11 \\ 4.32 & -1.95 & 0.00 & 2.08 \\ 1.00 & 0.333 & 1.50 & -0.333 \\ -2.01 & 1.45 & 0.50 & 2.95 \end{pmatrix} =$$

$$= \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.845 & 1.00 & 0 & 0 \\ 0.196 & 0.783 & 1.00 & 0 \\ -0.393 & -0.0839 & 0.515 & 1.00 \end{pmatrix} \begin{pmatrix} 5.11 & -4.00 & 3.33 & -1.11 \\ 0 & 1.43 & -2.81 & 3.02 \\ 0 & 0 & 3.05 & -2.47 \\ 0 & 0 & 0 & 4.04 \end{pmatrix}\ .$$

(Los pasos 11 y 12 resuelven el sistema triangular inferior $L\ \mathbf{z} = \mathbf{b}$.)

Paso 11: Tomar $z_1 = \frac{a_{1,5}}{l_{11}} = \frac{3.77}{1.00} = 3.77$.

Paso 12: Para $i = 2, 3, 4$ tomar

$$z_i = \frac{1}{l_{ii}}\ [a_{i,n+1} - \sum_{j=1}^{i-1} l_{ij}\ z_j]\ .$$

En nuestro caso:

$$\begin{aligned} z_2 &= \frac{1}{l_{22}}\ [a_{25} - l_{21}\ z_1] \\ &= 0.13 - (0.845)(3.77) = -3.06 \\ z_3 &= \frac{1}{l_{33}}\ [a_{35} - (l_{31}\ z_1 + l_{32}\ z_2)] \\ &= 3.00 - ((0.196)(3.77) + (0.783)(-3.06)) = 4.66 \\ z_4 &= \frac{1}{l_{44}}\ [a_{45} - (l_{41}\ z_1 + l_{42}\ z_2 + l_{43}\ z_3)] \\ &= 5.40 - ((-0.393)(3.77) + (-0.0839)(-3.06) + (0.515)(4.66)) = 4.22\ . \end{aligned}$$

(Los pasos 13 y 14 resuelven el sistema triangular superior $U\ \mathbf{x} = \mathbf{z}$.)

Paso 13: Tomar $x_4 = \frac{z_4}{u_{44}} = \frac{4.22}{4.04} = 1.04$.

Paso 14: Para $i = 3, 2, 1$ tomar

$$x_i = \frac{1}{u_{ii}}\ [z_i - \sum_{j=i+1}^{n} u_{ij}\ x_j]\ .$$

En nuestro caso:

$$\begin{aligned} x_3 &= \frac{1}{u_{33}} \left[z_3 - u_{34}\ x_4 \right] \\ &= \frac{1}{3.05} \left[4.66 - (-2.47)(1.04) \right] = 2.37 \\ x_2 &= \frac{1}{u_{22}} \left[z_2 - (u_{23}\ x_3 + u_{24}\ x_4) \right] \\ &= \frac{1}{1.43} \left[-3.06 - ((-2.81)(2.37) + (3.02)(1.04)) \right] = 0.322 \\ x_1 &= \frac{1}{u_{33}} \left[z_3 - (u_{12}\ x_2 + u_{13}\ x_3 + u_{14}\ x_4) \right] \\ &= \frac{1}{5.11} \left[3.77 - ((-4.00)(0.322) + (3.33)(2.37) + (-1.11)(1.04)) \right] = -0.329\ . \end{aligned}$$

Paso 15: SALIDA. La solución es

$$x_1 = -0.329\ ,\quad x_2 = 0.322\ ,\quad x_3 = 2.37\ ,\quad x_4 = 1.04\ .$$

(procedimiento completado satisfactoriamente) PARAR.

Una aplicación del algoritmo de factorización directa da lugar a la factorización

$$A = \begin{pmatrix} 1.00 & 0.333 & 1.50 & -0.333 \\ -2.01 & 1.45 & 0.50 & 2.95 \\ 4.32 & -1.95 & 0.00 & 2.08 \\ 5.11 & -4.00 & 3.33 & -1.11 \end{pmatrix} =$$

$$= \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ -2.01 & 1.00 & 0 & 0 \\ 4.32 & -1.60 & 1.00 & 0 \\ 5.11 & -2.69 & -6.04 & 1.00 \end{pmatrix} \begin{pmatrix} 1.00 & 0.333 & 1.50 & -0.333 \\ 0 & 2.12 & 3.52 & 2.28 \\ 0 & 0 & -0.85 & 7.17 \\ 0 & 0 & 0 & 50.0 \end{pmatrix}\ .$$

Aplicando entonces los pasos 11 hasta el 15 del algoritmo de factorización directa con pivoteo máximo de columna se obtiene la solución

$$x_1 = -0.370\ ,\quad x_2 = 0.236\ ,\quad x_3 = 2.42\ ,\quad x_4 = 1.03\ .$$

La siguiente tabla compara los resultados del algoritmo de factorización directa con pivoteo máximo de columna, del algoritmo de factorización directa y de la respuesta real a tres dígitos. Nótese la mejoría en la precisión cuando se incluyen intercambios de filas.

Tabla 1

	x_1	x_2	x_3	x_4
Alg. fact. pivoteo	−0.329	0.322	2.37	1.04
Alg. fact. directa	−0.370	0.236	2.42	1.03
Real	−0.324	0.321	2.37	1.04

3. EL ALGORITMO DE CHOLESKY

Cuando se sabe que la matriz real es simétrica y positiva definida, se puede mejorar significativamente la técnica de factorización de una matriz con respecto al número de operaciones aritméticas requeridas.

Teorema XVI.1

Si A es una matriz real de $n \times n$ simétrica y positiva definida, entonces A tiene una factorización de la forma $A = L\ L^t$, donde L es una matriz triangular inferior. La factorización se puede lograr aplicando el algoritmo de factorización directa con $l_{ii} = u_{ii}$ para cada $i = 1, 2, \ldots, n$.

Para una matriz simétrica y positiva definida, este Teorema se puede usar para simplificar el algoritmo de factorización directa. Además, si se tiene que resolver un sistema lineal representado por una matriz positiva definida, los pasos 1–6 del siguiente algoritmo (**algoritmo de Choleski**) pueden sustituirse por los pasos 1–10 del algoritmo de factorización directa con pivoteo máximo de columna para aprovechar la simplificación que resulta, siempre y cuando u_{ij} sea reemplazado por l_{ij} en los pasos 13 y 14. El procedimiento de factorización se describe en el siguiente algoritmo.

Algoritmo de Choleski.

==========

Para factorizar una matriz $n \times n$ simétrica y positiva definida $A = (a_{ij})$ como $A = L\ L^t$, donde L es triangular inferior.

Entrada: dimensión n; los elementos $a_{ij}, 1 \leq i, j \leq n$ de A.

Salida: los elementos l_{ij}, $1 \leq j \leq i$, $1 \leq i \leq n$ de L; (los elementos de $U = L^t$ son $u_{ij} = l_{ji}$, $i \leq j \leq n$, $1 \leq i \leq n$).

Paso 1: Tomar

$$l_{11} = \sqrt{a_{11}}\ .$$

Paso 2: Para $j = 2, 3, \ldots, n$ tomar

$$l_{j1} = \frac{a_{j1}}{l_{11}}\ .$$

Paso 3: Para $i = 2, 3, \ldots, n-1$ seguir los pasos 4 y 5.

Paso 4: Tomar

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}\ .$$

Paso 5: Para $j = i+1, i+2, \ldots, n$ tomar

$$l_{ji} = \frac{1}{l_{ii}}\ [a_{ji} - \sum_{k=1}^{i-1} l_{jk}\ l_{ik}]\ .$$

Paso 6: Tomar

$$l_{nn} = \sqrt{a_{nn} - \sum_{k=1}^{n-1} l_{nk}^2}\ .$$

Paso 7: SALIDA (l_{ij} para $j = 1, \ldots, i$ e $i = 1, \ldots, n$); PARAR.

==========

La solución de un sistema lineal típico representado por una matriz positiva definida usando el algoritmo de Choleski requiere de

raíces cuadradas

$$n$$

multiplicaciones/divisiones

$$\frac{n^3 + 9\ n^2 + 2\ n}{6}$$

sumas/restas

$$\frac{n^3 + 6\ n^2 - 7\ n}{6}\ .$$

Estas son alrededor de la mitad de las operaciones aritméticas requeridas en el algoritmo de eliminación Gaussiana. La vantaja computacional del método de Choleski depende del número de operaciones que se requieran para determinar los valores de las n raíces cuadradas, el cual, debido a que es un factor lineal con n, decrecerá significativamente conforme n crezca.

Ejemplo 3.

Para ilustrar el procedimiento seguido al aplicar el método de Choleski, consideremos la matriz real, simétrica y definida positiva

$$A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{pmatrix}\ .$$

Seguiendo los pasos del algoritmo de Choleski,

Paso 1: Tomamos

$$l_{11} = \sqrt{a_{11}} = \sqrt{4} = 2\ .$$

Paso 2: Para $j = 2, 3, \ldots, n$ debemos tomar

$$l_{j1} = \frac{a_{j1}}{l_{11}}\ .$$

En nuestro caso es:

$$l_{21} = \frac{a_{21}}{l_{11}} = \frac{-1}{2} = -0.5\ ,$$
$$l_{31} = \frac{a_{31}}{l_{11}} = \frac{1}{2} = 0.5\ .$$

Paso 3: Para $i = 2$ seguir los pasos 4 y 5.

Paso 4: Tomar

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}\ .$$

Es decir, en nuestro caso:

$$l_{22} = \sqrt{a_{22} - l_{21}^2} = \sqrt{4.25 - (-0.5)^2} = 2 \ .$$

Paso 5: Para $j = 3$ tomar

$$l_{ji} = \frac{1}{l_{ii}} \ [a_{ji} - \sum_{k=1}^{i-1} l_{jk} \ l_{ik}] \ .$$

Es decir,

$$l_{32} = \frac{1}{l_{22}} \ (a_{32} - l_{31} \ l_{21}) = \frac{1}{2} \ (2.75 - 0.5 \ (-0.5)) = 1.5 \ .$$

Paso 6: Tomar

$$l_{nn} = \sqrt{a_{nn} - \sum_{k=1}^{n-1} l_{nk}^2} \ .$$

Es decir,

$$l_{33} = \sqrt{a_{33} - (l_{31}^2 + l_{32}^2)} = \sqrt{3.5 - (0.5)^2 - (1.5)^2} = 1 \ .$$

Como L es triangular inferior y $U = L^t$, las matrices son:

$$L = \begin{pmatrix} 2 & 0 & 0 \\ -0.5 & 2 & 0 \\ 0.5 & 1.5 & 1 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 2 & -0.5 & 0.5 \\ 0 & 2 & 1.5 \\ 0 & 0 & 1 \end{pmatrix} \ .$$

4. EL ALGORITMO DE CROUT PARA SISTEMAS TRIDIAGONALES

Los algoritmos de factorización se pueden simplificar considerablemente en el caso de matrices de banda debido al gran número de ceros que aparecen en patrones regulares en estas matrices. Es particularmente interesante observar la forma que los métodos de Crout o Doolittle toman en este caso. Para ilustrar esta situación, supongamos que una matriz tridiagonal

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 & \dots & \dots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \dots & 0 \\ 0 & a_{32} & a_{33} & a_{34} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & \dots & \dots & 0 & a_{n,n-1} & a_{nn} \end{pmatrix} \ ,$$

pueda factorizarse en las matrices triangulares L y U.

Como A tiene solamente $(3 \ n - 2)$ elementos distintos de cero, habrá sólo $(3 \ n - 2)$ condiciones para determinar a los elementos de L y U siempre y cuando se obtengan

también los elementos cero de A. Supongamos que realmente es posible encontrar las matrices en la forma

$$L = \begin{pmatrix} l_{11} & 0 & \dots & \dots & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & \dots & 0 \\ 0 & l_{32} & l_{33} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & l_{n-1,n-2} & l_{n-1,n-1} & 0 \\ 0 & \dots & \dots & 0 & l_{n,n-1} & l_{nn} \end{pmatrix} ,$$

y

$$U = \begin{pmatrix} 1 & u_{12} & 0 & \dots & \dots & 0 \\ 0 & 1 & u_{23} & 0 & \dots & 0 \\ 0 & 0 & 1 & u_{34} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & 1 & u_{n-1,n} \\ 0 & \dots & \dots & 0 & 0 & 1 \end{pmatrix} .$$

De esta forma hay $(2\ n-1)$ elementos indeterminados de L y $(n-1)$ elementos indeterminados de U, que en total son iguales, en número, a las condiciones mencionadas arriba y además, los elementos cero de A se obtienen automáticamente.

La multiplicación $A = L\ U$ da, sin contar los elementos cero, las ecuaciones:

$$\begin{aligned} a_{11} &= l_{11} \ , \\ a_{i,i-1} &= l_{i,i-1} \ , \quad \text{para cada } \ i = 2, 3, \dots, n \ , \\ a_{ii} &= l_{i,i-1}\ u_{i-1,i} + l_{ii} \ , \quad \text{para cada } \ i = 2, 3, \dots, n \ , \\ a_{i,i+1} &= l_{ii}\ u_{i,i+1} \ , \quad \text{para cada } \ i = 1, 2, \dots, n-1 \ . \end{aligned}$$

Una solución a este sistema de ecuaciones puede encontarse obteniendo primero todos los términos no cero fuera de la diagonal de L, usando la segunda ecuación y luego usando la cuarta y la tercera para obtener alternadamente el resto de los elementos de U y L, los cuales se pueden ir guardando en los elementos correspondientes de A.

A continuación se da un algoritmo completo para resolver un sistema de ecuaciones lineales de $n \times n$ cuya matriz de coeficientes es tridiagonal.

Algoritmo de reducción de Crout para sistemas lineales tridiagonales.

==========

Para resolver el sistema lineal tridiagonal de $n \times n$

$$\begin{aligned} E_1 &: a_{11}\ x_1 + a_{12}\ x_2 && = a_{1,n+1} \ , \\ E_2 &: a_{21}\ x_1 + a_{22}\ x_2 + a_{23}\ x_3 && = a_{2,n+1} \ , \\ &\dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \\ E_{n-1} &: \quad a_{n-1,n-2}\ x_{n-2} + a_{n-1,n-1}\ x_{n-1} + a_{n-1,n}\ x_n = a_{n-1,n+1} \ , \\ E_n &: \quad a_{n,n-1}\ x_{n-1} + a_{nn}\ x_n = a_{n,n+1} \ . \end{aligned}$$

el cual se supone tiene solución única.

Entrada: dimensión n; los elementos a_{ij}, $1 \le i \le n$ y $1 \le j \le n+1$ de A_a.

Salida: solución $x_1, x_2, \ldots, x_n$.

Paso 1: Tomar
$$l_{11} = a_{11} \quad \text{y} \quad u_{12} = \frac{a_{12}}{l_{11}} \ .$$

Paso 2: Para $i = 2, 3, \ldots, n-1$ tomar
$l_{i,i-1} = a_{i,i-1}$; ($i-$*ésima fila de* L).
$l_{ii} = a_{ii} - l_{i,i-1}\ u_{i-1,i}$.
$u_{i,i+1} = \frac{a_{i,i+1}}{l_{ii}}$; ($(i+1)-$*ésima columna de* U).

Paso 3: Tomar $l_{n,n-1} = a_{n,n-1}$; ($n-$*ésima fila de* L).
$l_{nn} = a_{nn} - l_{n,n-1}\ u_{n-1,n}$.

(Los pasos 4 y 5 resuelven $L\ \mathbf{z} = \mathbf{b}$*).*

Paso 4: Tomar
$$z_1 = \frac{a_{1,n+1}}{l_{11}} \ .$$

Paso 5: Para $i = 2, 3, \ldots, n$ tomar
$$z_i = \frac{1}{l_{ii}} \left[a_{i,n+1} - l_{i,i-1}\ z_{i-1} \right] .$$

(Los pasos 6 y 7 resuelven $U\ \mathbf{x} = \mathbf{z}$*).*

Paso 6: Tomar
$$x_n = z_n \ .$$

Paso 7: Para $i = n-1, n-2, \ldots, 1$ tomar
$$x_i = z_i - u_{i,i+1}\ x_{i+1} \ .$$

Paso 8: SALIDA ($x_1, x_2, \ldots, x_n$); PARAR.

==========

Este algoritmo requiere sólo de $(5\ n-4)$ multiplicaciones/divisiones y de $(3\ n-3)$ sumas/restas, y consecuentemente tiene una ventaja computacional considerable sobre los métodos que no consideran la triadiagonalidad de la matriz, especialmente para valores grandes de n.

Ejemplo 4.

Para ilustrar el procedimiento llevado a cabo en el algoritmo de reducción de Crout, consideremos el sistema tridiagonal de ecuaciones

$$\begin{array}{rcrcrcrcl}
2\ x_1 & - & x_2 & & & & & = & 1 \\
-1\ x_1 & + & 2\ x_2 & - & x_3 & & & = & 0 \\
 & - & x_2 & + & 2\ x_3 & - & x_4 & = & 0 \\
 & & & - & x_3 & + & 2\ x_4 & = & 1
\end{array}$$

cuya matriz ampliada es:

$$[A, \mathbf{b}] = \left(\begin{array}{cccc|c}
2 & -1 & 0 & 0 & 1 \\
-1 & 2 & -1 & 0 & 0 \\
0 & -1 & 2 & -1 & 0 \\
0 & 0 & -1 & 2 & 1
\end{array} \right) \ .$$

Seguiendo los pasos del algoritmo de reducción de Crout,

Paso 1: Tomar

$$l_{11} = a_{11} = 2 \quad \text{y} \quad u_{12} = \frac{a_{12}}{l_{11}} = -\frac{1}{2} \ .$$

Paso 2: Para $i = 2, 3$ tomar

$l_{i,i-1} = a_{i,i-1}$; ($i-$*ésima fila de* L).

$l_{ii} = a_{ii} - l_{i,i-1}\ u_{i-1,i}$.

$u_{i,i+1} = \frac{a_{i,i+1}}{l_{ii}}$; ($(i+1)-$*ésima columna de* U).

Es decir, en nuestro caso:

$i = 2$

$$l_{21} = a_{21} = -1 \ ;$$

$$l_{22} = a_{22} - l_{21}\ u_{12} = 2 - (-1)(-\frac{1}{2}) = \frac{3}{2} \ ;$$

$$u_{23} = \frac{a_{23}}{l_{22}} = -\frac{1}{3/2} = -\frac{2}{3} \ ;$$

$i = 3$

$$l_{32} = a_{32} = -1 \ ;$$

$$l_{33} = a_{33} - l_{32}\ u_{23} = 2 - (-1)(-\frac{2}{3}) = \frac{4}{3} \ ;$$

$$u_{34} = \frac{a_{34}}{l_{33}} = -\frac{3}{4} \ .$$

Paso 3: Tomar

$$l_{43} = a_{43} = -1 \ ;$$

$$l_{44} = a_{44} - l_{43}\ u_{34} = 2 - (-1)(-\frac{3}{4}) = \frac{5}{4} \ .$$

(Los pasos 4 y 5 resuelven $L\ \mathbf{z} = \mathbf{b}$*).*

Paso 4: Tomar

$$z_1 = \frac{a_{1,5}}{l_{11}} = \frac{1}{2} \ .$$

Paso 5: Para $i = 2, 3, 4$ tomar

$$z_i = \frac{1}{l_{ii}}\ [a_{i,n+1} - l_{i,i-1}\ z_{i-1}] \ .$$

Es decir,

$$z_2 = \frac{1}{l_{22}}\ [a_{25} - l_{21}\ z_1] = \frac{0 - (-1)(\frac{1}{2})}{\frac{3}{2}} = \frac{1}{3} \ ,$$

$$z_3 = \frac{1}{l_{33}}\ [a_{35} - l_{32}\ z_2] = \frac{0 - (-1)(\frac{1}{3})}{\frac{4}{3}} = \frac{1}{4} \ ,$$

$$z_4 = \frac{1}{l_{44}}\ [a_{45} - l_{43}\ z_3] = \frac{1 - (-1)(\frac{1}{4})}{\frac{5}{4}} = 1 \ .$$

(Los pasos 6 y 7 resuelven $U\ \mathbf{x} = \mathbf{z}$*).*

Paso 6: Tomar

$$x_4 = z_4 = 1 \ .$$

Paso 7: Para $i = 3, 2, 1$ tomar

$$x_i = z_i - u_{i,i+1}\ x_{i+1} \ .$$

Es decir,

$$x_3 = z_3 - u_{34}\ x_4 = \frac{1}{4} - (-\frac{3}{4}) = 1 \ ,$$
$$x_2 = z_2 - u_{23}\ x_3 = \frac{1}{3} - (-\frac{2}{3}) = 1 \ ,$$
$$x_1 = z_1 - u_{12}\ x_2 = \frac{1}{2} - (-\frac{1}{2}) = 1 \ .$$

El algoritmo factorizó la matriz en

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ -1 & 3/2 & 0 & 0 \\ 0 & -1 & 4/3 & 0 \\ 0 & 0 & -1 & 5/4 \end{pmatrix} \begin{pmatrix} 1 & -1/2 & 0 & 0 \\ 0 & 1 & -2/3 & 0 \\ 0 & 0 & 1 & -3/4 \\ 0 & 0 & 0 & 1 \end{pmatrix} \ ,$$

y dió la solución correcta

$$x_1 = x_2 = x_3 = x_4 = 1 \ .$$

El algoritmo de reducción de Crout para sistemas lineales tridiagonales puede aplicarse cuando $l_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$. Dos condiciones, cualquiera de las cuales asegurará que esto es cierto, son que la matriz de coeficientes del sistema sea positiva definida o que sea estrictamente dominante diagonalmente. Una condición adicional que garantiza que este algoritmo se puede aplicar está dada en el siguiente Teorema.

Teorema XVI.2

Supóngase que $A = (a_{ij})$ es tridiagonal con $a_{i,i-1} \cdot a_{i,i+1} \neq 0$ para cada $i = 2, 3, \ldots, n-1$. Si $|a_{11}| > |a_{12}|$, $|a_{ii}| > |a_{i,i-1}| + |a_{i,i+1}|$ para cada $i = 2, 3, \ldots, n-1$, y $|a_{nn}| > |a_{n,n-1}|$, entonces A es no singular y los valores de l_{ii} descritos en el algoritmo de reducción de Crout son diferentes de cero para cada $i = 1, 2, \ldots, n$.

EJERCICIOS.

1. Factorizar las siguientes matrices en la descomposición $L\,U$ usando el algoritmo de Doolittle con $l_{ii} = 1$ para cada i:

$a)$ $\begin{pmatrix} 2 & -1 & 1 \\ 3 & 3 & 9 \\ 3 & 3 & 5 \end{pmatrix}$ $\qquad$ $b)$ $\begin{pmatrix} 2 & -1.5 & 3 \\ -1 & 0 & 3 \\ 4 & -4.5 & 5 \end{pmatrix}$

$c)$ $\begin{pmatrix} 1.012 & -2.132 & 3.104 \\ -2.132 & 4.096 & -7.013 \\ 3.104 & -7.013 & 0.014 \end{pmatrix}$ $\qquad$ $d)$ $\begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 1.5 & 0 & 0 \\ 0 & -3 & 0.5 & 0 \\ 2 & -2 & 1 & 1 \end{pmatrix}$

2. Resolver los siguientes sistemas lineales usando el algoritmo de factorización directa con pivoteo máximo de columna con $l_{ii} = 1$:

$a)$
$$\begin{array}{rcrcrcr} 2\,x_1 & - & x_2 & + & x_3 & = & -1\,, \\ 3\,x_1 & + & 3\,x_2 & + & 9\,x_3 & = & 0\,, \\ 3\,x_1 & + & 3\,x_2 & + & 5\,x_3 & = & 4\,. \end{array}$$

$b)$
$$\begin{array}{rcrcrcr} 2\,x_1 & - & 1.5\,x_2 & + & 3\,x_3 & = & 1\,, \\ -\,x_1 & & & + & 2\,x_3 & = & 3\,, \\ 4\,x_1 & - & 4.5\,x_2 & + & 5\,x_3 & = & -1\,. \end{array}$$

$c)$
$$\begin{array}{rcrcrcr} 1.012\,x_1 & - & 2.132\,x_2 & + & 3.104\,x_3 & = & 1.984\,, \\ -2.132\,x_1 & + & 4.096\,x_2 & - & 7.013\,x_3 & = & -5.049\,, \\ 3.104\,x_1 & - & 7.013\,x_2 & + & 0.014\,x_3 & = & -3.895\,. \end{array}$$

$d)$
$$\begin{array}{rcrcrcrcr} 2\,x_1 & & & & & & & = & 3\,, \\ x_1 & + & 1.5\,x_2 & & & & & = & 4.5\,, \\ & - & 3\,x_2 & + & 0.5\,x_3 & & & = & -6.6\,, \\ 2\,x_1 & - & 2\,x_2 & + & x_3 & + & x_4 & = & 0.8\,. \end{array}$$

3. Usar el algoritmo de Cholesky para encontrar una factorización de la forma $A = L\,L^t$ para las siguientes matrices:

$a)$ $\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ $\qquad$ $b)$ $\begin{pmatrix} 4 & 1 & 1 & 1 \\ 1 & 3 & -1 & 1 \\ 1 & -1 & 2 & 0 \\ 1 & 1 & 0 & 2 \end{pmatrix}$

$c)$ $\begin{pmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{pmatrix}$ $\qquad$ $d)$ $\begin{pmatrix} 4 & 1 & -1 & 0 \\ 1 & 3 & -1 & 0 \\ -1 & -1 & 5 & 2 \\ 0 & 0 & 2 & 4 \end{pmatrix}$

4. Resolver los siguientes sistemas lineales tridiagonales usando el algoritmo de reducción de Crout:

a)
$$\begin{array}{rcrcrcr} x_1 & - & x_2 & & & = & 0\,, \\ -2\,x_1 & + & 4\,x_2 & - & 2\,x_3 & = & -1\,, \\ & - & x_2 & + & 2\,x_3 & = & 1.5\,. \end{array}$$

b)
$$\begin{array}{rcrcrcr} 3\,x_1 & + & x_2 & & & = & -1\,, \\ 2\,x_1 & + & 4\,x_2 & + & x_3 & = & 7\,, \\ & & 2\,x_2 & + & 5\,x_3 & = & 9\,. \end{array}$$

c)
$$\begin{array}{rcrcrcr} 2\,x_1 & - & x_2 & & & = & 3\,, \\ -\,x_1 & + & 2\,x_2 & - & x_3 & = & -3\,, \\ & - & x_2 & + & 2\,x_3 & = & 1\,. \end{array}$$

d)
$$\begin{array}{rcrcrcrcr} 0.5\,x_1 & + & 0.25\,x_2 & & & & & = & 0.35\,, \\ 0.35\,x_1 & + & 0.8\,x_2 & + & 0.4\,x_3 & & & = & 0.77\,, \\ & & 0.25\,x_2 & + & x_3 & + & 0.5\,x_4 & = & -0.5\,, \\ & & & & x_3 & - & 2\,x_4 & = & -2.25\,. \end{array}$$

CAPITULO XVII. TECNICAS ITERATIVAS PARA RESOLVER SISTEMAS LINEALES

1. INTRODUCCION Y METODO

Una técnica iterativa para resolver un sistema lineal $A\,\mathbf{x} = \mathbf{b}$ de $n \times n$ empieza con una aproximación inicial $\mathbf{x}^{(0)}$ a la solución $\mathbf{x}$, y genera una sucesión de vectores $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ que converge a $\mathbf{x}$. La mayoría de estas técnicas iterativas involucran un proceso que convierte el sistema $A\,\mathbf{x} = \mathbf{b}$ en un sistema equivalente de la forma $\mathbf{x} = T\,\mathbf{x} + \mathbf{c}$ para alguna matriz T de $n \times n$ y un vector $\mathbf{c}$. Ya seleccionado el vector inicial $\mathbf{x}^{(0)}$ la sucesión de vectores de solución aproximada se genera calculando

$$\mathbf{x}^{(k)} = T\,\mathbf{x}^{(k-1)} + \mathbf{c} \qquad (XVII.1)$$

para cada $k = 1, 2, 3, \ldots$. Este tipo de procedimiento nos recuerda a la iteración del punto fijo estudiada en la tercera parte.

Las técnicas iterativas se emplean raras veces para resolver sistemas lineales de dimensión pequeña ya que el tiempo requerido para lograr una precisión suficiente excede al de las técnicas directas como el método de eliminación Gaussiana. Sin embargo, para sistemas grandes con un gran porcentaje de ceros, estas técnicas son eficientes en términos de almacenamiento en la computadora y del tiempo requerido. Los sistemas de este tipo surgen frecuentemente en la solución numérica de problemas de valores en la frontera y de ecuaciones diferenciales parciales.

Ejemplo 1.

El sistema lineal $A\,\mathbf{x} = \mathbf{b}$ dado por

$$\begin{array}{lrcrcrcrcr}
E_1: & 10\,x_1 & - & x_2 & + & 2\,x_3 & & & = & 6\,, \\
E_2: & -\,x_1 & + & 11\,x_2 & - & x_3 & + & 3\,x_4 & = & 25\,, \\
E_3: & 2\,x_1 & - & x_2 & + & 10\,x_3 & - & x_4 & = & -11\,, \\
E_4: & & & 3\,x_2 & - & x_3 & + & 8\,x_4 & = & 15\,,
\end{array}$$

tiene por solución a $\mathbf{x} = (1, 2, -1, 1)^t$. Para convertir $A\,\mathbf{x} = \mathbf{b}$ a la forma $\mathbf{x} = T\,\mathbf{x} + \mathbf{c}$, resolvemos la ecuación E_i para cada $i = 1, 2, 3, 4$, obteniendo:

$$\begin{array}{lrcrcrcrcr}
x_1 = & & & \frac{1}{10}x_2 & - & \frac{1}{5}x_3 & & & + & \frac{3}{5}\,, \\
x_2 = & \frac{1}{11}x_1 & & & + & \frac{1}{11}x_3 & - & \frac{3}{11}x_4 & + & \frac{25}{11}\,, \\
x_3 = & -\frac{1}{5}x_1 & + & \frac{1}{10}x_2 & & & + & \frac{1}{10}x_4 & - & \frac{11}{10}\,, \\
x_4 = & & - & \frac{3}{8}x_2 & + & \frac{1}{8}x_3 & & & + & \frac{15}{8}\,.
\end{array}$$

En este ejemplo,

$$T = \begin{pmatrix} 0 & \frac{1}{10} & -\frac{1}{5} & 0 \\ \frac{1}{11} & 0 & \frac{1}{11} & -\frac{3}{11} \\ -\frac{1}{5} & \frac{1}{10} & 0 & \frac{1}{10} \\ 0 & -\frac{3}{8} & \frac{1}{8} & 0 \end{pmatrix} \quad \text{y} \quad \mathbf{c} = \begin{pmatrix} \frac{3}{5} \\ \frac{25}{11} \\ -\frac{11}{10} \\ \frac{15}{8} \end{pmatrix}.$$

Como una aproximación inicial tomemos a $\mathbf{x}^{(0)} = (0,0,0,0)^t$ y generemos $\mathbf{x}^{(1)}$ mediante:

$$\begin{aligned}
x_1^{(1)} &= \tfrac{1}{10}x_2^{(0)} - \tfrac{1}{5}x_3^{(0)} + \tfrac{3}{5} = 0.6000\ ,\\
x_2^{(1)} &= \tfrac{1}{11}x_1^{(0)} + \tfrac{1}{11}x_3^{(0)} - \tfrac{3}{11}x_4^{(0)} + \tfrac{25}{11} = 2.2727\ ,\\
x_3^{(1)} &= -\tfrac{1}{5}x_1^{(0)} + \tfrac{1}{10}x_2^{(0)} + \tfrac{1}{10}x_4^{(0)} - \tfrac{11}{10} = -1.1000\ ,\\
x_4^{(1)} &= -\tfrac{3}{8}x_2^{(0)} + \tfrac{1}{8}x_3^{(0)} + \tfrac{15}{8} = 1.8750\ .
\end{aligned}$$

Las iteraciones adicionales $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^t$, se generan de manera similar y se presentan en la tabla siguiente.

Tabla 1

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$
0	0.0000	0.0000	0.0000	0.0000
1	0.6000	2.2727	−1.1000	1.8750
2	1.0473	1.7159	−0.80523	0.88524
3	0.93264	2.0533	−1.0493	1.1309
4	1.0152	1.9537	−0.96811	0.97385
5	0.98899	2.0114	−1.0103	1.0213
6	1.0032	1.9923	−0.99453	0.99444
7	0.99814	2.0023	−1.0020	1.0036
8	1.0006	1.9987	−0.99904	0.99889
9	0.99968	2.0004	−1.0004	1.0006
10	1.0001	1.9998	−0.99984	0.99980

La decisión de parar después de diez iteraciones está basada en el hecho de que

$$\frac{||\mathbf{x}^{(10)} - \mathbf{x}^{(9)}||_\infty}{||\mathbf{x}^{(10)}||_\infty} = \frac{8.0 \times 10^{-4}}{1.9998} < 10^{-3}\ .$$

En realidad, $||\mathbf{x}^{(10)} - \mathbf{x}||_\infty = 0.0002$.

El método del ejemplo 1 se llama **método iterativo de Jacobi**. Este consiste en resolver la $i-$ésima ecuación de $A\,\mathbf{x} = \mathbf{b}$ para x_i para obtener, siempre y cuando $a_{ii} \neq 0$, que

$$x_i = \sum_{\substack{j=1\\ j\neq i}}^{n} \Big(-\frac{a_{ij}\ x_j}{a_{ii}} \Big) + \frac{b_i}{a_{ii}} \qquad \text{para}\ \ i = 1,2,\ldots,n \tag{$XVII.2$}$$

y generar cada $x_i^{(k)}$ de las componentes de $\mathbf{x}^{(k-1)}$ para $k \geq 1$ con

$$x_i^{(k)} = \frac{1}{a_{ii}}[\sum_{\substack{j=1\\ j\neq i}}^{n}(-a_{ij}\ x_j^{(k-1)}) + b_i] \qquad \text{para}\ \ i = 1,2,\ldots,n\ . \tag{$XVII.3$}$$

El método puede escribirse en la forma $\mathbf{x}^{(k)} = T\ \mathbf{x}^{(k-1)} + \mathbf{c}$ dividiendo a A en su parte diagonal y no-diagonal. Para ver esto, sean D la matriz diagonal cuya diagonal es la

misma que la diagonal de A, $-L$ la parte triangular estrictamente inferior de A, y $-U$ la parte triangular estrictamente superior de A. Con esta notación, se separa en

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix} +$$
$$- \begin{pmatrix} 0 & 0 & \dots & 0 \\ -a_{21} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ -a_{n1} & \dots & -a_{n,n-1} & 0 \end{pmatrix} - \begin{pmatrix} 0 & -a_{12} & \dots & -a_{1n} \\ 0 & 0 & \dots & \dots \\ \dots & \dots & \dots & -a_{n-1,n} \\ 0 & 0 & \dots & 0 \end{pmatrix} =$$
$$= D - L - U .$$

La ecuación $A\,\mathbf{x} = \mathbf{b}$ ó $(D-L-U)\,\mathbf{x} = \mathbf{b}$ se transforma entonces en $D\,\mathbf{x} = (L+U)\,\mathbf{x}+\mathbf{b}$, y finalmente

$$\mathbf{x} = D^{-1}\,(L+U)\,\mathbf{x} + D^{-1}\,\mathbf{b}\,. \qquad (XVII.4)$$

Esto da lugar a la forma matricial de la técnica iterativa de Jacobi:

$$\mathbf{x}^{(k)} = D^{-1}\,(L+U)\,\mathbf{x}^{(k-1)} + D^{-1}\,\mathbf{b}\,, \qquad k = 1, 2, \dots . \qquad (XVII.5)$$

En la práctica, la ecuación $(XVII.3)$ es la que se usa para los cálculos, reservando a la ecuación $(XVII.5)$ para propósitos teóricos.

2. LOS ALGORITMOS DE JACOBI Y DE GAUSS-SEIDEL

Para resumir el método iterativo de Jacobi, presentamos el siguiente algoritmo:

Algoritmo iterativo de Jacobi.

==========

Para resolver el sistema lineal $A\,\mathbf{x} = \mathbf{b}$ con una aproximación inicial dada $\mathbf{x}^{(0)}$.

Entrada: número de incógnitas y de ecuaciones n; las componentes de la matriz $A = (a_{ij})$ donde $1 \le i, j \le n$; las componentes b_i, con $1 \le i \le n$, del término no homogéneo $\mathbf{b}$; las componentes XO_i, con $1 \le i \le n$, de la aproximación inicial $\mathbf{XO} = \mathbf{x}^{(0)}$; la tolerancia TOL; el número máximo de iteraciones N_0.

Salida: solución aproximada $x_1, x_2, \dots, x_n$ ó mensaje de que el número de iteraciones fue excedido.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \le N_0$ seguir los pasos 3–6.

Paso 3: Para $i = 1, 2, \dots, n$ tomar

$$x_i = \frac{1}{a_{ii}}[-\sum_{\substack{j=1 \\ j\neq i}}^{n}(a_{ij}\ XO_j) + b_i]\ .$$

Paso 4: Si $||\mathbf{x} - \mathbf{XO}|| < TOL$ entonces SALIDA $(x_1, x_2, \dots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

Paso 5: Tomar $k = k + 1$.

Paso 6: Para $i = 1, 2, \ldots, n$ tomar $XO_i = x_i$.

Paso 7: SALIDA *(número máximo de iteraciones excedido)*;
(procedimiento completado sin éxito) PARAR.

==========

El paso 3 del algoritmo requiere que $a_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$. Si éste no es el caso, se puede realizar un reordenamiento de las ecuaciones para que ningún $a_{ii} = 0$, a menos que el sistema sea singular. Se sugiere que las ecuaciones sean arregladas de tal manera que a_{ii} sea lo más grande posible para acelerar la convergencia.

