

Prologo a la segunda edición

Más de 10 años después de escrita la primera edición, de ellos seis como texto en varias universidades del país, pensamos que ha llegado el momento de mejorar algunas insuficiencias de la obra original. Durante ese tiempo hemos recogido suficientes experiencias propias y de muchos profesores, que pensamos redundará en un libro de mejor calidad.

Entre los cambios introducidos cabe destacar: Todas las fórmulas han sido re-escritas con una tipografía más adecuada, gracias a los avances de los software de edición de ecuaciones, la cantidad de ejercicios propuestos se ha incrementado considerablemente y de hecho casi todas las secciones cuentan con una colección de mas de 10 ejercicios, en particular muchos sobre algoritmos y modelación, se ha mejorado la calidad de figuras y tablas, cada capítulo cuenta con una lista de objetivos, una sección de otras lecturas recomendadas para aquellos lectores que desean profundizar o ampliar sobre algún tema, un resumen de las principales ideas del capítulo y un autoexamen, todo lo cual pensamos que puede hacer mas eficiente el proceso de aprendizaje.

Se han introducido nuevos temas que han sido sugeridos por varios lectores. Algunos, muy pocos, han sido simplificados o eliminados. Los principales cambios en cada capítulo son:

Capítulo 1: La notación empleada en el tema de errores se modificó para hacerla más clara y sencilla, se introdujeron los conceptos de problemas estables e inestables y de métodos estables e inestables, se amplió el seudocódigo incluyendo un símbolo para la asignación y la instrucción repeat – until, con lo cual los algoritmos quedan ahora mejor presentados.

Capítulo 2: Se incluyó el tema de solución numérica de sistemas de ecuaciones no lineales mediante el método de Newton – Raphson y la determinación de raíces complejas de ecuaciones algebraicas por el algoritmo de Newton – Bairstow.

Capítulo 3: Se introdujo el uso de las normas matriciales y vectoriales para hacer más completo el tratamiento de la estabilidad y de los métodos iterativos, se incluyó el tratamiento de los sistemas mal condicionados, el cálculo de determinantes, la inversión de matrices y la determinación de valores y vectores propios.

Capítulo 4: Se eliminó el método interpolación por diferencias finitas, poco ventajoso para el cálculo computacional, se amplió la sección sobre interpolación spline, incluyendo algoritmos para los spline cúbicos anclados, periódicos y paramétricos, además, se introdujeron cambios importantes en el tratamiento del ajuste de modelos no lineales.

Capítulo 5: El método de integración de los rectángulos, de poco valor práctico, fue eliminado, se ha dado un tratamiento mas completo al método de Gauss – Legendre (incluyendo una forma de acotar el error) y se incorporó una sección dedicada al cálculo numérico de integrales dobles mediante los métodos de Simpson y de Gauss.

Capítulo 6: Se dedica exclusivamente al tema de optimización numérica, que no fue tratado en la primera edición; en el se incluyen los algoritmos de optimización de funciones de una y de varias variables mas importantes por su valor teórico y didáctico.

Capítulo 7: Se fundió en uno solo los temas tratados en los capítulos seis y siete de la primera edición. Realmente, la temática de ecuaciones diferenciales fue completamente re-elaborada para mejorar la claridad. Se eliminó el tema de diferencias finitas para problemas de contorno que, por

su especificidad y el volumen de conceptos que requiere, hubiera necesitado un espacio demasiado amplio.

Por otra parte, hemos procurado mantener y reforzar el enfoque problemático que caracterizó a la primera edición, el énfasis en el tratamiento algorítmico de todos los métodos y la inclusión de problemas de modelación en todas las listas de ejercicios. Al igual que antes, el libro está concebido para estudiantes capaces de utilizar – o, mejor, confeccionar – programas computacionales de uso personal que le permitan probar, experimentar y comparar los métodos numéricos estudiados.

Los autores

ÍNDICE

CAPÍTULO 1: CONCEPTOS INICIALES	1
Objetivos	1
1.1 Introducción	1
¿Qué es la Matemática Numérica?	1
Una breve historia	2
1.2 Fuentes de error en la solución de un problema	3
La modelación y los errores de modelación	3
Métodos computacionales y errores de truncamiento	4
Errores en el proceso de cálculo	5
Ejercicios	10
1.3 Medidas del error	10
El mínimo error absoluto máximo	13
Relación entre $E_m(x)$ y $e_m(x)$	14
Hallando aproximaciones por defecto y por exceso	15
Ejercicios	16
1.4 Cifras significativas y cifras exactas	17
La notación científica	20
Cifra exacta	20
Contando las cifras exactas	22
El redondeo	23
Cifras decimales exactas y error absoluto	24
Cifras exactas y error relativo	24
Ejercicios	26
1.5 Los números en la computadora	27
Representación de números enteros	27
Representación de números reales	28
Recomendaciones	30
1.6 Propagación del error	30
Una ley general	30
Propagación del error en sumas y diferencias	32
Propagación del error en el producto	34
Propagación del error en el cociente	35
Propagación del error en la potencia y la exponencial	37
Ejercicios	39
1.7 Errores e inestabilidad	40
Problemas estables y problemas inestables	40
Pérdida de significación	42
Métodos inestables para problemas estables	44
Ejercicios	47
1.8 Seudo código para la escritura de algoritmos	48
El operador de asignación	48
La estructura alternativa	48
Estructuras repetitivas	49
Ejercicios	52
Otras lecturas recomendadas	53
Principales ideas del capítulo	53
Auto examen	55

CAPÍTULO 2: RAÍCES DE ECUACIONES	57
Objetivos	57
2.1 Introducción	57
El problema que se resolverá	57
2.2 Separación de raíces	62
Dos etapas	62
Separación gráfica de raíces	62
Algunos resultados importantes para ecuaciones algebraicas	63
Regla de Descartes	64
La fórmula de Lagrange para acotar raíces	64
Análisis de las raíces negativas de una ecuación algebraica	66
Combinando las técnicas	66
Ejercicios	70
2.3 El método de la bisección	72
Hipótesis	72
El método	72
Convergencia del método	73
Condición de terminación	74
Algoritmo en seudo código	75
Comentarios finales	75
Ejercicios	77
2.4 El método Regula Falsi	80
Hipótesis	81
El método	81
Convergencia del método	82
El error del método	84
Rapidez de la convergencia	86
Algoritmo en seudo código	87
Comentarios finales	87
Ejercicios	90
2.5 El método de Newton – Raphson	93
El método iterativo en general	93
El método de Newton – Raphson	97
Interpretación geométrica	98
Convergencia del método de Newton – Raphson	99
El error en el Método de Newton – Raphson	103
Algoritmo en seudo código	106
Comentarios finales	106
Ejercicios	110
2.6 El método de las secantes	113
Convergencia del método de las secantes	115
El error en el método de las secantes	116
Algoritmo en seudo código	117
Comentarios finales	118
Ejercicios	122
2.7 Extensiones del método de Newton – Raphson	125
El método de Newton – Raphson para sistemas de ecuaciones	125
Algoritmo en seudo código	127
El método de Newton – Bairstow	130
Algoritmo en seudo código	135
Ejercicios	136

Otras lecturas recomendadas	139
Principales ideas del capítulo	139
Auto examen	140
CAPÍTULO 3: SISTEMAS DE ECUACIONES LINEALES Y MATRICES	142
Objetivos	142
3.1 Introducción	143
Problemas a resolver	145
Métodos directos y métodos iterativos	146
Normas matriciales y vectoriales	147
3.2 El método de Gauss	149
Proceso directo	150
Proceso inverso	151
Estrategias de pivote	151
Algoritmo en seudo código	155
Cantidad de operaciones del método de Gauss	157
Ejercicios	159
3.3 Consecuencias del método de Gauss	160
El método de Gauss para sistemas tridiagonales	161
Algoritmo en seudo código para sistemas tridiagonales	164
Cálculo de determinantes	165
Algoritmo para calcular determinantes	167
Inversión de matrices mediante el método de Gauss	167
Algoritmo para invertir una matriz mediante el método de Gauss	169
Ejercicios	170
3.4 Sistemas mal condicionados	172
Una medida del mal condicionamiento	174
¿Qué hacer?	177
Ejercicios	178
3.5 Métodos iterativos para sistemas lineales	178
El método de Jacobi	179
Convergencia del método de Jacobi	182
El error en el método de Jacobi	187
Algoritmo del método de Jacobi	189
El método de Seidel	193
Convergencia del método de Seidel	195
Comparación entre la convergencia de los métodos de Jacobi y de Seidel	198
Algoritmo del método de Seidel	200
Comentarios finales sobre los métodos iterativos	203
Ejercicios	206
3.6 Cálculo de valores y vectores propios	207
Sub espacios propios	210
Polinomio característico	211
El caso especial de las matrices simétricas	211
Localización de los valores propios	212
Gráfica del polinomio característico	213
Solución numérica de la ecuación característica	215
Transformación de la matriz en una similar	215
El método de la potencia	215
Ejercicios	218

Otras lecturas recomendadas	220
Principales ideas del capítulo	220
Auto examen	221
CAPÍTULO 4: APROXIMACIÓN DE FUNCIONES	223
Objetivos	223
4.1 Introducción	223
El problema de la aproximación funcional	223
Conceptos básicos	227
4.2 Interpolación polinomial	228
Existencia y unicidad del polinomio interpolador	228
Error del polinomio interpolador	232
4.3 El método de Lagrange	237
Algoritmo en seudo código	239
Ejercicios	243
4.4 El método de Newton	245
Estimación del error de interpolación	248
Relación entre diferencias y derivadas	251
Algoritmo en seudo código	252
Ejercicios	253
4.5 Interpolación mediante splines	256
El problema de la interpolación global	256
Funciones spline	257
El spline cúbico de interpolación	257
El spline natural	261
Algoritmo en seudo código para el spline cúbico natural	263
El spline cúbico anclado	265
Algoritmo en seudo código para el spline cúbico anclado	266
El spline cúbico periódico	267
Algoritmo en seudo código para el spline cúbico periódico	270
Ejercicios	274
4.6 Ajuste de curvas	277
El problema del ajuste de curvas	278
Modelos lineales	279
Algoritmo en seudo código para ajustar modelos lineales	282
Ajuste de modelos no lineales mediante cambios de variables	287
Ajuste numérico de modelos no lineales	290
Ejercicios	291
Otras lecturas recomendadas	295
Principales ideas del capítulo	296
Auto examen	299
CAPÍTULO 5: INTEGRACIÓN NUMÉRICA	300
Objetivos	300
5.1 Introducción	300
5.2 El método de los trapecios	303
Algoritmo en seudo código	306
El error de truncamiento en el método de los trapecios	307
Una fórmula asintótica para el error de truncamiento	310
Estimación del error de truncamiento por doble cálculo	311
El error de redondeo en la fórmula de los trapecios	313

5.3 El método de Simpson	314
Algoritmo en seudo código	318
Error de truncamiento en el método de Simpson	319
Una fórmula asintótica para el error de truncamiento del método de Simpson	321
Estimación del error de truncamiento por doble cálculo	322
Fórmulas de Newton – Cotes	323
Ejercicios	324
5.4 El método de Gauss	327
Los polinomios de Legendre	329
Una base para el espacio P_5	330
Generalización para cualquier m	332
Generalización para cualquier intervalo	332
Algoritmo en seudo código	334
El error en el método de Gauss	335
Ejercicios	336
5.5 El método de Romberg	339
Algoritmo en seudo código	343
Ejercicios	344
5.6 Cálculo numérico de integrales dobles	346
Cálculo de integrales dobles por el método de Gauss	347
Algoritmo en seudo código del método de Gauss para integrales dobles	348
Cálculo de integrales dobles por el método de Simpson	350
Algoritmo en seudo código del método de Simpson para integrales dobles	352
Estimación del error	353
Ejercicios	356
Otras lecturas recomendadas	358
Principales ideas del capítulo	358
Auto examen	361
 CAPÍTULO 6: OPTIMIZACIÓN NUMÉRICA	 364
Objetivos	364
6.1 Introducción	364
Problemas de optimización	364
Funciones unimodales de una variable	366
Propiedad básica de la optimización unidimensional	368
Clasificación de los métodos de búsqueda unidimensional	368
6.2 Optimización unidimensional sin restricciones	369
Búsqueda simultánea	369
Búsqueda secuencial uniforme	369
Algoritmo en seudo código	370
Selección del sentido de la búsqueda	374
Ejercicios	375
6.3 Optimización en un intervalo	376
El método de bisección	376
Algoritmo en seudo código	378
El método de Fibonacci	380
El método de la sección de oro	382
Algoritmo en seudo código del método de la sección de oro	384

Refinamiento del resultado mediante interpolación	386
Ejercicios	388
6.4 Conceptos básicos para la optimización multidimensional	389
Notación y representación gráfica	390
Trayectoria lineal en R^n	391
Función linealmente unimodal	393
Funciones cuadráticas de n variables	394
Matrices positivas definidas y negativas definidas	396
Algoritmo para hallar el óptimo en una dirección	400
6.5 El método de búsqueda por coordenadas	401
Algoritmo en seudo código	402
Ejercicios	406
6.6 El método del gradiente	408
Algoritmo en seudo código	409
Ejercicios	412
6.7 El método de Powell	415
Interpretación geométrica	416
Propiedades fundamentales de las direcciones conjugadas	417
Método de las tangentes paralelas	420
El método de Powell	421
Algoritmo en seudo código	423
Ejercicios	426
6.8 El método del simplex secuencial	428
El simplex	428
El método del simplex para funciones de dos variables	429
Cálculo del simplex inicial	432
Determinación de los vértices de un nuevo simplex	434
Algoritmo en seudo código	436
Ejercicios	442
Otras lecturas recomendadas	444
Principales ideas del capítulo	444
Auto examen	446
 CAPÍTULO 7: ECUACIONES DIFERENCIALES ORDINARIAS	447
Objetivos	447
7.1 Introducción	447
Conceptos iniciales	447
Tipos de solución	448
Condiciones iniciales y de frontera	449
Limitaciones de los métodos analíticos	450
Ejercicios	454
7.2 Ecuaciones diferenciales de primer orden	456
El campo de direcciones	457
Isoclinas	462
La estabilidad de las ecuaciones diferenciales	464
Ecuación estable modelo	467
Ejercicios	467
7.3 Métodos de paso simple	469
Dos tipos de métodos	469
El método de Euler	469
Algoritmo en seudo código	471

Error en el método de Euler	473
Estimación del error por doble cálculo	475
Estabilidad del método de Euler	477
Los métodos de Taylor	479
Error en el método de Taylor	481
Método de Runge - Kutta de orden dos	482
Interpretación geométrica de RK-2	483
Algoritmo en seudo código para RK2	484
Estimación del error en RK2	486
Estabilidad de RK2	487
Método de Runge - Kutta de orden cuatro	489
Algoritmo en seudo código para RK4	489
Ejercicios	491
7.4 Métodos de paso múltiple	494
Comparación entre los métodos de paso simple y paso múltiple	494
Los métodos de Adams – Bashforth	494
Algoritmo en seudo código para AB4	498
Los métodos de Adams – Moulton	499
Métodos predictor – corrector	502
Algoritmo en seudo código para el método predictor – corrector de Adams	505
Estabilidad de los métodos de Adams	506
Ejercicios	509
7.5 Ecuaciones diferenciales con condiciones iniciales	511
Problema de Cauchy de orden m	511
Transformación de una ecuación de orden m en un problema de Cauchy	512
Solución numérica de un problema de Cauchy	517
El método RK2 para un problema de Cauchy de orden m	517
Algoritmo en seudo código	520
Estimación del error por doble cálculo	521
El método RK4 para un problema de Cauchy de orden m	522
Algoritmo en seudo código	525
Estimación del error por doble cálculo	525
Ejercicios	528
7.6 Ecuaciones diferenciales con condiciones de frontera	531
El método de los disparos	532
Ejercicios	539
Otras lecturas recomendadas	540
Principales ideas del capítulo	540
Auto examen	543
Respuestas a los ejercicios	544

CAPÍTULO 1

Matemática Numérica, 2da Edición

Manuel Álvarez, Alfredo Guerra, Rogelio Lau

CONCEPTOS INICIALES

Objetivos

Al finalizar el estudio y la ejercitación de este capítulo, el lector debe ser capaz de:

- Explicar brevemente las diferencias entre los métodos analíticos que ha estudiado hasta ahora y los métodos numéricos que aborda la Matemática Numérica.
- Enumerar las fuentes de error en la solución de un problema y argumentar acerca de la actitud que se debe asumir respecto a cada una de estas causas de errores.
- Utilizar el lenguaje de la teoría de errores y explicar el significado de los términos que se emplean en el mismo: error absoluto, error relativo, error absoluto máximo, error relativo máximo, cifras exactas.
- Transformar la información expresada en términos de cifras exactas a los conceptos de errores y viceversa.
- Explicar brevemente las características de los sistemas numéricos utilizados por una computadora digital, sus causas y las implicaciones prácticas que ellas provocan.
- Utilizar las leyes básicas de la propagación de errores para analizar la exactitud de los resultados obtenidos en algoritmos sencillos que emplean datos numéricos aproximados.
- Explicar el concepto de estabilidad de un problema e ilustrarlo mediante ejemplos sencillos.
- Enumerar y exemplificar las principales causas de inestabilidad en algoritmos computacionales.
- Utilizar los comandos del seudo código que se usará en este libro para comprender y escribir algoritmos numéricos simples.

1.1 Introducción

¿Qué es la Matemática Numérica?

En general, la Matemática está compuesta de diferentes ramas, cada una de las cuales se dedica al estudio de determinado objeto u objetos matemáticos; así, el Análisis Matemático se plantea, como objetivo central, el estudio de las funciones numéricas; el Álgebra Lineal se interesa en el análisis de los espacios vectoriales y las funciones lineales definidas en ellos; la Estadística trata sobre procesos aleatorios, etcétera. La Matemática Numérica, sin embargo, es una rama de la Matemática en la cual el objetivo no es el estudio de un ente matemático en particular; la Matemática Numérica tiene como propósito el desarrollo de métodos para la solución de los más diversos problemas matemáticos mediante una cantidad *finita* de operaciones *numéricas*. Es decir, lo que le da unidad a esta rama de la Matemática, no es el tipo de problema que se ha de resolver sino el método que se aplicará: operaciones *numéricas* en cantidad *finita*.

Está claro que, por regla general, los problemas matemáticos no pueden ser resueltos exactamente de esta manera. Por eso, la Matemática Numérica no se plantea llegar a resultados exactos; ni siquiera a resultados tan exactos como sea posible. El propósito aquí será obtener resultados tan exactos como sea *necesario*. Los métodos de solución que emplea la Matemática Numérica reciben el nombre genérico de métodos *numéricos* y, en contraposición, a los otros métodos matemáticos se les llamará métodos *analíticos*. A lo largo del libro se deducirán métodos numéricos para resolver ecuaciones, para aproximar funciones, para calcular integrales, para

optimizar funciones, para resolver ecuaciones diferenciales, para invertir matrices, etc. Nótese que se trata de problemas que ya antes han sido estudiados y, por tanto, se contará con mucha información acerca de los resultados teóricos fundamentales sobre estos temas; esto permitirá dedicar la atención, fundamentalmente, a encontrar métodos eficientes para resolver los problemas.

El carácter aproximado de los métodos que siguen, suele al principio decepcionar a algunos estudiantes que, en los 27 o 28 semestres de Matemática que ya han cursado desde que comenzaron en la escuela, siempre han admirado la exactitud de los resultados matemáticos. Sin embargo, en las aplicaciones de la Matemática, rara vez se necesitan resultados exactos. Por otra parte, el prescindir de la exactitud absoluta, permite a la Matemática Numérica elaborar métodos mucho más generales que los métodos analíticos exactos; por ejemplo: con un solo método numérico se pueden calcular de manera aproximada todas las integrales definidas vistas en los cursos de Cálculo y otras que se escapan a todos los métodos exactos; además, como se trata de métodos numéricos (solo se requieren operaciones aritméticas, no operaciones algebraicas de tipo simbólico) estos métodos pueden ser fácilmente implementados en una computadora digital. Por todas estas razones, la Matemática Numérica posee en la actualidad una gran importancia.

Una breve historia

Aunque, como ciencia estructurada y rigurosa, la Matemática Numérica es relativamente joven (siglos XIX y XX), desde tiempos muy remotos se emplearon métodos numéricos aproximados. En el papiro de Rhind (el documento matemático más antiguo que se conserva) que data de unos 2000 años a. n. e., fruto del desarrollo de la antigua civilización egipcia, aparecen, entre más de 80 problemas resueltos, métodos aproximados para calcular el volumen de un montón de frutos y el área de una circunferencia, tomándola como la de un cuadrado cuyo lado fuera $8/9$ del diámetro de la circunferencia. En Babilonia (siglos XX al III, a. n. e.) ya se conocían métodos aproximados para calcular raíces cuadradas. De la antigua Grecia, son famosos los trabajos de Arquímedes (siglo III a. n. e.) en la cuadratura del círculo que le permitió, aproximando una circunferencia mediante polígonos inscritos y circunscritos, llegar a la notable aproximación

$$3 + \frac{10}{71} < \pi < 3 + \frac{1}{7}$$

es decir:

$$3,14085 < \pi < 3,14286$$

El método de Arquímedes fue posteriormente aplicado por otros matemáticos y ya en la primera mitad del siglo XV el árabe Kashi había obtenido para π una aproximación de 17 cifras decimales utilizando polígonos de hasta 805 306 368 lados. Un notable ejemplo de cálculos numéricos son las tablas de logaritmos publicadas en 1614 por el holandés Neper en que aparecen, con 8 cifras exactas, los logaritmos de las funciones trigonométricas para ángulos desde 0 hasta 90 grados con paso de un minuto. Gracias al gigantesco trabajo numérico del propio Neper y de otros como el suizo Bürgi, el escocés Briggs y el holandés Vlacq ya en 1628 existían tablas de logaritmos decimales de los números desde 1 a 100 000 calculadas con 10 cifras decimales exactas.

Desde finales del siglo XVII comienza a perfilarse la teoría de las series infinitas, ligadas a matemáticos como el suizo Euler, el alemán Leibniz y los ingleses Newton y Taylor, sin las cuales hubiera sido imposible justificar o deducir muchos de los métodos numéricos que se estudiaran más adelante.

A principios del siglo XVIII se produce otro gran paso con la aparición del Cálculo de Diferencias Finitas (fundado por los ingleses Taylor y Stirling), el cual constituye la base teórica para fundamentar varios métodos numéricos.

El surgimiento y consolidación del Análisis Funcional desde finales del siglo XIX hasta principios del XX, permitió a la Matemática Numérica dar un salto cualitativo al lograrse esclarecer los conceptos básicos de la aproximación funcional.

Con el surgimiento de las computadoras digitales a mediados del siglo XX y su continuo desarrollo, la Matemática Numérica ha recibido un fuerte estímulo, ya que la computadora digital ha hecho posible la aplicación práctica de muchos métodos numéricos, que con el trabajo en forma manual, solo tendrían un valor teórico. Por otra parte, las computadoras digitales han traído la necesidad de desarrollar nuevos métodos numéricos para dar respuesta a nuevos problemas que antes no era posible siquiera imaginar.

1.2 Fuentes de error en la solución de un problema

El hecho de que la Matemática Numérica ponga su atención en métodos aproximados, no significa que en este libro los errores carezcan de importancia; todo lo contrario: un método aproximado solo tiene valor si permite, de alguna forma, tener una estimación de la magnitud del error que se comete con su aplicación. Por esta razón, es necesario dedicar un tiempo a estudiar los diversos tipos de errores que se pueden presentar en la solución de un problema real.

En la figura 1 se muestra esquemáticamente los pasos que suelen seguirse para llegar a la solución de un problema real y los errores que pueden introducirse en los diferentes pasos. Pudiera pensarse que el camino a seguir es tratar de eliminar todas las fuentes posibles de error, pero no es así: algunos tipos de error son inevitables y, como se verá, algunos resultan aconsejables.

La modelación y los errores de modelación

El primer paso en la solución de un problema consiste en pasar de una situación problémica del mundo real a un modelo matemático. Este modelo matemático consiste generalmente en un conjunto de objetos matemáticos relacionados entre sí, tales como ecuaciones diferenciales, integrales, ecuaciones e inecuaciones algebraicas, tablas, esquemas, etc. que intentan reflejar (no copiar) los aspectos *esenciales* del mundo real que constituyen la situación problémica del mundo natural. Este paso suele llamarse *modelación matemática* del problema. Nótese que el modelo no puede, ni debe, reflejar exactamente el mundo real sino sólo los aspectos de aquel que resultan importantes en el problema que se desea resolver, de acuerdo con el uso que se dará a los resultados obtenidos. Realmente, la modelación matemática tiene mucho de arte; si el modelo copia demasiados detalles de la realidad, es probable que el modelo matemático sea tan

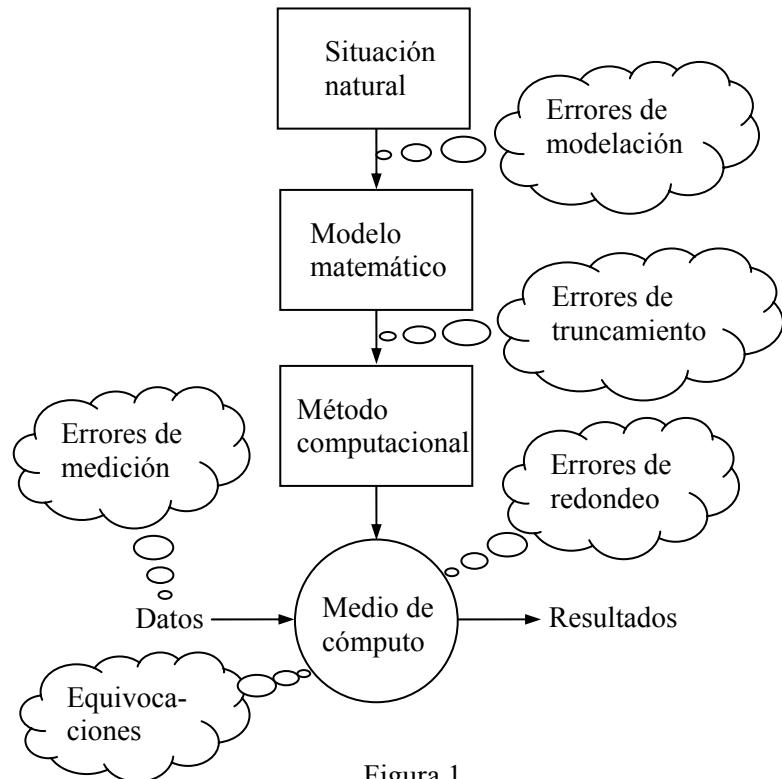


Figura 1

complicado que no ayude a comprender lo esencial del problema, e incluso, que no pueda ser resuelto posteriormente; si se ignoran aspectos importantes del mundo real entonces puede ocurrir que el modelo sea una aproximación demasiado grosera de la realidad y que se pueda llegar a conclusiones absurdas a partir del modelo. El arte consiste en decidir adecuadamente qué aspectos del mundo real deben estar reflejados en el modelo y cuáles no, de manera que los errores de modelación sean aceptables para los objetivos que se persigue. Por ejemplo, para la mayoría de los problemas de mecánica se suele suponer que la aceleración producida por la fuerza de la gravedad es $9,8 \text{ m/seg}^2$, independientemente del lugar de la tierra en que ocurra el fenómeno; en realidad esta aceleración varía desde 9,780 en el ecuador hasta 9,832 en las regiones polares; esta suposición no causa grandes errores en problemas comunes pero en el lanzamiento de cohetes propulsores de satélites artificiales, hay que tener en cuenta la aceleración gravitacional propia de cada lugar de La Tierra por donde vuela el cohete propulsor, pues de otra forma se introducen errores (de modelación) intolerables.

A los efectos de la Matemática Numérica, los errores de modelación suelen clasificarse (junto con los de medición) como errores *inherentes*, en el sentido de que no pueden ser eliminados o disminuidos por el tratamiento matemático del problema, ya que están presentes desde la misma formulación del problema.

Métodos computacionales y errores de truncamiento

La segunda etapa en la solución de un problema es establecer los métodos o algoritmos que se usarán para la solución del modelo matemático planteado. A veces estos métodos son exactos pero, casi siempre esto no es posible o no es práctico. La mayoría de los métodos exactos solamente se aplican a situaciones muy simples y específicas que raras veces se dan en los

problemas reales. Por ejemplo, las ecuaciones algebraicas de grado mayor que 4 solamente se pueden resolver por métodos exactos cuando poseen soluciones enteras o racionales (el método de Rufini) lo cual siempre sucede en los problemas escolares pero casi nunca en la realidad; las ecuaciones no algebraicas (es decir, las trigonométricas, logarítmicas, exponenciales, etc.) solo admiten soluciones por métodos exactos en casos muy triviales; los métodos de integración exactos, basados en hallar una primitiva del integrando, pueden aplicarse a un reducido número de integrales y en problemas tan simples como calcular la longitud de una elipse, fracasan rotundamente.

Por todas estas razones, en una gran cantidad de ocasiones hay que recurrir a métodos no exactos en la solución del modelo matemático obtenido. El error que se introduce en el proceso debido a la no exactitud del método de solución empleado se suele llamar error de *truncamiento*. Esta palabra se utiliza debido a que muchas veces la no exactitud del método utilizado proviene de utilizar en alguna parte de un proceso, solo una cantidad finita de términos de una serie infinita (es decir, de *truncar* una serie). Sin embargo, no es esta la única causa de que un método no sea exacto; a veces el error se produce por sustituir una derivada por un cociente finito de incrementos o una integral por una suma finita de muchos sumandos pequeños o por detener un proceso infinito convergente. A lo largo de este libro serán tratados muchos métodos aproximados y en cada caso se hará el estudio necesario del error de truncamiento cometido, que a veces se llama simplemente, error del método.

Errores en el proceso de cálculo

Una vez que está definido el algoritmo de solución del modelo matemático, se procede a la solución. En la actualidad, la solución se ejecuta, en su mayor parte, mediante calculadora electrónica o mediante una computadora digital con un programa adecuado. En esta etapa del proceso se pueden introducir tres tipos de errores:

- De medición u observación

Estos son los errores contenidos en los datos debido a la imperfección de los instrumentos de medición o los métodos de observación utilizados o a la poca información acerca del problema que se está resolviendo. A veces, los objetivos que se persiguen no justifican utilizar datos de mayor calidad, los cuales pueden ser muy costosos. Por ejemplo, medir una temperatura con un error menor que 0,01 grados centígrados requiere instrumentos sumamente costosos que pocos laboratorios en el mundo poseen en la actualidad y, posiblemente, el resultado que se desea obtener no se afecte grandemente por este pequeño error de medición. Como ya se mencionó, los errores de medición (junto con los de modelación) forman parte de los llamados errores inherentes, dado su carácter externo al procesamiento matemático del modelo.

- Equivocaciones

En el trabajo manual estos son esos frecuentes errores que se introducen, por ejemplo, cuando se dice que “tres por dos es cinco” o cuando se opriime una tecla equivocada en la calculadora. Con el uso de las computadoras las equivocaciones no suelen ocurrir en el momento en que se ejecuta el programa, pero sí pueden estar presentes en el programa elaborado y sus consecuencias pueden a veces pasar inadvertidas durante años.

Cuando el trabajo numérico se realizaba a mano, los algoritmos de cálculo se ejecutaban mediante tablas y en ellas siempre aparecían “columnas de comprobación”, destinadas a

realizar operaciones redundantes solamente con el objetivo de detectar las equivocaciones. Con el uso de las computadoras digitales, el proceso de detección y corrección de equivocaciones (que suele llamarse *debuging* en el argot de los programadores) es una etapa muy importante de la puesta a punto de un programa, pero se escapa a los objetivos de un curso de Matemática Numérica.

- Errores de redondeo

Estos errores se producen cuando se sustituye un número decimal por otro con menos cifras. Más adelante se profundizará en este tipo de errores pero por el momento puede adelantarse que ellos están constantemente presentes tanto si se trabaja a mano, como si se usa una calculadora o una computadora sofisticada. Son debidos a la naturaleza del sistema de numeración que se utiliza, el cual se basa en cifras y no permite representar todos los números reales mediante una cantidad finita de dígitos. Por ejemplo, cuando se utiliza $0,333333$ en lugar de $1/3$ ó $3,1416$ en lugar de π , se comenten errores de redondeo.

A diferencia de las equivocaciones ante las cuales todo lo que se puede hacer es tratar de evitarlas, con los errores de redondeo hay que aprender a convivir; ellos son inevitables y todo lo que se necesita es, por una parte, mantenerlos lo suficientemente pequeños de modo que no afecten significativamente los resultados que se desea obtener y, por otra parte, no intentar hacerlos exageradamente pequeños (por ejemplo, utilizando una cantidad muy grande de cifras decimales) porque ello se traduce en algoritmos innecesariamente lentos.

Para ilustrar los conceptos anteriores, considérese el siguiente ejemplo

Ejemplo 1

Desde un cierto punto del espacio se lanza un pequeño objeto al suelo (por ejemplo, una piedrecilla) y se desea saber qué distancia recorrerá en su trayectoria desde la mano hasta el piso.

Solución:

Primero es necesario modelar matemáticamente el problema. Como es un problema mecánico, hay que recurrir a las leyes de la Mecánica para elaborar un modelo adecuado. En el proceso de modelación habrá que realizar algunas aproximaciones que introducirán errores de modelación. Se tomarán las siguientes hipótesis:

- La partícula lanzada se tratará como si fuera un punto.
- Se considerará que la partícula solamente es atraída por la Tierra y no por la Luna o por otros cuerpos celestes.
- Se tomará el valor de $9,8 \text{ m/seg}^2$ como la aceleración de la gravedad.
- Se ignorará el efecto de la fuerza de empuje del aire sobre la partícula.
- No se tomará en cuenta la fricción entre el aire y la partícula.
- Se supondrá que el aire está en reposo, es decir, que no hay viento.
- Se aproximarán la superficie del suelo como un plano perfectamente horizontal.

Todas estas hipótesis son idealizaciones que permitirán llegar a un modelo matemático suficientemente sencillo, a costa de introducir errores de modelación. Se supone que el error de

modelación introducido es razonablemente pequeño a los efectos del resultado que se desea obtener.

Uno de los primeros pasos en la modelación matemática de un problema, es la definición de un sistema de referencia adecuado. Se tomará el instante en que la piedra es lanzada (en que abandona la mano) como el instante inicial ($t = 0$). En cuanto al sistema de referencia espacial, se tomará un plano coordenado con su eje x colocado en el suelo y el eje y vertical pasando por el punto en que la partícula abandona la mano. El eje x se toma en una dirección tal, que la trayectoria de la partícula se efectúa en plano xy . En la figura 2 se muestra el sistema de referencia, la partícula en su posición inicial (en blanco) y en un instante t posterior (en negro) y, en línea de puntos, la trayectoria que supuestamente seguirá.

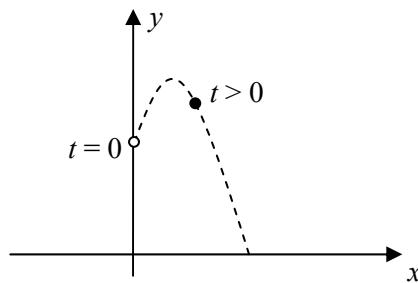


Figura 2

Para obtener el modelo matemático se hará uso de la conocida Segunda Ley de Newton de la Mecánica: “La suma de las fuerzas que actúan sobre una partícula es igual al producto de su masa por su aceleración”. Esta ley establece una igualdad vectorial, que puede expresarse como dos igualdades escalares, tomando las componentes respectivas de la fuerza resultante y de la aceleración:

$$F_x = ma_x \quad (1)$$

$$F_y = ma_y \quad (2)$$

En un instante $t \geq 0$ cualquiera, según la hipótesis iniciales, la única fuerza que actúa sobre la partícula es la debida a la atracción gravitacional, que es una fuerza vertical dirigida hacia abajo y de magnitud mg . De aquí resulta $F_x = 0$ y $F_y = -mg$. Sustituyendo en las ecuaciones (1) y (2):

$$ma_x = 0 \quad (3)$$

$$ma_y = -mg \quad (4)$$

De las ecuaciones (3) y (4) se obtienen de forma inmediata:

$$a_x = 0 \quad (5)$$

$$a_y = -g \quad (6)$$

Como la aceleración es la segunda derivada del desplazamiento respecto al tiempo, las ecuaciones (5) y (6) se traducen en:

$$\frac{d^2x}{dt^2} = 0 \quad (7)$$

$$\frac{d^2y}{dt^2} = -g \quad (8)$$

donde $t \geq 0$.

Para completar el modelo matemático hay que añadir a las ecuaciones (7) y (8) las condiciones iniciales del problema, algunas de las cuales son consecuencia del sistema de referencia definido:

$$x(0) = 0 \quad (9)$$

$$y(0) = h \quad (10)$$

$$v_x(0) = v_{0x} \quad (11)$$

$$v_y(0) = v_{0y} \quad (12)$$

donde h , v_{0x} y v_{0y} son datos del problema que habrá que obtener por medición.

Las ecuaciones (7) a (12) constituyen el modelo matemático del problema. Operando con este modelo matemático se pueden predecir muchas cosas: la trayectoria de la partícula, el lugar en que esta choca con el suelo, la máxima altura que alcanza en su recorrido, etcétera.

Como el problema que se desea investigar es la distancia que recorre la piedra en su trayectoria, será necesario trabajar con el modelo para obtener:

- La ecuación $y = f(x)$ de la trayectoria.
- Las coordenadas $(b, 0)$ del punto en que la partícula toca al piso.
- La longitud L de la trayectoria, que se calcula como:

$$L = \int_0^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx \quad (13)$$

Integrando la ecuación (7) respecto a t : $\frac{dx}{dt} = C_1$

Teniendo en cuenta la igualdad (11): $\frac{dx}{dt} = v_{0x}$

Integrando de nuevo respecto a t : $x = v_{0x}t + C_2$

Evaluando para $t = 0$ y utilizando (9): $x = v_{0x}t$ (14)

Integrando la ecuación (8) respecto a t : $\frac{dy}{dt} = -gt + C_3$

Evaluando en $t = 0$ y usando (12): $\frac{dy}{dt} = -gt + v_{0y}$

Integrando de nuevo respecto a t : $y = -\frac{gt^2}{2} + v_{0y}t + C_4$

Evaluando para $t = 0$ y empleando (10): $y = -\frac{gt^2}{2} + v_{0y}t + h$ (15)

Las ecuaciones (14) y (15) constituyen la forma paramétrica de la trayectoria de la partícula. La ecuación explícita se puede obtener eliminando el parámetro t . Para ello, se despeja t de (14) y se sustituye en (15):

$$y = -\frac{g}{2} \left(\frac{x}{v_{0x}} \right)^2 + v_{0y} \left(\frac{x}{v_{0x}} \right) + h$$

Es decir:

$$y = -\frac{g}{2v_{0x}^2} x^2 + \frac{v_{0y}}{v_{0x}} x + h \quad (16)$$

Que es la ecuación explícita de la trayectoria. Nótese que se trata de una parábola que abre hacia abajo. Para obtener la longitud de la trayectoria es necesario obtener la derivada de y respecto a x :

$$\frac{dy}{dx} = -\frac{g}{v_{0x}^2} x + \frac{v_{0y}}{v_{0x}}$$

Sustituyendo en la integral (13):

$$L = \int_0^b \sqrt{1 + \left(-\frac{g}{v_{0x}^2} x + \frac{v_{0y}}{v_{0x}} \right)^2} dx \quad (17)$$

El límite de integración, b , se obtiene haciendo $y = 0$ en (16) y despejando x . Aparecerán dos raíces, pero una de ellas es negativa y carece de sentido. Se obtiene:

$$b = \frac{v_{0x}}{g} \left(v_{0y} + \sqrt{v_{0y}^2 + 2gh} \right) \quad (18)$$

Hasta aquí, todos los pasos que se han dado en la solución del modelo matemático han sido exactos. Si la integral (17) se calcula por la regla de Newton – Leibniz entonces no habrá error de truncamiento. Nótese, sin embargo, que el cálculo de esta integral por esa vía es bastante engorroso. En el capítulo 5 de este libro se estudiarán varios métodos numéricos de integración que permiten calcular esta integral aproximadamente, con un error completamente controlado; este error, que puede hacerse tan pequeño como sea necesario, constituiría el error de truncamiento en este problema.

Para poder realizar los cálculos, en cualquier método que se utilice, se requiere conocer los datos: v_{0x} , v_{0y} y h , los cuales habrá que obtener por medición. En estas mediciones, en particular las de las velocidades, se introducirán errores de medición que afectarán también el resultado obtenido.

Por último, estarán presentes los errores de redondeo. Al calcular el valor de b y, posteriormente, la integral definida, se requerirá realizar cientos de operaciones aritméticas en una computadora digital; en cada una de esas operaciones se introducen pequeños errores de redondeo que también afectarán el resultado.

Dentro de todo este océano de errores, se obtiene, a pesar de todo, un resultado numérico que constituye una aproximación de la longitud de la trayectoria recorrida por la piedrecilla. Si las cosas se hacen bien, se logrará que esta aproximación sea aceptable y que pueda obtenerse sin un esfuerzo exagerado. Ese es el objetivo de la Matemática Numérica.

Ejercicios

1. En un relato breve debido al escritor argentino Jorge Luis Borges, se cuenta de un imaginario país en que los cartógrafos eran tan meticulosos y fieles a los detalles que para hacer el mapa de la nación habían necesitado todo un estado. Analice cómo se relaciona esta historia con el arte de modelar y los errores de modelación.
2. En el ejemplo 1 se hicieron varias hipótesis respecto al problema del lanzamiento de una partícula. Si se quisiera acercar un poco más el modelo a la realidad, analice qué hipótesis podría ser modificada razonablemente.
3. Aplique el modelo matemático elaborado en el ejemplo 1, a la caída de una gota de lluvia desde una nube que se encuentra a 5 000 metros de altura. Según este modelo, calcule la velocidad (en kilómetros por hora) de la gota de lluvia al llegar a la tierra. Suponga que la velocidad inicial de la gota es cero y todas las demás hipótesis del modelo. Analice si el resultado obtenido es razonable e indique qué hipótesis deberían ser modificadas para que el modelo se pueda utilizar en este caso.
4. La barras de acero para la construcción se fabrican en diámetros estándar de $\frac{1}{4}$ ", $\frac{3}{8}$ ", $\frac{1}{2}$ ", $\frac{5}{8}$ ", $\frac{3}{4}$ ", etc. En un modelo destinado a calcular el diámetro de las barras de acero que se colocarán en un elemento de hormigón armado, analice la magnitud de los errores que se deben tolerar.
5. Las resistencias eléctricas y los capacitores que se utilizan en el trabajo con circuitos electrónicos corrientes, se fabrican con errores de hasta 10%. En la modelación de circuitos electrónicos con vistas al diseño, analice la magnitud de los errores que se deben permitir.
6. En el diseño mecánico es usual trabajar con láminas de acero y tornillos. Estos elementos se fabrican industrialmente en espesores y diámetros discretos como 3mm, 4mm, 5mm, 6mm, etc. Analice con qué errores se puede realizar la modelación destinada al diseño de elementos mecánicos que utilizan elementos de este tipo.
7. En la pantalla de un monitor de una computadora personal los textos y los gráficos que aparecen se producen mediante puntos de colores llamados pixels. En la actualidad, la cantidad de pixels es del orden de 1000 en sentido horizontal y un poco menos en el sentido vertical. Si se está modelando un proceso con el objetivo final de mostrar geométricamente en una pantalla el resultado, analice la magnitud de los errores numéricos que se pueden permitir.
8. Los ingenieros hidráulicos realizan diseños con sistemas de tuberías. Si se tiene en cuenta que las tuberías se venden con diámetros interiores muy específicos como: $\frac{3}{8}$ ", $\frac{1}{2}$ ", $\frac{3}{4}$ ", 1", $1\frac{1}{4}$ ", $1\frac{1}{2}$ ", 2", etc., analice qué precisión se necesita en un modelo destinado a determinar el diámetro de un sistema complejo de tuberías.

1.3 Medidas del error

Independientemente de cual haya sido la fuente de un error, muchas veces se necesita medirlo. En lo que sigue se introducirán varias definiciones con este propósito. En todos los casos, se supone que x^* representa un número real cualquiera y x un número real aproximado a x^* .

Definición 1

El *error de x* en relación con el valor exacto x^* se denota $\text{error}(x)$ y se define como la diferencia:

$$\text{error}(x) = x^* - x$$

■

Cuando x es mayor que x^* es costumbre decir que se trata de una aproximación por exceso y en ese caso el error es negativo. Por el contrario, cuando x es menor que x^* el error es positivo y se dice que la aproximación es por defecto.

Definición 2

El *error absoluto de x* en relación con el valor exacto x^* se denota $E(x)$ y se define como

$$E(x) = |\text{error}(x)|$$

■

Si bien el error absoluto de un número aproximado da una idea de la magnitud del error, no siempre se puede juzgar la calidad de la aproximación utilizando el error absoluto. Por ejemplo, si x es el resultado de medir una longitud y $E(x)$ es 2 mm, no se sabe si se trata de una buena o mala aproximación; si x^* fuera el largo de una habitación, probablemente se considere que x es una medición aceptable, pero si x^* es el diámetro de un tornillo, la medición que se ha realizado es muy mala. Por esta causa se introduce el concepto de *error relativo*:

Definición 3

El *error relativo de x* en relación con el valor exacto $x^* \neq 0$ se denota $e(x)$ y se define como

$$e(x) = \frac{E(x)}{|x^*|}$$

■

Nótese que el error absoluto posee la misma dimensión física que los números x y x^* . El error relativo, sin embargo, es una cantidad adimensional y muchas veces se expresa en por ciento.

Ejemplo 1

- La fachada de una casa tiene un ancho de 9 540 mm. Al medirla se comete un error absoluto de 5 mm. ¿Cuál fue el error relativo cometido?
- Una batería de auto tiene entre sus bornes exactamente 11,5 volt. Al medirla se obtiene, sin embargo, 11,6 volt. Calcule el error de medición, el error absoluto y el error relativo.

Solución:

- En este caso es $x^* = 9\,540$ mm y $E(x) = 5$ mm. El error relativo será de

$$e(x) = \frac{E(x)}{|x^*|} = \frac{5}{9540} = 0,000524 = 0,0524 \%$$

- Aquí se tiene $x^* = 11,5$ volt y $x = 11,6$ volt. Por tanto:

$$\text{error}(x) = x^* - x = 11,5 - 11,6 = -0,1 \text{ volt}$$

$$E(x) = 0,1 \text{ volt}$$

$$e(x) = \frac{0,1}{11,5} = 0,008696 = 0,8696 \% \quad \blacksquare$$

Es muy frecuente que el error de un número aproximado no pueda ser conocido. Nótese que cuando se conoce un valor aproximado x y su error, entonces siempre se podría hallar el valor exacto como $x^* = x + \text{error}(x)$. La mayor parte de las veces hay que conformarse con una cota superior del error.

Definición 4

El *error absoluto máximo* de x en relación con x^* se denota $E_m(x)$ y se define como cualquier número real que satisfaga la condición:

$$E_m(x) \geq E(x) \quad \blacksquare$$

Obsérvese que, de acuerdo con la definición, el error absoluto máximo de un número aproximado no es un número preciso, sino cualquier número que no sea menor que el error absoluto. Es decir, el error absoluto máximo es cualquier número del cual se tenga la certeza de que nunca el error absoluto será mayor que él. Por supuesto, un error absoluto máximo muy grande no significa que el error absoluto sea grande, mientras que un error absoluto máximo pequeño sí garantiza que el error absoluto del número será pequeño.

Algo similar puede hacerse con los errores relativos:

Definición 5

El *error relativo máximo* de x en relación con x^* se denota $e_m(x)$ y se define como cualquier número real que satisfaga la condición:

$$e_m(x) \geq e(x) \quad \blacksquare$$

En la tabla 1 se resumen las cinco definiciones anteriores.

Concepto	Notación	Definición
Error de x	$\text{error}(x)$	$x^* - x$
Error absoluto de x	$E(x)$	$ \text{error}(x) $
Error relativo de x	$e(x)$	$\frac{E(x)}{ x^* }$
Error absoluto máximo de x	$E_m(x)$	$E_m(x) \geq E(x)$
Error relativo máximo de x	$e_m(x)$	$e_m(x) \geq e(x)$

Tabla 1

Vinculadas con los conceptos anteriores, existen algunas relaciones que resultan útiles para el trabajo con números aproximados. A continuación se exponen y fundamentan las mismas.

El mínimo error absoluto máximo

Sea x^- una aproximación por defecto de x^* y x^+ una aproximación por exceso del mismo número x^* . Si x es cualquier número del intervalo $[x^-, x^+]$, este número x será una aproximación de x^* cuyo error absoluto máximo puede ser determinado fácilmente. Para ello, considérese la figura 1. En ella se muestran las tres aproximaciones x^- , x y x^+ . Como el verdadero valor x^* pertenece al intervalo $[x^-, x^+]$, entonces el error absoluto $E(x)$ no puede exceder a la mayor de las dos distancias $(x^+ - x)$ y $(x - x^-)$ que determina x en el intervalo.

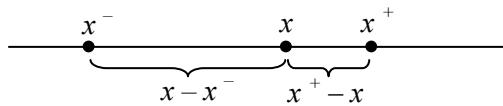


Figura 1

Es decir:

$$E_m(x) = \max \{x - x^-, x^+ - x\}$$

De la figura resulta obvio que el error absoluto máximo tomará su mínimo valor si se escoge x justo en el centro del intervalo $[x^-, x^+]$, en ese caso $E_m(x)$ será exactamente la semiamplitud del intervalo, esto es:

Si se toma

$$x = \frac{x^- + x^+}{2}$$

entonces se minimiza el error absoluto máximo, el cual será

$$E_m(x) = \frac{x^+ - x^-}{2}$$

Ejemplo 2

Se sabe que la raíz de una ecuación se encuentra en el intervalo $[5,25; 5,37]$. Si se toma como aproximación $x = 5,3$ ¿Cuál es el error absoluto máximo de esta aproximación? ¿Para qué valor de x se obtendría el menor error absoluto máximo?