En el paso 4, el criterio de paro ha sido $||\mathbf{x} - \mathbf{XO}|| < TOL$; otro criterio de paro es iterar hasta que

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

sea menor que alguna tolerancia predeterminada $\varepsilon > 0$. Para este propósito, se puede usar cualquier norma conveniente; la que más se usa es la norma l_∞.

Un análisis de la ecuación $(XVII.3)$ sugiere una posible mejora en el algoritmo iterativo de Jacobi. Para calcular $x_i^{(k)}$, se usan las componentes de $\mathbf{x}^{(k-1)}$. Como para $i > 1$, $x_1^{(k)}$, $x_2^{(k)}$, ..., $x_{i-1}^{(k)}$ ya han sido calculadas y supuestamente son mejores aproximaciones a la solución real x_1, x_2, ..., x_{i-1} que $x_1^{(k)}$, $x_2^{(k)}$, ..., $x_{i-1}^{(k)}$, parece razonable calcular $x_i^{(k)}$ usando los valores calculados más recientemente; es decir,

$$x_i^{(k)} = \frac{1}{a_{ii}}[-\sum_{j=1}^{i-1}(a_{ij}\ x_j^{(k)}) - \sum_{j=i+1}^{n}(a_{ij}\ x_j^{(k-1)}) + b_i]\ , \qquad (XVII.6)$$

para cada $i = 1, 2, \ldots, n$ en vez de la ecuación $(XVII.3)$.

Ejemplo 2.

El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{array}{lrcrcrcrcr}
E_1: & 10\ x_1 & - & x_2 & + & 2\ x_3 & & & = & 6\ , \\
E_2: & -\ x_1 & + & 11\ x_2 & - & x_3 & + & 3\ x_4 & = & 25\ , \\
E_3: & 2\ x_1 & - & x_2 & + & 10\ x_3 & - & x_4 & = & -11\ , \\
E_4: & & & 3\ x_2 & - & x_3 & + & 8\ x_4 & = & 15\ ,
\end{array}$$

fue resuelto en el ejemplo 1 con el método iterativo de Jacobi. Incorporando la ecuación $(XVII.6)$ en el algoritmo iterativo de Jacobi, se obtienen las ecuaciones que se usarán para cada $k = 1, 2, \ldots$:

$$\begin{array}{lcccccccccc}
x_1^{(k)} = & & & \frac{1}{10}x_2^{(k-1)} & - & \frac{1}{5}x_3^{(k-1)} & & & + & \frac{3}{5}\ , \\
x_2^{(k)} = & \frac{1}{11}x_1^{(k)} & & & + & \frac{1}{11}x_3^{(k-1)} & - & \frac{3}{11}x_4^{(k-1)} & + & \frac{25}{11}\ , \\
x_3^{(k)} = & -\frac{1}{5}x_1^{(k)} & + & \frac{1}{10}x_2^{(k)} & & & + & \frac{1}{10}x_4^{(k-1)} & - & \frac{11}{10}\ , \\
x_4^{(k)} = & & - & \frac{3}{8}x_2^{(k)} & + & \frac{1}{8}x_3^{(k)} & & & + & \frac{15}{8}\ .
\end{array}$$

Tomando $\mathbf{x}^{(0)} = (0, 0, 0, 0)^t$, generamos los vectores iterados de la tabla 2.

Tabla 2

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$
0	0.0000	0.0000	0.0000	0.0000
1	0.6000	2.3273	−0.98727	0.87885
2	1.0302	2.0369	−1.0145	0.98435
3	1.0066	2.0035	−1.0025	0.99838
4	1.0009	2.0003	−1.0003	0.99985
5	1.0001	2.0000	−1.0000	1.0000

Ya que

$$\frac{||\mathbf{x}^{(5)} - \mathbf{x}^{(4)}||_\infty}{||\mathbf{x}^{(4)}||_\infty} = \frac{0.0008}{2.000} = 4 \times 10^{-4} \ ,$$

se acepta $\mathbf{x}^{(5)}$ como una aproximación razonable a la solución. Es interesante notar que el método de Jacobi en el ejemplo 1 requiere el doble de iteraciones para la misma precisión.

La técnica presentada en el ejemplo 2 se llama **método iterativo de Gauss-Seidel**. Para escribir este método en la forma matricial ($XVII.1$) se multiplican ambos lados de la ecuación ($XVII.6$) por a_{ii} y se recolectan todos los $k-$ésimos términos iterados para dar

$$a_{i1}\ x_1^{(k)} + a_{i2}\ x_3^{(k)} + \ldots + a_{ii}\ x_i^{(k)} = -a_{i,i+1}\ x_{i+1}^{(k-1)} - \ldots - a_{in}\ x_n^{(k-1)} + b_i \ ,$$

para cada $i = 1, 2, \ldots, n$. Escribiendo las n ecuaciones tenemos:

$$a_{11}\ x_1^{(k)} = -a_{12}\ x_2^{(k-1)} - a_{13}\ x_3^{(k-1)} - \ldots - a_{1n}\ x_n^{(k-1)} + b_1 \ ,$$

$$a_{21}\ x_1^{(k)} + a_{22}\ x_2^{(k)} = -a_{23}\ x_3^{(k-1)} \ldots - a_{2n}\ x_n^{(k-1)} + b_2 \ ,$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$

$$a_{n1}\ x_1^{(k)} + a_{n2}\ x_2^{(k)} + \ldots + a_{nn}\ x_n^{(k)} = b_n \ ,$$

y se sigue que, en forma matricial, el método de Gauss-Seidel puede ser representado como $(D - L)\ \mathbf{x}^{(k)} = U\ \mathbf{x}^{(k-1)} + \mathbf{b}$, ó

$$\mathbf{x}^{(k)} = (D - L)^{-1}\ U\ \mathbf{x}^{(k-1)} + (D - L)^{-1}\ \mathbf{b} \ . \qquad (XVII.7)$$

Para que la matriz triangular inferior $(D - L)$ sea no singular, es necesario y suficiente que $a_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$.

Para resumir el método iterativo de Gauss-Seidel, presentamos el siguiente algoritmo:

Algoritmo iterativo de Gauss-Seidel.

==========

Para resolver el sistema lineal $A\ \mathbf{x} = \mathbf{b}$ con una aproximación inicial dada $\mathbf{x}^{(0)}$.

Entrada: número de incógnitas y de ecuaciones n; las componentes de la matriz $A = (a_{ij})$ donde $1 \leq i, j \leq n$; las componentes b_i, con $1 \leq i \leq n$, del término no homogéneo $\mathbf{b}$; las

componentes XO_i, con $1 \leq i \leq n$, de la aproximación inicial $\mathbf{XO} = \mathbf{x}^{(0)}$; la tolerancia TOL; el número máximo de iteraciones N_0.

Salida: solución aproximada $x_1, x_2, \ldots, x_n$ ó mensaje de que el número de iteraciones fue excedido.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \leq N_0$ seguir los pasos 3–6.

Paso 3: Para $i = 1, 2, \ldots, n$ tomar

$$x_i = \frac{1}{a_{ii}}[-\sum_{j=1}^{i-1}(a_{ij}\ x_j) - \sum_{j=i+1}^{n}(a_{ij}\ XO_j) + b_i]\ .$$

Paso 4: Si $||\mathbf{x} - \mathbf{XO}|| < TOL$ entonces SALIDA $(x_1, x_2, \ldots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

Paso 5: Tomar $k = k + 1$.

Paso 6: Para $i = 1, 2, \ldots, n$ tomar $XO_i = x_i$.

Paso 7: SALIDA *(número máximo de iteraciones excedido)*;
(procedimiento completado sin éxito) PARAR.

==========

Los resultados de los ejemplos 1 y 2 parecen implicar que el método de Gauss-Seidel es superior al método de Jacobi. Este es generalmente cierto, pero no siempre. En realidad, hay sistemas lineales para los cuales el método de Jacobi converge y el método de Gauss-Seidel no, y viceversa.

3. CONVERGENCIA DE LOS PROCESOS ITERATIVOS

Para estudiar la convergencia de las técnicas generales de iteración, consideramos la fórmula $(XVII.1)$

$$\mathbf{x}^{(k)} = T\ \mathbf{x}^{(k-1)} + \mathbf{c}$$

para cada $k = 1, 2, \ldots$, donde $\mathbf{x}^{(0)}$ es arbitrario. Este estudio requerirá del siguiente lema:

Lema XVII.1

Si el radio espectral $\rho(T)$ satisface que $\rho(T) < 1$, ó si la norma de la matriz T satisface que $||T|| < 1$, entonces $(I - T)^{-1}$ existe y

$$(I - T)^{-1} = I + T + T^2 + \ldots\ .$$

Teorema XVII.2

Para cualquier $\mathbf{x}^{(0)} \in \mathcal{R}^n$, la sucesión $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ definida por $(XVII.1)$

$$\mathbf{x}^{(k)} = T\ \mathbf{x}^{(k-1)} + \mathbf{c}$$

para cada $k \geq 1$ y $\mathbf{c} \neq \mathbf{0}$, converge a la solución única de $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$ si y sólo si $\rho(T) < 1$.

Demostración: de la ecuación $(XVII.1)$, se tiene que

$$\begin{aligned}\mathbf{x}^{(k)} &= T\ \mathbf{x}^{(k-1)} + \mathbf{c} = \\ &= T\ (T\ \mathbf{x}^{(k-2)} + \mathbf{c}) + \mathbf{c} = \\ &= T^2\ \mathbf{x}^{(k-2)} + (T + I)\ \mathbf{c} = \\ &\quad \ldots \\ &= T^k\ \mathbf{x}^{(0)} + (T^{k-1} + \ldots + T + I)\ \mathbf{c}\ .\end{aligned}$$

Suponiendo que $\rho(T) < 1$, podemos usar el Teorema XIII.15 y el Lema XVII.1 para obtener

$$\begin{aligned}\lim_{k\to\infty} \mathbf{x}^{(k)} &= \lim_{k\to\infty} T^k\ \mathbf{x}^{(0)} + \lim_{k\to\infty} \big(\sum_{j=0}^{k-1} T^j\big)\mathbf{c} \\ &= 0 \cdot \mathbf{x}^{(0)} + (I - T)^{-1}\mathbf{c} = (I - T)^{-1}\mathbf{c}\ .\end{aligned}$$

De $(XVII.1)$ $\mathbf{x} = \lim_{k\to\infty} \mathbf{x}^{(k)} = (I - T)^{-1}\mathbf{c}$ será la solución única de $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$.

Para probar el recíproco, sea $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ convergente a $\mathbf{x}$ para cualquier $\mathbf{x}^{(0)}$. De la ecuación $(XVII.1)$ sigue que $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$, así que para cada k,

$$\mathbf{x} - \mathbf{x}^{(k)} = T\ (\mathbf{x} - \mathbf{x}^{(k-1)}) = \ldots = T^k\ (\mathbf{x} - \mathbf{x}^{(0)})\ .$$

Por lo tanto, para cualquier vector $\mathbf{x}^{(0)}$,

$$\lim_{k\to\infty} T^k\ (\mathbf{x} - \mathbf{x}^{(0)}) = \lim_{k\to\infty} \mathbf{x} - \mathbf{x}^{(k)} = \mathbf{0}\ .$$

Consecuentemente, si $\mathbf{z}$ es un vector arbitrario y $\mathbf{x}^{(0)} = \mathbf{x} - \mathbf{z}$, entonces

$$\lim_{k\to\infty} T^k\ \mathbf{z} = \lim_{k\to\infty} T^k\ [\mathbf{x} - (\mathbf{x} - \mathbf{z})] = \mathbf{0}\ ,$$

lo cual, por el Teorema XIII.15, implica que $\rho(T) < 1$. **c.q.d.**

Un Teorema parecido nos dará condiciones de suficiencia para la convergencia de los procesos de iteración usando las normas en lugar del radio espectral.

Teorema XVII.3

Si $||T|| < 1$, para cualquier norma matricial natural, entonces la sucesión definida en la ecuación $(XVII.1)$, $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$, converge para cualquier $\mathbf{x}^{(0)} \in \mathcal{R}^n$, a un vector $\mathbf{x} \in \mathcal{R}^n$, y se satisfacen las siguientes cotas de error:

$$||\mathbf{x} - \mathbf{x}^{(k)}|| \leq ||T||^k\ ||\mathbf{x}^{(0)} - \mathbf{x}||\ , \qquad (XVII.8)$$

y

$$||\mathbf{x} - \mathbf{x}^{(k)}|| \leq \frac{||T||^k}{1 - ||T||}\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}||\ . \qquad (XVII.9)$$

Demostración: comenzando con un vector arbitrario $\mathbf{x}^{(0)}$, formaremos una secuencia de aproximaciones

$$\mathbf{x}^{(1)} = T\ \mathbf{x}^{(0)} + \mathbf{c}\ ,$$

$$\mathbf{x}^{(2)} = T\ \mathbf{x}^{(1)} + \mathbf{c}\ ,$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots$$

$$\mathbf{x}^{(k)} = T\ \mathbf{x}^{(k-1)} + \mathbf{c}\ ,$$

de donde

$$\mathbf{x}^{(k)} = T^k\ \mathbf{x}^{(0)} + (T^{k-1} + \ldots + T + I)\ \mathbf{c}\ .$$

Como para $||T|| < 1$ tenemos $||T^k|| \to 0$ cuando $k \to \infty$, se deduce que

$$\lim_{k\to\infty} T^k = 0 \quad \text{y} \quad \lim_{k\to\infty} (I + T + T^2 + \ldots + T^{k-1}) = \sum_{k=0}^{\infty} T^k = (I - T)^{-1}\ .$$

Y por tanto, pasando al límite cuando $k \to \infty$, tenemos

$$\mathbf{x} = \lim_{k\to\infty} \mathbf{x}^{(k)} = (I - T)^{-1}\mathbf{c}\ .$$

Esto prueba la convergencia del proceso iterativo. Además, tenemos $(I - T)\ \mathbf{x} = \mathbf{c}$ ó $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$, lo cual quiere decir que el vector $\mathbf{x}$ en el límite es una solución del sistema. Como la matriz $(I - T)$ no es singular, la solución $\mathbf{x}$ es única. Hemos así demostrado la primera parte del Teorema.

Demostramos ahora la cota de error $(XVII.8)$. Supongamos que $\mathbf{x}^{(k+p)}$ y $\mathbf{x}^{(k)}$ son dos aproximaciones de la solución del sistema lineal $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$; de la ecuación $(XVII.1)$, tenemos:

$$\begin{aligned} ||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| &= ||T\ \mathbf{x}^{(k+p-1)} - T\ \mathbf{x}^{(k-1)}|| = ||T\ (\mathbf{x}^{(k+p-1)} - \mathbf{x}^{(k-1)})|| = \ldots \\ &= ||T^k\ (\mathbf{x}^{(p)} - \mathbf{x}^{(0)})|| \leq ||T||^k\ ||\mathbf{x}^{(p)} - \mathbf{x}^{(0)}||\ . \end{aligned}$$

Ahora pasando al límite cuando $p \to \infty$, obtenemos

$$\lim_{p\to\infty} ||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| \leq \lim_{p\to\infty} ||T||^k\ ||\mathbf{x}^{(p)} - \mathbf{x}^{(0)}|| = ||T||^k\ \lim_{p\to\infty} ||\mathbf{x}^{(p)} - \mathbf{x}^{(0)}||$$

y entonces

$$||\mathbf{x} - \mathbf{x}^{(k)}|| \leq ||T||^k\ ||\mathbf{x} - \mathbf{x}^{(0)}||\ ,$$

que es la cota de error $(XVII.8)$

Finalmente demostramos la cota de error $(XVII.9)$. Como antes, supongamos que $\mathbf{x}^{(k+p)}$ y $\mathbf{x}^{(k)}$ son dos aproximaciones de la solución del sistema lineal $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$. Tenemos

$$||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| \leq ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| + ||\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}|| + \ldots + ||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k+p-1)}||\ .$$

Por lo visto antes:

$$||\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}|| \leq ||T||\ ||\mathbf{x}^{(m)} - \mathbf{x}^{(m-1)}|| \leq ||T||^{m-k}\ ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| \ ,$$

para $m > k \geq 1$. Entonces tenemos:

$$\begin{aligned} ||\mathbf{x}^{(p+k)} - \mathbf{x}^{(k)}|| &\leq ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| + ||T||\ ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| + \ldots + \\ &+ ||T||^{p-1}\ ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| \leq \frac{1}{1-||T||}\ ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| \leq \\ &\leq \frac{||T||}{1-||T||}\ ||\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}|| \leq \ldots \leq \frac{||T||^k}{1-||T||}\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| \ , \end{aligned}$$

de donde se deduce la cota de error $(XVII.9)$. **c.q.d.**

Notése que si en particular elegimos $\mathbf{x}^{(0)} = \mathbf{c}$, entonces $\mathbf{x}^{(1)} = T\ \mathbf{c} + \mathbf{c}$ y

$$||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| = ||T\ \mathbf{c}|| \leq ||T||\ ||\mathbf{c}|| \ ,$$

y la cota $(XVII.9)$ nos da:

$$||\mathbf{x} - \mathbf{x}^{(k)}|| \leq \frac{||T||^{k+1}}{1-||T||}\ ||\mathbf{c}|| \ . \qquad (XVII.9')$$

Ejemplo 3.

Demostrar que el proceso de iteración de Jacobi es convergente para el sistema lineal siguiente:

$$\begin{array}{lrcrcrcrcr} E_1: & 10\ x_1 & - & x_2 & + & 2\ x_3 & - & 3\ x_4 & = & 0\ , \\ E_2: & x_1 & + & 10\ x_2 & - & x_3 & + & 2\ x_4 & = & 5\ , \\ E_3: & 2\ x_1 & + & 3\ x_2 & + & 20\ x_3 & - & x_4 & = & -10\ , \\ E_4: & 3\ x_1 & + & 2\ x_2 & + & x_3 & + & 20\ x_4 & = & 15\ . \end{array}$$

¿Cuántas iteraciones han de efectuarse para hallar las raíces del sistema con un error menor de 10^{-4}?

Reduciendo el sistema a la forma especial para la iteración de Jacobi, tenemos

$$\begin{array}{lrcrcrcrcr} x_1 = & & & 0.1\ x_2 & - & 0.2\ x_3 & + & 0.3\ x_4\ , & \\ x_2 = & -0.1\ x_1 & & & + & 0.1\ x_3 & - & 0.2\ x_4 & + & 0.5\ , \\ x_3 = & -0.1\ x_1 & - & 0.15\ x_2 & & & + & 0.05\ x_4 & - & 0.5\ , \\ x_4 = & -0.15\ x_1 & - & 0.1\ x_2 & - & 0.05\ x_3 & & & + & 0.75\ . \end{array}$$

Entonces la matriz del sistema es:

$$T = \begin{pmatrix} 0 & 0.1 & -0.2 & 0.3 \\ -0.1 & 0 & 0.1 & -0.2 \\ -0.1 & -0.15 & 0 & 0.05 \\ -0.15 & -0.1 & -0.05 & 0 \end{pmatrix} \ .$$

Utilizando, por ejemplo, la norma l_1, tenemos:

$$||T||_1 = \max\{0.35, 0.35, 0.35, 0.55\} = 0.55 < 1 \ .$$

En consecuencia el proceso de iteración para el sistema dado es convergente. Si consideramos como aproximación inicial de la raíz $\mathbf{x}$ el vector

$$\mathbf{x}^{(0)} = \mathbf{c} = (0.0, 0.5, -0.5, 0.75)^t \ ,$$

entonces

$$||\mathbf{c}||_1 = 0.0 + 0.5 + 0.5 + 0.75 = 1.75 \ .$$

Sea ahora k el número de iteraciones requeridas para conseguir la exactitud especificada. Utilizando la fórmula $(XVII.9')$, tenemos:

$$||\mathbf{x} - \mathbf{x}^{(k)}||_1 \leq \frac{||T||_1^{k+1}}{1 - ||T||_1} \ ||\mathbf{c}||_1 = \frac{0.55^{k+1} \times 1.75}{0.45} < 10^{-4} \ .$$

De aquí,

$$0.55^{k+1} < \frac{45}{175} \ 10^{-4}$$

o sea

$$(k+1) \log_{10} 0.55 < \log_{10} 45 - \log_{10} 175 - 4$$

$$-(k+1) \ 0.25964 < 1.65321 - 2.24304 - 4 = -4.58983$$

y consecuentemente

$$k + 1 > \frac{4.58983}{0.25964} \approx 17.7 \qquad \Longrightarrow \qquad k > 16.7 \ .$$

Podemos tomar $k = 17$. Notése que la estimación teórica del número de iteraciones necesarias para asegurar la exactitud especificada es excesivamente alto. A menudo se obtiene la exactitud deseada en un número menor de iteraciones.

Para aplicar los resultados de arriba a las técnicas iterativas de Jacobi o Gauss-Seidel, necesitamos escribir las matrices de iteración del método de Jacobi, T_J, dadas en $(XVII.5)$ y del método de Gauss-Seidel, T_{GS}, dadas en $(XVII.7)$, como

$$T_J = D^{-1} \ (L + U) \qquad \text{y} \qquad T_{GS} = (D - L)^{-1} \ U \ .$$

De ser $\rho(T_J)$ ó $\rho(T_{GS})$ menores que uno, es claro que la sucesión $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converge a la solución $\mathbf{x}$ de $A \ \mathbf{x} = \mathbf{b}$. Por ejemplo, el esquema de Jacobi (ver ecuación $(XVII.5)$) tiene:

$$\mathbf{x}^{(k)} = D^{-1} \ (L + U) \ \mathbf{x}^{(k-1)} + D^{-1} \ \mathbf{b} \ ,$$

y si $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converge a $\mathbf{x}$, entonces

$$\mathbf{x} = D^{-1} \ (L + U) \ \mathbf{x} + D^{-1} \ \mathbf{b} \ .$$

Esto implica que

$$D\,\mathbf{x} = (L+U)\,\mathbf{x} + \mathbf{b} \qquad \text{y} \qquad (D-L-U)\,\mathbf{x} = \mathbf{b}\ .$$

Ya que $D-L-U=A$, luego $\mathbf{x}$ satisface $A\,\mathbf{x}=\mathbf{b}$. De manera parecida se procede con el esquema de Gauss-Seidel dado por la ecuación $(XVII.7)$.

Podemos dar ahora condiciones de suficiencia fáciles de verificar para la convergencia de los métodos de Jacobi y de Gauss-Seidel.

Teorema XVII.4

Si A es una matriz estrictamente dominante diagonalmente, entonces, para cualquier elección de $\mathbf{x}^{(0)} \in \mathcal{R}^n$ ambos métodos, el de Jacobi o el de Gauss-Seidel, dan lugar a sucesiones $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ que convergen a la solución de $A\,\mathbf{x}=\mathbf{b}$.

La relación entre la rapidez de convergencia y el radio espectral de la matriz de iteración T se puede ver de la desigualdad $(XVII.8)$. Como $(XVII.8)$ se satisface para cualquier norma matricial natural se sigue, de la afirmación que siguió al Teorema XIII.14, que

$$||\mathbf{x}^{(k)} - \mathbf{x}|| \approx \rho(T)^k\ ||\mathbf{x}^{(0)} - \mathbf{x}||\ . \qquad (XVII.10)$$

Supongamos que $\rho(T) < 1$ y que se va a usar $\mathbf{x}^{(0)} = \mathbf{0}$ en una técnica iterativa para aproximar $\mathbf{x}$ con un error relativo máximo de 10^{-t}. Por la estimación $(XVII.10)$, el error relativo después de k iteraciones es aproximadamente $\rho(T)^k$, así que se espera una precisión de 10^{-t} si

$$\rho(T)^k \leq 10^{-t}\ ,$$

esto es, si

$$k \geq \frac{t}{-\log_{10}\rho(T)}\ .$$

Por lo tanto, es deseable escoger la técnica iterativa con el menor $\rho(T) < 1$ para el sistema particular $A\,\mathbf{x}=\mathbf{b}$.

En general no se conoce cuál de las dos técnicas, la de Jacobi o la de Gauss-Seidel, debe usarse. Sin embargo, en un caso especial, sí se conoce la respuesta.

Teorema XVII.5 (Stein-Rosenberg)

Si $a_{ij} \leq 0$ para cada $i \neq j$ y $a_{ii} > 0$ para cada $i = 1, 2, \ldots, n$, entonces se satisface una y solamente una de las siguientes afirmaciones:

a) $0 < \rho(T_{GS}) < \rho(T_J) < 1$;
b) $1 < \rho(T_J) < \rho(T_{GS})$;
c) $\rho(T_{GS}) = \rho(T_J) = 0$;
d) $\rho(T_J) = \rho(T_{GS}) = 1$;

Para el caso especial descrito en el Teorema XVII.5, vemos que cuando un método converge, entonces ambos convergen, siendo el método de Gauss-Seidel más rápido que el método de Jacobi.

4. LOS METODOS DE RELAJACION

Como la razón de convergencia de un procedimiento depende del radio espectral de la matriz asociada con el método, una manera de seleccionar un procedimiento que nos lleve a una convergencia acelerada consiste en escoger un método cuya matriz asociada tenga un radio espectral mínimo. Estos procedimientos nos llevan a los métodos de relajación. Pero antes de formular la teoría de los métodos de relajación, veamos las ideas fundamentales de la forma más simple. Supongamos que se dispone de un sistema de ecuaciones lineales

$$\begin{aligned} E_1: &\quad a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = b_1\ , \\ E_2: &\quad a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = b_2\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = b_n\ . \end{aligned} \tag{$XVII.11$}$$

Transformaremos este sistema de la manera siguiente: pondremos los términos constantes a la izquierda y dividiremos la primera ecuación por $-a_{11}$, la segunda por $-a_{22}$, etc. Obtendremos entonces un sistema que está listo para la relajación:

$$\begin{aligned} E_1: &\quad -\ x_1 + b_{12}\ x_2 + \ldots + b_{1n}\ x_n + c_1 = 0\ , \\ E_2: &\quad b_{21}\ x_1 - \quad x_2 + \ldots + b_{2n}\ x_n + c_2 = 0\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad b_{n1}\ x_1 + b_{n2}\ x_2 + \ldots - \quad x_n + c_n = 0\ , \end{aligned} \tag{$XVII.12$}$$

donde

$$b_{ij} = -\frac{a_{ij}}{a_{ii}} \quad (i \neq j) \qquad \text{y} \qquad c_i = \frac{b_i}{a_{ii}}\ . \tag{$XVII.13$}$$

Supongamos que $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)})$ es la aproximación inicial a la solución del sistema dado. Sustituyendo estos valores en el sistema tendremos los **restos**

$$\begin{aligned} R_1^{(0)} &= c_1 - x_1^{(0)} + \sum_{j=2}^{n} b_{1j}\ x_j^{(0)} = x_1^{(1)} - x_1^{(0)}\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ R_k^{(0)} &= c_k - x_k^{(0)} + \sum_{\substack{j=1 \\ j\neq k}}^{n} b_{kj}\ x_j^{(0)} = x_k^{(1)} - x_k^{(0)}\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ R_n^{(0)} &= c_n - x_n^{(0)} + \sum_{j=1}^{n-1} b_{nj}\ x_j^{(0)} = x_n^{(1)} - x_n^{(0)}\ . \end{aligned} \tag{$XVII.14$}$$

Si damos un incremento $\delta x_s^{(0)}$ a una de las incógnitas $x_s^{(0)}$, el resto correspondiente $R_s^{(0)}$ quederá disminuido en $\delta x_s^{(0)}$ y todos los otros restos $R_i^{(0)}$ $(i \neq s)$ quedarán aumentados en

$b_{is}\ \delta x_s^{(0)}$. De este modo, para hacer que desaparezca el resto siguiente $R_i^{(1)}$ es suficiente dar a $x_s^{(1)}$ un incremento $\delta x_s^{(1)} = R_s^{(0)}$ y tendremos

$$R_s^{(1)} = 0 \qquad \text{y} \qquad R_i^{(1)} = R_i^{(0)} + b_{is}\ \delta x_s^{(0)} \quad \text{para} \quad i \neq s\ . \qquad (XVII.15)$$

Así el **método de relajación**, en su forma más simple, consiste en reducir el resto numéricamente más elevado a cero, en cada etapa, cambiando el valor del componente apropiado de la aproximación. El proceso acaba cuando todos los restos del último sistema transformado son iguales a cero con la exactitud requerida.

Ejemplo 4.

Vamos a resolver el sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{array}{lrcrcrcrcl} E_1: & 10\ x_1 & - & 2\ x_2 & - & 2\ x_3 & = & 6\ , \\ E_2: & -\ x_1 & + & 10\ x_2 & - & 2\ x_3 & = & 7\ , \\ E_3: & -\ x_1 & - & x_2 & + & 10\ x_3 & = & 8\ , \end{array}$$

con el método de relajación y aritmética de dos dígitos.

Reduzcamos el sistema a una forma conveniente para la relajación:

$$\begin{array}{lrcrcrcrcl} E_1: & -\ x_1 & + & 0.2\ x_2 & + & 0.2\ x_3 & + & 0.6 & = & 0\ , \\ E_2: & 0.1\ x_1 & - & x_2 & + & 0.2\ x_3 & + & 0.7 & = & 0\ , \\ E_3: & 0.1\ x_1 & + & 0.1\ x_2 & - & x_3 & + & 0.8 & = & 0\ . \end{array}$$

Eligiendo los valores $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$ como aproximaciones iniciales de las raíces, obtenemos los restos:

$$R_1^{(0)} = 0.60, \qquad R_2^{(0)} = 0.70, \qquad R_3^{(0)} = 0.80.$$

Siendo $R_3^{(0)} = 0.80$ el resto más grande, consideramos

$$\delta x_3^{(0)} = 0.80$$

de donde obtendremos los restos

$$\begin{aligned} R_1^{(1)} &= R_1^{(0)} + 0.2 \times 0.8 = 0.60 + 0.16 = 0.76\ , \\ R_2^{(1)} &= R_2^{(0)} + 0.2 \times 0.8 = 0.70 + 0.16 = 0.86\ , \\ R_3^{(1)} &= R_3^{(0)} - 0.8 = 0.80 - 0.80 = 0.0\ . \end{aligned}$$

Establezcamos ahora

$$\delta x_2^{(1)} = 0.86$$

y así sucesivamente. Los resultados de los cálculos se dan en la tabla 3.

Sumando todos los incrementos $\delta x_i^{(k)}$ $(i = 1, 2, 3; k = 0, 1, \ldots)$, tendremos los valores de las raíces:

$$\begin{aligned} x_1 &= 0.0 + 0.93 + 0.062 + 0.007 = 0.999 = 1.0\ , \\ x_2 &= 0.0 + 0.86 + 0.13 + 0.016 = 1.006 = 1.0\ , \\ x_3 &= 0.0 + 0.80 + 0.18 + 0.019 = 0.999 = 1.0\ . \end{aligned}$$

En este caso el sistema se ha resuelto exactamente.

Tabla 3

k	x_1	R_1	x_2	R_2	x_3	R_3
0	0.0	0.60	0.0	0.70	0.0	0.80
		0.16		0.16	0.80	−0.80
		0.76		0.86		0.0
1		0.17	0.86	−0.86		0.086
		0.93		0.0		0.086
2	0.93	−0.93		0.093		0.093
		0.0		0.093		0.18
3		0.036		0.036	0.18	−0.18
		0.036		0.13		0.0
4		0.026	0.13	−0.13		0.013
		0.062		0.0		0.013
5	0.062	−0.062		0.012		0.0062
		0.0		0.012		0.019
6		0.0038		0.0038	0.019	−0.019
		0.0038		0.016		0.0
7		0.0032	0.016	−0.016		0.0016
		0.007		0.0		0.0016
8	0.007	−0.007		0.0007		0.0007
		0.0		0.0007		0.0023
$\sum$	1.0		1.0		1.0	

Vamos ahora a describir los métodos de relajación. Antes de describir un procedimiento para seleccionar tales métodos, necesitamos introducir una manera nueva de medir la cantidad por la cual una aproximación a la solución de un sistema lineal difiere de la solución real del sistema. El método hace uso del denominado **vector residual**.

Definición. Si $\tilde{\mathbf{x}} \in \mathcal{R}^n$ es una aproximación a la solución del sistema lineal definido por $A\,\mathbf{x} = \mathbf{b}$, el **vector residual** de $\tilde{\mathbf{x}}$ con respecto a este sistema se define como $\mathbf{r} = \mathbf{b} - A\,\tilde{\mathbf{x}}$.

En procedimientos como los métodos de Jacobi o de Gauss-Seidel se asocia un vector residual con cada cálculo de una componente aproximada del vector solución. El objetivo del método consiste en generar una sucesión de aproximaciones que hagan que los vectores residuales asociados converjan a cero. Supongamos que tomamos

$$\mathbf{r}_i^{(k)} = (r_{1i}^{(k)}, r_{2i}^{(k)}, \ldots, r_{ni}^{(k)})^t$$

para denotar al vector residual para el método de Gauss-Seidel correspondiente al vector solución aproximado

$$(x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k-1)}, \ldots, x_n^{(k-1)})^t \ .$$

La $m-$ésima componente de $\mathbf{r}_i^{(k)}$ es

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj}\ x_j^{(k)} - \sum_{j=i}^{n} a_{mj}\ x_j^{(k-1)} \qquad (XVII.16)$$

ó

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{mj}\ x_j^{(k-1)} - a_{mi}\ x_i^{(k-1)}$$

para cada $m = 1, 2, \ldots, n$. En particular, la $i-$ésima componente de $\mathbf{r}_i^{(k)}$ es

$$r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)} - a_{ii}\ x_i^{(k-1)} \ ;$$

así que

$$a_{ii}\ x_i^{(k-1)} + r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)} \ . \qquad (XVII.17)$$

Recuérdese, sin embargo, que en el método de Gauss-Seidel $x_i^{(k)}$ se escoge como

$$x_i^{(k)} = \frac{-\sum_{j=1}^{i-1}(a_{ij}\ x_j^{(k)}) - \sum_{j=i+1}^{n}(a_{ij}\ x_j^{(k-1)}) + b_i}{a_{ii}} \ , \qquad (XVII.6)$$

así que la ecuación $(XVII.17)$ puede escribirse como $a_{ii}\ x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii}\ x_i^{(k)}$ ó

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}} \ . \qquad (XVII.18)$$

Podemos derivar otra conexión entre los vectores residuales y la técnica de Gauss-Seidel. De $(XVII.16)$, la $i-$ésima componente de $\mathbf{r}_{i+1}^{(k)}$ es

$$\begin{aligned} r_{i,i+1}^{(k)} &= b_i - \sum_{j=1}^{i} a_{ij}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)} \\ &= b_i - \sum_{j=1}^{i-1} a_{ij}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)} - a_{ii}\ x_i^{(k)} \ . \end{aligned} \qquad (XVII.19)$$

La ecuación $(XVII.6)$ implica que $r_{i,i+1}^{(k)} = 0$. Entonces, en cierto sentido, la técnica de Gauss-Seidel está ideada para requerir que la $i-$ésima componente de $\mathbf{r}_{i+1}^{(k)}$ sea cero.

Reducir una coordenada del vector residual a cero, sin embargo, no es necesariamente la manera más eficiente de reducir la norma del vector $\mathbf{r}_{i+1}^{(k)}$. En realidad, modificando el procedimiento de Gauss-Seidel en la forma de la ecuación $(XVII.18)$ a:

$$x_i^{(k)} = x_i^{(k-1)} + \omega \ \frac{r_{ii}^{(k)}}{a_{ii}} \qquad (XVII.20)$$

para ciertas elecciones de ω positivo nos llevará a una convergencia significativamente más rápida.

Los métodos que emplean la ecuación $(XVII.20)$ se conocen como **métodos de relajación**. Para $0 < \omega < 1$, los procedimientos se llaman **métodos de sub-relajación** y se pueden emplear para obtener la convergencia de algunos sistemas que no son convergentes por el método de Gauss-Seidel. Para $\omega > 1$, los procedimientos se llaman **métodos de sobre-relajación** y se pueden usar para acelerar la convergencia de sistemas que son convergentes por el método de Gauss-Seidel. Estos métodos se abrevian frecuentemente como **SOR** (de **Successive Over-Relaxation**) y son particularmente útiles para resolver los sistemas lineales que aparecen en la solución numérica de ciertas ecuaciones diferenciales parciales.

Antes de ilustrar las ventajas del método SOR notamos que usando la ecuación $(XVII.17)$, la ecuación $(XVII.20)$ se puede reformular para propósitos de cómputo como

$$x_i^{(k)} = (1-\omega)\ x_i^{(k-1)} + \frac{\omega}{a_{ii}}\Big[b_i - \sum_{j=1}^{i-1} a_{ij}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)}\Big]\ . \qquad (XVII.21)$$

Para determinar la forma matricial del método SOR reescribimos $(XVII.21)$ como

$$a_{ii}\ x_i^{(k)} + \omega \sum_{j=1}^{i-1} a_{ij}\ x_j^{(k)} = (1-\omega)\ a_{ii}\ x_i^{(k-1)} - \omega \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)} + \omega\ b_i$$

así que

$$(D - \omega\ L)\ \mathbf{x}^{(k)} = [(1-\omega)\ D + \omega\ U]\ \mathbf{x}^{(k-1)} + \omega\ \mathbf{b}$$

ó

$$\mathbf{x}^{(k)} = (D - \omega\ L)^{-1}\ [(1-\omega)\ D + \omega\ U]\ \mathbf{x}^{(k-1)} + \omega\ (D - \omega\ L)^{-1}\ \mathbf{b}\ .$$

Algoritmo iterativo Successive Over-Relaxation (SOR).

==========

Para resolver el sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dados el parámetro ω y una aproximación inicial $\mathbf{x}^{(0)}$.

Entrada: número de incógnitas y de ecuaciones n; las componentes de la matriz $A = (a_{ij})$ donde $1 \le i, j \le n$; las componentes b_i, con $1 \le i \le n$, del término no homogéneo $\mathbf{b}$; las componentes XO_i, con $1 \le i \le n$, de la aproximación inicial $\mathbf{XO} = \mathbf{x}^{(0)}$; el parámetro ω; la tolerancia TOL; el número máximo de iteraciones N_0.

Salida: solución aproximada $x_1, x_2, \ldots, x_n$ ó mensaje de que el número de iteraciones fue excedido.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \le N_0$ seguir los pasos 3–6.

Paso 3: Para $i = 1, 2, \ldots, n$ tomar

$$x_i = (1-\omega)\ XO_i + \frac{\omega}{a_{ii}}\Big[-\sum_{j=1}^{i-1}(a_{ij}\ x_j) - \sum_{j=i+1}^{n}(a_{ij}\ XO_j) + b_i\Big]\ .$$

Paso 4: Si $||\mathbf{x} - \mathbf{XO}|| < TOL$ entonces SALIDA $(x_1, x_2, \ldots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

Paso 5: Tomar $k = k + 1$.

Paso 6: Para $i = 1, 2, \ldots, n$ tomar $XO_i = x_i$.

Paso 7: SALIDA *(número máximo de iteraciones excedido)*;
(procedimiento completado sin éxito) PARAR.

==========

Ejemplo 5.

El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{array}{lrcrcrcr} E_1: & 4\,x_1 & + & 3\,x_2 & & & = & 24\ , \\ E_2: & 3\,x_1 & + & 4\,x_2 & - & x_3 & = & 30\ , \\ E_3: & & - & x_2 & + & 4\,x_3 & = & -24\ , \end{array}$$

tiene por solución $\mathbf{x} = (3, 4, -5)^t$. Se usarán los métodos de Gauss-Seidel y el SOR con $\omega = 1.25$ para resolver este sistema usando $\mathbf{x}^{(0)} = (1, 1, 1)^t$ para ambos métodos. Las ecuaciones para el método de Gauss-Seidel son

$$\begin{aligned} x_1^{(k)} &= -0.75\ x_2^{(k-1)} + 6\ , \\ x_2^{(k)} &= -0.75\ x_1^{(k)} + 0.25\ x_3^{(k-1)} + 7.5\ , \\ x_3^{(k)} &= 0.25\ x_2^{(k)} - 6\ , \end{aligned}$$

para cada $k = 1, 2, \ldots$, y las ecuaciones para el método SOR con $\omega = 1.25$ son

$$\begin{aligned} x_1^{(k)} &= -0.25\ x_1^{(k-1)} - 0.9375\ x_2^{(k-1)} + 7.5\ , \\ x_2^{(k)} &= -0.9375\ x_1^{(k)} - 0.25\ x_2^{(k-1)} + 0.3125\ x_3^{(k-1)} + 9.375\ , \\ x_3^{(k)} &= 0.3125\ x_2^{(k)} - 0.25\ x_3^{(k-1)} - 7.5\ . \end{aligned}$$

Las primeras siete iteraciones de cada método se muestran en las tablas 4 y 5.

Para obtener una precisión de siete lugares decimales el método de Gauss-Seidel requiere de 34 iteraciones en contra de las 14 que se necesitan en el método de sobrerelajación con $\omega = 1.25$.

Tabla 4

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	1.000000	1.000000	1.000000
1	5.250000	3.812500	−5.046875
2	3.1406250	3.8828125	−5.0292969
3	3.0878906	3.9267578	−5.0183105
4	3.0549317	3.9542236	−5.0114441
5	3.0343323	3.9713898	−5.0071526
6	3.0214577	3.9821186	−5.0044703
7	3.0134111	3.9888241	−5.0027940

Tabla 5

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	1.000000	1.000000	1.000000
1	6.312500	3.5195313	−6.6501465
2	2.6223144	3.9585266	−4.6004238
3	3.1333027	4.0102646	−5.0966864
4	2.9570513	4.0074838	−4.9734897
5	3.0037211	4.0029250	−5.0057135
6	2.9963275	4.0009263	−4.9982822
7	3.0000498	4.0002586	−5.0003486

Un problema que se presenta al usar el método SOR, es cómo escoger el valor apropiado de ω. Aún cuando no se conoce una respuesta completa a esta pregunta para un sistema lineal general $n \times n$, los siguientes resultados pueden usarse en ciertas situaciones.

Teorema XVII.6 (Kahan)

Si $a_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$, entonces $\rho(T_\omega) \geq |\omega - 1|$. Esto implica que $\rho(T_\omega) < 1$ sólo si $0 < \omega < 2$, donde $T_\omega = (D - \omega\ L)^{-1}\ [(1 - \omega)\ D + \omega\ U]$ es la matriz de iteración del método SOR.

Teorema XVII.7 (Ostrowski-Reich)

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

Teorema XVII.8

Si A es una matriz positiva definida y tridiagonal, entonces $\rho(T_{GS}) = [\rho(T_J)]^2 < 1$, la elección óptima de ω para el método SOR es

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(T_J)]^2}}\ , \qquad (XVII.22)$$

y con este valor de ω, $\rho(T_\omega) = \omega - 1$.

Ejemplo 6.

En el ejemplo 5 la matriz A estaba dada por

$$A = \begin{pmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}\ .$$

Esta matriz es positiva definida y tridiagonal, así que se aplica el Teorema XVII.8. Como

$$\begin{aligned} T_J = D^{-1}\ (L + U) = \\ = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{pmatrix} \begin{pmatrix} 0 & -3 & 0 \\ -3 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \\ = \begin{pmatrix} 0 & -0.75 & 0 \\ -0.75 & 0 & 0.25 \\ 0 & 0.25 & 0 \end{pmatrix}\ , \end{aligned}$$

tenemos que

$$T_J - \lambda I = \begin{pmatrix} -\lambda & -0.75 & 0 \\ -0.75 & -\lambda & 0.25 \\ 0 & 0.25 & -\lambda \end{pmatrix} ,$$

con lo que

$$\det(T_J - \lambda I) = -\lambda(\lambda^2 - 0.625) .$$

Por lo tanto,

$$\rho(T_J) = \sqrt{0.625} \approx 0.79057$$

y la ecuación $(XVII.22)$ nos da

$$\omega = \frac{2}{1 + \sqrt{1 - \rho(T_{GS})}} = \frac{2}{1 + \sqrt{1 - [\rho(T_J)]^2}} = \frac{2}{1 + \sqrt{1 - 0.625}} \approx 1.24 .$$

Esto explica la rápida convergencia obtenida usando $\omega = 1.25$ en el ejemplo 5.

5. ELECCION DEL METODO PARA RESOLVER SISTEMAS LINEALES

Cuando el sistema lineal es lo suficientemente pequeño para que sea fácilmente acomodado en la memoria principal de un ordenador, es en general más eficaz usar una técnica directa que minimice el efecto del error de redondeo. Específicamente, es adecuado el algoritmo de eliminación Gaussiana con pivoteo escalado de columna.

Los sistemas lineales grandes cuyos coeficientes son entradas básicamente de ceros y que aparecen en patrones regulares se pueden resolver generalmente de una manera eficiente usando un procedimiento iterativo como el discutido en este capítulo. Los sistemas de este tipo aparecen naturalmente, por ejemplo, cuando se usan técnicas de diferencias finitas para resolver problemas de valor en la frontera, una aplicación común en la solución numérica de ecuaciones diferenciales parciales.

EJERCICIOS.

1. Encontrar las dos primeras iteraciones del método de Jacobi, del método de Gauss-Seidel y del método SOR para los siguientes sistemas lineales, usando $\mathbf{x}^{(0)} = \mathbf{0}$.

a)
$$\begin{aligned} 2\,x_1 - x_2 + x_3 &= -1\,, \\ 3\,x_1 + 3\,x_2 + 9\,x_3 &= 0\,, \\ 3\,x_1 + 3\,x_2 + 5\,x_3 &= 4\,. \end{aligned}$$

b)
$$\begin{aligned} 2\,x_2 + 4\,x_3 &= 0\,, \\ x_1 - x_2 - x_3 &= 0.375\,, \\ x_1 - x_2 + 2\,x_3 &= 0\,. \end{aligned}$$

c)
$$\begin{aligned} 10\,x_1 - x_2 &= 9\,, \\ -\,x_1 + 10\,x_2 - 2\,x_3 &= 7\,, \\ -\,2\,x_2 + 10\,x_3 &= 6\,. \end{aligned}$$

d)
$$\begin{aligned} 2\,x_1 &= 3\,, \\ x_1 + 1.5\,x_2 &= 4.5\,, \\ -\,3\,x_2 + 0.5\,x_3 &= -6.6\,, \\ 2\,x_1 - 2\,x_2 + x_3 + x_4 &= 0.8\,. \end{aligned}$$

e)
$$\begin{aligned} 2\,x_1 - x_2 + 10\,x_3 &= -11\,, \\ 3\,x_2 - x_3 + 8\,x_4 &= -11\,, \\ 10\,x_1 - x_2 + 2\,x_3 &= 6\,, \\ -\,x_1 + 11\,x_2 - x_3 + 3\,x_4 &= 25\,. \end{aligned}$$

f)
$$\begin{aligned} 10\,x_1 - x_2 + 2\,x_3 &= 6\,, \\ -\,x_1 + 11\,x_2 - x_3 + 3\,x_4 &= 25\,, \\ 2\,x_1 - x_2 + 10\,x_3 &= -11\,, \\ 3\,x_2 - x_3 + 8\,x_4 &= -11\,. \end{aligned}$$

g)
$$\begin{aligned} 4\,x_1 - 2\,x_2 &= 0\,, \\ -2\,x_1 + 5\,x_2 - x_3 &= 2\,, \\ -\,x_2 + 4\,x_3 + 2\,x_4 &= 3\,, \\ 2\,x_3 + 3\,x_4 &= -2\,. \end{aligned}$$

2. Aplicar, si es posible, los algoritmos iterativos de Jacobi, de Gauss-Seidel y el algoritmo SOR, con $\omega = 1.2$, para resolver los sistemas lineales del ejercicio 1. Usar $TOL = 10^{-2}$ y el número máximo de iteraciones $N = 25$.

CAPITULO XVIII. ESTIMACIONES DE ERROR Y REFINAMIENTO ITERATIVO

1. ESTIMACIONES DE ERROR

Parece razonable intuitivamente que si $\tilde{\mathbf{x}}$ es una aproximación a la solución $\mathbf{x}$ de $A\ \mathbf{x} = \mathbf{b}$ y el vector residual $\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}}$ tiene la propiedad de que $||\mathbf{r}||$ es pequeño, entonces $||\mathbf{x} - \tilde{\mathbf{x}}||$ será también peque˜no. Aún cuando éste es frecuentemente el caso, ciertos sistemas especiales, que aparecen bastante en la práctica, no tienen esta propiedad.

Ejemplo 1.

El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3.0001 \end{pmatrix} ,$$

tiene la solución única $\mathbf{x} = (1, 1)^t$. La aproximación a esta solución $\tilde{\mathbf{x}} = (3, 0)^t$ tiene vector residual

$$\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}} = \begin{pmatrix} 3 \\ 3.0001 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -0.0002 \end{pmatrix} ,$$

así que $||\mathbf{r}||_\infty = 0.0002$.

Aunque la norma del vector residual es pequeña, la aproximación $\tilde{\mathbf{x}} = (3, 0)^t$ es obviamente bastante pobre; en realidad, $||\mathbf{x} - \tilde{\mathbf{x}}||_\infty = 2$.

Esta dificultad se puede explicar muy simplemente si se observa que la solución del sistema representa la intersección de las rectas

$$l_1 : \quad x_1 + 2\ x_2 = 3 \quad \text{y} \quad l_2 : \quad 1.0001\ x_1 + 2\ x_2 = 3.0001\ .$$

El punto $(3, 0)$ se encuentra en l_1 y las rectas son casi paralelas. Esto implica que $(3, 0)$ se encuentra también cerca de l_2, aún cuando difiere significativamente del punto de intersección $(1, 1)$. Si las rectas no hubieran sido casi paralelas, se esperería que un vector residual pequeño implicara una aproximación precisa.

En general, no podemos depender de la geometría del sistema para obtener una indicación de cúando pueden presentarse problemas. Sin embargo, podemos extraer esta información considerando las normas de la matriz A y de su inversa.

Definición. El **número de condición** $K(A)$ de la matriz no singular A relativo a la norma $||\ \cdot\ ||$ se define como

$$K(A) = ||A||\ ||A^{-1}|| .$$

Teorema XVIII.1

Si $\tilde{\mathbf{x}}$ es una aproximación a la solución de $A\ \mathbf{x} = \mathbf{b}$ y A es una matriz no singular, entonces para cualquier norma natural,

$$||\mathbf{x} - \tilde{\mathbf{x}}|| \leq ||\mathbf{r}||\ ||A^{-1}|| = K(A)\frac{||\mathbf{r}||}{||A||} \qquad (XVIII.1)$$

y

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||}{||\mathbf{x}||} \leq ||A^{-1}||\ ||A||\ \frac{||\mathbf{r}||}{||\mathbf{b}||} = K(A)\ \frac{||\mathbf{r}||}{||\mathbf{b}||}\ , \qquad (XVIII.2)$$

siempre que $\mathbf{x} \neq 0$ y $\mathbf{b} \neq 0$, donde $\mathbf{r}$ es el vector residual de $\tilde{\mathbf{x}}$ con respecto al sistema $A\ \mathbf{x} = \mathbf{b}$.

Demostración: como $\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}} = A\ \mathbf{x} - A\ \tilde{\mathbf{x}}$ y A no es singular:

$$||\mathbf{x}-\tilde{\mathbf{x}}|| = ||A^{-1}\ \mathbf{r}|| \leq ||A^{-1}||\ ||\mathbf{r}||\ .$$

Además, como $\mathbf{b} = A\ \mathbf{x}$, $||\mathbf{b}|| \leq ||A||\ ||\mathbf{x}||$; así que

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||}{||\mathbf{x}||} \leq ||A^{-1}||\ ||A||\ \frac{||\mathbf{r}||}{||\mathbf{b}||}\ .$$

c.q.d.

Las desigualdades $(XVIII.1)$ y $(XVIII.2)$ implican que las cantidades $||A^{-1}||$ y $K(A) = ||A||\ ||A^{-1}||$ pueden ser usadas para dar una indicación de la conexión entre el vector residual y la precisión de la aproximación. En general, el error relativo $||\mathbf{x} - \tilde{\mathbf{x}}||/||\mathbf{x}||$ es de mayor interés y por la desigualdad $(XVIII.2)$ este error está acotado por el producto del número de condición $K(A) = ||A||\ ||A^{-1}||$ con el residual relativo para esta aproximación $||\mathbf{r}||/||\mathbf{b}||$. Para esta aproximación puede usarse cualquier norma que sea conveniente, el único requisito es que se use consistentemente desde el principio hasta el final.

Ya que para cualquier matriz no singular A

$$1 = ||I|| = ||A \cdot A^{-1}|| \leq ||A||\ ||A^{-1}|| = K(A)\ ,$$

se espera que la matriz A tenga un buen comportamiento (llamada formalmente una **matriz bien condicionada**) si $K(A)$ está cerca de uno y un comportamiento defectuoso (llamada una **matriz mal condicionada**) cuando $K(A)$ sea significativamente mayor que uno. El comportamiento en esta situación se refiere a la relativa seguridad de que un vector residual peque˜no implique correspondientemente una solución aproximada precisa.

Ejemplo 2.

La matriz del sistema considerado en el ejemplo 1 es

$$A = \begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix}\ ,$$

que tiene $||A||_\infty = 3.0001$. Esta norma no se considera grande, sin embargo

$$A^{-1} = \begin{pmatrix} -10000 & 10000 \\ 5000.5 & -5000 \end{pmatrix}\ ,$$

y $||A^{-1}||_\infty = 20000$ y para la norma infinita $K(A) = 20000 \times 3.0001 = 60002$. El tama˜no del número de condición para este ejemplo seguramente nos detendría al tomar decisiones apresuradas acerca de la precisión, basadas en el residual de la aproximación.

Mientras que, en teoría, el número de condición de una matriz depende totalmente de las normas de la matriz y de su inversa, en la práctica, el cálculo de la inversa está sujeto a errores de redondeo y es dependiente de la exactitud con la que se estén haciendo los cálculos. Si hacemos la suposición de que la solución aproximada al sistema lineal $A\ \mathbf{x} = \mathbf{b}$ se determina usando aritmética de t dígitos y eliminación Gaussiana, se puede demostrar que el vector residual $\mathbf{r}$ para la aproximación $\tilde{\mathbf{x}}$ tiene la propiedad

$$||\mathbf{r}|| \approx 10^{-t}\ ||A||\ ||\tilde{\mathbf{x}}|| \ . \qquad (XVIII.3)$$

De esta ecuación aproximada, se puede obtener una estimación del número de condición efectivo para la aritmética de t dígitos, sin la necesidad de invertir la matriz A. [La aproximación en la ecuación $(XVIII.3)$ supone que todas las operaciones aritméticas en la técnica de eliminación Gaussiana se efectúan usando aritmética de t dígitos, pero que las operaciones que se necesitan para determinar el residual se hacen en doble precisión, es decir, $2t$ dígitos, para eliminar la pérdida de precisión involucrada en la sustracción de números casi iguales que ocurre en los cálculos del residual].

La aproximación del número de condición $K(A)$ a t dígitos viene de considerar el sistema lineal $A\ \mathbf{y} = \mathbf{r}$. La solución de este sistema puede aproximarse fácilmente ya que los multiplicadores para el método de eliminación Gaussiana han sido ya calculados y supuestamente retenidos. De hecho $\tilde{\mathbf{y}}$, la solución aproximada de $A\ \mathbf{y} = \mathbf{r}$, satisface que

$$\tilde{\mathbf{y}} \approx A^{-1}\ \mathbf{r} = A^{-1}\ (\mathbf{b} - A\ \tilde{\mathbf{x}}) = A^{-1}\ \mathbf{b} - A^{-1}\ A\ \tilde{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{x}}\ ; \qquad (XVIII.4)$$

así que $\tilde{\mathbf{y}}$ es una estimación del error cometido al aproximar la solución del sistema original. Consecuentemente la ecuación $(XVIII.3)$ puede usarse para deducir que

$$\begin{aligned} ||\tilde{\mathbf{y}}|| &\approx ||\mathbf{x} - \tilde{\mathbf{x}}|| = ||A^{-1}\ \mathbf{r}|| \leq \\ &\leq ||A^{-1}||\ ||\mathbf{r}|| \approx ||A^{-1}||\ (10^{-t}\ ||A||\ ||\tilde{\mathbf{x}}||) = 10^{-t}\ ||\tilde{\mathbf{x}}||\ K(A)\ . \end{aligned}$$

Esto proporciona una aproximación para el número de condición involucrado en la solución del sistema $A\ \mathbf{x} = \mathbf{b}$ usando eliminación Gaussiana y el tipo de aritmética de t dígitos descrito anteriormente:

$$K(A) \approx 10^t\ \frac{||\tilde{\mathbf{y}}||}{||\tilde{\mathbf{x}}||}\ . \qquad (XVIII.5)$$

Ejemplo 3.

El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{pmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.71 & 9.612 \\ 1.5611 & 5.1791 & 1.6852 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 15913 \\ 28.544 \\ 8.4254 \end{pmatrix}\ ,$$

tiene la solución exacta $\mathbf{x} = (1, 1, 1)^t$.

Usando eliminación Gaussiana y aritmética de redondeo de 5 dígitos llegamos sucesivamente a las matrices ampliadas

$$\left(\begin{array}{ccc|c} 3.3330 & 15920 & -10.333 & 15913 \\ 0 & -10596 & 16.501 & -10580 \\ 0 & -7451.4 & 6.525 & -7444.9 \end{array}\right)$$

y

$$\left(\begin{array}{ccc|c} 3.3330 & 15920 & -10.333 & 15913 \\ 0 & -10596 & 16.501 & -10580 \\ 0 & 0 & -5.079 & -4.7 \end{array}\right) .$$

La solución aproximada a este sistema es

$$\tilde{\mathbf{x}} = (1.2001, 0.99991, 0.92538)^t .$$

El vector residual correspondiente a $\tilde{\mathbf{x}}$ calculado con doble precisión (y luego redondeado a cinco dígitos) es

$$\begin{aligned} \mathbf{r} = \mathbf{b} - A\,\tilde{\mathbf{x}} = \\ = \begin{pmatrix} 15913 \\ 28.544 \\ 8.4254 \end{pmatrix} - \begin{pmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.71 & 9.612 \\ 1.5611 & 5.1791 & 1.6852 \end{pmatrix} \begin{pmatrix} 1.2001 \\ 0.99991 \\ 0.92538 \end{pmatrix} = \\ = \begin{pmatrix} -0.0051818 \\ 0.27413 \\ -0.18616 \end{pmatrix} ; \end{aligned}$$

así que

$$||\mathbf{r}||_\infty = 0.27413 .$$

La estimación del número de condición dada en la discusión anterior se obtiene resolviendo primero el sistema $A\,\mathbf{y} = \mathbf{r}$:

$$\begin{pmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.71 & 9.612 \\ 1.5611 & 5.1791 & 1.6852 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -0.0051818 \\ 0.27413 \\ -0.18616 \end{pmatrix} ,$$

lo cual implica que $\tilde{\mathbf{y}} = (-0.20008, 8.9989 \times 10^{-5}, 0.074607)^t$. Usando la estimación dada por la ecuación $(XVIII.5)$:

$$K(A) \approx 10^5\ \frac{||\tilde{\mathbf{y}}||_\infty}{||\tilde{\mathbf{x}}||_\infty} = \frac{10^5\ (0.20008)}{1.2001} = 16672 .$$

Las cotas de error dadas en el Teorema XVIII.1 para estos valores son

$$||\mathbf{x} - \tilde{\mathbf{x}}||_\infty \leq K(A)\frac{||\mathbf{r}||_\infty}{||A||_\infty} = \frac{(16672)(0.27413)}{15934} = 0.28683$$

y

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||_\infty}{||\mathbf{x}||_\infty} \leq K(A)\ \frac{||\mathbf{r}||_\infty}{||\mathbf{b}||_\infty} = \frac{(16672)(0.27413)}{15913} = 0.28721\ .$$

Para determinar el número de condición exacto de A, necesitamos construir primero A^{-1}. Usando aritmética de redondeo de 5 dígitos para los cálculos se obtiene la aproximación:

$$A^{-1} = \begin{pmatrix} -1.1701 \times 10^{-4} & -1.4983 \times 10^{-1} & 8.5416 \times 10^{-1} \\ 6.2782 \times 10^{-5} & 1.2124 \times 10^{-4} & -3.0662 \times 10^{-4} \\ -8.6631 \times 10^{-5} & 1.3846 \times 10^{-1} & -1.9689 \times 10^{-1} \end{pmatrix}\ .$$

El Teorema XIII.13 puede usarse para demostrar que $||A^{-1}||_\infty = 1.0041$ y $||A||_\infty = 15934$. Como consecuencia la matriz A mal condicionada tiene

$$K(A) = (1.0041)\ (15934) = 15999\ .$$

La aproximación que habíamos obtenido antes está bastante cerca de este $K(A)$ y ha requerido un esfuerzo computacional considerablemente menor.

Como la solución real $\mathbf{x} = (1, 1, 1)^t$ de este sistema es conocida, podemos calcular ambos

$$||\mathbf{x}-\tilde{\mathbf{x}}||_\infty = 0.2001$$

y

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||_\infty}{||\mathbf{x}||_\infty} = 0.2001\ .$$

Las cotas de error dadas en el Teorema XVIII.1 para estos valores son

$$||\mathbf{x}-\tilde{\mathbf{x}}||_\infty \leq K(A)\frac{||\mathbf{r}||_\infty}{||A||_\infty} = \frac{(15999)(0.27413)}{15934} = 0.27525$$

y

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||_\infty}{||\mathbf{x}||_\infty} \leq K(A)\ \frac{||\mathbf{r}||_\infty}{||\mathbf{b}||_\infty} = \frac{(15999)(0.27413)}{15913} = 0.27561\ .$$

2. REFINAMIENTO ITERATIVO

En la ecuación $(XVIII.4)$ usamos la estimación $\tilde{\mathbf{y}} \approx \mathbf{x}-\tilde{\mathbf{x}}$, en la que $\tilde{\mathbf{y}}$ es la solución aproximada al sistema $A\ \mathbf{y} = \mathbf{r}$. Sería razonable sospechar, a partir de este resultado, que $\tilde{\mathbf{x}} + \tilde{\mathbf{y}}$ fuese una mejor aproximación a la solución del sistema lineal $A\ \mathbf{x} = \mathbf{b}$ que la aproximación inicial $\tilde{\mathbf{x}}$.

El método que usa esta suposición se llama **refinamiento iterativo**, o mejora iterativa y consiste en llevar a cabo iteraciones sobre el sistema cuyo lado derecho es el vector residual para las aproximaciones sucesivas, hasta que se obtiene una precisión satisfactoria. El procedimiento se usa generalmente sólo en los sistemas en que se sospecha que la matriz involucrada es mal condicionada, debido a que esta técnica no mejora mucho la aproximación para un sistema bien condicionado.

Algoritmo de refinamiento iterativo.

==========

Para aproximar la solución al sistema lineal $A\,\mathbf{x} = \mathbf{b}$ cuando se sospecha que A sea mal condicionada.

Entrada: número de incógnitas y de ecuaciones n; las componentes de la matriz $A = (a_{ij})$ donde $1 \leq i, j \leq n$; las componentes b_i, con $1 \leq i \leq n$, del término no homogéneo $\mathbf{b}$; la tolerancia TOL; el número máximo de iteraciones N_0.

Salida: solución aproximada $\mathbf{xx} = (xx_1, xx_2, \ldots, xx_n)$ ó mensaje de que el número de iteraciones fue excedido.

Paso 0: Resolver el sistema $A\,\mathbf{x} = \mathbf{b}$ para $x_1, x_2, \ldots, x_n$ por eliminación Gaussiana guardando los multiplicadores m_{ji}, $j = i+1, i+2, \ldots, n$, $i = 1, 2, \ldots, n-1$ y haciendo notar los intercambios de filas.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \leq N_0$ seguir los pasos 3–8.

Paso 3: Para $i = 1, 2, \ldots, n$ (*calcular* $\mathbf{r}$*, realizando los cálculos con doble precisión aritmética*), tomar

$$r_i = b_i - \sum_{j=1}^{n} (a_{ij}\ x_j).$$

Paso 4: Resolver el sistema lineal $A\,\mathbf{y} = \mathbf{r}$ usando eliminación Gaussiana en el mismo orden que en el paso 0.

Paso 5: Para $i = 1, 2, \ldots, n$ tomar

$$xx_i = x_i + y_i.$$

Paso 6: Si $||\mathbf{x} - \mathbf{xx}|| < TOL$ entonces SALIDA $(xx_1, xx_2, \ldots, xx_n)$; *(procedimiento completado satisfactoriamente)* PARAR.

Paso 7: Tomar $k = k + 1$.

Paso 8: Para $i = 1, 2, \ldots, n$ tomar $x_i = xx_i$.

Paso 9: SALIDA *(número máximo de iteraciones excedido)*; *(procedimiento completado sin éxito)* PARAR.

==========

Si se está usando aritmética de t dígitos, un procedimiento recomendable para parar en el paso 6 consiste en iterar hasta que $|y_i^{(k)}| \leq 10^{-t}$ para cada $i = 1, 2, \ldots, n$.

Debe enfatizarse que la técnica de refinamiento iterativo no da resultados satisfactorios para todos los sistemas que contienen matrices mal condicionadas. En particular, si $K(A) \geq 10^t$, es probable que el procedimiento falle y que la única alternativa sea el uso de mayor precisión en los cálculos.

Ejemplo 4.

En el ejemplo 3 encontramos que la aproximación al problema que habíamos estado considerando, usando aritmética de cinco dígitos y la eliminación Gaussiana, era

$$\tilde{\mathbf{x}}^{(1)} = (1.2001, 0.99991, 0.92538)^t$$

y que la solución a $A\,\mathbf{y}^{(1)} = \mathbf{r}^{(1)}$ era

$$\tilde{\mathbf{y}}^{(1)} = (-0.20008, 8.9989 \times 10^{-5}, 0.074607)^t\ .$$

Usando el paso 5 del algoritmo, tenemos que

$$\tilde{\mathbf{x}}^{(2)} = \tilde{\mathbf{x}}^{(1)} + \tilde{\mathbf{y}}^{(1)} = (1.0000, 1.0000, 0.99999)^t$$

y el error real en esta aproximación es

$$||\mathbf{x} - \tilde{\mathbf{x}}^{(2)}||_\infty = 1.0 \times 10^{-5} \ .$$

Usando la técnica de paro sugerida para el algoritmo, calculamos $\mathbf{r}^{(2)} = \mathbf{b} - A\ \tilde{\mathbf{x}}^{(2)}$, y resolvemos el sistema $A\ \mathbf{y}^{(2)} = \mathbf{r}^{(2)}$, obteniéndose

$$\tilde{\mathbf{y}}^{(2)} = (-2.7003 \times 10^{-8}, 1.2973 \times 10^{-8}, 9.9817 \times 10^{-6})^t \ .$$

Puesto que $||\tilde{\mathbf{y}}^{(2)}||_\infty \leq 10^{-5}$, concluímos que

$$\tilde{\mathbf{x}}^{(3)} = \tilde{\mathbf{x}}^{(2)} + \tilde{\mathbf{y}}^{(2)} = (1.0000, 1.0000, 1.0000)^t$$

es suficientemente preciso. De hecho es claramente correcto.

EJERCICIOS.

1. Resolver los siguientes sistemas lineales usando eliminación Gaussiana y refinamiento iterativo (con aritmética a dos dígitos para el problema *a*) y a tres dígitos para *b*) y *c*)). Calcular $K(A)$.

$$a)\quad \begin{aligned} 3.9\,x_1 &+ 1.6\,x_2 = 5.5\,,\\ 6.8\,x_1 &+ 2.9\,x_2 = 9.7\,. \end{aligned}$$

$$b)\quad \begin{aligned} 4.56\,x_1 &+ 2.18\,x_2 = 6.74\,,\\ 2.79\,x_1 &+ 1.38\,x_2 = 4.13\,. \end{aligned}$$

$$c)\quad \begin{aligned} x_1 &+ \tfrac{1}{2}\,x_2 + \tfrac{1}{3}\,x_3 = \tfrac{11}{6}\,,\\ 5\,x_1 &+ \tfrac{10}{3}\,x_2 + \tfrac{5}{2}\,x_3 = \tfrac{65}{6}\,,\\ \tfrac{100}{3}\,x_1 &+ 25\,x_2 + 20\,x_3 = \tfrac{235}{3}\,. \end{aligned}$$

2. Calcular los números de condición de las siguientes matrices, relativos a $||\cdot||_\infty$.

$$a)\ \begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix}\,,$$

$$b)\ \begin{pmatrix} 3.9 & 1.6 \\ 6.8 & 2.9 \end{pmatrix}\,,$$

$$c)\ \begin{pmatrix} 4.56 & 2.18 \\ 2.79 & 1.38 \end{pmatrix}\,,$$

$$d)\ \begin{pmatrix} 1.003 & 58.09 \\ 5.550 & 321.8 \end{pmatrix}\,.$$

3. El sistema lineal

$$\begin{pmatrix} 1. & 2. \\ 1.0001 & 2. \end{pmatrix}\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3. \\ 3.0001 \end{pmatrix}$$

tiene solución $(1,1)^t$. Cambiar la matriz A por

$$\begin{pmatrix} 1. & 2. \\ 0.9999 & 2. \end{pmatrix}$$

y considerar el sistema lineal

$$\begin{pmatrix} 1. & 2. \\ 0.9999 & 2. \end{pmatrix}\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3. \\ 3.0001 \end{pmatrix}\,.$$

Calcular la nueva solución $\tilde{\mathbf{x}}$ usando aritmética de cinco dígitos.

4. El sistema lineal

$$\begin{pmatrix} 1. & 2. \\ 1.00001 & 2. \end{pmatrix}\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3. \\ 3.00001 \end{pmatrix}$$

tiene solución $(1,1)^t$. Encontrar la solución del sistema lineal perturbado

$$\begin{pmatrix} 1. & 2. \\ 1.000011 & 2. \end{pmatrix}\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3.00001 \\ 3.00003 \end{pmatrix}\,,$$

usando aritmética de siete dígitos.

CURSO DE MÉTODOS NUMÉRICOS

CUARTA PARTE

MÉTODOS DE MÍNIMO CUADRADOS

CAPITULO XIX. EL PROBLEMA DE LOS MINIMOS CUADRADOS: PRELIMINARES

1. SISTEMAS LINEALES DE ECUACIONES SOBREDETERMINADOS

La discusión de los problemas algebráicos de la parte anterior se había centrado exclusivamente en la ecuación $A\,\mathbf{x} = \mathbf{b}$, en la cual se ha intentado hacer coincidir exactamente el vector $\mathbf{b}$ que aparece a la derecha con el vector $A\,\mathbf{x}$ que aparece a la izquierda, es decir hacer que $\mathbf{b}$ sea exactamente una combinación lineal de las columnas de A. En muchos casos, el vector $\mathbf{b}$ no se puede expresar de la forma $A\,\mathbf{x}$, y el sistema $A\,\mathbf{x} = \mathbf{b}$ es **incompatible**. El problema lineal de mínimos cuadrados se acerca a la optimización si se encuentra un vector $\mathbf{x}$ que en algún sentido da la **mejor** (aunque no la perfecta) aproximación entre $A\,\mathbf{x}$ y $\mathbf{b}$.

Considérese un sistema de m ecuaciones con n incógnitas escrito en la forma

$$A\,\mathbf{x} = \mathbf{b}\ . \qquad (XIX.1)$$

Aquí, A es de $m \times n$, $\mathbf{x}$ es de $n \times 1$ y $\mathbf{b}$ es de $m \times 1$. Supondremos que el rango de A es n, de lo que resulta que $m \geq n$. Si $m > n$, se dice que el sistema lineal de ecuaciones es **sobredeterminado**. Los sistemas sobredeterminados son generalmente no compatibles.

Ejemplo 1.

a) Sea dado el sistema de ecuaciones lineales

$$\begin{aligned} x_1 + x_2 &= 1\ , \\ x_1 - x_2 &= 3\ , \\ -x_1 + 2\,x_2 &= -2\ . \end{aligned}$$

Las tres ecuaciones del sistema representano rectas en el plano. Las primeras dos se intersecan en el punto $(2, -1)$, mientras que la tercera no pasa por este punto. Entonces, no existe ningun punto común entre las tres líneas. El sistema es incompatible.

b) Sea dado el sistema de ecuaciones lineales

$$\begin{aligned} x_1 + 2\,x_2 + x_3 &= 1\ , \\ 2\,x_1 - x_2 + x_3 &= 2\ , \\ 4\,x_1 + 3\,x_2 + 3\,x_3 &= 4\ , \\ 2\,x_1 - x_2 + 3\,x_3 &= 5\ , \end{aligned}$$

Usando eliminación Gaussiana se deduce que el sistema tiene exactamente una única solución $(0.1, -0.3, 1.5)$. El sistema es compatible determinado.

c) Sea dado el sistema de ecuaciones lineales

$$\begin{aligned} x_1 + 2\,x_2 + x_3 &= 1\ , \\ 2\,x_1 - x_2 + x_3 &= 2\ , \\ 4\,x_1 + 3\,x_2 + 3\,x_3 &= 4\ , \\ 3\,x_1 + x_2 + 2\,x_3 &= 3\ , \end{aligned}$$

Usando eliminación Gaussiana y resolviendo x_2 y x_1 en función de x_3 se deduce que el sistema tiene una infinidad de soluciones de la forma $(1 - 0.6\alpha, -0.2\alpha, \alpha)$. El sistema es compatible indeterminado.

2. EL VECTOR RESIDUAL Y EL PROBLEMA DE LOS MINIMOS CUADRADOS

Por lo general, el sistema $(XIX.1)$ no tiene solución debido a que $\mathbf{b}$ no pertenece al subespacio de $\mathcal{R}^n$ de dimensión n, generado por las columnas de A. Es frecuente en tales casos que se requiera encontrar un $\mathbf{x}$ tal que $\mathbf{b}$ esté cerca de $A\ \mathbf{x}$, es decir que minimice una norma del vector residual $\mathbf{r} = \mathbf{b} - A\ \mathbf{x}$. El primer problema será determinar el sentido de cercanía, es decir la norma que mide la diferencia entre $\mathbf{b}$ y $A\ \mathbf{x}$. La interpretación y la naturaleza de la solución de este problema varía dependiendo de la norma que se use. La norma más comúnmente usada es la norma euclidiana, la norma l_2. Entonces, la *solución* en mínimos cuadrados de $(XIX.1)$ es el vector $\mathbf{x}$ que hace de $||\mathbf{r}||_2 = ||\mathbf{b} - A\ \mathbf{x}||_2$ un mínimo

$$\min_{\mathbf{x} \in \mathcal{R}^n} ||\mathbf{b} - A\ \mathbf{x}||_2^2\ , \qquad (XIX.2)$$

Dado que las normas no pueden ser negativas, minimizar una norma es equivalente a minimizar su cuadrado; hableremos entonces libremente de normas cuadradas y no. Según lo que se ha supuesto acerca del rango de A, la solución $\mathbf{x}$ será única.

La formulación de los mínimos cuadrados es popular sobre todo por el hecho de que resolver un problema de mínimos cuadrados lineal es más fácil que minimizar otras normas de $\mathbf{r}$. Veremos que el problema de mínimos cuadrados se puede resolver transformando la matriz A y resolviendo un sistema compatible a ella relacionado. Al contrario, minimizar la norma l_1 o la norma l_∞ es equivalente a un problema de programación lineal no diferenciable, cuya solución necesita una técnica de iteración.