Solución:

De acuerdo con lo explicado anteriormente, para $x = 5,3$ se tiene:

$$E_m(x) = \max \{5,3 - 5,25; 5,37 - 5,3\} = \max \{0,05; 0,07\} = 0,07$$

El error absoluto máximo se minimiza tomando x en el centro del intervalo $[5,25; 5,37]$, es decir:

$$x = \frac{5,25 + 5,37}{2} = 5,31$$

Para ese valor de x , el error absoluto máximo será:

$$E_m(x) = \frac{5,37 - 5,25}{2} = 0,06$$

Obsérvese que esto no significa que 5,31 esté más próximo a x^* que 5,3. Tan solo puede afirmarse que si se toma $x = 5,3$ como aproximación, el error absoluto *podría* llegar hasta 0,07 pero tomando $x = 5,31$ el error absoluto no puede pasar de 0,06.

Relación entre $E_m(x)$ y $e_m(x)$

El error absoluto máximo y el error relativo máximo de un número aproximado x con respecto a un número exacto x^* , están estrechamente relacionados. En efecto, como, según la definición:

$$e(x) = \frac{E(x)}{|x^*|} \quad (1)$$

se tiene que:

$$e(x) \leq \frac{E_m(x)}{|x^*|}$$

Así que el miembro derecho de esta desigualdad puede tomarse como error relativo máximo, es decir:

$$e_m(x) = \frac{E_m(x)}{|x^*|} \quad (2)$$

Similarmente, la igualdad (1) se puede escribir:

$$E(x) = |x^*| e(x)$$

de donde resulta:

$$E(x) \leq |x^*| e_m(x)$$

por lo cual, el miembro de la derecha de la desigualdad puede escogerse como error absoluto máximo de x , esto es:

$$E_m(x) = |x^*| e_m(x) \quad (3)$$

En muchos casos prácticos, las formulas (2) y (3) no pueden aplicarse por no conocer el número exacto x^* . En ese caso se utiliza en lugar de x^* una aproximación del mismo. Para utilizar la fórmula (2) es preferible, si se tiene, utilizar en lugar de $|x^*|$ una aproximación por defecto, ya que se trata de hallar una cota superior del error relativo. Al aplicar la fórmula (3), sin embargo, debe tomarse (si se posee) una aproximación por exceso de $|x^*|$.

Ejemplo 3

Arquímedes obtuvo para el número π la desigualdad:

$$3,14085 < \pi < 3,14286$$

Obtenga, a partir de aquí, una aproximación de π con su error absoluto máximo y su error relativo máximo.

Solución:

Se tomará como valor aproximado π_a el punto medio del intervalo:

$$\pi_a = \frac{3,14085 + 3,14286}{2} = 3,14186$$

El error absoluto máximo viene dado por:

$$E_m(\pi_a) = \frac{3,14286 - 3,14085}{2} = 0,001$$

Para hallar el error relativo máximo se usará la relación:

$$e_m(x) = \frac{E_m(x)}{|x^*|}$$

Suponiendo que no se conoce el valor verdadero de π , se tomará una aproximación por defecto:

$$e_m(x) \leq \frac{0,001}{3,14085} = 0,000318 < 0,00032$$

Puede tomarse $e_m(\pi_a) = 0,00032$ ó, utilizando por cientos, $e_m(\pi_a) = 0,032\%$

Hallando aproximaciones por defecto y por exceso

Si se conoce el error absoluto máximo de un número x aproximado a un número x^* se pueden hallar aproximaciones por defecto y por exceso con facilidad. En efecto, como

$$E(x) = |x^* - x| \leq E_m(x)$$

Se tiene que:

$$-E_m(x) \leq x^* - x \leq E_m(x)$$

y, sumando x en los tres miembros:

$$x - E_m(x) \leq x^* \leq x + E_m(x) \quad (4)$$

Es decir,

$$x^- = x - E_m(x) \quad \text{y} \quad x^+ = x + E_m(x)$$

son aproximaciones por defecto y por exceso respectivamente de x^* . Esto frecuentemente se expresa:

$$x^* = x \pm E_m(x)$$

Ejemplo 4

Las componentes eléctricas no se fabrican con valores exactos. Por ejemplo, las resistencias para circuitos electrónicos, se distribuyen con errores relativos máximos de 10%, 5% ó 1% de acuerdo

con su calidad (y su precio). Si una resistencia tiene un valor nominal de $47 \text{ k}\Omega$ y tiene un error relativo máximo de 5%, ¿en qué intervalo se encuentra el valor de la resistencia?

Solución:

Se tiene: $R = 47 \text{ k}\Omega$,
 $e_m(R) = 5\% = 0,05$

y, de ahí, $E_m(R) = |R|e_m(R) \approx 47\text{k}\Omega \cdot 0,05 = 2,35\text{k}\Omega$

Por tanto, según (4): $R - E_m(R) \leq R^* \leq R + E_m(R)$

Es decir: $44,65\text{k}\Omega \leq R^* \leq 49,35\text{k}\Omega$

lo cual puede expresarse también como: $R^* = 47\text{k}\Omega \pm 2,35\text{k}\Omega$

Ejercicios

1. Calcule los errores absolutos y relativos que se cometan al aproximar las siguientes constantes matemáticas y físicas por los valores indicados a su derecha.

a) $e = 2,7182818\dots$ (base de los logaritmos neperianos)	$e_A = 2,7$
b) $c = 2,99793 \cdot 10^8 \text{ m/s}$ (velocidad de la luz en el vacío)	$c_A = 3 \cdot 10^8 \text{ m/s}$
c) $g = 9,8 \text{ m/s}^2$ (aceleración de la gravedad)	$g_A = 10 \text{ m/s}^2$
d) $C = 0,577216$ (constante de Euler)	$C_A = 0,58$
e) $m_p = 1,67239 \cdot 10^{-24} \text{ g}$ (masa del protón)	$m_{pA} = 1,7 \cdot 10^{-24} \text{ g}$
f) $m_e = 9,1983 \cdot 10^{-28} \text{ g}$ (masa del electrón)	$m_{eA} = 10^{-27} \text{ g}$
2. Suponga que usted no conoce el valor de $\sqrt{2}$. Como $1,4^2 = 1,96 < 2$ y $1,5^2 = 2,25 > 2$, se puede asegurar que $1,4 < \sqrt{2} < 1,5$. A partir de esta conclusión obtenga una aproximación para $\sqrt{2}$ que tenga el mínimo error absoluto máximo. Halle también el error relativo máximo.
3. Si la aproximación que usted halló en el ejercicio anterior se eleva al cuadrado, se puede saber si ella es menor o mayor que $\sqrt{2}$. De ese modo usted puede determinar un intervalo de menor amplitud que el ofrecido en ese ejercicio, donde se encuentre $\sqrt{2}$. A partir de este nuevo intervalo se puede encontrar una nueva aproximación y su error absoluto máximo. Siguiendo esta idea, calcule una aproximación para $\sqrt{2}$ que posea un error absoluto menor que 0,001.
4. En el capacitor de arranque de un motor eléctrico aparece su capacidad como: $32 \pm 3 \mu\text{F}$. Determine el error absoluto máximo y el error relativo máximo del valor nominal de $32 \mu\text{F}$.
5. Se quiere calcular la distancia entre dos puntos de un territorio a partir de un mapa de escala $1 \text{ Km} = 1 \text{ cm}$. Midiendo con una cinta métrica se obtuvo una distancia de 18,3 cm. Si los editores del mapa garantizan un error relativo menor que 1% en la confección del mapa y el error en la medición pudiera haber sido hasta de 1mm, calcule entre qué valores debe hallarse la distancia real que se busca.

6. En un programa que produce una gráfica sobre la pantalla de una computadora, se calculan las coordenadas (x, y) de un punto del display. Estas coordenadas son números reales, pero para dibujarlas en la pantalla primero hay que redondearlas al valor entero más cercano obteniendo (x_p, y_p) donde x_p y y_p son pixels. Si se está utilizando una resolución de 1024 por 768 pixels y la pantalla mide 28 cm de ancho y 21 cm de altura, halle el error absoluto máximo que se produce en dirección vertical y en dirección horizontal cuando el punto (x, y) se representa en el display.
 7. Se mide el voltaje en un circuito eléctrico con un voltímetro, cuyo fabricante garantiza un error relativo máximo de 0,1 %. Si el voltaje medido es de 225 v, determine el error absoluto máximo de la medición realizada y halle un intervalo donde se encuentra el verdadero valor con toda seguridad.
 8. A veces se utiliza la aproximación $\sin x \approx x$ para valores pequeños de x . Grafique (mejor si usa algún programa para realizar la gráfica) las funciones $y^* = \sin x$ y $y = x$ en un mismo sistema coordenado y compárelas. Determine el máximo valor x_{max} que puede tomar x de modo que el error absoluto que se comete en la aproximación sea menor que 0,001.
 9. Una mejor aproximación que en el ejercicio anterior se alcanza si se toma $\sin x \approx x - \frac{x^3}{6}$. Repita en este caso el enunciado del ejercicio anterior.
 10. En la antigua babilonia se conocía una forma para hallar aproximadamente la raíz cuadrada de un número que fuera cercano a un cuadrado perfecto. En la notación actual, se escribiría así:
- $$\sqrt{a^2 + x} \approx a + \frac{x}{2a}$$
- Halle el error absoluto y el error relativo al calcular $\sqrt{10}$ por esta vía.
11. En el papiro de Rihn aparece una fórmula para calcular aproximadamente el área de un círculo como la de un cuadrado cuyo lado fuera $8/9$ del diámetro del círculo. Demuestre que esto equivale a tomar para π la aproximación $256/81$. Calcule el error absoluto y el error relativo de esta aproximación.
 12. De un sobre con resistencias eléctricas corrientes (10% de error relativo máximo) correspondientes a un valor nominal de $56 \text{ K}\Omega$ se midieron algunos ejemplares y se encontró una resistencia de $50 \text{ K}\Omega$ y otra de $60 \text{ K}\Omega$. Determine si alguna de ellas constituye una equivocación del fabricante.

1.4 Cifras significativas y cifras exactas

En el trabajo con números aproximados es muy frecuente utilizar el lenguaje de las cifras. En este epígrafe se harán las definiciones necesarias y se estudiarán las relaciones entre esta manera de hablar y los conceptos introducidos en el epígrafe anterior.

Cifras significativas de un número

El sistema de numeración que se emplea hoy en todo el mundo, salvo en cuestiones muy específicas, es el creado por la antigua civilización hindú y difundido posteriormente por los

árabes en Europa durante la edad media. Es un sistema posicional de base 10 y, por la simplicidad de los algoritmos que utiliza para las operaciones aritméticas, desplazó rápidamente a otros sistemas usados en aquella época, tales como el romano y el griego.

En este sistema, cualquier número real puede expresarse utilizando solamente 10 símbolos (6 dígitos): 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Cuando un dígito aparece formando parte de un número, representa un valor que depende de su figura y de su posición. Para simplificar la exposición que sigue, se introduce el concepto de valor posicional de un número.

Definición 1

Si el dígito d ocupa en un número real la posición k -sima según la siguiente tabla:

Lugar decimal	k
⋮	
Milésimas	-3
Centésimas	-2
Décimas	-1
Unidades	0
Decenas	1
Centenas	2
⋮	

se denota el *valor posicional* de d como $p(d)$ y se define como $p(d) = 10^k$

■

Nótese que el valor posicional de un dígito dentro de un número no es más que el valor que tendría la unidad colocada en esa misma posición. El valor de un dígito d dentro de un número, que se abreviará como $v(d)$, se obtiene multiplicando el dígito por su valor posicional y el valor del número es la suma de los valores de cada uno de sus dígitos. En el ejemplo que sigue se aclaran estas ideas.

Ejemplo 1

Determine el valor posicional y el valor de cada dígito en el número real 65,403

Solución:

$$\begin{array}{ll}
 p(6) = 10^1 = 10 & v(6) = 6 \quad p(6) = 60 \\
 p(5) = 10^0 = 1 & v(5) = 5 \quad p(5) = 5 \\
 p(4) = 10^{-1} = 0,1 & v(4) = 4 \quad p(4) = 0,4 \\
 p(0) = 10^{-2} = 0,01 & v(0) = 0 \quad p(0) = 0 \\
 p(3) = 10^{-3} = 0,001 & v(3) = 3 \quad p(3) = 0,003
 \end{array}$$

Obsérvese que la suma $v(6) + v(5) + v(4) + v(0) + v(3)$ coincide con el valor del número real, esto es, 65,403.

■

Aunque el valor de cualquier dígito 0 es cero, independientemente de su posición, en la expresión del número no se pueden omitir los ceros porque ello afectaría la posición de los dígitos restantes,

así, por ejemplo, los números 65,403 y 65,43 no significan lo mismo ya que al omitir el dígito 0 la posición del 3 es -2 y no -3 como en el primer caso.

Definición 2

Cuando un dígito 0 se incluye en un número con el único propósito de ocupar una posición dentro del número, ese dígito se llaman cero no significativo. En los demás casos, se dice que el 0 es significativo. Todos los dígitos que no son 0 son significativos.

Ejemplo 2

En el número 0,0002030 ¿Qué dígitos son significativos?

Solución:

Los primeros cuatro ceros del número no son significativos, solo sirven para informar que el dígito 2 ocupa la posición – 4. El quinto y sexto ceros son ambos significativos, en ambos casos se desea hacer notar que el valor de esa posición decimal debe ser cero. En conclusión, a continuación se muestran subrayados los dígitos significativos del número:

0,0002030



En general, todos los ceros que aparecen entre dígitos significativos, son significativos. En algunos casos, solamente el contexto donde se encuentra el número permite determinar si un 0 es significativo o no, de acuerdo con la *intención* de la persona que lo escribió.

Ejemplo 3

A continuación aparece el número 120 000 en varios contextos distintos. Determine en cada caso qué ceros son significativos.

- a) En un titular de una periódico aparece “120 000 personas participaron en la concentración de ayer”.
- b) En el sorteo efectuado ayer resultó premiado el número 120 000.
- c) En el informe de un cajero de un banco al gerente. “En el día de ayer fueron depositados en caja un total de 120 000 USD”.
- d) En un libro de Zoología: “El cuerpo de este animal está cubierto por unos 120 000 pelos”

Solución:

- a) Como es muy difícil que alguien haya podido contar exactamente las personas que participaron en una concentración, se sobre entiende que los ceros que aparecen son no significativos.
- b) Todos los ceros son significativos.
- c) En el trabajo bancario no suelen hacerse aproximaciones, así que seguramente todos los ceros son significativos.
- d) A nadie le interesaría conocer exactamente cuantos pelos cubren el cuerpo de un animal, por otra parte, difícilmente todos los animales de esta especie poseerán la misma cantidad de pelos, además, la palabra “unos” indica que se trata de una aproximación; de todo esto se infiere que seguramente los cuatro ceros de este número son no significativos.

La notación científica

En los trabajos científicos, donde es importante no dejar a la interpretación de cada persona el saber si un dígito es significativo o no y donde, además, suelen aparecer cantidades muy grandes y muy pequeñas, se utiliza la llamada notación científica, que consiste en expresar los números como un producto de un número (llamado *mantisa*) mayor o igual que 1 y menor que 10 por una potencia de 10. En la mantisa se incluyen todos los dígitos significativos del número y solamente ellos.

Ejemplo 4

A continuación se muestran algunas constantes físicas en notación científica. Expréselas sin utilizar esta notación y observe lo inconveniente de hacerlo.

Número de Avogadro: $N = 6,02497 \cdot 10^{23} \text{ mol}^{-1}$
 Masa del electrón: $m_e = 9,1083 \cdot 10^{-28} \text{ g}$
 Velocidad de la luz en el vacío $c_0 = 2,99793 \cdot 10^{10} \text{ cm/s}$

Solución:

Además de lo extenso y confuso de la escritura, en el caso de números enteros grandes como N y c_0 , no se podría determinar por el contexto qué ceros son significativos y cuáles no. ■

Nótese que cuando se habla de cifras significativas no se tiene en cuenta la veracidad o no del número, sino solamente la intención. Así, si alguien afirma que “ayer desfilaron 236 703 personas por la avenida principal de la ciudad” cualquiera comprende que este número seguramente es inexacto, pero todas sus cifras son significativas. Algo muy diferente sucede con el concepto de cifra exacta que se verá a continuación.

Cifra exacta

Un dígito d de un número x se dice que es un *dígito exacto* o una *cifra exacta* si el error absoluto de x es menor o igual que la mitad del valor posicional de d . Esto es, si

$$E(x) \leq \frac{1}{2} p(d)$$

En caso contrario, la cifra d se dice que no es exacta.

Ejemplo 5

A continuación se dan varios números x aproximados. Determine en cada caso qué cifras de x son exactas.

- a) $x = 3,1416$. Se sabe que $x^* = \pi = 3,141592653\dots$
 b) $x = 3,99999$. Se sabe que $x^* = 4$
 c) $x = 4,20457$. Se sabe que $x^* = 4,20451$
 d) $x = 0,00046384$. Se sabe que $E(x) = 0,0000002$
 e) $x = 23,01241$. Se sabe que $E_m(x) = 0,04$

Solución:

- a) Antes que todo hay que hallar $E(x) = |x^* - x| = |3,141592653 - 3,1416| = 0,00000734\dots$

Para determinar si una cifra d es exacta hay que comprobar si este error satisface que

$$E(x) \leq \frac{1}{2} p(d)$$

En este caso, procediendo de izquierda a derecha, se tiene que:

$E(x) \leq \frac{1}{2} p(3) = 0,5$	3 es una cifra exacta
$E(x) \leq \frac{1}{2} p(1) = 0,05$	1 es una cifra exacta
$E(x) \leq \frac{1}{2} p(4) = 0,005$	4 es una cifra exacta
$E(x) \leq \frac{1}{2} p(1) = 0,0005$	1 es una cifra exacta
$E(x) \leq \frac{1}{2} p(6) = 0,00005$	6 es una cifra exacta

Por razones didácticas se ha procedido analizando todos los dígitos de izquierda a derecha, pero resulta obvio que habría bastado probar que la cifra 6, que es la de menor valor posicional, era exacta para afirmar que ella y todas las que se encuentran a su izquierda, son exactas. En resumen, en este caso los 5 dígitos del número x son exactos.

- b) En este caso es $E(x) = |x^* - x| = |4 - 3,99999| = 0,00001$

Si d representa al quinto 9 de x , $\frac{1}{2} p(d) = 0,000005$ y no se satisface $E(x) \leq \frac{1}{2} p(d)$

Si d representa al cuarto 9 de x , $\frac{1}{2} p(d) = 0,00005$ y se satisface $E(x) \leq \frac{1}{2} p(d)$

Como los demás dígitos de x poseen mayor valor posicional, ellos también serán exactos. En resumen, en este caso las cifras exactas de x son las que aparecen subrayadas a continuación: 3,9999. El último 9 no es una cifra exacta.

- c) Como se conoce x^* se puede calcular el error absoluto:

$$E(x) = |x^* - x| = |4,20451 - 4,20457| = 0,00006$$

El dígito 4 que se encuentra en las milésimas tiene valor posicional 0,001. Se cumple que:

$$E(x) \leq \frac{1}{2} p(4) = 0,0005$$

Luego, las milésimas y todas las cifras que aparecen a su izquierda, son exactas. En cuanto al dígito 5, que aparece en la cuarta cifra decimal:

$$E(x) > \frac{1}{2} p(5) = 0,00005$$

El dígito 5 y, con mayor razón, el 7 que aparece a su derecha son cifras no exactas. Como resumen, a continuación aparecen subrayadas las cifras exactas del número x : 4,20457.

- d) En este caso se ilustra el hecho de que no es necesario conocer el valor exacto x^* para determinar los dígitos exactos de x , basta conocer el error absoluto. Como el error contiene un 2 en la sexta cifra decimal, está claro que la sexta cifra de x no puede ser exacta.

Considérese el 6 que se halla en el quinto lugar decimal, su valor posicional es

$$p(6) = 0,00001$$

y se cumple que: $E(x) = 0,000002 \leq \frac{1}{2} p(6) = 0,000005$

Se concluye que son exactas las cifras 6 y las que se encuentran a su izquierda, las que aparecen subrayadas a continuación: 0,00046384.

- e) En este caso se desconoce el error absoluto de x , solo se tiene el error absoluto máximo, que es una cota superior del error absoluto. No obstante, esa información resulta útil: indica que el error absoluto podría llegar a ser 0,04. En ese caso, es evidente que la segunda cifra decimal (un 1) pudiera no ser exacta. En cuanto al dígito 0 que se halla en el primer lugar decimal, se cumple que:

$$E(x) \leq 0,04 \leq \frac{1}{2} p(0) = 0,05$$

De modo que este dígito 0 es una cifra exacta y, con mayor razón, las que se encuentran a la izquierda. Estas cifras exactas se muestran subrayadas a continuación: 23,01241. Como el valor preciso del error absoluto no se tiene en este ejemplo, los dígitos que no aparecen subrayados tienen un carácter desconocido y pudieran ser exactos o no; usualmente a estas cifras se les llama *dudosas*.

Contando las cifras exactas

Una vez que se conoce qué cifras de un número aproximado son exactas, un asunto mucho más simple es contarlas. Para ello se siguen dos criterios.

Definición 3

La *cantidad de cifras exactas* de un número aproximado es la cantidad de dígitos *significativos* exactos de dicho número. La *cantidad de cifras decimales exactas* de un número aproximado es la cantidad de cifras exactas que están después de la coma decimal.

Ejemplo 6

A continuación se muestran los cinco números aproximados del ejemplo 5 en los cuales las cifras exactas aparecen subrayadas. Determine en cada caso la cantidad de cifras exactas y la cantidad de cifras decimales exactas.

- a) $x = \underline{3,1}416$
- b) $x = \underline{3,9999}9$
- c) $x = 4,\underline{2}0457$
- d) $x = \underline{0,00046384}$
- e) $x = \underline{23,0}1241$

Solución:

- a) 5 cifras exactas; 4 cifras decimales exactas.
- b) 5 cifras exactas; 4 cifras decimales exactas.
- c) 4 cifras exactas; 3 cifras decimales exactas.
- d) 2 cifras exactas; 5 cifras decimales exactas.
- e) 3 cifras exactas; 1 cifra decimal exacta.

El redondeo

Cuando un número posee una cantidad demasiado grande de cifras significativas y, sobretodo si no son exactas, las cifras excedentes se redondean. Se supone que el lector conoce las reglas de redondeo, que se estudian en cursos anteriores, así que no se entrará a verlas en detalle. Estas reglas están diseñadas de tal modo que cuando se redondea un número exacto, el número aproximado que resulta tiene todas sus cifras exactas, ya que el error absoluto que se introduce al redondear es menor que la mitad del valor posicional del último dígito conservado.

Cuando se redondea un número aproximado, sin embargo, deben tenerse algunas precauciones. Un número aproximado posee siempre algún error, al redondear el número se introduce un error adicional que puede agregarse al error que existía. Este fenómeno puede causar que al redondear un número aproximado eliminando todas las cifras no exactas y conservando solamente las exactas, ocurra que alguna cifra que originalmente era exacta, deje de serlo debido al incremento del error. Esto se ilustra en el siguiente ejemplo.

Ejemplo 7

Considérese el número exacto

$$x^* = e = 2,718\,281\,828\,459\,045\dots$$

y el valor aproximado

$$x = 2,718\,286\,325\,411$$

El error absoluto es

$$E(x) = |x^* - x| = |2,718\,281\,828\,459\,045\dots - 2,718\,286\,325\,411| = 0,000\,004\,49\dots$$

Es obvio que su sexta cifra decimal no es exacta. En cuanto a la quinta (un 8) se cumple que

$$E(x) \leq \frac{1}{2} p(8) = 0,000\,005$$

así que se trata de una cifra exacta. Los dígitos exactos de x son 6 y aparecen subrayados a continuación:

$$x = \underline{2,718286325411}$$

Obsérvese que este número aproximado posee un error por exceso. Si se decidiera ahora redondear este número conservando solamente las cifras exactas, se introduciría un nuevo error (que en este caso, casualmente, también es por exceso). El número obtenido sería:

$$x_1 = 2,71829$$

El error absoluto de x_1 es

$$E(x_1) = |x^* - x_1| = |2,718\,281\,828\,459\,045\dots - 2,718\,29| = 0,000\,008\,17\dots$$

de manera que ahora la quinta cifra decimal ya no es exacta. El número x_1 posee solamente 5 cifras exactas que aparecen subrayadas a continuación: $x_1 = \underline{2,71829}$. ■

Por esta razón se acostumbra redondear los números aproximados conservando una o dos de sus cifras no exactas (o dudosas). Por cierto, cuando se trata de cifras dudosas, es frecuente que algunas de ellas sean realmente exactas y esta es otra razón para esta regla.

Cifras decimales exactas y error absoluto

Existe una relación muy directa entre la cantidad de cifras decimales exactas y el error absoluto de un número. En efecto, como la k -sima cifra decimal tiene un valor posicional $\frac{1}{2}10^{-k}$, un número aproximado posee k cifras decimales exactas si y solo si su error absoluto es menor o igual que $\frac{1}{2}10^{-k}$. Como ejemplo se relaciona a continuación algunos casos:

Cifras decimales exactas	Error absoluto menor o igual que:
2	0,005
3	0,0005
4	0,00005
5	0,000005

Cifras exactas y error relativo

La cantidad de cifras exactas de un número no está relacionada con el error absoluto sino con el error relativo. Para ello, supóngase que el número aproximado x posee k cifras exactas, es decir, que sus k primeras cifras significativas son exactas. Si el número x se expresa en notación científica se escribiría:

$$x = m \cdot 10^q$$

donde m representa a la mantisa y el exponente q es algún número entero (positivo, negativo o cero). Como m posee 1 dígito entero que es significativo y exacto y, por tanto, $k - 1$ cifras decimales exactas, su error absoluto satisface

$$E(m) \leq \frac{1}{2}10^{-(k-1)}$$

El error absoluto máximo de x se puede obtener fácilmente, multiplicando por el factor 10^q que es un número exacto, es decir:

$$E(x) \leq \frac{1}{2}10^{-(k-1)} \cdot 10^q$$

Nótese que, el error absoluto de x depende no solo del número de cifras exactas, sino también de la cantidad q . Algo distinto sucede con el error relativo de x . Para calcularlo, basta dividir por el valor absoluto de x :

$$e(x) = \frac{E(x)}{|x|} \leq \frac{\frac{1}{2}10^{-(k-1)} \cdot 10^q}{|m \cdot 10^q|} = \frac{10^{-(k-1)}}{2|m|}$$

donde se observa que el error relativo de un número no depende de la magnitud (q) del número sino del número k de cifras exactas. De la expresión anterior se puede obtener una fórmula que a

veces se emplea para hallar el error relativo máximo a partir de la cantidad de cifras exactas. En efecto, como $|m| \geq 1$, se tiene:

$$e(x) \leq \frac{10^{-(k-1)}}{2|m|} \leq \frac{1}{2} \cdot 10^{-(k-1)}$$

Por tanto, se puede tomar como error relativo máximo la expresión:

$$e_m(x) = \frac{1}{2} \cdot 10^{-(k-1)} \quad (1)$$

Pero esta fórmula suele producir cotas exageradas del error relativo, pues, como se ha visto, se obtiene suponiendo $m = 1$, cuando realmente m puede ser hasta 10. En la práctica es preferible calcular el error relativo hallando primero el absoluto, como se ilustra en el ejemplo que sigue, en el cual se aprecia, tal como se ha dicho, la relación del error relativo con la cantidad de cifras exactas del número y no con su magnitud.

Ejemplo 13

Los tres números que siguen poseen cuatro cifras exactas. Determine en cada caso el error absoluto máximo y el error relativo máximo.

- a) $x = 673\ 500$
- b) $x = 67,35$
- c) $x = 0,000\ 673\ 5$

Solución:

- a) $x = 673\ 500$. Como 5 es la última cifra exacta y posee un valor posicional de 100, el error absoluto es menor o igual que 50. Por tanto, $E_m(x) = 50$. El error relativo máximo se puede obtener como:

$$e_m(x) = \frac{E_m(x)}{|x|} = \frac{50}{673500} = 0,0000074$$

- b) $x = 67,35$. Como 5 es la última cifra exacta y posee un valor posicional de 0,01, el error absoluto es menor o igual que 0,005. Por tanto, $E_m(x) = 0,005$. El error relativo máximo se puede obtener como:

$$e_m(x) = \frac{E_m(x)}{|x|} = \frac{0,005}{67,35} = 0,0000074$$

- c) $x = 0,000\ 673\ 5$. Como 5 es la última cifra exacta y posee un valor posicional de 10^{-7} , el error absoluto es menor o igual que $0,5 \cdot 10^{-7}$. Por tanto, $E_m(x) = 0,5 \cdot 10^{-7}$. El error relativo máximo se puede obtener como:

$$e_m(x) = \frac{E_m(x)}{|x|} = \frac{0,5 \cdot 10^{-7}}{0,0006735} = 0,0000074$$

En resumen, con cuatro cifras exactas, los números poseen los siguientes errores:

x	$E_m(x)$	$e_m(x)$
673 500	50	0,000074
67,35	0,005	0,000074
0,0006735	$0,5 \cdot 10^{-7}$	0,000074

Por cierto, al aplicar la fórmula (1) se obtiene $e_m(x) = \frac{1}{2} \cdot 10^{-(k-1)} = \frac{1}{2} \cdot 10^{-3} = 0,0005$ que es unas 7 veces mayor que el hallado por la otra vía.

Ejercicios

- Diga qué dígitos de los números que siguen no son significativos: a) 0,00048003; b) 12004; c) 4,54600.
- A continuación aparecen, dentro de determinados contextos, números enteros con varios ceros finales. Determine, si fuera posible, a partir del contexto, cuales de dichos ceros son significativos y cuales no.
 - De un folleto de Datos sobre Cuba: "... Su longitud es de 1250 kilómetros y sus costas suman alrededor de 7000 kilómetros..."
 - De la leyenda de un mapa: Escala 1:30 000 000.
 - De un plegable sobre la ciudad de Baracoa: "El municipio tiene una población de 80000 habitantes, de ella 35000 en la ciudad cabecera".
 - De una revista turística: "En los alrededores de la ciudad de Sucre (Bolivia) se encuentra la mayor cantidad (5000) de huellas de dinosaurios del mundo"
 - El año 2000 se considera el último del siglo XX.
- En los siguientes casos se dan números x aproximados a números exactos x^* . Determine cuales cifras de x son exactas y diga la cantidad de cifras exactas y de cifras decimales exactas de x .
 - $x^* = 2,718281\dots$ $x = 2,71$
 - $x^* = 0,000436$ $x = 0,00044$
 - $x^* = 236\ 578$ $x = 236\ 000$
 - $x^* = 0,066666\dots$ $x = 0,067$
 - $x^* = \pi$ $x = 3,14$
- En los siguientes casos se dan números aproximados y alguna información sobre su error. Determine en cada caso la cantidad de cifras exactas y de cifras decimales exactas del número aproximado. Diga qué cifras son dudosas.
 - $x = 5,839\ 362$ $E(x) = 0,0002$
 - $x = 0,0045387$ $e(x) = 0,03\%$
 - $x = 7,8876 \cdot 10^{-5}$ $E(x) = 4 \cdot 10^{-7}$
 - $x = 54,831$ $e_m(x) = 0,001$
 - $x = 45\ 846\ 352$ $E_m(x) = 300$
 - $x = 33\ 786$ $e(x) = 0,002$
 - $x = 0,055763$ $E_m(x) = 0,00005$
 - $x = 0,0000785$ $e_m(x) = 0,1\%$
- A continuación aparecen números aproximados con su cantidad de cifras exactas o de cifras decimales exactas. Determine en cada caso el error absoluto máximo y el error relativo máximo del número y diga qué cifras son dudosas.

- | | |
|---------------------|--------------------------------|
| a) $x = 697,6587$ | con 2 cifras decimales exactas |
| b) $x = 50,006$ | con 4 cifras exactas |
| c) $x = 0,0054768$ | con 3 cifras exactas |
| d) $x = 0,0005973$ | con 5 cifras decimales exactas |
| e) $x = 4,99999$ | con 3 cifras decimales exactas |
| f) $x = 390\ 785$ | con 4 cifras exactas |
| g) $x = 0,09878$ | con 2 cifras exactas |
| h) $x = 0,00003543$ | con 6 cifras decimales exactas |
6. Redondee los números que aparecen a continuación de acuerdo con las especificaciones que se indican. Determine en cada caso el error absoluto y el error relativo introducido por el redondeo.
- | | |
|----------------------|---|
| a) $x = 58,54654$ | Conservar 5 cifras significativas |
| b) $x = 0,045365$ | Conservar 4 cifras decimales |
| c) $x = 6,549873$ | Conservar hasta las milésimas |
| d) $x = 67\ 845,675$ | Conservar hasta las centenas |
| e) $x = 0,00657873$ | Redondear a partir de las diezmilésimas |

1.5 Los números en la computadora

Los números se representan en la computadora en forma binaria, es decir, utilizando dispositivos binarios de memoria, los cuales pueden guardar ceros y unos. Para almacenar un número se utilizan n bits de memoria y se le asocia a cada uno de los estados posibles (secuencia de ceros y unos) un número. Esto significa que la cantidad de números diferentes que se puede representar será 2^n . Dependiendo del valor de n esta cantidad será mayor o menor, pero siempre será finita. Como se trata de un conjunto finito de números, será necesariamente acotado. Esto es algo muy diferente de lo que sucede en Matemática, donde, por lo general, se trabaja con conjuntos infinitos y no acotados. Por su importancia, esta conclusión será referida con un número:

Propiedad 1

Los conjuntos numéricos que utiliza cualquier computadora son finitos y acotados inferior y superiormente.

En la representación de números enteros y de números reales se siguen convenciones muy diferentes que vale la pena estudiar por separado.

Representación de números enteros

Por su naturaleza, los números naturales y enteros no suelen aproximarse. Para representarlos, la máquina utiliza la notación de punto fijo. Esto se hace de diferentes formas, pero lo esencial es que en todas ellas cada número entero dentro de un cierto rango, se representa con una secuencia de bits. La cantidad de enteros que contiene dicho rango depende de n . En la tabla 1 se muestra la información que aparece en los manuales de los lenguajes de programación referida a los conjuntos numéricos de punto fijo obtenidos para diferentes cantidades de bits.

En los diferentes lenguajes estos conjuntos reciben diferentes nombres, tales como: byte, integer, word, long integer, etc.

Cantidad de bits para representar los números	Rango de valores	
	Enteros con signo	Enteros sin signo
8 (1 byte)	- 128 a 127	0 a 255
16 (2 bytes)	- 32 768 a 32 767	0 a 65 535
32 (4 bytes)	- 2 147 483 648 a 2 147 483 647	0 a 4 294 967 295

Tabla 1

Representación de números reales

Para representar internamente los números reales, se emplea la notación de punto flotante. Aunque los detalles varían de una arquitectura a otra, la idea es expresar un número real como:

$$x = \pm m \cdot 2^{\pm k}$$

Entonces, para almacenar el número x , se utiliza:

- 1 bit para guardar el signo de la mantisa.
- n bits para guardar el valor absoluto m de la mantisa
- 1 bit para guardar el signo del exponente
- q bits para guardar el valor absoluto k del exponente

Cada número real requiere de $n + q + 2$ bits para ser almacenado. Diferentes combinaciones de n y de q se han utilizado. n determina la cantidad de cifras binarias significativas de los números de la computadora, mientras q está relacionado con la magnitud del mayor número positivo y del menor número positivo que se puede representar. El total de bits para cada número toma valores muy variados dependiendo del fabricante y de la precisión (cifras significativas) que se quiera lograr. Algunos valores que se han usado son: 24, 32, 60, 64, 120.

Como los números tienen un máximo de n cifras binarias significativas, está claro que los números irracionales no pueden representarse (tienen infinitas cifras en su notación decimal), así que, otra propiedad importante se puede agregar a continuación:

Propiedad 2

Los números reales que se pueden representar internamente en la computadora son siempre racionales. ■

Por otra parte, la forma de representación en punto flotante añade una característica más a los números de la computadora. Como para cada valor del exponente $\pm k$ existen 2^n números posibles, resulta que los números reales de la computadora no están distribuidos de forma pareja en el eje real, por ejemplo, entre 4 y 8 (que corresponden con el exponente $k = 2$) hay tantos números como entre 1024 y 2048 (que corresponde al exponente $k = 10$). Esto es, los números están mucho más densamente distribuidos a medida que se acercan a cero y su distribución se enrarece a medida que crecen. En la figura 4 se trata de dar una idea gráfica de los números reales de una

computadora. Para evitar confusiones, dado que los números irracionales quedan excluidos, y todos los racionales que requieran en el sistema binario más de n dígitos, el conjunto de los números representables en la computadora será llamado Q_c . Esta propiedad es también importante:

Propiedad 3

Los números reales que se pueden representar internamente en la computadora están mucho más densamente distribuidos a medida que se acercan a cero y su distribución se enrarece a medida que crecen. ■

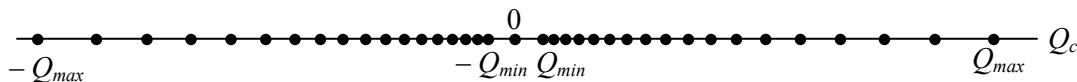


Figura 1

Aunque la propiedad 2 establece que los elementos de Q_c son racionales, nótese que no todos los racionales forman parte de Q_c ; solo aquellos cuya mantisa se puede expresar con n cifras *binarias* significativas. Este hecho suele sorprender a los principiantes, pues números con una expresión decimal muy simple, pueden no ser representables en la máquina. Un ejemplo típico es el número racional $0,1$ (es decir, $1/10$) el cual, expresado en el sistema binario, da lugar a una fracción periódica y, por tanto, no puede ser representado con una cantidad finita de dígitos binarios.

El hecho de que Q_c es un conjunto discreto (es decir, no continuo) hace que ciertas propiedades de las operaciones en el conjunto R , no se cumplan en Q_c . Así sucede con la asociatividad de la suma y del producto y la distributividad del producto respecto a la suma.

Para que se comprenda mejor esta característica, considérese una máquina hipotética que posee aritmética decimal (trabaja internamente en el sistema de base 10) y en la cual los números reales se representan mediante una mantisa de dos cifras mientras que el exponente se representa con una cifra decimal. En esta máquina, $Q_{min} = 1,0 \cdot 10^{-9}$ y $Q_{max} = 9,9 \cdot 10^9$. Considérese ahora tres elementos que pertenecen a Q_c :

$$\begin{aligned} a &= 5,0 = 5,0 \cdot 10^0 \\ b &= 0,4 = 4,0 \cdot 10^{-1} \\ c &= 3,1 = 3,1 \cdot 10^0 \end{aligned}$$

Considérense las operaciones *i*) $(ab)c$ y *ii*) $a(bc)$, que en el conjunto R dan idénticos resultados. Obsérvese qué sucede cuando se trabaja en Q_c :

$$i) ab = 5,0 \cdot 0,4 = 2,0; \quad (ab)c = 2,0 \cdot 3,1 = 6,2$$

$$ii) bc = 0,4 \cdot 3,1 = 1,2; \quad a(bc) = 5,0 \cdot 1,2 = 6,0$$

Se obtienen resultados distintos. La causa es obvia, al realizar operaciones intermedias entre números de la máquina, pueden obtenerse a veces números que no son de la máquina (aquí sucedió al multiplicar $0,4 \cdot 3,1$, cuyo verdadero resultado $1,24$ posee tres cifras significativas) y son automáticamente redondeados. Algo análogo ocurre en las computadoras reales, aunque al trabajar con más cifras significativas los errores de redondeo introducidos son mucho menores.

Como conclusión de este epígrafe, cuando elabore algoritmos que se implementarán en una computadora, tenga en cuenta las siguientes recomendaciones.

Recomendaciones

- Si en alguna operación de la máquina, se obtienen números fuera del rango $[-Q_{\max}, Q_{\max}]$, se producirá un error en la ejecución del programa, el cual se detendrá. Este tipo de error suele denominarse *overflow*.
- Los números en el intervalo $(0, Q_{\min})$, no pueden ser representados en la computadora y serán aproximados a Q_{\min} ó cero, según la arquitectura de la máquina y según el número de que se trate. Incluso, en algunas configuraciones, se produce un error de *underflow*.
- Los números reales que se encuentran en el rango permisible, es decir, que son cero o están en $[-Q_{\max}, -Q_{\min}]$ o en $[Q_{\min}, Q_{\max}]$ pueden ser tratados por la computadora, aunque, en la mayoría de los casos, sufrirán una aproximación para sustituirlos por elementos de Q_c . El error introducido en esta aproximación dependerá de la precisión de la representación numérica utilizada en el programa. Los lenguajes actuales de programación ofrecen diferentes precisiones (al menos, simple y doble precisión) para que el usuario seleccione la que estime adecuada. Tenga en cuenta al seleccionar la precisión con que trabajará, que una mayor precisión significa errores de redondeo más pequeños pero también utilizar más bits de memoria para representar a cada número real y mayor tiempo de ejecución.
- Aun cuando en un algoritmo matemático tenga sentido verificar si los números reales x y y son iguales, tenga presente que, debido a los errores de redondeo que se producen en la máquina, es casi imposible que los números x_c y y_c que contiene la memoria de la máquina puedan ser exactamente iguales. En lugar de verificar si $x = y$, verifique si

$$|x - y| < \varepsilon$$

donde ε es un número pequeño, pero grande en comparación con los errores de redondeo que pueda haber introducido la imprecisión de la máquina.

- Piense siempre que los números que almacena la máquina *no son los mismos* con los que trabaja su algoritmo manual sino aproximaciones de aquellos, aun cuando usted no haya realizado ninguna operación aritmética con ellos.

1.6 Propagación del error

Una vez que en algún paso de un algoritmo se introducen errores por una causa cualquiera, estos errores incidirán en los pasos siguientes. A este proceso se le denomina propagación del error y en esta sección se estudiarán algunas leyes básicas que permiten, en ciertos casos sencillos, comprender la forma en que ella se produce y evitar resultados indeseables.

Una ley general

Considérese el caso en que dos datos x y y se utilizan para calcular un resultado R mediante una función f conocida:

$$R = f(x, y)$$

El problema que se desea analizar es: ¿de qué forma los errores en x y y afectarán al resultado R ?

Sean x^* y y^* los valores exactos de x y y respectivamente, es decir, no afectados por el error. El valor exacto del resultado sería entonces:

$$R^* = f(x^*, y^*)$$

El error en el resultado es la diferencia:

$$\text{error}(R) = R^* - R = f(x^*, y^*) - f(x, y)$$

Si se limita el análisis al caso en que la función f es diferenciable y los errores absolutos de x y y son pequeños, se puede aproximar el incremento funcional mediante su diferencial, esto es:

$$f(x, y) - f(x^*, y^*) = f_x(x^*, y^*)(x - x^*) + f_y(x^*, y^*)(y - y^*)$$

multiplicando ambos miembros de la igualdad por -1 :

$$f(x^*, y^*) - f(x, y) = f_x(x^*, y^*)(x^* - x) + f_y(x^*, y^*)(y^* - y)$$

que se puede escribir en términos de errores como:

$$\text{error}(R) = f_x(x^*, y^*)\text{error}(x) + f_y(x^*, y^*)\text{error}(y) \quad (1)$$

Si los valores exactos x^* y y^* no se conocen, que es lo más frecuente, las derivadas parciales se pueden evaluar en los valores conocidos x y y . Se obtiene la fórmula:

$$\text{error}(R) = f_x(x, y)\text{error}(x) + f_y(x, y)\text{error}(y)$$

Si se toma valor absoluto en cada miembro queda:

$$|\text{error}(R)| = |f_x(x, y)\text{error}(x) + f_y(x, y)\text{error}(y)|$$

Como el módulo de una suma es menor o igual que la suma de los módulos:

$$|\text{error}(R)| \leq |f_x(x, y)| \cdot |\text{error}(x)| + |f_y(x, y)| \cdot |\text{error}(y)|$$

Es decir:

$$E(R) \leq |f_x(x, y)| \cdot E(x) + |f_y(x, y)| \cdot E(y)$$

Si en el segundo miembro se cambian los errores absolutos por los errores absolutos máximos, la desigualdad se satisface con mayor razón:

$$E(R) \leq |f_x(x, y)| \cdot E_m(x) + |f_y(x, y)| \cdot E_m(y)$$

Esta desigualdad significa que el miembro de la derecha puede ser una cota superior de $E(R)$ y puede tomarse como el error absoluto máximo:

$$E_m(R) = |f_x(x, y)| \cdot E_m(x) + |f_y(x, y)| \cdot E_m(y) \quad (2)$$

que se utilizará varias veces en lo que sigue. Las ecuaciones (1) y (2) se extienden sin dificultad a funciones de mayor cantidad de variables independientes.

Propagación del error en sumas y diferencias

Si el resultado R se obtiene como la suma de dos números reales (positivos o negativos)

$$R = x + y$$

las derivadas parciales de la fórmula (2) valen ambas 1 y se obtiene la ecuación:

$$E_m(R) = E_m(x) + E_m(y) \quad (3)$$

Como los números x y y pueden ser positivos o negativos, la fórmula (3) es válida para sumas o diferencias. Es decir:

$$E_m(x \pm y) = E_m(x) + E_m(y) \quad (4)$$

La fórmula (3) se puede generalizar a una suma algebraica con cualquier cantidad finita de sumandos:

$$E_m\left(\sum_{i=1}^n x_i\right) = \sum_{i=1}^n E_m(x_i) \quad (5)$$

Es decir, que el error absoluto máximo de una suma puede tomarse como la suma de los errores absolutos máximos de los sumandos.

Una vez que se conoce el error absoluto máximo del resultado, es fácil buscar el error relativo máximo, si es que se requiere.

Ejemplo 1

Los números aproximados a , b y c tienen los siguientes valores: $a = 26,868$ (con 4 cifras exactas), $b = 39,63$ (con error menor que 3%) y $c = 54,875$ con error absoluto máximo de 0,002. Determine la suma $S = a + b + c$, su error absoluto máximo, su error relativo máximo y un intervalo de seguridad para el resultado.

Solución:

Como la cifra exacta menos significativa de a es el 6, cuyo valor posicional es 0,01, se tiene que:

$$E_m(a) = 0,005$$

El error relativo máximo de b es 0,03 (3%), así que su error absoluto máximo será:

$$E_m(b) = b \cdot e_m(b) = (39,63)(0,03) = 1,189$$

Según el enunciado:

$$E_m(c) = 0,002$$

Por tanto:

$$S = 121,373$$

$$E_m(S) = 0,005 + 1,189 + 0,002 = 1,196$$

No es usual, por su carácter aproximado, dar los errores con más de dos o tres cifras significativas. En general se aproxima por exceso, para no comprometer la veracidad del resultado. Así, se tomará:

$$E_m(S) = 1,2$$

El error relativo máximo de S se puede ahora calcular como:

$$e_m(S) = \frac{E_m(S)}{S} = \frac{1,2}{121,373} = 0,0099$$

Redondeando:

$$e_m(S) = 0,01 = 1\%$$

Sumando y restando al resultado el error absoluto máximo, se obtienen aproximaciones por exceso y por defecto:

$$121,373 \pm 1,2$$

Por tanto:

$$120,173 \leq S \leq 122,573$$

■

Una consecuencia inmediata de (5) es que en una suma donde intervienen sumandos con diferente exactitud (diferentes errores absolutos máximos) el error absoluto máximo del resultado estará muy poco influenciado por el error de los sumandos más exactos (con menor error absoluto máximo) y dependerá fundamentalmente de los errores de los sumandos con mayores errores, por lo tanto, para mejorar el resultado de la suma, lo más importante es tratar de disminuir el error de los sumandos menos exactos y no preocuparse mucho de los sumandos más exactos. Así, en el ejemplo anterior, si los sumandos a y b hubieran sido números mucho más exactos, por ejemplo, con 20 cifras decimales exactas, el error absoluto máximo de S hubiera sido prácticamente el mismo.

Ejemplo 2

Se desea conocer el grosor de las paredes de un gran tanque de base circular. Ante la imposibilidad de medirlo directamente, se decide medir el diámetro interior introduciendo un operario dentro del tanque y calcular el diámetro exterior midiendo el perímetro. Los valores obtenidos fueron de $D_i = 12,36$ m y $P_e = 40,43$ m en ambos casos con un error máximo de 0,1% debido a la calidad de la cinta métrica utilizada. Calcule el grosor aproximado de las paredes del tanque y el error relativo máximo del resultado obtenido.

Solución

Como

$$P_e = \pi D_e$$

El diámetro exterior se puede obtener como: $D_e = \frac{P_e}{\pi} = \frac{40,43}{\pi} = 12,869$ m

En cuanto al cálculo del error, π puede considerarse como un número exacto ya que la división anterior se efectuó tomando π con las 31 cifras decimales exactas de una calculadora científica. Así, el error absoluto de D_e es el de P_e dividido por π .

$$E_m(D_i) = e_m(D_i) \cdot D_i = 0,001 \cdot 12,36 = 0,013 \text{ m}$$

$$E_m(P_e) = e_m(P_e) \cdot P_e = 0,001 \cdot 40,43 = 0,041 \text{ m}$$

$$E_m(D_e) = \frac{E_m(P_e)}{\pi} = \frac{0,041}{\pi} = 0,013 \text{ m}$$

El grosor aproximado x de la pared será:

$$x = \frac{D_e - D_i}{2} = \frac{12,869 - 12,36}{2} = 0,2545 \text{ m}$$

El error absoluto máximo de x se puede calcular hallando el error absoluto máximo del numerador anterior y dividiendo por 2, que es un número exacto.

$$E_m(x) = \frac{E_m(D_e - D_i)}{2} = \frac{E_m(D_e) + E_m(D_i)}{2} = \frac{0,013 + 0,013}{2} = 0,013 \text{ m}$$

$$e_m(x) = \frac{E_m(x)}{x} = \frac{0,013}{0,2545} = 0,051 = 5,1\%$$

El grosor de la pared puede estimarse en 0,2545 m con un error hasta de 5,1 % ■

Nótese algo interesante: a pesar de que en el problema anterior las mediciones se hicieron con un error máximo de 0,1%, el resultado que se obtuvo contiene un error relativo de hasta 5,1%, más de 50 veces mayor que los errores originales. Este fenómeno, que ocurre con más frecuencia de lo que se desearía, se llama *pérdida de significación* y será estudiado en la próxima sección.