Es útil pensar en el problema de los mínimos cudrados en términos de los subespacios definidos por la matriz A. El subespacio rango de A, rango(A), consiste en todos los vectores m-dimensionales que son combinaciones lineales de las columnas de A y el subespacio complementario, el subespacio nulo de A^t, nulo(A), contiene todos los vectores m-dimensionales que son $\mathbf{z}$-ortogonales a las columnas de A, es decir tales que $A^t\ \mathbf{z} = 0$. Si r denota el rango de la matriz A, entonces r es la dimensión del subespacio rango de A y $(m - r)$ es la dimensión del subespacio nulo de A^t.

Dada una matriz no nula A, cualquier vector m-dimensional $\mathbf{c}$ se puede expresar como la suma de un vector $\mathbf{c_R}$ en el rango de A, $\mathbf{c_R} \in$ rango(A), y de un vector $\mathbf{c_N}$ en el espacio nulo de A^t, $\mathbf{c_N} \in$ nulo(A^t), es decir

$$\mathbf{c} = \mathbf{c_R} + \mathbf{c_N}\ . \qquad (XIX.3)$$

Los vectores $\mathbf{c_R}$ y $\mathbf{c_N}$ son únicos y satisfacen

$$\mathbf{c_R} = A\ \mathbf{c_A}\ , \qquad A^t\ \mathbf{c_N} = 0\ , \qquad \mathbf{c_R}^t \cdot \mathbf{c_N} = 0\ , \qquad (XIX.4)$$

donde $\mathbf{c_A}$ se refiere a un vector n-dimensional cualquiera tal que $\mathbf{c_R} = A\ \mathbf{c_A}$. La unicidad de $\mathbf{c_R}$ y $\mathbf{c_N}$ implica que las componentes en el espacio rango y en el espacio nulo de vectores iguales tienen que ser iguales. En efecto,

$$\mathbf{c} = \mathbf{d} \quad \text{si y sólo si} \quad \mathbf{c_R} = \mathbf{d_R} \quad \text{y} \quad \mathbf{c_N} = \mathbf{d_N}\ . \tag{XIX.5}$$

Aunque $\mathbf{c_R}$ es único, el vector $\mathbf{c_A}$ es único sólo si las columnas de A son linealmente independientes.

Debido al hecho de que la norma euclídea está definida en términos del producto escalar, las relaciones (XIX.3) y (XIX.4) tienen una consecuencia muy importante:

$$||\mathbf{c}||_2^2 = \mathbf{c^t} \cdot \mathbf{c} = ||\mathbf{c_R}||_2^2 + ||\mathbf{c_N}||_2^2\ , \tag{XIX.6}$$

de manera que la representación de un vector en términos de sus componentes en el espacio rango y nulo separa el cuadrado de la norma euclídea en dos partes independientes. Esta propiedad no vale para otras normas.

Estas observaciones son relevantes para el problema de los mínimos cuadrados dado que permiten determinar la más pequeña posible norma euclídea del vector residual $\mathbf{r} = \mathbf{b} - A\ \mathbf{x}$ para todos vectores $\mathbf{x}$. Dado que ambos $\mathbf{b}$ y el vector residual $\mathbf{r}$ son vectores m-dimensionales, ambos pueden escribirse en la representación (XIX.3):

$$\begin{aligned} \mathbf{b} &= \mathbf{b_R} + \mathbf{b_N} \quad \text{con} \quad \mathbf{b_R} = A\ \mathbf{b_A}\ , \\ \mathbf{r} &= \mathbf{r_R} + \mathbf{r_N} \quad \text{con} \quad \mathbf{r_R} = A\ \mathbf{r_A}\ . \end{aligned} \tag{XIX.7}$$

Combinando la definición de $\mathbf{r}$ como $\mathbf{b} - A\ \mathbf{x}$ con estas relaciones, se obtiene

$$\begin{aligned} \mathbf{r} = \mathbf{r_R} + \mathbf{r_N} &= \mathbf{b} - A\ \mathbf{x} = \mathbf{b_R} + \mathbf{b_N} - A\ \mathbf{x} \\ &= \mathbf{b_R} - A\ \mathbf{x} + \mathbf{b_N} \end{aligned} \tag{XIX.8}$$

Un punto obvio pero crucial en esta expresión es que el vector $A\ \mathbf{x}$ está enteramente en el rango de A. Cuando se resta $A\ \mathbf{x}$ del vector $\mathbf{b}$ para crear el residual, sigue de (XIX.5) que la componente en el espacio rango de $\mathbf{b} - A\ \mathbf{x}$ tiene que ser $\mathbf{b_R} - A\ \mathbf{x}$. En contraste, el restar $A\ \mathbf{x}$ no puede eliminar nada a la componente de b en el espacio nulo, de manera que la componente en el espacio nulo de $\mathbf{b} - A\ \mathbf{x}$ es igual a $\mathbf{b_N}$. Entonces, las componentes en el espacio rango y en el nulo del vector residual satisfacen

$$\mathbf{r_R} = \mathbf{b_R} - A\ \mathbf{x} \quad \text{y} \quad \mathbf{r_N} = \mathbf{b_N}\ . \tag{XIX.9}$$

Como ya dicho, el problema de mínimos cuadrados consiste en minimizar la norma l_2 del vector residual. De (XIX.6) y (XIX.9) sigue que la norma euclídea del vector residual $\mathbf{r}$ para cualquier vector $\mathbf{x}$ satisface

$$||\mathbf{r}||_2^2 = ||\mathbf{b_R} - A\ \mathbf{x}||_2^2 + ||\mathbf{b_N}||_2^2 \geq ||\mathbf{b_N}||_2^2\ . \tag{XIX.10}$$

Dado que $\mathbf{b_N}$ se mantiene en el residual, se encontrará el mínimo de $||\mathbf{b} - A\ \mathbf{x}||_2$ cuando la norma de la componente en el subespacio rango, $||\mathbf{b_R} - A\ \mathbf{x}||_2$, sea la más pequeña

posible. Además, por definición $\mathbf{b_R}$ está en el rango de A, luego tiene que existir un vector $\mathbf{x}$ tal que $A\,\mathbf{x} = \mathbf{b_R}$. Para este vector particular $\mathbf{x}$, se elimina la componente de $\mathbf{b}$ en el espacio rango cuando se le resta $A\,\mathbf{x}$, que significa que $\mathbf{b} - A\,\mathbf{x} = \mathbf{b_N}$ y que la norma euclídea del vector residual es igual a su cota inferior $||\mathbf{b_N}||_2$.

Escoger $\mathbf{x}$ de esta manera, $A\,\mathbf{x} = \mathbf{b_R}$, no sólo minimiza la norma del vector residual, sino que además fuerza al residual a estar enteramente en el espacio nulo de A^t. De esta manera se llega a dos caracterizaciones equivalentes de la solución optimal del problema de mínimos cuadrados:

$$\mathbf{x} \quad \text{minimiza} \quad ||\mathbf{b} - A\,\mathbf{x}||_2 \quad \Longleftrightarrow \quad A^t\,(\mathbf{b} - A\,\mathbf{x}) = \mathbf{0}\;, \tag{XIX.11}$$

$$\mathbf{x} \quad \text{minimiza} \quad ||\mathbf{b} - A\,\mathbf{x}||_2 \quad \Longleftrightarrow \quad \begin{cases} A\,\mathbf{x} = \mathbf{b_R}\;, \\ \mathbf{b} - A\,\mathbf{x} = \mathbf{b_N}\;. \end{cases} \tag{XIX.12}$$

Un problema de mínimos cuadrados se dice **compatible** si su vector residual optimal es cero, en este caso $\mathbf{b_N} = 0$ y $\mathbf{b} = \mathbf{b_R}$; si éste no es el caso se dice **incompatible**.

La ortogonalidad del vector residual respecto de las filas de la matriz A corresponde al principio geométrico familiar de que la distancia más corta de un punto a un plano es la longitud de la perpendicular que los une. Aquí el punto es el vector $\mathbf{b}$ y el plano es el subespacio rango de A.



(a) $\mathbf{b} \in R^2$ and A is a 2 × 1 matrix of rank 1. (b) $\mathbf{b} \in R^3$ and A is a 3 × 2 matrix of rank 2.

La unicidad del vector $\mathbf{b_N}$ implica que el vector residual optimal para el problema de mínimos cuadrados es único. El vector $\mathbf{b_R}$ es también único, y el sistema $A\,\mathbf{x} = \mathbf{b_R}$ es compatible por definición. Sin embargo, el vector $\mathbf{x}$ que satisface $A\,\mathbf{x} = \mathbf{b_R}$ es único si y sólo si las columnas de la matriz A son linealmente independientes. Dado que cualquier vector $\mathbf{x}$ que satisface $A\,\mathbf{x} = \mathbf{b_R}$ es una solución de mínimos cuadrados, sigue una importante conclusión: **la solución del problema de mínimos cuadrados es única si y sólo si la matriz A es de rango n.**

Ejemplo 2.

Sean

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix}$$

con las columnas de A linealmente independientes. Por definición, cualquier vector $\mathbf{y}$ en el rango de A tiene la forma

$$\mathbf{y} = A\mathbf{x} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + x_2 \\ x_1 \\ x_2 \end{pmatrix}\;.$$

El espacio nulo de A^t consiste de vectores $\mathbf{z}$ que satisfacen

$$A^t\mathbf{z} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} z_1 + z_2 \\ z_1 + z_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} .$$

Las relaciones $z_1 + z_2 = 0$ y $z_1 + z_3 = 0$ implican que cualquier vector $\mathbf{z}$ del espacio nulo de A^t tiene que ser de la forma

$$\mathbf{z} = \begin{pmatrix} x_3 \\ -x_3 \\ -x_3 \end{pmatrix}$$

para algún escalar x_3. Entonces, para buscar la representación del vector $\mathbf{b}$ como suma de vectores en el subespacio rango de A y en el subespacio nulo de A^t tiene que ser

$$\mathbf{b} = \mathbf{y} + \mathbf{z} = \begin{pmatrix} x_1 + x_2 \\ x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} x_3 \\ -x_3 \\ -x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix}$$

que nos da $x_1 = x_3 = 2$ y $x_2 = -3$.

Se deduce entonces, que la representación del vector $\mathbf{b}$ como suma de vectores en el subespacio rango de A y en el subespacio nulo de A^t viene dada por

$$\mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix} = \mathbf{b_R} + \mathbf{b_N} = \begin{pmatrix} -1 \\ 2 \\ -3 \end{pmatrix} + \begin{pmatrix} 2 \\ -2 \\ -2 \end{pmatrix} =$$

$$= A\ \mathbf{b_A} + \mathbf{b_N} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ -3 \end{pmatrix} + \begin{pmatrix} 2 \\ -2 \\ -2 \end{pmatrix} .$$

Aplicando $(XIX.12)$, se obtiene que la solución de mínimos cuadrados y el vector residual optimal son

$$\mathbf{x} = \begin{pmatrix} 2 \\ -3 \end{pmatrix} \quad \text{y} \quad \mathbf{r} = \begin{pmatrix} 2 \\ -2 \\ -2 \end{pmatrix} .$$

En este ejemplo, la solución de mínimos cuadrados es única porque A es de rango lleno.

Un aspecto crucial de la relación $(XIX.12)$ es que revela cómo cambios en el vector $\mathbf{b}$ afectan a la solución $\mathbf{x}$ y al problema de mínimos cuadrados $\min ||\mathbf{b} - A\ \mathbf{x}||_2^2$. Para sistemas de ecuaciones lineales no singulares y compatibles, cambios en $\mathbf{b}$ implicaban encontrar una solución $\mathbf{x}$ diferente. Esta propiedad no se verifica para el problema de los mínimos cuadrados, cuyo analisis es más complicado.

Se supone por simplicidad que A tiene rango lleno, entonces $\mathbf{x}$ es la única solución del problema de mínimos cuadrados $\min ||\mathbf{b} - A\ \mathbf{x}||_2^2$ y $\mathbf{r}$ es el vector residual optimal, $\mathbf{r} = \mathbf{b} - A\ \mathbf{x}$. Si $\mathbf{b}$ viene perturbado por un vector que está enteramente en el subespacio rango de A, es decir $\tilde{\mathbf{b}} = \mathbf{b} + \delta\mathbf{b_R}$, donde $\delta\mathbf{b_R} = A\ \delta\mathbf{x}$ para un único n-vector $\delta\mathbf{x}$, entonces

$$\tilde{\mathbf{b}}_\mathbf{R} = \mathbf{b_R} + \delta\mathbf{b_R} = \mathbf{b_R} + A\ \delta\mathbf{x} \quad \text{y} \quad \tilde{\mathbf{b}}_\mathbf{N} = \mathbf{b_N} . \qquad (XIX.13)$$

Sigue entonces de $(XIX.12)$ que la solución de mínimos cuadrados $\tilde{\mathbf{x}}$ y el residual $\tilde{\mathbf{r}}$ correspondientes a $\tilde{\mathbf{b}}$ satisfacen

$$\tilde{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x} \quad \text{y} \quad \tilde{\mathbf{r}} = \mathbf{r} \ , \tag{$XIX.14$}$$

mostrando que cambios en $\mathbf{b}$ en el espacio rango modifican el vector solución $\mathbf{x}$ pero no el vector residual $\mathbf{r}$. Por el contrario, si $\mathbf{b}$ viene perturbado por un vector que está enteramente en el subespacio nulo de A^t, es decir $\overline{\mathbf{b}} = \mathbf{b} + \mathbf{z}$, donde $A^t\ \mathbf{z} = \mathbf{0}$, entonces

$$\overline{\mathbf{x}} = \mathbf{x} \quad \text{y} \quad \overline{\mathbf{r}} = \mathbf{r} + \mathbf{z} \ . \tag{$XIX.15$}$$

mostrando que cambios en $\mathbf{b}$ en el espacio nulo modifican el vector residual $\mathbf{r}$, mientras que el vector solución $\mathbf{x}$ permanece inalterado.

Ejemplo 3.

Para ver un ejemplo especifico de estos efectos, reconsideremos el ejemplo 2:

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix}$$

con las columnas de A linealmente independientes.

Primero perturbamos el vector $\mathbf{b}$ con un vector $\delta\mathbf{b_R}$ que está enteramente el en subespacio rango de A, es decir

$$\delta\mathbf{b_R} = A\ \delta\mathbf{x} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \quad \text{con} \quad \delta\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

y sea

$$\tilde{\mathbf{b}} = \mathbf{b} + \delta\mathbf{b_R} = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ -4 \end{pmatrix} \ .$$

La solución de mínimos cuadrados $\tilde{\mathbf{x}}$ asociada con $\tilde{\mathbf{b}}$ puede eliminar completamente el cambio hecho en $\mathbf{b}$ y se cambia por un vector $\delta\mathbf{x}$; mientras que el vector residual optimal se queda el mismo dado que $\tilde{\mathbf{b}}_\mathbf{N} = \mathbf{b_N}$. Los resultados son:

$$\tilde{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x} = \begin{pmatrix} 2 \\ -3 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ -2 \end{pmatrix} \quad \text{y} \quad \tilde{\mathbf{r}} = \mathbf{r} = \begin{pmatrix} 2 \\ -2 \\ -2 \end{pmatrix}$$

Por otro lado, supongamos que se perturbe el vector $\mathbf{b}$ con un vector en el subespacio nulo de A^t, por ejemplo

$$\mathbf{z} = \begin{pmatrix} 10 \\ -10 \\ -10 \end{pmatrix}$$

de manera que

$$\overline{\mathbf{b}} = \mathbf{b} + \mathbf{z} = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix} + \begin{pmatrix} 10 \\ -10 \\ -10 \end{pmatrix} = \begin{pmatrix} 11 \\ -10 \\ -15 \end{pmatrix} .$$

Aunque $\overline{\mathbf{b}}$ es significativamente distinto de $\mathbf{b}$, ningun posible vector $\mathbf{x}$ puede capturar la perturbación $\mathbf{z}$, que tiene que ser absorbida completamente en el vector residual optimal. La solución de mínimos cuadrados $\overline{\mathbf{x}}$ correspondiente a $\overline{\mathbf{b}}$ es entonces

$$\overline{\mathbf{x}} = \mathbf{x} = \begin{pmatrix} 2 \\ -3 \end{pmatrix} \quad \text{y} \quad \overline{\mathbf{r}} = \mathbf{r} + \mathbf{z} = \begin{pmatrix} 2 \\ -2 \\ -2 \end{pmatrix} + \begin{pmatrix} 10 \\ -10 \\ -10 \end{pmatrix} = \begin{pmatrix} 12 \\ -12 \\ -12 \end{pmatrix} .$$

3. LAS ECUACIONES NORMALES

La propiedad más significativa del vector residual optimal del problema de mínimos cuadrados es que está enteramente en el espacio nulo de la matriz A^t. En términos algebráicos, una solución $\mathbf{x}$ satisface

$$A^t \ (\mathbf{b} - A \ \mathbf{x}) = \mathbf{0} \ , \tag{XIX.16}$$

que es equivalente a

$$A^t A \ \mathbf{x} = A^t \ \mathbf{b} \ . \tag{XIX.17}$$

Este sistema de ecuaciones se conoce como el de las **ecuaciones normales**.

La matriz simétrica $A^t A$, conocida como la **matriz de la ecuación normal**, es semidefinida positiva para una matriz cualquiera A, y es definida positiva si y sólo si las columnas de A son linealmente independientes. Las ecuaciones normales son siempre compatibles, es decir existe siempre una solución de (XIX.17) aunque $A^t A$ sea singular. Cualquier solución $\mathbf{x}$ tiene que satisfacer las ecuaciones normales, y cualquier vector $\mathbf{x}$ que satisface las ecuaciones normales tiene que ser una solución de mínimos cuadrados.

Las ecuaciones normales son de gran importancia práctica por el hecho de que ofrecen una manera directa para calcular la solución de mínimos cuadrados. Si $A^t A$ es definida positiva, es decir si A tiene rango lleno, las ecuaciones normales (XIX.17) tienen solución única, que se puede encontrar usando el algoritmo de factorización de Cholesky. Si $A^t A$ es singular, se puede hallar una solución de las ecuaciones normales usando una versión de la factorización de Cholesky que incluya un intercambio simétrico. Entonces, el problema de mínimos cuadrados se puede resolver siguiendo los siguientes pasos:
(i) formar la matriz de la ecuación normal $A^t A$ y el vector $A^t \ \mathbf{b}$;
(ii) calcular la factorización de Cholesky $A^t A = L^t L$, con L triangular superior;
(iii) resolver los dos sistemas $L^t \ \mathbf{y} = A^t \ \mathbf{b}$ y $L \ \mathbf{x} = \mathbf{y}$. El vector $\mathbf{x}$ es la solución deseada.

Ejemplo 4.

Como ejemplo especifico de matriz de rango lleno, reconsideremos el ejemplo 2:

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix}$$

con las columnas de A linealmente independientes. La matriz de la ecuación normal A^tA y el vector $A^t\ \mathbf{b}$ para el vector asociado $\mathbf{b}$ son:

$$A^t\ A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad \text{y} \quad A^t\ \mathbf{b} = \begin{pmatrix} 1 \\ -4 \end{pmatrix}\ ,$$

y la solución de las ecuaciones normales

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -4 \end{pmatrix}$$

es

$$\mathbf{x} = \begin{pmatrix} 2 \\ -3 \end{pmatrix}$$

que es la solución de mínimos cuadrados obtenida anteriormente.

Ejemplo 5.

Para ver un ejemplo especifico de matriz de rango deficiente, consideremos una matriz variación de la matriz usada en el ejemplo anterior, creada añadiendo una columna dependente a la matriz A, y el mismo vector $\mathbf{b}$:

$$\overline{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix}$$

ahora las columnas de $\overline{A}$ son linealmente dependientes. La matriz de la ecuación normal $\overline{A}^t\overline{A}$ y el vector $\overline{A}^t\ \mathbf{b}$ para el vector asociado $\mathbf{b}$ son:

$$\overline{A}^t\overline{A} = \begin{pmatrix} 2 & 1 & \frac{3}{2} \\ 1 & 2 & \frac{3}{2} \\ \frac{3}{2} & \frac{3}{2} & \frac{3}{2} \end{pmatrix} \quad \text{y} \quad \overline{A}^t\mathbf{b} = \begin{pmatrix} 1 \\ -4 \\ -\frac{3}{2} \end{pmatrix}\ .$$

Aunque $\overline{A}^t\overline{A}$ es obviamente singular (la última columna es la media de las primeras dos columnas), las ecuaciones normales son compatibles, y el vector $\mathbf{x} = (2, -3, 0)^t$ (la solución anterior ampliada con una componente nula correspondiente a la última columna de $\overline{A}$) resuelve la ecuaciones normales. La solución de mínimos cuadrados ya no es única, y cualquiera del conjunto infinito de vectores que satisfacen las ecuaciones normales singulares per compatibles es una solución de mínimos cuadrados. Por ejemplo, $\mathbf{x_1} = (3, -2, -2)^t$ y $\mathbf{x_2} = (1, -4, 2)^t$ resuelven las ecuaciones normales y son consecuentemente soluciones de mínimos cuadrados.

Desde el punto de vista computacional, resolver el problema de mínimos cuadrados con las ecuaciones normales es eficiente: para formar A^tA y $A^t\mathbf{b}$ se necesitan aproximadamente $m \cdot n^2/2$ operaciones, calcular la factorización de Cholesky usa del orden de $n^3/6$ operaciones y resolver los dos sistemas triangulares implica aproximadamente n^2 operaciones. Desafortunadamente, el uso de las ecuaciones normales presenta el problema de la estabilidad numérica, dado que el número de condición de la matriz A^tA es el cuadrado

del número de condición de A. Consecuentemente, la matriz de la ecuación normal está seriamente mal-condicionada si la matriz A misma está ligeramente mal-condicionada.

Ejemplo 6.

Sea

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 10^{-4} \end{pmatrix}$$

para la cual

$$\text{cond}(A) = 2.8 \times 10^4 \ .$$

La matriz de la ecuación normal asociada A^tA es

$$A^tA = \begin{pmatrix} 2 & 2 \\ 2 & 2 + 10^{-8} \end{pmatrix}$$

para la cual

$$\text{cond}(A^tA) = 7.8 \times 10^8 \ .$$

El mal condicionamiento de la matriz de la ecuación normal asociada puede conducir no solo a una aproximación no precisa de la solución calculada de las ecuaciones normales, sino también a una perdida de información cuando el rango numérico de A es marginal.

4. APLICACIONES

Los científicos a menudo coleccionan datos e intentan encontrar una relación funcional entre las variables. Si los datos son $n+1$ puntos del plano, es posible encontrar un polinomio de grado n o inferior que pasa por todos los puntos. Este polinomio se llama **polinomio de interpolación**. Dado que los datos tienen en general error experimental, no hay razón para pedir que las funciones pasen por todos los puntos. De hecho, polinomios de grado inferior que no pasan por los puntos de manera exacta, dan una mejor descripción de la relación entre variables. Por ejemplo, si la relación entre variables es actualmente lineal y los datos tienen un pequeño error, sería desastroso usar un polinomio de interpolación.

Dada una tabla de datos

x	x_1	x_2	$\ldots$	x_m
y	y_1	y_2	$\ldots$	y_m

$(XIX.18)$

deseamos encontrar una función lineal

$$y = c_0 + c_1 \ x \qquad (XIX.19)$$

que mejor aproxima los datos en el sentido de mínimos cuadrados. Si se pide que

$$y_i = c_0 + c_1 \ x_i \qquad \text{para} \quad i = 1, 2, \ldots, m \qquad (XIX.20)$$

obtenemos un sistema de m ecuaciones en dos incognitas

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} . \qquad (XIX.21)$$

La función lineal cuyos coeficientes son la solución de mínimos cuadrados de $(XIX.21)$ viene a ser la aproximación de mínimos cuadrados a los datos con una función lineal.

Ejemplo 7.

Queremos encontrar la mejor aproximación de mínimos cuadrados con una función lineal a los siguientes datos:

x	0	3	6
y	1	4	5

Para estos datos el sistema $(XIX.21)$ es $A\ \mathbf{c} = \mathbf{y}$ con

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 6 \end{pmatrix} , \qquad \mathbf{c} = \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} \quad \text{y} \quad \mathbf{y} = \begin{pmatrix} 1 \\ 4 \\ 5 \end{pmatrix} .$$

Las ecuaciones normales, $A^t A\ \mathbf{c} = A^t\ \mathbf{y}$, se reducen a

$$\begin{pmatrix} 3 & 9 \\ 9 & 45 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} 10 \\ 42 \end{pmatrix} .$$

La solución a este sistema es $(\frac{4}{3}, \frac{2}{3})$. Entonces la mejor aproximación de mínimos cuadrados linear es

$$y = \frac{4}{3} + \frac{2}{3}\ x\ .$$

Podriamos reconsiderar este ejemplo usando el vector residual $\mathbf{r}(\mathbf{c}) = \mathbf{y} - A\ \mathbf{c}$ y

$$\begin{aligned} ||\mathbf{r}(\mathbf{c})||_2^2 &= ||\mathbf{y} - A\ \mathbf{c}||_2^2 = \\ &= [1 - (c_0 + 0\ c_1)]^2 + [4 - (c_0 + 3\ c_1)]^2 + [5 - (c_0 + 6\ c_1)]^2 = \\ &= f(c_0, c_1)\ . \end{aligned}$$

Entonces $||\mathbf{r}(\mathbf{c})||_2^2$ puede pensarse como una función $f(c_0, c_1)$ de dos variables. El minímo de tale función se presentará cuando sus derivadas parciales son nulas:

$$\frac{\partial f}{\partial c_0} = -2(10 - 3\ c_0 - 9\ c_1) = 0$$

$$\frac{\partial f}{\partial c_1} = -6(14 - 3\ c_0 - 15\ c_1) = 0$$

que resolviendo da el resultado anterior.

Si los datos no aparecen en relación lineal, se podria usar un polinomio de grado mayor. Es decir, para encontrar los coeficientes c_0, $c_1, \ldots$, c_n de la mejor aproximación por mínimos cuadrados a los datos

x	x_1	x_2	$\ldots$	x_m
y	y_1	y_2	$\ldots$	y_m

$(XIX.18)$

con un polinomio de grado n, tenemos que encontrar la solución de mínimos cuadrados al sistema

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 & \ldots & x_m^n \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} . \qquad (XIX.22)$$

Ejemplo 8.

Queremos encontrar la mejor aproximación de mínimos cuadrados con una función cuadrática a los siguientes datos:

x	0	1	2	3
y	3	2	4	4

Para estos datos el sistema $(XIX.22)$ es $A\ \mathbf{c} = \mathbf{y}$ con

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix} , \qquad \mathbf{c} = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} \quad \text{y} \quad \mathbf{y} = \begin{pmatrix} 3 \\ 2 \\ 4 \\ 4 \end{pmatrix} .$$

Las ecuaciones normales, $A^t A\ \mathbf{c} = A^t\ \mathbf{y}$, se reducen a

$$\begin{pmatrix} 4 & 6 & 14 \\ 6 & 14 & 36 \\ 14 & 36 & 98 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 13 \\ 22 \\ 54 \end{pmatrix} .$$

La solución a este sistema es $(2.75, -0.25, 0.25)$. Entonces el polinomio cuadrático que da la mejor aproximación de mínimos cuadrados es

$$y = p(x) = 2.75 - 0.25\ x + 0.25\ x^2 .$$

EJERCICIOS.

1. Encontrar la solución de mínimos cuadrados a cada uno de los siguientes sistemas $A\,\mathbf{x} = \mathbf{b}$:

$$A = \begin{pmatrix} 1 & 1 \\ 2 & -3 \\ 0 & 0 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$$

$$A = \begin{pmatrix} -1 & 1 \\ 2 & 1 \\ 1 & -2 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 10 \\ 5 \\ 20 \end{pmatrix}$$

$$A = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 4 \\ 0 \\ 1 \\ 2 \end{pmatrix}$$

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ -1 & -2 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

$$A = \begin{pmatrix} 1 & 1 & 3 \\ -1 & 3 & 1 \\ 1 & 2 & 4 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} -2 \\ 0 \\ 8 \end{pmatrix}$$

2. Encontrar la mejor aproximación de mínimos cuadrados con una función lineal a los siguientes datos:

x	-1	0	1	2
y	0	1	3	9

Dibujar la función lineal y los datos en un sistema de coordenadas cartesiano.

3. Encontrar la mejor aproximación de mínimos cuadrados con una función polinomica cuadrática a los siguientes datos:

x	-1	0	1	2
y	0	1	3	9

Dibujar la función cuadrática y los datos en un sistema de coordenadas cartesiano.

CAPITULO XX. LOS METODOS DE TRANSFORMACION ORTOGONAL

1. LAS TRANSFORMACIONES DE HOUSEHOLDER

Debido a las posibles dificultades numéricas del uso de las ecuaciones normales, los modernos métodos de mínimos cuadrados se han desarrollado basándose en las **transformaciones ortogonales**, que preservan las distancias euclídeas y no empeoran las condiciones de la matriz A. La idea es transformar un problema de mínimos cuadrados de manera que sea fácil de resolver, reduciendo la matriz A a la forma que revela el rango. El término forma triangular que revela el rango denota una matriz genérica $m \times n$ de rango r correspondiente a la matriz $\tilde{T} = \begin{pmatrix} T \\ 0 \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ 0 & 0 \end{pmatrix}$, con T_{11} matriz $r \times r$ no singular triangular superior y T_{12} matriz $r \times (n-r)$. Cuando $\tilde{T}$ tiene rango lleno de filas entonces no aparecen ambos bloques de ceros; si $\tilde{T}$ es de rango lleno de columna, la matriz T_{12} y el bloque de ceros derecho no aparecerán. Más en general, el rango de una matriz $m \times n$, $\tilde{F}$, es r si $\tilde{F}$ es de la forma $\tilde{F} = \begin{pmatrix} F \\ 0 \end{pmatrix}$, con F matriz $r \times n$ y las filas de F son linealmente independientes.

Dado que ya no estamos resolviendo ecuaciones, el conjunto de transformaciones que se puede aplicar a A sin cambiar la solución está restringido. Las transformaciones ortogonales son obvias en este contexto dado que estas no alteran la norma l_2.

Sea A una matriz no nula $m \times n$, con rango igual a r. Supóngase que se pueda encontrar una matriz $m \times m$ ortogonal Q que produzca la forma que revela el rango

$$Q^t \; A = \begin{pmatrix} F \\ 0 \end{pmatrix} \; , \tag{XX.1}$$

donde F es una matriz $r \times n$ y tiene las filas linealmente independientes; el bloque de ceros no aparece si $r = m$.

La forma (XX.1) nos permite resolver el problema de mínimos cuadrados usando las matrices Q y F. Sea $\mathbf{d}$ el vector transformado $Q^t \; \mathbf{b}$, descompuesto en

$$\mathbf{d} = Q^t \; \mathbf{b} = \begin{pmatrix} \mathbf{d_r} \\ \mathbf{d_{m-r}} \end{pmatrix} \; . \tag{XX.2}$$

Usando esta definición y la estructura mostrada en (XX.1), se puede escribir el vector residual transformado como

$$Q^t \; (\mathbf{b} - A \; \mathbf{x}) = \begin{pmatrix} \mathbf{d_r} \\ \mathbf{d_{m-r}} \end{pmatrix} - \begin{pmatrix} F \; \mathbf{x} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{d_r} - F \; \mathbf{x} \\ \mathbf{d_{m-r}} \end{pmatrix} \; . \tag{XX.3}$$

Dado que la matriz Q^t es ortogonal, su aplicación al vector residual no altera la norma euclídea, y

$$||\mathbf{b} - A \; \mathbf{x}||_2^2 = ||Q^t \; (\mathbf{b} - A \; \mathbf{x})||_2^2 \; . \tag{XX.4}$$

Combinando esta relación con la forma especial del vector residual transformado, se concluye que

$$||\mathbf{b} - A\ \mathbf{x}||_2^2 = ||Q^t\ (\mathbf{b} - A\ \mathbf{x})||_2^2 = ||\mathbf{d_r} - F\ \mathbf{x}||_2^2 + ||\mathbf{d_{m-r}}||_2^2 \geq ||\mathbf{d_{m-r}}||_2^2\ , \qquad (XX.5)$$

que significa que $||\mathbf{b} - A\ \mathbf{x}||_2$ no puede ser menor que $||\mathbf{d_{m-r}}||_2$ para cualquier vector $\mathbf{x}$. El valor más pequeño posible para $||\mathbf{b} - A\ \mathbf{x}||_2^2$ es la cota inferior. Igualdades con la cota inferior se obtienen si y sólo si el vector $\mathbf{x}$ satisface

$$F\ \mathbf{x} = \mathbf{d_r}\ . \qquad (XX.6)$$

Dado que el rango de la matriz F es igual a r, el sistema $F\ \mathbf{x} = \mathbf{d_r}$ tiene que ser compatible. Cuando $F\ \mathbf{x} = \mathbf{d_r}$, el vector residual transformado satisface $Q^t\ (\mathbf{b} - A\ \mathbf{x}) = \begin{pmatrix} 0 \\ \mathbf{d_{m-r}} \end{pmatrix}$ y $||\mathbf{b} - A\ \mathbf{x}||_2 = ||\mathbf{d_{m-r}}||_2$. Esto demuestra que cualquier vector $\mathbf{x}$ que satisfaga $F\ \mathbf{x} = \mathbf{d_r}$ será una solución de mínimos cuadrados. Suponiendo que los sistemas en los cuales aparecen las matrices F son fáciles de resolver, se sigue que el problema de mínimos cuadrados se puede resolver encontrando una matriz ortogonal Q^t que nos dé la forma $(XX.1)$.

Antes de presentar la **factorización QR** para resolver el problema de mínimos cuadrados, es necesario introducir las **transformaciones de Householder**.

La técnica más popular para construir una matriz ortogonal que reduzca la matriz A en forma triangular usa una clase especial de matrices que son simultaneamente simétricas, elementales y ortogonales. Para cualquier vector no nulo $\mathbf{u}$, la correspondiente **transformación de Householder** (o **matriz de Householder**, o **reflector de Householder**) es una matriz de la forma

$$H = H(u) = I - \frac{2\mathbf{u}\mathbf{u}^t}{||\mathbf{u}||_2^2} = I - \frac{\mathbf{u}\mathbf{u}^t}{\beta}\ , \quad \text{con}\ \ \beta = \frac{||\mathbf{u}||_2^2}{2}\ , \qquad (XX.7)$$

donde el vector $\mathbf{u}$ es el **vector de Householder**

Teorema XX.1

Si H es la matriz definida en $(XX.7)$, entonces

1) $H = H^t$,

2) $H = H^{-1}$,

que es lo mismo que decir que la matriz H es simétrica y ortogonal.

Demostración: Antes de todos, volvemos a escribir la definición de la matriz de Householder usando el vector normalizado de norma uno

$$\mathbf{w} = \frac{\mathbf{u}}{||\mathbf{u}||_2}\ .$$

Entonces, 1) usando las propiedades básicas de las matrices:

$$H^t = I^t - 2\ (\mathbf{w}\ \mathbf{w}^t)^t = I - 2\ (\mathbf{w}^t)^t\ \mathbf{w}^t = I - 2\ \mathbf{w}\ \mathbf{w}^t = H\ .$$

2) usando el hecho que $\mathbf{w}^t\ \mathbf{w} = 1$, se sigue que

$$H^t\ H = H^2 = (I - 2\ \mathbf{w}\ \mathbf{w}^t)\ (I - 2\ \mathbf{w}\ \mathbf{w}^t) = I - 4\ \mathbf{w}\ \mathbf{w}^t + 4\ \mathbf{w}\ (\mathbf{w}^t\ \mathbf{w})\ \mathbf{w}^t = I\ .$$

c.q.d.

Entonces, las matrices de Householder son simétricas y ortogonales, y dependen sólo de la dirección del vector $\mathbf{u}$.

En el contexto de las reducciones a matrices triangulares, las matrices de Householder poseen dos propiedades cruciales:

- para cualquier par de vectores distintos de igual norma l_2, existe una transformación de Householder que transforma el uno en el otro,

$$H\ \mathbf{a} = (I - \frac{\mathbf{u}\mathbf{u}^t}{\beta})\ \mathbf{a} = \mathbf{b}\ , \qquad (XX.8)$$

con $||\mathbf{a}||_2 = ||\mathbf{b}||_2$. Esto implica que el vector $\mathbf{u}$ tiene que satisfacer la condición

$$-\frac{\mathbf{u}^t\ \mathbf{a}}{\beta}\ \mathbf{u} = \mathbf{b} - \mathbf{a}\ , \qquad (XX.9)$$

es decir, $\mathbf{u}$ es un múltiplo de $\mathbf{b} - \mathbf{a}$;

- cualquier vector $\mathbf{c}$ transformado por una matriz de Householder posee una forma especial:

$$H\ \mathbf{c} = (I - \frac{\mathbf{u}\mathbf{u}^t}{\beta})\ \mathbf{c} = \mathbf{c} - \frac{\mathbf{u}^t\mathbf{c}}{\beta}\ \mathbf{u}\ , \qquad (XX.10)$$

de manera que $H\ \mathbf{c}$ es la diferencia entre el vector original $\mathbf{c}$ y un múltiplo especial del vector de Householder $\mathbf{u}$.

Claramente el vector $\mathbf{c}$ no varía si $\mathbf{u}^t\ \mathbf{c} = 0$. Además, calcular el producto $H\ \mathbf{c}$ no necesita los elementos explícitos de H, sino sólo el vector $\mathbf{u}$ y el escalar β.

Ejemplo 1.