Propagación del error en el producto

Considérese ahora que el resultado R se obtiene como el producto de los números aproximados x y y :

$$R = xy$$

Utilizando la ecuación (2) con $f(x, y) = xy$

$$E_m(R) = |f_x(x, y)| \cdot E_m(x) + |f_y(x, y)| \cdot E_m(y)$$

$$E_m(R) = |y| \cdot E_m(x) + |x| \cdot E_m(y) \quad (6)$$

Esta fórmula, sin embargo, es un poco complicada y aun más cuando se generaliza a más de dos factores. A partir de ella se obtiene una mucho más notable que es la que se utiliza en la práctica. Para ello, se divide en ambos miembros por $|R| = |x| \cdot |y|$ y se llega a:

$$\frac{E_m(R)}{|R|} = \frac{E_m(x)}{|x|} + \frac{E_m(y)}{|y|}$$

Es decir:

$$e_m(R) = e_m(x) + e_m(y)$$

En resumen,

$$e_m(xy) = e_m(x) + e_m(y) \quad (7)$$

La fórmula anterior se puede extender sin dificultad a un producto de más factores:

$$e_m\left(\prod_{i=1}^n x_i\right) = \sum_{i=1}^n e_m(x_i) \quad (8)$$

Es decir, el error relativo máximo de un producto de números aproximados, se puede tomar como la suma de los errores relativos máximos de los factores.

Una consecuencia inmediata de este resultado es que, cuando se multiplican números aproximados, el error relativo máximo del resultado siempre será mayor que el del factor con mayor error relativo (menos cifras exactas). Así, si tiene que multiplicar números con diferentes cantidades de cifras exactas, tenga en cuenta que las cifras exactas del resultado nunca sobrepasarán a las del factor con menos cifras exactas; si desea disminuir el error del resultado, trate de aumentar las cifras exactas de dicho factor.

Otra consecuencia inmediata de (7) es que, si uno de los factores es exacto, es decir, no contiene error, entonces, el error relativo máximo del producto es el mismo que el error relativo máximo del factor aproximado. En símbolos:

Si k es exacto:

$$e_m(kx) = e_m(x) \quad (9)$$

Ejemplo 3

Para estimar la ganancia de una cierta empresa durante el próximo año, se hace el pronóstico de la cantidad de artículos C que venderá y de la ganancia unitaria g que se obtendrá en cada producto. La ganancia se calcula como $G = Cg$. El valor de C se ha estimado por un grupo de expertos como $C = 6\,000 \pm 500$ y la ganancia unitaria como $g = (48 \pm 6)$ USD. Calcule cual es la ganancia aproximada que se obtendrá y halle su error absoluto máximo.

Solución:

Como $G = Cg$ se tiene que: $G = 6000 \cdot 48 = 288\,000$ USD

Además: $e_m(G) = e_m(C) + e_m(g) = \frac{500}{6000} + \frac{6}{48} = 0,21 = 21\%$

$$E_m(G) = G \cdot e_m(G) = 288\,000 \cdot 0,21 = 60\,480 \text{ USD}$$

Es decir, se puede pronosticar: $G = 288\,000 \pm 60\,480$ USD

Propagación del error en el cociente

En este caso, será:

$$R = \frac{x}{y}$$

donde se supone que el divisor está lo suficientemente distante de 0 como para que tanto y como y^* sean números del mismo signo.

Utilizando la ecuación (2) con $f(x, y) = \frac{x}{y}$ se obtiene:

$$E_m(R) = |f_x(x, y)| \cdot E_m(x) + |f_y(x, y)| \cdot E_m(y)$$

$$E_m(R) = \frac{1}{|y|} \cdot E_m(x) + \frac{|x|}{|y|^2} \cdot E_m(y)$$

La fórmula anterior es poco adecuada por su complejidad. Si en ambos miembros se divide por

$$|R| = \frac{|x|}{|y|} \text{ se obtiene: } \frac{E_m(R)}{|R|} = \frac{1}{|x|} \cdot E_m(x) + \frac{1}{|y|} \cdot E_m(y)$$

Es decir:

$$e_m\left(\frac{x}{y}\right) = e_m(x) + e_m(y) \quad (10)$$

Que es una forma muy fácil de recordar, sobre todo, si se observa que es idéntica a la fórmula del producto.

Ejemplo 4

Se sabe que $w = \frac{x-y}{x+y}$ y se tienen valores aproximados $x = 51,254$ y $y = 23,978$ ambos con todas sus cifras exactas. Calcular w y hallar cuales de sus dígitos son exactos.

Solución:

El algoritmo para calcular w consiste en sumas diferencias y cocientes, por tanto, con las formulas estudiadas se puede ir analizando la propagación del error. Es conveniente organizar las operaciones en una especie de tabla de tres columnas:

Valor	Error absoluto máximo	Error relativo máximo
$x = 51,254$	0,0005	
$y = 23,978$	0,0005	
$x - y = 27,276$	0,001	$\rightarrow \frac{0,001}{27,276} = 0,000037$
$x + y = 75,232$	0,001	$\rightarrow \frac{0,001}{75,232} = 0,000014$
$w = \frac{x-y}{x+y} = \frac{27,276}{75,232}$	$(0,3625585)(0,000051) =$	$\leftarrow 0,000037 + 0,000014 = 0,000051$
	$= 0,3625585$	$0,000019$

Como $E_m(w) = 0,000019$ el resultado posee cuatro cifras decimales exactas; las que le siguen son dudosas. Redondeando hasta la quinta cifra decimal: $w = 0,36256$ con cuatro cifras exactas.

Otro modo de solución:

En expresiones como esta o más complicadas, puede ser preferible utilizar la ley general (2) para calcular directamente el error absoluto máximo:

$$E_m(R) = |f_x(x, y)|E_m(x) + |f_y(x, y)|E_m(y)$$

Utilizando esta igualdad:

$$E_m(w) = \left| \frac{2y}{(x+y)^2} \right| E_m(x) + \left| \frac{-2x}{(x+y)^2} \right| E_m(y)$$

De donde:

$$E_m(w) = \left| \frac{2 \cdot 23,978}{(51,254 + 23,978)^2} \right| 0,0005 + \left| \frac{-2 \cdot 51,254}{(51,254 + 23,978)^2} \right| 0,0005$$

$$E_m(w) = 0,000014$$

Valor similar al obtenido anteriormente, un poco menor debido a que se realizaron menos redondeos intermedios. Nótese, sin embargo, que esta forma de proceder, aunque es más directa, requiere calcular derivadas y efectuar operaciones aritméticas engorrosas.

Propagación del error en la potencia y la exponencial

Como último caso particular, considérese que

$$R = x^y$$

donde $x > 0$ y y es cualquier real.

En este caso la fórmula (2):

$$E_m(R) = |f_x(x, y)| \cdot E_m(x) + |f_y(x, y)| \cdot E_m(y)$$

se convierte en: $E_m(R) = |y \cdot x^{y-1}| \cdot E_m(x) + |x^y \ln x| \cdot E_m(y) \quad (11)$

Las dos situaciones más importantes de esta fórmula son aquellos en que o bien x o bien y son exactos. A continuación se analizan ambos casos.

Si el exponente x es un número exacto, llámese k , entonces su error es cero y la fórmula (11) se transforma en:

$$E_m(x^k) = |k \cdot x^{k-1}| \cdot E_m(x)$$

Dividiendo ambos miembros por $|x^k|$ se obtiene:

$$\frac{E_m(x^k)}{|x^k|} = \frac{|k \cdot x^{k-1}|}{|x^k|} \cdot E_m(x) = |k| \frac{E_m(x)}{|x|}$$

O sea:

$$e_m(x^k) = |k| e_m(x) \quad (12)$$

Es decir, el error relativo máximo de una potencia con exponente exacto es el módulo del exponente por el error relativo máximo de la base.

En el caso en que la base es un número exacto, llámese b , se tiene una función exponencial de base b . Como el error en b es cero, la ecuación (11) conduce a:

$$E_m(b^y) = |b^y \ln b| \cdot E_m(y)$$

Si se divide en ambos miembros por b^y se obtiene:

$$\frac{E_m(b^y)}{b^y} = |\ln b| \cdot E_m(y)$$

Es decir:

$$e_m(b^y) = |\ln b| \cdot E_m(y) \quad (13)$$

El caso más importante es la exponencial de base e , para la cual resulta:

$$e_m(e^y) = E_m(y) \quad (14)$$

Ejemplo 5

Se quiere calcular la superficie exterior de un cilindro circular recto. Suponiendo que r y h se medirán con el mismo error relativo máximo, se quiere conocer cual debe ser este error para que la superficie se pueda obtener con un error inferior al 0,5%.

Solución:

Se trata de un problema inverso en que se desea saber con que error tomar los datos para obtener en el resultado un cierto error máximo. La superficie de un cilindro circular recto con radio de la base r y altura h , viene dada por:

$$S = 2\pi r^2 + 2\pi r h$$

A los efectos prácticos, puede suponerse que en el número π no se cometerán errores, ya que se puede calcular con cualquier número de cifras exactas. Aplicando la fórmula de propagación en la suma:

$$E_m(S) = E_m(2\pi r^2) + E_m(2\pi r h)$$

Como 2π es un número exacto: $E_m(S) = 2\pi E_m(r^2) + 2\pi E_m(rh)$

Los errores absolutos se pueden poner en términos de los relativos:

$$E_m(S) = 2\pi \cdot r^2 \cdot e_m(r^2) + 2\pi \cdot rh \cdot e_m(rh)$$

Teniendo en cuenta la propagación del error en la potencia y en el producto:

$$E_m(S) = 2\pi \cdot r^2 \cdot 2e_m(r) + 2\pi \cdot rh \cdot [e_m(r) + e_m(h)]$$

En el enunciado se indica suponer que los errores relativos máximos para ambos datos son iguales. Llamando $u = e_m(r) = e_m(h)$ y simplificando:

$$E_m(S) = 4\pi \cdot r^2 \cdot u + 4\pi \cdot rh \cdot u = 4\pi r(r+h) \cdot u$$

Sustituyendo $E_m(S) = S \cdot e_m(S)$ y despejando u :

$$u = \frac{S \cdot e_m(S)}{4\pi r(r+h)}$$

Tomando $S = 2\pi r^2 + 2\pi rh = 2\pi r(r+h)$

$$u = \frac{2\pi r(r+h) \cdot e_m(S)}{4\pi r(r+h)} = \frac{e_m(S)}{2} = \frac{0,005}{2} = 0,0025$$

Es decir, el radio y la altura deben medirse con error relativo máximo de $0,0025 = 0,25\%$

■

Ejercicios

1. Si $x = 23,76$; $y = 45,74$ y $z = 65,272$ todos con error relativo máximo de $0,5\%$, halle el valor de $S = x - y + z$, su error relativo máximo y la cantidad de cifras exactas que posee.
2. Se sabe que $L = xy - uz$ y se conocen valores aproximados $x = 0,5487$; $y = 6,7855$; $z = 0,07824$; $u = 2,76803$ todos con 3 cifras decimales exactas. Calcule el valor de L y determine cuántas cifras decimales exactas posee este resultado.
3. Se tiene la ecuación de segundo grado $4,3276x^2 - 9,776x + 1,8655 = 0$ y se sabe que los coeficientes están calculados con 4 cifras exactas. Determine el valor de la mayor de las raíces, aplicando la fórmula general para la ecuación de segundo grado, y diga cuáles de sus cifras son exactas.
4. Se conocen aproximadamente los tres lados de un triángulo: $a = 435,87$ (error relativo menor que $0,0002$) $b = 355,75$ (con todas sus cifras exactas) y $c = 532,31$ (con error menor que $0,1\%$). Calcule el área del triángulo mediante la fórmula $S = \sqrt{p(p-a)(p-b)(p-c)}$ donde p es el semiperímetro del triángulo y determine el error relativo máximo del resultado.
5. Calcule el valor de $\frac{34,785 - 22,873}{65,872 - 12,543}$ si todos los números son aproximados y poseen cuatro cifras exactas. Halle las cifras exactas que posee el resultado.
6. Para calcular el volumen de un cilindro se miden el radio y la altura. Se obtienen las mediciones: $r = 45,6 \pm 1$ cm y $h = 142,7 \pm 2$. Determine el volumen, su error absoluto

máximo y su error relativo máximo. Proponga que errores máximos se podría permitir si se quisiera calcular el volumen con un error absoluto dos veces menor que el obtenido.

7. Suponiendo la tierra como una esfera de unos 6400 km de radio, proponga cuántas cifras exactas tomar de π y del radio para calcular el volumen de la tierra con un error menor que 1%.
8. Para calcular el valor de $e^{-\pi}$ se utilizarán los primeros 8 términos de la serie de Maclaurin para la exponencial: $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$. Si el valor de π se toma con cuatro cifras decimales exactas, calcule $e^{-\pi}$, el error de truncamiento, el error debido al redondeo y el error total del resultado obtenido.
9. Se conoce que $x = 4,7684$ con todas sus cifras exactas. Determine el error absoluto máximo que se comete si se calcula y como:

$$\text{a) } y = \frac{3x^4 + x^2}{2x^5 - x^3} \quad \text{b) } y = \frac{3x^2 + 1}{2x^3 - x} \quad \text{c) } y = \frac{3x + \frac{1}{x}}{2x^2 - 1}$$

1.7 Errores e inestabilidad

El problema de la inestabilidad está íntimamente ligado con los errores numéricos. Se trata de un fenómeno muy importante que aparece en muchas ocasiones. En varios temas de este libro se volverá a hablar de la inestabilidad. Aquí solamente se tocarán los aspectos más generales y básicos.

Problemas estables y problemas inestables

En términos no muy precisos, se entiende por un problema estable aquel en el cual pequeños cambios en los datos producen pequeños cambios en los resultados. Por el contrario, problemas inestables son aquellos en que pequeños cambios en los datos pueden causar grandes cambios en los resultados. En determinados tipos de problemas estos conceptos se pueden hacer más precisos, e incluso se puede medir cuantitativamente la inestabilidad. Por el momento, solo se quiere enfatizar en el concepto. Para hacerlo más claro, se han incluido los dos ejemplos que siguen:

Ejemplo 1

Sea la ecuación algebraica de grado 10:

$$(x-1)(x-2)\cdots(x-10)=0$$

cuyos ceros son todos reales: $x = 1, x = 2, \dots, x = 10$. Si el producto indicado se efectúa se obtiene:

$$x^{10} - 55x^9 + \cdots + 3628800 = 0$$

Considérese ahora el problema consistente en hallar las raíces de la ecuación tomando como datos sus coeficientes. Si el coeficiente de x^9 se cambia en 0,001 (lo cual representa un error relativo de 0,00002), se obtiene la ecuación:

$$x^{10} - 55,001x^9 + \dots + 3628800 = 0$$

Los ceros reales de esta ecuación, calculados con 15 cifras exactas, son:

$$\begin{aligned}x &= 1,000\,000\,006\,008\,28 \\x &= 2,000\,012\,150\,455\,79 \\x &= 2,998\,069\,854\,683\,42 \\x &= 4,075\,898\,053\,001\,94 \\x &= 4,616\,487\,711\,668\,10 \\x &= 10,809\,887\,979\,566\,3\end{aligned}$$

Las otras cuatro raíces son complejas. Como se ve, los ceros no solo cambiaron en forma significativa (algunos hasta en un 8%) sino que cuatro de ellos desaparecieron como raíces reales. Se trata, evidentemente, de un problema sumamente inestable.

Ejemplo 2

Entre problemas tan simples como resolver un sistema lineal de dos ecuaciones con dos incógnitas, se pueden encontrar problemas inestables. Considérese el sistema de ecuaciones:

$$\begin{cases} x + y = 3 \\ x + 1,01y = 3,01 \end{cases}$$

cuya solución puede fácilmente comprobarse que es $x = 2$ y $y = 1$. En este problema, los datos son los 6 coeficientes de las ecuaciones y el resultado los valores de x y y que forman la solución.

Nótese cómo el simple cambio del coeficiente 3,01 por 3,02 (un cambio inferior al 1%) transforma el sistema en

$$\begin{cases} x + y = 3 \\ x + 1,01y = 3,02 \end{cases}$$

cuya solución es $x = 1$ y $y = 2$. El resultado sufrió cambios del orden de 100%. Se trata obviamente de un problema muy inestable. ■

Si se utilizaran exclusivamente números exactos, la inestabilidad de un problema no tendría mayor importancia. Pero ese no es el caso. Los errores por redondeo, por truncamiento y por medición están presentes constantemente y se propagan a lo largo de todos los algoritmos matemáticos; cuando un error alcanza a los datos de un problema inestable, este error se amplifica repentinamente y, a partir de ahí, este error desproporcionado contamina todos los pasos siguientes del algoritmo.

Entonces, ¿qué hacer con los problemas inestables? Primeramente, detectarlos. Una vez descubiertos, tratar de sustituir el modelo matemático por otro que no sea inestable; si esto no es posible, minimizar los errores en sus datos, por ejemplo, con mediciones más exactas, utilizando más cifras exactas en los números, etcétera.

Cuando se realizan algoritmos generales, que se deben ejecutar con diferentes juegos de datos, el problema es más complicado, ya que en muchos casos la inestabilidad se presenta solo para determinados juegos de datos. Por ejemplo, los sistemas de dos ecuaciones lineales con dos incógnitas solo pueden ser problemas inestables cuando el determinante de la matriz del sistema es pequeño en comparación con la magnitud de los coeficientes. Sin embargo, no siempre se puede prever esta inestabilidad condicional.

Pérdida de significación

Uno de los problemas condicionalmente inestables mejor conocidos se presenta al restar números reales o, en forma más general, al efectuar sumas de números positivos y negativos. Este tipo de problemas ya apareció en el ejemplo 2 de la sección 1.6. En ese ejemplo se calculó el grosor de la pared de un tanque cilíndrico a partir de los diámetros exterior e interior del tanque; los datos del problema contenían un error máximo de 0,1% y el resultado se obtuvo con error máximo de más de 5%, 50 veces mayor que el de los datos.

En el caso de la resta de dos números reales, este problema se presenta cuando los números son muy similares (es el caso de los dos diámetros del tanque). Sea por ejemplo:

$$d = x - y \quad (1)$$

Como se sabe, el error absoluto máximo de d viene dado por:

$$E_m(d) = E_m(x) + E_m(y)$$

El error relativo máximo puede obtenerse dividiendo por d y tomando en cuenta (1):

$$e_m(d) = \frac{E_m(x) + E_m(y)}{|x - y|}$$

Resulta claro que, si x y y son similares, el denominador de este cociente se hace pequeño y el error relativo crece y alcanza valores tanto más grandes cuanto más cercanos sean x y y . El nombre de pérdida de significación proviene del hecho de que, al calcular la diferencia de dos números muy similares, la cantidad de dígitos significativos exactos se reduce considerablemente, lo cual equivale a un aumento del error relativo.

Ejemplo 3

Sean $x = 3,2548547$ y $y = 3,2546675$, ambos números aproximados con cinco cifras exactas. Calcule $d = x - y$, su error absoluto máximo, su error relativo máximo y la cantidad de cifras exactas.

Solución:

$$d = 3,2548547 - 3,2546675 = 0.0001872$$

Como ambos números tienen exactas las primeras cuatro cifras decimales, su error absoluto máximo es 0,00005:

$$E_m(d) = E_m(x) + E_m(y) = 0,00005 + 0,00005 = 0,0001$$

Nótese que el error absoluto no ha crecido demasiado. El problema está en el error relativo:

$$e_m(d) = \frac{E_m(d)}{|d|} = \frac{0,0001}{0,0001872} = 0,534$$

Es decir, más de un 53% de error relativo máximo. Compárese con el error relativo de los datos que era menor que 0,00002. Este mismo efecto se aprecia analizando las cifras exactas: los datos tenían cinco cifras exactas; el resultado 0,0001872 no posee ninguna cifra significativa exacta, pues la primera cifra significativa, que es la cuarta cifra decimal, está afectada por el error absoluto 0,0001.

■

En el caso de algoritmos donde se suman varios números de diferentes signos, el problema se suele hacer inestable cuando los datos conducen a una suma próxima a cero, debido a que en el error absoluto se suman los errores absolutos de los sumandos y el error relativo se hace muy grande al dividir por una suma muy pequeña.

Ejemplo 4

Como se sabe, para cualquier x real (o complejo) la serie:

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

converge hacia e^x . Este hecho se puede utilizar para calcular valores de esta función de manera sencilla. En particular, cuando x es negativa, la serie se hace alterna y puede aplicarse el teorema de Leibniz para acotar el error de truncamiento que se produce; si se trunca la serie en el término de exponente n puede tomarse como error absoluto máximo el término de exponente $n+1$, esto es:

$$\text{para } x < 0, \quad S = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!} \quad \text{con} \quad E_m(S) = \frac{|x^{n+1}|}{(n+1)!}$$

En un algoritmo muy sencillo, la suma S se inicia en 1 y se van calculando y agregando nuevos sumandos hasta llegar a alguno que sea menor que una cierta tolerancia, el cual ya no es necesario sumar; con ello se obtiene el valor de la exponencial con un error de truncamiento menor que la tolerancia que se haya tomado. A continuación se muestra el algoritmo en detalle. Como es muy simple se ha utilizado el seudo código que se explicará en el próximo epígrafe. Se ha tomado una tolerancia de 0,000005, de modo que el error de truncamiento no afecte la quinta cifra decimal del resultado.

```

Leer  $x < 0$ 
 $n := 1$ ,  $Suma := 1$  y  $Sumando := 1$ 
do while  $|Sumando| > 0,000005$ 
     $Sumando := Sumando \cdot x/n$ 
     $Suma := Suma + Sumando$ 
end
Mostrar  $Suma$ 

```

Para valores de x próximos a cero, el algoritmo muestra resultados esperados. Sin embargo, a medida que x se aleja, van apareciendo errores importantes en el resultado. Por ejemplo, para $x = -9$ el resultado del algoritmo es 0,000179 cuando realmente $e^{-9} = 0,000123$. Aunque los cálculos fueron realizados utilizando una máquina con 7 u 8 cifras exactas, para $x = -9$, el resultado no posee ninguna cifra significativa exacta. Un análisis más detallado muestra que, para $x = -9$ los valores de los primeros 11 sumandos que forman el resultado son (todos con 8 cifras exactas):

$$\begin{array}{r}
 1,000\,000\,0 \\
 - 9,000\,000\,0 \\
 40,500\,000 \\
 - 121,500\,00 \\
 273,375\,00 \\
 - 492,075\,00 \\
 738,112\,50 \\
 - 949,001\,79 \\
 1067,627\,0 \\
 - 1067,627\,0 \\
 960,364\,31
 \end{array}$$

En particular, algunos de los sumandos poseen solamente cuatro cifras decimales exactas, lo cual representa un error absoluto máximo de 0,00005. Tan solo el noveno y décimo sumandos pueden producir un error conjunto de 0,0001 que ya es del orden de la suma que se debería obtener. En este caso se ha producido una pérdida de significación por realizar una suma de números positivos y negativos cuyo resultado es pequeño en relación con los errores (de redondeo) que contienen algunos de los sumandos.

Métodos inestables para problemas estables

A veces para resolver un problema se recurre a otro problema más sencillo cuyo resultado coincide o se aproxima mucho; a la solución de este nuevo problema se le llama un método de solución del primero. Así por ejemplo, el problema de medir el grosor de una pared se sustituye por el de medir dos diámetros y restarlos o el de calcular una exponencial se sustituye por el de sumar algunos términos en una serie. Por supuesto que, si el problema original es inestable, cualquier método que se utilice para resolverlo será también un problema inestable. La situación más sorprendente sucede cuando, para resolver un problema estable se introduce un método que constituye en sí mismo un problema inestable. Eso es lo que ha sucedido con los dos ejemplos citados: medir el grosor de una pared es un problema estable pero el método de restar los dos diámetros es un problema inestable; calcular e^{-9} es un problema estable, pero hacerlo con la serie alterna es un método inestable. En estos casos, la solución consiste en buscar otro método que sí sea estable. Por ejemplo, el grosor de la pared se podría medir haciendo una pequeña perforación por la que se introduzca una varilla que después se mide; el valor de e^{-9} se puede calcular multiplicando e nueve veces por si mismo y hallando después su recíproco.

A continuación se incluyen dos ejemplos más de problemas estables resueltos primeramente por métodos inestables y después por procedimientos estables.

Ejemplo 5

Resuelva la ecuación $x^2 - 6x + 0,001 = 0$ utilizando en los cálculos 5 cifras exactas.

Solución 1

La fórmula usual para resolver la ecuación de segundo grado: $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ contiene en el numerador una suma y una diferencia. Para la raíz que se obtenga sumando el algoritmo es estable; cuando se halla la raíz que requiere restar y sucede que el producto $4ac$ es muy pequeño, entonces el numerador es una diferencia de números aproximados muy similares y se produce pérdida de significación. En este caso, la aplicación de la fórmula conduce a:

$$x_1 = \frac{6 + \sqrt{6^2 - 0,004}}{2} \quad y \quad x_2 = \frac{6 - \sqrt{6^2 - 0,004}}{2}$$

Trabajando con 5 cifras exactas, se obtiene para x_1 :

$$x_1 = \frac{6 + \sqrt{35,996}}{2} = \frac{6 + 5,9997}{2} = 5,9999$$

y para x_2 : $x_2 = \frac{6 - \sqrt{35,996}}{2} = \frac{6 - 5,9997}{2} = 0,00015$

Nótese que para x_2 solamente han quedado 2 cifras significativas, de las cuales como se verá enseguida, solo una es exacta.

Solución 2

El valor de x_1 , donde no hubo pérdida de significación se calcula del mismo modo. En cambio para calcular x_2 se procede así:

$$x_2 = \frac{6 - \sqrt{35,996}}{2} = \frac{(6 - \sqrt{35,996})(6 + \sqrt{35,996})}{2(6 + \sqrt{35,996})} = \frac{36 - 35,996}{2(11,999)} = \frac{0,004}{23,998} = 0,00016668$$

que posee 5 cifras exactas. ■

El próximo ejemplo ha sido tomado de “Computer Methods for Mathematical Computations” de Forsythe et al.

Ejemplo 6

Calcular la integral definida $\int_0^1 x^9 e^{x-1} dx$

Solución 1:

Primero el problema se generaliza de la siguiente manera:

$$I_n = \int_0^1 x^n e^{x-1} dx \quad \text{para } n \geq 0$$

Es claro que lo que se desea es calcular I_9 . Para $n = 1$ la integral se calcula fácilmente por partes:

$$I_1 = \int_0^1 x e^{x-1} dx = x e^{x-1} \Big|_0^1 - \int_0^1 e^{x-1} dx = 1 - e^{-1} \Big|_0^1 = e^{-1} = 0.367879 \quad (2)$$

Si se aplica la fórmula de integración por partes a I_n , se obtiene:

$$\begin{aligned} \text{para } I_n &= \int_0^1 x^n e^{x-1} dx = x^n e^{x-1} \Big|_0^1 - \int_0^1 n x^{n-1} e^{x-1} dx \\ I_n &= \int_0^1 x^n e^{x-1} dx = 1 - n \int_0^1 x^{n-1} e^{x-1} dx \end{aligned}$$

Esto es:

$$I_n = 1 - n I_{n-1} \quad \text{para } n = 2, 3, 4, \dots \quad (3)$$

Como I_1 es conocido, la aplicación reiterada de la fórmula recursiva (3), permite calcular, mediante dos operaciones aritméticas en cada paso, los valores de I_2, I_3, I_4, \dots hasta llegar a I_9 , que es el valor deseado. Al realizar los cálculos con un programa de computadora se obtuvo:

$$I_9 = -0,06848$$

Resultado completamente absurdo, ya que, por ser el integrando positivo en el intervalo de integración, el resultado de la integral debe ser positivo.

El desastre de este elegante procedimiento, está en que se trata de un algoritmo inestable. Obsérvese que en cada paso iterativo, el valor de la integral precedente se multiplica por el valor correspondiente de n . Sin la presencia de errores, esto no traería dificultades pero el pequeño error que contiene I_1 , cuyo error absoluto máximo es $0,5 \cdot 10^{-6}$ es multiplicado más y más en cada paso, primero por 2, después por 3, después por 4 y finalmente por 9. Es decir, ese pequeño error ha sido amplificado $9! = 362\,880$ veces y se ha producido un error final que supera al valor que se deseaba calcular. Este elegante procedimiento, lamentablemente es muy inestable.

Solución 2:

La ecuación (3) puede ser escrita de otra manera, si se despeja I_{n-1} :

$$I_{n-1} = \frac{1 - I_n}{n} \quad \text{para } n = 2, 3, 4, \dots \quad (4)$$

Ahora, si se conoce, por ejemplo, I_{20} la ecuación (4) permite calcular sucesivamente $I_{19}, I_{18}, I_{17}, \dots$ hasta obtener finalmente I_9 . Lo más importante es que este algoritmo es sumamente estable. El error inicial que existe en I_{20} quedará dividido por 20, después por 19, después por 18, etc. y es obvio que, al llegar a I_9 , el error inicial, a todos los efectos prácticos, habrá desaparecido.

Tomando para I_{20} cualquier valor, y trabajando en todos los pasos con 6 dígitos exactos, se obtiene para I_9 :

$$I_9 = 0,091\,612$$

que tiene todas sus cifras exactas.

Ejercicios

1. La ecuación $\sin x = kx$ para $k < 1$ no se puede resolver por métodos exactos. Para valores de k próximos a 1 se muestran a continuación la raíz positiva de la ecuación con 4 cifras decimales exactas (en el capítulo 2 se verá cómo hacerlo).

Ecuación:	Raíz positiva:
$\sin x = 0,999x$	$x = 0,0775$
$\sin x = 0,998x$	$x = 0,1096$
$\sin x = 0,997x$	$x = 0,1342$
$\sin x = 0,996x$	$x = 0,1550$
$\sin x = 0,995x$	$x = 0,1733$

Como se observa, pequeños cambios en el coeficiente k conducen a cambios casi 20 veces mayores en la solución. Grafique las funciones $y = \sin x$ y las rectas $y = kx$ para diferentes valores de k y explique a qué se debe la inestabilidad del problema.

2. A continuación se muestra un sistema lineal de ecuaciones con su solución y otro sistema lineal ligeramente cambiado, también con su solución. Diga si se trata de un problema inestable. En caso afirmativo, explique cual es la causa de la inestabilidad.

Sistema original:

$$\begin{cases} 30x_1 + 15x_2 + 10x_3 = 55 \\ 15x_1 + 10x_2 + 7.5x_3 = 32.5 \\ 10x_1 + 7.5x_2 + 6x_3 = 23.5 \end{cases}$$

Sistema modificado:

$$\begin{cases} 30x_1 + 15x_2 + 10x_3 = 55 \\ 15x_1 + 10x_2 + 7.5x_3 = 32.5 \\ 10x_1 + 7.5x_2 + 6x_3 = 23.4 \end{cases}$$

Solución: $x_1 = 1$
 $x_2 = 1$
 $x_3 = 1$

Solución: $x_1 = 0,9$
 $x_2 = 1,6$
 $x_3 = 0,4$

3. La función $f(x) = \frac{x \cos x - \sin x}{x^3}$ no está definida para $x = 0$ pero debe tender a $-\frac{1}{3}$ cuando x tiende hacia cero. Evalúela para valores cercanos a $x = 0$, mediante su calculadora o con un asistente matemático, y observe lo que ocurre cuando le asigna a x valores del orden de 0,00001. Explique el comportamiento que observe.
4. Exprese la función $f(x)$ del ejercicio anterior de una forma más conveniente para valores muy pequeños de x . (Sugerencia: exprese las funciones $\sin x$ y $\cos x$ mediante sus series de Maclaurin y simplifique la expresión teniendo en cuenta que x tomará valores muy pequeños.

5. Muestre que la función $g(x) = \frac{\ln(1-x) + xe^{\frac{x}{2}}}{x^3}$ presenta pérdida de significación cuando se evalúa para valores de x cercanos a cero, a pesar de que aparentemente, no se presentan diferencias de números parecidos. Proponga una forma alternativa estable para evaluar la función para valores de x próximos a cero.
6. Para trazar en el display de una computadora personal la recta tangente a la gráfica de la función $y = x^7$ en el punto de abscisa $x = 4$, se calcula aproximadamente el valor de la pendiente de la recta como:

$$f'(4) \approx \frac{(4+h)^7 - 4^7}{h}$$

para valores pequeños de h . Según la teoría, mientras menor sea h en valor absoluto mejor será la aproximación. Sin embargo, utilice una calculadora o una computadora y emplee la aproximación anterior para calcular $f'(4)$ para valores de h : 0,001; 0,00001; 0,0000001. Explique el comportamiento observado.

7. En el ejemplo 4 se observó que la serie de Maclaurin de la función e^x presenta inestabilidad para valores de x negativos alejados de cero. Proponga un algoritmo que permita evaluar dicha función para valores negativos de cualquier tamaño sin problemas de pérdida de significación. (Sugerencia: Considere $x = -n - q$ donde n es entero positivo y $0 < q < 1$, entonces $e^x = e^{-n} \cdot e^{-q}$).

1.8 Seudo código para la escritura de algoritmos

A lo largo de este libro serán estudiados muchos algoritmos numéricos. Para expresarlos claramente y sin ambigüedades será utilizado el seudo código que se describe a continuación. Utilizar un seudo código en lugar de un lenguaje completamente formal (Como Fortran, Basic, Pascal, C, Matlab, etc.) tiene varias ventajas: por una parte, un seudo código no está sujeto a reglas de sintaxis tan estrictas y no hay que sacrificar la preferencia del lector por un lenguaje u otro. Por otra parte, el seudo código que se utilizará es cercano a la mayoría de los lenguajes de computación en uso, de manera que no será difícil, para el lector interesado, traducir los algoritmos al lenguaje que deseé.

El operador de asignación

Se utilizará el signo $:=$ para indicar que la expresión que aparezca a la derecha debe ser asignada a la variable que aparece a la izquierda. Por ejemplo, $\text{Área} := 45$ significa que a la variable llamada Área se le debe asignar el valor 45; $n := n - 1$, indica que a la variable llamada n debe asignársele el valor que tome la expresión $n - 1$. Observe que este símbolo indica una orden, no una relación de igualdad.

La estructura alternativa

Se utiliza para indicar que, si se cumple una condición, se ejecute una secuencia de acciones y en caso contrario, se ejecute otra. Su estructura general es la siguiente:

```

If <condición> then
    <secuencia 1 de acciones>
else
    <secuencia 2 de acciones>
end

```

A veces, la secuencia 2 de acciones no se necesita y, en ese caso se omite la palabra **else**.

Estructuras repetitivas

Sirven para que se ejecute repetidamente una secuencia de acciones. La secuencia se repite hasta que se satisface una cierta condición. Se emplearán tres tipos de estructuras repetitivas: **do – while**; **repeat – until** y **for**.

La estructura **do – while** tiene la forma:

```

do while <condición>
    <secuencia de acciones>
end

```

La secuencia de acciones se ejecuta una y otra vez mientras la condición siga siendo cierta. Tan pronto como deje de cumplirse la condición, la ejecución del algoritmo pasa a la instrucción que siga a la palabra **end**. A diferencia de la estructura **do – while**, en la cual la se analiza la veracidad de la condición como requisito previo para entrar dentro del ciclo, la estructura **repeat – until** permite entrar directamente en la secuencia de acciones correspondiente y, después de ejecutadas estas acciones, se analiza la validez de una condición para permitir o no repetir la secuencia de acciones.

La estructura **repeat – until** tiene la forma:

```

repeat
    <secuencia de acciones>
until <condición>

```

Como se observa, aquí no se necesita la palabra **end** para indicar el final de la secuencia de acciones, ya que la palabra **until** realiza esta función. Después de ejecutadas las acciones de la secuencia se analiza si la condición es cierta y, si lo es, se pasa a ejecutar la instrucción que siga a la palabra **until**. En otras palabras, la secuencia de acciones se ejecuta una y otra vez, hasta que la condición se cumple.

La estructura repetitiva **for**, permite repetir una secuencia de acciones un número de veces que está previamente fijado, bajo el control de una variable que toma valores enteros consecutivos desde un número inicial hasta un número final, ambos prefijados.

La estructura **for** tiene la forma:

```

for <variable> = <valor inicial> to <valor final>
    <secuencia de acciones>
end

```

Realmente las tres estructuras repetitivas no son imprescindibles. Un mismo algoritmo, por lo general, puede expresarse utilizando una u otra. El objetivo de utilizar las tres formas es buscar en cada caso mayor claridad y brevedad.

A continuación se muestran ejemplos de algoritmos escritos con el seudo código adoptado.

Ejemplo 1

Escriba un seudo código para obtener las raíces reales de una ecuación de segundo grado:

$$ax^2 + bx + c = 0 \quad (a \neq 0)$$

utilizando la fórmula general.

Solución:

Como se sabe, si el discriminante $D = b^2 - 4ac$ es negativo, las raíces son imaginarias. En caso contrario las raíces son reales y se obtienen como:

$$x_1 = \frac{-b + \sqrt{D}}{2a} \quad \text{y} \quad x_2 = \frac{-b - \sqrt{D}}{2a}$$

El algoritmo de cálculo será como sigue:

Se supone conocidos los coeficientes de la ecuación: a , b y c

if $a \neq 0$ **then**

$$D := b^2 - 4ac$$

if $D \geq 0$ **then**

$$x_1 := \frac{-b + \sqrt{D}}{2a}$$

$$x_2 := \frac{-b - \sqrt{D}}{2a}$$

else

No hay raíces reales

end

else

La ecuación no es de segundo grado

end

Ejemplo 2

Escriba un algoritmo en seudo código para evaluar una función $y = f(x)$ en n puntos igualmente espaciados de un intervalo $[a, b]$ de su dominio, siendo a el primer punto y b el último.

Solución:

Los n puntos dividen al intervalo $[a, b]$ en $n - 1$ partes iguales, cada una de longitud h , dada por

$$h = \frac{b-a}{n-1}$$

Para obtener los n puntos $a = x_1, x_2, x_3, \dots, x_n = b$ se utilizará la fórmula:

$$x_i = a + (i-1)h \quad \text{para } i = 1, 2, 3, \dots, n$$

El algoritmo de cálculo adoptará la forma:

Se supone conocidos la función f y los números a, b y n

$$h := \frac{b-a}{n-1}$$

for $i = 1$ **to** n

$$x_i := a + (i-1)h$$

$$y_i := f(x_i)$$

end

Ejemplo 3

Realice en seudo código un algoritmo que permita obtener los elementos de una progresión aritmética de incremento $d > 0$ y valor inicial a , que sean menores o iguales que un cierto valor conocido x_{max} .

Solución:

Los elementos de la progresión se obtendrán mediante la fórmula:

$$\begin{aligned} x_k &= x_{k-1} + d \quad \text{para } k = 2, 3, 4, \dots \\ \text{con} \quad x_1 &= a \end{aligned}$$

El algoritmo de cálculo puede expresarse así:

Se supone conocidos a, d y x_{max}

$$k := 1$$

$$x_1 := a$$

do while $x_k \leq x_{max} - d$

$$k := k + 1$$

$$x_k = x_{k-1} + d$$

end

Ejemplo 4

Una cierta sucesión convergente está definida mediante la fórmula recurrente:

$$\begin{aligned} x_0 &= a \\ x_i &= \phi(x_{i-1}) \quad \text{para } i = 1, 2, 3, \dots \end{aligned}$$

Se desea hallar aproximadamente el límite L de la sucesión. Desarrolle un algoritmo que genere valores de la sucesión hasta llegar a dos valores sucesivos que difieran en menos que un número positivo pero muy pequeño, ε . Se tomará el límite como el último valor hallado.

Solución

Aunque el algoritmo también podría escribirse utilizando la estructura **do – while**, se utilizará **repeat – until**, con la cual, en este caso, resulta más claro. Nótese que no es necesario recordar todos los valores de la sucesión, por ello se guarda solamente el último, en la variable $x_{anterior}$. El algoritmo que resulta es:

Se supone conocidos la función ϕ y los números a y ε

```

 $x_{anterior} := a$ 
repeat
     $x_{actual} := \phi(x_{anterior})$ 
     $Diferencia := |x_{actual} - x_{anterior}|$ 
     $x_{anterior} := x_{actual}$ 
until  $Diferencia \leq \varepsilon$ 
 $L := x_{actual}$ 

```

Ejercicios

En cada uno de los problemas que siguen, elabore un algoritmo en seudo código que lleve a cabo la tarea pedida.

1. Evaluar la función $f(x)$ que se da a continuación en un punto x cualquiera.

$$f(x) = \begin{cases} \sqrt{-x} & \text{si } x \leq 0 \\ \frac{1}{x} & \text{si } 0 < x < 100 \\ \ln(x-90) & \text{si } x \geq 100 \end{cases}$$

2. Obtener todos los términos x_n de una progresión geométrica de razón $r > 0$ y primer término $a > 0$ que sean menores que un valor $M > a$. Los valores de a , r y M son conocidos.
3. Evaluar un polinomio $p(x)$ de grado n escrito en la forma $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$. Los coeficientes del polinomio, el grado n y el valor de x son conocidos.
4. Repita el ejercicio 3 pero ahora con el polinomio escrito en la forma de Horner:

$$a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + xa_n) \cdots))$$

Analice la ventaja de efectuar las operaciones de esta manera.

5. Efectuar el producto escalar de dos vectores de n componentes: $x = [x_1, x_2, \dots, x_n]$ y $y = [y_1, y_2, \dots, y_n]$. Tanto n como las componentes de los vectores son conocidos.

6. Multiplicar una matriz cuadrada $A = \{a_{ij}\}$ de orden n por una matriz columna $B = \{b_i\}$ de orden n . Tanto n como las componentes de las matrices son conocidos.
7. Dadas dos matrices $A = \{a_{ij}\}$ de orden $p \times n$ y $B = \{b_{ij}\}$ de orden $q \times m$, determinar si son conformes para el producto. En caso negativo escribir: “Operación imposible” y en caso positivo hallar la matriz $A \cdot B$. Los números p, n, q, m y los elementos de ambas matrices son conocidos.

Otras lecturas recomendadas

Respecto a temas históricos (en particular, los métodos numéricos) el libro “Historia de las Matemáticas” de K. Ríbnikov, editorial Mir, 1987, está muy completo y lleno de detalles interesantes. La traducción del ruso al español posee una gran calidad. Mucho más ameno, pero poco abarcador es el clásico “Grandes Matemáticos” de H. W. Turnbull que se ha editado en Cuba por la Editorial Científico Técnica y que con más de 50 años de escrita, no ha perdido su actualidad y belleza.

Acerca de la teoría de errores y el cálculo con números aproximados la exposición que se hace en “Computational Mathematics” de B. P. Demidovich e I. A. Maron es una de las mas amplias y completas. En Cuba es bastante conocida la traducción al inglés de la Editorial Mir. Existe una traducción al español de la editorial Aguilar pero muy poco difundida en Cuba.

Los temas relacionados con la inestabilidad numérica y su influencia en los algoritmos computacionales está muy bien tratado en el libro “Computer Methods for Mathematical Computations” de G. E. Forsythe, M. A. Malcolm y C. B. Moler, de la Editorial Prentice – Hall, 1977, en idioma inglés. También se abordan con profundidad en “An Introduction to Numerical Analysis” de K. E. Atkinson, Editorial John Wiley and Sons, 1989 y a un nivel más sencillo en “Elementary Numerical Analysis”, del mismo autor y la misma editorial, publicado en 1993, ambos en idioma inglés.

Principales ideas del capítulo

- La Matemática Numérica tiene como propósito el desarrollo de métodos para la solución de los más diversos problemas matemáticos mediante una cantidad *finita* de operaciones *numéricas*.
- La Matemática Numérica no se plantea llegar a resultados exactos; ni siquiera a resultados tan exactos como sea posible. El propósito aquí será obtener resultados tan exactos como sea *necesario*.
- Prescindir de la exactitud absoluta, permite a la Matemática Numérica elaborar métodos mucho más generales que los métodos analíticos exactos.
- La computadora digital ha hecho posible la aplicación práctica de muchos métodos numéricos, que con el trabajo en forma manual, solo tendrían un valor teórico. Por otra parte, las computadoras digitales han traído la necesidad de desarrollar nuevos métodos numéricos para dar respuesta a nuevos problemas.
- Un método aproximado solo tiene valor si permite, de alguna forma, tener una estimación de la magnitud del error que se comete con su aplicación.

- Un modelo matemático no puede, ni debe, reflejar exactamente el mundo real sino sólo los aspectos de aquel que resultan importantes en el problema que se desea resolver, de acuerdo con el uso que se dará a los resultados obtenidos.
- La mayoría de los métodos exactos solamente se aplican a situaciones muy simples y específicas que raras veces se dan en los problemas reales.
- El error que se introduce en el proceso debido a la no exactitud del método de solución empleado se suele llamar *error de truncamiento*.
- A diferencia de las equivocaciones ante las cuales todo lo que se puede hacer es tratar de evitarlas, con los errores de redondeo hay que aprender a convivir.
- El *error de x* en relación con el valor exacto x^* se denota $\text{error}(x)$ y se define como la diferencia: $\text{error}(x) = x^* - x$
- El *error absoluto de x* en relación con el valor exacto x^* se denota $E(x)$ y se define como $E(x) = |\text{error}(x)|$. El *error relativo de x* en relación con el valor exacto $x^* \neq 0$ se denota $e(x)$ y se define como $e(x) = \frac{E(x)}{|x^*|}$
- El *error absoluto máximo de x* en relación con x^* se denota $E_m(x)$ y se define como cualquier número real que satisfaga la condición: $E_m(x) \geq E(x)$. El *error relativo máximo de x* en relación con x^* se denota $e_m(x)$ y se define como cualquier número real que satisfaga la condición: $e_m(x) \geq e(x)$. Están ligados por la relación: $E_m(x) = |x^*|e_m(x)$
- El error absoluto máximo permite acotar a x^* : $x - E_m(x) \leq x^* \leq x + E_m(x)$
- Si el dígito d ocupa en un número real la posición k -sima se denota el *valor posicional de d* como $p(d)$ y se define como $p(d) = 10^k$
- Cuando un dígito 0 se incluye en un número con el único propósito de ocupar una posición dentro del número, ese dígito se llaman cero no significativo. En los demás casos, se dice que el 0 es significativo. Todos los dígitos que no son 0 son significativos.
- Un dígito d de un número x se dice que es un *dígito exacto* o una *cifra exacta* si el error absoluto de x es menor o igual que la mitad del valor posicional de d . Esto es, si $E(x) \leq \frac{1}{2} p(d)$. En caso contrario, la cifra d se dice que no es exacta.
- La *cantidad de cifras exactas* de un número aproximado es la cantidad de dígitos *significativos* exactos de dicho número. La *cantidad de cifras decimales exactas* de un número aproximado es la cantidad de cifras exactas que están después de la coma decimal.
- Se acostumbra redondear los números aproximados conservando una o dos de sus cifras no exactas (o dudosas).
- El error absoluto máximo está vinculado con las cifras decimales exactas: un número aproximado posee k cifras decimales exactas si y solo si su error absoluto es menor o igual que $\frac{1}{2} 10^{-k}$
- El error relativo de un número no depende de la magnitud del número sino de la cantidad de cifras exactas
- Los conjuntos numéricos que utiliza cualquier computadora son finitos y acotados inferior y superiormente.
- Los números reales que se pueden representar internamente en la computadora son siempre racionales y están mucho más densamente distribuidos a medida que se acercan a cero y su distribución se enrarece a medida que crecen.
- El hecho de que \mathbb{Q}_c es un conjunto discreto (es decir, no continuo) hace que ciertas propiedades de las operaciones en el conjunto \mathbb{R} , no se cumplan en \mathbb{Q}_c . Así sucede con la asociatividad de la suma y del producto y la distributividad del producto respecto a la suma.

- Los números reales que se encuentran en el rango permisible, es decir, que son cero o están en $[-Q_{max}, -Q_{min}]$ o en $[Q_{min}, Q_{max}]$ pueden ser tratados por la computadora, aunque, en la mayoría de los casos, sufrirán una aproximación para sustituirlos por elementos de Q_c
- En el trabajo con números reales en una máquina, en lugar de verificar si $x = y$, verifique si $|x - y| < \varepsilon$ donde ε es un número pequeño, pero grande en comparación con los errores de redondeo que pueda haber introducido la imprecisión de la máquina.
- Si $R = f(x, y)$ entonces $E_m(R) = |f_x(x, y)| \cdot E_m(x) + |f_y(x, y)| \cdot E_m(y)$
- Las leyes principales que rigen la propagación de errores son: $E_m(x \pm y) = E_m(x) + E_m(y)$
 $e_m(xy) = e_m(x) + e_m(y); e_m\left(\frac{x}{y}\right) = e_m(x) + e_m(y); e_m(x^k) = |k|e_m(x); e_m(e^y) = E_m(y)$
- Se entiende por un problema inestable aquel en el cual pequeños cambios en los datos producen grandes cambios en los resultados. Cuando un error alcanza a los datos de un problema inestable, este error se amplifica repentinamente y, a partir de ahí, este error desproporcionado contamina todos los pasos siguientes del algoritmo.
- En el caso de algoritmos donde se suman varios números de diferentes signos, el problema se suele hacer inestable cuando los datos conducen a una suma próxima a cero, debido que en el error absoluto se suman los errores absolutos de los sumandos y el error relativo se hace muy grande al dividir por una suma muy pequeña.
- Se utilizará el signo $:=$ para indicar que la expresión que aparezca a la derecha debe ser asignada a la variable que aparece a la izquierda
- La estructura **if – then – else** se utiliza para indicar que, si se cumple una condición, se ejecute una secuencia de acciones y en caso contrario, se ejecute otra.
- Las estructuras repetitivas sirven para que se ejecute repetidamente una secuencia de acciones. La secuencia se repite hasta que se satisface una cierta condición. Se emplean tres tipos de estructuras repetitivas: **do – while; repeat – until** y **for**.

Auto examen

1. Explique las diferencias esenciales entre los métodos analíticos y los métodos numéricos en cuanto a:
 - a) Tipo de operaciones que utilizan.
 - b) Exactitud de sus resultados.
 - c) Generalidad de sus métodos.
 - d) Posibilidad de implementarse en una computadora.
2. ¿Qué son los errores de truncamiento y por qué reciben este nombre?
3. Se sabe que el número aproximado $x = 26,8768$ tiene un error relativo máximo de 2%. Determine un intervalo donde se encuentre con toda seguridad el número exacto x^* .
4. Al resolver una ecuación por un procedimiento numérico se llegó a la conclusión de que la raíz buscada estaba comprendida en el intervalo $(2,56482; 2,56494)$. Si se toma como aproximación de la raíz el valor $x = 2,5649$ halle el error absoluto máximo de x , su error relativo máximo y la cantidad de cifras decimales exactas que posee.
5. Explique por qué se puede afirmar que en la memoria de una computadora no se pueden almacenar ningún número irracional ni todos los números racionales. ¿Cuáles son, en definitiva, los números que sí se pueden guardar?

6. Se sabe que $a = 55,24$ con 3 cifras exactas, $b = 0,85674$ con error absoluto menor que $0,00001$ y $c = 0,045386$ con un error relativo máximo de $0,05\%$. Si con estos datos se calcula $S = ab + bc + ca$ halle el error relativo máximo de S y determine cuales de sus dígitos son exactos.
7. ¿Qué significa la afirmación de que calcular las raíces de un polinomio de grado alto suele ser un problema inestable?
8. ¿Por qué en los algoritmos numéricos debe evitarse la operación de restar números similares?
9. Se sabe que la función $\sin x$ puede ser aproximada por su polinomio de Taylor de grado $2n+1$:

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \quad n = 0, 1, 2, \dots$$

con error absoluto de truncamiento menor que $\frac{|x|^{2n+3}}{(2n+3)!}$. Escriba un algoritmo en seudo código que para un valor de x y una tolerancia ε conocidas, determine el valor de $\sin x$ con error de truncamiento menor que ε .

CAPÍTULO 2

Matemática Numérica, 2da Edición

Manuel Álvarez, Alfredo Guerra, Rogelio Lau

RAÍCES DE ECUACIONES

Objetivos

Al finalizar el estudio y la ejercitación de este capítulo, el lector debe ser capaz de:

- Separar las raíces reales de una ecuación utilizando en su ayuda la graficación manual en los casos sencillos o un programa graficador en los casos más complicados.
- Utilizar algunos resultados del Álgebra Superior, tales como el teorema de las n raíces, la regla de los signos de Descartes y la Fórmula de Lagrange para establecer cotas en cuanto a la cantidad y localización de las raíces de las ecuaciones algebraicas.
- Describir brevemente los métodos de bisección, Regula Falsi, iterativo general, Newton – Raphson y el método de las secantes, las hipótesis necesarias en cada caso, su interpretación gráfica, sus ventajas e inconvenientes.
- Describir mediante un seudo código cada uno de los algoritmos antes mencionados.
- Describir el concepto de rapidez de convergencia y su relación con el orden de la convergencia de cada uno de los métodos estudiados y comparar los métodos desde este punto de vista.
- Utilizar los métodos de bisección, Regula Falsi, Newton – Raphson y secantes para resolver ecuaciones en forma manual o utilizando algún programa personal si posee los conocimientos necesarios de programación.
- Modelar problemas sencillos que conducen a ecuaciones no lineales y resolverlos, seleccionando en cada caso el método más conveniente.
- Describir el método de Newton para sistemas de dos ecuaciones no lineales y su extensión a un número mayor de ecuaciones y utilizar este método en casos sencillos y con alguna información sobre la posición de la solución buscada.
- Describir el método de Newton – Bairstow para hallar raíces imaginarias de ecuaciones algebraicas y utilizar el método cuando posea alguna información sobre la ubicación de las raíces imaginarias que se desea hallar.

2.1 Introducción

El problema que se resolverá

En este capítulo se tratará acerca del problema de encontrar las raíces reales de una ecuación con una incógnita; en términos más precisos: se tiene una ecuación del tipo:

$$f(x) = 0$$

y se desea hallar los números reales (si es que existe alguno) r_1, r_2, \dots, r_n comprendidos en un cierto intervalo I que satisfacen la ecuación, es decir tales que:

$$f(r_i) = 0 \quad \text{para } i = 1, 2, \dots, n$$

El intervalo I en que se busca las raíces puede ser cerrado o no y su amplitud puede incluso ser infinita.

La importancia de los métodos numéricos que resuelven este problema, viene dada por dos hechos: por una parte, la frecuencia con que se presenta el problema en cualquier rama de la ciencia o de la técnica; por otra parte, lo limitados que resultan los métodos analíticos de solución.

En los ejemplos que siguen se podrá apreciar la sencillez de algunos problemas de los cuales aparecen ecuaciones cuya solución por métodos analíticos no es posible.

Ejemplo 1

Bajo ciertas suposiciones, muchas poblaciones de animales y plantas crecen según el modelo logístico:

$$p(t) = \frac{P_L}{1 - ce^{-kt}}$$

cuya gráfica se muestra en la figura 1 y donde P_L , (población límite), c y k son parámetros; t es el tiempo y $p(t)$ la población en el instante t . Si se conoce la población en tres instantes t_1 , t_2 y t_3 , se puede determinar los parámetros y con ello la función logística correspondiente. Sean: $p(t_1) = p_1$, $p(t_2) = p_2$ y $p(t_3) = p_3$. Entonces, se deberá cumplir que:

$$\frac{P_L}{1 - ce^{-kt_1}} = p_1 \quad (1)$$

$$\frac{P_L}{1 - ce^{-kt_2}} = p_2 \quad (2)$$

$$\frac{P_L}{1 - ce^{-kt_3}} = p_3 \quad (3)$$

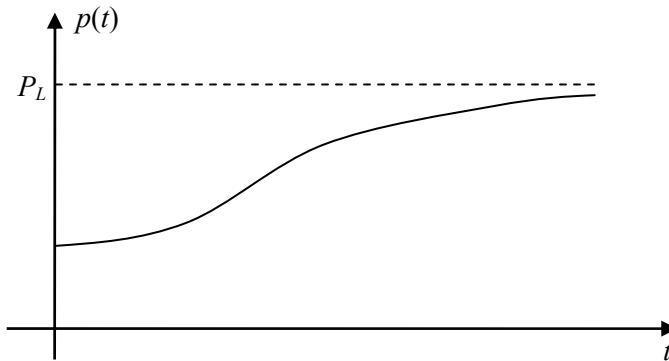


Figura 1

Las ecuaciones (1), (2) y (3) forman un sistema que contiene las tres incógnitas P_L , c y k . Eliminando de este sistema las incógnitas P_L y c se obtiene:

$$(p_2 - p_1)(p_3 e^{-kt_3} - p_2 e^{-kt_2}) - (p_3 - p_2)(p_2 e^{-kt_2} - p_1 e^{-kt_1}) = 0 \quad (4)$$

que es una ecuación con la única incógnita k . La resolución de esta ecuación por métodos algebraicos es, en general, imposible. Más adelante será resuelta numéricamente.

Ejemplo 2

En el momento ($t = 0$) en que el voltaje sinusoidal de la fuente de voltaje de la figura 2 alcanza su máximo valor, se cierra el interruptor y comienza a circular una corriente. Se quiere saber en qué instante la corriente $i(t)$ tomará por primera vez el valor cero.

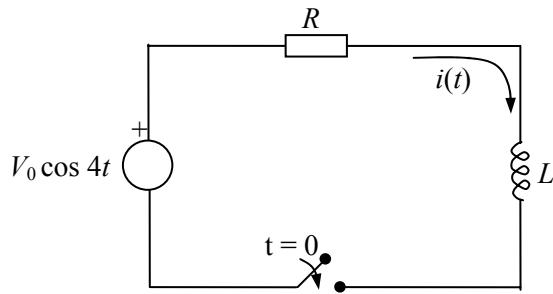


Figura 2

Solución:

Para determinar la expresión analítica de la corriente eléctrica hay que resolver la ecuación diferencial lineal de primer orden:

$$L \frac{di}{dt} + ri = V_0 \cos 4t$$

con la condición inicial: $i(t) = 0$

Al resolver esta ecuación analíticamente se obtiene:

$$i(t) = ae^{-\frac{R}{L}t} + a \cos 4t + b \sin 4t$$

donde

$$a = \frac{V_0 R}{16L^2 + R^2} \quad y \quad b = \frac{4V_0 L}{16L^2 + R^2} \quad (5)$$

Haciendo $i(t) = 0$, se obtiene la ecuación:

$$ae^{-\frac{R}{L}t} + a \cos 4t + b \sin 4t = 0 \quad (6)$$

cuya primera raíz positiva es la solución del problema. Nótese que la resolución de la misma por los métodos tradicionales del álgebra, es imposible. Posteriormente, esta ecuación será resuelta.

Ejemplo 3

Un cable que cuelga de sus extremos adquiere la forma que se muestra en la figura 3 y se demuestra que la ecuación de esta curva (llamada *catenaria*) viene dada por

$$y = L + a \cosh \frac{x}{a}$$

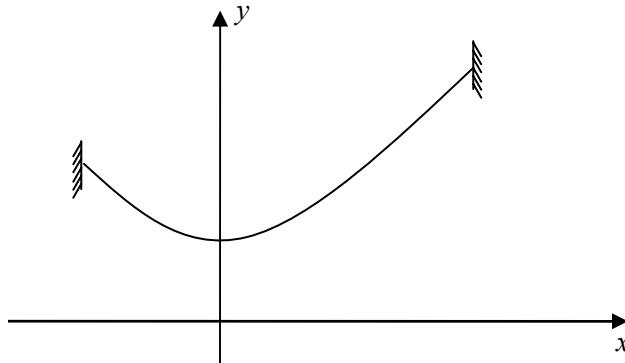


Figura 3

donde el parámetro a depende del peso por unidad de longitud del cable y de la tensión a que es sometido y L de la posición del sistema de referencia, el cual se encuentra colocado de manera que el origen de coordenadas se encuentra justamente debajo del punto de altura mínima. Como la ecuación solo contiene dos parámetros, midiendo la altura de un cable en dos puntos, se pueden determinar ambos parámetros. Por ejemplo, si la altura mínima es de 15 m y si 10 m más allá, la altura del cable es de 17 m, entonces:

Para $x = 0$ es $y = 15$ y, por tanto: $L + a = 15$

Para $x = 10$ es $y = 17$ y, será: $L + a \cosh \frac{10}{a} = 17$

Basta restar ambas ecuaciones para eliminar la incógnita L y obtener:

$$a \cosh \frac{10}{a} - a = 2 \quad (7)$$

ecuación cuya raíz da el valor de a , del cual resulta fácil determinar posteriormente el valor de L . Sin embargo, solo mediante métodos como los que se verán en este capítulo, puede resolverse esta ecuación, en apariencia sencilla.

Ejemplo 4

Se quiere construir un recipiente cilíndrico de 1000 cm^3 de capacidad, utilizando la mínima cantidad de material. Teniendo en cuenta que es necesario un sobrante de $0,25 \text{ cm}$ para poder doblar y soldar el material, entonces, si las dimensiones del recipiente son $r \text{ cm}$ de radio y $h \text{ cm}$ de altura, se tiene (ver figura 4):

$$S = 2\pi(r + 0,25)^2 + (2\pi r + 0,25)h$$

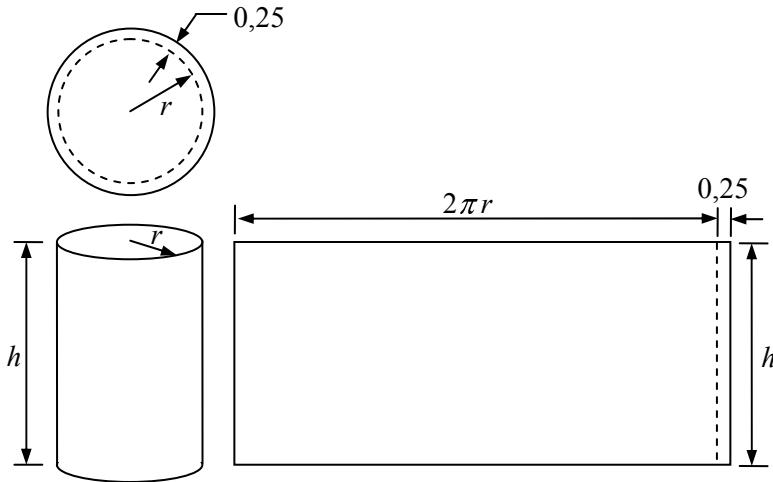


Figura 4

donde S es la superficie en cm^2 de material necesario para fabricar la lata. Como el volumen de la misma tiene que ser de 1000 cm^3 , se tendrá:

$$\pi r^2 h = 1000$$

de donde, despejando h y sustituyendo en la expresión de S se obtiene:

$$S = 2\pi(r + 0,25)^2 + (2\pi r + 0,25) \frac{1000}{\pi r^2}$$

Como se sabe, el valor de r que hace mínimo a S se obtiene derivando respecto a r e igualando a cero:

$$\frac{dS}{dr} = 4\pi(r + 0,25) - \frac{2000}{r^2} - \frac{500}{\pi r^3} = 0$$

Multiplicando ambos miembros por r^3 (se supone $r \neq 0$)

$$4\pi r^4 + \pi r^3 - 2000r - \frac{500}{\pi} = 0 \quad (8)$$

que es una ecuación algebraica de cuarto grado, cuyas raíces se puede calcular por los métodos algebraicos, pero con una gran dificultad. Próximamente, esta ecuación será resuelta numéricamente.

2.2 Separación de raíces

Dos etapas

En todos los métodos de resolución de ecuaciones que serán considerados en este capítulo, se parte del conocimiento de un intervalo (a, b) en el que la ecuación tiene exactamente una raíz. Lo más frecuente, sin embargo, es que en el intervalo de interés, lo más que se pueda asegurar es la existencia de una o más raíces. Por esta razón, el cálculo de las raíces de una ecuación consta, en general, de dos etapas; la primera, llamada separación de las raíces, persigue precisamente determinar intervalos como el (a, b) que contengan una raíz; la segunda es ya la aplicación de un algoritmo para hallar una aproximación de la raíz deseada con la aproximación requerida.

Separación gráfica de raíces

La técnica más elemental para la separación de raíces es el método gráfico que utiliza el conocido hecho de que las raíces de $f(x) = 0$ son las abscisas de los puntos en que la gráfica de la función $y = f(x)$ corta al eje x .

Es obvio que de esta forma no se pueden determinar las raíces con una precisión aceptable, pero sí se pueden acotar dentro de intervalos de separación suficientemente pequeños. Como en la actualidad existe una enorme cantidad de programas que grafican funciones, el método gráfico es sumamente atractivo y eficiente. Por lo general, se comienza graficando la función $f(x)$ en un intervalo grande y se van precisando intervalos de búsqueda más pequeños en los cuales se ordena de nuevo graficar la función, hasta obtener un intervalo en que solamente esté contenida la raíz que interese. El método gráfico brinda, además, valiosa información acerca de las características de la función $f(x)$, tales como los signos de la función y de su primera y segunda derivadas en diferentes puntos del intervalo de separación; esta información puede ser de gran importancia para la aplicación posterior de los métodos de cálculo de raíces.

Ejemplo 1

Separar las raíces de la ecuación: $\ln x + 4x - x^2 - 2 = 0$

Solución:

En la figura 1 se muestra la gráfica que aparece en la pantalla correspondiente a la función:

$$f(x) = \ln x + 4x - x^2 - 2$$

en el intervalo $[0,001; 100]$. Es evidente que fuera de este intervalo no puede encontrarse ninguna raíz, ya que para $x \leq 0$ el logaritmo de x no está definido y para $x > 100$ el término x^2 predomina obviamente sobre los demás, haciendo a $f(x)$ tomar valores negativos enormes. De la gráfica se observa que las raíces se hallan relativamente próximas a cero, por lo cual se repite el proceso con un intervalo más reducido tal como $[0,001; 8]$ (figura 2), en el cual ya es posible apreciar dos raíces, una en $(0, 1)$ y otra en $(3, 4)$.

$$f(x) = \ln x + 4x - x^2 - 2$$

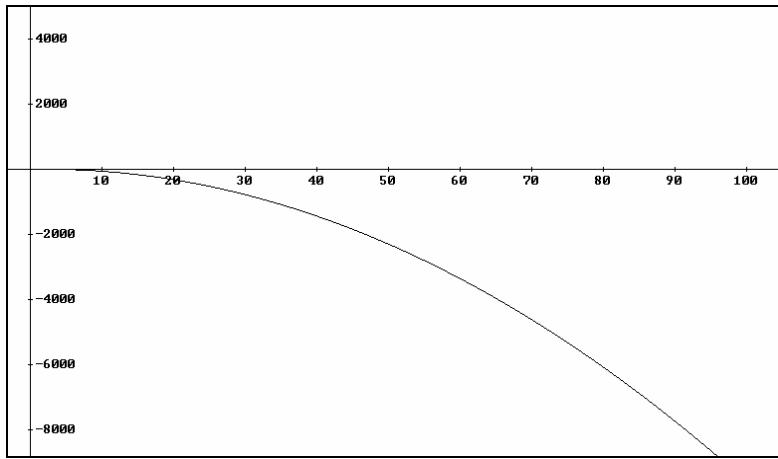


Figura 1

$$f(x) = \ln x + 4x - x^2 - 2$$

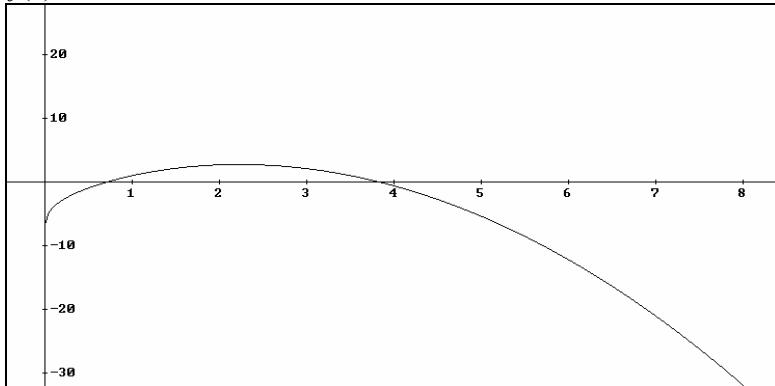


Figura 2

Algunos resultados importantes para ecuaciones algebraicas

En particular para las ecuaciones algebraicas, es decir, de la forma:

$$a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n = 0$$

donde n es un número natural y a_0, a_1, \dots, a_n son constantes reales, existen resultados muy importantes del Álgebra Superior que hacen más simple la separación de raíces. A continuación se verán algunos. En todos los casos se excluyen las demostraciones, ya que no constituye objetivos de este texto. Al final del capítulo, en “Otras lecturas recomendadas”, se indica dónde puede encontrarlas el lector interesado en ellas.

Teorema 1 (de las n raíces)

Una ecuación algebraica de grado n tiene n raíces, reales o imaginarias, si cada raíz se cuenta tantas veces según sea su multiplicidad. ■

El interés de este teorema en el proceso de separación de raíces reales, es que hace posible establecer una cota superior del número de estas, pues, evidentemente, del teorema puede extraerse la siguiente consecuencia:

Una ecuación algebraica de grado n tiene como máximo n raíces reales.

Ejemplo 2

Se puede afirmar que la ecuación: $x^5 + 4x^4 - x^3 - 10x^2 - 6x - 36 = 0$ tiene a lo más cinco raíces reales, ya que es de quinto grado.

Regla de Descartes

Sea m el número de cambios de signo que se presentan en la sucesión de coeficientes de la ecuación $a_0x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n = 0$. Entonces el número de raíces positivas de la ecuación es menor o igual que m y tiene su misma paridad.

Para la aplicación de esta sencilla regla, cuya demostración no es nada trivial, los coeficientes nulos no se tienen en cuenta. En el número de raíces, por supuesto, cada una se considera tantas veces como sea su multiplicidad. Obsérvese que la regla de Descartes establece una *cota superior* para la cantidad de raíces *positivas* de una ecuación *algebraica*; no obstante en dos casos particulares se puede llegar a conclusiones precisas: Si $m = 0$ se puede asegurar que no hay raíces positivas y si $m = 1$ se puede asegurar que existe exactamente una raíz positiva.

Ejemplo 3

Aplique la regla de Descartes para analizar el número de raíces positivas de las ecuaciones:

- $x^4 + 3x^2 - 5x + 8 = 0$
- $x^5 - x^4 + 3x^3 - 8x^2 + x - 9 = 0$
- $x^5 + 4x^4 - 3x^3 - 10x^2 - 6x - 36 = 0$

Solución:

- La sucesión de los coeficientes (los coeficientes nulos, como el de x^3 , no se consideran) es:

$$\begin{array}{ccccccc} 1 & & 3 & & -5 & & 8 \\ & & \underbrace{}_{\text{cambio}} & & \underbrace{}_{\text{cambio}} & & \end{array}$$

Como hay $m = 2$ cambios de signo en la sucesión de los coeficientes, el número de raíces positivas tiene que ser menor o igual a 2 y de su misma paridad, es decir, 0 ó 2. En conclusión la ecuación ó tiene 2 raíces positivas ó no tiene raíces positivas. En este caso, la regla de Descartes solo deja estas dos posibilidades.

- b) Como hay $m = 5$ cambios de signo en la sucesión de los coeficientes, el número de raíces positivas tiene que ser menor o igual a 5 y de su misma paridad, es decir, impar. La cantidad de raíces positivas será entonces: 5 ó 3 ó 1.
- c) Como existe un solo cambio de signo en la sucesión de los coeficientes ($m = 1$) se puede asegurar que la ecuación tiene exactamente una raíz positiva.

La fórmula de Lagrange para acotar raíces

Sea la ecuación $a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$ con $a_0 > 0$. Si B es el valor absoluto del coeficiente negativo con mayor valor absoluto y a_k es el primer coeficiente negativo contando desde la izquierda, entonces todas las raíces positivas de la ecuación, si existen, son menores que el número

$$R = 1 + \sqrt[k]{\frac{B}{a_0}}$$

Nótese que el número R da una cota superior de las raíces *positivas* de la ecuación; no dice nada acerca de las negativas. Obsérvese también que el índice k de la raíz se obtiene como el subíndice que le corresponde al primer coeficiente negativo.

Ejemplo 4

Mediante la fórmula de Lagrange, halle una cota superior de las raíces positivas de la ecuación:

$$x^4 + 3x^2 - 5x + 8 = 0$$

Solución:

Los coeficientes de la ecuación son: $a_0 = 1$; $a_1 = 0$; $a_2 = 3$; $a_3 = -5$; $a_4 = 8$

El coeficiente negativo de mayor valor absoluto (en este caso, el único) vale -5 , así que $B = 5$. Como el primer coeficiente negativo es a_3 entonces $k = 3$. La fórmula da para R :

$$R = 1 + \sqrt[3]{\frac{B}{a_0}} = 1 + \sqrt[3]{\frac{5}{1}} = 1 + \sqrt[3]{5} = 2,71$$

Puede entonces asegurarse que todas las raíces positivas que tenga la ecuación (que, por la regla de Descartes, son dos o ninguna) están comprendidas en el intervalo $[0; 2,71]$.

Análisis de las raíces negativas de una ecuación algebraica

Las reglas de Descartes y de Lagrange solo se refieren a raíces positivas. Cuando se deseé investigar las raíces negativas, bastará cambiar en la ecuación la variable x por $-x$ y analizar las raíces positivas de la nueva ecuación obtenida, ya que, evidentemente, si r es una raíz positiva de la nueva ecuación $f(-x) = 0$ entonces $-r$ es una raíz negativa de la ecuación original $f(x) = 0$.

Ejemplo 5

Analice las raíces reales negativas de las ecuaciones

- a) $x^5 + 4x^4 - x^3 - 10x^2 - 6x - 36 = 0$
- b) $x^4 + 3x^2 - 5x + 8 = 0$

Solución:

- a) Designese la ecuación original con (1):

$$x^5 + 4x^4 - x^3 - 10x^2 - 6x - 36 = 0 \quad (1)$$

Cambiando x por $-x$ se obtiene la nueva ecuación:

$$-x^5 + 4x^4 + x^3 - 10x^2 + 6x - 36 = 0 \quad (2)$$

o, lo que es igual: $x^5 - 4x^4 - x^3 + 10x^2 - 6x + 36 = 0$

Como se observan cuatro cambios de signo en la sucesión de coeficientes, se puede afirmar que la ecuación (2) tiene cuatro, dos o ninguna raíces positivas. Según el teorema de Lagrange, las mismas están acotadas por:

$$R = 1 + \sqrt[k]{\frac{B}{a_0}} = 1 + \sqrt[1]{\frac{6}{1}} = 1 + \sqrt[1]{6} = 7$$

De acuerdo con esto, las raíces negativas de la ecuación (1) serán cuatro, dos o ninguna y estarán comprendidas en el intervalo $[-7, 0]$.

- b) La ecuación original es: $x^4 + 3x^2 - 5x + 8 = 0 \quad (3)$

Cambiando x por $-x$ se obtiene: $x^4 + 3x^2 + 5x + 8 = 0 \quad (4)$

Cuyos coeficientes son todos no negativos. En este caso la regla de Descartes es concluyente: la ecuación (4) no posee raíces positivas. Por lo tanto, la ecuación original (3), no tiene raíces negativas.

Combinando las técnicas

Aunque en las páginas anteriores, algunas de las diversas técnicas de separación de raíces han sido utilizadas en forma aislada para facilitar su comprensión, es claro que combinándolas adecuadamente es como se logra efectuar con mayor seguridad y rapidez la separación de las

raíces de una ecuación. En los ejemplos que siguen se ilustran algunas posibilidades en este sentido.

Ejemplo 6

Separar las raíces de la ecuación $x^3 - 9x^2 - 9x + 19 = 0$ en intervalos de amplitud 0,5.

Solución:

Como se trata de una ecuación algebraica de tercer grado, se sabe que el número de raíces reales será como máximo tres. De ellas, las positivas serán, según la regla de Descartes, dos o ninguna. Con esta información resulta fácil aplicar el método gráfico en la computadora para separar las raíces. No obstante, si se conoce una cota superior R de las posibles raíces positivas, entonces se simplifica el proceso gráfico pues la búsqueda está limitada al intervalo $[0, R]$. Según la regla de Lagrange:

$$R = 1 + \sqrt[k]{\frac{B}{a_0}} = 1 + \sqrt[3]{\frac{9}{1}} = 1 + \sqrt[3]{9} = 10$$

Al graficar la función $f(x) = x^3 - 9x^2 - 9x + 19$ en el intervalo $[0, 10]$ se observa la gráfica de la figura 3:

$$f(x) = x^3 - 9x^2 - 9x + 19$$

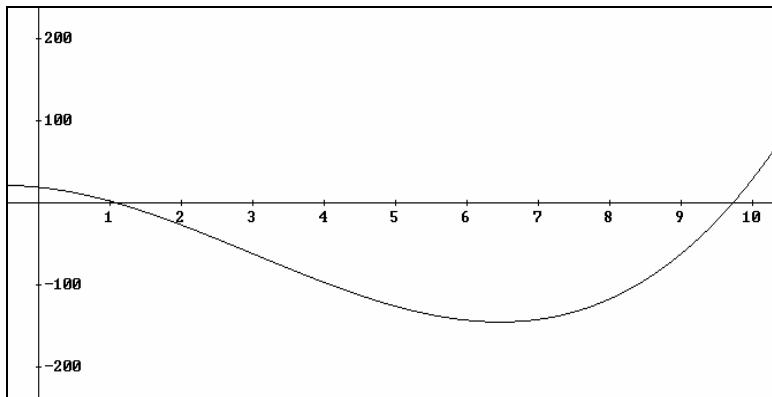


Figura 3

Resulta evidente la presencia de dos raíces positivas, una en el intervalo $[1; 1,5]$ y otra en $[9,5; 10]$. Una vez separadas las dos raíces reales positivas se sabe que hay una raíz negativa, ya que las raíces imaginarias de las ecuaciones algebraicas con coeficientes reales solo se pueden presentar por pares. Para acotar esta raíz, se obtiene la ecuación resultante de cambiar x por $-x$:

$$-x^3 - 9x^2 + 9x + 19 = 0$$

esto es:

$$x^3 + 9x^2 - 9x - 19 = 0$$

Aplicando la fórmula de Lagrange: $R = 1 + \sqrt[k]{\frac{B}{a_0}} = 1 + \sqrt[3]{\frac{19}{1}} = 1 + \sqrt[3]{19} \approx 5,4$

Por lo tanto, la raíz negativa de la ecuación original se halla en el intervalo $[-5,4; 0]$. Utilizando el procedimiento gráfico en el intervalo $[-6, 0]$ (figura 4) resulta evidente que la raíz se encuentra en el intervalo $[-2; -1,5]$.

$$f(x) = x^3 - 9x^2 - 9x + 19$$

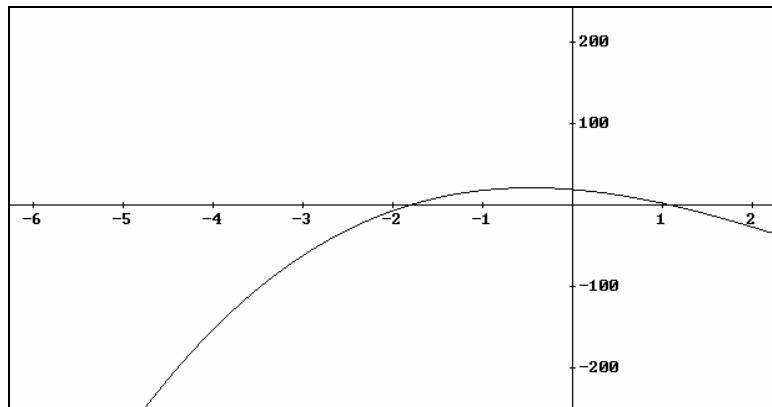


Figura 4

En conclusión, las raíces de la ecuación $x^3 - 9x^2 - 9x + 19 = 0$ son tres, las tres son reales y se encuentran en los intervalos: $[-2; -1,5]$, $[1; 1,5]$ y $[9,5; 10]$.

Ejemplo 7

Separé las raíces de la ecuación $x^2 + 10 \cos x = 0$ en intervalos de amplitud 0,5.

Solución:

Nótese primeramente que esta no es una ecuación algebraica y, por tanto, no se pueden aplicar algunas de las herramientas, específicas para ese tipo de ecuaciones, como las reglas de Descartes y de Lagrange. No obstante, siempre se debe auxiliar al método puramente gráfico con un análisis de la ecuación que permita establecer, al menos, cotas groseras de las raíces reales. La función

$$f(x) = x^2 + 10 \cos x$$

cuyos ceros se quiere hallar, está formada por la suma de dos funciones pares y es, por tanto, par; luego sus ceros estarán situados simétricamente alrededor de $x = 0$. Basta entonces ocuparse de las raíces positivas. La función x^2 es positiva y crece ilimitadamente en R^+ , mientras que $10 \cos x$ solo toma valores entre -10 y 10 . Resulta entonces que, una vez que x^2 alcance valores mayores que 10 , la función $f(x)$ no puede tener más ceros. La búsqueda puede entonces reducirse al intervalo $[0, 4]$. En la figura 5 se muestra la gráfica de la función en ese intervalo. Resulta evidente que la ecuación tiene una raíz en cada uno de los intervalos $[1,5; 2]$ y $[3; 3,5]$. Por la simetría de la función, existen raíces negativas en los intervalos $[-3,5; -3]$ y $[-2; -1,5]$.

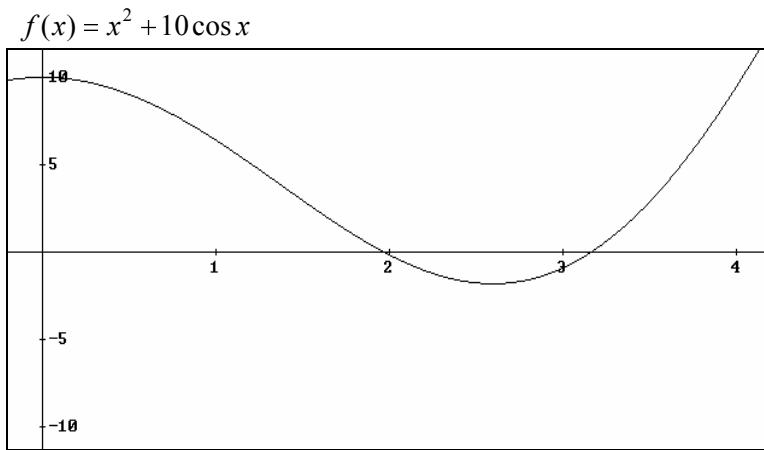


Figura 5

Ejemplo 8

Separé las raíces positivas de la ecuación $\frac{1}{x+1} - x^2 + 100x - 100 = 0$

Solución:

Con este ejemplo se pretende ilustrar el peligro de aplicar el método gráfico sin un análisis de la ecuación, que permita asegurar que se han considerado todas las raíces. La gráfica de la función

$$f(x) = \frac{1}{x+1} - x^2 + 100x - 100$$

en el intervalo $[0, 10]$ se muestra en la figura 6; se observa una raíz en $[0,5; 1,5]$; se ve además que la función crece rápidamente alcanzando en $x = 10$ un valor mayor que 800. Tomando un intervalo mucho mayor: $[0, 40]$ (figura 7) se observa el mismo comportamiento y $f(x)$ toma ya un valor superior a 2 300 para $x = 40$. Si, a partir de aquí, se concluyera que la ecuación solo posee una raíz positiva, se estaría cometiendo un error. Basta observar la gráfica en un intervalo aun mayor, por ejemplo $[0, 120]$ (figura 8) para comprobar la existencia de una segunda raíz en las proximidades de $x = 100$.

$$f(x) = \frac{1}{x+1} - x^2 + 100x - 100$$

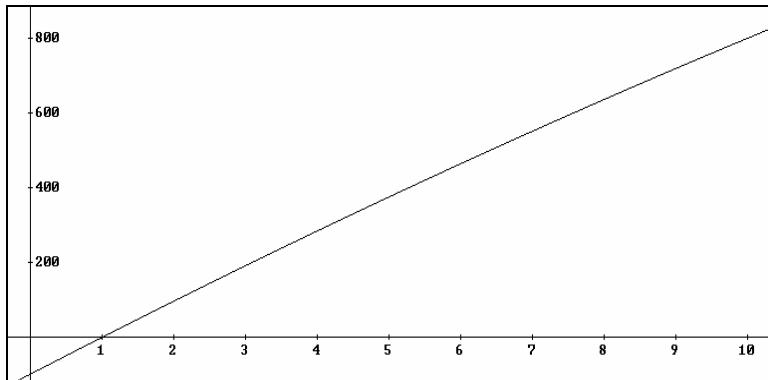


Figura 6

$$f(x) = \frac{1}{x+1} - x^2 + 100x - 100$$

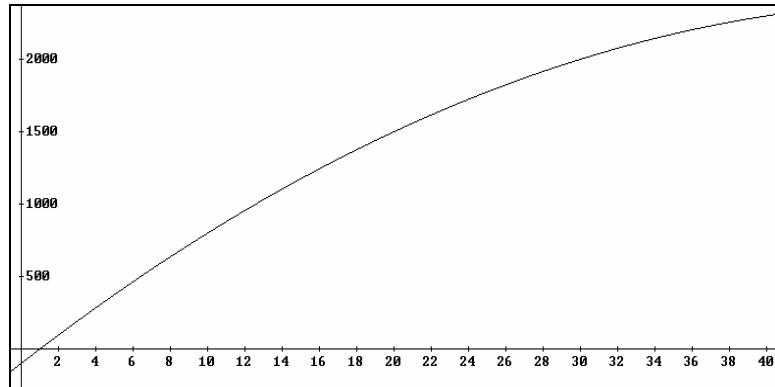


Figura 7

$$f(x) = \frac{1}{x+1} - x^2 + 100x - 100$$

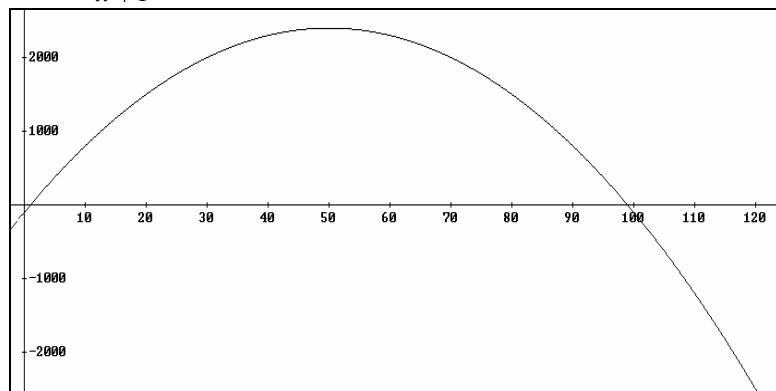


Figura 8

El error se podría haber evitado considerando que

$$\lim_{x \rightarrow \infty} f(x) = -\infty$$

ya que el término $-x^2$ predomina para x suficientemente grande y, por tanto, la función $f(x)$ tendría que tomar definitivamente los valores negativos.

Ejercicios

En los siguientes ejercicios utilice un programa graficador en la computadora cuando esto sea necesario.

1. Aplique las reglas de Descartes y de Lagrange para acotar el número y la posición de las raíces positivas y negativas de las ecuaciones que siguen. Compruebe sus resultados mediante el método gráfico.

a) $x^4 + x^3 - x^2 + x - 2 = 0$
b) $x^4 - 11x^3 + 41x^2 - 60x + 30 = 0$
c) $x^4 - 3x^3 + 10x^2 - 13x + 5 = 0$
d) $x^4 + 3x^2 + 2 = 0$
e) $x^3 + 7x^2 + 14x + 9 = 0$

2. Separe las raíces de las ecuaciones anteriores en intervalos de amplitud 0,5.
3. Separe los ceros, los puntos de extremo y los puntos de inflexión de la función

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

en intervalos de amplitud 0,3.

4. Separe las raíces de la ecuación $\sin x - \log x = 0$ en intervalos donde la derivada de la función $\sin x - \log x$ no cambie de signo.
5. Separe las raíces de la ecuación $5e^x - 2x - 10 = 0$ en intervalos de amplitud 0,5.
6. Separe las raíces de la ecuación $(x^2 + 1)\cos x = 1$ comprendidos en el intervalo $[-10, 10]$.
7. Dada la función $f(x) = 2 \tanh x - \sin x - 0,3$, separe las raíces de la ecuación $f(x) = 0$ en intervalos donde las derivadas primera y segunda de la función $f(x)$ no se anulen.
8. Separe las raíces de la ecuación $x^4 - 4x^2 - 4x - 16 - \ln|x| = 0$ en intervalos de amplitud 0,2.
9. Separe las raíces de la ecuación $e^x \sin x - 2e^x + 3 = 0$ en intervalos de amplitud 0,5.
10. Separe en intervalos de amplitud 0,2 las raíces de la ecuación $x = \tan^2 x$ comprendidas en el intervalo $[0, 2\pi]$.
11. Separe en un intervalo de amplitud 0,2 la mayor raíz de la ecuación $\sin \frac{1}{x} = x$.
12. La hipérbola equilátera $xy = 1$ y la circunferencia $x^2 + y^2 = 4$ se cortan en cuatro puntos. Determine intervalos de amplitud 0,1 que contengan las abscisas de esos puntos.
13. Separe las raíces de las ecuaciones $\pm \sqrt{x} = 2 \ln x$

2.3 El método de la bisección

Hipótesis

Sea la ecuación $f(x) = 0$ y un intervalo $[a, b]$ tales que:

1. En el intervalo la ecuación tiene una sola raíz $x = r$.
2. $f(x)$ es continua en $[a, b]$
3. $f(x)$ posee signos diferentes en a y en b , es decir, $f(a) \cdot f(b) < 0$

El método

Tal como indica su nombre, el método consiste en aproximar la raíz de la ecuación como el punto medio del intervalo $[a, b]$. Evaluando la función en este punto se decide si la raíz se encuentra en la mitad izquierda del intervalo o en la mitad derecha. De esta manera, una de las dos mitades queda descartada y la amplitud del nuevo intervalo de búsqueda es exactamente un medio de la anterior. A medida que este proceso se repite, el intervalo de búsqueda va disminuyendo en amplitud. Si se conviene en llamar $[a_1, b_1]$ al intervalo inicial, entonces, en la iteración número n del método se tiene:

$$x_n = \frac{a_n + b_n}{2} \quad n = 1, 2, 3, \dots \quad (1)$$

con error absoluto máximo: $E_m(x_n) = \frac{b_n - a_n}{2}$ (2)

Al evaluar la función $f(x)$ en x_n se selecciona el nuevo intervalo de búsqueda $[a_{n+1}, b_{n+1}]$ de acuerdo con la siguiente regla:

Si $f(a_n) \cdot f(x_n) < 0$ entonces la raíz se encuentra en $[a_n, x_n]$ y se escoge $a_{n+1} = a_n$ y $b_{n+1} = x_n$

Si $f(x_n) \cdot f(b_n) < 0$ entonces la raíz se encuentra en $[x_n, b_n]$ y se escoge $a_{n+1} = x_n$ y $b_{n+1} = b_n$

Si $f(x_n) \cdot f(b_n) = 0$ entonces x_n es la raíz buscada y no hay que continuar.

En la figura 1 se ilustra gráficamente los primeros pasos del procedimiento. Para poder tener un algoritmo se requieren todavía varias cosas. Primero, hay que probar que este procedimiento converge hacia la raíz r ; esto posee importancia teórica y práctica, ya que la convergencia garantiza que la aproximación x_n se acercará a la raíz r tanto como se quiera tomando una n suficientemente grande y eso significa que se podrá obtener una aproximación a r tan buena como se desee. También se necesita un criterio práctico para detener el proceso iterativo, el cual se obtiene a partir del error absoluto máximo en cada aproximación.

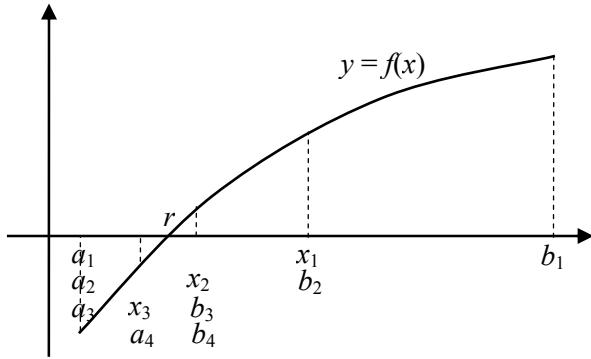


Figura 1

Convergencia del método

Como en cada paso del algoritmo el intervalo de búsqueda se reduce a la mitad, se tiene que:

$$b_n - a_n = \frac{b_{n-1} - a_{n-1}}{2} \quad \text{para } n = 2, 3, 4, \dots$$

Como, según (2) el error absoluto máximo se obtiene mediante: $E_m(x_n) = \frac{b_n - a_n}{2}$ resulta:

$$E_m(x_n) = \frac{b_n - a_n}{2} = \frac{1}{2} \left(\frac{b_{n-1} - a_{n-1}}{2} \right)$$

Esto es:

$$E_m(x_n) = \frac{1}{2} E_m(x_{n-1}) \quad (3)$$

Aplicando reiteradamente la ecuación (3):

$$E_m(x_n) = \frac{1}{2} E_m(x_{n-1}) = \frac{1}{4} E_m(x_{n-2}) = \frac{1}{8} E_m(x_{n-3}) = \dots = \frac{E_m(x_1)}{2^{n-1}}$$

y como $E_m(x_1) = \frac{b-a}{2}$ resulta:

$$E_m(x_n) = \frac{b-a}{2^n} \quad (4)$$

Cuando la variable entera n tiende hacia infinito, el miembro de la derecha de la igualdad (4) tiende hacia cero, así que:

$$\lim_{n \rightarrow \infty} E_m(x_n) = 0 \quad (5)$$

Según su definición, el error absoluto máximo de x_n satisface:

$$|r - x_n| \leq E_m(x_n)$$

Entonces:

$$\lim_{n \rightarrow \infty} |r - x_n| = 0$$

y esto significa que

$$\lim_{n \rightarrow \infty} x_n = r \quad (6)$$

La conclusión anterior se puede redactar como un teorema:

Teorema 1

Si en el intervalo $[a, b]$ la función $f(x)$ satisface las hipótesis del método de la bisección y las aproximaciones x_n ($n = 1, 2, 3, \dots$) son halladas aplicando dicho procedimiento, entonces la sucesión x_1, x_2, x_3, \dots converge hacia la solución de la ecuación $f(x) = 0$ en $[a, b]$.

Condición de terminación

Una vez que se tiene una forma de hallar el error absoluto máximo en cada paso de un método iterativo el proceso se puede detener tan pronto como dicho error sea suficientemente pequeño. Para precisar ideas, si el número $\varepsilon > 0$ es la tolerancia con que se necesita la raíz de la ecuación, entonces el proceso iterativo se debe detener cuando $E_m(x_n)$ sea menor o igual que ε . El error absoluto máximo se halla en cada paso mediante la ecuación (2). Entonces la condición de terminación del proceso iterativo será:

Condición de terminación:

Si se desea obtener la raíz de la ecuación con un error absoluto menor que ε , el método de bisección se llevará a cabo hasta la aproximación x_n para la cual

$$E_m(x_n) = \frac{b_n - a_n}{2} \leq \varepsilon$$

En el método de bisección se puede determinar, antes de comenzar, el número de iteraciones que será necesario realizar para alcanzar una cierta precisión; en efecto, de acuerdo con la ecuación (4) se tiene:

$$E_m(x_n) = \frac{b - a}{2^n}$$

Si se desea un error absoluto menor que ε será necesario iterar hasta un n tal que:

$$\frac{b - a}{2^n} \leq \varepsilon$$

de donde se puede despejar n :

$$\frac{b - a}{\varepsilon} \leq 2^n$$

$$n \ln 2 \geq \ln\left(\frac{b-a}{\varepsilon}\right)$$

$$n \geq \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln 2} \quad (7)$$

Esta fórmula resulta un poco complicada para recordar, así que es preferible en cada caso utilizar la fórmula (4) y realizar el razonamiento que condujo a (7). De esta ecuación se aprecia, por otra parte, un hecho muy importante: la cantidad de iteraciones que se requiere en el método de bisección para obtener una raíz con error absoluto menor o igual que ε , solamente depende de ε y de la amplitud $(b-a)$ del intervalo inicial pero no de las características de la función $f(x)$. De esto se volverá a hablar más adelante.

Ejemplo 1

Determine cuántas iteraciones del método de bisección se necesitan para obtener con cinco cifras decimales exactas una raíz de una ecuación, si dicha raíz está separada dentro de un intervalo de amplitud menor que 10.

Solución:

Tomando

$$b-a = 10 \quad y \quad \varepsilon = 0,5 \cdot 10^{-5}$$

la fórmula (7) da:

$$n \geq \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln 2} = \frac{\ln\left(\frac{10}{0,5 \cdot 10^{-5}}\right)}{\ln 2} = \frac{\ln(2 \cdot 10^6)}{\ln 2} = 19,93157$$

Como se ve, con 20 iteraciones se logra la exactitud deseada.

Algoritmo en seudo código

Se supone que la ecuación a resolver es $f(x) = 0$, que la raíz que se quiere hallar está separada dentro de un intervalo $[a, b]$ en el cual $f(x)$ es continua y que $f(a)f(b) < 0$. Se suponen conocidas la función $f(x)$, los extremos a y b del intervalo de separación y la tolerancia ε que se permitirá.

```

repeat
     $x := \frac{a+b}{2}$ 
     $Error := \frac{b-a}{2}$ 
    if  $f(x) = 0$  then
         $x$  es exactamente la raíz buscada
        Terminar
    else
        if  $f(a)f(x) < 0$  then
             $b := x$ 
        else
    
```

```

    a := x
end
end
until Error ≤ ε
La raíz buscada es  $x$  y su error absoluto máximo es  $Error$ 
Terminar

```

Comentarios finales

El método de bisección es el más sencillo de los métodos para determinar raíces reales de ecuaciones. Es un método poco eficiente en comparación con los que se verá más adelante, por lo cual no es recomendable si los cálculos (sobre todo, la evaluación de $f(x)$) hay que realizarlos a mano. Con el uso masivo de las computadoras, este problema no posee tanta trascendencia, a menos que se requiera resolver un enorme volumen de ecuaciones como parte de un algoritmo mayor. Por otra parte el método de bisección posee varios atractivos importantes:

- Las condiciones que se requiere para su aplicación son mínimas, de hecho $f(x)$ solo requiere ser continua en el intervalo de separación.
- El algoritmo posee una lógica muy simple y es muy fácil de programar.
- La rapidez de la convergencia es independiente de la función $f(x)$, por lo cual no existen temores de que se presenten casos patológicos, cuestión presente en casi todos los métodos más eficaces.
- La acotación del error es muy simple y segura y es también independiente de las características que posea la función $f(x)$.

Todo lo anterior se puede resumir en una sola frase: No es un método rápido pero es el más robusto de todos los procedimientos para hallar raíces reales de ecuaciones.

Ejemplo 2

Halle, con cuatro cifras decimales exactas, las raíces de la ecuación $e^{-x} = x$

Solución:

Primero es necesario separar las raíces. Si no se dispone de una computadora, lo mejor sería graficar sobre el mismo sistema de ejes las funciones $y = e^{-x}$ y $y = x$, que son muy sencillas, y determinar las abscisas de los puntos en que las gráficas se intersecan. Si se dispone de un programa graficador, seguramente es preferible escribir la ecuación como $e^{-x} - x = 0$ y graficar la función $f(x) = e^{-x} - x$. Aquí se utilizará esta variante. Como la exponencial es positiva para toda x , la ecuación solo podrá tener raíces para $x > 0$. Por otra parte, cuando x crece el término e^{-x} tiende hacia cero y, por lo tanto, $f(x)$ se hará negativa; todo esto significa que las raíces de la ecuación serán positivas y próximas a $x = 0$.

En la figura 2 se muestra la gráfica de $f(x)$ en el intervalo $[0, 4]$. Obviamente solo existe una raíz y se halla en el intervalo $[0, 1]$. De la gráfica se observa, además, que las hipótesis requeridas se satisfacen:

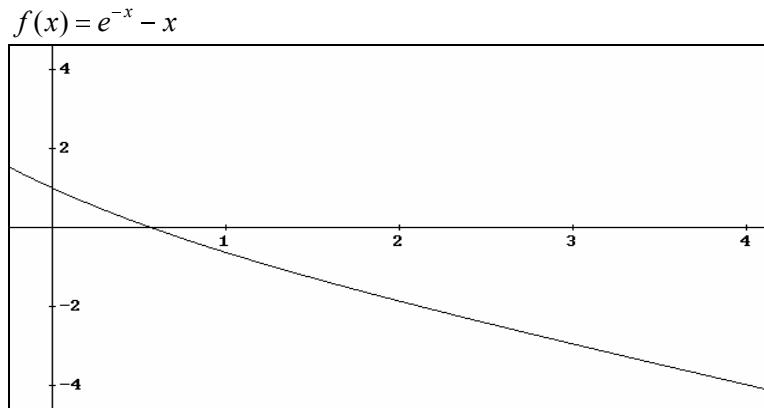


Figura 2

- En $[0, 1]$ la ecuación pose una sola raíz.
- $f(x)$ es continua en $[0, 1]$
- $f(0)$ y $f(1)$ tienen signos opuestos.