Sea $\mathbf{u} = (-1, -2)^t$, de manera que $||\mathbf{u}||_2 = \sqrt{5}$. La transformación de Householder asociada es la matriz dada por

$$H = H(u) = I - \frac{2\mathbf{u}\mathbf{u}^t}{||\mathbf{u}||_2^2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{2}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{8}{5} \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & -\frac{4}{5} \\ -\frac{4}{5} & -\frac{3}{5} \end{pmatrix}\ .$$

Ahora, consideremos los dos vectores tales que $||\mathbf{a}||_2 = ||\mathbf{b}||_2$

$$\mathbf{a} = \begin{pmatrix} \frac{9}{2} \\ -1 \end{pmatrix}\ , \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} \frac{7}{2} \\ -3 \end{pmatrix}\ , \quad \text{para los cuales} \quad \mathbf{b} - \mathbf{a} = \begin{pmatrix} -1 \\ -2 \end{pmatrix}\ .$$

Usando la expresión $(XX.9)$, podemos construir una transformación de Householder H tal que $H\ \mathbf{a} = (I - \frac{\mathbf{u}\mathbf{u}^t}{\beta})\ \mathbf{a} = \mathbf{b}$ escogiendo el vector $\mathbf{u}$ como un múltiplo de $\mathbf{b} - \mathbf{a}$. Dado

que, en nuestro caso $\mathbf{b}-\mathbf{a}=\mathbf{u}=(-1,-2)^t$, es decir, el vector $\mathbf{b}-\mathbf{a}$ coincide con el vector $\mathbf{u}$ usado para escribir una transformación de Householder, sigue que

$$H\mathbf{a}=\begin{pmatrix} \frac{3}{5} & -\frac{4}{5} \\ -\frac{4}{5} & -\frac{3}{5} \end{pmatrix} \begin{pmatrix} \frac{9}{2} \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{7}{2} \\ -3 \end{pmatrix} = \mathbf{b} \ .$$

2. LA FACTORIZACION QR CON MATRICES DE HOUSEHOLDER

Para una matriz genérica A, $n \times n$, no singular, las propiedades que se acaban de describir permiten construir una sucesión de $n-1$ matrices de Householder tales que

$$H_{n-1} \ \ldots \ H_2 \ H_1 \ A = R \ , \qquad (XX.11)$$

donde R es una matriz $n \times n$ no singular y triangular superior.

El primer paso de este proceso es construir una matriz de Householder H_1 que transforma $\mathbf{a_1}$ (la primera columna de A) en un múltiplo de $\mathbf{e_1}$, es decir, se desean crear ceros en las componentes 2 hasta n del vector $\mathbf{a_1}$. La norma euclídea se conserva bajo transformaciones ortogonales, de manera que

$$H_1 \ \mathbf{a_1} = (I - \frac{\mathbf{u_1}\mathbf{u_1}^t}{\beta_1}) \ \mathbf{a_1} = \pm||\mathbf{a_1}||_2 \ \mathbf{e_1} = \begin{pmatrix} r_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \ , \qquad (XX.12)$$

donde $|r_{11}| = ||\mathbf{a_1}||_2$. De la expresión $(XX.9)$ sabemos que el vector $\mathbf{u_1}$ tiene que ser un múltiplo de $||\mathbf{a_1}||_2 \ \mathbf{e_1} - \mathbf{a_1}$, y dado que H_1 depende sólo de la dirección de $\mathbf{u_1}$, podemos elegir el vector $\mathbf{u_1}$ como

$$\mathbf{u_1} = \begin{pmatrix} a_{11} - r_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} \ . \qquad (XX.13)$$

Al definir $\mathbf{u_1}$, el signo de r_{11} se puede eligir positivo o negativo (excepto cuando $\mathbf{a_1}$ es ya un múltiplo de $\mathbf{e_1}$), y para evitar el problema de la cancelación de términos parecidos, usualmente se escoge el signo opuesto al signo de a_{11}, de manera que

$$\text{signo}(r_{11}) = -\text{signo}(a_{11}) \ . \qquad (XX.14)$$

Después de la primera aplicación de las transformaciones de Householder, la primera columna de la matriz parcialmente reducida $A^{(2)} = H_1 \ A$ es un múltiplo de $\mathbf{e_1}$, y los demás elementos han sido alterados

$$A^{(2)} = H_1 \ A = \begin{pmatrix} r_{11} & a_{12}^{(2)} & \ldots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \ldots & a_{nn}^{(2)} \end{pmatrix} \ . \qquad (XX.15)$$

En muy importante notar que a diferencia de la eliminación Gaussiana, la primera fila de la matriz A viene modificada por efecto de la transformación de Householder H.

Al construir la segunda transformación de Householder, el objetivo principal es reducir la segunda columna a la forma correcta, sin alterar la primera fila y la primera columna de la matriz parcialmente reducida. Debido a la propiedad $(XX.10)$, se puede obtener este resultado definiendo el segundo vector de Householder $\mathbf{u_2}$ con la primera componente nula. Habiendo escogido así el vector $\mathbf{u_2}$, la aplicación de la matriz de Householde H_2 a un vector genérico no cambia la primera componente, y la aplicación a un múltiplo de $\mathbf{e_1}$ deja el vector entero como está.

Si A es una matriz no singular, se pueden efectuar $n-1$ pasos de reducción de Householder, para obtener $H_{n-1}\ \ldots\ H_2\ H_1\ A = R$, con R una matriz triangular superior no singular. Si se denota con Q^t la matriz ortogonal $n \times n$

$$Q^t = H_{n-1}\ \ldots\ H_2\ H_1 \qquad \Longrightarrow \qquad Q = H_1\ H_2\ \ldots\ H_{n-1}\ . \tag{$XX.16$}$$

Cualquiera de las dos formas

$$Q^t\ A = R \qquad \text{o} \qquad A = Q\ R \tag{$XX.17$}$$

se conoce como la factorización QR de la matriz A.

Una vez conocida la factorización QR de A, ecuación $(XX.17)$, la solución al sistema $A\ \mathbf{x} = \mathbf{b}$ se obtiene resolviendo el sistema triangular superior $R\ \mathbf{x} = Q^t\ \mathbf{b}$.

Ejemplo 2.

Sean

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 3 & 1 \\ 3 & -1 & -1 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 3 \\ 2 \\ 6 \end{pmatrix}\ .$$

Aplicando la $(XX.13)$, obtenemos

$$||\mathbf{a_1}||_2 = \sqrt{14}\ , \quad r_{11} = -\sqrt{14}\ , \quad \text{y} \quad \mathbf{u_1} = \begin{pmatrix} 1+\sqrt{14} \\ 2 \\ 3 \end{pmatrix}\ .$$

Entonces al primer paso, en aritmética de cuatro dígitos,

$$H_1 = \begin{pmatrix} -0.2673 & -0.5345 & -0.8018 \\ -0.5345 & 0.7745 & -0.3382 \\ -0.8018 & -0.3382 & 0.4927 \end{pmatrix} \quad \text{y} \quad A^{(2)} = \begin{pmatrix} -0.2673 & -1.069 & -0.2673 \\ & 2.127 & 0.04368 \\ & -2.309 & -2.434 \end{pmatrix}\ .$$

Efectuando un segundo paso, se obtiene

$$R = \begin{pmatrix} -3.742 & -1.069 & -0.2673 \\ & -3.140 & -1.820 \\ & & -1.617 \end{pmatrix} \quad \text{y} \quad Q^t\ \mathbf{b} = \begin{pmatrix} -6.682 \\ 1.320 \\ -1.617 \end{pmatrix}\ .$$

Con sustitución regresiva se obtiene la solución $\mathbf{x} = (2, -1, 1)^t$.

En general, es necesario un intercambio de columnas para asegurar que el rango está plenamente revelado. Sin el intercambio de columnas la reducción de Householder terminaría inmediatamente con una matriz no nula cuya primera columna es cero. Si las demás columnas son también ceros, la reducción termina. De otra manera, existe por lo menos una columna no nula, la columna pivote, que se puede eligir como candidata para la siguiente reducción. Como en la eliminación gaussiana, una estrategia de pivoteo pide que se escoja la columna pivote como la primera columna de norma máxima (otra posibilidad es escoger la columna "menos reducida").

En general, si A es $m \times n$ de rango r, se necesitarán r permutaciones $\{P_k\}$ y r matrices de Householder $\{H_k\}$, $k = 1, \ldots, r$. Después de estos r pasos, la configuración final será

$$H_r \ \ldots \ H_1 \ A \ P_1 \ \ldots \ P_r = \tilde{R} \ , \tag{XX.18}$$

donde

$$\tilde{R} = \begin{pmatrix} R \\ 0 \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix} \tag{XX.19}$$

es triangular superior que nos revela el rango, y R_{11} es una matriz $r \times r$ no singular triangular superior. Con una correcta estrategia de pivoteo, dependiente del problema original, y aritmética exacta, este procedimiento de Householder terminará después de r pasos, cuando la matriz restante se transformará en cero. Combinando los intercambios de columnas en una sola matriz de permutación P y las transformaciones de Householder en una sola matriz ortogonal Q^t, se obtiene

$$Q^t \ A \ P = \tilde{R} \ , \tag{XX.20}$$

o equivalentemente,

$$A \ P = Q \ \tilde{R} \ , \tag{XX.21}$$

y

$$Q^t \ A = \tilde{R} \ P^t = \begin{pmatrix} R \\ 0 \end{pmatrix} \ P^t = \begin{pmatrix} R \ P^t \\ 0 \end{pmatrix} \ . \tag{XX.22}$$

Las filas de la matriz $R \ P^t$ son linealmente independientes, de manera que la matriz transformada $Q^t \ A$ tiene la forma deseada (XX.1), con $F = R \ P^t$. El problema de mínimos cuadrados se puede entonces resolver con el siguiente algoritmo:
(i) Calcular la factorización QR de la matriz A, usando la reducción de Householder (que determina el rango r, y las matrices P, Q y R);
(ii) Formar el vector $\mathbf{d} = Q^t \ \mathbf{b}$, y denotar con $\mathbf{d_r}$ las primeras r componentes de $\mathbf{d}$;
(iii) Calcular cualquier solución $\mathbf{y}$ del sistema $R \ \mathbf{y} = \mathbf{d_r}$;
(iv) Formar $\mathbf{x} = P \ \mathbf{y}$.

El número de operaciones requerido para resolver el problema de mínimos cuadrados de esta manera es del orden de $2 \cdot m \cdot n \cdot r - r^2 \ (m + n) + \frac{2}{3} n^3$.

Si la matriz A posee columnas linealmente independientes ($r = n$), la matriz R es no singular, la solución de $R \ \mathbf{y} = \mathbf{d_r}$ es única, y la solución del problema de mínimos cuadrados es única. En este caso resolver el problema con el método QR es aproximadamente

dos veces más costoso que resolviendolo con las ecuaciones normales. Cuando las columnas de A son linealmente dependientes, de manera que $r < n$, el sistema $r \times n$ $R\ \mathbf{y} = \mathbf{d_r}$ posee un número infinito de soluciones. Dado que R es de la forma $R = (R_{11}R_{12})$, con R_{11} triangular superior no singular, es posible escoger un vector $\mathbf{y} = (\mathbf{y_B}, 0)^t$ tal que $R\ \mathbf{y} = \mathbf{d_r}$ y con la propiedad de que $R_{11}\ \mathbf{y_B} = \mathbf{d_r}$. Si $\mathbf{y}$ posee esta forma, la solución $\mathbf{x} = P\ \mathbf{y}$ es sencillamente una reordenación de $\mathbf{y}$, y es llamada una **solución básica** del problema de mínimos cuadrados.

Ejemplo 3.

Sean

$$A = \begin{pmatrix} 1 & 2 & 2 \\ 7 & 6 & 10 \\ 4 & 4 & 6 \\ 1 & 0 & 1 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 6 \\ 6 \\ 8 \\ 3 \end{pmatrix} .$$

La representación única de $\mathbf{b}$ en términos de los espacios rango y nulo de A es

$$\mathbf{b} = \mathbf{b_R} + \mathbf{b_N} , \quad \text{con} \quad \mathbf{b_R} = \begin{pmatrix} 4 \\ 8 \\ 6 \\ -1 \end{pmatrix} \quad \text{y} \quad \mathbf{b_N} = \begin{pmatrix} 2 \\ -2 \\ 2 \\ 4 \end{pmatrix} ,$$

con $||\mathbf{b_N}||_2 = 5.292$. Aplicando la reducción de Householder con intercambio de columnas en A, obtenemos

$$R = \begin{pmatrix} -11.87 & -7.411 & -8.169 \\ & 1.038 & -0.5191 \end{pmatrix} \quad \text{y} \quad P = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} .$$

El vector transformado $\mathbf{d} = Q^t\ \mathbf{b}$ es

$$\mathbf{d} = Q^t\ \mathbf{b} = \begin{pmatrix} d_r \\ d_{m-r} \end{pmatrix} = \begin{pmatrix} -10.36 \\ 3.115 \\ 3.419 \\ 4.038 \end{pmatrix} ,$$

con $||\mathbf{d_{m-r}}||_2 = ||\mathbf{b_N}||_2 = 5.292$. Ahora, el vector $\mathbf{y_B}$ que satisface $R_{11}\ \mathbf{y_B} = \mathbf{d_r}$ es

$$\mathbf{y_B} = \begin{pmatrix} -1 \\ 3 \end{pmatrix} ,$$

de manera que la solución básica del problema de mínimos cuadrados $\mathbf{x}$ es

$$\mathbf{x} = P \begin{pmatrix} y_B \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ -1 \end{pmatrix} .$$

Finalmente, el vector residual optimal es

$$\mathbf{r} = \mathbf{b} - A\ \mathbf{x} = \begin{pmatrix} 2 \\ -2 \\ 2 \\ 4 \end{pmatrix} .$$

3. LA FACTORIZACION QR CON MATRICES DE ROTACION DE GIVENS

Vamos a presentar ahora un método más efectivo que los estudiados anteriormente para resolver el problema de mínimos cuadrados. Como el método de Householder, el **método de las rotaciones de Givens** consiste en factorizar la matriz A como producto de una matriz ortogonal Q y una matriz triangular R, $A = Q\,R$. La razón de estos esquemas es que las matrices ortogonales cumplen unas propiedades que permiten ser manipuladas computacionalmente sin que los errores de redondeo crezcan sin control.

Consideramos antes el caso del plano R^2, y sea $\mathbf{v} = (v_1, v_2) \in R^2$. Si deseamos rotar el vector $\mathbf{v}$ de un ángulo θ, multiplicaremos el vector por una matriz Q del siguiente tipo:

$$Q = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \qquad Q\,\mathbf{v} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1\,\cos\theta - v_2\,\sin\theta \\ v_1\,\sin\theta + v_2\,\cos\theta \end{pmatrix}.$$

La multiplicación de un vector por la transpuesta de la matriz Q produce el efecto contrario, esto es rotar el vector en sentido opuesto al ángulo θ.

Como ya dicho en el método de Householder, el objetivo de los métodos de transformación ortogonales para resolver sistemas lineales es el de conseguir una factorización del tipo $A = Q\,R$, esto es $Q^t\,A = R$, donde R es triangular superior. Para ello se necesita que la multiplicación de la matriz Q^t por A rote o convierta el primer vector de esta última $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$ en un vector de la forma $Q^t\,\mathbf{v} = (w_1, 0, \ldots, 0)^t$.

En el caso de matrices de dimensión 2×2 se tiene

$$Q^t\,\mathbf{v} = \begin{pmatrix} v_1\,\cos\theta + v_2\,\sin\theta \\ -v_1\,\sin\theta + v_2\,\cos\theta \end{pmatrix} \quad \Rightarrow \quad -v_1\,\sin\theta + v_2\,\cos\theta = 0$$

y para que se cumpla esta última igualdad es suficiente que sea

$$\sin\theta = \frac{v_2}{\sqrt{v_1^2 + v_2^2}}, \qquad \cos\theta = \frac{v_1}{\sqrt{v_1^2 + v_2^2}}.$$

En el caso general de matrices de dimensión $n \times n$, a las matrices $Q_{j\,i}$ que, multiplicadas por una matriz A, convierten el elemento a_{ji} de esta última en 0, se les denomina **rotaciones de Givens**. Su estructura es la siguiente

$$Q_{ji} = \begin{cases} q_{kk} = 1 & \text{si} \quad k \neq i, j \\ q_{ii} = q_{jj} = c & \\ q_{ji} = -q_{ij} = s & \\ 0 & \text{en las otras posiciones} \end{cases}$$

donde los coeficientes c y s se definen como

$$c = \frac{a_{ii}}{\sqrt{a_{ii}^2 + a_{ji}^2}} \qquad \text{y} \qquad s = \frac{a_{ji}}{\sqrt{a_{ii}^2 + a_{ji}^2}}.$$

Al multiplicar la matriz Q_{ji}^t por la matriz A, se modifican sólo las componentes de las filas $i-$ésima y $j-$ésima, dado que sobre el resto de filas se aplica el mismo efecto que

multiplicar por la matriz identidad. Se puede entonces pensar en un proceso de multiplicaciones ordenadas y sistematicas de matrices ortogonales Q_{ji}^t por la matriz A hasta reducir la matriz original A a una matriz triangular superior R.

La primera matriz Q_{21}^t que se multiplica por A, transforma el primer vector columna de A, $\mathbf{a}_1 = (a_{11}, a_{21}, a_{31}, \ldots, a_{n1})^t$, en $Q_{21}^t\,\mathbf{a}_1 = (a_{11}^{(1)}, 0, a_{31}^{(1)}, \ldots, a_{n1}^{(1)})^t$. Sucesivamente $Q_{31}^t\,Q_{21}^t\,\mathbf{a}_1 = (a_{11}^{(2)}, 0, 0, a_{41}^{(2)}, \ldots, a_{n1}^{(2)})^t$ y después de repetir $n-1$ veces el mismo proceso tenemos $Q_{n1}^t \ldots Q_{21}^t\,\mathbf{a}_1 = (a_{11}^{(n-1)}, 0, \ldots, 0)^t$. Como solamente se han utilizado rotaciones, la norma de este último vector es la misma que la del vector original $\mathbf{a}_1$.

Una vez acabado el proceso con la primera columna de A, se repite el proceso con la segunda, luego con todas las siguientes hasta acabar con la $(n-1)$-ésima. La multiplicación sucesiva de las $\frac{(n-1)\,n}{2}$ matrices Q_{ji}^t constituye una matriz ortogonal Q^t que completa la factorización de A en una matriz R triangular superior

$$Q_{n\,n-1}^t \ldots Q_{n1}^t \ldots Q_{21}^t = Q^t \quad \Rightarrow \quad Q^t\,M = R.$$

El coste operativo de la factorización $M = Q\,R$ es de $2\,\frac{n^3}{3}$ operaciones, esto es aproximadamente el doble de una facorización $L\,U$. La ventaja es que las descomposiciones $Q\,R$ no amplifican el error de redondeo surgidos en las operaciones.

Una vez obtenida la factorización $M = Q\,R$ se puede resolver el sistema lineal original $M\,\mathbf{x} = \mathbf{b}$, multiplicando Q^t por $\mathbf{b}$ y resoviendo directamente, mediante sustitución regresiva, el sistema triangular $R\,\mathbf{x} = Q^t\,\mathbf{b}$.

Ejemplo 4.

Sean A la matriz y $\mathbf{b}$ el vector del problema $A\,\mathbf{x} = \mathbf{b}$

$$A = \begin{pmatrix} 3 & 15 & -1 \\ -4 & 1 & 2 \\ 9 & 9 & 5 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 17 \\ -1 \\ 23 \end{pmatrix}.$$

Empezamos definiendo la primera matriz ortogonal Q_{21} definida por

$$Q_{21} = \begin{pmatrix} c & -s & 0 \\ s & c & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

con

$$c = \frac{3}{\sqrt{3^2 + (-4)^2}} = \frac{3}{5} = 0.6 \quad \text{y} \quad s = -\frac{4}{\sqrt{3^2 + (-4)^2}} = -\frac{4}{5} = -0.8.$$

Esto es

$$Q_{21} = \begin{pmatrix} 0.6 & 0.8 & 0 \\ -0.8 & 0.6 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

y por lo tanto

$$Q_{21}^t\,A = \begin{pmatrix} 0.6 & -0.8 & 0 \\ 0.8 & 0.6 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 15 & -1 \\ -4 & 1 & 2 \\ 9 & 9 & 5 \end{pmatrix} = \begin{pmatrix} 5 & 8.2 & -2.2 \\ 0 & 12.6 & 0.4 \\ 9 & 9 & 5 \end{pmatrix}.$$

Ahora pasamos a definir la segunda matriz ortogonal

$$Q_{31} = \begin{pmatrix} c & 0 & -s \\ 0 & 1 & 0 \\ s & 0 & c \end{pmatrix},$$

con

$$c = \frac{5}{\sqrt{5^2+9^2}} = 0.48564 \quad \text{y} \quad s = \frac{9}{\sqrt{5^2+9^2}} = 0.87416.$$

Esto es

$$Q_{31} = \begin{pmatrix} 0.48564 & 0 & -0.87416 \\ 0 & 1 & 0 \\ 0.87416 & 0 & 0.48564 \end{pmatrix},$$

y por lo tanto

$$Q_{31}^t\, Q_{21}^t\, A = \begin{pmatrix} 0.48564 & 0 & 0.87416 \\ 0 & 1 & 0 \\ -0.87416 & 0 & 0.48564 \end{pmatrix} \begin{pmatrix} 5 & 8.2 & -2.2 \\ 0 & 12.6 & 0.4 \\ 9 & 9 & 5 \end{pmatrix} =$$

$$= \begin{pmatrix} 10.2956 & 11.8497 & 3.30237 \\ 0 & 12.6 & 0.4 \\ 0 & -2.7973 & 4.35136 \end{pmatrix}.$$

Finalmente la tercera matriz ortogonal es

$$Q_{32} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c & -s \\ 0 & s & c \end{pmatrix},$$

con

$$c = \frac{12.6}{\sqrt{12.6^2+(-2.7973)^2}} = 0.97621 \quad \text{y} \quad s = \frac{-2.7973}{\sqrt{12.6^2+(-2.7973)^2}} = -0.21674.$$

Esto es

$$Q_{32} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.97621 & 0.21674 \\ 0 & -0.21674 & 0.97621 \end{pmatrix},$$

y por lo tanto

$$R = Q_{32}^t\, Q_{31}^t\, Q_{21}^t\, A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.97621 & -0.21674 \\ 0 & 0.21674 & 0.97621 \end{pmatrix} \begin{pmatrix} 10.2956 & 11.8497 & 3.30237 \\ 0 & 12.6 & 0.4 \\ 0 & -2.7973 & 4.35136 \end{pmatrix} =$$

$$= \begin{pmatrix} 10.2956 & 11.8497 & 3.30237 \\ 0 & 12.9068 & -0.552584 \\ 0 & 0 & 4.33463 \end{pmatrix}.$$

Además,

$$Q^t = Q_{32}^t\, Q_{31}^t\, Q_{21}^t = \begin{pmatrix} 0.291386 & -0.388514 & 0.874157 \\ 0.894659 & 0.434173 & -0.105254 \\ -0.338643 & 0.812743 & 0.4741 \end{pmatrix},$$

o que es lo mismo

$$Q = \begin{pmatrix} 0.291386 & 0.894659 & -0.338643 \\ -0.388514 & 0.434173 & 0.812743 \\ 0.874157 & -0.105254 & 0.4741 \end{pmatrix},$$

y resulta que

$$A = Q\,R.$$

Como dicho antes, una vez obtenida la factorización $A = Q\,R$ se puede resolver el sistema lineal original $A\,\mathbf{x} = \mathbf{b}$, multiplicando Q^t por $\mathbf{b}$ y resoviendo directamente, mediante sustitución regresiva, el sistema triangular $R\,\mathbf{x} = Q^t\,\mathbf{b}$. En este ejemplo,

$$Q^t\,\mathbf{b} = \begin{pmatrix} 0.291386 & -0.388514 & 0.874157 \\ 0.894659 & 0.434173 & -0.105254 \\ -0.338643 & 0.812743 & 0.4741 \end{pmatrix} \begin{pmatrix} 17 \\ -1 \\ 23 \end{pmatrix} = \begin{pmatrix} 25.4477 \\ 12.3542 \\ 4.33463 \end{pmatrix},$$

y por lo tanto el sistema a resolver es

$$\begin{pmatrix} 10.2956 & 11.8497 & 3.30237 \\ 0 & 12.9068 & -0.552584 \\ 0 & 0 & 4.33463 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 25.4477 \\ 12.3542 \\ 4.33463 \end{pmatrix}$$

que tiene solución $(1, 1, 1)^t$.

Ejemplo 5.

Sean

$$A = \begin{pmatrix} 1 & 2 & 2 \\ 7 & 6 & 10 \\ 4 & 4 & 6 \\ 1 & 0 & 1 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 6 \\ 6 \\ 8 \\ 3 \end{pmatrix}$$

la matriz A y el vector $\mathbf{b}$ considerados en el ejemplo 3 donde se usaron las transformaciones de Householder. Usaremos ahora el método de las rotaciones de Givens para hallar una solución de mínimos cuadratos del problema.

Empezamos definiendo la primera matriz ortogonal Q_{21}, con $c = 1/\sqrt{1^2 + 7^2} = 0.141421$ y $s = 7/\sqrt{1^2 + 7^2} = 0.989949$, definida por

$$Q_{21} = \begin{pmatrix} c & -s & 0 & 0 \\ s & c & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.141421 & -0.989949 & 0 & 0 \\ 0.989949 & 0.141421 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

y por lo tanto

$$Q_{21}^t\,A = \begin{pmatrix} 0.141421 & 0.989949 & 0 & 0 \\ -0.989949 & 0.141421 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 2 \\ 7 & 6 & 10 \\ 4 & 4 & 6 \\ 1 & 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 7.07107 & 6.22254 & 10.1823 \\ 0 & -1.13137 & -0.565685 \\ 4 & 4 & 6 \\ 1 & 0 & 1 \end{pmatrix}.$$

Ahora pasamos a definir la segunda matriz ortogonal, siendo $c = 0.870388$ y $s = 0.492366$

$$Q_{31} = \begin{pmatrix} c & 0 & -s & 0 \\ 0 & 1 & 0 & 0 \\ s & 0 & c & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.870388 & 0 & -0.492366 & 0 \\ 0 & 1 & 0 & 0 \\ 0.492366 & 0 & 0.870388 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

y por lo tanto

$$Q_{31}^t\, Q_{21}^t\, A = \begin{pmatrix} 0.870388 & 0 & 0.492366 & 0 \\ 0 & 1 & 0 & 0 \\ -0.492366 & 0 & 0.870388 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 7.07107 & 6.22254 & 10.1823 \\ 0 & -1.13137 & -0.565685 \\ 4 & 4 & 6 \\ 1 & 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 8.12404 & 7.38549 & 11.8168 \\ 0 & -1.13137 & -0.565685 \\ 0 & 0.417786 & 0.208893 \\ 1 & 0 & 1 \end{pmatrix}.$$

Siguiendo en la misma línea se calculan las demás matrices

$$Q_{41} = \begin{pmatrix} c & 0 & 0 & -s \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ s & 0 & 0 & c \end{pmatrix} = \begin{pmatrix} 0.992509 & 0 & 0 & -0.122169 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.122169 & 0 & 0 & 0.992509 \end{pmatrix},$$

$$Q_{41}^t\, Q_{31}^t\, Q_{21}^t\, A = \begin{pmatrix} 8.18535 & 7.33017 & 11.8504 \\ 0 & -1.13137 & -0.565685 \\ 0 & 0.417786 & 0.208893 \\ 0 & -0.902281 & -0.451141 \end{pmatrix};$$

$$Q_{32} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & c & -s & 0 \\ 0 & s & c & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.938083 & -0.34641 & 0 \\ 0 & 0.34641 & -0.938083 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$Q_{32}^t\, Q_{41}^t\, Q_{31}^t\, Q_{21}^t\, A = \begin{pmatrix} 8.18535 & 7.33017 & 11.8504 \\ 0 & 1.20605 & 0.603023 \\ 0 & 0 & 0 \\ 0 & -0.902281 & -0.451141 \end{pmatrix};$$

$$Q_{42} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & c & 0 & -s \\ 0 & 0 & 1 & 0 \\ 0 & s & 0 & c \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.800717 & 0 & 0.599042 \\ 0 & 0 & 1 & 0 \\ 0 & -0.599042 & 0 & 0.800717 \end{pmatrix},$$

y finalmente

$$R = Q_{42}^t\, Q_{32}^t\, Q_{41}^t\, Q_{31}^t\, Q_{21}^t\, A = \begin{pmatrix} 8.18535 & 7.33017 & 11.8504 \\ 0 & 1.50621 & 0.753103 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Por otro lado

$$Q^t = Q_{42}^t\, Q_{32}^t\, Q_{41}^t\, Q_{31}^t\, Q_{21}^t = \begin{pmatrix} 0.122169 & 0.855186 & 0.488678 & 0.122169 \\ 0.733285 & -0.178367 & 0.277459 & -0.594555 \\ 0.408248 & 0.408248 & -0.816497 & 0 \\ 0.529813 & -0.264906 & 0.132453 & 0.794719 \end{pmatrix}$$

y

$$Q^t\, \mathbf{b} = \begin{pmatrix} 10.1401 \\ 3.76552 \\ -1.63299 \\ 5.03322 \end{pmatrix}.$$

Finalmente resolviendo las dos primeras ecuaciones del problemas $R\,\mathbf{x} = Q^t\,\mathbf{b}$ en las incognitas (x_1, x_2, x_3) se obtiene la solución de mínimos cuadratos

$$\mathbf{x} = (-1 - x_3, 2.5 - 0.5\, x_3, x_3)$$

cuyo vector residual optimal es, como ya visto en el ejemplo 3,

$$\mathbf{r} = \mathbf{b} - A\ \mathbf{x} = \begin{pmatrix} 2 \\ -2 \\ 2 \\ 4 \end{pmatrix}\ .$$

EJERCICIOS.

1. Encontrar la solución $\mathbf{x}$ al problema de mínimos cuadrados, dado en la forma $A = Q\ R$ en cadauno de los casos siguientes:

$$Q = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 3 \\ 1 \\ 2 \end{pmatrix}.$$

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 1 \\ \sqrt{2} \\ -\sqrt{2} \end{pmatrix}.$$

$$Q = \begin{pmatrix} 1/2 & 1/\sqrt{2} & 0 & 1/2 \\ 1/2 & 0 & 1/\sqrt{2} & -1/2 \\ 1/2 & 0 & -1/\sqrt{2} & -1/2 \\ 1/2 & -1/\sqrt{2} & 0 & 1/2 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 2 \\ -2 \\ 0 \\ 2 \end{pmatrix}.$$

CURSO DE MÉTODOS NUMÉRICOS

QUINTA PARTE

SOLUCIONES NUMÉRICAS DE SISTEMAS NO LINEALES

CAPITULO XXI. PUNTOS FIJOS PARA FUNCIONES DE VARIAS VARIABLES

1. PRELIMINARES

La forma general de un sistema de ecuaciones no lineales es:

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \ , \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \ , \\ \ldots \quad \ldots \quad \ldots & \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \ , \end{aligned} \tag{$XXI.1$}$$

donde cada función f_i puede tomarse como una aplicación de un vector $\mathbf{x}$ del espacio $n-$dimensional $\mathcal{R}^n$, $\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_n)^t$, en la recta real $\mathcal{R}$. Alternativamente, el sistema puede representarse definiendo una función $\mathbf{F}$, de $\mathcal{R}^n$ en $\mathcal{R}^n$ por

$$\mathbf{F}(x_1, x_2, \ldots, x_n) = (f_1(x_1, \ldots, x_n), f_2(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n))^t = \mathbf{0} \ .$$

Usando notación vectorial para representar las variables x_i, el sistema $(XXI.1)$ asume la forma:

$$\mathbf{F}(\mathbf{x}) = \mathbf{0} \ . \tag{$XXI.2$}$$

Las funciones $f_1, f_2, \ldots, f_n$ se llaman **funciones coordenadas de F**.

Antes de discutir la solución de un sistema dado en las formas $(XXI.1)$ ó $(XXI.2)$, necesitamos considerar algunos resultados concernientes a la continuidad y a la diferenciabilidad de funciones de $\mathcal{R}^n$ en $\mathcal{R}^n$.

Definición. Sea f una función definida en un conjunto $D \subset \mathcal{R}^n$ y con valores en $\mathcal{R}$. Se dice que la función f tiene **límite** L en $\mathbf{x}_0$, denotado por $\lim\limits_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = L$, si, dado cualquier $\varepsilon > 0$, existe un número $\delta > 0$ con la propiedad de que $|f(\mathbf{x}) - L| < \varepsilon$ siempre que $x \in D$ y $0 < ||\mathbf{x} - \mathbf{x}_0|| < \delta$.

Debe hacerse notar que la existencia de un límite es independiente de la norma vectorial particular usada debido a la equivalencia de las normas vectoriales en $\mathcal{R}^n$.

Definición. Sea f una función del conjunto $D \subset \mathcal{R}^n$ en $\mathcal{R}$. Se dice que la función f es **continua** en $\mathbf{x}_0 \in D$ siempre y cuando el $\lim\limits_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})$ exista y $\lim\limits_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0)$. Se dice, además que f es **continua** en un conjunto D si f es continua en cada punto de D. Este concepto se expresa escribiendo $f \in \mathcal{C}(D)$.

Definición. Sea $\mathbf{F}$ una función del conjunto $D \subset \mathcal{R}^n$ en $\mathcal{R}^n$ y supongamos que $\mathbf{F}$ tiene la representación $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x}))^t$, donde f_i para cada i es una aplicación de $\mathcal{R}^n$ en $\mathcal{R}$. Definimos $\lim\limits_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{L} = (L_1, L_2, \ldots, L_n)^t$ si y sólo si $\lim\limits_{\mathbf{x}\to\mathbf{x}_0} f_i(\mathbf{x}) = L_i$ para cada $i = 1, 2, \ldots, n$.

Definición. Sea $\mathbf{F}$ una función del conjunto $D \subset \mathcal{R}^n$ en $\mathcal{R}^n$ con la representación $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x}))^t$. Se dice que la función $\mathbf{F}$ es **continua** en $\mathbf{x}_0 \in D$ si

$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x})$ existe y $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0)$. Se dice que $\mathbf{F}$ es **continua** en el conjunto D si $\mathbf{F}$ es continua en cada $\mathbf{x}$ de D. Este concepto se expresa escribiendo $\mathbf{F} \in \mathcal{C}(D)$.

Enunciaremos un Teorema que relaciona la continuidad de una función de n variables en un punto con las derivadas parciales de la función en ese punto.

Teorema XXI.1

Sea f una función del conjunto $D \subset \mathcal{R}^n$ en $\mathcal{R}^n$ y $\mathbf{x}_0 \in D$. Si existen constantes $\delta > 0$ y $K > 0$ con

$$\Big|\frac{\partial f(\mathbf{x})}{\partial x_j}\Big| \leq K \qquad \text{para cada } \; j = 1, 2, \ldots, n \; ,$$

siempre que $0 < ||\mathbf{x} - \mathbf{x}_0|| < \delta$ y $\mathbf{x} \in D$, entonces f es continua en $\mathbf{x}_0$.

2. METODO DE ITERACION Y EJEMPLOS

En el capítulo XVI, se desarrolló un proceso iterativo para resolver la ecuación $f(x) = 0$ transformando primero esta ecuación en una ecuación de la forma $x = g(x)$. La función g tiene sus puntos fijos precisamente en las soluciones de la ecuación original. Aquí se investigará un procedimiento similar para funciones de $\mathcal{R}^n$ en $\mathcal{R}^n$.

Sea dado un sistema de ecuaciones no lineales de un tipo especial:

$$\begin{aligned} x_1 &= g_1(x_1, x_2, \ldots, x_n) \; , \\ x_2 &= g_2(x_1, x_2, \ldots, x_n) \; , \\ \ldots \quad &\quad \ldots \quad \ldots \quad \ldots \\ x_n &= g_n(x_1, x_2, \ldots, x_n) \; , \end{aligned} \tag{$XXI.3$}$$

donde las funciones $g_1,\ g_2, \ldots, g_n$ son reales, definidas y continuas en un conjunto D, vecindad de una solución separada $(x_1^*, x_2^*, \ldots, x_n^*)$ del sistema $(XXI.3)$. De forma más compacta el sistema $(XXI.3)$ se puede escribir como:

$$\mathbf{x} = \mathbf{G}(\mathbf{x}) \; . \tag{$XXI.4$}$$

Definición. Se dice que una función $\mathbf{G} : D \subset \mathcal{R}^n \to \mathcal{R}^n$ tiene un **punto fijo** en $\mathbf{p} \in D$ si $\mathbf{G}(\mathbf{p}) = \mathbf{p}$.

Para hallar la raíz vectorial $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ de la ecuación $(XXI.4)$, frecuentemente resulta conveniente utilizar el **método de iteración**

$$\mathbf{x}^{(k+1)} = \mathbf{G}(\mathbf{x}^{(k)}) \; , \tag{$XXI.5$}$$

$k = 0, 1, 2, \ldots$, donde la aproximación inicial $\mathbf{x}^{(0)} \approx \mathbf{p}$. Obsérvese que si el proceso de iteración $(XXI.5)$ converge, entonces el valor límite $\xi = \lim_{k\to\infty} \mathbf{x}^{(k)}$ es definitivamente una raíz de la ecuación $(XXI.4)$, y pasando al límite en $(XXI.5)$ para $k \to \infty$ tendremos, en virtud de la continuidad de la función $\mathbf{G}(\mathbf{x})$,

$$\xi = \lim_{k\to\infty} \mathbf{x}^{(k+1)} = \mathbf{G}(\lim_{k\to\infty} \mathbf{x}^{(k)}) = \mathbf{G}(\xi) \; .$$

De este modo, ξ es una raíz de la ecuación vectorial $(XXI.4)$. Si, además, todas las aproximaciones $\mathbf{x}^{(k)}$ $(k = 0, 1, 2, \ldots)$ pertenecen al dominio D y $\mathbf{p}$ es una raíz única del sistema en D, entonces $\xi = \mathbf{p}$.