Con un programa confeccionado a partir del algoritmo del método de bisección, se obtienen los resultados que muestra la tabla 1. Se utilizaron los datos: $f(x) = e^{-x} - x$; intervalo inicial: $[0, 1]$; tolerancia: $\varepsilon = 0,00005$.

Iteración	a	b	x	$E_m(x)$
1	0	1	0,5	0,5
2	0,5	1	0,75	0,25
3	0,5	0,75	0,625	0,125
4	0,5	0,625	0,5625	0,0625
5	0,5625	0,625	0,59375	0,03125
6	0,5625	0,59375	0,578125	0,015625
7	0,5625	0,578125	0,570313	0,007813
8	0,5625	0,570313	0,566406	0,003906
9	0,566406	0,570313	0,568359	0,001953
10	0,566406	0,568359	0,567383	0,000977
11	0,566406	0,567383	0,566895	0,000488
12	0,566895	0,567383	0,567139	0,000244
13	0,567139	0,567383	0,567261	0,000122
14	0,567139	0,567261	0,567200	0,000061
15	0,567139	0,567200	0,567169	0,000031

Tabla 1

Resultado: La única raíz real de la ecuación es 0,567169 con cuatro cifras decimales exactas.

Ejercicios

En todos los ejercicios que siguen se debe utilizar, siempre que se necesite, un programa graficador tanto para separar las raíces como para verificar las hipótesis del método utilizado. También se supone que el algoritmo de bisección se utilice mediante un programa computacional, preferiblemente confeccionado por usted. Si no cuenta con un programa adecuado, realice los cálculos a mano y obtenga las raíces con solo dos o tres cifras decimales exactas.

1. Calcule, con cinco cifras decimales exactas, las raíces reales de las siguientes ecuaciones algebraicas. Posiblemente, ya usted separó las raíces de estas ecuaciones en los ejercicios de la sección 2.2.

- a) $x^4 + x^3 - x^2 + x - 2 = 0$
- b) $x^4 - 11x^3 + 41x^2 - 60x + 30 = 0$
- c) $x^4 - 3x^3 + 10x^2 - 13x + 5 = 0$
- d) $x^4 + 3x^2 + 2 = 0$
- e) $x^3 + 7x^2 + 14x + 9 = 0$

2. Halle los ceros, los puntos de extremo y los puntos de inflexión de la función

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

con cinco cifras decimales exactas. Es probable que ya usted haya separado las raíces necesarias en los ejercicios de la sección 2.2.

3. Halle, con cinco cifras decimales exactas, las raíces de las siguientes ecuaciones trascendentes. Posiblemente, ya usted separó las raíces de estas ecuaciones en los ejercicios de la sección 2.2.

- a) $\sin x - \log x = 0$
- b) $5e^x - 2x - 10 = 0$
- c) $(x^2 + 1)\cos x = 1; -10 \leq x \leq 10$
- d) $2 \tanh x - \sin x - 0,3 = 0$
- e) $x^4 - 4x^2 - 4x - 16 - \ln|x| = 0$
- f) $e^x \sin x - 2e^x + 3 = 0$
- g) $x = \tan^2 x; 0 \leq x \leq 2\pi$
- h) $\sqrt{x} = 2 \ln x$

4. Halle, con cinco cifras decimales exactas la mayor raíz de la ecuación $\sin \frac{1}{x} = x$.

5. Halle, con cinco cifras decimales exactas, los ceros, los puntos de extremo y los puntos de inflexión de la función

$$f(x) = x^3 - 7x^2 + 4x + 1 + \sin x$$

6. Determine, con error menor que 0,0001 los puntos de intersección de la circunferencia $x^2 + y^2 = 4$ y la exponencial $y = e^x$

7. Las curvas $y = \frac{1}{x}$, $y = x$ y $y = x^2$ forman un triángulo curvilíneo en el primer cuadrante uno de cuyos vértices es $(1, 1)$. Halle, con cuatro cifras decimales exactas los otros dos vértices de cada este triángulo.
8. Dada la función $y = x^x$ halle, con cinco cifras exactas el valor de x para el cual la función toma el valor 2 y el punto de la gráfica que tiene pendiente 2.
9. La figura 2 muestra la gráfica de la función $y = x^2 \sin x$ y la recta L que es tangente a la gráfica en el punto P y la corta en ese punto. Halle la ecuación de L . Realice los cálculos necesarios con cinco cifras decimales exactas.

$$f(x) = x^2 \sin x$$

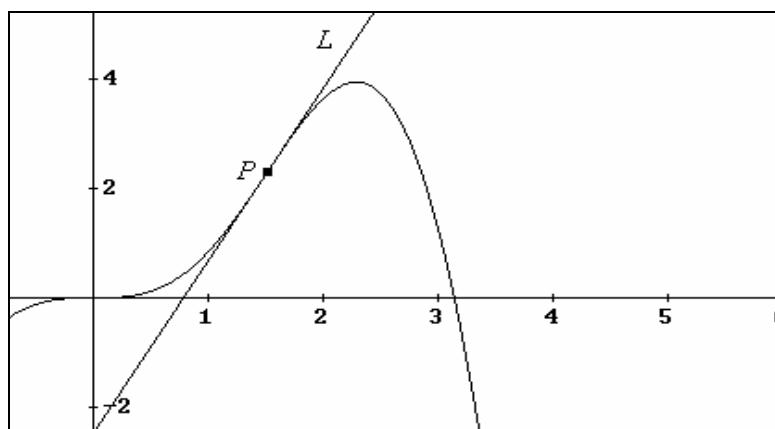


Figura 2

10. Dos escaleras de madera, una de tres metros y otra de cuatro metros de largo, están colocadas contra las paredes de dos edificios que limitan un pasillo, como muestra la figura 3. El punto P en que ambas se cruzan está a 1,5 metros del suelo. Determine el ancho del pasillo con una precisión de 1 milímetro.

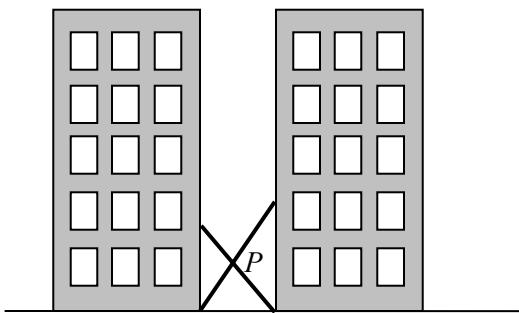


Figura 3

11. La esfera B tiene su radio un centímetro mayor que la esfera A y un centímetro menor que la esfera C . Se sabe que el volumen de la esfera C es igual a la suma de los volúmenes de A y B . Determine el radio de cada esfera con 5 cifras decimales exactas.

12. El algoritmo en seudo código que sigue es una ligera modificación del que se ofreció dentro de la sección; se han suprimido algunas instrucciones. Analice qué sucede con este algoritmo si, casualmente, $f(x)$ se anula en alguna de las aproximaciones que se obtiene durante el proceso.

```

repeat
     $x := \frac{a + b}{2}$ 
     $Error := \frac{b - a}{2}$ 
    if  $f(a)f(x) < 0$  then
         $b := x$ 
    else
         $a := x$ 
    end
until  $Error \leq \epsilon$ 
La raíz buscada es  $x$  y su error absoluto máximo es  $Error$ 
Terminar

```

13. Analice qué sucede en el algoritmo de la bisección si no se satisface la hipótesis $f(a)f(b) < 0$.
14. Analice qué sucede con el algoritmo de la bisección si siendo $f(a)f(b) < 0$ la ecuación $f(x) = 0$ no se satisface en el intervalo $[a, b]$ debido a que la función presenta una discontinuidad para $x = c$ en el intervalo $[a, b]$, tal como se muestra en la figura 4.

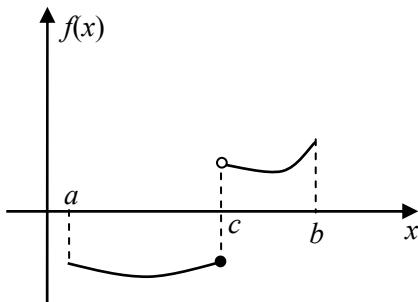


Figura 4

2.4 El método Regula Falsi

En el método de la bisección se approxima la raíz en cada iteración como el punto medio del intervalo de búsqueda; la lógica de esta opción es que, de esa forma se minimiza el error absoluto máximo. Sin embargo, puede haber otras opciones. Una de ellas es la que sigue el método Regula Falsi: si la raíz se encuentra en el intervalo $[a, b]$ es de suponer que esté más cerca de aquel extremo del intervalo donde la función $f(x)$ tome un valor más cercano a cero. En la figura 1 se muestran dos funciones para las cuales el valor de la función en b es mucho más próximo a cero que el valor en a ; en la primera función la raíz se encuentra próxima a b , que es lo esperado, pero en la segunda función, que muestra un comportamiento extraño, la raíz se halla más cercana a a .

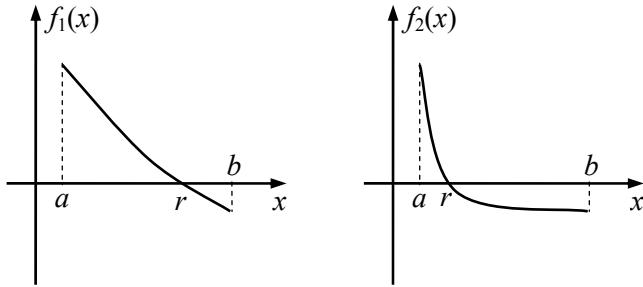


Figura 1

Como se ve, esta lógica no siempre funciona pero, la mayor parte de las veces sí y, en esos casos, se logra en cada iteración una aproximación mejor de la raíz y se llega más rápido a la exactitud deseada.

Hipótesis

Se desea hallar la raíz r de una ecuación $f(x) = 0$ que se encuentra en el intervalo $[a, b]$ y se cumplen las mismas hipótesis que en el método de bisección:

1. En el intervalo la ecuación tiene una sola raíz $x = r$.
2. $f(x)$ es continua en $[a, b]$
3. $f(x)$ posee signos diferentes en a y en b , es decir, $f(a) \cdot f(b) < 0$

El método

El nombre de este método proviene de una frase latina que significa *regla inclinada* y geométricamente consiste en tomar como aproximación de la raíz en el intervalo $[a_n, b_n]$ el punto de intersección con el eje x de un segmento que une los extremos del arco de la gráfica en ese intervalo. Por esta razón, también se le conoce como *método de las cuerdas*. En la figura 2 se muestra esta idea. Como el segmento AB determina con el eje x dos triángulos rectángulos semejantes, se puede establecer la proporcionalidad entre sus lados:

$$\frac{x_n - a_n}{|f(a_n)|} = \frac{b_n - x_n}{|f(b_n)|}$$

Es decir, que la distancia de x a cada extremo del intervalo es proporcional al valor absoluto de la función en ese extremo.

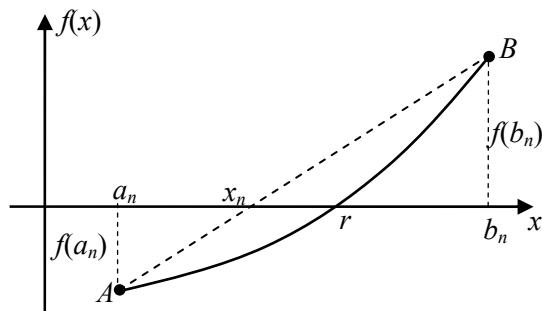


Figura 2

Para determinar el punto x_n se halla la ecuación de la recta que contiene al segmento AB :

$$y - f(a_n) = \frac{f(b_n) - f(a_n)}{b_n - a_n} (x - a_n)$$

y como x_n es la abscisa que corresponde a $y = 0$:

$$0 - f(a_n) = \frac{f(b_n) - f(a_n)}{b_n - a_n} (x_n - a_n)$$

Despejando x_n resulta:

$$x_n = a_n - \frac{b_n - a_n}{f(b_n) - f(a_n)} f(a_n) \quad (1)$$

Tal como en el método de bisección, una vez obtenido el valor x_n , se analiza el signo de $f(x_n)$ y de acuerdo con él se determina si la raíz r se encuentra en $[a_n, x_n]$ o en $[x_n, b_n]$ y el proceso se repite sucesivamente.

Convergencia del método

Antes de entrar en demostraciones, es conveniente ver gráficamente como se produce la convergencia en este proceso. En las figuras 3 y 4 se muestran dos de las situaciones más frecuentes: en la primera, $f(x)$ posee segunda derivada positiva (concavidad hacia arriba); en la figura 4 la segunda derivada es negativa (concavidad hacia abajo). Nótese que en ambos casos x_n tiende hacia r pero que, en la figura 3, $b_1 = b_2 = b_3 = \dots$ mientras a_n tiende hacia la raíz r , pero que en la figura 4, es b_n quien tiende hacia r y el extremo de la izquierda permanece fijo, esto es: $a_1 = a_2 = a_3 = \dots$

Aunque, excepcionalmente es posible que ambos extremos del intervalo tiendan hacia la raíz, y en ese caso la amplitud del intervalo $[a_n, b_n]$ tendería hacia cero, lo más frecuente es que uno de los extremos permanezca fijo mientras el otro tiende hacia la raíz; en este caso, la longitud del intervalo $[a_n, b_n]$ no tiende hacia cero. Esta es una importante diferencia entre este método y el de la bisección, en el cual ambos extremos del intervalo tienden hacia la raíz y su amplitud, por tanto, tiende hacia cero en cualquier caso.

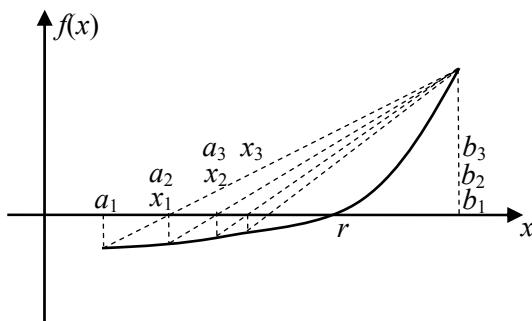


Figura 3

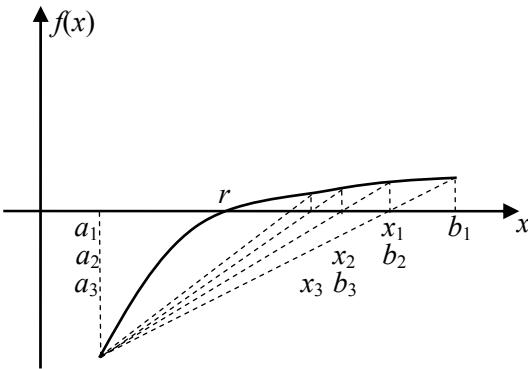


Figura 4

A pesar de que la longitud del intervalo $[a_n, b_n]$ no tiende en general hacia cero, se puede demostrar que, bajo las hipótesis hechas,

$$\lim_{n \rightarrow \infty} x_n = r$$

pero la demostración es muy larga, pues requiere considerar varias posibilidades diferentes. Si se imponen hipótesis más fuertes (que en la práctica generalmente se satisfacen) las demostraciones se simplifican y se obtienen expresiones más útiles. Estas hipótesis adicionales son:

4. $f(x)$ es derivable y $f'(x)$ no cambia de signo en $[a, b]$
5. Existe $f''(x)$ y no cambia de signo en $[a, b]$

Para precisar, considérese el caso en que $f'(x) > 0$ y $f''(x) > 0$ en $[a, b]$, que es el caso que ilustra la figura 3. Para las demás combinaciones de signos, la deducción siguiente se lleva cabo de forma similar.

Como se ve en la figura 3, la sucesión de aproximaciones en este caso satisface que:

$$\begin{aligned} x_n &> x_{n-1} & (x_n \text{ es creciente}) \\ x_n &\leq r & (x_n \text{ está acotada superiormente}) \end{aligned}$$

Por ser creciente y acotada superiormente, la sucesión x_n posee límite. Sea

$$\bar{x} = \lim_{n \rightarrow \infty} x_n$$

Esto prueba que el proceso iterativo es convergente, ahora hay que probar que converge hacia la solución de la ecuación $f(x) = 0$. En efecto, bajo las hipótesis realizadas, el extremo b_n se mantiene fijo en b mientras que $a_n = x_{n-1}$ para todo n . Sustituyendo en la ecuación (1):

$$x_n = x_{n-1} - \frac{b - x_{n-1}}{f(b) - f(x_{n-1})} f(x_{n-1}) \quad (2)$$

Pasando al límite para $n \rightarrow \infty$ y recordando que por ser $f(x)$ continua

$$\lim_{n \rightarrow \infty} f(x_n) = f(\lim_{n \rightarrow \infty} x_n) = f(\bar{x})$$

resulta:

$$\bar{x} = \bar{x} - \frac{b - \bar{x}}{f(b) - f(\bar{x})} f(\bar{x})$$

es decir:

$$\frac{b - \bar{x}}{f(b) - f(\bar{x})} f(\bar{x}) = 0$$

y esto solamente es posible si $f(\bar{x}) = 0$, o sea si \bar{x} es la raíz de la ecuación.

El error del método

Tan importante o más que el hecho de que el proceso iterativo converge hacia la solución es conocer con qué rapidez lo hace y poseer una estimación del error en cada iteración que permita un criterio práctico acerca de cuándo debe ser detenido el proceso.

Igual que antes, se supone que se cumplen las hipótesis 1, 2, 3, 4 y 5 y se considerará solamente el caso en que $f'(x) > 0$ y $f''(x) > 0$ en $[a, b]$. Otras combinaciones de signos llevan a los mismos resultados. Bajo estas suposiciones se obtuvo la ecuación (2), que se puede escribir:

$$x_n - x_{n-1} = -\frac{f(x_{n-1})}{f(b) - f(x_{n-1})} (b - x_{n-1})$$

Como $f(r) = 0$, se puede sumar a $f(x_{n-1})$:

$$x_n - x_{n-1} = -\frac{f(x_{n-1}) - f(r)}{f(b) - f(x_{n-1})} (b - x_{n-1})$$

Aplicando el teorema del valor medio a cada término del cociente:

$$x_n - x_{n-1} = -\frac{f'(\alpha)(x_{n-1} - r)}{f'(\beta)(b - x_{n-1})} (b - x_{n-1})$$

donde $\alpha \in (x_{n-1}, r)$ y $\beta \in (x_{n-1}, b)$. Simplificando

$$x_n - x_{n-1} = -\frac{f'(\alpha)}{f'(\beta)} (x_{n-1} - r) \quad (3)$$

y trasponiendo:

$$f'(\beta)(x_n - x_{n-1}) = f'(\alpha)(r - x_{n-1})$$

Sumando y restando x_n :

$$f'(\beta)(x_n - x_{n-1}) = f'(\alpha)(r - x_n + x_n - x_{n-1})$$

Esto es:

$$f'(\beta)(x_n - x_{n-1}) = f'(\alpha)(r - x_n) + f'(\alpha)(x_n - x_{n-1})$$

Despejando $r - x_n$:

$$r - x_n = \frac{f'(\beta) - f'(\alpha)}{f'(\alpha)} (x_n - x_{n-1}) \quad (4)$$

y tomando módulos:

$$|r - x_n| = \frac{|f'(\beta) - f'(\alpha)|}{|f'(\alpha)|} |x_n - x_{n-1}| \quad (5)$$

Sean ahora d y D cotas inferior y superior respectivamente de $|f'(x)|$ en $[a, b]$, esto es:

$$d \leq |f'(x)| \leq D \text{ en } [a, b]$$

Entonces la ecuación (5) implica que:

$$|r - x_n| \leq \frac{D - d}{d} |x_n - x_{n-1}| \quad (6)$$

Cuando el intervalo $[a, b]$ se selecciona lo suficientemente pequeño de modo que $|f'(x)|$ no sufra cambios grandes (esto significa que la gráfica no presenta pendientes muy diferentes) entonces sucede que

$$2d > D \quad (7)$$

en ese caso:

$$\frac{D - d}{d} < \frac{2d - d}{d} = 1$$

y la ecuación (6) se puede escribir: $|r - x_n| \leq |x_n - x_{n-1}|$

En otras palabras, la diferencia $|x_n - x_{n-1}|$ puede tomarse como error absoluto máximo de x_n y sirve como condición para terminar el proceso iterativo del método Regula Falsi. Nótese sin embargo, que esta condición está sujeta al cumplimiento de muchas hipótesis y el incumplimiento de cualquiera de ellas la invalida. Para su referencia posterior, es conveniente resumir el desarrollo anterior en un teorema:

Teorema 1

Si $f(x)$ es continua y dos veces derivable en $[a, b]$, $f(x)$ posee en $[a, b]$ una sola raíz siendo $f(a)$ y $f(b)$ de signos opuestos, $f'(x)$ y $f''(x)$ no cambian de signo en $[a, b]$ y existen números d y D tales que $d \leq |f'(x)| \leq D$ en $[a, b]$ y se cumple que $2d > D$, entonces el error absoluto máximo de la aproximación x_n obtenida por el método Regula Falsi puede tomarse como:

$$E_m(x_n) = |x_n - x_{n-1}|$$

■

Condición de terminación:

Si se desea obtener la raíz de la ecuación con un error absoluto menor que ε y se satisfacen las hipótesis del teorema 1, el método Regula Falsi se llevará a cabo hasta la aproximación x_n para la cual

$$E_m(x_n) = |x_n - x_{n-1}| \leq \varepsilon$$

Rapidez de la convergencia

Está claro que la mayor complejidad del método Regula Falsi y la necesidad de verificar muchas más hipótesis lo hacen menos atractivo que el método de bisección. Por ello, solo se justifica su uso si se obtiene una velocidad de convergencia mucho mayor que la que logra el método de bisección. La ecuación (3) de la sección anterior caracteriza la forma en que converge el método de bisección:

$$E_m(x_{n+1}) = \frac{1}{2} E_m(x_n)$$

Es decir, en cada iteración el error absoluto máximo disminuye a la mitad. Para Regula Falsi una expresión similar se puede obtener de la siguiente forma. Al analizar el error del método se obtuvo la expresión (3):

$$x_n - x_{n-1} = -\frac{f'(\alpha)}{f'(\beta)}(x_{n-1} - r)$$

de la cual:

$$r - x_{n-1} = \frac{f'(\beta)}{f'(\alpha)}(x_n - x_{n-1}) \quad (8)$$

Más adelante, se llegó a la ecuación (4):

$$r - x_n = \frac{f'(\beta) - f'(\alpha)}{f'(\alpha)}(x_n - x_{n-1})$$

Si esta ecuación se divide miembro a miembro por la igualdad (8) resulta:

$$\frac{r - x_n}{r - x_{n-1}} = \frac{f'(\beta) - f'(\alpha)}{f'(\beta)}$$

Esto es, en término de errores:

$$E(x_n) = \left| \frac{f'(\beta) - f'(\alpha)}{f'(\beta)} \right| E(x_{n-1})$$

Teniendo en cuenta que $d \leq |f'(x)| \leq D$:

$$E(x_n) \leq \frac{D - d}{d} E_m(x_{n-1})$$

Puede entonces tomarse:

$$E_m(x_n) = \frac{D - d}{d} E_m(x_{n-1}) \quad (9)$$

En el caso en que el término $\frac{D - d}{d}$ es menor que 0,5 el método Regula Falsi presenta una rapidez de convergencia mayor que bisección. Esto sucederá si la diferencia entre la menor pendiente d y la mayor pendiente D de la gráfica de $f(x)$ es pequeña, lo cual en palabras simples significa que en el intervalo $[a, b]$ dicha gráfica presenta poca curvatura. Si por el contrario, el término

$\frac{D - d}{d}$ toma valores mayores que 0,5 la convergencia es peor que en bisección y, de hecho, se presentan ecuaciones en las que la convergencia del método es extraordinariamente lenta.

Algoritmo en seudo código

Se supone que la ecuación a resolver es $f(x) = 0$, que la raíz que se quiere hallar está separada dentro de un intervalo $[a, b]$ en el cual $f(x)$ es continua y que $f(a)f(b) < 0$ y se cumplen las demás hipótesis del teorema 1. Se suponen conocidas la función $f(x)$, los extremos a y b del intervalo de separación y la tolerancia ε que se permitirá.

```
xanterior := 106
repeat
    x := a -  $\frac{b-a}{f(b)-f(a)} f(a)$ 
    Error := |x - xanterior|
    if f(x) = 0 then
        x es exactamente la raíz buscada
        Terminar
    else
        if f(a)f(x) < 0 then
            b := x
        else
            a := x
        end
    end
    xanterior := x
until Error ≤ ε
La raíz buscada es x y su error absoluto máximo es Error
Terminar
```

Nótese que se necesita guardar el valor de la aproximación anterior con vistas a calcular el error. La primera vez que se calcula x , sin embargo, no existe aproximación anterior y por eso se le ha dado un valor grande de modo que el primer error obtenido sea tan grande que el algoritmo no se detenga en la primera iteración.

Comentarios finales

El método Regula Falsi puede considerarse como una codificación del método de bisección para mejorar la velocidad de convergencia. Aunque para que esta se produzca basta con que se cumplan hipótesis muy sencillas, para lograr una buena velocidad se requieren condiciones más fuertes, en particular que la gráfica de la función $f(x)$ presente poca curvatura. En el caso extremo en que la gráfica es lineal, la convergencia se produce en una sola iteración.

La condición de curvatura pequeña en el intervalo de búsqueda siempre puede lograrse reduciendo la amplitud de $[a, b]$, lo cual es muy sencillo si se puede visualizar la gráfica de $f(x)$ en un display. Si la condición $2d > D$ no se puede garantizar, se corre el riesgo no solo de obtener una pobre velocidad de convergencia sino que la fórmula para acotar el error deja de ser válida y da cotas del error que no son ciertas.

El algoritmo es bastante sencillo de programar y difiere del de bisección solamente en algunos detalles.

Ejemplo 1

Halle, con cinco cifras decimales exactas mediante el método Regula Falsi, la mayor raíz positiva de la ecuación:

$$x^4 + x^3 - 8x^2 - 12x - 1 = 0$$

Solución:

Primero se necesita separar la raíz buscada. Como se trata de una ecuación algebraica, es fácil acotar el número y la posición de las raíces. La regla de Descartes asegura que hay solo una raíz positiva, que es precisamente la que se desea hallar. Una cota superior para esa raíz la da la regla de Lagrange:

$$R = 1 + \sqrt[k]{\frac{B}{a_0}} = 1 + \sqrt[4]{\frac{12}{1}} < 4,5$$

La gráfica en un intervalo $[0, 5]$ se muestra en la figura 5. Se aprecia una raíz muy próxima a 3. Para garantizar las condiciones de poca curvatura, se tomará un intervalo de pequeña amplitud que contenga la raíz. En la figura 6 se muestra la gráfica en $[2,8; 3,2]$ donde se puede ver que las hipótesis de continuidad, derivadas y curvatura se satisfacen. En la tabla 1 se muestran los resultados obtenidos con un programa confeccionado a partir del algoritmo en seudo código.

Observe que, utilizando el método de bisección con este mismo intervalo inicial, para obtener cinco cifras decimales exactas, es decir $\epsilon = 0,000005$ se habría necesitado

$$n \geq \frac{\ln\left(\frac{b-a}{\epsilon}\right)}{\ln 2} = \frac{\ln\left(\frac{3,2-2,8}{0,000005}\right)}{\ln 2} = 16,29$$

es decir, no menos de 17 iteraciones, en lugar de las 7 que se necesitó con Regula Falsi.

$$f(x) = x^4 + x^3 - 8x^2 - 12x - 1$$

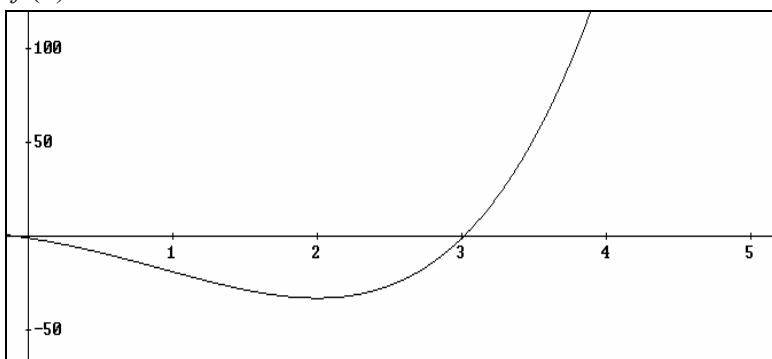


Figura 5

$$f(x) = x^4 + x^3 - 8x^2 - 12x - 1$$

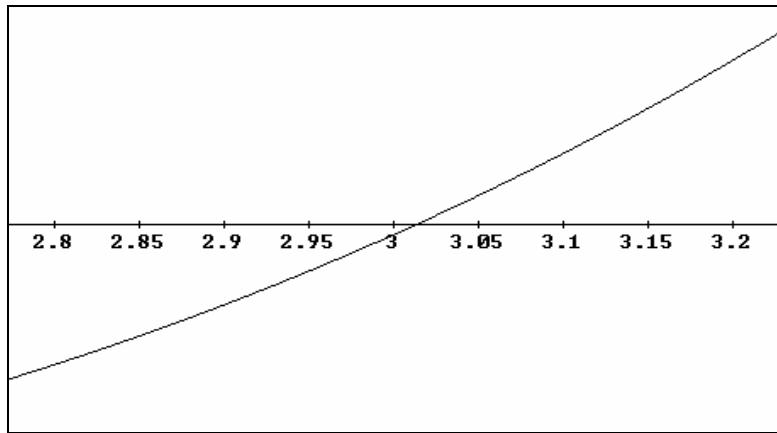


Figura 6

Iteración	a	b	x	$E_m(x)$
1	2,8	3,2	2,984089	
2	2,984089	3,2	3,009546	0,025457
3	3,009546	3,2	3,012750	0,003205
4	3,012750	3,2	3,013149	0,000298
5	3,013149	3,2	3,013198	0,000049
6	3,013198	3,2	3,013204	0,000006
7	3,013204	3,2	3,013205	0,000001

Tabla 1

Resultado: La mayor raíz real de la ecuación es 3,013205 con cinco cifras decimales exactas.

Ejemplo 2

Aplique el método Regula Falsi para resolver la ecuación $\frac{1}{x} - \frac{1}{5}$ con error absoluto menor que 0,005. Tome como intervalo de búsqueda [1,5; 7]. Analice los resultados obtenidos.

Solución:

En el intervalo dado, la función $f(x) = \frac{1}{x} - \frac{1}{5}$ cambia de signo y es continua y derivable cualquier número de veces. Si se observa la gráfica (figura 7), se aprecia, sin embargo que en este intervalo la pendiente toma valores muy diferentes, lo cual hace prever un pobre comportamiento. En efecto, la tabla 2 muestra los resultados obtenidos.

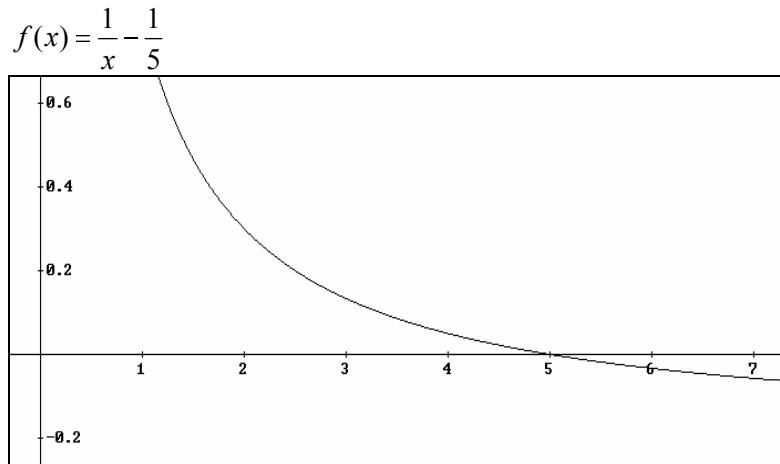


Figura 7

Iteración	a	b	x	$E_m(x)$
1	1,5	7	6,4	
2	1,5	6,4	5,98	0,42
3	1,5	5,98	5,686	0,294
4	1,5	5,686	5,4807	0,2058
5	1,5	5,4807	5,33614	0,14406
6	1,5	5,33614	5,235298	0,100842
7	1,5	5,235298	5,164709	0,070589
8	1,5	5,164709	5,115296	0,049413
9	1,5	5,115296	5,080707	0,034589
10	1,5	5,080707	5,056495	0,024212
11	1,5	5,056495	5,039547	0,016949
12	1,5	5,039547	5,027683	0,011864
13	1,5	5,027683	5,019378	0,008303
14	1,5	5,019378	5,013564	0,005813
15	1,5	5,013564	5,009495	0,004069

Tabla 2

Observe que en 15 iteraciones la aproximación que se obtiene es 5,013564 de la solución exacta, que es $x = 5$, esto significa un error de 0,013564. Con 15 iteraciones, el método de bisección habría logrado un error:

$$E_m(x_n) = \frac{b-a}{2^n} = \frac{7-1,5}{2^{15}} = 0,000168$$

Por otra parte, las cotas de los errores que se muestran en la tabla 2 no son ciertas. Todo este comportamiento se debe a que la condición $2d > D$ no se satisface en el intervalo escogido. Bajo estas condiciones el método Regula Falsi posee una convergencia muy lenta y la fórmula $E_m(x_n) = |x_n - x_{n-1}|$ para obtener el error, pierde su validez.

Ejercicios

En todos los ejercicios que siguen se debe utilizar, siempre que se necesite, un programa graficador tanto para separar las raíces como para verificar las hipótesis del método utilizado. También se supone que el algoritmo Regula Falsi se utilice mediante un programa computacional, preferiblemente confeccionado por usted. Si no cuenta con un programa adecuado, realice los cálculos a mano y obtenga las raíces con solo dos o tres cifras decimales exactas.

1. Calcule, con cinco cifras decimales exactas, las raíces reales de las siguientes ecuaciones algebraicas. Posiblemente, ya usted separó las raíces de estas ecuaciones en los ejercicios de la sección 2.2 y las resolvió por bisección en los ejercicios de la sección 2.3. Compare la cantidad de iteraciones que necesitó con las que necesitaría el método de bisección.
 - a) $x^4 + x^3 - x^2 + x - 2 = 0$
 - b) $x^4 - 11x^3 + 41x^2 - 60x + 30 = 0$
 - c) $x^4 - 3x^3 + 10x^2 - 13x + 5 = 0$
 - d) $x^4 + 3x^2 + 2 = 0$
 - e) $x^3 + 7x^2 + 14x + 9 = 0$
2. Utilice el método Regula Falsi para calcular con cinco cifras decimales exactas, las raíces de las siguientes ecuaciones trascendentes. Posiblemente, ya usted separó las raíces de estas ecuaciones en los ejercicios de la sección 2.2 y las resolvió por el método de bisección. En ese caso, compare la cantidad de iteraciones que necesito con cada método.
 - a) $\sin x - \log x = 0$
 - b) $5e^x - 2x - 10 = 0$
 - c) $(x^2 + 1)\cos x = 1; -10 \leq x \leq 10$
 - d) $2 \tanh x - \sin x - 0,3 = 0$
 - e) $x^4 - 4x^2 - 4x - 16 - \ln|x| = 0$
 - f) $e^x \sin x - 2e^x + 3 = 0$
 - g) $x = \tan^2 x; 0 \leq x \leq 2\pi$
 - h) $\sqrt{x} = 2 \ln x$
3. Aplique el método Regula Falsi para determinar, con cuatro cifras decimales exactas las coordenadas cartesianas del punto del primer cuadrante donde se intersecan la parábola cúbica $y = x^3$ y la conchoide $r = 3 - \cos \theta$.
4. Halle, con error menor que 0,0001 la pendiente de una recta que pasa por el punto (5, 1) y es tangente en el primer cuadrante a la elipse $4x^2 + 9y^2 = 36$. Utilice el método Regula Falsi.
5. El cuadrado de un número positivo menos la raíz cuadrada del número da 1. Determine el número con 4 cifras decimales exactas empleando el método Regula Falsi
6. Halle con cinco cifras decimales exactas el valor de a de manera que el área sombreada de la figura 8 tome el valor 2.

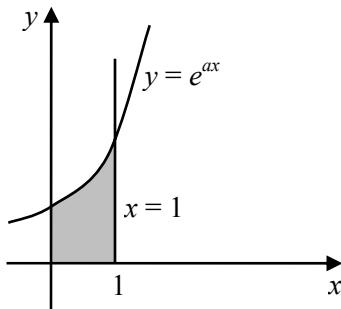


Figura 8

7. Al comienzo de este capítulo se analizó (vea el ejemplo 3 de la sección 2.1) un problema en que se requería hallar la ecuación de un cable colgante conociendo las coordenadas de dos de sus puntos. Se llegó a la siguiente ecuación cuya solución fue pospuesta:

$$a \cosh \frac{10}{a} - a = 2$$

Halle mediante Regula Falsi la raíz positiva de esta ecuación con error menor que 10^{-4} .

8. La ecuación $x \tan x = 1$ tiene una sola raíz en el intervalo $[0,5; 1,4]$. Aplique el método Regula Falsi (utilice este intervalo) para hallar esta raíz con cuatro cifras decimales exactas. Calcule cuántas iteraciones del método de Bisección se necesitarían para resolver el mismo problema. Explique.
9. A continuación aparece un algoritmo en seudo código que es una modificación del algoritmo Regula Falsi que se mostró en esta sección. Se ha tomado como cota del error la longitud del intervalo $[a, b]$, lo cual es cierto; sin embargo este algoritmo no funciona. Explique por qué.

```

repeat
     $x := a - \frac{b-a}{f(b)-f(a)} f(a)$ 
     $Error := b - a$ 
    if  $f(x) = 0$  then
         $x$  es exactamente la raíz buscada
        Terminar
    else
        if  $f(a)f(x) < 0$  then
             $b := x$ 
        else
             $a := x$ 
        end
    end
until  $Error \leq \epsilon$ 
La raíz buscada es  $x$  y su error absoluto máximo es  $Error$ 
Terminar.

```

10. En el algoritmo Regula Falsi se utiliza la fórmula $x := a - \frac{b-a}{f(b)-f(a)} f(a)$. ¿La presencia de la diferencia $f(b) - f(a)$ puede provocar pérdida de significación si $f(a)$ y $f(b)$ son números muy parecidos? Explique.
11. En el algoritmo Regula Falsi usado en esta sección no se prestó atención a la cantidad de veces que debe ser evaluada la función $f(x)$ en cada iteración. Observe que en la fórmula $x := a - \frac{b-a}{f(b)-f(a)} f(a)$ se requieren 3 evaluaciones y otras 2 para analizar si $f(a)f(x) < 0$. Perfeccione el algoritmo en seudo código de manera que en cada iteración solo haya que evaluar $f(x)$ una vez.

2.5 El método de Newton – Raphson

Los dos algoritmos para resolver ecuaciones vistos hasta aquí tienen en común el hecho de que se trata de métodos de intervalos. En ellos se comienza con un intervalo de búsqueda y todo el proceso ocurre dentro de este intervalo. Los métodos que ahora se estudiarán son métodos de puntos, no de intervalos. Todos estos métodos funcionan de manera similar: se tiene una aproximación inicial x_0 de la raíz de la ecuación y, mediante un proceso más o menos simple, se obtiene otra aproximación x_1 ; este mismo proceso aplicado sobre x_1 da lugar a la aproximación x_2 y sucesivamente, se obtienen los elementos de una sucesión de aproximaciones. Bajo ciertas condiciones, esta sucesión converge hacia la raíz buscada.

Antes de pasar al método de Newton – Raphson, es conveniente analizar algunos aspectos generales de los procesos de este tipo.

El método iterativo en general

Los métodos iterativos de punto, o más brevemente, iterativos, pueden representarse así:

$$\begin{aligned} x_0 &\text{ es conocido} \\ x_n &= g(x_{n-1}) \quad n = 1, 2, 3, \dots \end{aligned} \tag{1}$$

Es fácil demostrar que, si g es continua y la sucesión x_n generada de esta forma posee un límite finito \bar{x} entonces este límite es una raíz de la ecuación $x = g(x)$. En efecto, tomando límites en ambos miembros de (1):

$$\lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} g(x_{n-1})$$

Como g es continua: $\lim_{n \rightarrow \infty} x_n = g(\lim_{n \rightarrow \infty} x_{n-1})$

Como ambos límites existen y valen \bar{x} :

$$\bar{x} = g(\bar{x})$$

lo cual significa que \bar{x} es solución de la ecuación $x = g(x)$.

La idea anterior es realmente muy sugestiva: si se quiere hallar una raíz de la ecuación $f(x) = 0$, bastaría escribirla de la forma $x = g(x)$ (lo cual siempre es posible y de muchas formas) y definir un proceso iterativo del tipo (1) a partir de un cierto valor x_0 inicial; si la sucesión así generada converge hacia un límite finito, ese límite es una solución de la ecuación original.

Ejemplo 1

Intente resolver las siguientes ecuaciones definiendo el proceso iterativo que resulta de escribir la ecuación como se indica y tomar la aproximación inicial dada

a) $\cos x - x = 0$ escrita como: $x = \cos x$ tomando $x_0 = 0$

b) $x^2 = 2$ mediante: $x = \frac{2}{x}$ tomando $x_0 = 1$

c) $x^2 = 2$ mediante: $x = \frac{x + \frac{2}{x}}{2}$ tomando $x_0 = 1$

Solución:

a) Se escoge $x_0 = 0$ y $x_n = \cos x_{n-1}$ para $n = 1, 2, 3, \dots$ Se obtienen sucesivamente los valores:

$$x_0 = 0,0000; x_1 = 1,0000; x_2 = 0,5403; x_3 = 0,8576; x_4 = 0,6543; x_5 = 0,7935; \dots x_{15} = 0,7401;$$

$$x_{16} = 0,7384; x_{17} = 0,7396; x_{18} = 0,7388; \dots x_{25} = 0,739106; x_{26} = 0,739071; \dots$$

Se observa que la sucesión converge hacia un límite cuyas primeras cifras son 0,7390. Después de 45 iteraciones la convergencia se hace evidente:

$$x_{45} = 0,739085140; x_{46} = 0,739085126; x_{47} = 0,739085130 \dots$$

Con seis cifras decimales exactas, el límite de la sucesión es $\bar{x} = 0,739085$. Esta es una solución de la ecuación $\cos x - x = 0$ con seis cifras decimales exactas.

b) Se escoge $x_0 = 1$ y $x_n = \frac{2}{x_{n-1}}$ para $n = 1, 2, 3, \dots$ Se obtienen los valores:

$$x_0 = 1; x_1 = 2; x_2 = 1; x_3 = 2; x_4 = 1; x_5 = 2; \dots$$

Evidentemente, en este caso la sucesión obtenida no converge hacia ningún límite.

c) Observe primero que la ecuación $x = \frac{x+2}{2}$ se obtiene de $x^2 = 2$ por transformaciones de equivalencia. Partiendo de $x^2 = 2$

Se divide por $x \neq 0$:

$$x = \frac{2}{x}$$

Sumando x a cada miembro:

$$2x = x + \frac{2}{x}$$

y dividiendo por 2:

$$x = \frac{x + \frac{2}{x}}{2}$$

Para obtener la sucesión se hace: $x_0 = 1$ y $x_n = \frac{x_{n-1} + \frac{2}{x_{n-1}}}{2}$ para $n = 1, 2, 3, \dots$

Se obtienen sucesivamente los valores: $x_0 = 1$; $x_1 = 1,5$; $x_2 = 1,416666666$; $x_3 = 1,414215686$; $x_4 = 1,414213562$; $x_5 = 1,414213562 \dots$

Como se observa, después de la cuarta aproximación, las primeras 9 cifras decimales se mantienen fijas. Se ha obtenido una solución de la ecuación $x^2 = 2$ con una gran exactitud en muy pocos pasos. ■

Del ejemplo anterior se puede concluir varias cosas:

- El proceso iterativo a veces converge y otras no. Incluso una misma ecuación puede generar un proceso convergente o divergente.
- Cuando el proceso iterativo converge puede hacerlo lentamente, como en el inciso a) o rápidamente como en el c).
- Si la ecuación original tiene más de una raíz (como en el inciso c) y el proceso iterativo converge lo hace hacia una de las raíces, las otras no se obtienen.

Evidentemente, se necesita investigar varios aspectos: ¿Por qué unas veces el proceso converge y otras no? ¿De qué depende la rapidez de convergencia? ¿Cómo saber a qué raíz convergerá el proceso? Todas estas preguntas las responde un solo teorema:

Teorema 1

Sea r una raíz de la ecuación $x = g(x)$ e I un entorno de r en el cual g y g' son continuas y se cumple que, para alguna constante $K < 1$, $|g'(x)| \leq K$ para todo x de I . Entonces, la sucesión generada por el proceso iterativo:

$$x_0 \in I; \quad x_n = g(x_{n-1}) \text{ para } n = 1, 2, 3, \dots$$

converge hacia r , es decir:

$$\lim_{n \rightarrow \infty} x_n = r$$

Demostración:

Como r es una raíz de $x = g(x)$, se cumplirá que $r = g(r)$. Si esta igualdad se resta miembro a miembro de $x_n = g(x_{n-1})$ se obtiene:

$$x_n - r = g(x_{n-1}) - g(r) \tag{2}$$

Primero se utilizará el principio de inducción completa para probar que todos los x_n están en el entorno I . La aproximación inicial cumple la condición por hipótesis. Supóngase ahora que $x_{n-1} \in I$. Como $r \in I$ (ya que I es un entorno de r), entonces la función g es continua y derivable

en el intervalo que determinen x_{n-1} y r . Bajo estas condiciones, se puede aplicar el teorema del valor medio en el miembro derecho de la ecuación (2) y resulta:

$$x_n - r = g'(c)(x_{n-1} - r) \quad (3)$$

para algún c entre x_{n-1} y r . Puede entonces afirmarse que $c \in I$ y, por tanto, se cumple que

$$|g'(c)| \leq K < 1$$

Tomando módulos en (3) y sustituyendo $|g'(c)|$ por K :

$$|x_n - r| \leq K|x_{n-1} - r| \quad (4)$$

Como $K < 1$:

$$|x_n - r| < |x_{n-1} - r|$$

Esto indica que la distancia entre x_n y r es menor que la que hay entre x_{n-1} y r . Como $x_{n-1} \in I$ entonces, con más razón $x_n \in I$. Esto demuestra que, para todo n , $x_n \in I$.

Ahora se demuestra que la sucesión converge hacia r . Para ello observe que, por estar en I todos los términos x_n , la ecuación (4) es válida para todo n :

$$|x_n - r| \leq K|x_{n-1} - r| \quad \text{Para } n = 1, 2, 3, \dots$$

Esta ecuación se puede escribir en términos de errores absolutos:

$$E(x_n) \leq KE(x_{n-1}) \quad \text{Para } n = 1, 2, 3, \dots \quad (5)$$

o también:

$$E_m(x_n) = KE_m(x_{n-1}) \quad \text{Para } n = 1, 2, 3, \dots \quad (6)$$

Aplicando la ecuación (5) n veces se tiene:

$$E(x_n) \leq KE(x_{n-1}) \leq K^2 E(x_{n-2}) \leq \dots \leq K^n E(x_0)$$

Esto es:

$$E(x_n) \leq K^n E(x_0)$$

Cuando n tiende hacia infinito, el término de la derecha tiende hacia cero por ser $K < 1$ y se tiene:

$$\lim_{n \rightarrow \infty} E(x_n) = 0$$

es decir:

$$\lim_{n \rightarrow \infty} x_n = r \quad \text{como se quería probar.} \quad \blacksquare$$

Como parte de la demostración, queda aclarado el problema de la velocidad de convergencia. En la ecuación (6) se muestra que el error absoluto máximo en cada paso se reduce a K veces el error absoluto máximo del paso anterior. Como K es una cota superior de $|g'(x)|$ resulta claro que, si la derivada de $g(x)$ es pequeña en valor absoluto en un entorno de r , se puede acotar con una K

pequeña y la rapidez de convergencia será alta. Por el contrario valores de $|g'(x)|$ próximos a 1 en las cercanías de r significarán una velocidad de convergencia baja.

Ejemplo 2

Analice los valores que toman las derivadas de las funciones $g(x)$ del ejemplo 1 en un entorno de la raíz de la ecuación y verifique el cumplimiento del teorema 1.

Solución:

a) $g(x) = \cos x$, así que $g'(x) = -\operatorname{sen} x$. Como la raíz está muy próxima a 0,74, se tiene que:

$$|g'(r)| \approx |\operatorname{sen}(0,74)| = 0,674$$

Es de suponer que en un entorno de r los valores de K sean de este orden, es decir, números próximos a 1. Esto explica la convergencia lenta del proceso iterativo del inciso a).

b) $g(x) = \frac{2}{x}$. La derivada es: $g'(x) = -\frac{2}{x^2}$. En la raíz de la ecuación $r = \sqrt{2}$, se tiene $g'(r) = -1$. En cualquier entorno de esta raíz, la derivada tomará valores absolutos mayores que 1, por lo cual no se pueden satisfacer las condiciones del teorema de convergencia. Como el teorema solo da condiciones suficientes para la convergencia (no necesarias), esto no asegura la divergencia del proceso pero no permite asegurar la convergencia.

c) $g(x) = \frac{1}{2} \left(x + \frac{2}{x} \right) = \frac{x}{2} + \frac{1}{x}$. Derivando: $g'(x) = \frac{1}{2} - \frac{1}{x^2}$

En la raíz $r = \sqrt{2}$ de la ecuación se tiene: $g'(r) = \frac{1}{2} - \frac{1}{2} = 0$. Esto asegura valores muy próximos a cero de $g'(x)$ en un entorno de la raíz y esto determina una alta velocidad de convergencia, como se vio en el ejemplo 1 c). ■

El método iterativo general es muy importante desde un punto de vista teórico pero su utilización práctica es limitada debido a que la manera en que la ecuación $f(x) = 0$ se transforma en $x = g(x)$ decide si se obtendrá un buen algoritmo iterativo o si será uno de convergencia lenta o incluso divergente. Claro que siempre se podría probar varias alternativas y, con el análisis de la derivada, tomar alguna que garantice una convergencia rápida. A menos que se trate de una ecuación que deba ser resuelta muchas veces para diferentes valores de algunos parámetros, no vale la pena emplear tanto tiempo en un análisis complicado cuando existen otros procedimientos numéricos más sencillos para determinar raíces.

El método de Newton – Raphson

Este importante método numérico puede concebirse como una forma sistemática de aplicar el método iterativo de manera que se obtenga una rápida convergencia.

Considérese la ecuación

$$f(x) = 0$$

cuya raíz r se desea hallar. La función $f(x)$ se supone derivable todas las veces necesarias en las proximidades de r . Si la ecuación se multiplica por una constante $A \neq 0$ y se suma x en cada miembro, se obtiene la ecuación equivalente:

$$x = x + Af(x) \quad (7)$$

La ecuación se ha escrito de la forma $x = g(x)$, donde $g(x) = x + Af(x)$

La idea es hallar un valor de A tal que la derivada de $g(x)$ sea muy pequeña en valor absoluto en las proximidades de r . Derivando se obtiene:

$$g'(x) = 1 + Af'(x)$$

para $x = r$ se tiene: $g'(r) = 1 + Af'(r)$

A menos que $f'(r)$ se anule, se puede seleccionar un valor de A para el cual sea $g'(r) = 0$. En efecto:

$$g'(r) = 0 \Leftrightarrow 1 + Af'(r) = 0 \Leftrightarrow A = -\frac{1}{f'(r)} \quad \text{donde } f'(r) \neq 0$$

La ecuación (7) quedaría entonces como: $x = x - \frac{f(x)}{f'(r)}$

la cual define el proceso iterativo: $x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(r)}$ (8)

el cual posee una velocidad de convergencia muy alta, debido a que los valores de la derivada de $g(x)$ en un entorno de r son muy próximos a cero. El tamaño del entorno estaría limitado por la proximidad de x_0 a r .

El proceso iterativo (8), teóricamente impecable, presenta una dificultad práctica muy importante. Se supone conocida la raíz r que es precisamente lo que se desea hallar. Si en lugar de evaluar la derivada en r se evalúa en x_{n-1} , que es la mejor aproximación que se tiene para r , se obtiene un proceso iterativo quizás no tan rápido como (8) pero que si resulta aplicable:

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \quad (9)$$

Este algoritmo recibe el nombre de método de Newton – Raphson y es posiblemente uno de los más importantes de toda la Matemática Numérica por la cantidad de aplicaciones directas y la gran cantidad de generalizaciones, modificaciones y aplicaciones que se han hecho de él.

Desde ahora se puede ver que la selección de x_0 será muy importante para determinar la convergencia del algoritmo a la raíz deseada con una velocidad alta. Este asunto será tratado posteriormente con más detalle.

Interpretación geométrica

Sea $f(x) = 0$ la ecuación cuya raíz r se desea hallar. En la figura 1 se muestra la gráfica de la función $y = f(x)$ donde r es el intercepto con el eje x . Se supone que x_{n-1} es un valor de x próximo a r . Este valor determina un punto sobre la gráfica, cuyas coordenadas son $(x_{n-1}, f(x_{n-1}))$ y que en la figura se denota por P_{n-1} . Por ese punto se ha trazado una recta L tangente a la gráfica cuya pendiente es $f'(x_{n-1})$. La ecuación de L es:

$$y - f(x_{n-1}) = f'(x_{n-1})(x - x_{n-1})$$

haciendo $y = 0$ se obtiene el intercepto de L con el eje x :

$$x = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$$

Como el intercepto es el mismo valor de x_n que brinda el método de Newton – Raphson, resulta la interpretación geométrica que se muestra en la figura 1: el algoritmo consiste en tomar como aproximación x_n el intercepto con el eje x de la recta tangente a la gráfica de $f(x)$ en el punto que determina x_{n-1} ; una vez obtenido x_n se determina una nueva recta tangente cuyo intercepto es x_{n+1} y el proceso continúa de tal manera que las aproximaciones sucesivas convergen rápidamente hacia r .

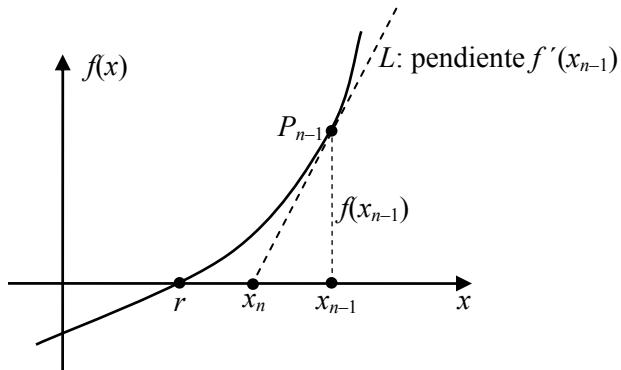


Figura 1

Convergencia del método de Newton – Raphson

Antes de pasar a enunciar y demostrar el teorema correspondiente, es útil ver, desde un punto de vista geométrico algunas situaciones en las que la convergencia del proceso iterativo no se produciría.

En la figura 2 se observa como la presencia en las cercanías de r de un punto donde la derivada se anula y cambia de signo puede provocar o bien la aparición de una tangente horizontal, que no cortará al eje x (y provocará en el algoritmo una división por cero) o de una aproximación x_n muy alejada de la raíz y donde incluso podría no estar definida la función $f(x)$.

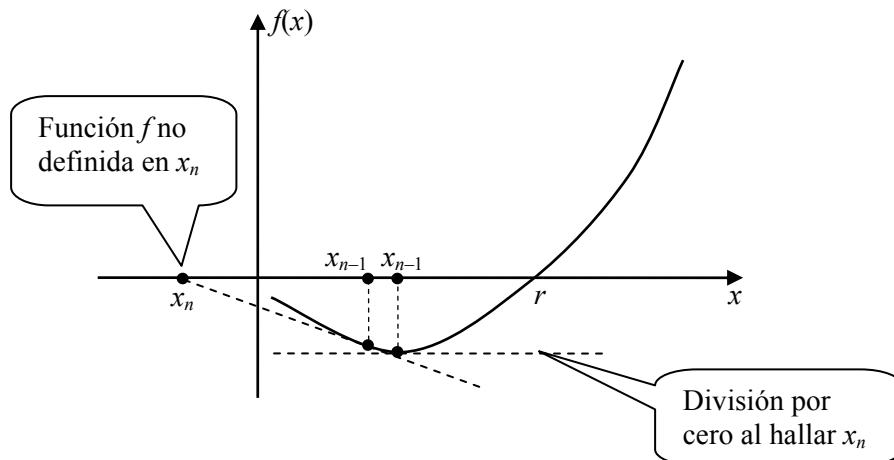


Figura 2

En la figura 3 se ilustra lo que puede suceder si en las cercanías de la raíz existe algún punto donde la segunda derivada cambia de signo (punto de inflexión). Nótese como al tomar x_{n-1} como P se obtiene $x_n = Q$ y viceversa, tomando Q como una aproximación se obtiene P como la siguiente. De esta manera, la sucesión de aproximaciones sería ... P, Q, P, Q, P, Q, \dots indefinidamente.

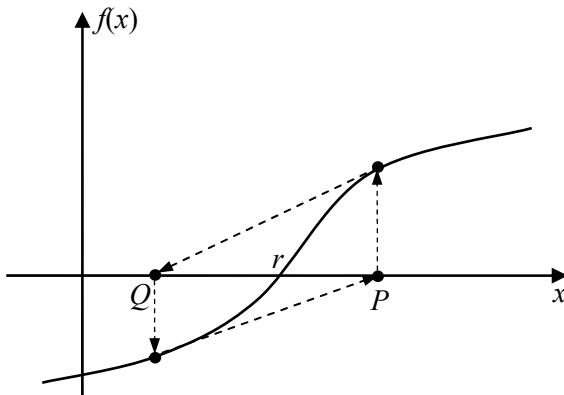


Figura 3

Por último, la figura 4 indica la importancia de seleccionar acertadamente la aproximación inicial x_0 ; véase cómo para esta ecuación al tomar como x_0 el punto a se obtiene como x_1 un punto muy alejado de la raíz r , en el que $f(x)$ puede no estar definida o no cumplir las condiciones necesarias para que el algoritmo converja. Sin embargo, al seleccionar x_0 como el punto b se asegura una convergencia rápida hacia la raíz r .

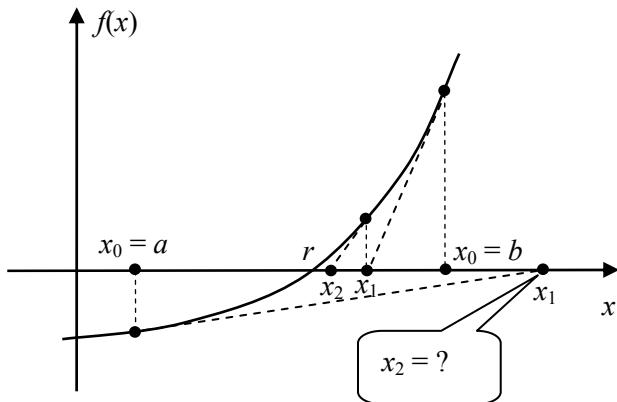


Figura 4

Teorema 2

Sea r la única raíz de $f(x) = 0$ en $[a, b]$. Sean $f'(x)$ y $f''(x)$ continuas y no nulas en $[a, b]$. Sea x_0 un elemento de $[a, b]$ tal que $f(x_0)f''(x_0) > 0$. Entonces, si

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \quad \text{para } n = 1, 2, 3, \dots$$

se cumple que $\lim_{x \rightarrow \infty} x_n = r$.

Demostración:

Como las derivadas son funciones continuas que no se anulan en $[a, b]$, entonces ellas no pueden cambiar su signo en ese intervalo. La demostración que sigue debe considerar las cuatro combinaciones posibles de signo para f' y f'' . Por razones de espacio solo se verá el caso en que ambas derivadas son positivas; las otras tres combinaciones se llevan a cabo de manera similar. Se supone entonces en todo lo que sigue que en $[a, b]$ $f'(x) > 0$ y $f''(x) > 0$, que es un caso como el que muestra la figura 4.

Está claro que por ser $f'(x) > 0$, $f(x)$ es creciente en $[a, b]$, así que tiene signos distintos a uno y otro lados de r . Por otra parte, como $f''(x)$ mantiene un solo signo, existirán muchos puntos (en este caso todos los que se encuentran a la derecha de r) donde se cumple que $f(x)f''(x) > 0$.

Esto significa que es posible seleccionar x_0 de manera que satisfaga $f(x_0)f''(x_0) > 0$, basta con que se tome $x_0 > r$. A continuación se demuestra que, si es así, entonces $x_n > r$ para todo n . Se utilizará el principio de inducción completa. La propiedad se cumple para $n = 0$. Partiendo de que se cumpla para algún n entonces, utilizando el polinomio de Taylor de primer grado alrededor de x_n con resto de Lagrange, se puede escribir:

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(c)}{2!}(x - x_n)^2$$

donde c se encuentra entre x y x_n . En particular, para $x = r$, se tiene que:

$$f(r) = 0 = f(x_n) + f'(x_n)(r - x_n) + \frac{f''(c)}{2!}(r - x_n)^2 \quad (10)$$

Como el sumando $\frac{f''(c)}{2!}(r - x_n)^2$ es positivo y la suma es cero,

$$f(x_n) + f'(x_n)(r - x_n) < 0$$

despejando r (recuérdese que se ha supuesto la derivada positiva) se tiene:

$$r < x_n - \frac{f(x_n)}{f'(x_n)}$$

Pero el miembro de la derecha de esta desigualdad es x_{n+1} , así que se ha llegado a

$$x_{n+1} > r$$

Es decir, que si $x_n > r$ entonces también se cumple $x_{n+1} > r$ y como $x_0 > r$ se tiene que, para todos los valores de n , es $x_n > r$. Esto significa que la sucesión de las x_n está acotada inferiormente.

Ahora se demostrará que es una sucesión decreciente. En efecto, como

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \quad (11)$$

y $x_{n-1} > r$, entonces $f(x_{n-1}) > 0$; por otra parte $f'(x_{n-1}) > 0$ porque se ha supuesto $f'(x) > 0$ en todo el intervalo $[a, b]$. Por estas razones se tiene:

$$\frac{f(x_{n-1})}{f'(x_{n-1})} > 0$$

y ello significa que $x_n < x_{n-1}$ para $n = 1, 2, 3, \dots$ Es decir, la sucesión es decreciente.

Como se trata de una sucesión decreciente y acotada inferiormente es obligatoriamente convergente. Sea entonces:

$$\bar{x} = \lim_{n \rightarrow \infty} x_n$$

Ahora hay que probar que este valor límite es la raíz buscada. En efecto, aplicando límite en cada término de la igualdad (11) y teniendo en cuenta la continuidad de las funciones que en ella aparecen, se obtiene:

$$\lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} x_{n-1} - \frac{\lim_{n \rightarrow \infty} f(x_{n-1})}{\lim_{n \rightarrow \infty} f'(x_{n-1})} = \lim_{n \rightarrow \infty} x_{n-1} - \frac{f(\lim_{n \rightarrow \infty} x_{n-1})}{f'(\lim_{n \rightarrow \infty} x_{n-1})}$$

Esto es:

$$\bar{x} = \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})}$$

De donde $f(\bar{x}) = 0$, lo cual significa que el valor límite es la raíz buscada. Esto prueba el teorema. ■

El error en el Método de Newton – Raphson

Al tratar del error, serán abordados dos aspectos: con qué rapidez tiende este error hacia cero y como se puede hallar en cada paso una cota del error absoluto que permita detener el proceso iterativo en el momento adecuado. En ambos se requiere plantear el polinomio de Taylor de primer grado con resto de Lagrange alrededor de x_{n-1} :

$$f(x) = f(x_{n-1}) + f'(x_{n-1})(x - x_{n-1}) + \frac{f''(c)}{2!}(x - x_{n-1})^2 \quad (12)$$

donde c es un número entre x y x_{n-1}

Si en esta expresión hacemos $x = r$, como $f(r) = 0$, se obtiene:

$$0 = f(x_{n-1}) + f'(x_{n-1})(r - x_{n-1}) + \frac{f''(c)}{2!}(r - x_{n-1})^2$$

Dividiendo en cada término por $f'(x_{n-1})$ se llega a:

$$0 = \frac{f(x_{n-1})}{f'(x_{n-1})} + (r - x_{n-1}) + \frac{1}{2} \frac{f''(c)}{f'(x_{n-1})}(r - x_{n-1})^2$$

Agrupando de otra forma: $r - \left[x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \right] = -\frac{1}{2} \frac{f''(c)}{f'(x_{n-1})}(r - x_{n-1})^2$

Nótese que la expresión entre corchetes es x_n , por tanto resulta:

$$r - x_n = -\frac{1}{2} \frac{f''(c)}{f'(x_{n-1})}(r - x_{n-1})^2$$

Tomando módulos en ambos miembros esta igualdad se puede expresar en términos de errores absolutos:

$$E(x_n) = \frac{1}{2} \left| \frac{f''(c)}{f'(x_{n-1})} \right| [E(x_{n-1})]^2$$

Una expresión mucho más interesante se obtiene si se llama:

M : una cota superior de $|f''(x)|$ para x en $[a, b]$

d : una cota inferior de $|f'(x)|$ para x en $[a, b]$

Resulta: $E(x_n) \leq \frac{M}{2d} [E(x_{n-1})]^2$

O, en términos de errores absolutos máximos:

$$E_m(x_n) = \frac{M}{2d} [E_m(x_{n-1})]^2$$

Si se compara esta expresión con las similares obtenidas para los métodos de bisección y Regula Falsi se verá que en aquellos el error absoluto máximo de una iteración era igual a una parte (menor que 1) del error absoluto máximo de la aproximación anterior; es decir, eran expresiones del tipo:

$$E_m(x_n) = k \cdot E_m(x_{n-1})$$

Para el método de Newton – Raphson, el error en x_{n-1} aparece al cuadrado y esto posee una gran importancia. Por ejemplo, esto implica que si en la iteración $n - 1$, el error es del orden de 0,001, entonces en la iteración siguiente será del orden de $(0,001)^2 = 0,000001$, lo cual significa que en este método, una vez que el error ha alcanzado valores pequeños, con cada nueva iteración se duplica el número de las cifras decimales exactas obtenidas en la iteración anterior. Como se ve, se trata de una velocidad de convergencia muy alta. Este hecho se expresa en términos más técnicos diciendo que los métodos de bisección y Regula Falsi presentan convergencia lineal mientras el de Newton – Raphson posee convergencia cuadrática.

Para hallar una forma sencilla de acotar el error absoluto del método, hágase $x = x_n$ en la ecuación (12):

$$f(x_n) = f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) + \frac{f''(c)}{2!}(x_n - x_{n-1})^2 \quad (13)$$

donde c es un número entre x_n y x_{n-1} .

Sin embargo, recuérdese que en este método se basa en que

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$$

Trasponiendo algunos términos esto se puede escribir como:

$$f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) = 0$$

Como $f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1})$ son los dos primeros términos a la derecha de la ecuación (13), ella se transforma en:

$$f(x_n) = \frac{f''(c)}{2!}(x_n - x_{n-1})^2$$

Esta ecuación se puede escribir, teniendo en cuenta que $f(r) = 0$, como:

$$f(x_n) - f(r) = \frac{f''(c)}{2!}(x_n - x_{n-1})^2$$

Aplicando el teorema del valor medio al primer miembro de esta ecuación se tiene:

$$f'(r)(x_n - r) = \frac{f''(c)}{2!}(x_n - x_{n-1})^2$$

Esto es:

$$r - x_n = -\frac{f''(c)}{2f'(\alpha)}(x_n - x_{n-1})^2$$

Tomando módulos y sustituyendo las derivadas por sus cotas M y d :

$$|r - x_n| \leq \left| \frac{M}{2d} \right| (x_n - x_{n-1})^2 \quad (14)$$

Para no tener que hallar las cotas M y d , se puede razonar así. Sea δ un número positivo muy pequeño y supóngase que se ha llegado a una iteración para la cual:

$$|x_n - x_{n-1}| < \delta$$

Entonces la ecuación (14) implica que:

$$|r - x_n| \leq \left| \frac{M\delta}{2d} \right| |x_n - x_{n-1}| \quad (15)$$

Como el número δ es muy pequeño, la expresión $\left| \frac{M\delta}{2d} \right|$ es mucho menor que 1, así que la ecuación (15) implica que: $|r - x_n| \leq |x_n - x_{n-1}|$

Como el primer miembro de esta desigualdad es el error absoluto, resulta que el miembro de la derecha puede tomarse como error absoluto máximo de x_n :

$$E_m(x_n) = |x_n - x_{n-1}| \quad (16)$$

Nótese que (16) puede no ser cierto en los primeros pasos del algoritmo pero sí lo será una vez que la diferencia entre dos iteraciones sucesivas se haya hecho pequeña.

En el método de Newton – Raphson se toma la siguiente condición de terminación:

Condición de terminación:

Si se desea obtener la raíz de la ecuación con un error absoluto menor que ε el método de Newton – Raphson se llevará a cabo hasta la aproximación x_n para la cual

$$E_m(x_n) = |x_n - x_{n-1}| \leq \varepsilon$$

Algoritmo en seudo código

Se supone que la ecuación a resolver es $f(x) = 0$, que la raíz que se quiere hallar está separada dentro de un intervalo $[a, b]$ en el cual $f(x)$ y sus dos primeras derivadas son continuas y que $f'(x)$ y $f''(x)$ no se anulan en $[a, b]$. Se supone que x_0 se ha seleccionado dentro del intervalo $[a, b]$ de modo que se cumple $f(x_0)f''(x_0) > 0$ (es decir, el signo de la función coincidiendo con el sentido de la concavidad). Se suponen conocidas la funciones $f(x)$ y $f'(x)$, la aproximación inicial x_0 y la tolerancia ε que se permitirá.

```
 $x_{anterior} := x_0$ 
repeat
     $x := x_{anterior} - \frac{f(x_{anterior})}{f'(x_{anterior})}$ 
     $Error := |x - x_{anterior}|$ 
     $x_{anterior} := x$ 
until  $Error < \varepsilon$ 
La raíz buscada es  $x$  y su error absoluto máximo es  $Error$ 
Terminar
```

Comentarios finales

El método de Newton – Raphson es, generalmente, un método de convergencia rápida aunque esta rapidez depende de la función $f(x)$ y de la aproximación inicial que se elija; usualmente con cuatro o cinco iteraciones se obtiene la raíz con más de cuatro cifras decimales exactas. Esta característica hace aconsejable el empleo de este algoritmo en el trabajo a mano o cuando las limitaciones de tiempo obliguen a utilizar un método muy eficiente para el cálculo de raíces. Su mayor inconveniente es la necesidad de hallar la primera derivada de $f(x)$, lo cual puede ser muy engoroso y hay que hacerlo casi siempre fuera de la máquina.

Aunque las condiciones de convergencia son más exigentes que en otros métodos, las mismas se satisfacen si el intervalo $[a, b]$ se toma suficientemente pequeño; además se puede verificar fácilmente su cumplimiento con solo mirar en la pantalla la gráfica de $f(x)$, teniendo en cuenta que el crecimiento, positivo o negativo, de la función indica el signo de la primera derivada en tanto que el sentido de la concavidad de la curva indica el signo de la segunda derivada.

Cuando el intervalo $[a, b]$ donde está separada la raíz es pequeño, la selección del punto x_0 de partida no es muy importante, pero cuando el intervalo es mayor, debe tenerse cuidado de seleccionar x_0 de manera que $f(x_0)$ coincida con el signo de $f''(x_0)$. Observando la gráfica de f es muy fácil realizar esta selección.

Como se ha visto, el método de Newton – Raphson pose un algoritmo muy simple para su programación y, como todos los métodos numéricos iterativos es prácticamente inmune a los errores de redondeo que ocurran a lo largo del proceso.

Ejemplo 3

Utilice el método de Newton – Raphson para determinar con cinco cifras decimales exactas la menor raíz positiva de la ecuación: $2 \operatorname{sen} \pi x + x = 0$

Solución:

Primero se requiere aislar la raíz en un intervalo en que se cumplan las condiciones suficientes para la convergencia del método. Como la función $2 \operatorname{sen} \pi x$ solo toma valores entre -2 y 2, la función:

$$f(x) = 2 \operatorname{sen} \pi x + x$$

no se puede anular para $x > 2$. Basta entonces graficar $f(x)$ en el intervalo $[0, 2]$. En la figura 5 se muestra la gráfica. En ella se observa que la menor raíz positiva se encuentra en el intervalo $[1; 1,25]$ en el cual se cumple además que la primera derivada es negativa (función decreciente) y la segunda derivada es positiva (concavidad hacia arriba). Para seleccionar x_0 , como la segunda derivada es positiva en el intervalo (un análisis más detallado muestra que en $x = 1$ la segunda derivada se anula), se selecciona un punto en el que la función también sea positiva. En este caso se ha seleccionado $x_0 = 1,1$ aunque cualquier punto del intervalo a la izquierda de la raíz serviría igual. Utilizando un programa realizado a partir del algoritmo de Newton – Raphson y tomando como tolerancia para el error $\epsilon = 0,000005$ se obtienen los resultados que muestra la tabla 1. La raíz buscada es, por tanto, 1,206035 con cinco cifras decimales exactas en solo cuatro iteraciones.

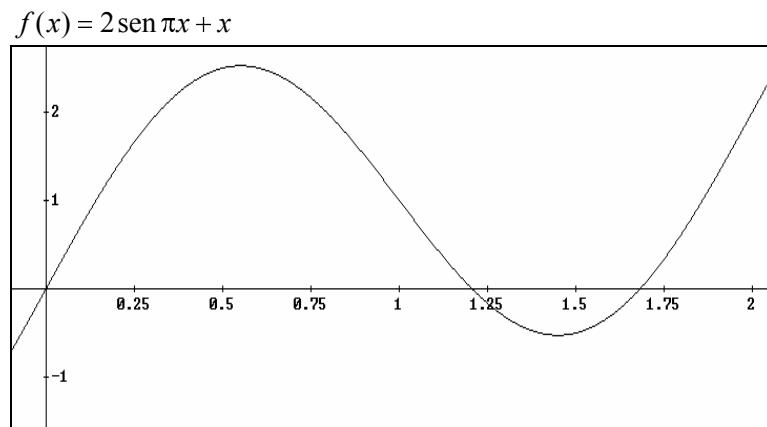


Figura 5

Iteración	x	$E_m(x)$
0	1,1	
1	1,19686466	0,09686466
2	1,20591673	0,00905207
3	1,20603510	0,00011837
4	1,20603512	0,00000002

Tabla 1

Ejemplo 4

Se sabe que en el intervalo $[1, 2]$ la ecuación $x^3 + 3x^2 + x - 6 = 0$ tiene una raíz.

- a) Determine cuantas iteraciones se requerirían para hallar dicha raíz por el método de bisección.
- b) Halle la raíz por el método Regula Falsi.
- c) Halle la raíz por el método de Newton – Raphson.
- d) Compare la rapidez de la convergencia y la eficiencia relativa de los tres métodos en este caso.

Solución:

- a) La cantidad de iteraciones necesarias en el método de bisección es independiente de $f(x)$ y se calcula a partir del intervalo de separación inicial y la tolerancia deseada como:

$$n \geq \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln 2}$$

Como en este caso es $a = 1$, $b = 2$ y $\varepsilon = 0,000005$, se obtiene:

$$n \geq \frac{\ln\left(\frac{2-1}{0,000005}\right)}{\ln 2} = 17,6$$

Es decir, se requiere como mínimo, de 18 iteraciones.

- b) Antes de proceder al cálculo, es conveniente verificar si se satisfacen las hipótesis de continuidad y de poca curvatura en la función $f(x)$ para el intervalo $[1, 2]$. En la figura 6 se muestra la gráfica de la función $f(x) = x^3 + 3x^2 + x - 6$ en dicho intervalo. Se aprecia que la función es continua y que la derivada no se anula y sufre poco cambio en el intervalo. Esto garantiza el buen funcionamiento del método Regula Falsi y la validez de la fórmula para acotar el error absoluto. En la tabla 2 se muestran los resultados numéricos. La raíz buscada es, con cinco cifras decimales exactas: 1,094550

$$f(x) = x^3 + 3x^2 + x - 6$$

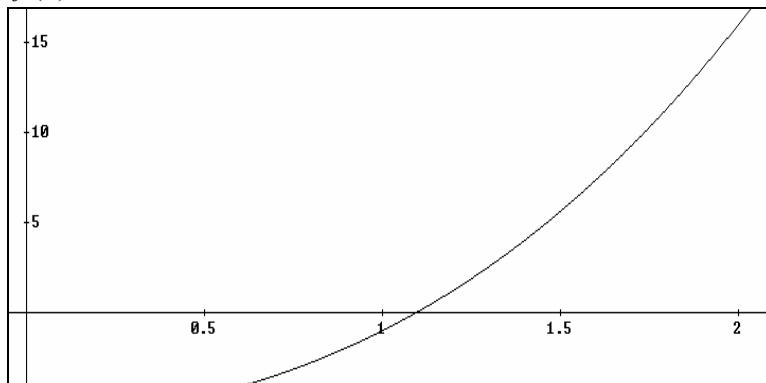


Figura 6

Iteración	a	b	x	$E_m(x)$
1	1	2	1,058824	
2	1,058824	2	1,081264	0,022440
3	1,081264	2	1,089639	0,008376
4	1,089639	2	1,092740	0,003100
5	1,092740	2	1,093884	0,001144
6	1,093884	2	1,094305	0,000422
7	1,094305	2	1,094461	0,000155
8	1,094461	2	1,094518	0,000057
9	1,094518	2	1,094539	0,000021
10	1,094539	2	1,094547	0,000008
11	1,094547	2	1,094550	0,000003

Tabla 2

- c) Para la aplicación del método de Newton – Raphson se verifican las hipótesis adicionales acerca de las derivadas. En el intervalo [1, 2] la primera derivada y la segunda son positivas. La aproximación inicial debe tomarse en algún punto a la derecha de la raíz, donde $f(x)$ es positiva como la segunda derivada. En este caso se seleccionó $x_0 = 1,5$. Los resultados numéricos se muestran en la tabla 3. La raíz deseada, calculada con 5 cifras decimales exactas es 1.094551.

Iteración	x	$E_m(x)$
0	1,5	
1	1,16417910	0,33582090
2	1,09713536	0,06704375
3	1,09455523	0,00258012
4	1,09455148	0,00000375

Tabla 3

- d) Como era de esperar, la cantidad de iteraciones requeridas para el cálculo fue mayor en el método de bisección (18), menor en Regula Falsi (11) y muy inferior en Newton – Raphson (4). En cuanto a la eficiencia, hay que tener en cuenta que, mientras en el método de bisección y en Regula Falsi en cada iteración se requiere evaluar solamente una función, en el método de Newton – Raphson se requiere evaluar dos funciones: $f(x)$ y su derivada. Así, resulta que este método ha requerido 8 evaluaciones en comparación con 11 de Regula Falsi. Como se ve, la eficiencia en este caso es mayor pero no tan desproporcionada.

Ejercicios

En todos los ejercicios que siguen se debe utilizar, siempre que se necesite, un programa graficador tanto para separar las raíces como para verificar las hipótesis del método utilizado. También se supone que el algoritmo Newton – Raphson se utilice mediante un programa computacional, preferiblemente confeccionado por usted. Si no cuenta con un programa adecuado, realice los cálculos a mano y obtenga las raíces con solo dos o tres cifras decimales exactas.

1. En la figura 7 se muestra la gráfica de la función $f(x)$. Proponga valores adecuados de x_0 para calcular las cuatro raíces de la ecuación $f(x) = 0$ mediante el método de Newton – Raphson.

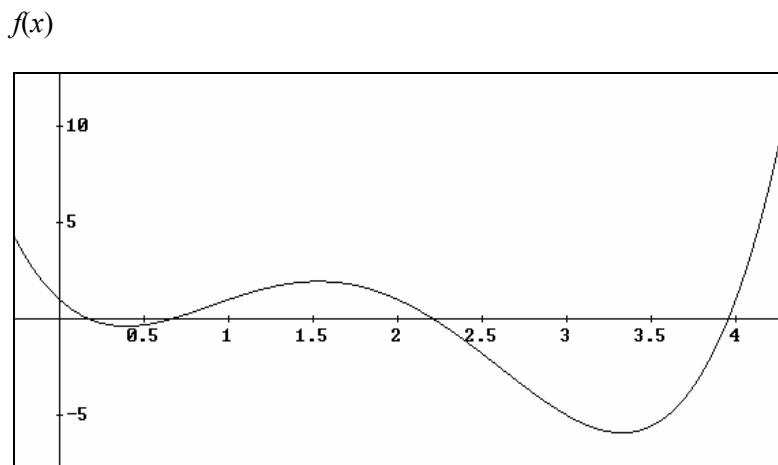


Figura 7

2. Calcule, con cinco cifras decimales exactas, las raíces reales de las siguientes ecuaciones algebraicas mediante el método de Newton – Raphson. Posiblemente, ya usted separó las raíces de estas ecuaciones en los ejercicios de la sección 2.2 y las resolvió por alguno de los métodos anteriores. Compare la cantidad de iteraciones que necesitó con cada uno de los métodos que haya utilizado.
 - $x^4 + x^3 - x^2 + x - 2 = 0$
 - $x^4 - 11x^3 + 41x^2 - 60x + 30 = 0$
 - $x^4 - 3x^3 + 10x^2 - 13x + 5 = 0$
 - $x^4 + 3x^2 + 2 = 0$
 - $x^3 + 7x^2 + 14x + 9 = 0$
3. Utilice el método Newton – Raphson para calcular con cinco cifras decimales exactas, las raíces de las siguientes ecuaciones trascendentes. Posiblemente, ya usted separó las raíces de estas ecuaciones en los ejercicios de la sección 2.2 y las resolvió usando los algoritmos anteriores. En ese caso, compare la cantidad de iteraciones que necesito con cada método usado.
 - $\sin x - \log x = 0$
 - $5e^x - 2x - 10 = 0$
 - $(x^2 + 1)\cos x = 1; -10 \leq x \leq 10$

- d) $2 \tanh x - \sin x - 0,3 = 0$
e) $x^4 - 4x^2 - 4x - 16 - \ln|x| = 0$
f) $e^x \sin x - 2e^x + 3 = 0$
g) $x = \tan^2 x; \quad 0 \leq x \leq 2\pi$
h) $\sqrt{x} = 2 \ln x$
4. Al comienzo de este capítulo, en el ejemplo 2 de la sección 2.1, se planteó el problema de hallar las dimensiones de un recipiente cilíndrico de 1000 cm^3 de capacidad que hiciera mínima la cantidad de material a utilizar. Se llegó a la ecuación:

$$4\pi r^4 + \pi r^3 - 2000r - \frac{500}{\pi} = 0$$

Calcule el valor de r con cinco cifras decimales exactas mediante Newton – Raphson.

5. En la figura 8, el área sombreada es 0,6 veces el área del rectángulo $ABCD$. Halle el área de dicho rectángulo. Utilice el método de Newton – Raphson y determine las raíces con cinco cifras decimales exactas.

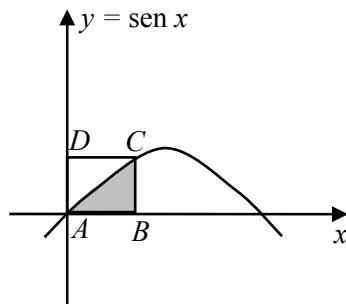


Figura 8

6. La recta L de la figura 9 es tangente a la sinusoida en dos puntos. Determine la ecuación de la recta. Utilice el método de Newton – Raphson y realice los cálculos con cuatro cifras decimales exactas.

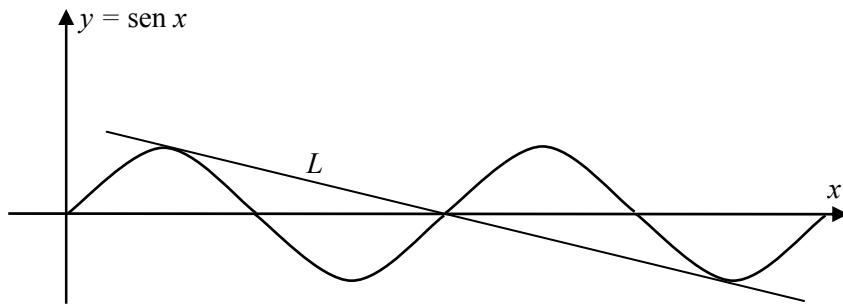


Figura 9

7. Una pieza cilíndrica de queso tiene 30 cm de diámetro. Se quiere separar la quinta parte mediante un corte, como muestra la figura 10. Determine, con error menor que 0,5 mm a qué distancia del centro debe darse el corte. Utilice el método de Newton – Raphson.

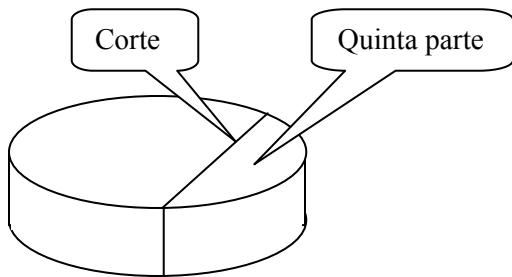


Figura 10

8. La circunferencia de la figura 11 tiene su centro en el punto $(0, 1)$ y es tangente a la cosinusoide. Determine su radio con cinco cifras decimales exactas. Utilice el método de Newton – Raphson.

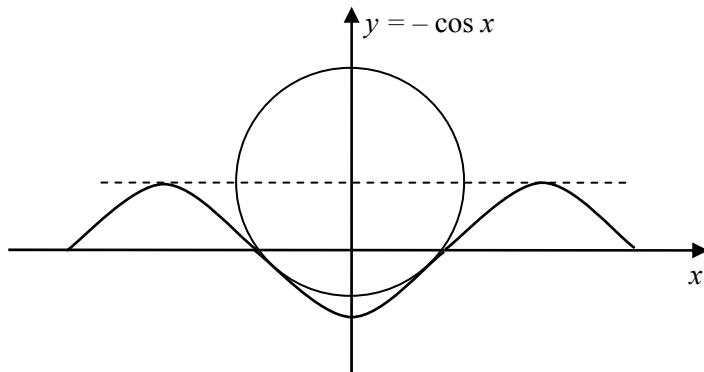


Figura 11

9. Al comienzo de esta sección se mostró un algoritmo iterativo que converge muy rápidamente hacia la raíz cuadrada de 2. Se basa en la fórmula recursiva:

$$x_0 = 2; \quad x_n = \frac{x_{n-1} + \frac{2}{x_{n-1}}}{2} \quad n = 1, 2, 3, \dots$$

Demuestre que esta fórmula se obtiene aplicando el método de Newton – Raphson a la ecuación $x^2 = 2$.

10. Generalice la idea del problema anterior para obtener un algoritmo recursivo que permita calcular la raíz cuadrada positiva de cualquier número positivo N .
11. Suponga que usted necesita un algoritmo que permita hallar el recíproco de un número positivo N sin efectuar divisiones. Plantee el problema en términos de resolver la ecuación

$$N - \frac{1}{x} = 0$$

y aplique el método de Newton – Raphson para obtener un algoritmo iterativo adecuado. Recuerde que la selección de x_0 es parte del algoritmo. Ensaye el algoritmo, hallando el recíproco de 3.

12. El algoritmo en seudo código que aparece a continuación es una simplificación del método de Newton – Raphson que suele llamarse Newton Modificado. Interprétilo gráficamente y analice las ventajas y desventajas que presenta.

```

 $x_{anterior} := x_0$ 
 $P := f'(x_0)$ 
repeat
     $x := x_{anterior} - \frac{f(x_{anterior})}{P}$ 
     $Error := |x - x_{anterior}|$ 
     $x_{anterior} := x$ 
until  $Error < \epsilon$ 
La raíz buscada es  $x$  y su error absoluto máximo es  $Error$ 
Terminar

```

2.6 El método de las secantes

La principal desventaja del método de Newton – Raphson es la necesidad de trabajar con la función derivada de $f(x)$. En muchos casos prácticos esto es un gran inconveniente. Por ejemplo, recuerdes la ecuación (4) de la sección 2.1 que surgió al tratar de determinar los parámetros de la función logística

$$p(t) = \frac{P_L}{1 - ce^{-kt}}$$

a partir de los valores p_1 , p_2 y p_3 correspondientes a tres instantes t_1 , t_2 y t_3 . Para calcular el parámetro k se necesitaba resolver la ecuación:

$$(p_2 - p_1)(p_3 e^{-kt_3} - p_2 e^{-kt_2}) - (p_3 - p_2)(p_2 e^{-kt_2} - p_1 e^{-kt_1}) = 0$$

En este caso resulta obvio lo problemático de utilizar dicho método.

El método de las secantes es una modificación del método de Newton – Raphson dirigida a eliminar la necesidad de utilizar la función derivada. Para ello, se sustituye la pendiente de la recta tangente por la pendiente de una recta secante a la gráfica de $f(x)$. El método de las secantes requiere de dos aproximaciones iniciales de la raíz r ya que una secante se determina por dos puntos de la curva.

En la figura 1 se muestra gráficamente cómo funciona el algoritmo. La recta L_1 , secante a la curva $y = f(x)$ por los puntos de abscisas x_0 y x_1 corta al eje x en un punto cuya abscisa se toma como x_2 ; con x_1 y x_2 se determina una nueva secante L_2 la cual interseca al eje x en x_3 , etcétera; este proceso da lugar a la sucesión $\{x_0, x_1, x_2, x_3, \dots\}$ la cual, bajo condiciones apropiadas, converge hacia la raíz r .

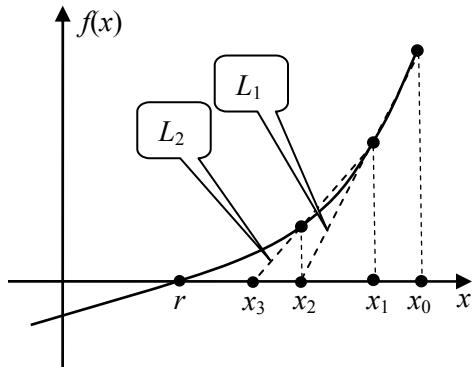


Figura 1

Como la pendiente de la recta que pasa por los puntos $(x_{n-2}, f(x_{n-2}))$ y $(x_{n-1}, f(x_{n-1}))$ es

$$\frac{f(x_{n-1}) - f(x_{n-2})}{x_{n-1} - x_{n-2}} \quad (1)$$

la ecuación que permite determinar x_n se puede obtener si en la fórmula de Newton – Raphson

$$x = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$$

se sustituye la pendiente de la recta tangente $f'(x_{n-1})$ por la expresión (1). De ahí:

$$x_n = x_{n-1} - \frac{x_{n-1} - x_{n-2}}{f(x_{n-1}) - f(x_{n-2})} f(x_{n-1}) \quad \text{para } n = 2, 3, 4, \dots \quad (2)$$

Las aproximaciones x_0 y x_1 no hay que tomarlas obligatoriamente a un mismo lado de r ni en un orden específico y no siempre la sucesión x_n converge monótonamente a la raíz r ; en muchas ocasiones se obtienen unas aproximaciones por defecto y otras por exceso. Obsérvese la forma en que se produce la convergencia en el caso de la figura 2.

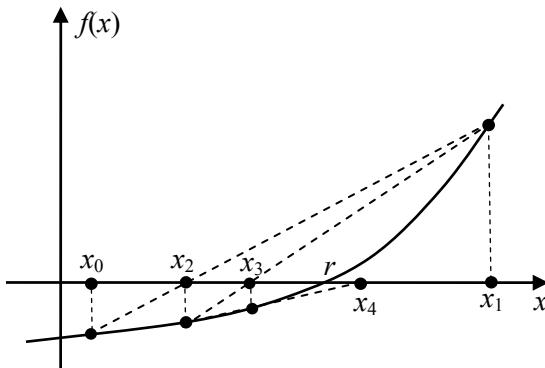


Figura 2

Tal como en el método de Newton – Raphson del cual es una modificación, en el método de las secantes pueden suceder comportamientos indeseables del proceso de convergencia si la primera

o la segunda derivadas de $f(x)$ se anulan en las cercanías de la raíz r o si las aproximaciones iniciales no se seleccionan con cuidado.

Convergencia del método de las secantes

Al analizar el error del método de Newton – Raphson se llegó a la relación:

$$r - x_n = -\frac{1}{2} \frac{f''(c)}{f'(x_{n-1})} (r - x_{n-1})^2$$

de la cual se obtuvo posteriormente importantes consecuencias. Mediante algunas manipulaciones algebraicas en cuyos detalles no se entrará, en el método de las secantes se puede llegar a una relación análoga:

$$r - x_n = -\frac{1}{2} \frac{f''(\alpha_n)}{f'(\beta_n)} (r - x_{n-1})(r - x_{n-2}) \quad (3)$$

Esta fórmula relaciona los errores en tres iteraciones consecutivas obtenidas mediante la fórmula de la secante. A partir de ella, se puede analizar cómo es la convergencia en este método.

Usando la misma notación que en Newton – Raphson, sea

M : una cota superior de $|f''(x)|$ para x en un entorno I de la raíz r .

d : una cota inferior de $|f'(x)|$ para x en I .

Tomando módulos en ambos miembros de (3) se puede escribir dicha ecuación en términos de errores absolutos:

$$E(x_n) \leq \frac{M}{2d} E(x_{n-1}) E(x_{n-2}) \quad (4)$$

Para simplificar las notaciones, se llamará: $\lambda = \frac{M}{2d}$ de modo que la ecuación (4) se puede escribir:

$$E(x_n) \leq \lambda E(x_{n-1}) E(x_{n-2})$$

o, multiplicando por λ : $\lambda E(x_n) \leq \lambda E(x_{n-1}) \lambda E(x_{n-2})$ (5)

Si ahora se supone que las iteraciones iniciales x_0 y x_1 se seleccionan lo suficientemente próximas a r de manera que:

$$\lambda E(x_0) \leq \delta < 1 \quad \text{y} \quad \lambda E(x_1) \leq \delta < 1$$

y se aplica la desigualdad (5), se obtiene para las siguientes iteraciones

$$\begin{aligned} \lambda E(x_2) &\leq \lambda E(x_1) \lambda E(x_0) \leq \delta \cdot \delta = \delta^2 \\ \lambda E(x_3) &\leq \lambda E(x_2) \lambda E(x_1) \leq \delta^2 \cdot \delta = \delta^3 \end{aligned}$$

$$\begin{aligned}\lambda E(x_4) &\leq \lambda E(x_3)\lambda E(x_2) \leq \delta^3 \cdot \delta^2 = \delta^5 \\ \lambda E(x_5) &\leq \lambda E(x_4)\lambda E(x_3) \leq \delta^5 \cdot \delta^3 = \delta^8 \\ &\text{etcétera.}\end{aligned}$$

De manera que los errores absolutos van siendo cada vez menores y convergen rápidamente hacia cero. Nótese que el exponente a que aparece elevada δ viene dado por los términos de la sucesión de Fibonacci, los cuales tienden a infinito con rapidez exponencial.

El resultado anterior se puede establecer como un teorema:

Teorema 1

Sea $f(x) = 0$ una ecuación con una sola raíz $x = r$ en un intervalo $[a, b]$. Sean $f(x)$, $f'(x)$ y $f''(x)$ continuas en $[a, b]$ y no nulas en dicho intervalo. Si M representa una cota superior de $|f''(x)|$ para x en $[a, b]$ y d una cota inferior de $|f'(x)|$ para x en $[a, b]$. Entonces, tomando aproximaciones iniciales x_0 y x_1 en $[a, b]$ que satisfagan la condición: $\lambda E(x_0) \leq \delta < 1$ y $\lambda E(x_1) \leq \delta < 1$ con $\lambda = M/2d$, la sucesión de valores obtenida con la aplicación de la fórmula de las secantes converge hacia r .

■

En la práctica, es muy difícil contar con el parámetro λ , pero lo fundamental es que el teorema garantiza que, si se toman x_0 y x_1 suficientemente próximos a r , el algoritmo convergerá hacia r .

El error en el método de las secantes

El método de Newton – Raphson posee una notable velocidad de convergencia debido a que es un método de convergencia cuadrática, es decir, el error absoluto máximo se comporta según la ecuación:

$$E_m(x_n) = \frac{M}{2d} [E_m(x_{n-1})]^2$$

cabe entonces preguntarse si el método de las secantes poseerá una convergencia del mismo orden.

Supóngase que en el método de las secantes el error se comporta según la ecuación:

$$E_m(x_n) = k [E_m(x_{n-1})]^p \quad (5)$$

donde k y p son parámetros por determinar. La ecuación (4)

$$E(x_n) \leq \frac{M}{2d} E(x_{n-1}) E(x_{n-2})$$

se puede escribir fácilmente en términos de errores absolutos máximos:

$$E_m(x_n) = \frac{M}{2d} E_m(x_{n-1}) E_m(x_{n-2}) \quad (6)$$

Según la ecuación (5): $E_m(x_{n-1}) = k [E_m(x_{n-2})]^p$

$$y \quad E_m(x_n) = k[E_m(x_{n-1})]^p = k[k[E_m(x_{n-2})]^p]^p = k^{p+1}[E_m(x_{n-2})]^{p^2}$$

Sustituyendo estas expresiones en la ecuación (6), resulta:

$$k^{p+1}[E_m(x_{n-2})]^{p^2} = \frac{M}{2d} k[E_m(x_{n-2})]^p E_m(x_{n-2})$$

es decir: $k^p [E_m(x_{n-2})]^{p^2} = \frac{M}{2d} [E_m(x_{n-2})]^{p+1}$

De aquí se obtienen las ecuaciones: $k^p = \frac{M}{2d}$ y $p^2 = p + 1$

La segunda de ellas posee dos raíces, una negativa, que carece de sentido, y una positiva:

$$p = \frac{1 + \sqrt{5}}{2} = 1,618$$

en cuanto al valor de k : $k = \left(\frac{M}{2d}\right)^{\frac{1}{p}} = \left(\frac{M}{2d}\right)^{0,618}$

Sustituyendo en la ecuación (5) los valores hallados para p y k :

$$E_m(x_n) = \left(\frac{M}{2d}\right)^{0,618} [E_m(x_{n-1})]^{1,618}$$

Puede concluirse entonces que el método de las secantes posee una convergencia de orden 1,618 (un poco menor que 2), de modo que no se obtendrá una velocidad de convergencia tan alta como en el método de Newton – Raphson.

En cuanto a la estimación del error en el método, dado que su convergencia es comparable con la Newton – Raphson, se utiliza el mismo criterio que en ese método:

$$E_m(x_n) = |x_n - x_{n-1}|$$

y por tanto, la misma condición de terminación:

Condición de terminación:

Si se desea obtener la raíz de la ecuación con un error absoluto menor que ε el método de las secantes se llevará a cabo hasta la aproximación x_n para la cual

$$E_m(x_n) = |x_n - x_{n-1}| \leq \varepsilon$$

Algoritmo en seudo código

Se supone que la ecuación a resolver es $f(x) = 0$, que la raíz que se quiere hallar está separada dentro de un intervalo $[a, b]$ en el cual $f(x)$ y sus dos primeras derivadas son continuas y que $f'(x)$ y $f''(x)$ no se anulan en $[a, b]$. Se supone que x_0 y x_1 se han seleccionado dentro del intervalo $[a, b]$ lo suficientemente próximos a r para que el algoritmo converja. Se suponen conocidas la función $f(x)$, las aproximaciones iniciales x_0 y x_1 y la tolerancia ε que se permitirá.

```
xa := x0
ya := f(xa)
xb := x1
yb := f(xb)
repeat
    xc := xb -  $\frac{x_b - x_a}{y_b - y_a} y_b$ 
    yc := f(xc)
    Error := |xc - xb|
    xa := xb
    ya := yb
    xb := xc
    yb := yc
until Error < ε
La raíz buscada es xc y su error absoluto máximo es Error
Terminar
```

Comentarios finales

El método de las secantes posee varias características muy positivas, principalmente su rapidez de convergencia. Si esta se mide en cantidad de iteraciones necesarias para obtener la raíz con cierta exactitud, entonces esta velocidad es ligeramente menor que la de Newton – Raphson; sin embargo, como este método requiere evaluar dos funciones en cada paso mientras que en el algoritmo de las secantes solo se evalúa una función, cuando la rapidez se mide en tiempo necesario para alcanzar la solución, el método de las secantes es casi siempre más rápido.