El método de iteración puede aplicarse también al sistema general $(XXI.1)$ ó $(XXI.2)$, donde $\mathbf{F}(\mathbf{x})$ es una función vectorial, definida y continua en la vecindad D de una raíz separada $\mathbf{p}$. Por ejemplo, escribamos nuevamente este sistema de la forma

$$\mathbf{x} = \mathbf{x} + \Lambda\ \mathbf{F}(\mathbf{x}) = \mathbf{G}(\mathbf{x})\ , \qquad (XXI.6)$$

donde Λ es una matriz no singular. El método ordinario de iteración $(XXI.5)$ es directamente aplicable a esta ecuación. Si la función $\mathbf{F}(\mathbf{x})$ tiene una derivada continua $\mathbf{F}'(\mathbf{x})$ en D, se sigue que

$$\mathbf{G}'(\mathbf{x}) = I + \Lambda\ \mathbf{F}'(\mathbf{x}) = I + \Lambda\ J(\mathbf{x})\ ,$$

donde la matriz $J(\mathbf{x}) = \mathbf{F}'(\mathbf{x})$ es la **matriz Jacobiana**, definida como $J(\mathbf{x}) = J_{ij} = (\frac{\partial f_i(\mathbf{x})}{\partial x_j})$. Entonces elegiremos (debido a lo que se demostrará más adelante) la matriz Λ de forma que

$$\mathbf{G}'(\mathbf{x}^{(0)}) = I + \Lambda\ J(\mathbf{x}^{(0)}) = 0\ ,$$

de donde, si la matriz $J(\mathbf{x}^{(0)}) = \mathbf{F}'(\mathbf{x}^{(0)})$ es no singular, tendremos

$$\Lambda = -[J(\mathbf{x}^{(0)})]^{-1}\ .$$

Si $\det[J(\mathbf{x}^{(0)})] = 0$, debe eligirse una aproximación inicial $\mathbf{x}^{(0)}$ diferente.

Es decir, es esencialmente el proceso de Newton modificado (que veremos en un próximo capítulo) aplicado a la ecuación $(XXI.2)$.

Ejemplo 1.

Utilícese el método de iteración para dar una solución aproximada del sistema

$$x_1^2 + x_2^2 = 1\ ,$$
$$x_1^3 - x_2 = 0\ .$$

A partir de una construcción gráfica, puede verse que el sistema dado tiene dos soluciones que difieren únicamente en el signo. Nos limitaremos a la solución positiva. Se puede tomar $\mathbf{x}^{(0)} = (0.9, 0.5)^t$ para la aproximación inicial de la solución positiva. Estableciendo

$$\mathbf{F}(\mathbf{x}) = \begin{pmatrix} x_1^2 + x_2^2 - 1 \\ x_1^3 - x_2 \end{pmatrix}$$

tenemos

$$J(\mathbf{x}) = \mathbf{F}'(\mathbf{x}) = \begin{pmatrix} 2\ x_1 & 2\ x_2 \\ 3\ x_1^2 & -1 \end{pmatrix}$$

de donde

$$J(\mathbf{x}^{(0)}) = \begin{pmatrix} 1.8 & 1 \\ 2.43 & -1 \end{pmatrix}\ , \qquad \text{y} \qquad \det[J(\mathbf{x}^{(0)})] = -1.8 - 2.43 = -4.23\ .$$

Como la matriz $J(\mathbf{x}^{(0)})$ es no singular, existe una matriz inversa, y de este modo

$$\Lambda = -[J(\mathbf{x}^{(0)})]^{-1} = \frac{1}{4.23}\begin{pmatrix} -1 & -1 \\ -2.43 & 1.8 \end{pmatrix}.$$

Ahora hacemos

$$\begin{aligned}\mathbf{G}(\mathbf{x}) &= \mathbf{x} + \Lambda\ \mathbf{F}(\mathbf{x}) = \\ &= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \frac{1}{4.23}\begin{pmatrix} 1 & 1 \\ 2.43 & -1.8 \end{pmatrix}\begin{pmatrix} x_1^2 + x_2^2 - 1 \\ x_1^3 - x_2 \end{pmatrix}.\end{aligned}$$

Usando esta última fórmula hallamos las aproximaciones sucesivas siguientes para la solución del sistema dado (con aritmética de cuatro dígitos significativos):

$$\begin{aligned}\mathbf{x}^{(1)} &= \begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix} - \frac{1}{4.23}\begin{pmatrix} 1 & 1 \\ 2.43 & -1.8 \end{pmatrix}\begin{pmatrix} {x_1^{(0)}}^2 + {x_2^{(0)}}^2 - 1 \\ {x_1^{(0)}}^3 - x_2^{(0)} \end{pmatrix} \\ &= \begin{pmatrix} 0.9 \\ 0.5 \end{pmatrix} - \frac{1}{4.23}\begin{pmatrix} 1 & 1 \\ 2.43 & -1.8 \end{pmatrix}\begin{pmatrix} 0.060 \\ 0.229 \end{pmatrix} \\ &= \begin{pmatrix} 0.9 \\ 0.5 \end{pmatrix} - \begin{pmatrix} 0.06832 \\ -0.06298 \end{pmatrix} = \begin{pmatrix} 0.8317 \\ 0.5630 \end{pmatrix} \\ \mathbf{x}^{(2)} &= \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} - \frac{1}{4.23}\begin{pmatrix} 1 & 1 \\ 2.43 & -1.8 \end{pmatrix}\begin{pmatrix} {x_1^{(1)}}^2 + {x_2^{(1)}}^2 - 1 \\ {x_1^{(1)}}^3 - x_2^{(1)} \end{pmatrix} \\ &= \begin{pmatrix} 0.8317 \\ 0.5630 \end{pmatrix} - \frac{1}{4.23}\begin{pmatrix} 1 & 1 \\ 2.43 & -1.8 \end{pmatrix}\begin{pmatrix} 0.0090 \\ 0.0123 \end{pmatrix} \\ &= \begin{pmatrix} 0.8317 \\ 0.5630 \end{pmatrix} - \begin{pmatrix} 0.005035 \\ -0.0002700 \end{pmatrix} = \begin{pmatrix} 0.8267 \\ 0.5633 \end{pmatrix} \\ \mathbf{x}^{(3)} &= \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \end{pmatrix} - \frac{1}{4.23}\begin{pmatrix} 1 & 1 \\ 2.43 & -1.8 \end{pmatrix}\begin{pmatrix} {x_1^{(2)}}^2 + {x_2^{(2)}}^2 - 1 \\ {x_1^{(2)}}^3 - x_2^{(2)} \end{pmatrix} \\ &= \begin{pmatrix} 0.8267 \\ 0.5633 \end{pmatrix} - \frac{1}{4.23}\begin{pmatrix} 1 & 1 \\ 2.43 & -1.8 \end{pmatrix}\begin{pmatrix} 0.0010 \\ 0.0017 \end{pmatrix} \\ &= \begin{pmatrix} 0.8267 \\ 0.5633 \end{pmatrix} - \begin{pmatrix} 0.0006383 \\ -0.0001489 \end{pmatrix} = \begin{pmatrix} 0.8261 \\ 0.5634 \end{pmatrix} \\ \mathbf{x}^{(4)} &= \begin{pmatrix} x_1^{(3)} \\ x_2^{(3)} \end{pmatrix} - \frac{1}{4.23}\begin{pmatrix} 1 & 1 \\ 2.43 & -1.8 \end{pmatrix}\begin{pmatrix} {x_1^{(3)}}^2 + {x_2^{(3)}}^2 - 1 \\ {x_1^{(3)}}^3 - x_2^{(3)} \end{pmatrix} \\ &= \begin{pmatrix} 0.8261 \\ 0.5634 \end{pmatrix} - \frac{1}{4.23}\begin{pmatrix} 1 & 1 \\ 2.43 & -1.8 \end{pmatrix}\begin{pmatrix} -0.0002 \\ 0.0003 \end{pmatrix} \\ &= \begin{pmatrix} 0.8261 \\ 0.5634 \end{pmatrix} - \begin{pmatrix} 0.00002364 \\ -0.0002426 \end{pmatrix} = \begin{pmatrix} 0.8261 \\ 0.5636 \end{pmatrix}\end{aligned}$$

y así sucesivamente. Finalizando con la cuarta aproximación, tenemos las raíces

$$x_1 = 0.8261, \quad x_2 = 0.5636$$

y $\mathbf{F}(\mathbf{x}) = (0.0000, 0.0001)^t$, con una exactitud de 10^{-4}.

3. CONDICIONES PARA LA CONVERGENCIA DEL PROCESO DE ITERACION

Definición. La aplicación $\mathbf{y} = \mathbf{G}(\mathbf{x})$ se denomina **aplicación sobreyectiva** o **de contracción** en el dominio D si existe una fracción propia q, $(0 \leq q < 1)$, tal que para dos puntos cualesquiera $\mathbf{x}_1,\ \mathbf{x}_2 \in D$ sus imágenes $\mathbf{y}_1 = \mathbf{G}(\mathbf{x}_1)$ y $\mathbf{y}_2 = \mathbf{G}(\mathbf{x}_2)$ satisfagan la condición

$$||\mathbf{y}_1 - \mathbf{y}_2|| \leq q\ ||\mathbf{x}_1 - \mathbf{x}_2|| \ ,$$

para cualquier norma canónica.

Teorema XXI.2

Sea D un dominio cerrado y la aplicación $\mathbf{y} = \mathbf{G}(\mathbf{x})$ una aplicación sobreyectiva. En tal caso, si para el proceso de iteración

$$\mathbf{x}^{(k+1)} = \mathbf{G}(\mathbf{x}^{(k)}) \ , \qquad (XXI.5)$$

todas las aproximaciones sucesivas $\mathbf{x}^{(k)}$ están en D $(k = 0, 1, 2, \ldots)$, quiere decir que:

(a) independientemente de la elección de la aproximación inicial $\mathbf{x}^{(0)}$ el proceso de iteración $(XXI.5)$ converge, es decir existe el límite $\mathbf{p} = \lim\limits_{k\to\infty} \mathbf{x}^{(k)}$;

(b) el vector límite $\mathbf{p}$ es la única solución de la ecuación $\mathbf{x} = \mathbf{G}(\mathbf{x})$ en el dominio D;

(c) se cumple la estimación

$$||\mathbf{p} - \mathbf{x}^{(k)}|| \leq \frac{q^k}{1-q}\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| \ . \qquad (XXI.7)$$

Demostración: (a) para probar la convergencia de la secuencia de aproximaciones $\mathbf{x}^{(k)}$ $(k = 0, 1, 2, \ldots)$, apliquemos el criterio de Cauchy para obtener

$$\begin{aligned} ||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| &= ||(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) + (\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}) + \ldots + (\mathbf{x}^{(k+p)} - \mathbf{x}^{(k+p-1)})|| \\ &\leq ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| + ||\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}|| + \ldots + ||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k+p-1)}|| \end{aligned}$$

Utilizando la relación de iteración y la condición de contracción, obtendremos sucesivamente:

$$\begin{aligned} ||\mathbf{x}^{(s+1)} - \mathbf{x}^{(s)}|| &= ||\mathbf{G}(\mathbf{x}^{(s)}) - \mathbf{G}(\mathbf{x}^{(s-1)})|| \\ &\leq q\ ||\mathbf{x}^{(s)} - \mathbf{x}^{(s-1)}|| \leq q^2\ ||\mathbf{x}^{(s-1)} - \mathbf{x}^{(s-2)}|| \leq \ldots \\ &\leq q^s\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| \ , \end{aligned}$$

donde $s \geq 0$. Por consiguiente, hallamos

$$||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| \leq q^k\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| + q^{k+1}\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| + \ldots + q^{k+p-1}\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| \ ,$$

ó, utilizando la fórmula de la suma de los términos de una progresión geométrica, hallaremos

$$||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| \leq \frac{q^k - q^{k+p}}{1-q}\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| \leq \frac{q^k}{1-q}\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| \ .$$

Como $0 \leq q < 1$ y, como $q^k \to 0$ para $k \to \infty$, se deduce que para cualquier $\varepsilon > 0$ existe $N = N(\varepsilon)$ tal que para $k > N(\varepsilon)$ y $p > 0$ se cumple la desigualdad

$$||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| < \varepsilon$$

esto es, el criterio de Cauchy es válido para la secuencia $\mathbf{x}^{(k)}$ $(k = 0, 1, 2, \ldots)$. Por consiguiente, $\mathbf{p} = \lim_{k\to\infty} \mathbf{x}^{(k)}$ y $\mathbf{p} \in D$ ya que el dominio D es cerrado.

(b) El vector $\mathbf{p}$ es una solución de la ecuación $\mathbf{x} = \mathbf{G}(\mathbf{x})$ ya que, pasando al límite en la relación de iteración $(XXI.5)$ para $k \to \infty$ y teniendo en cuenta la continuidad en D de la función vectorial $\mathbf{G}(\mathbf{x})$, tenemos

$$\mathbf{p} = \lim_{k\to\infty} \mathbf{x}^{(k)} = \lim_{k\to\infty} \mathbf{G}(\mathbf{x}^{(k-1)}) = \mathbf{G}(\mathbf{p}) \ .$$

Esta solución es única en D. En efecto, sea $\mathbf{p}'$ otra solución, es decir $\mathbf{p}' = \mathbf{G}(\mathbf{p}')$. Restando tenemos:

$$\mathbf{p} - \mathbf{p}' = \mathbf{G}(\mathbf{p}) - \mathbf{G}(\mathbf{p}')$$

de donde

$$||\mathbf{p} - \mathbf{p}'|| = ||\mathbf{G}(\mathbf{p}) - \mathbf{G}(\mathbf{p}')|| \leq q \ ||\mathbf{p} - \mathbf{p}'||$$

ó

$$(1 - q) \ ||\mathbf{p} - \mathbf{p}'|| \leq 0 \ .$$

Como $(1-q) > 0$, la desigualdad obtenida puede mantenerse únicamente si $||\mathbf{p}-\mathbf{p}'|| = 0$, esto es cuando $\mathbf{p} = \mathbf{p}'$. Por lo tanto no puede haber otra solución de la ecuación $\mathbf{x} = \mathbf{G}(\mathbf{x})$ en el dominio D.

(c) Pasando al límite en la desigualdad

$$||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| \leq \frac{q^k}{1-q} \ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}||$$

para $p \to \infty$ obtenemos la estimación $(XXI.7)$. Esto completa la demostración del Teorema. **c.q.d.**

En el capítulo XIII hemos definido las normas l_1, l_2 y l_∞ para el vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ como

$$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i| \ , \qquad ||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \quad \text{y} \quad ||\mathbf{x}||_\infty = \max_{1\leq i\leq n} |x_i| \ .$$

Consecuentemente, las normas matriciales asociadas tienen las formas

$$||A||_1 = \max_{||\mathbf{x}||_1=1} ||A\mathbf{x}||_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \ , \quad \text{norma} \ \ l_1 \ ,$$

$$||A||_2 = \max_{||\mathbf{x}||_2=1} ||A\mathbf{x}||_2 \ , \quad \text{norma } \ l_2 \ ,$$

y

$$||A||_\infty = \max_{||\mathbf{x}||_\infty=1} ||A\mathbf{x}||_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ , \quad \text{norma } \ l_\infty \ .$$

Además con respecto al dominio D definiremos las normas:

$$||\mathbf{G}'(\mathbf{x})||_I = \max_{\mathbf{x}\in D} ||\mathbf{G}'(\mathbf{x})||_1 \qquad \text{y} \qquad ||\mathbf{G}'(\mathbf{x})||_{II} = \max_{\mathbf{x}\in D} ||\mathbf{G}'(\mathbf{x})||_\infty \ ,$$

donde, obviamente,

$$||\mathbf{G}'(\mathbf{x})||_1 = \max_{1\le j\le n} \sum_{i=1}^{n} \Big|\frac{\partial g_i(\mathbf{x})}{\partial x_j}\Big| \qquad \text{y} \qquad ||\mathbf{G}'(\mathbf{x})||_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} \Big|\frac{\partial g_i(\mathbf{x})}{\partial x_j}\Big| \ .$$

Lema XXI.3

Si $\mathbf{F}(\mathbf{x}) = (f_{ij}(\mathbf{x}))$ es una función matricial $n \times r$, donde las $f_{ij}(\mathbf{x})$ son continuas juntamente con sus derivadas parciales de primer orden en un dominio convexo que contenga los punto $\mathbf{x}$ y $\mathbf{x} + \boldsymbol{\Delta}\mathbf{x}$, entonces,

$$||\mathbf{F}(\mathbf{x} + \boldsymbol{\Delta}\mathbf{x}) - \mathbf{F}(\mathbf{x})||_\infty \le r \ ||\Delta x||_\infty \ ||\mathbf{F}'(\xi)||_\infty$$

donde $\xi = \mathbf{x} + \theta \ \boldsymbol{\Delta}\mathbf{x}$, $0 < \theta < 1$.

Teorema XXI.4

Sean las funciones $\mathbf{G}(\mathbf{x})$ y $\mathbf{G}'(\mathbf{x})$ continuas en el dominio D, cumpliéndose en D la desigualdad

$$||\mathbf{G}'(\mathbf{x})||_{II} \le q < 1 \ , \tag{$XXI.8$}$$

donde q es una constante. Si las aproximaciones sucesivas

$$\mathbf{x}^{(k+1)} = \mathbf{G}(\mathbf{x}^{(k)}) \tag{$XXI.5$}$$

$(k = 0, 1, 2, \ldots)$ están contenidas en D, el proceso de iteración $(XXI.5)$ converge y el vector límite

$$\mathbf{p} = \lim_{k\to\infty} \mathbf{x}^{(k)}$$

es la única solución del sistema $\mathbf{y} = \mathbf{G}(\mathbf{x})$ en el dominio D.

Demostración: en virtud del Teorema XXI.2, es suficiente demostrar que la aplicación $\mathbf{y} = \mathbf{G}(\mathbf{x})$ es, dada la condición $(XXI.8)$, una aplicación sobreyectiva en D para la norma l_∞.

Sean $\mathbf{x}_1, \mathbf{x}_2 \in D$ e $\mathbf{y}_i = \mathbf{G}(\mathbf{x}_i)$ $(i = 1, 2)$. Por el Lema XXI.3 tenemos:

$$\begin{aligned} ||\mathbf{y}_1 - \mathbf{y}_2||_\infty &= ||\mathbf{G}(\mathbf{x}_1) - \mathbf{G}(\mathbf{x}_2)||_\infty \le \\ &\le ||\mathbf{x}_1 - \mathbf{x}_2||_\infty \ ||\mathbf{G}'(\xi)||_\infty \le ||\mathbf{x}_1 - \mathbf{x}_2||_\infty \ ||\mathbf{G}'(\mathbf{x})||_{II} \ . \end{aligned}$$

De donde

$$||\mathbf{y}_1 - \mathbf{y}_2||_\infty \leq q \; ||\mathbf{x}_1 - \mathbf{x}_2||_\infty$$

con $0 \leq q < 1$, lo que completa la demostración. **c.q.d.**

Corolario XXI.5

El proceso de iteración

$$\mathbf{x}^{(k+1)} = \mathbf{G}(\mathbf{x}^{(k)}) \qquad (XXI.5)$$

$(k = 0, 1, 2, \ldots)$ converge si

$$\sum_{j=1}^{n} \Big| \frac{\partial g_i(\mathbf{x})}{\partial x_j} \Big| \leq q_i < 1 \qquad (XXI.9)$$

$(i = 1, 2, \ldots, n)$ cuando $\mathbf{x} \in D$.

Demostración: Evidentemente, del sistema de desigualdades $(XXI.9)$ sigue la condición $(XXI.8)$ del Teorema XXI.4. **c.q.d.**

Notése que en razón del Teorema XXI.2 obtenemos el siguiente estimado para la aproximación $\mathbf{x}^{(k)}$:

$$||\mathbf{p} - \mathbf{x}^{(k)}||_\infty \leq \frac{q^k}{1-q} \; ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}||_\infty \; ,$$

$(k = 0, 1, 2, \ldots)$.

Lema XXI.6

Si la función vectorial $\mathbf{F}(\mathbf{x}) = (f_i(\mathbf{x}))$ es continua juntamente con su derivada $\mathbf{F}'(\mathbf{x})$, en un dominio convexo que contenga los puntos $\mathbf{x}$ y $\mathbf{x} + \mathbf{\Delta x}$, entonces,

$$||\mathbf{F}(\mathbf{x} + \mathbf{\Delta x}) - \mathbf{F}(\mathbf{x})||_1 \leq ||\Delta x||_1 \; ||\mathbf{F}'(\xi)||_1$$

donde $\xi = \mathbf{x} + \theta \; \mathbf{\Delta x}$, $0 < \theta < 1$.

Teorema XXI.7

Sean las funciones $\mathbf{G}(\mathbf{x})$ y $\mathbf{G}'(\mathbf{x})$ continuas en el dominio D cerrado, convexo y acotado, cumpliéndose en D la desigualdad

$$||\mathbf{G}'(\mathbf{x})||_I \leq q < 1 \; , \qquad (XXI.10)$$

donde q es una constante. Si $\mathbf{x}^{(0)} \in D$ y todas las aproximaciones sucesivas

$$\mathbf{x}^{(k+1)} = \mathbf{G}(\mathbf{x}^{(k)}) \qquad (XXI.5)$$

$(k = 0, 1, 2, \ldots)$ pertenecen a D, el proceso de iteración $(XXI.5)$ converge y el vector límite

$$\mathbf{p} = \lim_{k \to \infty} \mathbf{x}^{(k)}$$

es la única solución del sistema $\mathbf{y} = \mathbf{G}(\mathbf{x})$ en el dominio D.

Demostración: demostraremos que $\mathbf{y} = \mathbf{G}(\mathbf{x})$ es aplicación sobreyectiva en D para la norma l_1.

Supongamos $\mathbf{x}_1, \mathbf{x}_2 \in D$ y $\mathbf{y}_i = \mathbf{G}(\mathbf{x}_i)$ $(i = 1, 2)$. Por el Lema XXI.6 tenemos:

$$||\mathbf{y}_1 - \mathbf{y}_2||_1 = ||\mathbf{G}(\mathbf{x}_1) - \mathbf{G}(\mathbf{x}_2)||_1 \leq ||\mathbf{x}_1 - \mathbf{x}_2||_1 \; ||\mathbf{G}'(\xi)||_1$$

donde $\xi \in D$. Como

$$||\mathbf{G}'(\xi)||_1 \leq \max_{\mathbf{x} \in D} ||\mathbf{G}'(\mathbf{x})||_1 = ||\mathbf{G}'(\mathbf{x})||_{II} \leq q$$

se deduce que

$$||\mathbf{y}_1 - \mathbf{y}_2||_1 \leq q \; ||\mathbf{x}_1 - \mathbf{x}_2||_1$$

donde $0 \leq q < 1$, lo que completa la demostración. **c.q.d.**

Corolario XXI.8

El proceso de iteración

$$\mathbf{x}^{(k+1)} = \mathbf{G}(\mathbf{x}^{(k)}) \qquad (XXI.5)$$

$(k = 0, 1, 2, \ldots)$ converge si

$$\sum_{j=1}^{n} \Big| \frac{\partial g_j(\mathbf{x})}{\partial x_i} \Big| \leq q_i < 1 \qquad (XXI.11)$$

$(i = 1, 2, \ldots, n)$ cuando $\mathbf{x} \in D$.

Demostración: evidentemente, del sistema de desigualdades $(XXI.11)$ se sigue la condición $(XXI.10)$ del Teorema XXI.7. **c.q.d.**

Nótese que por el Teorema XXI.2 obtenemos la siguiente estimación para la aproximación $\mathbf{x}^{(k)}$:

$$||\mathbf{p} - \mathbf{x}^{(k)}||_1 \leq \frac{q^k}{1 - q} \; ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}||_1 \; ,$$

$(k = 0, 1, 2, \ldots)$.

El siguiente Teorema generaliza los Teoremas de punto fijo VII.1 y VII.2 al caso $n-$dimensional. Este Teorema es un caso especial del bien conocido **Teorema de la Aplicación Contractiva**.

Teorema XXI.9

Sea $D = \{(x_1, x_2, \ldots, x_n)^t \mid a_i \leq x_i \leq b_i \;\text{ para cada }\; i = 1, 2, \ldots, n\}$ para alguna colección de constantes $a_1, a_2, \ldots, a_n$ y $b_1, b_2, \ldots, b_n$. Supongamos que $\mathbf{G}$ es una función continua con primeras derivadas parciales continuas de $D \subset \mathcal{R}^n$ a $\mathcal{R}^n$ con la propiedad de que $\mathbf{G}(\mathbf{x}) \in D$ para cada $\mathbf{x} \in D$. Entonces, $\mathbf{G}$ tiene un punto fijo en D. Además, supóngase que existe una constante $K < 1$ con

$$\Big| \frac{\partial g_i(\mathbf{x})}{\partial x_j} \Big| \leq \frac{K}{n} \qquad \text{siempre que } \; \mathbf{x} \in D \; ,$$

para cada $j = 1, 2, \ldots, n$ y cada función componente g_i. Entonces la sucesión $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ definida por un $\mathbf{x}_0$ en D seleccionada arbitrariamente y generada por

$$\mathbf{x}^{(k)} = \mathbf{G}(\mathbf{x}^{(k-1)}) \qquad \text{para cada } k \geq 1$$

converge al punto fijo único $\mathbf{p} \in D$ y

$$||\mathbf{x}^{(k)} - \mathbf{p}||_\infty \leq \frac{K^k}{1-K} \, ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}||_\infty \, . \qquad (XXI.12)$$

Ejemplo 2.

Considérese el sistema no lineal dado por

$$\begin{aligned} 3\, x_1 - \cos(x_2 x_3) - \frac{1}{2} &= 0 \, , \\ x_1^2 - 81\, (x_2 + 0.1)^2 + \sin x_3 + 1.06 &= 0 \, , \\ e^{-x_1 x_2} + 20\, x_3 + \frac{10\pi - 3}{3} &= 0 \, . \end{aligned}$$

Si de la $i-$ésima ecuación se despeja x_i, el sistema puede cambiarse a un problema de punto fijo

$$\begin{aligned} x_1 &= \frac{1}{3} \, \cos(x_2 x_3) + \frac{1}{6} \, , \\ x_2 &= \frac{1}{9} \, \sqrt{x_1^2 + \sin x_3 + 1.06} - 0.1 \, , \\ x_3 &= -\frac{1}{20} \, e^{-x_1 x_2} - \frac{10\pi - 3}{60} \, . \end{aligned}$$

Sea $\mathbf{G} : \mathcal{R}^3 \to \mathcal{R}^3$ definida por $\mathbf{G}(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), g_3(\mathbf{x}))^t$ donde

$$\begin{aligned} g_1(x_1, x_2, x_3) &= \frac{1}{3} \, \cos(x_2 x_3) + \frac{1}{6} \, , \\ g_2(x_1, x_2, x_3) &= \frac{1}{9} \, \sqrt{x_1^2 + \sin x_3 + 1.06} - 0.1 \, , \\ g_3(x_1, x_2, x_3) &= -\frac{1}{20} \, e^{-x_1 x_2} - \frac{10\pi - 3}{60} \, . \end{aligned}$$

Se usarán los Teoremas XXI.1 y XXI.9 para demostrar que $\mathbf{G}$ tiene un único punto fijo en

$$D = \{(x_1, x_2, x_3)^t \mid -1 \leq x_i \leq 1 \text{ para cada } i = 1, 2, 3\} \, .$$

Para $\mathbf{x} = (x_1, x_2, x_3)^t \in D$,

$$\begin{aligned} |g_1(x_1, x_2, x_3)| &\leq \frac{1}{3} \, |\cos(x_2 x_3)| + \frac{1}{6} \leq 0.50 \, , \\ |g_2(x_1, x_2, x_3)| &= |\frac{1}{9} \, \sqrt{x_1^2 + \sin x_3 + 1.06} - 0.1| \leq \\ &\leq \frac{1}{9} \, \sqrt{1 + \sin 1 + 1.06} - 0.1 < 0.09 \, , \\ |g_3(x_1, x_2, x_3)| &= \frac{1}{20} \, e^{-x_1 x_2} + \frac{10\pi - 3}{60} \leq \\ &\leq \frac{1}{20} \, e + \frac{10\pi - 3}{60} < 0.61 \, ; \end{aligned}$$

así que $-1 \leq g_i(x_1, x_2, x_3) \leq 1$, para cada $i = 1, 2, 3$. Por lo tanto, $\mathbf{G}(\mathbf{x}) \in D$ siempre que $\mathbf{x} \in D$.

Para las cotas de las derivadas parciales en D se obtiene lo siguiente:

$$\left|\frac{\partial g_1}{\partial x_1}\right| = 0 \ , \quad \left|\frac{\partial g_2}{\partial x_2}\right| = 0 \ , \quad \text{y} \quad \left|\frac{\partial g_3}{\partial x_3}\right| = 0 \ ,$$

mientras que

$$\begin{aligned}
\left|\frac{\partial g_1}{\partial x_2}\right| &\leq \frac{1}{3}\ |x_3|\ |\sin x_2 x_3| \leq \frac{1}{3}\ \sin 1 < 0.281 \ , \\
\left|\frac{\partial g_1}{\partial x_3}\right| &\leq \frac{1}{3}\ |x_2|\ |\sin x_2 x_3| \leq \frac{1}{3}\ \sin 1 < 0.281 \ , \\
\left|\frac{\partial g_2}{\partial x_1}\right| &\leq \frac{|x_1|}{9\ \sqrt{x_1^2 + \sin x_3 + 1.06}} \leq \frac{1}{9\ \sqrt{0.218}} < 0.238 \ , \\
\left|\frac{\partial g_2}{\partial x_3}\right| &\leq \frac{|\cos x_3|}{18\ \sqrt{x_1^2 + \sin x_3 + 1.06}} \leq \frac{1}{18\ \sqrt{0.218}} < 0.119 \ , \\
\left|\frac{\partial g_3}{\partial x_1}\right| &\leq \frac{|x_2|}{20}\ e^{-x_1 x_2} \leq \frac{1}{20}\ e < 0.14 \ , \\
\left|\frac{\partial g_3}{\partial x_2}\right| &\leq \frac{|x_1|}{20}\ e^{-x_1 x_2} \leq \frac{1}{20}\ e < 0.14 \ .
\end{aligned}$$

Como las derivadas parciales de g_1, g_2 y g_3 están acotadas en D, el Teorema XXI.1 implica que estas funciones son continuas en D. Consecuentemente, $\mathbf{G}$ es continua en D. Además, para todo $\mathbf{x} \in D$

$$\left|\frac{\partial g_i(\mathbf{x})}{\partial x_j}\right| \leq 0.281 \quad \text{para cada } \ i = 1, 2, 3 \quad \text{y} \quad j = 1, 2, 3 \ ,$$

y de aquí la condición de la segunda parte del Teorema XXI.9 se satisface con $K = 0.843 < 1$.

De la misma manera puede demostrarse que $\frac{\partial g_i}{\partial x_j}$ es continua en D para cada $i = 1, 2, 3$ y $j = 1, 2, 3$. Consecuentemente, $\mathbf{G}$ tiene un punto fijo único en D y el sistema no lineal tiene una solución en D.

Es necesario hacer notar que, el que $\mathbf{G}$ tenga una solución única en D no implica que la solución al sistema original sea única en este dominio. En nuestro caso, la solución para x_2 implica la elección de la raíz cuadrada principal (positiva).

Para aproximar el punto fijo $\mathbf{p}$ escogemos $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$. La sucesión de vectores generada por

$$\begin{aligned}
x_1^{(k)} &= \frac{1}{3}\ \cos(x_2^{(k-1)} x_3^{(k-1)}) + \frac{1}{6} \ , \\
x_2^{(k)} &= \frac{1}{9}\ \sqrt{(x_1^{(k-1)})^2 + \sin x_3^{(k-1)} + 1.06} - 0.1 \ , \\
x_3^{(k)} &= -\frac{1}{20}\ e^{-x_1^{(k-1)} x_2^{(k-1)}} - \frac{10\pi - 3}{60} \ ,
\end{aligned}$$

convergerá a la solución única del problema dado en forma de punto fijo. En este ejemplo se ha generado la sucesión hasta que k fuera tal que

$$||\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}||_\infty < 10^{-5} \ .$$

Tabla 1

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|\|_\infty$
0	0.100000000	0.100000000	−0.100000000	—
1	0.499983333	0.009441150	−0.523101267	0.4231
2	0.499995935	0.000025568	−0.523363316	9.4×10^{-3}
3	0.500000000	0.000012337	−0.523598136	2.3×10^{-4}
4	0.500000000	0.000000034	−0.523598467	1.2×10^{-5}
5	0.500000000	0.000000016	−0.523598775	3.1×10^{-7}

Usando la cota de error $(XXI.12)$ con $K = 0.843$ da

$$||\mathbf{x}^{(5)} - \mathbf{p}||_\infty \leq \frac{K^5}{1-K}\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}||_\infty \leq \frac{(0.843)^5}{1-0.8431}\ (0.423) < 1.15 \ ,$$

lo cual no indica la precisión real de $\mathbf{x}^{(5)}$, debido a la inexacta aproximación inicial. La solución real es

$$\mathbf{p} = \big(0.5, 0, -\frac{\pi}{6}\big)^t \approx (0.5, 0, -0.5235987756)^t \ ,$$

así que el error real es

$$||\mathbf{x}^{(5)} - \mathbf{p}||_\infty \leq 2 \times 10^{-8} \ .$$

Una manera general de acelerar la convergencia de la iteración de punto fijo en la solución de sistemas no lineales es usar las últimas estimaciones $x_1^{(k)}, \ldots, x_{i-1}^{(k)}$, en lugar de $x_1^{(k-1)}, \ldots, x_{i-1}^{(k-1)}$, para calcular $x_i^{(k)}$, como se hizo en el método de Gauss-Seidel para sistemas lineales. Las ecuaciones son entonces:

$$\begin{aligned}
x_1^{(k)} &= \frac{1}{3}\ \cos(x_2^{(k-1)} x_3^{(k-1)}) + \frac{1}{6} \ , \\
x_2^{(k)} &= \frac{1}{9}\ \sqrt{(x_1^{(k)})^2 + \sin x_3^{(k-1)} + 1.06} - 0.1 \ , \\
x_3^{(k)} &= -\frac{1}{20}\ e^{-x_1^{(k)} x_2^{(k)}} - \frac{10\pi - 3}{60} \ .
\end{aligned}$$

Con $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$, los resultados de estos cálculos se muestran en la tabla 2.

Tabla 2

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|\|_\infty$
0	0.100000000	0.100000000	−0.100000000	—
1	0.499983333	0.022229794	−0.523046126	0.4230
2	0.499977468	0.000028154	−0.523598072	2.2×10^{-2}
3	0.500000000	0.000000038	−0.523598775	2.8×10^{-5}
4	0.500000000	0.000000000	−0.523598776	3.8×10^{-8}

La iteración $\mathbf{x}^{(4)}$ es exacta dentro de 10^{-7} en la norma l_∞; usando el método de Seidel se ha acelerado realmente la convergencia para este problema. Se debe observar, sin embargo, que el método de Seidel no *siempre* acelera la convergencia.

EJERCICIOS.

1. El sistema no lineal

$$\begin{aligned} x_1^2 - 10\ x_1 + x_2^2 + 8 = 0\ , \\ x_1\ x_2^2 + x_1 - 10\ x_2 + 8 = 0\ , \end{aligned}$$

puede transformarse al problema de punto fijo

$$\begin{aligned} x_1 = g_1(x_1, x_2) = \frac{x_1^2 + x_2^2 + 8}{10}\ , \\ x_2 = g_2(x_1, x_2) = \frac{x_1\ x_2^2 + x_1 + 8}{10}\ . \end{aligned}$$

Usar el Teorema XXI.9 para demostrar que $\mathbf{G} = (g_1, g_2)^t : D \subset \mathcal{R}^2 \to \mathcal{R}^2$ tiene un único punto fijo en $D = \{(x_1, x_2)^t \mid 0 \leq x_1, x_2 \leq 1.5\}$. Aplicar la iteración funcional para aproximar la solución. ¿Acelera el método de Seidel la convergencia?

2. El sistema no lineal

$$\begin{aligned} 3\ x_1^2 - x_2^2 = 0\ , \\ 3\ x_1\ x_2^2 - x_1^3 - 1 = 0\ , \end{aligned}$$

tiene una solución cerca de $(1/2, 3/4)^t$.
Encontrar una función $\mathbf{G}$ y un conjunto $D \subset \mathcal{R}^2$ tales que $\mathbf{G} = (g_1, g_2)^t : D \subset \mathcal{R}^2 \to \mathcal{R}^2$ tenga un único punto fijo en D. Aplicar la iteración funcional para aproximar la solución. ¿Acelera el método de Seidel la convergencia?

3. Usar iteración funcional para encontrar soluciones a los siguientes sistemas no lineales:

$$(a) \quad \begin{aligned} x_1^2 + 2\ x_2^2 - x_2 - 2\ x_3 = 0\ , \\ x_1^2 - 8\ x_2^2 + 10\ x_3 = 0\ , \\ \frac{x_1^2}{7\ x_2\ x_3} - 1 = 0\ . \end{aligned}$$

$$(b) \quad \begin{aligned} 2\ x_1 + x_2 + x_3 - 4 = 0\ , \\ x_1 + 2\ x_2 + x_3 - 4 = 0\ , \\ x_1\ x_2\ x_3 - 1 = 0\ . \end{aligned}$$

4. Usar el Teorema XXI.9 para demostrar que las siguientes funciones $\mathbf{G} = (g_1, g_2, g_3)^t : D \subset \mathcal{R}^3 \to \mathcal{R}^3$ tienen un único punto fijo en D. Aplicar la iteración funcional para aproximar las soluciones dentro de 10^{-4}, usando la norma l_∞. Luego usar el método de Seidel para aproximar los puntos fijos

dentro de 10^{-4}, usando la norma l_∞.