Otro aspecto favorable es el hecho de que no se requiere conocer la primera ni la segunda derivadas de la da función; sólo se necesita que estas sean continuas y no se anulen, lo cual puede verificarse con la observación de la gráfica de $f(x)$ obtenida en la pantalla.

Un inconveniente del método es la posibilidad de no convergencia a la raíz si las aproximaciones iniciales no están suficientemente cerca de ella; sin embargo, cuando el cociente $f''(x)/f'(x)$ toma valores pequeños en las proximidades de la raíz (esto es, la gráfica de $f(x)$ tiene pendiente pronunciada y curvatura pequeña), entonces la convergencia es casi segura aun para valores de x_0 y x_1 no tan próximos a r .

Como todos los métodos de puntos (no de intervalos) el método de las secantes no produce intervalos donde esté encerrada la raíz buscada y el acotamiento del error no es completamente seguro; téngase en cuenta, por tanto, que el criterio de terminación que se ha dado es valido solamente si las aproximaciones obtenidas están en un entorno de la raíz pequeño, en que la derivada de $f(x)$ no sufre grandes cambios. Siempre que sea posible, es aconsejable seleccionar las aproximaciones iniciales mediante una observación de la gráfica de $f(x)$ y teniendo en cuenta el fundamento geométrico del método.

En el método de las secantes aparece un cociente donde hay una diferencia de valores de x entre una diferencia de valores de la función; cuando los valores de la función son ya muy similares y con el mismo signo (lo cual no sucede en la fórmula de Regula Falsi), puede presentarse pérdida de significación. Por este motivo, no se debe buscar una exactitud que exceda la precisión del sistema numérico de la computadora que se esté empleando.

Ejemplo 1

Determine una función logística $p(t) = \frac{P_L}{1 - ce^{-kt}}$ que satisfaga las condiciones:

$$\begin{aligned} p(2001) &= 1,2 \text{ millones} \\ p(2003) &= 2,4 \text{ millones} \\ p(2004) &= 2,8 \text{ millones} \end{aligned}$$

Solución:

Conviene primero seleccionar adecuadamente las unidades de tiempo y de población. Sea la variable t definida por:

$$t = \text{año} - 2000$$

Si la variable p se mide en millones de habitantes, entonces:

$$\begin{aligned} p(1) &= 1,2 \\ p(3) &= 2,4 \\ p(4) &= 2,8 \end{aligned}$$

El valor de k será entonces la raíz de la ecuación:

$$(p_2 - p_1)(p_3 e^{-kt_3} - p_2 e^{-kt_2}) - (p_3 - p_2)(p_2 e^{-kt_2} - p_1 e^{-kt_1}) = 0$$

Haciendo	$t_1 = 1$	$p_1 = 1,2$
	$t_1 = 3$	$p_1 = 2,4$
	$t_1 = 4$	$p_1 = 2,8$

se obtiene; $3,36 e^{-3k} - 3,84 e^{-2k} + 0,48 = 0$ (7)

Como el parámetro k es positivo por definición, al separar las raíces de esta ecuación bastará considerar $k > 0$. La gráfica de la función $f(k) = 3,36 e^{-3k} - 3,84 e^{-2k} + 0,48$ se muestra en la figura 3 en el intervalo $[0, 4]$.

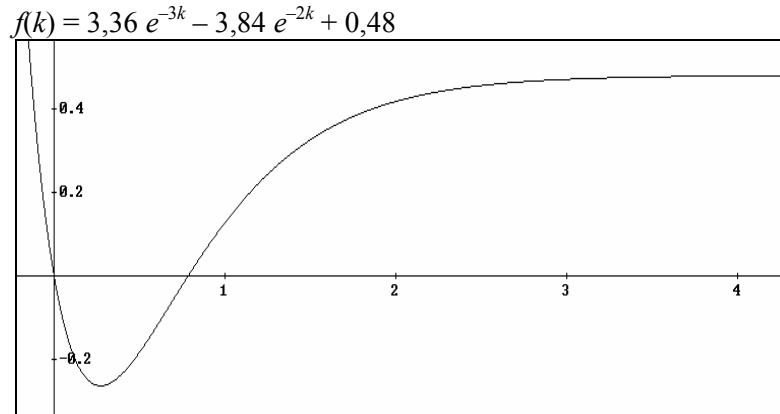


Figura 3

Nótese que la función se aproxima rápidamente a su valor asintótico 0,48 debido a que los sumandos exponenciales decrecen con rapidez. Se puede descartar la presencia de raíces para valores de k más allá de 4. De la gráfica se observa la presencia de una raíz positiva en el intervalo $[0,5; 1]$. Como en ese intervalo la gráfica presenta una curvatura muy reducida la segunda derivada es pequeña y los valores de x_0 y x_1 no tiene que estar muy próximo a r .

Utilizando un programa basado en el algoritmo del método de las secantes con $x_0 = 0,5$ y $x_1 = 1$ y tomando una tolerancia de 0,000005, se obtiene la raíz con cinco cifras decimales exactas. Los resultados se muestran en la tabla 1.

Iteración	x	$E_m(x)$
5	0,5	
6	1	
7	0,79455403	0,20544597
8	0,78431003	0,01024400
9	0,78508192	0,00077189
10	0,78508003	0,00000190

Tabla 1

Luego, con cinco cifras decimales exactas $k = 0,78508$. Los otros dos parámetros se hallan fácilmente. En efecto, como

$$p_1 = \frac{P_L}{1 - ce^{-kt_1}} \quad \text{y} \quad p_2 = \frac{P_L}{1 - ce^{-kt_2}}$$

resulta: $p_1 - p_1 ce^{-kt_1} = P_L \quad \text{y} \quad p_2 - p_2 ce^{-kt_2} = P_L$

Restando miembro a miembro: $p_2 - p_1 = c(p_2 e^{-kt_2} - p_1 e^{-kt_1})$

de donde

$$c = \frac{p_2 - p_1}{p_2 e^{-kt_2} - p_1 e^{-kt_1}} = \frac{2,4 - 1,2}{2,4e^{-3k} - 1,2e^{-k}} = -3,75457$$

y

$$P_L = p_1(1 - ce^{-kt_1}) = 1,2(1 - ce^{-k}) = 3,25488$$

$$\text{La función logística será entonces: } p(t) = \frac{3,25488}{1 + 3,75457e^{-0,78508t}}$$

donde $t = 0$ corresponde al año 2000 y la población se refiere a millones de habitantes. La gráfica de la función en el intervalo de 2000 a 2005 se muestra en la figura 4. Se han agregado los tres puntos correspondientes a los datos que permitieron hallar los parámetros de la curva.

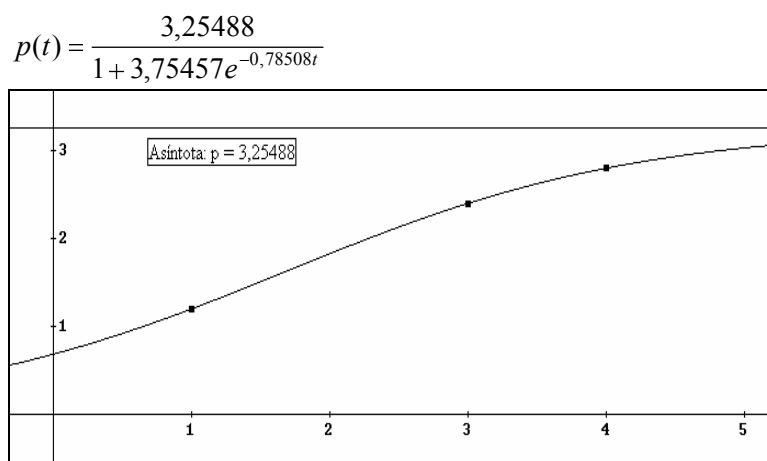


Figura 4

Ejemplo 2

La ecuación $x^3 + 3x^2 + x - 6 = 0$ posee una raíz en el intervalo $[1, 2]$. Para hallar esta raíz con cinco cifras decimales exactas se utilizaron varios métodos en el ejemplo 4 de la sección 2.5. Usando el método de bisección se necesitarían 18 iteraciones; mediante Regula Falsi, 11 iteraciones y mediante Newton – Raphson solamente 4. Halle la raíz por el método de las secantes utilizando $x_0 = 1$ y $x_1 = 2$.

Solución:

En la tabla 2 se muestran los resultados obtenidos. En cinco iteraciones se obtienen las cinco cifras decimales exactas que se deseaban.

Iteración	x	$E_m(x)$
0	1	
1	2	
2	1.05882353	0.94117647
3	1.08126366	0.02244013
4	1.09482415	0.01356049
5	1.09454943	0.00027472
6	1.09455148	0.00000205

Tabla 2

La rapidez de convergencia solo se compara con la de Newton – Raphson. Sin embargo, si se cuenta la cantidad de veces que se necesita evaluar funciones, que es el paso de estos algoritmo que más tiempo consume, el método de Newton – Raphson requirió realizar 8 evaluaciones (dos evaluaciones en cada iteración) mientras que el método de las secantes necesitó evaluar $f(x)$ en las dos aproximaciones iniciales y, a partir de ahí, una evaluación por cada iteración, es decir un total de siete evaluaciones, una menos que el método de Newton – Raphson.

Ejercicios

En todos los ejercicios que siguen se debe utilizar, siempre que se necesite, un programa graficador tanto para separar las raíces como para verificar las hipótesis del método utilizado. También se supone que el algoritmo de las secantes se utilice mediante un programa computacional, preferiblemente confeccionado por usted. Si no cuenta con un programa adecuado, realice los cálculos a mano y obtenga las raíces con solo dos o tres cifras decimales exactas.

1. Se quiere resolver la ecuación $f(x) = 0$ mediante el método de las secantes, donde $f(x)$ es la función cuya gráfica muestra la figura 5. Diga qué aproximaciones iniciales permitirían, con seguridad, obtener la raíz.

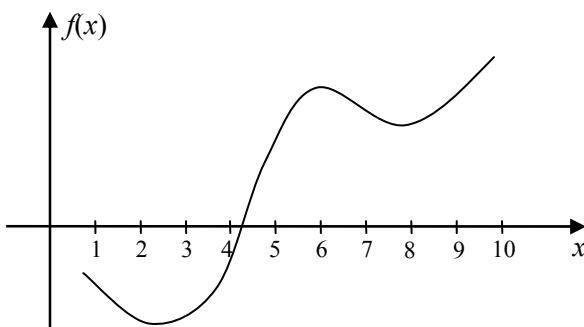


Figura 5

2. Calcule, con cinco cifras decimales exactas, las raíces reales de las siguientes ecuaciones algebraicas mediante el método de las secantes. Posiblemente, ya usted separó las raíces de

estas ecuaciones en los ejercicios de la sección 2.2 y las resolvió por alguno de los métodos anteriores. Compare la cantidad de iteraciones que necesitó con cada uno de los métodos que haya utilizado y la cantidad de veces que hubo que evaluar funciones.

- a) $x^4 + x^3 - x^2 + x - 2 = 0$
 - b) $x^4 - 11x^3 + 41x^2 - 60x + 30 = 0$
 - c) $x^4 - 3x^3 + 10x^2 - 13x + 5 = 0$
 - d) $x^4 + 3x^2 + 2 = 0$
 - e) $x^3 + 7x^2 + 14x + 9 = 0$
3. Utilice el método de las secantes para calcular con cinco cifras decimales exactas, las raíces de las siguientes ecuaciones trascendentes. Posiblemente, ya usted separó las raíces de estas ecuaciones en los ejercicios de la sección 2.2 y las resolvió usando los algoritmos anteriores. En ese caso, compare la cantidad de iteraciones que necesito con cada método usado y la cantidad de veces que hubo que evaluar funciones.
- a) $\sin x - \log x = 0$
 - b) $5e^x - 2x - 10 = 0$
 - c) $(x^2 + 1)\cos x = 1; -10 \leq x \leq 10$
 - d) $2 \tanh x - \sin x - 0,3 = 0$
 - e) $x^4 - 4x^2 - 4x - 16 - \ln|x| = 0$
 - f) $e^x \sin x - 2e^x + 3 = 0$
 - g) $x = \tan^2 x; 0 \leq x \leq 2\pi$
 - h) $\sqrt{x} = 2 \ln x$
4. Una circunferencia tiene su centro en el origen de coordenadas y es tangente a la gráfica de la función $y = e^x$. Halle el radio de la circunferencia con cinco cifras decimales exactas mediante el método de las secantes.
5. Se tiene un tanque cilíndrico de 2 m de diámetro y 10 000 litros de capacidad, colocado con su eje horizontal. Se quiere construir una vara de madera con 10 marcas que indiquen las alturas del líquido en el tanque cuando el mismo contiene 1000, 2000, 3000, ..., 10 000. Determine con precisión de un mm la posición de cada marca. Utilice el método de las secantes.
6. Resuelva el problema anterior para un tanque esférico de la misma capacidad.
7. Se sabe que el área sombreada de la figura 6 es de 10 unidades cuadradas. Halle el valor de a con cuatro cifras decimales exactas. Utilice el método de las secantes.

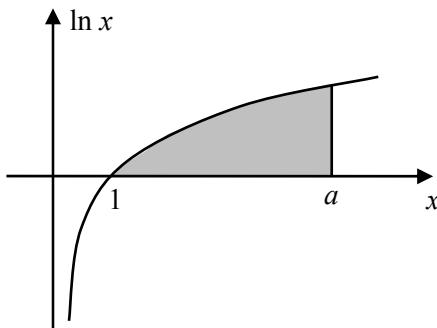


Figura 6

8. Un poste de 10 m de altura estaba situado junto a un muro de un metro de altura y un metro de ancho. El poste se ha partido, como muestra la figura 7, de manera que ha quedado tocando el suelo y el borde del muro. Utilice el método de las secantes para calcular, con precisión de 1 mm, a qué altura del suelo se partió el poste.

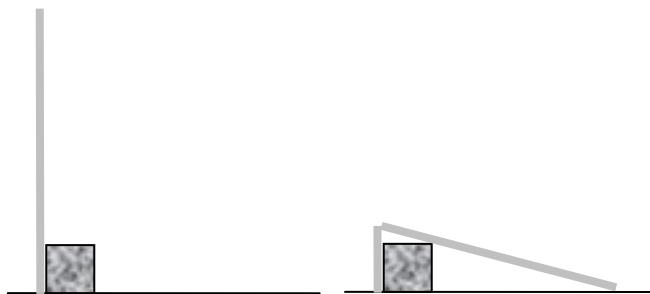


Figura 7

9. Un cuadro de 2 m de altura está colocado a 3 m del piso en una pared, como muestra la figura 8. ¿A qué distancia de la pared deberá colocarse un observador cuyos ojos se encuentran a 1,70 m del piso para contemplar el cuadro bajo un ángulo vertical de 20° ? Utilice el método de las secantes y obtenga la respuesta con un error menor que 1 mm.

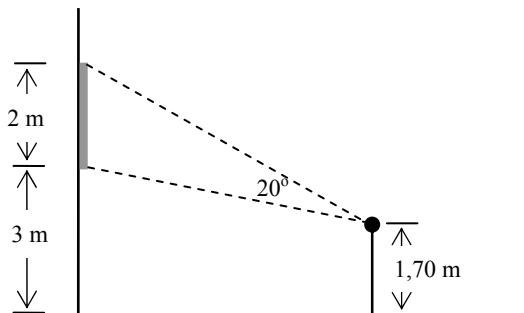


Figura 8

10. En una copa de sección parabólica cayó un palillo de 5 cm de longitud, como se muestra en la figura 9. Uno de sus extremos está en el punto $(-1, 1)$. ¿Dónde está el otro extremo? Utilice el método de las secantes y obtenga la respuesta con cuatro cifras decimales exactas.

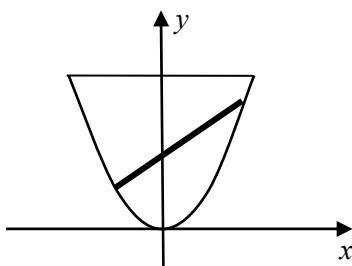


Figura 9

11. En el método Regula Falsi se aproxima la raíz de una ecuación como el punto en que una recta secante interseca al eje x . Explique cuál es entonces la diferencia entre Regula Falsi y el método de las secantes.
12. Muestre la gráfica de una función $f(x)$ que ilustre el hecho de que en el método de las secantes no es indiferente el orden en que se tomen x_0 y x_1 . Para ello, haga que en el ejemplo que usted proponga, al tomar $x_0 = 3$ y $x_1 = 2$ el algoritmo converja a la raíz deseada, pero tomando $x_0 = 2$ y $x_1 = 3$ no converja.
13. Se necesita ajustar el modelo de dos parámetros $f(x) = e^{bx} + e^{ax}$ de manera que $f(1) = 2.6$ y $f'(1) = 1.1$. Determine los valores de los parámetros a y b con cuatro cifras decimales exactas. Utilice el método de las secantes.

2.7 Extensiones del método de Newton – Raphson

El método de Newton – Raphson puede ser extendido en varios sentidos. En esta sección se estudiarán dos de estas extensiones. La aplicación del método a sistemas de dos o más ecuaciones y su adaptación para calcular raíces complejas de polinomios.

El método de Newton – Raphson para sistemas de ecuaciones

Para simplificar la notación solo se considerarán sistemas de dos ecuaciones con dos incógnitas pero la extensión a problemas de mayor dimensión es inmediata. Sea entonces el sistema:

$$\begin{cases} f(x, y) = 0 \\ g(x, y) = 0 \end{cases} \quad (1)$$

Si se utiliza la notación matricial, el par (x, y) se puede representar como la matriz columna:

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$$

así que el sistema se expresaría como: $\begin{cases} f(\mathbf{x}) = 0 \\ g(\mathbf{x}) = 0 \end{cases}$

Empleando la función vectorial $\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$

el sistema de ecuaciones adquiere el aspecto de una ecuación matricial:

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

donde $\mathbf{0}$ representa a la matriz columna nula de orden 2.

El análogo matricial de la derivada $f'(x)$ se define como la matriz jacobiana de $\mathbf{f}(\mathbf{x})$ respecto a \mathbf{x} , y será denotada por $\mathbf{W}(\mathbf{x})$, es decir:

$$\mathbf{W}(\mathbf{x}) = \begin{bmatrix} f_x(x, y) & f_y(x, y) \\ g_x(x, y) & g_y(x, y) \end{bmatrix}$$

Entonces la extensión del algoritmo de Newton – Raphson para un sistema de ecuaciones sería:

$$\mathbf{x}_n = \mathbf{x}_{n-1} - \mathbf{W}^{-1}(\mathbf{x}_{n-1})\mathbf{f}(\mathbf{x}_{n-1}) \quad n = 1, 2, 3, \dots \quad (3)$$

donde: $\mathbf{x}_0 = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$ es la aproximación inicial a la solución buscada.

El teorema que sigue, en cuya demostración no se entrará, establece condiciones suficientes para la convergencia.

Teorema 1

Sea \mathbf{x}^* una raíz del sistema $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ y R una región que contiene a \mathbf{x}^* tal que: f, g y sus primeras y segundas derivadas parciales son continuas y acotadas en R y \mathbf{W} es no singular en R . Entonces, si \mathbf{x}_0 se escoge en R y lo suficientemente cerca de \mathbf{x}^* , la sucesión de vectores $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots$ generada por el algoritmo $\mathbf{x}_n = \mathbf{x}_{n-1} - \mathbf{W}^{-1}(\mathbf{x}_{n-1})\mathbf{f}(\mathbf{x}_{n-1})$ converge hacia \mathbf{x}^* .

■

Cuando el proceso iterativo converge lo hace en forma cuadrática lo cual garantiza una rápida aproximación hacia la raíz.

En el caso de los sistemas de ecuaciones, el error en la aproximación \mathbf{x}_n , se denota $E(\mathbf{x}_n)$ y se define como la norma- ∞ del vector $\mathbf{x}^* - \mathbf{x}_n$, esto es:

$$E(\mathbf{x}_n) = \|\mathbf{x}^* - \mathbf{x}_n\|_\infty = \max \{|x^* - x_n|, |y^* - y_n|\}$$

Puede probarse que, si la norma- ∞ de la diferencia de dos iteraciones sucesivas es pequeña, puede tomarse como cota del error, es decir:

$$E_m(\mathbf{x}_n) = \|\mathbf{x}_n - \mathbf{x}_{n-1}\|_\infty = \max \{|x_n - x_{n-1}|, |y_n - y_{n-1}|\}$$

y, por lo tanto el proceso iterativo se debe detener tan pronto como ambos, $|x_n - x_{n-1}|$ y $|y_n - y_{n-1}|$ son menores que la tolerancia ε que se permitirá.

Como la operación de invertir la matriz \mathbf{W} es muy costosa en tiempo, es preferible transformar la expresión (3) de la siguiente manera.

Trasponiendo el término \mathbf{x}_{n-1} : $\mathbf{x}_n - \mathbf{x}_{n-1} = -\mathbf{W}^{-1}(\mathbf{x}_{n-1})\mathbf{f}(\mathbf{x}_{n-1})$

Llamando $\Delta\mathbf{x}_n$ al vector diferencia: $\Delta\mathbf{x}_n = -\mathbf{W}^{-1}(\mathbf{x}_{n-1})\mathbf{f}(\mathbf{x}_{n-1})$

Si se premultiplica en ambos miembros por la matriz jacobiana:

$$\mathbf{W}(\mathbf{x}_{n-1})\Delta\mathbf{x}_n = -\mathbf{f}(\mathbf{x}_{n-1}) \quad (4)$$

Ahora, el vector $\Delta\mathbf{x}_n$ se halla resolviendo el sistema lineal (4) por algún método numérico eficiente, por ejemplo, el método de Gauss. En el caso de sistemas de dos ecuaciones, se pueden aplicar incluso métodos más elementales, como Cramer.

El sistema (4), si se prefiere, puede ser expresado en forma escalar, escribiendo las matrices en términos de sus componentes:

$$\begin{aligned} & \begin{bmatrix} f_x(x_{n-1}, y_{n-1}) & f_y(x_{n-1}, y_{n-1}) \\ g_x(x_{n-1}, y_{n-1}) & g_y(x_{n-1}, y_{n-1}) \end{bmatrix} \begin{bmatrix} \Delta x_n \\ \Delta y_n \end{bmatrix} = - \begin{bmatrix} f(x_{n-1}, y_{n-1}) \\ g(x_{n-1}, y_{n-1}) \end{bmatrix} \\ \text{de donde: } & \begin{cases} f_x(x_{n-1}, y_{n-1})\Delta x_n + f_y(x_{n-1}, y_{n-1})\Delta y_n = -f(x_{n-1}, y_{n-1}) \\ g_x(x_{n-1}, y_{n-1})\Delta x_n + g_y(x_{n-1}, y_{n-1})\Delta y_n = -g(x_{n-1}, y_{n-1}) \end{cases} \end{aligned} \quad (5)$$

Resolviendo el sistema (5) se obtienen Δx_n y Δy_n y con ellos se determinan x_n y y_n :

$$\begin{aligned} x_n &= x_{n-1} + \Delta x_n \\ y_n &= y_{n-1} + \Delta y_n \end{aligned}$$

El algoritmo se detiene cuando $|\Delta x_n|$ y $|\Delta y_n|$ son ambos menores que ε .

Algoritmo en seudo código

Se desea resolver el sistema de ecuaciones

$$\begin{cases} f(x, y) = 0 \\ g(x, y) = 0 \end{cases}$$

el cual posee una raíz $\mathbf{x}^* = (x^*, y^*)$ en una región R del plano xy . Se supone que las funciones y sus derivadas de primer y segundo orden son continuas y acotadas en R , que \mathbf{W} no es singular en R y que el par (x_0, y_0) , formado por las aproximaciones iniciales, está en R y suficientemente próximo a \mathbf{x}^* , de modo que el proceso iterativo converge.

El algoritmo requiere como datos: las funciones f, g, f_x, f_y, g_x, g_y , las aproximaciones iniciales x_0 , y y_0 y la tolerancia ε que se permitirá.

```

 $x := x_0$ 
 $y := y_0$ 
repeat
    Resolver el sistema  $\begin{cases} f_x(x, y)\Delta x + f_y(x, y)\Delta y = -f(x, y) \\ g_x(x, y)\Delta x + g_y(x, y)\Delta y = -g(x, y) \end{cases}$ 
     $Error := \max \{|\Delta x|, |\Delta y|\}$ 
     $x := x + \Delta x$ 

```

$y := y + \Delta y$
until $Error < \varepsilon$
 La solución del sistema es (x, y) con error menor que $Error$
 Terminar

Ejemplo 1

Halle la solución del sistema de ecuaciones

$$\begin{cases} xy = 1 \\ x = y^2 \end{cases}$$

mediante el método de Newton – Raphson tomando aproximación inicial $x_0 = 0,7$ y $y_0 = 1,5$.
 Obtenga el resultado con cinco cifras decimales exactas.

Solución:

Primero el sistema debe escribirse de manera que en cada ecuación el segundo miembro sea cero:

$$\begin{cases} xy - 1 = 0 \\ x - y^2 = 0 \end{cases}$$

las funciones $f(x, y) = xy - 1$ y $g(x, y) = x - y^2$ son continuas y poseen primeras y segundas derivadas continuas y acotadas en cualquier región R acotada. La aproximación inicial se ignora si está o no cerca de la solución, así que todo lo que se puede hacer es utilizar el proceso iterativo y ver si converge. Este ejemplo, que solo posee un valor didáctico, tiene una solución evidente en $x = 1$ y $y = 1$, pero en un caso real, esta información no se tiene.

El sistema de ecuaciones lineales que habrá que resolver en cada iteración es:

$$\begin{cases} f_x(x, y)\Delta x + f_y(x, y)\Delta y = -f(x, y) \\ g_x(x, y)\Delta x + g_y(x, y)\Delta y = -g(x, y) \end{cases}$$

en este caso:

$$\begin{cases} y\Delta x + x\Delta y = 1 - xy \\ \Delta x - 2y\Delta y = y^2 - x \end{cases}$$

$$\begin{aligned} x &= x_0 = 0,7 \\ y &= y_0 = 1,5 \end{aligned}$$

Iteración 1:

Sistema lineal:

$$\begin{cases} 1,5\Delta x + 0,7\Delta y = -0,05 \\ \Delta x - 3\Delta y = 1,55 \end{cases}$$

Solución del sistema lineal:

$$\begin{aligned} \Delta x &= 0,179808 \\ \Delta y &= -0,456731 \end{aligned}$$

$$\begin{aligned} x &= 0,7 + \Delta x = 0,879808 \\ y &= 1,5 + \Delta y = 1,043269 \end{aligned}$$

$$E_m(\mathbf{x}_1) = \max \{|\Delta x|, |\Delta y|\} = 0,457$$

Iteración 2:

Sistema lineal:

$$\begin{cases} 1,043269\Delta x + 0,879808\Delta y = 0,082124 \\ \Delta x - 2,086538\Delta y = 0,208602 \end{cases}$$

Solución del sistema lineal:

$$\begin{aligned} \Delta x &= 0,116103 \\ \Delta y &= -0,044331 \end{aligned}$$

$$\begin{aligned} x &= 0,879808 + \Delta x = 0,995911 \\ y &= 1,043269 + \Delta y = 0,998938 \end{aligned}$$

$$E_m(\mathbf{x}_2) = \max \{|\Delta x|, |\Delta y|\} = 0,117$$

Iteración 3:

Sistema lineal:

$$\begin{cases} 0,998938\Delta x + 0,995911\Delta y = 0,005147 \\ \Delta x - 1,997876\Delta y = 0,001966 \end{cases}$$

Solución del sistema lineal:

$$\begin{aligned} \Delta x &= 0,004092 \\ \Delta y &= 0,001064 \end{aligned}$$

$$\begin{aligned} x &= 0,995911 + \Delta x = 1,000003 \\ y &= 0,998938 + \Delta y = 1,000002 \end{aligned}$$

$$E_m(\mathbf{x}_3) = \max \{|\Delta x|, |\Delta y|\} = 0,000409$$

Iteración 4:

Sistema lineal:

$$\begin{cases} 1,000002\Delta x + 1,000003\Delta y = -0,000005 \\ \Delta x - 2,000004\Delta y = 0,000001 \end{cases}$$

Solución del sistema lineal:

$$\begin{aligned} \Delta x &= -0,000003 \\ \Delta y &= -0,000002 \end{aligned}$$

$$\begin{aligned} x &= 1,000003 + \Delta x = 1,000000 \\ y &= 1,000002 + \Delta y = 1,000000 \end{aligned}$$

$$E_m(\mathbf{x}_4) = \max \{|\Delta x|, |\Delta y|\} = 0,000003$$

El proceso se detiene por ser $E_m(\mathbf{x}_4) < \varepsilon = 0,000005$

Resultado: Con cinco cifras decimales exactas la solución es $x = 1,000000$ $y = 1,000000$

El método de Newton – Bairstow

El método de Newton – Raphson puede utilizarse para buscar raíces imaginarias de una función. En ese caso la función y sus derivadas hay que considerarlas como funciones de variable compleja y todas las operaciones deben ser realizadas en el campo complejo, lo cual genera un problema bastante complicado cuando debe ser implementado en una computadora.

El método de Newton – Bairstow, que es la segunda de las extensiones de Newton – Raphson que se estudiarán en esta sección, es una aplicación del algoritmo para la solución de sistemas de dos ecuaciones, al problema de hallar las raíces imaginarias de una ecuación polinomial que posea todos sus coeficientes reales. El mayor atractivo del algoritmo es que permite hallar un par de raíces imaginarias realizando todas las operaciones en el campo real.

Considérese entonces la ecuación polinomial de grado $n > 2$:

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0 \quad (6)$$

con todos sus coeficientes reales y $a_n \neq 0$. Nótese que los coeficientes se han numerado en orden diferente a como se hizo al principios de este capítulo (Sección 2.2); en este método es preferible seguir este criterio.

Como se sabe, en las ecuaciones polinomiales con coeficientes reales las raíces imaginarias se presentan en pares conjugados. Supóngase que la ecuación (6) posee un par de raíces imaginarias conjugadas:

$$r_1 = a + bi \quad y \quad r_2 = a - bi$$

Si r_1 y r_2 son raíces de (5), esto significa que $(x - r_1)$ y $(x - r_2)$ son factores del polinomio:

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

En este caso, el polinomio de segundo grado que resulta de multiplicar ambos factores de primer grado, también será un factor de $p(x)$. Es decir que $p(x)$ contiene como factor al polinomio:

$$(x - r_1)(x - r_2) = (x - a - bi)(x - a + bi) = x^2 - 2ax + a^2 + b^2$$

En lo que sigue, para simplificar la exposición, se llamará:

$$\begin{aligned} r &= 2a \\ s &= -(a^2 + b^2) \end{aligned} \quad (7)$$

Con esto el factor cuadrático queda de la forma: $x^2 - rx - s$.

Ahora el problema de encontrar las raíces imaginarias r_1 y r_2 se transforma en este equivalente: Hallar coeficientes r y s tales que el polinomio $x^2 - rx - s$ sea un factor de $p(x)$.

Para valores arbitrarios de r y de s , el polinomio $x^2 - rx - s$ no será, en general, un factor de $p(x)$, sino que se tendrá:

$$p(x) = (x^2 - rx - s)q(x) + \text{Resto}$$

donde $q(x)$ es un polinomio de grado $n - 2$ que es el cociente de dividir $p(x)$ entre $x^2 - rx - s$ y Resto es un polinomio de grado menor o igual que 1. El problema original de encontrar las raíces imaginarias r_1 y r_2 puede enunciarse de esta otra forma alternativa:

Hallar coeficientes r y s tales que el Resto sea cero.

El siguiente teorema del Álgebra Superior, que es una generalización del algoritmo de división sintética, simplifica el proceso de encontrar el cociente $q(x)$ y el Resto de la división.

Teorema 2

Sea $p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$ ($n > 2$) y el polinomio $q(x)$, de grado $n - 2$, el cociente de dividir $p(x)$ por $x^2 - rx - s$. Si se denota:

$$q(x) = b_n x^{n-2} + b_{n-1} x^{n-3} + \dots + b_3 x + b_2 \quad \text{y} \quad \text{Resto} = b_1(x - r) + b_0$$

entonces:

$$\begin{aligned} b_n &= a_n \\ b_{n-1} &= a_{n-1} + rb_n \\ b_{n-2} &= a_{n-2} + rb_{n-1} + sb_n \\ b_{n-3} &= a_{n-3} + rb_{n-2} + sb_{n-1} \\ &\vdots \\ b_1 &= a_1 + rb_2 + sb_3 \\ b_0 &= a_0 + rb_1 + sb_2 \end{aligned} \tag{8}$$

La demostración consiste en verificar que, con los coeficientes dados en (8), se cumple

$$p(x) = (x^2 - rx - s)q(x) + \text{Resto}$$

■

Las relaciones (8) se pueden representar como un algoritmo que permite, conociendo valores r y s y el conjunto de coeficientes $A = \{a_0, a_1, a_2, \dots, a_n\}$ obtener el conjunto de coeficiente $B = \{b_0, b_1, b_2, \dots, b_n\}$. Si este algoritmo de división sintética se representa formalmente como DivSint , se tiene:

$$B = \text{DivSint}(r, s, A)$$

El Resto de la división depende de los valores que se asigne a r y s , por tanto puede escribirse:

$$\text{Resto}(r, s) = b_1(r, s)(x - r) + b_0(r, s)$$

Como se necesita que el resto sea cero, el problema se ha convertido en hallar los valores de r y s que satisfagan el sistema:

$$\begin{cases} b_1(r, s) = 0 \\ b_0(r, s) = 0 \end{cases} \quad (9)$$

La dependencia de b_1 y b_0 respecto r y s es bastante complicada y viene dada por el conjunto (8) de ecuaciones. El sistema (9) se resolverá iterativamente mediante el método de Newton – Raphson para sistemas de dos ecuaciones, desarrollado en las páginas anteriores. Para ello, se parte de aproximaciones iniciales r_0 y s_0 , y, a partir de ellas, en cada paso del algoritmo se calcula

$$\begin{aligned} r_n &= r_{n-1} + \Delta r \\ s_n &= s_{n-1} + \Delta s \end{aligned} \quad n = 1, 2, 3, \dots$$

donde los incrementos Δr y Δs se obtienen resolviendo el sistema:

$$\begin{cases} \left. \frac{\partial b_1}{\partial r} \right|_{n-1} \Delta r + \left. \frac{\partial b_1}{\partial s} \right|_{n-1} \Delta s = -b_1(r_{n-1}, s_{n-1}) \\ \left. \frac{\partial b_0}{\partial r} \right|_{n-1} \Delta r + \left. \frac{\partial b_0}{\partial s} \right|_{n-1} \Delta s = -b_0(r_{n-1}, s_{n-1}) \end{cases}$$

Donde el subíndice $n - 1$ en las derivadas indica que se deben evaluar en la aproximación r_{n-1} y s_{n-1} . Para hallar las derivadas parciales respecto a r , se deriva en cada una de las expresiones (8) respecto a r . Se obtiene:

$$\begin{aligned} \frac{\partial b_n}{\partial r} &= 0 \\ \frac{\partial b_{n-1}}{\partial r} &= b_n \\ \frac{\partial b_{n-2}}{\partial r} &= b_{n-1} + r \frac{\partial b_{n-1}}{\partial r} \\ \frac{\partial b_{n-3}}{\partial r} &= b_{n-2} + r \frac{\partial b_{n-2}}{\partial r} + s \frac{\partial b_{n-1}}{\partial r} \\ &\vdots \\ \frac{\partial b_1}{\partial r} &= b_2 + r \frac{\partial b_2}{\partial r} + s \frac{\partial b_3}{\partial r} \\ \frac{\partial b_0}{\partial r} &= b_1 + r \frac{\partial b_1}{\partial r} + s \frac{\partial b_2}{\partial r} \end{aligned} \quad (10)$$

Para simplificar la notación, conviene llamar las derivadas respecto a r de la siguiente forma:

$$\begin{aligned}
c_n &= \frac{\partial b_{n-1}}{\partial r} \\
c_{n-1} &= \frac{\partial b_{n-2}}{\partial r} \\
c_{n-2} &= \frac{\partial b_{n-3}}{\partial r} \\
&\vdots \\
c_2 &= \frac{\partial b_1}{\partial r} \\
c_1 &= \frac{\partial b_0}{\partial r}
\end{aligned}$$

y el sistema (10) se convierte en:

$$\begin{aligned}
c_n &= b_n \\
c_{n-1} &= b_{n-1} + rc_n \\
c_{n-2} &= b_{n-2} + rc_{n-1} + sc_n \\
&\vdots \\
c_2 &= b_2 + rc_2 + sc_4 \\
c_1 &= b_1 + rc_2 + sc_3
\end{aligned} \tag{11}$$

Si se compara el sistema (11) con el (8) se observará que el algoritmo para hallar el conjunto de derivadas $C = \{c_0, c_1, c_2, \dots, c_n\}$ vuelve a ser DivSint salvo que el coeficiente c_0 no interesa. Esto es:

$$C = \text{DivSint}(r, s, B)$$

Para hallar las derivadas parciales respecto a s , se deriva en cada una de las ecuaciones (8) respecto a esa variable y se obtiene:

$$\begin{aligned}
\frac{\partial b_n}{\partial s} &= 0 \\
\frac{\partial b_{n-1}}{\partial s} &= 0 \\
\frac{\partial b_{n-2}}{\partial s} &= b_n \\
\frac{\partial b_{n-3}}{\partial s} &= r \frac{\partial b_{n-2}}{\partial s} + b_{n-1} + s \frac{\partial b_{n-1}}{\partial s} = r \frac{\partial b_{n-2}}{\partial s} + b_{n-1} \\
&\vdots \\
\frac{\partial b_1}{\partial s} &= r \frac{\partial b_2}{\partial s} + b_3 + s \frac{\partial b_3}{\partial s} \\
\frac{\partial b_0}{\partial s} &= r \frac{\partial b_1}{\partial s} + b_2 + s \frac{\partial b_2}{\partial s}
\end{aligned} \tag{12}$$

Introduciendo ahora las variables $D = \{d_0, d_1, d_2, \dots, d_n\}$ como:

$$\begin{aligned}
d_n &= \frac{\partial b_{n-2}}{\partial s} \\
d_{n-1} &= \frac{\partial b_{n-3}}{\partial s} \\
d_{n-2} &= \frac{\partial b_{n-4}}{\partial s} \\
&\vdots \\
d_3 &= \frac{\partial b_1}{\partial s} \\
d_2 &= \frac{\partial b_0}{\partial s}
\end{aligned}$$

el sistema (12) se reduce a:

$$\begin{aligned}
d_n &= b_n \\
d_{n-1} &= b_{n-1} + rd_n \\
d_{n-2} &= b_{n-2} + rd_{n-1} + sd_n \\
&\vdots \\
d_3 &= b_3 + rd_4 + sd_5 \\
d_2 &= b_2 + rd_3 + sd_4
\end{aligned} \tag{13}$$

Si se compara las ecuaciones (13) con las (11), se comprobará que los coeficientes D y C coinciden hasta $d_2 = c_2$; por esta causa, no será necesario utilizar las expresiones (13) ya que: $d_3 = c_3$ y $d_2 = c_2$. Volviendo al sistema de ecuaciones del método de Newton – Raphson

$$\begin{cases} \left. \frac{\partial b_1}{\partial r} \right|_{n-1} \Delta r + \left. \frac{\partial b_1}{\partial s} \right|_{n-1} \Delta s = -b_1(r_{n-1}, s_{n-1}) \\ \left. \frac{\partial b_0}{\partial r} \right|_{n-1} \Delta r + \left. \frac{\partial b_0}{\partial s} \right|_{n-1} \Delta s = -b_0(r_{n-1}, s_{n-1}) \end{cases}$$

y sustituyendo las derivadas por los coeficientes correspondientes:

$$\begin{aligned}
\frac{\partial b_1}{\partial r} &= c_2; \quad \frac{\partial b_0}{\partial r} = c_1; \quad \frac{\partial b_1}{\partial s} = d_3 = c_3; \quad \frac{\partial b_0}{\partial s} = d_2 = c_2 \\
\begin{cases} c_2 \Delta r + c_3 \Delta s = -b_1 \\ c_1 \Delta r + c_2 \Delta s = -b_0 \end{cases} \tag{14}
\end{aligned}$$

Resolviendo el sistema (14) se determinan los Δr y Δs y, según el método de Newton – Raphson,

$$\begin{aligned}
r_n &= r_{n-1} + \Delta r \\
s_n &= s_{n-1} + \Delta s
\end{aligned}$$

Los valores iniciales de r y s se determinan de alguna aproximación inicial $\alpha \pm \beta i$ a las raíces buscadas, para ello, aplicando (7):

$$\begin{aligned} r_0 &= 2\alpha \\ s_0 &= -(\alpha^2 + \beta^2) \end{aligned}$$

Una vez hallados los valores de r y s , las mismas ecuaciones (7) permiten hallar los valores de a y b . En efecto:

Como $r = 2a$ y $s = -(a^2 + b^2)$

Se obtiene: $a = \frac{r}{2}$ y $b = \sqrt{-a^2 - s}$

Algoritmo en seudo código

Se desea hallar un par de raíces imaginarias $a \pm bi$ de la ecuación algebraica

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0$$

con coeficientes reales, $n > 2$ y $a_n \neq 0$ que están próximas a $\alpha \pm \beta i$, con error menor que ε . Se supone que las aproximaciones iniciales $\alpha \pm \beta i$ se encuentran lo suficientemente cerca de las raíces $a \pm bi$ de modo que el algoritmo iterativo converja. El algoritmo supone conocidos, n , los coeficientes del polinomio: $a_0, a_1, a_2, \dots, a_n$, la aproximación inicial $\alpha \pm \beta i$ y la tolerancia ε con que se desea hallar la raíz.

```

 $r := 2\alpha$ 
 $s := -(\alpha^2 + \beta^2)$ 
repeat
    Calcular los coeficientes  $B$  mediante:  $B := \text{DivSint}(r, s, A)$ 
    Calcular los coeficientes  $C$  mediante:  $C := \text{DivSint}(r, s, B)$ 
    Resolver el sistema lineal:  $\begin{cases} c_2 \Delta r + c_3 \Delta s = -b_1 \\ c_1 \Delta r + c_2 \Delta s = -b_0 \end{cases}$ 
     $Error := \max \{|\Delta r|, |\Delta s|\}$ 
     $r := r + \Delta r$ 
     $s := s + \Delta s$ 
until  $Error < \varepsilon$ 
 $a := \frac{r}{2}$ 
 $b := \sqrt{-a^2 - s}$ 
Las raíces buscadas son  $a \pm bi$  con error menor que  $Error$ 
Terminar

```

El seudo código del algoritmo DivSint se ofrece a continuación:

El algoritmo DivSint supone conocidos los valores de r y s y los coeficientes de entrada: $a_0, a_1, a_2, \dots, a_n$. Ofrece como resultado los coeficientes $b_0, b_1, b_2, \dots, b_n$.

```

 $b_n := a_n$ 
 $b_{n-1} := a_{n-1} + rb_n$ 
 $k := n - 2$ 
do while  $k \geq 0$ 
     $b_k := a_k + rb_{k+1} + sb_{k+2}$ 
     $k := k - 1$ 
end
Los coeficientes resultantes son  $b_0, b_1, b_2, \dots, b_n$ 
Terminar

```

Ejemplo 2

Halle el par de raíces imaginarias de la ecuación $x^4 + 5,2x^3 + 6,5x^2 - 0,2x + 13 = 0$ que están próximas a $0,5 \pm 0,9i$ con cinco cifras decimales exactas.

Solución:

$$r = r_0 = 2\alpha = 2(0,5) = 1$$

$$s = s_0 = -(\alpha^2 + \beta^2) = -(0,5^2 + 0,9^2) = -1,06 \approx -1$$

Iteración 1:

Coeficientes:	$A:$	1	5,2	6,5	-0,2	13
	$B:$	1	6,2	11,7	5,3	6,6
	$C:$	1	7,2	17,9	16,0	---

$$b_1 = 5,3; \quad b_0 = 6,6 \quad c_3 = 7,2 \quad c_2 = 17,9 \quad c_1 = 16,0$$

Sistema lineal:

$$\begin{cases} 17,9\Delta r + 7,2\Delta s = -5,3 \\ 16,0\Delta r + 17,9\Delta s = -6,6 \end{cases}$$

Solución del sistema lineal:

$$\begin{aligned} \Delta r &= -0,2307 \\ \Delta s &= -0,1625 \end{aligned}$$

$$\begin{aligned} r &= 1 + \Delta r = 0,7693 \\ s &= -1 + \Delta s = -1,1625 \end{aligned}$$

$$E_m(\mathbf{X}_1) = \max \{|\Delta r|, |\Delta s|\} = 0,2307$$

Iteración 2:

Coeficientes:	$A:$	1	5,2	6,5	-0,2	13
	$B:$	1	5,9693	9,92968	0,49959	1,84108
	$C:$	1	6,7386	13,95118	2,89902	---

$$b_1 = 0,49959 \quad b_0 = 1,84108 \quad c_3 = 6,7386 \quad c_2 = 13,95118 \quad c_1 = 2,89902$$

Sistema lineal:

$$\begin{cases} 13,95118\Delta r + 6,7386\Delta s = -0,49959 \\ 2,89902\Delta r + 13,95118\Delta s = -1,84108 \end{cases}$$

Solución del sistema lineal:

$$\begin{aligned} \Delta r &= 0,03105 \\ \Delta s &= -0,13842 \end{aligned}$$

$$\begin{aligned} r &= 0,7693 + \Delta r = 0,80035 \\ s &= -1,1625 + \Delta s = -1,30092 \end{aligned}$$

$$E_m(\mathbf{X}_1) = \max \{|\Delta r|, |\Delta s|\} = 0,13842$$

Iteración 3: (En esta iteración y las siguientes, solo se muestran los resultados finales)

Solución del sistema lineal:

$$\begin{aligned} \Delta r &= -0,000349 \\ \Delta s &= 0,000919 \end{aligned}$$

$$\begin{aligned} r &= 0,80035 + \Delta r = 0,800001 \\ s &= -1,30092 + \Delta s = -1,300001 \end{aligned}$$

$$E_m(\mathbf{X}_1) = \max \{|\Delta r|, |\Delta s|\} = 0,000919$$

Iteración 4:

Solución del sistema lineal:

$$\begin{aligned} \Delta r &= -0,000001 \\ \Delta s &= 0,000001 \end{aligned}$$

$$\begin{aligned} r &= 0,800001 + \Delta r = 0,800000 \\ s &= -1,300001 + \Delta s = -1,300000 \end{aligned}$$

$$E_m(\mathbf{X}_1) = \max \{|\Delta r|, |\Delta s|\} = 0,000001$$

El proceso se detiene por ser $E_m(\mathbf{X}_4) < \varepsilon = 0,000005$

Con cinco cifras decimales exactas se obtuvo: $r = 0,800000 \quad s = -1,300000$
Por tanto, el polinomio dado posee un factor cuadrático $x^2 - 0,800000x + 1,300000$

Este factor posee las raíces imaginarias $a \pm bi$ donde:

$$\begin{aligned} a &:= \frac{r}{2} = 0,400000 \\ b &:= \sqrt{-a^2 - s} = \sqrt{-0,160000 + 1,300000} = \sqrt{1,140000} = 1,067708 \end{aligned}$$

Resultado: Con cinco cifras decimales exactas, las raíces imaginarias buscadas son:

$$0,400000 \pm 1,067708 i$$

Ejercicios

1. Resuelva el sistema de dos ecuaciones que sigue mediante Newton – Raphson para sistemas, con cuatro cifras decimales exactas. Grafique las ecuaciones correspondientes para hallar la aproximación inicial necesaria.

$$\begin{cases} x^2 + y^2 = 4 \\ 4(x - 1)^2 + 9(y - 2)^2 = 36 \end{cases}$$

2. Resuelva el sistema que sigue utilizando el método de Newton – Raphson para sistemas. Tome como aproximación inicial: $x = 1$ y $y = 0,8$ y obtenga cuatro cifras decimales exactas.

$$\begin{cases} 5x^2 + 2y^2 = 7 \\ \sin x + \cos 2y = 1 \end{cases}$$

3. En la figura 1 se muestran las graficas de las ecuaciones 1) $x^2y + 3xy^2 + 2x - 3y = 4$ y 2) $2xy + 4y^2 + x^3 = 2$. Utilice el método de Newton – Raphson para sistemas para hallar los puntos de intersección con cuatro cifras decimales exactas.

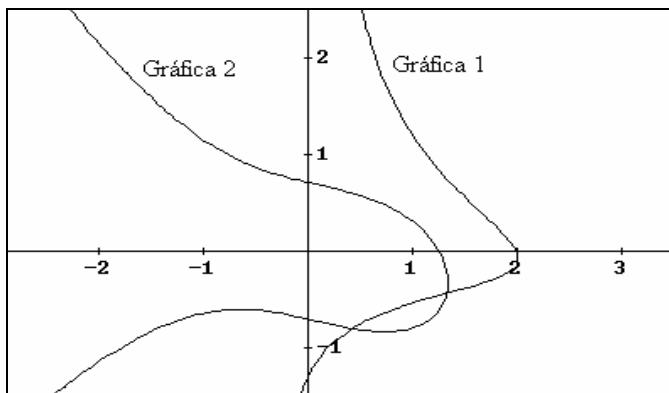


Figura 1

4. Para ajustar el modelo de dos parámetros $f(x) = e^{bx} + e^{ax}$ (vea el ejercicio 13 de la sección anterior) de manera que $f(1) = 2,6$ y $f'(1) = 1,1$, se necesita resolver un sistema de dos ecuaciones. Resuévalo mediante el método de Newton – Raphson para sistemas, con cuatro cifras decimales exactas. Tome como aproximación inicial $a = 0,9$ y $b = -0,6$.
5. La ecuación $z^3 - 8z^2 + 5z - 2 = 0$ tiene una raíz compleja cercana a $0,3 + 0,4i$. Hállela de la siguiente forma: sustituya $z = x + yi$ y descomponga la ecuación original en el sistema: $\operatorname{Re}(z^3 - 8z^2 + 5z - 2) = 0$ y $\operatorname{Im}(z^3 - 8z^2 + 5z - 2) = 0$, después, resuelva el sistema mediante el método de Newton – Raphson para sistemas de ecuaciones. Halle las soluciones con cuatro cifras decimales exactas.
6. Resuelva el problema anterior aplicando el método de Newton – Bairstow.