(a) $g_1(x_1, x_2, x_3) = \dfrac{\cos(x_2\ x_3) + 0.5}{3}$,

$g_2(x_1, x_2, x_3) = \dfrac{1}{25}\ \sqrt{x_1^2 + 0.3125} - 0.03$,

$g_3(x_1, x_2, x_3) = -\dfrac{1}{20}\ e^{-x_1\ x_2} - \dfrac{10\ \pi - 3}{60}$,

$D = \{(x_1, x_2, x_3)^t \mid 0 \leq x_i \leq 1,\ i = 1, 2, 3\}$.

(b) $g_1(x_1, x_2, x_3) = \dfrac{7.17 + 3\ x_2^2 + 4\ x_3}{12}$,

$g_2(x_1, x_2, x_3) = \dfrac{11.54 + x_3 - x_1^2}{10}$,

$g_3(x_1, x_2, x_3) = \dfrac{7.631 - x_2^3}{7}$,

$D = \{(x_1, x_2, x_3)^t \mid 0 \leq x_i \leq 1.5,\ i = 1, 2, 3\}$.

(c) $g_1(x_1, x_2, x_3) = 1 - \cos(x_1\ x_2\ x_3)$,

$g_2(x_1, x_2, x_3) = 1 - (1 - x_1)^{1/4} - 0.05\ x_3^2 + 0.15\ x_3$,

$g_3(x_1, x_2, x_3) = x_1^2 + 0.1\ x_2^2 - 0.01\ x_2 + 1$,

$D = \{(x_1, x_2, x_3)^t \mid\ -0.1 \leq x_1 \leq 0.1, -0.1 \leq x_2 \leq 0.3, 0.5 \leq x_3 \leq 1.1, \}$.

CAPITULO XXII. METODO DE NEWTON

1. INTRODUCCION Y METODO

Aún cuando el problema presentado en el ejemplo 2 del Capítulo XXI puede transformarse fácilmente en un formato convergente de punto fijo, esto no sucede frecuentemente. En este capítulo consideraremos un procedimiento algorítmico que puede usarse para realizar la transformación para un problema general.

Así como en el Capítulo IX hemos introducido el método de Newton-Raphson para la resolución del problema de la búsqueda de raíces de la ecuación $f(x) = 0$ usando un enfoque intuitivo basado en el polinomio de Taylor, aquí también usaremos este procedimiento para introducir el método.

La forma general de un sistema de ecuaciones no lineales es:

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \ , \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \ , \\ \cdots \quad &\cdots \quad \cdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \ . \end{aligned} \tag{XXII.1}$$

Usando notación vectorial, el sistema (XXII.1) asume la forma:

$$\mathbf{F}(\mathbf{x}) = \mathbf{0} \ . \tag{XXII.2}$$

Resolveremos el sistema (XXII.2) por el método de aproximaciones sucesivas. Supongamos que hemos hallado la aproximación k-ésima, $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})^t$, de una de las raíces separadas $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ de la ecuación vectorial (XXII.2). La raíz exacta de (XXII.2) puede representarse entonces como

$$\mathbf{x} = \mathbf{x}^{(k)} + \varepsilon^{(k)} \ , \tag{XXII.3}$$

donde $\varepsilon^{(k)} = (\varepsilon_1^{(k)}, \varepsilon_2^{(k)}, \ldots, \varepsilon_n^{(k)})^t$ es la corrección (**error de la raíz**).

Sustituyendo (XXII.3) en (XXII.2), tenemos

$$\mathbf{F}(\mathbf{x}^{(k)} + \varepsilon^{(k)}) = \mathbf{0} \ . \tag{XXII.4}$$

Suponiendo que la función $\mathbf{F}(\mathbf{x})$ es continuamente diferenciable en un cierto dominio convexo que contiene a $\mathbf{x}$ y $\mathbf{x}^{(k)}$, desarrollemos el primer miembro de la ecuación (XXII.4) en potencias del pequeño vector $\varepsilon^{(k)}$ limitándonos a los términos lineales,

$$\mathbf{F}(\mathbf{x}^{(k)} + \varepsilon^{(k)}) = \mathbf{F}(\mathbf{x}^{(k)}) + \mathbf{F}'(\mathbf{x}^{(k)}) \ \varepsilon^{(k)} = \mathbf{0} \ . \tag{XXII.5}$$

De la fórmula (XXII.5) se deduce que la derivada $\mathbf{F}'(\mathbf{x})$ ha de ser considerada como la **matriz Jacobiana**, $J(\mathbf{x})$, del conjunto de funciones $f_1, f_2, \ldots, f_n$ con respecto a las

variables $x_1, x_2, \ldots, x_n$; esto es,

$$J(\mathbf{x}) = \mathbf{F}'(\mathbf{x}) = \left(\frac{\partial f_i}{\partial x_j}\right) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n(\mathbf{x})}{\partial x_1} & \frac{\partial f_n(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_n(\mathbf{x})}{\partial x_n} \end{pmatrix} \qquad (XXII.6)$$

$(i, j = 1, 2, \ldots, n)$.

El sistema $(XXII.5)$ es un sistema lineal en las correcciones $\varepsilon_i^{(k)}$ $(i = 1, 2, \ldots, n)$ con la matriz $J(\mathbf{x}^{(k)})$, y de aquí que la fórmula $(XXII.5)$ pueda escribirse como:

$$\mathbf{F}(\mathbf{x}^{(k)}) + J(\mathbf{x}^{(k)})\ \varepsilon^{(k)} = \mathbf{0}\ ,$$

de donde, dando por supuesto que la matriz $J(\mathbf{x}^{(k)})$ es no singular, tenemos

$$\varepsilon^{(k)} = -J^{-1}(\mathbf{x}^{(k)})\ \mathbf{F}(\mathbf{x}^{(k)})\ .$$

En consecuencia

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - J^{-1}(\mathbf{x}^{(k)})\ \mathbf{F}(\mathbf{x}^{(k)})\ , \qquad (XXII.7)$$

$(k = 0, 1, 2, \ldots)$. Para la aproximación de orden cero $\mathbf{x}^{(0)}$ podemos tomar un valor aproximado de la raíz deseada.

Ejemplo 1.

Usamos el método de Newton, con aproximación inicial $x_0 = y_0 = z_0 = 0.5$, para aproximar la solución positiva del sistema de ecuaciones

$$\begin{aligned} x^2 + y^2 + z^2 &= 1\ , \\ 2\ x^2 + y^2 - 4\ z &= 0\ , \\ 3\ x^2 - 4\ y + z^2 &= 0\ . \end{aligned}$$

Tenemos:

$$\mathbf{F}(\mathbf{x}) = \begin{pmatrix} x^2 + y^2 + z^2 - 1 \\ 2\ x^2 + y^2 - 4\ z \\ 3\ x^2 - 4\ y + z^2 \end{pmatrix}$$

de donde

$$\mathbf{F}(\mathbf{x}^{(0)}) = \begin{pmatrix} 0.25 + 0.25 + 0.25 - 1 \\ 0.50 + 0.25 - 2.00 \\ 0.75 - 2.00 + 0.25 \end{pmatrix} = \begin{pmatrix} -0.25 \\ -1.25 \\ -1.00 \end{pmatrix}\ .$$

Calculemos ahora la matriz Jacobiana:

$$J(\mathbf{x}) = \mathbf{F}'(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \frac{\partial f_1(\mathbf{x})}{\partial x_3} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \frac{\partial f_2(\mathbf{x})}{\partial x_3} \\ \frac{\partial f_3(\mathbf{x})}{\partial x_1} & \frac{\partial f_3(\mathbf{x})}{\partial x_2} & \frac{\partial f_3(\mathbf{x})}{\partial x_3} \end{pmatrix} = \begin{pmatrix} 2\ x & 2\ y & 2\ z \\ 4\ x & 2\ y & -4 \\ 6\ x & -4 & 2\ z \end{pmatrix}\ .$$

Por consiguiente tenemos

$$J(\mathbf{x}^{(0)}) = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & -4 \\ 3 & -4 & 1 \end{pmatrix} ,$$

y

$$\Delta = \det J(\mathbf{x}^{(0)}) = -40.0 .$$

De este modo sabemos ya que la matriz $J(\mathbf{x}^{(0)})$ es no singular. Calculemos la inversa

$$J^{-1}(\mathbf{x}^{(0)}) = \frac{1}{\Delta} \begin{pmatrix} -15 & -5 & -5 \\ -14 & -2 & 6 \\ -11 & 7 & -1 \end{pmatrix} .$$

Utilizando la fórmula $(XXII.7)$, obtendremos la primera aproximación:

$$\begin{aligned} \mathbf{x}^{(1)} &= \mathbf{x}^{(0)} - J^{-1}(\mathbf{x}^{(0)})\ \mathbf{F}(\mathbf{x}^{(0)}) = \\ &= \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} + \frac{1}{40} \begin{pmatrix} -15 & -5 & -5 \\ -14 & -2 & 6 \\ -11 & 7 & -1 \end{pmatrix} \begin{pmatrix} -0.25 \\ -1.25 \\ -1.00 \end{pmatrix} = \\ &= \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} + \begin{pmatrix} 0.375 \\ 0.0 \\ -0.125 \end{pmatrix} = \begin{pmatrix} 0.875 \\ 0.50 \\ 0.375 \end{pmatrix} . \end{aligned}$$

Para calcular la segunda aproximación $\mathbf{x}^{(2)}$, calculemos previamente:

$$\mathbf{F}(\mathbf{x}^{(1)}) = \begin{pmatrix} 0.875^2 + 0.5^2 + 0.375^2 - 1 \\ 2\ 0.875^2 + 0.5^2 - 4\ 0.375 \\ 3\ 0.875^2 - 4\ 0.5 + 0.375^2 \end{pmatrix} = \begin{pmatrix} 0.15625 \\ 0.28125 \\ 0.43750 \end{pmatrix} ,$$

y

$$J(\mathbf{x}^{(1)}) = \begin{pmatrix} 1.75 & 1 & 0.75 \\ 3.50 & 1 & -4 \\ 5.25 & -4 & 0.75 \end{pmatrix} .$$

Por consiguiente

$$\Delta = \det J(\mathbf{x}^{(1)}) = -64.75 ,$$

y

$$J^{-1}(\mathbf{x}^{(1)}) = \frac{1}{\Delta} \begin{pmatrix} -15.25 & -3.75 & -4.75 \\ -23.625 & -2.625 & 9.625 \\ -19.25 & 12.25 & -1.75 \end{pmatrix} .$$

Utilizando la fórmula $(XXII.7)$, obtendremos la segunda aproximación:

$$\begin{aligned} \mathbf{x}^{(2)} &= \mathbf{x}^{(1)} - J^{-1}(\mathbf{x}^{(1)})\ \mathbf{F}(\mathbf{x}^{(1)}) = \\ &= \begin{pmatrix} 0.875 \\ 0.50 \\ 0.375 \end{pmatrix} + \frac{1}{64.75} \begin{pmatrix} -15.25 & -3.75 & -4.75 \\ -23.625 & -2.625 & 9.625 \\ -19.25 & 12.25 & -1.75 \end{pmatrix} \begin{pmatrix} 0.15625 \\ 0.28125 \\ 0.43750 \end{pmatrix} = \\ &= \begin{pmatrix} 0.875 \\ 0.50 \\ 0.375 \end{pmatrix} - \begin{pmatrix} 0.08519 \\ 0.00338 \\ 0.00507 \end{pmatrix} = \begin{pmatrix} 0.78981 \\ 0.49662 \\ 0.36993 \end{pmatrix} . \end{aligned}$$

Las aproximaciones siguientes se hallan análogamente:

$$\mathbf{x}^{(3)} = \begin{pmatrix} 0.78521 \\ 0.49662 \\ 0.36992 \end{pmatrix} \qquad \text{y} \qquad \mathbf{F}(\mathbf{x}^{(3)}) = \begin{pmatrix} 0.00001 \\ 0.00004 \\ 0.00005 \end{pmatrix} ,$$

y así sucesivamente. Finalizando con la tercera aproximación, tenemos:

$$x = 0.78521, \qquad y = 0.49662, \qquad z = 0.36992 \ .$$

Ejemplo 2.

Queremos aproximar las soluciones positivas del sistema de ecuaciones no lineales dado por

$$\begin{aligned} x_1 + 3\ \log_{10} x_1 - x_2^2 &= 0 \ , \\ 2\ x_1^2 - x_1\ x_2 - 5\ x_1 + 1 &= 0 \ . \end{aligned}$$

Las curvas definidas en el sistema dado se cortan aproximadamente en los puntos $M_1\ (1.4, -1.5)$ y $M_2\ (3.4, 2.2)$. Comenzando con la aproximación inicial $\mathbf{x}^{(0)} = (3.4, 2.2)^t$ calculemos las segundas aproximaciones de la raíz, efectuando los cálculos con cinco digítos significativos. Tenemos:

$$\mathbf{F}(\mathbf{x}^{(0)}) = \begin{pmatrix} 3.4 + 3\ \log_{10} 3.4 - 2.2^2 \\ 2\ 3.4^2 - 3.4\ 2.2 - 5\ 3.4 + 1 \end{pmatrix} = \begin{pmatrix} 0.15440 \\ -0.36000 \end{pmatrix} \ .$$

Calculemos ahora la matriz Jacobiana:

$$J(\mathbf{x}) = \mathbf{F}'(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 1 + \frac{3\ M}{x_1} & -2\ x_2 \\ 4\ x_1 - x_2 - 5 & -x_1 \end{pmatrix} ,$$

donde $M = 0.43429$; por consiguiente

$$J(\mathbf{x}^{(0)}) = \begin{pmatrix} 1 + \frac{3\ 0.43429}{3.4} & -2\ 2.2 \\ 4\ 3.4 - 2.2 - 5 & -3.4 \end{pmatrix} = \begin{pmatrix} 1.3832 & -4.4 \\ 6.4 & -3.4 \end{pmatrix} ,$$

y

$$\Delta = \det J(\mathbf{x}^{(0)}) = 23.457 \ .$$

De este modo, la matriz $J(\mathbf{x}^{(0)})$ es no singular. Calculemos la inversa

$$J^{-1}(\mathbf{x}^{(0)}) = \frac{1}{\Delta} \begin{pmatrix} -3.4 & 4.4 \\ -6.4 & 1.3832 \end{pmatrix} \ .$$

Utilizando la fórmula $(XXII.7)$, obtendremos:

$$\begin{aligned} \mathbf{x}^{(1)} &= \mathbf{x}^{(0)} - J^{-1}(\mathbf{x}^{(0)})\ \mathbf{F}(\mathbf{x}^{(0)}) = \\ &= \begin{pmatrix} 3.4 \\ 2.2 \end{pmatrix} - \frac{1}{23.457} \begin{pmatrix} -3.4 & 4.4 \\ -6.4 & 1.3832 \end{pmatrix} \begin{pmatrix} 0.1544 \\ -0.3600 \end{pmatrix} = \\ &= \begin{pmatrix} 3.4 \\ 2.2 \end{pmatrix} + \begin{pmatrix} 0.089909 \\ 0.063354 \end{pmatrix} = \begin{pmatrix} 3.4899 \\ 2.2634 \end{pmatrix} \ . \end{aligned}$$

Las aproximaciones siguientes se hallan de forma análoga; los resultados se dan en la tabla siguiente:

Tabla 1

k	$x_1^{(k)}$	$\varepsilon_1 = \Delta x_1$	$x_2^{(k)}$	$\varepsilon_2 = \Delta x_2$
0	3.4	+0.089909	2.2	+0.063354
1	3.4899	−0.0008	2.2634	−0.0012
2	3.4891	−0.0016	2.2621	−0.0005
3	3.4875		2.2616	

Finalizando con la aproximación $\mathbf{x}^{(3)}$, tenemos

$$x_1 = 3.4875, \quad x_2 = 2.2616 \qquad \text{y} \qquad \mathbf{F}(\mathbf{x}^{(3)}) = (0.0002, 0.0)^t \ .$$

Además del enfoque intuitivo basado en el polinomio de Taylor, para construir el algoritmo que nos llevó a un método apropiado de punto fijo en el caso unidimensional, en el Capítulo X, tratamos de encontrar una función ϕ con la propiedad de que

$$g(x) = x - \phi(x)\ f(x) \ , \tag{$XXII.8$}$$

dando convergencia cuadrática al punto fijo p de g. De esta condición surgió el método de Newton, escogiendo $\phi(x) = 1/f'(x)$.

Usando un enfoque similar para el caso $n-$dimensional, es necesaria una matriz

$$A(\mathbf{x}) = \begin{pmatrix} a_{11}(\mathbf{x}) & a_{12}(\mathbf{x}) & \dots & a_{1n}(\mathbf{x}) \\ a_{21}(\mathbf{x}) & a_{22}(\mathbf{x}) & \dots & a_{2n}(\mathbf{x}) \\ \dots & \dots & \dots & \dots \\ a_{n1}(\mathbf{x}) & a_{n2}(\mathbf{x}) & \dots & a_{nn}(\mathbf{x}) \end{pmatrix} \tag{$XXII.9$}$$

donde cada una de las componentes $a_{ij}(\mathbf{x})$ sea una función de $\mathcal{R}^n$ a $\mathcal{R}$. El procedimiento requiere encontrar una matriz $A(\mathbf{x})$ tal que

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - A(\mathbf{x})^{-1}\ \mathbf{F}(\mathbf{x}) \ , \tag{$XXII.10$}$$

dé convergencia cuadrática a la solución de $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, siempre que, desde luego, $A(\mathbf{x})$ sea no singular en el punto fijo de $\mathbf{G}$. El siguiente Teorema verifica que este enfoque puede usarse para justificar la elección de A.

Teorema XXII.1

Supóngase que $\mathbf{p}$ es una solución de $\mathbf{G}(\mathbf{x}) = \mathbf{x}$, para alguna función $\mathbf{G} = (g_1, g_2, \ldots, g_n)^t$, que manda a $\mathcal{R}^n$ en $\mathcal{R}^n$. Si existe un número $\delta > 0$ con la propiedad de que

i) $\partial g_i/\partial x_j$ es continua en $N_\delta = \{\mathbf{x} \mid ||\mathbf{x} - \mathbf{p}|| < \delta\}$ para cada $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, n$,

ii) $\partial^2 g_i(\mathbf{x})/(\partial x_j \partial x_k)$ es continua y $|\partial^2 g_i(\mathbf{x})/(\partial x_j \partial x_k)| \leq M$ para alguna constante M, siempre que $\mathbf{x} \in N_\delta$ para cada $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$ y $k = 1, 2, \ldots, n$,

iii) $\partial g_i(\mathbf{p})/\partial x_j = 0$ para cada $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, n$.

Entonces existe un número $\tilde{\delta} \leq \delta$ tal que la sucesión generada por $\mathbf{x}^{(k)} = \mathbf{G}(\mathbf{x}^{(k-1)})$ converge cuadráticamente a $\mathbf{p}$ para cualquier $\mathbf{x}^{(0)}$ siempre que $||\mathbf{x}^{(0)} - \mathbf{p}|| < \tilde{\delta}$. Además,

$$||\mathbf{x}^{(k)} - \mathbf{p}||_\infty \leq \frac{n^2 M}{2} \; ||\mathbf{x}^{(k-1)} - \mathbf{p}||_\infty^2 \qquad \text{para cada } k \geq 1 \; .$$

Este Teorema es extensión del Teorema X.1 y su demostración requiere que se pueda expresar $\mathbf{G}$ en términos de su serie de Taylor en n variables alrededor del punto $\mathbf{p}$.

Para utilizar el Teorema XXII.1, supongamos que $A(\mathbf{x})$ es una matriz $n \times n$ de funciones de $\mathcal{R}^n$ a $\mathcal{R}$ en la forma de la ecuación $(XXII.8)$, donde las componentes específicas se escogerán más adelante. Supongamos además, que $A(\mathbf{x})$ es no singular cerca de una solución $\mathbf{p}$ de $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ y denotemos por $b_{ij}(\mathbf{x})$ a la componente de $A(\mathbf{x})^{-1}$ en el $i-$ésima fila y en la $j-$ésima columna. Como $\mathbf{G}(\mathbf{x}) = \mathbf{x} - A(\mathbf{x})^{-1} \; \mathbf{F}(\mathbf{x})$,

$$g_i(\mathbf{x}) = x_i - \sum_{j=1}^{n} b_{ij}(\mathbf{x}) \; f_j(\mathbf{x}) \; ;$$

así que

$$\frac{\partial g_i(\mathbf{x})}{\partial x_k} = \begin{cases} 1 - \sum\limits_{j=1}^{n} b_{ij}(\mathbf{x}) \; \frac{\partial f_j(\mathbf{x})}{\partial x_k} + \frac{\partial b_{ij}(\mathbf{x})}{\partial x_k} \; f_j(\mathbf{x}) \; , & \text{si } k = i \; ; \\ \quad - \sum\limits_{j=1}^{n} b_{ij}(\mathbf{x}) \; \frac{\partial f_j(\mathbf{x})}{\partial x_k} + \frac{\partial b_{ij}(\mathbf{x})}{\partial x_k} \; f_j(\mathbf{x}) \; , & \text{si } k \neq i \; . \end{cases}$$

El Teorema XXII.1 implica que necesitamos tener $\partial g_i(\mathbf{p})/\partial x_k = 0$ para cada $i = 1, 2, \ldots, n$ y $k = 1, 2, \ldots, n$. Esto significa que para $k = i$,

$$0 = 1 - \sum_{j=1}^{n} b_{ij}(\mathbf{p}) \; \frac{\partial f_j(\mathbf{p})}{\partial x_i} \; ,$$

con lo que

$$\sum_{j=1}^{n} b_{ij}(\mathbf{p}) \; \frac{\partial f_j(\mathbf{p})}{\partial x_i} = 1 \; , \qquad (XXII.11)$$

y cuando $k \neq i$,

$$0 = - \sum_{j=1}^{n} b_{ij}(\mathbf{p}) \; \frac{\partial f_j(\mathbf{p})}{\partial x_k} \; ,$$

con lo que

$$\sum_{j=1}^{n} b_{ij}(\mathbf{p}) \; \frac{\partial f_j(\mathbf{p})}{\partial x_k} = 0 \; . \qquad (XXII.12)$$

Usando la matriz Jacobiana $J(\mathbf{x})$, vemos que las condiciones $(XXII.11)$ y $(XXII.12)$ requieren que

$$A(\mathbf{p})^{-1} \; J(\mathbf{p}) = I \; ,$$

y por lo tanto

$$A(\mathbf{p}) = J(\mathbf{p}) \ .$$

Una elección apropiada para $A(\mathbf{x})$ es consecuentemente $A(\mathbf{x}) = J(\mathbf{x})$, ya que la condición (iii) del Teorema XXII.1 se satisface con esta elección.

La función $\mathbf{G}$ se define como

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - J(\mathbf{x})^{-1} \ \mathbf{F}(\mathbf{x}) \ ,$$

y el procedimiento de iteración funcional surge de seleccionar $\mathbf{x}^{(0)}$ y de generar, para $k \geq 1$,

$$\mathbf{x}^{(k)} = \mathbf{G}(\mathbf{x}^{(k-1)}) = \mathbf{x}^{(k-1)} - J(\mathbf{x}^{(k-1)})^{-1} \ \mathbf{F}(\mathbf{x}^{(k-1)}) \ . \qquad (XXII.13)$$

Este método se llama **método de Newton para sistemas no lineales** y se espera que generalmente dé convergencia cuadrática siempre y cuando se conozca un valor inicial lo suficientemente exacto y que $J(\mathbf{p})^{-1}$ exista.

2. ALGORITMO Y EJEMPLOS

Una debilidad clara del procedimiento del método de Newton surge de la necesidad de invertir la matriz $J(\mathbf{x})$ en cada paso. En la práctica, el método se realiza generalmente en una forma de dos pasos. Primero, se encuentra un vector $\mathbf{y}$ que satisfaga

$$J(\mathbf{x}^{(k)}) \ \mathbf{y} = -\mathbf{F}(\mathbf{x}^{(k)}) \ .$$

Después de que se ha logrado esto, la nueva aproximación $\mathbf{x}^{(k+1)}$ se puede obtener sumando $\mathbf{y}$ a $\mathbf{x}^{(k)}$; es decir,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{y} \ .$$

El siguiente algoritmo usa este procedimiento de dos pasos.

Algoritmo del método de Newton para sistemas no lineales.

==========

Para aproximar una solución del sistema no lineal $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, dada una aproximación inicial $\mathbf{x}$:

Entrada: número n de ecuaciones e incógnitas; aproximación inicial $\mathbf{x} = (x_1, \ x_2, \ \ldots, \ x_n)^t$; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada $\mathbf{x} = (x_1, \ x_2, \ \ldots, \ x_n)^t$ o mensaje de que el número de iteraciones fue excedido.

Paso 1: tomar $k = 1$;

Paso 2: mientras que $k \leq N_0$ seguir pasos 3–7;

Paso 3: calcular $\mathbf{F}(\mathbf{x})$ y $J(\mathbf{x})$, donde $(J(\mathbf{x}))_{ij} = (\partial f_i(\mathbf{x})/\partial x_j)$ para $1 \leq i, j \leq n$;

Paso 4: resolver el sistema lineal de $J(\mathbf{x}) \ \mathbf{y} = -\mathbf{F}(\mathbf{x})$;

Paso 5: tomar $\mathbf{x} = \mathbf{x} + \mathbf{y}$;

Paso 6: si $||\mathbf{y}|| < TOL$ entonces SALIDA ($\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$);
(procedimiento completado satisfactoriamente) PARAR;

Paso 7: tomar $k = k + 1$;

Paso 8: SALIDA ($'$*Número máximo* N_0 *de iteraciones excedido,* $N_0 =$ $', N_0$);
PARAR.

==========

Ejemplo 3.

Consideremos el sistema no lineal dado por

$$\begin{aligned} 3\ x_1 - \cos(x_2 x_3) - \frac{1}{2} &= 0\ , \\ x_1^2 - 81\ (x_2 + 0.1)^2 + \sin x_3 + 1.06 &= 0\ , \\ e^{-x_1 x_2} + 20\ x_3 + \frac{10\pi - 3}{3} &= 0\ . \end{aligned}$$

En el ejemplo 2 del capítulo XXI se demostró que este sistema no lineal tenía una solución aproximada en $(0.5, 0, -0.52359877)^t$. Se usará el método de Newton para obtener esta aproximación cuando la aproximación inicial es $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$.

La matriz Jacobiana $J(\mathbf{x})$ para este sistema está dada por

$$J(\mathbf{x}) = \begin{pmatrix} 3 & x_3\ \sin(x_2 x_3) & x_2\ \sin(x_2 x_3) \\ 2\ x_1 & -162\ (x_2 + 0.1) & \cos x_3 \\ -x_2\ e^{-x_1 x_2} & -x_1\ e^{-x_1 x_2} & 20 \end{pmatrix}\ ,$$

y

$$\begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{pmatrix} = \begin{pmatrix} x_1^{(k-1)} \\ x_2^{(k-1)} \\ x_3^{(k-1)} \end{pmatrix} + \begin{pmatrix} y_1^{(k-1)} \\ y_2^{(k-1)} \\ y_3^{(k-1)} \end{pmatrix}\ ,$$

donde

$$\begin{pmatrix} y_1^{(k-1)} \\ y_2^{(k-1)} \\ y_3^{(k-1)} \end{pmatrix} = -[J(x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)})]^{-1}\ \mathbf{F}(x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)})\ .$$

Por lo tanto, en el $k-$ésimo paso, se debe resolver el sistema lineal

$$\begin{pmatrix} 3 & x_3^{(k-1)}\ \sin(x_2^{(k-1)} x_3^{(k-1)}) & x_2^{(k-1)}\ \sin(x_2^{(k-1)} x_3^{(k-1)}) \\ 2\ x_1^{(k-1)} & -162\ (x_2^{(k-1)} + 0.1) & \cos x_3^{(k-1)} \\ -x_2^{(k-1)}\ e^{-x_1^{(k-1)} x_2^{(k-1)}} & -x_1^{(k-1)}\ e^{-x_1^{(k-1)} x_2^{(k-1)}} & 20 \end{pmatrix} \times$$

$$\times \begin{pmatrix} x_1^{(k)} - x_1^{(k-1)} \\ x_2^{(k)} - x_2^{(k-1)} \\ x_3^{(k)} - x_3^{(k-1)} \end{pmatrix} = - \begin{pmatrix} 3\ x_1^{(k-1)} - \cos(x_2^{(k-1)} x_3^{(k-1)}) - \frac{1}{2} \\ (x_1^{(k-1)})^2 - 81\ (x_2^{(k-1)} + 0.1)^2 + \sin x_3^{(k-1)} + 1.06 \\ e^{-x_1^{(k-1)} x_2^{(k-1)}} + 20\ x_3^{(k-1)} + \frac{10\pi - 3}{3} \end{pmatrix}\ .$$

Los resultados obtenidos usando este procedimiento iterativo se muestran en la tabla siguiente:

Tabla 2

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|\|_\infty$
0	0.10000000	0.10000000	−0.10000000	—
1	0.50003702	0.01946686	−0.52152047	0.422
2	0.50004593	0.00158859	−0.52355711	1.79×10^{-2}
3	0.50000034	0.00001244	−0.52359845	1.58×10^{-3}
4	0.50000000	0.00000004	−0.52359877	1.24×10^{-5}
5	0.50000000	0.00000000	−0.52359877	0

Nótese que la convergencia del método de Newton se hace muy rápida una vez que estamos cerca de la solución **p**. Esto ilustra la convergencia cuadrática del método cerca de la solución.

EJERCICIOS.

1. Encontrar una solución a los siguientes sistemas no lineales usando el método de Newton, y calcular la cota de error dada en el Teorema XXII.1. Iterar hasta que $||\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}||_\infty < 10^{-5}$.

$$(a) \qquad \begin{aligned} & x_1^2 - 10\ x_1 + x_2^2 + 8 = 0\ , \\ & x_1\ x_2^2 + x_1 - 10\ x_2 + 8 = 0\ . \end{aligned}$$

$$(b) \qquad \begin{aligned} & 3\ x_1^2 - x_2^2 = 0\ , \\ & 3\ x_1\ x_2^2 - x_1^3 - 1 = 0\ . \end{aligned}$$

Comparar la convergencia con la de los ejercicios 1 y 2 del Capítulo XXI.

2. Encontrar una solución a los siguientes sistemas no lineales usando el método de Newton, y calcular la cota de error dada en el Teorema XXII.1. Iterar hasta que $||\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}||_\infty < 10^{-5}$.

$$(a) \qquad \begin{aligned} & x_1^2 + x_2 - 37 = 0\ , \\ & x_1 - x_2^2 - 5 = 0\ , \\ & x_1 + x_2 + x_3 - 3 = 0\ . \end{aligned}$$

$$(b) \qquad \begin{aligned} & 12\ x_1 - 3\ x_2^2 - 4\ x_3 = 7.17\ , \\ & x_1^2 + 10\ x_2 - x_3 = 11.54\ , \\ & x_2^3 + 7\ x_3 = 7.631\ . \end{aligned}$$

3. Encontrar una solución diferente de $(1, 1, 1)^t$ a los siguientes sistemas no lineales usando el método de Newton, y calcular la cota de error dada en el Teorema XXII.1. Iterar hasta que $||\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}||_\infty < 10^{-5}$.

$$(a) \qquad \begin{aligned} & 2\ x_1 + x_2 + x_3 + x_4 - 5 = 0\ , \\ & x_1 + 2\ x_2 + x_3 + x_4 - 5 = 0\ , \\ & x_1 + x_2 + 2\ x_3 + x_4 - 5 = 0\ , \\ & x_1\ x_2\ x_3\ x_4 - 1 = 0\ . \end{aligned}$$

$$(b) \qquad \begin{aligned} & 2\ x_1 + x_2 + x_3 - 4 = 0\ , \\ & x_1 + 2\ x_2 + x_3 - 4 = 0\ , \\ & x_1\ x_2\ x_3 - 1 = 0\ . \end{aligned}$$

CAPITULO XXIII. METODOS CUASI-NEWTON

1. EL METODO DE NEWTON MODIFICADO

El mayor inconveniente al realizar el proceso de Newton

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - J^{-1}(\mathbf{x}^{(k)})\ \mathbf{F}(\mathbf{x}^{(k)})\ , \qquad (k = 0, 1, 2, \ldots) \tag{XXIII.1}$$

es la necesidad de calcular la matriz inversa $J^{-1}(\mathbf{x}^{(k)})$ en cada paso. Si la matriz inversa $J^{-1}(\mathbf{x}^{(k)})$ es continua en la proximidad de la solución deseada $\mathbf{p}$ y la aproximación inicial $\mathbf{x}^{(0)}$ está suficientemente próxima a $\mathbf{p}$, entonces podemos escribir de forma aproximada $J^{-1}(\mathbf{x}^{(k)}) \approx J^{-1}(\mathbf{x}^{(0)})$. Llegamos de este modo al **proceso de Newton modificado**:

$$\xi^{(k+1)} = \xi^{(k)} - J^{-1}(\mathbf{x}^{(0)})\ \mathbf{F}(\xi^{(k)})\ , \qquad (k = 0, 1, 2, \ldots)\ , \tag{XXIII.2}$$

donde $\xi^{(0)} = \mathbf{x}^{(0)}$. Obsérvese que las primeras aproximaciones $\xi^{(1)}$ y $\mathbf{x}^{(1)}$ coinciden para los procesos (XXIII.1) y (XXIII.2).

2. EL METODO DE BROYDEN

Una debilidad importante del método de Newton para resolver sistemas de ecuaciones no lineales es la necesidad de calcular la matriz Jacobiana en cada iteración y resolver un sistema lineal $n \times n$ asociado a esta matriz.

Para ilustrar la magnitud de esta debilidad, consideremos la cantidad de cálculos necesarios para llevar a cabo una iteración del método de Newton. La matriz Jacobiana asociada a un sistema de n ecuaciones no lineal escrita en la forma $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ requiere la determinación y evaluación de las n^2 derivadas parciales de las n componentes de $\mathbf{F}$. En la mayoría de los casos, la evaluación exacta de las derivadas parciales es complicada y en muchas ocasiones imposible. Para superar esta dificultad se pueden usar aproximaciones de diferencia finita a las derivadas parciales. Por ejemplo,

$$\frac{\partial f_j}{\partial x_k}(\mathbf{x}^{(i)}) \approx \frac{f_j(\mathbf{x}^{(i)} + \mathbf{e}_k\ h) - f_j(\mathbf{x}^{(i)})}{h}\ , \tag{XXIII.3}$$

donde h es pequeño en valor absoluto y $\mathbf{e}_k$ es el vector cuyo único elemento diferente de cero es 1 en la $k-$ésima coordenada. Esta aproximación, sin embargo, requiere aún la realización de por lo menos n^2 evaluaciones funcionales escalares para aproximar el Jacobiano y no reduce la cantidad de cálculos, que es en general $O(n^3)$, para resolver el sistema lineal que contiene al Jacobiano aproximado. El esfuerzo computacional total para sólo una iteración del método de Newton es entonces de por lo menos, $(n^2 + n)$ evaluaciones funcionales escalares (n^2 para la evaluación de la matriz Jacobiana y n para la evaluación de $\mathbf{F}$) junto con $O(n^3)$ operaciones aritméticas para resolver el sistema lineal. Esta cantidad de esfuerzo computacional es prohibitiva excepto para valores relativamente pequeños de n y para funciones escalares fáciles de evaluar.

En este apartado consideraremos una generalización del método de la secante en sistemas de ecuaciones no lineales y, en particular, consideraremos una técnica conocida

como el **método de Broyden**. El método requiere solamente de n evaluaciones funcionales escalares por iteración y reduce también el número de cálculos aritméticos a $O(n^2)$. Es uno de los método conocidos como *renovaciones de secante de mínimo cambio* que producen los algoritmos llamados cuasi-Newton. Estos métodos reemplazan la matriz Jacobiana en el método de Newton por una matriz de aproximación que se renueva en cada iteración. La desventaja de estos métodos consiste en que se pierde la convergencia cuadrática del método de Newton y se reemplaza por una convergencia llamada **superlineal**. Esto implica que $\lim_{i\to\infty} \frac{||\mathbf{x}^{(i+1)}-\mathbf{p}||}{||\mathbf{x}^{(i)}-\mathbf{p}||} = 0$, donde $\mathbf{p}$ denota la solución a $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ y $\mathbf{x}^{(i)}$ y $\mathbf{x}^{(i+1)}$ son aproximaciones consecutivas.

En la mayoría de las aplicaciones, la reducción a la convergencia superlineal es un cambio más que aceptable por la reducción en la cantidad de cómputo. Otra desventaja adicional consiste en que, a diferencia del método de Newton, los métodos cuasi-Newton no son autocorregibles. El método de Newton, por ejemplo, generalmente corrige errores de redondeo mediante iteraciones sucesivas; en cambio, el método de Broyden no lo hace.

Supongamos que se da una aproximación inicial $\mathbf{x}^{(0)}$ a la solución $\mathbf{p}$ de $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. Calculamos la siguiente aproximación $\mathbf{x}^{(1)}$ de la misma manera que en el método de Newton ($\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - J^{-1}(\mathbf{x}^{(0)})\ \mathbf{F}(\mathbf{x}^{(0)})$), o, si no conviene determinar $J(\mathbf{x}^{(0)})$ exactamente, usamos la ecuación de diferencia dada en $(XXIII.3)$ para aproximar las derivadas parciales. Sin embargo, para calcular $\mathbf{x}^{(2)}$, nos apartamos del método de Newton y examinamos el método de la secante para una sola ecuación no lineal. El método de la secante usa la aproximación $\frac{f(x_1)-f(x_0)}{x_1-x_0}$ para reemplazarla por $f'(x_1)$ en el método de Newton. Para sistemas no lineales, $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ es un vector y el cociente correspondiente está indefinido. Sin embargo, el método procede de manera similar reemplazando la matriz $J(\mathbf{x}^{(1)})$ en el método de Newton por una matriz A_1 con la propiedad de que

$$A_1(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}) = \mathbf{F}(\mathbf{x}^{(1)}) - \mathbf{F}(\mathbf{x}^{(0)})\ . \qquad (XXIII.4)$$

La ecuación $(XXIII.4)$ no define a una matriz sóla debido a que no describe cómo actua A_1 en vectores ortogonales a $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$. Como no se tiene información disponible acerca de los cambios de $\mathbf{F}$ en dirección ortogonal a $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ se pide además que

$$A_1\ \mathbf{z} = J(\mathbf{x}^{(0)})\ \mathbf{z} \quad \text{siempre que } (\mathbf{x}^{(1)} - \mathbf{x}^{(0)})^t\ \cdot\ \mathbf{z} = 0\ . \qquad (XXIII.5)$$

Esta condición especifica que cualquier vector ortogonal a $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ no será afectado por la renovación de $J(\mathbf{x}^{(0)})$, que se utilizó para calcular $\mathbf{x}^{(1)}$, a A_1, y que se usará en la determinación de $\mathbf{x}^{(2)}$.

Las condiciones $(XXIII.4)$ y $(XXIII.5)$ definen únicamente a A_1 como

$$A_1 = J(\mathbf{x}^{(0)}) + \frac{[\mathbf{F}(\mathbf{x}^{(1)}) - \mathbf{F}(\mathbf{x}^{(0)}) - J(\mathbf{x}^{(0)})\ (\mathbf{x}^{(1)} - \mathbf{x}^{(0)})]\ (\mathbf{x}^{(1)} - \mathbf{x}^{(0)})^t}{||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}||_2^2}\ .$$

Es esta matriz la que se usa en lugar de $J(\mathbf{x}^{(1)})$ para determinar $\mathbf{x}^{(2)}$:

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - A_1^{-1}\ \mathbf{F}(\mathbf{x}^{(1)})\ .$$

El método se puede repetir entonces para encontrar $\mathbf{x}^{(3)}$ usando A_1 en lugar de $A_0 = J(\mathbf{x}^{(0)})$ y con $\mathbf{x}^{(2)}$ y $\mathbf{x}^{(1)}$ en lugar de $\mathbf{x}^{(1)}$ y $\mathbf{x}^{(0)}$. En general, una vez que se ha encontrado $\mathbf{x}^{(i)}$, se calcula $\mathbf{x}^{(i+1)}$ mediante

$$A_i = A_{i-1} + \frac{[\mathbf{F}(\mathbf{x}^{(i)}) - \mathbf{F}(\mathbf{x}^{(i-1)}) - A_{i-1}\ (\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)})]\ (\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)})^t}{||\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}||_2^2}\ , \quad (XXIII.6)$$

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - A_i^{-1}\ \mathbf{F}(\mathbf{x}^{(i)})\ . \quad (XXIII.7)$$

Introduciendo la notación $\mathbf{y}_i = \mathbf{F}(\mathbf{x}^{(i)}) - \mathbf{F}(\mathbf{x}^{(i-1)})$ y $\mathbf{s}_i = \mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}$ la fórmula $(XXIII.6)$ se reescribe como

$$A_i = A_{i-1} + \frac{(\mathbf{y}_i - A_{i-1}\ \mathbf{s}_i)}{||\mathbf{s}_i||_2^2}\ \mathbf{s}_i^t\ . \quad (XXIII.6')$$

Si el método se lleva a cabo como se indica en las ecuaciones $(XXIII.6)$ y $(XXIII.7)$, el número de evaluaciones funcionales escalares se reduce de (n^2+n) a n, pero el método aún requiere de $O(n^3)$ cálculos para resolver el sistema lineal $n \times n$ asociado $A_i\ \mathbf{y}_i = -\mathbf{F}(\mathbf{x}^{(i)})$. Y entonces no tendría justificación emplear el método en esta forma debido a la reducción de la convergencia cuadrática del método de Newton a la convergencia superlineal. Sin embargo, se puede incorporar una mejoría considerable empleando la fórmula de inversión matricial de Sherman y Morrison. Este resultado dice que si A es una matriz no singular y $\mathbf{x}$ e $\mathbf{y}$ son vectores, entonces $A + \mathbf{x}\ \mathbf{y}^t$ es no singular siempre que $\mathbf{y}^t\ A^{-1}\ \mathbf{x} \neq -1$, y

$$(A + \mathbf{x}\ \mathbf{y}^t)^{-1} = A^{-1} - \frac{A^{-1}\ \mathbf{x}\ \mathbf{y}^t\ A^{-1}}{1 + \mathbf{y}^t\ A^{-1}\ \mathbf{x}}\ . \quad (XXIII.8)$$

Esta fórmula permite calcular A_i^{-1} directamente de A_{i-1}^{-1}, eliminando la necesidad de invertir una matriz en cada iteración. Tomando $A = A_{i-1}$, $\mathbf{x} = \frac{(\mathbf{y}_i - A_{i-1}\ \mathbf{s}_i)}{||\mathbf{s}_i||_2^2}$, e $\mathbf{y} = \mathbf{s}_i$, las fórmulas $(XXIII.6')$ junto con la $(XXIII.8)$ implican que

$$A_i^{-1} = \left(A_{i-1} + \frac{(\mathbf{y}_i - A_{i-1}\ \mathbf{s}_i)}{||\mathbf{s}_i||_2^2}\ \mathbf{s}_i^t\right)^{-1} = A_{i-1}^{-1} - \frac{A_{i-1}^{-1}\left(\frac{(\mathbf{y}_i - A_{i-1}\ \mathbf{s}_i)}{||\mathbf{s}_i||_2^2}\ \mathbf{s}_i^t\right)\ A_{i-1}^{-1}}{1 + \mathbf{s}_i^t\ A_{i-1}^{-1}\left(\frac{(\mathbf{y}_i - A_{i-1}\ \mathbf{s}_i)}{||\mathbf{s}_i||_2^2}\right)} =$$

$$= A_{i-1}^{-1} - \frac{\left(A_{i-1}^{-1}\ \mathbf{y}_i - \mathbf{s}_i\right)\mathbf{s}_i^t A_{i-1}^{-1}}{||\mathbf{s}_i||_2^2 + \mathbf{s}_i^t\ A_{i-1}^{-1}\ \mathbf{y}_i - ||\mathbf{s}_i||_2^2} = A_{i-1}^{-1} + \frac{\left(\mathbf{s}_i - A_{i-1}^{-1}\ \mathbf{y}_i\right)\mathbf{s}_i^t A_{i-1}^{-1}}{\mathbf{s}_i^t\ A_{i-1}^{-1}\ \mathbf{y}_i}\ .$$

Esta fórmula implica sólo multiplicación de matrices en cada paso y requiere de $O(n^2)$ cálculos aritméticos.

Algoritmo de Broyden para sistemas no lineales.

==========

Para aproximar una solución del sistema no lineal $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, dada una aproximación inicial $\mathbf{x}$:

Entrada: número n de ecuaciones e incógnitas; aproximación inicial $\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_n)^t$; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada $\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_n)^t$ o mensaje de que el número de iteraciones fue excedido.

Paso 1: tomar $A_0 = J(\mathbf{x})$ donde $J(\mathbf{x})_{ij} = \frac{\partial f_i(\mathbf{x})}{\partial x_j}$ para $1 \leq i, j \leq n$;

$$\mathbf{v} = \mathbf{F}(\mathbf{x})\ ; \qquad (Nota: \quad \mathbf{v} = \mathbf{F}(\mathbf{x}^{(0)})\ ;)$$

Paso 2: tomar $A = A^{-1}$;

Paso 3: tomar $k = 1$;

$$\mathbf{s} = -A\ \mathbf{v}\ ; \qquad (Nota: \quad \mathbf{s} = \mathbf{s}_1\ ;)$$

$$\mathbf{x} = \mathbf{x} + \mathbf{s}\ ; \qquad (Nota: \quad \mathbf{x} = \mathbf{x}^{(1)}\ ;)$$

Paso 4: mientras que $k \leq N_0$ seguir pasos 5–13;

Paso 5: tomar

$$\mathbf{w} = \mathbf{v}\ ; \qquad (Guardar \quad \mathbf{v}\ ;)$$

$$\mathbf{v} = \mathbf{F}(\mathbf{x})\ ; \qquad (Nota: \quad \mathbf{v} = \mathbf{F}(\mathbf{x}^{(k)})\ ;)$$

$$\mathbf{y} = \mathbf{v} - \mathbf{w}\ ; \qquad (Nota: \quad \mathbf{y} = \mathbf{y}^{(k)}\ ;)$$

Paso 6: tomar

$$\mathbf{z} = -A\ \mathbf{y}\ ; \qquad (Nota: \quad \mathbf{z} = -A_{k-1}^{-1}\ \mathbf{y}_k\ ;)$$

Paso 7: tomar

$$p = -\mathbf{s}^t\ \mathbf{z}\ ; \qquad (Nota: \quad p = \mathbf{s}_k^t\ A_{k-1}^{-1}\ \mathbf{y}_k\ ;)$$

Paso 8: tomar

$$C = pI + (\mathbf{s} + \mathbf{z})\ \mathbf{s}^t\ ; \qquad (Nota: \quad C = \mathbf{s}_k^t\ A_{k-1}^{-1}\ \mathbf{y}_k\ I + (\mathbf{s}_k + A_{k-1}^{-1}\ \mathbf{y}_k)\ \mathbf{s}_k^t\ ;)$$

Paso 9: tomar

$$A = C\ A/p\ ; \qquad (Nota: \quad A = A_k^{-1}\ ;)$$

Paso 10: tomar

$$\mathbf{s} = -A\ \mathbf{v}\ ; \qquad (Nota: \quad \mathbf{s} = -A_k^{-1}\ \mathbf{F}(\mathbf{x}^{(k)})\ ;)$$

Paso 11: tomar

$$\mathbf{x} = \mathbf{x} + \mathbf{s}\ ; \qquad (Nota: \quad \mathbf{x} = \mathbf{x}^{(k)}\ ;)$$

Paso 12: si $||\mathbf{s}|| < TOL$ entonces SALIDA ($\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$);
(procedimiento completado satisfactoriamente) PARAR;

Paso 13: tomar $k = k + 1$;

Paso 14: SALIDA ($'$*Número máximo* N_0 *de iteraciones excedido,* $N_0 =\ '$, N_0);
PARAR.

==========

Veamos ahora una aplicación del método de Broyden, reconsiderando el sistema no lineal del ejemplo 1 del Capítulo XXXI, que se resolvió entonces por el método de Newton.

Ejemplo 1.

Consideremos el sistema no lineal dado por

$$\begin{aligned} 3\ x_1 - \cos(x_2 x_3) - \frac{1}{2} &= 0\ , \\ x_1^2 - 81\ (x_2 + 0.1)^2 + \sin x_3 + 1.06 &= 0\ , \\ e^{-x_1 x_2} + 20\ x_3 + \frac{10\pi - 3}{3} &= 0\ . \end{aligned}$$

La matriz Jacobiana para este sistema es

$$J(x_1, x_2, x_3) = \begin{pmatrix} 3 & x_3\ \sin(x_2 x_3) & x_2\ \sin(x_2 x_3) \\ 2\ x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2\ e^{-x_1 x_2} & -x_1\ e^{-x_1 x_2} & 20 \end{pmatrix}\ .$$

Con $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$, tenemos $\mathbf{F}(\mathbf{x}^{(0)}) = (-1.199949, -2.269832, 8.462026)^t$.

$$A_0 = J(\mathbf{x}^{(0)}) = \begin{pmatrix} 3 & 9.999836 \times 10^{-4} & -9.999836 \times 10^{-4} \\ 0.2 & -323.9999 & 0.9950041 \\ -9.900498 \times 10^{-2} & -9.900498 \times 10^{-2} & 20 \end{pmatrix}\ ,$$

$$A_0^{-1} = J(\mathbf{x}^{(0)})^{-1} = \begin{pmatrix} 0.3333331 & 1.023852 \times 10^{-5} & 1.615703 \times 10^{-5} \\ 2.108606 \times 10^{-3} & -3.086882 \times 10^{-2} & 1.535838 \times 10^{-3} \\ 1.660522 \times 10^{-3} & -1.527579 \times 10^{-4} & 5.000774 \times 10^{-2} \end{pmatrix}\ ,$$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - A_0^{-1}\ \mathbf{F}(\mathbf{x}^{(0)}) = (0.4998693, 1.946693 \times 10^{-2}, -0.5215209)^t\ ,$$

$$\mathbf{F}(\mathbf{x}^{(1)}) = (-3.404021 \times 10^{-4}, -0.3443899, 3.18737 \times 10^{-2})^t\ ,$$

$$\mathbf{y}^{(1)} = \mathbf{F}(\mathbf{x}^{(1)}) - \mathbf{F}(\mathbf{x}^{(0)}) = (1.199608, 1.925442, -8.430152)^t\ ,$$

$$\mathbf{s}_1 = (0.3998693, -8.053307 \times 10^{-2}, -0.4215209)^t\ ,$$

$$\mathbf{s}_1^t\ A_0^{-1}\ \mathbf{y}_1 = 0.3424604\ ,$$

$$A_1^{-1} = A_0^{-1} + \frac{1.0}{0.3424604}\ [(\mathbf{s}_1 - A_0^{-1}\ \mathbf{y}_1)\ s_1^t\ A_0^{-1}] =$$

$$= \begin{pmatrix} 0.3333781 & 1.11077 \times 10^{-5} & 8.944584 \times 10^{-6} \\ -2.021271 \times 10^{-3} & -3.094847 \times 10^{-2} & 2.196909 \times 10^{-3} \\ 1.022381 \times 10^{-3} & -1.650679 \times 10^{-4} & 5.010987 \times 10^{-2} \end{pmatrix}\ ,$$

y

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - A_1^{-1}\ \mathbf{F}(\mathbf{x}^{(1)}) = (0.4999863, 8.737888 \times 10^{-3}, -0.5231746)^t\ .$$

Los resultados de las otras iteraciones se muestran en la tabla siguiente:

Tabla 1

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|\|_\infty$
3	0.5000066	8.672215×10^{-4}	-0.5236918	7.88×10^{-3}
4	0.5000005	6.087473×10^{-5}	-0.5235954	8.12×10^{-4}
5	0.5000002	-1.445223×10^{-6}	-0.5235989	6.24×10^{-5}

EJERCICIOS.

1. Usar el método de Newton modificado para aproximar las soluciones a los sistemas no lineales dados en los dos capítulos anteriores. Iterar hasta que $||\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}||_\infty < 10^{-5}$. Comparar el número de iteraciones requerido para esta precisión con el número de iteraciones requerido por la aproximación correspondiente del método de Newton calculada en los ejercicios del Capítulo XXII.

2. Usar el algoritmo de Broyden para aproximar las soluciones a los sistemas no lineales dados en los dos capítulos anteriores. Iterar hasta que $||\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}||_\infty < 10^{-5}$. Comparar el número de iteraciones requerido para esta precisión con el número de iteraciones requerido por la aproximación correspondiente del método de Newton calculada en los ejercicios del Capítulo XXII.

CAPITULO XXIV. TECNICAS DE DESCENSO MAS RAPIDO

1. INTRODUCCION Y METODO

La ventaja de los métodos de Newton y de los cuasi-Newton para la solución de sistemas no lineales de ecuaciones es su rapidez de convergencia, una vez que se conoce una aproximación suficientemente precisa. La debilidad de estos métodos es que frecuentemente es necesario tener una aproximación inicial precisa para asegurar la convergencia. El método considerado en este capítulo, conocido como **método del descenso más rápido**, convergerá a la solución generalmente sólo en forma lineal, pero se puede utilizar para encontrar una aproximación inicial suficientemente precisa como para ser usada en las técnicas de Newton, de la misma manera que el método de bisección se usa para una sola ecuación.

El método de descenso más rápido determina un mínimo local para una función multivariada (de valores múltiples) de la forma $G : \mathcal{R}^n \to \mathcal{R}$. La conexión entre la minimización de una función de $\mathcal{R}^n$ a $\mathcal{R}$ y la solución de un sistema de ecuaciones no lineales se debe al hecho de que un sistema de la forma

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \ , \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \ , \\ \cdots\cdots\cdots \qquad & \ldots \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \ , \end{aligned}$$

tiene una solución en $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ precisamente cuando la función G definida como

$$G(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} [f_i(x_1, x_2, \ldots, x_n)]^2$$

tiene el valor mínimo cero.

Antes de describir el método de descenso más rápido, necesitamos revisar algunos resultados del cálculo.

Definición. Si es $G : \mathcal{R}^n \to \mathcal{R}$, el **gradiente** de G en $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ se denota por $\nabla G(\mathbf{x})$ y se define como

$$\nabla G(\mathbf{x}) = \Big(\frac{\partial G}{\partial x_1}(\mathbf{x}), \frac{\partial G}{\partial x_2}(\mathbf{x}), \ldots, \frac{\partial G}{\partial x_n}(\mathbf{x}) \Big)^t \ .$$

El gradiente de una función multivariada es análogo a la derivada de una función de una sola variable en el sentido de que una función multivariada diferenciable puede tener un mínimo en $\mathbf{x}$ sólo cuando el gradiente es cero.

El gradiente tiene otra propiedad importante relacionada con la minimización de funciones multivariadas.

Definición. Supongamos que $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$ es un vector unitario en $\mathcal{R}^n$ ($||\mathbf{v}||_2^2 = 1$). La **derivada direccional** de G en la dirección de $\mathbf{v}$ se define por

$$D_{\mathbf{v}}G(\mathbf{x}) = \lim_{h \to 0} \frac{G(\mathbf{x} + h\ \mathbf{v}) - G(\mathbf{x})}{h}\ .$$

La derivada direccional de G en $\mathbf{x}$ en la dirección $\mathbf{v}$ mide el cambio en el valor de la función G relativo al cambio de la variable en la dirección de $\mathbf{v}$.

Un resultado estándar del cálculo de funciones multivariadas establece que la dirección que produce el valor máximo de la derivada direccional se presenta cuando se escoge $\mathbf{v}$ de manera que sea paralelo a $\nabla G(\mathbf{x})$ siempre y cuando $\nabla G(\mathbf{x}) \neq 0$. Como consecuencia, la dirección en la que el valor de G en $\mathbf{x}$ decrece más es aquélla dada por $-\nabla G(\mathbf{x})$.

Volvamos ahora al método de descenso más rápido.

El método del descenso más rápido para encontrar un mínimo local de una función G de $\mathcal{R}^n$ a $\mathcal{R}$ puede describirse intuitivamente de la manera siguiente:

(i) evaluar G en una aproximación inicial $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^t$;

(ii) encontrar una dirección a partir de $\mathbf{x}^{(0)}$ que resulte en un descenso en el valor de G;

(iii) decidir cuánto debe moverse en esta dirección y llamar a este nuevo valor $\mathbf{x}^{(1)}$;

(iv) repetir los pasos (i) a (iii) reemplezando a $\mathbf{x}^{(0)}$ por $\mathbf{x}^{(1)}$.

Como la dirección en la que el valor de G en $\mathbf{x}$ decrece más es aquélla dada por $-\nabla G(\mathbf{x})$, entonces, una elección apropiada para $\mathbf{x}$ es

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha\ \nabla G(\mathbf{x}^{(0)}) \qquad (XXIV.1)$$

para alguna constante $\alpha > 0$.

El problema se reduce ahora a escoger α tal que $G(\mathbf{x}^{(1)})$ sea significativamente menor que $G(\mathbf{x}^{(0)})$. Para encontrar cómo eligir el valor de α de una forma apropiada, consideramos la función de una variable

$$h(\alpha) = G(\mathbf{x}^{(0)} - \alpha\ \nabla G(\mathbf{x}^{(0)}))\ .$$

El valor de α que minimiza a h es el valor que se necesita en la ecuación $(XXIV.1)$.

Para encontrar el valor mínimo de h directamente se necesitaría derivar h y encontrar los puntos críticos; éste es un procedimiento muy costoso generalmente. En su lugar podemos empezar con tres aproximaciones no negativas α_1, α_2 y α_3 a α, el número que produce el valor mínimo de h. Después encontramos el polinomio cuadrático que interpola a h en α_1, α_2 y α_3. Se define α como el número que minimiza a este polinomio cuadrático. Se usa ese valor para determinar la nueva iteración para aproximar el valor mínimo de G

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha\ \nabla G(\mathbf{x}^{(0)})\ .$$

2. ALGORITMO Y EJEMPLOS

El siguiente algoritmo usa el método del descenso más rápido para encontrar una aproximación al valor mínimo de $G(\mathbf{x})$. Para emplear este método en la aproximación de la solución al sistema

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \ , \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \ , \\ \cdots\cdots\cdots & \quad \ldots \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \ , \end{aligned}$$

simplemente se reemplaza la función G por $\sum_{i=1}^{n} [f_i(x_1, x_2, \ldots, x_n)]^2$.

Algoritmo del descenso más rápido.

==========

Para aproximar una solución $\mathbf{p}$ al problema de minimización $G(\mathbf{p}) = \min_{\mathbf{x} \in \mathcal{R}^n} G(\mathbf{x})$, dada una aproximación inicial $\mathbf{x}$.

Entrada: número n de variables; aproximación inicial $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$; tolerancia TOL; número máximo de iteraciones N_0;

Salida: solución aproximada $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ o mensaje de fracaso.

Paso 1: tomar $k = 0$;

Paso 2: mientras que $k \leq N_0$ seguir pasos 3–15;

Paso 3: sean

$$g_1 = G(x_1, x_2, \ldots, x_n) \ ; \qquad (Nota: \quad g_1 = G(\mathbf{x}^{(k)}) \ ;)$$

$$\mathbf{z} = -\nabla G(x_1, x_2, \ldots, x_n) \ ; \qquad (Nota: \quad \mathbf{z} = -\nabla G(\mathbf{x}^{(k)}) \ ;)$$

$$z_0 = ||\mathbf{z}||_2 \ ;$$

Paso 4: si $z_0 = 0$ entonces SALIDA ('GRADIENTE CERO');
SALIDA $(x_1, x_2, \ldots, x_n, g_1)$;
(procedimiento completado, puede tener un mínimo; verificar más aún) PARAR;

Paso 5: tomar

$$\mathbf{z} = \mathbf{z}/z_0 \ ; \qquad (Normalizar \ \ \mathbf{z} \ ;)$$

$$\alpha_1 = 0, \quad \alpha_3 = 1, \quad g_3 = G(\mathbf{x} + \alpha_3 \ \mathbf{z}) \ ;$$

Paso 6: mientras que $|g_3| \geq |g_1|$ seguir pasos 7 y 8;

Paso 7: tomar

$$\alpha_3 = \alpha_3/2, \quad g_3 = G(\mathbf{x} + \alpha_3 \ \mathbf{z}) \ ;$$

Paso 8: si $\alpha_3 < TOL \ / \ 2$ entonces SALIDA ('NO HUBO MEJORA');
SALIDA $(x_1, x_2, \ldots, x_n, g_1)$; *(procedimiento completado, puede tener un mínimo; verificar más aún)* PARAR;

Paso 9: tomar

$$\alpha_2 = \alpha_3/2, \quad g_2 = G(\mathbf{x} + \alpha_2 \ \mathbf{z}) \ ;$$

Paso 10: tomar

$$h_1 = (g_2 - g_1)/(\alpha_2 - \alpha_1) \ ;$$
$$h_2 = (g_3 - g_2)/(\alpha_3 - \alpha_2) \ ;$$
$$h_3 = (h_2 - h_1)/(\alpha_3 - \alpha_1) \ ;$$

(*Nota : el polinomio cuadrático* $P(\alpha) = g_1 + h_1\ \alpha + h_3\ \alpha\ (\alpha - \alpha_2)$ *interpola a* $h(\alpha)$ *en* $\alpha = 0, \alpha = \alpha_2, \alpha = \alpha_3$)

Paso 11: tomar (*Nota : en* α_0 *hay un punto crítico de* P)

$$\alpha_0 = 0.5\ (\alpha_2 - h_1/h_3) \ ;$$

$$g_0 = G(\mathbf{x} + \alpha_0\ \mathbf{z}) \ ;$$

Paso 12: encontrar α de $\{\alpha_0, \alpha_1, \alpha_2, \alpha_3\}$ tal que

$$g = G(\mathbf{x} + \alpha\ \mathbf{z}) = \min\{g_0, g_1, g_2, g_3\} \ ;$$

Paso 13: tomar

$$\mathbf{x} = \mathbf{x} + \alpha\ \mathbf{z} \ ;$$

Paso 14: si $|g - g_1| < TOL$ entonces SALIDA $((x_1, x_2, \ldots, x_n, g)$; *(procedimiento completado satisfactoriamente)* PARAR;

Paso 15: tomar $k = k + 1$;

Paso 16: SALIDA ($'$*Número máximo* N_0 *de iteraciones excedido,* $N_0 =\ '$, N_0); PARAR.

==========

Ejemplo 1.

Para encontrar una aproximación inicial razonable a la solución del sistema no lineal

$$f_1(x_1, x_2, x_3) = 3\ x_1 - \cos(x_2 x_3) - \frac{1}{2} = 0 \ ,$$
$$f_2(x_1, x_2, x_3) = x_1^2 - 81\ (x_2 + 0.1)^2 + \sin x_3 + 1.06 = 0 \ ,$$
$$f_3(x_1, x_2, x_3) = e^{-x_1 x_2} + 20\ x_3 + \frac{10\pi - 3}{3} = 0 \ ,$$

usamos el algoritmo del descenso más rápido con $TOL = 0.005$, $N_0 = 10$ y $\mathbf{x}^{(0)} = (0.5, 0.5, 0.5)^t$.

Sea $G(x_1, x_2, x_3) = [f_1(x_1, x_2, x_3)]^2 + [f_2(x_1, x_2, x_3)]^2 + [f_3(x_1, x_2, x_3)]^2$; entonces

$$\begin{aligned}
\nabla G(x_1, x_2, x_3) = \nabla G(\mathbf{x}) = \\
&= \Big(2\ f_1(\mathbf{x})\ \frac{\partial f_1}{\partial x_1}(\mathbf{x}) + 2\ f_2(\mathbf{x})\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) + 2\ f_3(\mathbf{x})\ \frac{\partial f_3}{\partial x_1}(\mathbf{x}), \\
&\quad 2\ f_1(\mathbf{x})\ \frac{\partial f_1}{\partial x_2}(\mathbf{x}) + 2\ f_2(\mathbf{x})\ \frac{\partial f_2}{\partial x_2}(\mathbf{x}) + 2\ f_3(\mathbf{x})\ \frac{\partial f_3}{\partial x_2}(\mathbf{x}), \\
&\quad 2\ f_1(\mathbf{x})\ \frac{\partial f_1}{\partial x_3}(\mathbf{x}) + 2\ f_2(\mathbf{x})\ \frac{\partial f_2}{\partial x_3}(\mathbf{x}) + 2\ f_3(\mathbf{x})\ \frac{\partial f_3}{\partial x_3}(\mathbf{x})\Big)^t \\
&= 2\ J(\mathbf{x})^t\ \mathbf{F}(\mathbf{x}) \ .
\end{aligned}$$

Para $\mathbf{x}^{(0)} = (0.5, 0.5, 0.5)^t$,

$$G(\mathbf{x}^{(0)}) = 1159.24 \ ,$$

$$\mathbf{z} = (-0.0131206, 0.9897569, 0.1421648)^t \ ,$$

$$z_0 = ||\mathbf{z}||_2 = 5359.964 \ ,$$

$$\alpha_1 = 0.00, \quad g_1 = 1159.24,$$

$$\alpha_2 = 0.25, \quad g_2 = 454.8059, \quad h_1 = -2817.738,$$

$$\alpha_3 = 0.50, \quad g_3 = 363.5173, \quad h_2 = -365.1542, \quad h_3 = 4905.167,$$

así que $P(\alpha) = 1159.24 - 2817.173\ \alpha + 4905.167\ \alpha\ (\alpha - 0.25)$. Por lo que,

$$\alpha_0 = 0.4122214, \quad g_0 = 372.2808.$$

Entonces $\alpha = 0.5$ y

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - 0.5\ \mathbf{z} = (0.5065602, 0.005121529, 0.4289176)^t \ ,$$

$$G(\mathbf{x}^{(1)}) = 363.5173 \ .$$

La siguiente tabla contiene los restantes resultados.

Tabla 1

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$G(x_1^{(k)}, x_2^{(k)}, x_3^{(k)})$
0	0.5	0.5	0.5	1159.240
1	0.5065602	0.005121529	0.4289176	363.5173
2	0.5045557	0.06421983	−0.5155808	1.874157
3	0.5068673	0.001808132	−0.5179668	0.01322149
4	0.5067566	0.001208410	−0.5235992	0.0005774418

En el ejemplo 2 del capítulo XXVIII se dio como solución al sistema no lineal $(0.5, 0, -0.5235988)^t$, así que los resultados obtenidos aquí pueden servir como una aproximación inicial adecuada para los métodos de Newton y de Broyden.

EJERCICIOS.

1. Usar el método del descenso más rápido con $TOL = 0.005$ para aproximar las soluciones a los sistemas no lineales dados en los dos capítulos anteriores.

2. Usar el método de descenso más rápido para aproximar los mínimos de las funciones siguientes con una precisión de 0.005:

(a) $G(x_1, x_2) = \cos(x_1 + x_2) + \sin x_1 + \cos x_2$,

(b) $G(x_1, x_2) = 100\ (x_1^2 - x_2)^2 + (1 - x_1)^2$,

(c) $G(x_1, x_2, x_3) = x_1^2 + 2\ x_2^2 + x_3^2 - 2\ x_1\ x_2 + 2\ x_1 - 2.5\ x_2 - x_3 + 2$.

CURSO DE MÉTODOS NUMÉRICOS

SEXTA PARTE

BIBLIOGRAFÍA

BIBLIOGRAFIA BASICA

- R.L. Burden & J.D. Faires: *Análisis Numérico*, Grupo Editorial Iberoamérica, 1985; traducción de *Numerical Analysis*, Prindle, Weber and Schmidt, 1985.
- D. Kincaid & W. Cheney: *Análisis Numérico: las matemáticas del cálculo científico*, Addison-Wesley, 1994; traducción de *Numerical Mathematics and Computing*, Brooks/Cole Publ. Comp., 1991.
- B.P. Demidovich & I.A. Maron: *Cálculo Numérico Fundamental*, Paraninfo, 1988.
- J. Stoer & R. Bulirsch: *Introduction to Numerical Analysis*, Springer-Verlag, 1980.
- K.E. Atkinson: *An Introduction to Numerical Analysis*, John Wiley & Sons, 1978.
- P. Henrici: *Elementos de Análisis Numérico*, Editorial Trillas, 1972; traducción de *Elements of Numerical Analysis*, John Wiley & Sons, 1964.
- A. Ralston: *Introducción al Análisis Numérico*, Editorial Limusa-Wiles, 1970; traducción de *A first course in Numerical Analysis*, McGraw-Hill, 1965.

BIBLIOGRAFIA COMPLEMENTARIA

- A. Aubanell, A. Benseny & A. Delshams: *Útiles básicos de cálculo numérico*, Editorial Labor, 1993.
- A. Andronico, U. Bozzo, A. De Giorgio, I. Galligani, R. Marconi, E. Nervegna, G. Provenzano, M. Refice: *Principi di informatica*, Zanichelli, 1973.
- H.C. Bezner: *FORTRAN 77*, Prentice-Hall, 1989.
- P.M. Birns, P.B. Brown & J.C.C. Muster: *UNIX for People*, Prentice-Hall, 1985.
- E.K. Blum: *Numerical Analysis and Computation: Theory and Practise*, Addison-Wesley, 1972.
- M.P. Cherkasova: *Problems on Numerical Methods*, Wolters-Noordhoff Publ., 1972.
- C. Conde Lazaro & G. Winter Althaus: *Métodos y algoritmos básicos del álgebra numérica*, Editorial Reverté, 1990.
- G.B. Davis & T.R. Hoffmann: *FORTRAN: a structured disciplined style*, McGraw-Hill, 1978.
- G. Dahlquist & Å. Björck: *Numerical Methods*, Prentice-Hall, 1974.
- J.E. Dennis & R.B. Schnabel: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, 1983.
- D.K. Faddeev & V.M. Faddeeva: *Computational Methods of Linear Algebra*, W.H. Freeman & Comp., 1963.
- A.I. Forsythe, T.A. Keenan, E.I. Organick & W. Stenberg: *Lenguajes de diagramas de flujo*, Editorial Limusa-Wiley, 1973; traducción de *Computer Science: a first course*, John Wiley & Sons, 1969.

- G.E. Forsythe & C.B. Moler: *Computer solution of Linear Algebraic Systems*, Prentice-Hall, 1967.
- L. Fox: *An Introduction to Numerical Linear Algebra*, Clarendon Press, 1964.
- L. Fox & D.F. Mayers: *Computing methods for Scientists and Engineers*, Clarendon Press, 1968.
- C.E. Fröberg: *Introduction to Numerical Analysis*, Addison-Wesley, 1970.
- F. Garcia Merayo & A. Nevot Luna: *Análisis Numérico*, Editorial Paraninfo, 1992.
- C.F. Gerald & P.O. Wheatley: *Applied Numerical Analysis*, Addison-Wesley, 1984.
- P.E. Gill, W. Murray & M.H. Wright: *Numerical Linear Algebra and Optimization*, Addison-Wesley, 1991.
- H.H. Goldstine: *A Hystory of Numerical Analysis*, Springer-Verlag, 1977.
- G.H. Golub & C.F. Van Loan: *Matrix Computations*, North Oxford Academic, 1983.
- A.R. Gourlay & G.A. Watson: *Computational Methods for Matrix Eigenproblems*, John Wiley & Sons, 1973.
- W. Hackbusch: *Iterative Solution of Large Sparse systems of equations*, Springer-Verlag, Inc., 1994.
- V. Hernández, E. Ramos, R. Vélez & I. Yáñez: *Matemáticas Básicas*, Universidad Nacional de Educación a Distancia, 1992.
- J.D. Hoffman: *Numerical Methods for Engineers and Scientists*, McGraw-Hill, 1992.
- A.S. Householder: *The theory of Matrices in Numerical Analysis*, Blaisdell, 1965.
- A.S. Householder: *The Numerical Treatment of a single Nonlinear Equation*, McGraw-Hill, 1970.
- E. Isaacson & H.B. Keller: *Analysis of Numerical Methods*, John Wiley & Sons, 1966.
- J.A. Jacquez: *A First Course in Computing and Numerical Methods*, Addison-Wesley, 1970.
- I. Jacques & C. Judd: *Numerical Analysis*, Chapman and Hall Ltd., 1987.
- L.W. Johnson & R.D. Riess: *Numerical Analysis*, Addison-Wesley, 1982.
- E.B. Koffman & F.L. Friedman: *Problem solving and structured programming in Fortran 77*, Addison-Wesley, 1990.
- C. Lawson & R. Hanson: *Solving Least Squares Problems*, Prentice-Hall, 1974.
- S.J. Leon: *Linear Algebra with applications*, MacMillan, 1986.
- P. Lévy: *La invención del ordenador*, en *Historia de la Ciencias*, dirigida por M. Serres, Cátedra, 1991.
- M. El Lozy: *Editing in a Unix environment: the vi/ex Editor*, Prentice-Hall, 1985.

- M. Loukides: *UNIX for FORTRAN programmers*, O'Reilly & Associates, 1990.

- J.H. Mathews: *Numerical Methods for Computer Science, Engineering and Mathematics*, Prentice-Hall, 1987.

- J.M. McCormick & M.G. Salvadori: *Numerical Methods in Fortran*, Prentice-Hall, 1964.

- S.S. McNeary: *Introduction to computational methods for students of calculus*, Prentice-Hall, 1973.

- F. Michavila & L. Gavete: *Programación y cálculo numérico*, Edit. Reverté, 1992.

- D.G. Moursund & C.S. Duris: *Elementary Theory and Application of Numerical Analysis*, Dover Pubb., Inc., 1967.

- J.C. Nash: *Compact Numerical Methods for computers: linear algebra and function minimisation*, Adam Hilger, 1990.

- B. Noble & J.W. Daniel: *Applied Linear Algebra*, Prentice-Hall, 1977.

- T.R.F. Nonweiler: *Computational Mathematics: an introduction to numerical aproximations*, Ellis Horwood Lim. (John Wiley & Sons), 1984.

- J.P. Nougier: *Méthodes de calcul numérique*, Masson, 1993.

- J.M. Ortega: *Numerical Analysis*, Academic Press, 1972.

- J.M. Ortega & W.C. Rheinboldt: *Iterative solutions of Nonlinear Equations in several variables*, Academic Press, 1970.

- G.M. Phillips & P.J. Taylor: *Theory and Applications of Numerical Analysis*, Academic Press, 1973.

- G.W. Stewart: *Introduction to Matrix Computations*, Academic Press, 1973.

- E. Stiefel: *Introducción a la Matemática Numérica*, Editorial Labor, 1966.

- G. Strang: *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, 1986.

- A.N. Tijonov & D.P. Kostomárov: *Algo acerca de la Matemática Aplicada*, Editorial MIR, 1979.

- J.C. Vaissiére & J.P. Nougier: *Programmes et exercices sur les méthodes numériques*, Masson, 1991.

- R. Wait: *The Numerical solutions of Algebraic Equations*, John Wiley & Sons, 1979.

- J.H. Wilkinson: *The Algebraic Eigenvalue Problem*, Clarendon Press, 1965.

- J.H. Wilkinson & C. Reinsch (Eds.): *Handbook for Automatic Computations. Vol.2: Linear Algebra*, 1971.

- D.M. Young & R.T. Gregory: *A survey of Numerical Mathematics*, Vol I–II, Addison-Wesley, 1972-1973.