7. Compruebe que la ecuación $x^4 - 9x^3 + 26x^2 - 24x + 8 = 0$ no posee raíces reales. Halle todas sus raíces con cuatro cifras decimales exactas. Utilice el método de Newton – Bairstow. Una de sus raíces está próxima a $4 + i$.
8. Resuelva la ecuación $x^4 + 3x^2 + x + 2 = 0$ con 4 cifras decimales exactas. Una de sus raíces complejas está cerca de $0,3 + 1,6 i$. Emplee el método de Newton – Bairstow.

Otras lecturas recomendadas

Varios de los resultados algebraicos útiles para la separación de raíces de ecuaciones algebraicas no han sido demostrados por razones de brevedad. El lector interesado en estos temas puede dirigirse a varios clásicos en esta materia, tal como “Álgebra Superior” de Pablo Miquel o al capítulo sobre ecuaciones algebraicas del Análisis Matemático de Rey Pastor.

Algunos métodos importantes para la solución de ecuaciones tales como los métodos de Graeffe, de Bernoulli, el modificado de Newton (que solo fue mencionado en un ejercicio) y el de Muller, no han sido tratados por razones de espacio. Todos ellos aparecen con un enfoque claro en el texto “Computational Mathematics” de B. P. Demidovich e I. A. Maron, de la editorial MIR y también en “An Introduction to Numerical Analysis” de K. E. Atkinson.

En cuanto a los algoritmos computacionales, algunas cuestiones importantes han sido pasadas por alto debido a que el interés de este texto no está dirigido a la elaboración de algoritmos de carácter profesional. Los lectores interesados en estos aspectos encontrarán valiosas informaciones en “Computer Methods for Mathematical Computations” G. E. Forsythe, M. A. Malcolm y C. B. Moler, editado por Prentice-Hall.

El método iterativo general es uno de los aspectos más centrales de la teoría de los procesos iterativos; aquí solo algunos aspectos han sido tratados. Un enfoque serio y riguroso puede encontrarse en “Analysis of Numerical Methods” de E. Isaacson y H. B. Keller.

Principales ideas del capítulo

- Con una gran frecuencia aparecen en la práctica ecuaciones y sistemas de ecuaciones que no pueden ser resueltos por los métodos analíticos exactos.
- Antes de intentar resolver una ecuación hay que separar sus raíces, es decir, hallar intervalos cada uno de los cuales contenga solamente una raíz.
- La forma más simple de separar las raíces de la ecuación $f(x) = 0$ es graficar la función $f(x)$ con un programa graficador.
- Para las ecuaciones algebraicas existen varios resultados fáciles de aplicar como el teorema de las n raíces, la regla de los signos de Descartes y la fórmula de Lagrange que permiten acotar la cantidad y la localización de las raíces reales.
- El método de bisección es el algoritmo más simple y seguro para hallar las raíces reales de una ecuación. Descansa en hipótesis muy poco exigentes y ofrece una cota del error muy segura. Es un algoritmo poco eficiente pero muy robusto y su rapidez de convergencia no depende más que del intervalo inicial y de la tolerancia que se exija.
- El método Regula Falsi puede considerarse como una modificación del método de bisección para mejorar la velocidad de convergencia, lo cual casi siempre consigue. El algoritmo es

sencillo y funciona bien cuando la función $f(x)$ no presenta cambios grandes en su derivada en el intervalo $[a, b]$; esto siempre se puede conseguir tomando un intervalo de partida suficientemente pequeño.

- El método iterativo general descansa en una idea muy simple: expresar la ecuación $f(x) = 0$ en la forma $x = g(x)$ y definir el proceso iterativo $x_n = g(x_{n-1})$ el cual converge a la raíz de la ecuación original si en un entorno de ella la derivada de g es modularmente menor que un número $K < 1$ y si la aproximación inicial x_0 se toma en ese entorno.
- El método iterativo general es muy importante desde un punto de vista teórico pero su utilización práctica es limitada debido a que la manera en que la ecuación $f(x) = 0$ se transforma en $x = g(x)$ decide si se obtendrá un buen algoritmo iterativo o si será uno de convergencia lenta o incluso divergente.
- El método de Newton – Raphson o método de las tangentes está basado en la idea geométrica de aproximar la gráfica de $f(x)$ por la de su tangente y buscar el intercepto de la tangente en lugar del de $f(x)$. Es una forma efectiva de definir un proceso iterativo asegurando una alta rapidez de convergencia.
- El método de Newton – Raphson requiere del cumplimiento de hipótesis más fuertes que Bisección o Regula Falsi, ya que $f(x)$ debe ser derivable dos veces y sus dos primeras derivadas no se pueden anular en un entorno de la raíz buscada. Requiere, además, como dato la derivada de $f(x)$, lo cual puede ser un trabajo engorroso y que hay que realizar, por lo general, manualmente.
- El método de Newton – Raphson se puede extender en varios sentidos: solución de sistemas de ecuaciones no lineales, cálculo de raíces imaginarias o determinación de factores cuadráticos en ecuaciones algebraicas.
- El método de las secantes es una modificación del método de Newton – Raphson en el cual se utilizan secantes en lugar de tangentes para aproximar la función $f(x)$. Esto evita tener que conocer la derivada de $f(x)$ que es uno de los principales problemas de aquel método.
- El método de las secantes posee requisitos similares al método de Newton – Raphson para su convergencia.
- El orden de convergencia permite cuantificar la eficiencia de los métodos iterativos tratados. En el método de bisección y Regula Falsi la convergencia es de primer orden: $E_m(x_n) = kE_m(x_{n-1})$ aunque en el caso de bisección $k = 0,5$ y en Regula Falsi se logra generalmente valores más pequeños de k , que garantizan una convergencia más rápida. En el método de Newton – Raphson la convergencia es de segundo orden: $E_m(x_n) = k [E_m(x_{n-1})]^2$ y en el método de las secantes el orden es 1,618.
- Aunque el método de las secantes posee una rapidez de convergencia ligeramente menor que el de Newton – Raphson, como solo requiere evaluar una función en cada paso del proceso iterativo, resulta en general más eficiente, si se mide en tiempo de cómputo.
- El método de Newton – Raphson para sistemas de ecuaciones es una extensión del problema escalar $f(x) = 0$ al problema matricial $\mathbf{F}(\mathbf{X}) = \mathbf{0}$. En lugar de aparecer una división por la derivada de f aparece un producto por la inversa de la matriz jacobiana de \mathbf{F} . Se trata de un algoritmo mucho más complejo que su análogo escalar pues en cada iteración debe resolverse un sistema lineal de ecuaciones.
- El método de Newton – Bairstow es una aplicación del método de Newton – Raphson para sistemas que permite encontrar un factor cuadrático de un polinomio, en forma iterativa. Esto permite hallar las raíces imaginarias de una ecuación algebraica realizando todas las operaciones con aritmética real.

Auto examen

1. Separe las raíces de las siguientes ecuaciones en intervalos de longitud 0,5. Cuando sea posible, aplique previamente las reglas de Descartes y de Lagrange.
 - a) $x^2 = \cos 2x$
 - b) $x^5 - 4x^3 + 3x^2 - 2x + 1 = 0$
2. Desde un punto de vista geométrico, ¿en qué se parecen y en qué difieren los métodos Regula Falsi y de las secantes?
3. ¿Qué condiciones debe cumplir la función $f(x)$ para resolver la ecuación $f(x) = 0$ mediante el método de Newton – Raphson?
4. A continuación se muestra un algoritmo en seudo código. Identifique qué problema se está resolviendo y qué método se está aplicando y repare un error que se ha introducido en el algoritmo.

$f(x)$, a , b y T son datos del problema

$y_a := f(a)$

$y_b := f(b)$

$x_{previa} := 10^{10}$

repeat

$$x := a - \frac{b-a}{y_b - y_a} y_a$$

$y_x := f(x)$

$Error := |x - x_{previa}|$

if $y_a y_x < 0$ **then**

$b := x$

$y_b := y_x$

else

$a := x$

$y_a := y_x$

end

$x_{previa} := x$

until $Error > T$

La solución es x con error absoluto menor que $Error$

Terminar.

5. ¿Qué significa la afirmación de que el método de Newton – Raphson posee una convergencia cuadrática? ¿Qué implica este hecho desde un punto de vista computacional?
6. Resuelva las ecuaciones del primer problema de este examen. Utilice para cada una un método numérico diferente. Obtenga la solución con 4 cifras decimales exactas.
7. Acerca de un cilindro circular se conoce que su volumen es de 500 cm^3 y su superficie total de 350 cm^2 . Determine las dimensiones del cilindro con error menor que $0,1 \text{ mm}$.
8. La ecuación $x^4 + 3x^3 - 3x^2 - 2x + 5 = 0$ posee una raíz imaginaria cerca de $1 + i$. Halle esta raíz mediante el método de Newton – Bairstow con cuatro cifras decimales exactas.

CAPÍTULO 3 **Matemática Numérica, 2da Edición**
Manuel Álvarez, Alfredo Guerra, Rogelio Lau
SISTEMAS DE ECUACIONES LINEALES Y MATRICES

Objetivos

Al finalizar el estudio y la ejercitación de este capítulo, el lector debe ser capaz de:

- Describir las características de los métodos directos y los métodos iterativos para la resolución de sistemas de ecuaciones lineales.
- Describir los conceptos de norma de un vector y de norma de una matriz y explicar las propiedades (axiomas) de las mismas.
- Describir el algoritmo de Gauss para resolver sistemas de ecuaciones lineales.
- Describir las diferentes estrategias de pivote que se utilizan en el método de Gauss y compararlas entre sí.
- Utilizar el método de Gauss manualmente o mediante un programa personal en algún lenguaje computacional que conozca.
- Analizar el volumen de operaciones que requerirá resolver un sistema de ecuaciones por el método de Gauss.
- Describir el algoritmo de Gauss especializado en sistemas tridiagonales y computar el orden de la cantidad de operaciones que se requerirá para su aplicación.
- Utilizar el método de Gauss especializado en sistemas tridiagonales, ya sea manualmente o mediante un programa computacional en algún leguaje que domine.
- Describir el algoritmo para calcular determinantes basado en el proceso directo del método de Gauss y utilizarlo para calcular determinantes, ya sea manualmente o mediante un programa que confeccione.
- Describir el algoritmo para invertir matrices basado en el método de Gauss y utilizarlo, ya sea manualmente o mediante un programa que confeccione.
- Describir el concepto de sistema lineal mal condicionado.
- Establecer el concepto de número de condición de una matriz y explicar por que éste mide el mal condicionamiento de una matriz.
- Calcular el número de condición de la matriz de un sistema lineal y decidir si se trata o no de un sistema mal condicionado.
- Describir los métodos de Jacobi y de Seidel para resolver sistemas lineales en forma iterativa y las condiciones bajo las cuales los mismos convergen.
- Resolver sistemas de ecuaciones lineales mediante los algoritmos de Jacobi y de Seidel manualmente o mediante programas computacionales escritos al efecto.
- Realizar un análisis comparativo entre los algoritmos iterativos de Jacobi y de Seidel.
- Decidir, basándose en la cantidad de operaciones necesarias, cual de los algoritmos estudiados es preferible para resolver un sistema de ecuaciones lineales.
- Describir los conceptos de valores y vectores propios de una matriz.
- Describir el concepto de polinomio característico y las dificultades de su obtención en problemas reales.
- Graficar grosso modo el polinomio característico de un matriz trazando una cantidad grande de puntos de la gráfica mediante un programa capaz de calcular determinantes eficientemente.
- Elaborar programas computacionales en algún lenguaje que usted domine, que permitan hallar valores propios resolviendo en forma numérica la ecuación característica de la matriz.

- Describir el método de la potencia para hallar el valor propio de mayor valor absoluto de una matriz y utilizarlo manualmente o mediante un programa computacional confeccionado al efecto.

3.1 Introducción

En la modelación de problemas reales surgen con frecuencia sistemas de ecuaciones lineales de orden apreciable. Los dos ejemplos que siguen muestran solamente dos situaciones concretas y pueden dar una idea de cuan grande puede ser el tamaño de estos sistemas.

Ejemplo 1

En la resolución de circuitos eléctricos, uno de los métodos empleados consiste en definir las llamadas corrientes de malla y plantear la ley de Kirchoff para voltajes en las diferentes mallas del circuito. En la figura 1 se muestra un circuito con cuatro mallas donde se han definido las corrientes i_1, i_2, i_3 e i_4 . Al plantear la ley de Kirchoff para cada malla se obtienen las ecuaciones:

$$\begin{aligned} R_1 i_1 + R_5(i_1 - i_4) &= E_1 - E_2 \\ R_2 i_2 + R_3(i_2 - i_3) &= E_2 \\ R_3(i_3 - i_2) + R_6 i_3 + R_4(i_3 - i_4) &= E_3 \\ R_5(i_4 - i_1) + R_4(i_4 - i_3) &= 0 \end{aligned}$$

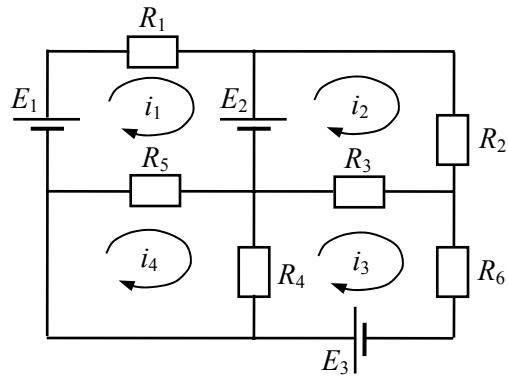


Figura 1

Si se conocen los valores de las resistencia y de las fuentes de voltaje, los valores de intensidad de corriente pueden ser obtenidos resolviendo el sistema anterior, que es un sistema de cuatro ecuaciones lineales con cuatro incógnitas i_1, i_2, i_3 e i_4 . Después de ordenado, el sistema toma la forma:

$$\begin{array}{lll} (R_1 + R_5)i_1 & & -R_5i_4 = E_1 - E_2 \\ (R_2 + R_3)i_2 & -R_3i_3 & = E_2 \\ -R_3i_2 + (R_3 + R_4 + R_6)i_3 & -R_4i_4 & = E_3 \\ -R_5i_1 & -R_4i_3 + (R_4 + R_5)i_4 & = 0 \end{array}$$

Es fácil imaginar que en un circuito real pueden aparecer decenas de mallas que dan lugar a sistemas con esa misma cantidad de ecuaciones y de incógnitas, los cuales no pueden ser resueltos por vía manual. ■

La importancia de los sistemas lineales no viene dada solamente porque ellos aparezcan directamente en la modelación matemática de infinidad de problemas del mundo real; además de esto, la solución de muchos problemas matemáticos conduce a resolver sistemas lineales; así sucede con el ajuste de curvas, en varias técnicas de interpolación, en la solución de ecuaciones diferenciales parciales y en muchos otros campos. El ejemplo que sigue muestra, en forma elemental, los sistemas lineales que surgen cuando se aplica la técnica de diferencias finitas en la solución de una ecuación diferencial parcial de tipo elíptico.

Ejemplo 2

Una lámina cuadrada de metal de un metro de lado es sometida en sus bordes a temperaturas fijas (ver figura 2) hasta que la temperatura alcanza valores estables en cada punto de la lámina. Se demuestra que si $u = u(x, y)$ es la temperatura en el punto (x, y) entonces, para los puntos de la lámina rige la ecuación diferencial:

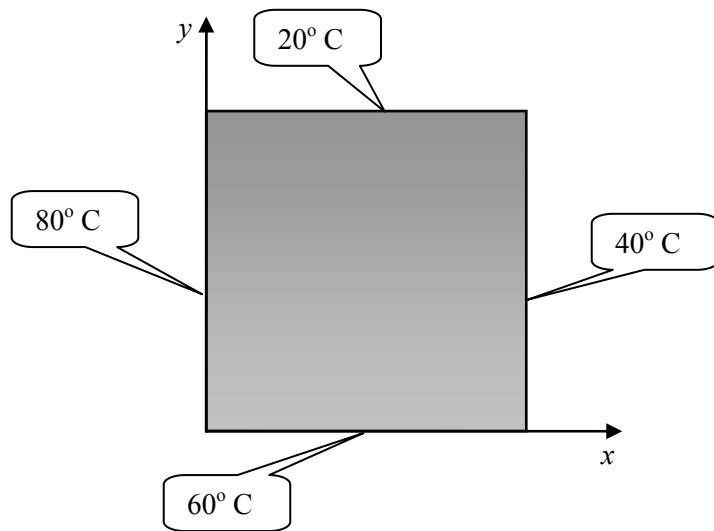


Figura 2

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

con las condiciones de frontera:

$$\begin{aligned} u(0, y) &= 80 & \text{para } 0 < y < 1 \\ u(1, y) &= 40 & \text{para } 0 < y < 1 \\ u(x, 0) &= 60 & \text{para } 0 < x < 1 \\ u(x, 1) &= 20 & \text{para } 0 < x < 1 \end{aligned}$$

Si se define sobre la lámina una red de puntos como la que muestra la figura 3 y se llama u_{ij} ($i = 0, 1, 2, 3, 4, 5$; $j = 0, 1, 2, 3, 4, 5$) a la temperatura en los puntos de la red, entonces la temperatura en cada uno de los puntos que no están en la frontera está relacionada aproximadamente con la temperatura de sus puntos vecinos mediante la ecuación lineal:

$$4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = 0 \quad (1)$$

donde: $i = 1, 2, 3, 4$; $j = 1, 2, 3, 4$

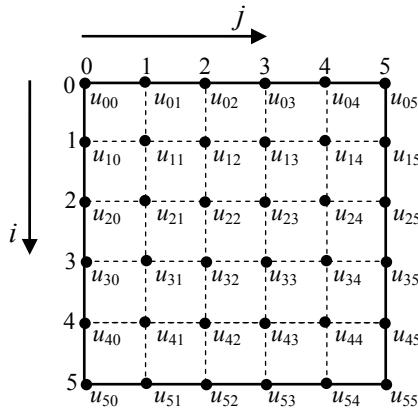


Figura 3

Nótese que la expresión (1) constituye realmente un sistema de 16 ecuaciones lineales con 16 incógnitas que corresponden a las temperaturas en los 16 puntos interiores, ya que las temperaturas de los puntos que están en el borde son conocidas por las condiciones de frontera. Resolviendo este sistema de 16 ecuaciones se obtienen las temperaturas deseadas.

La aproximación se hace mejor mientras más tupida sea la red de puntos, sin embargo, la cantidad de ecuaciones crece considerablemente. Por ejemplo, si en lugar de formar una red con distancias entre puntos de $h = 0,2$, se hubiera tomado $h = 0,01$, la exactitud mejoraría y la temperatura sería conocida casi en cualquier punto de la lámina; sin embargo, el sistema lineal que habría que resolver en ese caso sería de $(99)(99) = 9801$ ecuaciones con 9801 incógnitas.

Problemas a resolver

Relacionados con los sistemas lineales existen varios problemas numéricos importantes. El primero es la solución de grandes sistemas, formados por cientos y hasta decenas de miles de ecuaciones. Cuando el tamaño del problema crece en ese grado es fundamental analizar la eficiencia del método empleado, de otro modo la cantidad de operaciones aritméticas puede crecer a tal punto que ni aun en la máquina más eficiente sea factible su solución en un tiempo razonable. Por ejemplo, si se decide resolver un sistema lineal de 30 ecuaciones con 30 incógnitas por el método de Cramer, será necesario calcular 31 determinantes de orden 30; si para calcular cada uno de estos determinantes se utiliza el algoritmo recursivo conocido como desarrollo por menores, el cual transforma el determinante de orden 30 en una combinación lineal de 30 determinantes de orden 29, cada determinante de orden 29 en una combinación lineal de 29 determinantes de orden 28, etcétera, es obvio que cada determinante de orden 30 requerirá 30! operaciones de multiplicación. Como son 31 determinantes, se necesitarán $31 \cdot 30! = 31!$

multiplicaciones (sin contar las sumas). ¿Puede alguna máquina efectuar un algoritmo que requiera $31!$ operaciones? Note el lector que $31!$ es aproximadamente $8,22 \cdot 10^{33}$. Una máquina que efectúe 10^9 operaciones por segundo y que trabaje los 31 536 000 segundos que tiene un año será capaz de realizar unas $3,15 \cdot 10^{16}$ operaciones en el año. Si se contara con 10 000 millones de computadoras de este tipo (más de una computadora por cada habitante del planeta) y se pudiera tenerlas a todas trabajando en el mismo problema simultáneamente, en este descomunal sistema paralelo se podrían realizar $3,15 \cdot 10^{26}$ operaciones en un año. Se necesitaría entonces de

$$\frac{8,22 \cdot 10^{33}}{3,15 \cdot 10^{26}} = 2,61 \cdot 10^7 \text{ años}$$

o sea, unos 26 millones de años; claro, es de suponer que mucho antes, el incauto analista numérico de esta historia se habría percatado de que, con un método más razonable una computadora personal requiere menos de un segundo para resolver este problema.

Además de la solución de grandes sistemas lineales con algoritmos eficientes, en este capítulo se abordará el problema del cálculo de determinantes numéricos, operación que, aunque no es aconsejable como forma de resolver sistemas lineales, en ocasiones se requiere realizar con otros fines.

Otro problema menos frecuente pero que suele presentarse es la inversión de matrices. Aunque en los cursos de Álgebra Lineal se calculan las matrices inversas mediante el algoritmo usual de dividir la traspuesta de la matriz de cofactores entre el determinante de la matriz, este procedimiento, salvo para matrices de orden 3 ó 4, es extraordinariamente ineficiente. La inversión de matrices numéricas será abordada también en este capítulo.

El problema de hallar los valores y vectores propios de una matriz es uno de los más complejos de la matemática numérica y en ocasiones se necesita resolver, al menos parcialmente. Aquí solo se hará una introducción elemental a este tema y en la bibliografía recomendada se encuentran algunas referencias útiles.

Relacionados con los sistemas lineales se encuentran frecuentemente problemas inestables (que en este ámbito se suelen llamar, *mal condicionados*). Aunque el carácter elemental de este texto no permite un tratamiento profundo de este asunto, en la sección 3.3 se abordará esta cuestión.

Métodos directos y métodos iterativos

La mayor parte de los métodos utilizados para la solución de sistemas lineales están concentrados en dos grandes tipos: métodos directos o “exactos” y métodos iterativos o de aproximaciones sucesivas. La primera clase incluye a aquellos métodos en que, si se pudieran evitar todos los errores por redondeo, se obtendría la respuesta exacta en un número finito de pasos; muchos de los métodos elementales para resolver sistemas (Cramer, sustitución, suma y resta) y para calcular determinantes e invertir matrices pertenecen a esta categoría; aquí se estudiará el método de Gauss, que es también un método directo aplicable a varios de los problemas que se analizarán. Por otra parte, los métodos iterativos son todos aquellos que producen una sucesión infinita de respuestas aproximadas, sucesión que, bajo ciertas condiciones, converge hacia la solución exacta del problema; este tipo de método estará aquí representado por los métodos de Jacobi y de Seidel para resolver ciertos sistemas lineales y el método de la potencia para el cálculo de valores y

vectores propios. Más adelante, cuando se haya apreciado las características de cada uno de los métodos estudiados, se podrá valorar en qué situaciones es cada uno más conveniente.

Además de los métodos directos y los iterativos, se pueden mencionar los llamados métodos de Montecarlo, en los que, con la utilización de procesos aleatorios, se puede llegar a una solución aproximada, en general con muy poca exactitud, de sistemas lineales. La utilización de la teoría de probabilidades que ello requiere impide su tratamiento en un libro como este.

Normas matriciales y vectoriales

Se supone que el lector tiene conocimientos acerca de las operaciones con matrices y vectores al nivel que se estudia en los cursos de pregrado de Álgebra Lineal. No obstante, como en muchos de estos cursos se obvia el tema de las normas matriciales, aquí se han incluido los elementos que más adelante se requieren.

En general, si se tiene un espacio vectorial V , una norma en V es cualquier función $\|\cdot\|$ que a cada vector v de V le haga corresponder un número real denotado como $\|v\|$, y que posea las siguientes propiedades, que se llaman axiomas de norma:

1. Para todo v de V , $\|v\| \geq 0$
2. Para todo v de V y todo k real, $\|kv\| = |k| \cdot \|v\|$
3. Para todos u y v de V , $\|u + v\| \leq \|u\| + \|v\|$
4. Si $\|v\| = 0$ entonces v es el vector nulo de V

En los espacios de tipo R^n se definen normas de diversas formas. Probablemente el lector esté familiarizado con la norma cuadrática que se define como la raíz cuadrada de la suma de los cuadrados de las componentes del vector, que en algunos temas es muy importante pero que aquí no resulta la más conveniente. En todo este capítulo se utilizará para los vectores de R^n la siguiente norma.

Definición 1

Si x es un vector de R^n : $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, se define la norma de x como: $\|x\| = \max_i |x_i|$.

Se puede probar sin dificultad que esta definición satisface los cuatro axiomas de norma.

En el caso de las matrices cuadradas, ellas, como se sabe, constituyen un espacio vectorial y por tanto, su norma debe satisfacer los cuatro axiomas generales. No obstante, como en el espacio de las matrices cuadradas están definidas también las operaciones de producto entre matrices y producto de una matriz por un vector, se requiere de axiomas especiales que establezcan las propiedades que debe poseer la norma matricial respecto a estas operaciones. Así, para el espacio vectorial M_n de las matrices cuadradas de orden n , cualquier norma que se utilice deberá cumplir las siguientes 6 propiedades (axiomas de norma matricial):

1. Para toda matriz \mathbf{A} de M_n $\|\mathbf{A}\| \geq 0$
2. Para todo \mathbf{A} de M_n y todo k real, $\|k\mathbf{A}\| = |k| \cdot \|\mathbf{A}\|$
3. Para todos \mathbf{A} y \mathbf{B} de M_n , $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$
4. Si $\|\mathbf{A}\| = 0$ entonces \mathbf{A} es la matriz nula de M_n
5. Para todos \mathbf{A} y \mathbf{B} de M_n , $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$
6. Si $\|\cdot\|_v$ es la norma de un vector de R^n , entonces, para toda \mathbf{A} de M_n y todo \mathbf{x} de R^n ,

$$\|\mathbf{Ax}\|_v \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\|_v$$

Nótese que el axioma 6 significa que el tipo de norma utilizado para las matrices debe ser compatible con la norma utilizada en R^n . La siguiente definición de norma matricial satisface estos 6 axiomas si se toma la norma en R^n de acuerdo con la definición 1.

Definición 2

Si \mathbf{A} es una matriz cuadrada de orden n , $\mathbf{A} = [a_{ij}]$, se define la norma de \mathbf{A} como:

$$\|\mathbf{A}\| = \max_i \sum_{j=1}^n |a_{ij}|$$

Aunque se pueden utilizar otras combinaciones de normas matriciales y vectoriales, la que aquí se ha seleccionado es la que más se emplea en el tema que se tratará, debido a la simplicidad de los cálculos que se requieren.

Ejemplo 3

Dadas las matrices $\mathbf{A} = \begin{bmatrix} 3 & -1 & 2 \\ 0 & 5 & 1 \\ 8 & -3 & 2 \end{bmatrix}$ y $\mathbf{B} = \begin{bmatrix} 0 & 3 & 7 \\ 1 & -1 & 3 \\ 2 & 0 & -2 \end{bmatrix}$ y los vectores $\mathbf{x} = \begin{bmatrix} 2 \\ -1 \\ 4 \end{bmatrix}$ y $\mathbf{y} = \begin{bmatrix} 0 \\ 2 \\ -5 \end{bmatrix}$,

calcule: a) $\|\mathbf{A}\|$, b) $\|\mathbf{B}\|$, c) $\|\mathbf{A} + \mathbf{B}\|$, d) $\|\mathbf{AB}\|$, e) $\|\mathbf{x}\|$, f) $\|\mathbf{y}\|$, g) $\|\mathbf{x} + \mathbf{y}\|$, h) $\|\mathbf{Ax}\|$ y verifique que se satisfacen los axiomas correspondientes.

Solución:

a) $\|\mathbf{A}\| = \max\{6, 6, 13\} = 13$

b) $\|\mathbf{B}\| = \max\{10, 5, 4\} = 10$

c) $\mathbf{A} + \mathbf{B} = \begin{bmatrix} 3 & 2 & 9 \\ 1 & 4 & 4 \\ 10 & -3 & 0 \end{bmatrix}$, así que $\|\mathbf{A} + \mathbf{B}\| = \max\{14, 9, 13\} = 14$. Se cumple el axioma 3 de la

norma matricial, es decir: $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\| = 23$

d) $\mathbf{AB} = \begin{bmatrix} 3 & 10 & 14 \\ 7 & -5 & 13 \\ 1 & 27 & 43 \end{bmatrix}$, por tanto, $\|\mathbf{AB}\| = \max\{27, 25, 71\} = 71$. Se satisface el axioma 5 de la norma matricial, esto es, $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\| = 130$.

e) $\|\mathbf{x}\| = 4$

f) $\|\mathbf{y}\| = 5$

g) $\mathbf{x} + \mathbf{y} = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$, luego, $\|\mathbf{x} + \mathbf{y}\| = 2$. Se cumple que $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| = 9$, que es el axioma 3 de los vectores.

h) $\mathbf{Ax} = \begin{bmatrix} 15 \\ -1 \\ 27 \end{bmatrix}$, por lo tanto, $\|\mathbf{Ax}\| = 27$. Se satisface el axioma 6 de la norma matricial, esto es: $\|\mathbf{Ax}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\| = (13)(4) = 52$.

3.2 El método de Gauss

Aunque el método de Gauss puede ser aplicado a sistemas lineales de cualquier orden, aquí solo interesan los sistemas cuadrados, es decir, con igual número de ecuaciones e incógnitas. Se supone además que se trata de sistemas determinados, es decir, con solución única. Sea entonces el sistema a resolver:

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

donde todos los coeficientes a_{ij} son números reales conocidos, al igual que los términos independientes. Este sistema, como se sabe, se puede representar en forma matricial de la forma:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{2}$$

Tomando el convenio de representar la matriz de los coeficientes por \mathbf{A} , el vector de incógnitas por \mathbf{x} y el vector de términos independientes como \mathbf{b} , la ecuación matricial (2) se reduce a:

$$\mathbf{Ax} = \mathbf{b} \tag{3}$$

$$\text{donde } \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ y } \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Como se supone que el sistema es determinado, eso significa que \mathbf{A} es no singular, es decir, que su determinante no es cero.

El algoritmo de Gauss consta de dos etapas bien diferenciadas que se llaman *proceso directo* y *proceso inverso*.

Proceso directo

El proceso directo consiste en realizar sobre el sistema de ecuaciones transformaciones elementales de dos tipos:

- I. Intercambiar dos ecuaciones del sistema.
- II. Sumar a una ecuación, miembro a miembro, otra ecuación (pivot) del sistema multiplicada por un número real cualquiera.

Es claro que la transformación elemental tipo I no cambia la solución del sistema, aunque si afecta el determinante de \mathbf{A} , ya que equivale a la permutación de dos de sus filas, lo cual cambia el signo del determinante. La transformación elemental de tipo II no altera ni la solución del sistema ni el valor del determinante. Estas transformaciones elementales se realizan con el objetivo de transformar el sistema a la forma triangular:

$$\begin{aligned} c_{11}x_1 + c_{12}x_2 + \cdots + c_{1n}x_n &= d_1 \\ c_{22}x_2 + \cdots + c_{2n}x_n &= d_2 \\ &\vdots \\ c_{nn}x_n &= d_n \end{aligned} \tag{4}$$

Cuando se trabaja en forma manual (y con más razón cuando se realiza un algoritmo computacional) estas transformaciones elementales se efectúan sobre los coeficientes de las ecuaciones, los cuales se escriben en forma de una matriz de orden n por $n+1$ que agrupa los elementos de \mathbf{A} y los de \mathbf{b} y que suele llamarse *matriz ampliada* del sistema:

$$\mathbf{A}|\mathbf{b} = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array} \right] \tag{5}$$

la cual, efectuado el proceso directo, toma la forma escalonada:

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} & d_1 \\ 0 & c_{22} & \cdots & c_{2n} & d_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & c_{nn} & d_n \end{bmatrix} \quad (6)$$

Para pasar de la matriz inicial a su forma escalonada, primero se efectúan $n - 1$ transformaciones elementales de tipo II, utilizando la fila 1 como pivote y afectando a las filas debajo de ella, de modo que se anulen todos los elementos de la primera columna que están debajo de la diagonal, después $n - 2$ transformaciones tipo II con la fila 2 como pivote, para modificar a las filas debajo de ella de manera que se anulen todos los elementos de la segunda columna debajo de la diagonal; el proceso continua hasta llegar a la fila $n - 1$ como pivote para eliminar al coeficiente de la columna $n - 1$ debajo de la diagonal.

En el paso número i del proceso directo, en el cual la fila i -sima actúa como pivote, la fila k ($k > i$) se cambia de acuerdo con la fórmula:

$$\text{fila } k := \text{fila } k - \frac{a'_{ki}}{a'_{ii}} (\text{fila } i) \quad k = i + 1, i + 2, \dots, n \quad (7)$$

donde los coeficientes se han afectado con un apóstrofe para indicar que no son necesariamente los coeficientes originales. Nótese que puede suceder que el coeficiente a'_{ii} de la fila pivote tome valor cero o muy pequeño. En ese caso se procede a permutar la fila i por alguna de las inferiores, de acuerdo con algunas estrategias disponibles y que se analizarán más adelante.

Proceso inverso

Una vez obtenido el sistema triangular (4) o su forma compacta (6) se puede calcular x_n de la última ecuación, después x_{n-1} a partir de la penúltima ecuación y así sucesivamente hasta encontrar x_1 de la primera ecuación. La forma de obtener la solución, de abajo hacia arriba, le da nombre a esta parte del algoritmo. Resulta:

$$\begin{aligned} x_n &= \frac{1}{c_{nn}} d_n \\ x_{n-1} &= \frac{1}{c_{n-1,n-1}} (d_{n-1} - c_{n-1,n} x_n) \\ &\vdots \\ x_1 &= \frac{1}{c_{11}} (d_1 - c_{12} x_2 - c_{13} x_3 - \cdots - c_{1n} x_n) \end{aligned}$$

Estrategias de pivote

La estrategia de pivote es el criterio que se utiliza para seleccionar la fila que se ha de utilizar como pivote en cada paso del proceso directo. Se utilizan en la práctica tres estrategias llamadas *elemental*, *parcial* y *total*. En la estrategia *elemental*, cuando se va a realizar el paso número i del proceso directo, se utiliza la fila i -sima como pivote a menos que el elemento de la diagonal a'_{ii} sea cero o menor que un número ϵ muy pequeño; en ese caso, se procede a analizar las filas por debajo de la i -sima, se escoge la primera que contenga el elemento de la i -sima posición distinto de cero (o modularmente mayor que ϵ) y esta fila se intercambia con la fila i ; si todas las filas tuvieran nulo su elemento i , entonces la matriz original sería singular, contrario a lo que se

ha supuesto. En la estrategia parcial de pivote, al realizar el paso i del proceso directo, se analizan todas las filas desde la i -sima hasta la última para seleccionar aquella que posea el elemento en posición i con el mayor valor absoluto (no todos pueden ser cero) y esta fila se intercambia con la i -sima fila. En la estrategia total, en el paso número i del proceso directo se busca el elemento de mayor valor absoluto de la submatriz que aun no se encuentra escalonada y se intercambian dos filas y dos columnas de manera que ese elemento pase a ocupar la posición de a'_{ii} .

La estrategia elemental es muy fácil de implementar pero se corre el riesgo de que el elemento a'_{ii} que forma el denominador de la expresión (7) sea muy pequeño. En ese caso la fila i -sima puede ser multiplicada por un número muy grande y esto produce una ampliación de los errores absolutos que contiene esta fila debido a los redondeos y su propagación hasta ese momento. En sistemas un poco grandes esta acumulación y amplificación de errores puede llevar a resultados desastrosos.

La estrategia parcial es también fácil de implementar y garantiza que el factor que se utiliza para multiplicar a la fila pivote sea siempre menor que 1, con lo cual los errores acumulados en la fila i -sima, lejos de ampliarse, se disminuyen.

La estrategia total da aun mejores resultados en cuanto a la propagación de los errores de redondeo pero es mucho más difícil de implementar debido al intercambio de columnas; nótese que intercambiar dos filas no trae afectaciones a la solución del sistema, pero cuando se intercambian dos columnas esto significa que las variables correspondientes han cambiado su posición y se necesita mantener un registro de todos los intercambios de este tipo que se hayan efectuado a la hora de realizar el proceso inverso. Por esta razón, la estrategia total no se incluirá en el algoritmo en seudo código para el método de Gauss.

Ejemplo 1

Resuelva el siguiente sistema lineal utilizando a) estrategia elemental de pivote y b) estrategia parcial de pivote. En ambos casos conserve solamente tres cifras exactas en los resultados.

$$\begin{aligned} 3x_1 + 3x_2 - x_3 - x_4 &= 4 \\ x_1 - x_2 + 3x_3 + x_4 &= 2 \\ -2x_1 - x_2 + x_3 + 3x_4 &= 4 \\ 2x_1 - 2x_2 - x_3 - 2x_4 &= -1 \end{aligned}$$

Solución:

Formando la matriz ampliada del sistema:

$$\left[\begin{array}{ccccc|c} 3 & 3 & -1 & -1 & 4 \\ 1 & -1 & 3 & 1 & 2 \\ -2 & -1 & 1 & 3 & 4 \\ 2 & -2 & -1 & -2 & -1 \end{array} \right]$$

a) Utilizando estrategia elemental de pivote

Primer paso del proceso directo:

Como el elemento 1 de la fila 1 es $3.00 \neq 0$, se conserva esta fila como pivote.

fila 2 := fila 2 - (0,333) fila 1

fila 3 := fila 3 + (0,667) fila 1

fila 4 := fila 4 - (0,667) fila 1:

$$\begin{bmatrix} 3.00 & 3.00 & -1.00 & -1.00 & 4.00 \\ -2.00 & 3.33 & 1.33 & 0.668 \\ 1.00 & 0.333 & 2.33 & 6.67 \\ -4.00 & -0.333 & -1.33 & -3.67 \end{bmatrix}$$

Segundo paso del proceso directo:

Como el segundo elemento de la fila 2 es $-2.00 \neq 0$, se conserva esta fila pivote.

fila 3 := fila 3 + (0,500) fila 2

fila 4 := fila 4 - (2,00) fila 2:

$$\begin{bmatrix} 3.00 & 3.00 & -1.00 & -1.00 & 4.00 \\ -2.00 & 3.33 & 1.33 & 0.668 \\ 2.00 & 3.00 & 7.00 \\ -6.99 & -3.99 & -5.01 \end{bmatrix}$$

Tercer paso del proceso directo:

Como el tercer elemento de la fila 3 es $2.00 \neq 0$, se conserva esta fila pivote.

fila 4 := fila 4 + (3,50) fila 3:

$$\begin{bmatrix} 3.00 & 3.00 & -1.00 & -1.00 & 4.00 \\ -2.00 & 3.33 & 1.33 & 0.668 \\ 2.00 & 3.00 & 7.00 \\ 6.51 & 19.5 \end{bmatrix}$$

Resulta entonces el sistema escalonado:

$$\begin{aligned} 3.00x_1 + 3.00x_2 - 1.00x_3 - 1.00x_4 &= 4.00 \\ -2.00x_2 + 3.33x_3 + 1.33x_4 &= 0.668 \\ 2.00x_3 + 3.00x_4 &= 7.00 \\ 6.51x_4 &= 19.5 \end{aligned}$$

Proceso inverso:

$$x_4 = \frac{19.5}{6.51} = 3.00$$

$$x_3 = \frac{7.00 - (3.00)(3.00)}{2.00} = -1.00$$

$$x_2 = \frac{0.668 - (3.33)(-1.00) - (1.33)(3.00)}{-2.00} = -0.004$$

$$x_1 = \frac{4.00 - (3.00)(-0.004) + (1.00)(-1.00) + (1.00)(3.00)}{3.00} = 2.00$$

b) Utilizando estrategia parcial de pivote

$$\begin{bmatrix} 3 & 3 & -1 & -1 & 4 \\ 1 & -1 & 3 & 1 & 2 \\ -2 & -1 & 1 & 3 & 4 \\ 2 & -2 & -1 & -2 & -1 \end{bmatrix}$$

Primer paso del proceso directo:

Como el elemento de mayor valor absoluto esta en la fila 1, se conserva esta fila pivote.

fila 2 := fila 2 - (0,333) fila 1

fila 3 := fila 3 + (0,667) fila 1

fila 4 := fila 4 - (0,667) fila 1:

$$\begin{bmatrix} 3.00 & 3.00 & -1.00 & -1.00 & 4.00 \\ -2.00 & 3.33 & 1.33 & 0.668 & \\ 1.00 & 0.333 & 2.33 & 6.67 & \\ -4.00 & -0.333 & -1.33 & -3.67 & \end{bmatrix}$$

Segundo paso del proceso directo:

Como el elemento de mayor valor absoluto esta en la fila 4, se intercambian las filas 2 y 4:

$$\begin{bmatrix} 3.00 & 3.00 & -1.00 & -1.00 & 4.00 \\ -4.00 & -0.333 & -1.33 & -3.67 & \\ 1.00 & 0.333 & 2.33 & 6.67 & \\ -2.00 & 3.33 & 1.33 & 0.668 & \end{bmatrix}$$

fila 3 := fila 3 + (0,250) fila 2:

fila 4 := fila 4 - (0,500) fila 2:

$$\begin{bmatrix} 3.00 & 3.00 & -1.00 & -1.00 & 4.00 \\ -4.00 & -0.333 & -1.33 & -3.67 & \\ 0.250 & 2.00 & 5.75 & & \\ 3.50 & 2.00 & 2.50 & & \end{bmatrix}$$

Tercer paso del proceso directo:

Como el elemento de mayor valor absoluto esta en la fila 4, se intercambian las filas 3 y 4:

$$\left[\begin{array}{ccccc} 3.00 & 3.00 & -1.00 & -1.00 & 4.00 \\ -4.00 & -0.333 & -1.33 & -3.67 & \\ 3.50 & 2.00 & 2.50 & & \\ 0.250 & 2.00 & 5.75 & & \end{array} \right]$$

fila 4 := fila 4 – (0,0714) fila 3:

$$\left[\begin{array}{ccccc} 3.00 & 3.00 & -1.00 & -1.00 & 4.00 \\ -4.00 & -0.333 & -1.33 & -3.67 & \\ 3.50 & 2.00 & 2.50 & & \\ 1.86 & & 5.57 & & \end{array} \right]$$

Resulta entonces el sistema escalonado:

$$\begin{aligned} 3.00x_1 + 3.00x_2 - 1.00x_3 - 1.00x_4 &= 4.00 \\ -4.00x_2 - 0.333x_3 - 1.33x_4 &= -3.67 \\ 3.50x_3 + 2.00x_4 &= 2.50 \\ 1.86x_4 &= 5.57 \end{aligned}$$

Proceso inverso: $x_4 = \frac{5.57}{1.86} = 2.99$

$$x_3 = \frac{2.50 - (2.00)(2.99)}{3.50} = -0.994$$

$$x_2 = \frac{-3.67 + (0.333)(-0.994) + (1.33)(2.99)}{-4.00} = 0.00608$$

$$x_1 = \frac{4.00 - (3.00)(0.00608) + (1.00)(-0.994) + (1.00)(2.99)}{3.00} = 1.99$$

Como en este ejemplo se trata de un sistema pequeño (cuarto orden) no hay una diferencia apreciable entre ambas estrategias de pivote y en ambos casos se obtienen pequeños errores debido a que solo se ha trabajado con tres cifras exactas.

Algoritmo en seudo código

Se trata de resolver el sistema de n ecuaciones lineales con n incógnitas $\mathbf{Ax} = \mathbf{b}$ donde se supone que \mathbf{A} es no singular. Se consideran datos del problema el número n y los coeficiente a_{ij} ($i = 1, 2, \dots, n; j = 1, 2, \dots, n$) de \mathbf{A} y b_i ($i = 1, 2, \dots, n$) de \mathbf{b} . El algoritmo se ha acompañado de algunos comentarios entre llaves, para darle más claridad. La orden “Seleccionar la fila pivote i – sima” se especificará después.

$\mathbf{C} := [\mathbf{A}|\mathbf{b}]$ {Se forma la matriz ampliada \mathbf{C} , de n filas y $n+1$ columnas}
 {Aquí comienza el proceso directo}
for $i = 1$ **to** $n-1$
 Seleccionar la fila pivote i -sima {Depende de la estrategia de pivote usada}
 for $k = i+1$ **to** n
 $m := \frac{c_{ki}}{c_{ii}}$
 $fila(k) := fila(k) - m fila(i)$
 end
 end
 {Aquí comienza el proceso inverso}
 $i := n$
repeat
 $x_i := c_{i,n+1}$
 for $j = i+1$ **to** n {Cuando $i = n$ este lazo no se ejecuta}
 $x_i := x_i - c_{ij}x_j$
 end
 $x_i := \frac{x_i}{c_{ii}}$
 $i := i - 1$
until $i = 0$
 La solución es $\mathbf{x} = [x_1 \quad x_2 \quad \cdots \quad x_n]^T$ {El supraíndice T indica traspuesta}
 Terminar

En el caso de usar la estrategia elemental de pivote, el algoritmo “Seleccionar la fila pivote i -sima” sería:

```

if  $c_{ii} = 0$  then
     $k := i$ 
    repeat
         $k := k + 1$ 
    until  $c_{ki} \neq 0$  or  $k > n$  {Si se llega a  $k > n$  y todos los  $c_{ki}$  han sido cero es porque la matriz del sistema es singular}
    if  $k > n$  then
        El sistema no tiene solución única pues la matriz es singular
        Terminar
    else
        Intercambiar  $fila(k)$  con  $fila(i)$ 
    end
end
  
```

En el caso de usar la estrategia parcial de pivote, el algoritmo “Seleccionar la fila pivote i -sima” sería:

$max := c_{ii} $	{La variable max guarda el mayor coeficiente obtenido hasta el momento}
$Fila\ de\ max := i$	{La variable $Fila\ de\ max$ guarda el número de la fila donde ocurrió el máximo}

```

for  $k = i + 1$  to  $n$ 
    if  $|c_{ki}| > max$  then
         $max := |c_{ki}|$ 
         $Fila\ de\ max := k$ 
    end
end
if  $k > i$  then {Si  $k = i$  el intercambio no tiene sentido}
    Intercambiar  $fila(k)$  con  $fila(i)$ 
end

```

Cantidad de operaciones del método de Gauss

Es muy importante conocer de que orden es el número de operaciones aritméticas que se necesita para ejecutar el algoritmo de Gauss para un sistema de orden n , ya que el interés principal ahora es obtener métodos eficientes para resolver sistemas con un número alto de ecuaciones. Primero se analizará el proceso directo y después el inverso.

Proceso directo:

Con vistas a simplificar y como el intercambio de filas no siempre se produce y solo sería necesario a lo más en $n - 1$ ocasiones, no se tomará en cuenta esta operación. Con el mismo objetivo, solo se contarán las multiplicaciones y divisiones, no las sumas, las cuales consumen mucho menos tiempo en la máquina.

En el paso 1 se realizan $n - 1$ transformaciones elementales de tipo II que requieren cada una un cociente (para hallar m) y n productos (recuérdese que las filas tienen $n + 1$ elementos). Un total de $(n - 1)(n + 1)$ multiplicaciones y divisiones.

En el paso 2 se realizan $n - 2$ transformaciones elementales de tipo II que requieren cada una un cociente y $n - 1$ productos. Un total de $(n - 2)(n)$ multiplicaciones y divisiones.

En general, en el paso i ($i = 1, 2, \dots, n - 1$) del proceso directo se necesitan $(n - i)(n - i + 2)$ multiplicaciones y divisiones.

Entonces, en todo el proceso directo la cantidad de operaciones de multiplicar y dividir será:

$$\sum_{i=1}^{n-1} (n - i)(n - i + 2)$$

En los libros de Álgebra Superior se estudian formas de calcular este tipo de sumas. Aquí no se entrará en los detalles. El resultado de la suma es:

$$\frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n \quad (8)$$

Proceso inverso:

Para hallar la variable x_n se requiere una división y ninguna multiplicación.

Para hallar el valor de x_{n-1} hacen falta una multiplicación y una división.

Para hallar el valor de x_{n-2} hacen falta dos productos y una división.

En general, cada variable requiere de una división y la cantidad de productos aumenta desde 0 hasta $n - 1$ que se necesitan para calcular x_1 . El número total de multiplicaciones y divisiones del proceso inverso será de:

$$\sum_{i=1}^n i = 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$$

Es decir:

$$\frac{1}{2}n^2 + \frac{1}{2}n \quad (9)$$

Si se comparan las expresiones (8) y (9) se comprenderá que el proceso inverso requiere muy pocas operaciones en comparación con el proceso directo. Por ejemplo, un sistema de 30 ecuaciones requerirá, para el proceso directo, de:

$$\frac{1}{3}(30)^3 + \frac{1}{2}(30)^2 - \frac{5}{6}(30) = 9000 + 450 - 25 = 9425 \text{ productos y cocientes}$$

Mientras el proceso inverso necesitará:

$$\frac{1}{2}(30)^2 + \frac{1}{2}(30) = 450 + 15 = 465 \text{ productos y cocientes}$$

El total de operaciones es de unas 10000 y se realiza en una computadora personal en fracciones de segundo. Compárese con la astronómica cifra de $8,22 \cdot 10^{33}$ obtenida al principio de este capítulo mediante un procedimiento ineficiente.

La expresión para el número total de multiplicaciones y divisiones será la suma de (8) y (9):

$$\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$$

Como usualmente no se necesita sino una idea del orden de este número y no su valor exacto, se suele utilizar solamente su sumando más significativo, que es el de exponente 3. Se concluye que:

La cantidad de multiplicaciones y divisiones que requiere el método de Gauss para resolver un sistema de n ecuaciones con n incógnitas es del orden de

$$\frac{1}{3}n^3$$

Aunque aquí no se hayan tenido en cuenta las sumas, su número es prácticamente el mismo que el de multiplicaciones, así que no aumenta significativamente cualquier estimado de tiempo que se haga con la fórmula obtenida.

Ejemplo 2

Haga un estimado del tiempo que requerirá una computadora capaz de realizar 50 000 multiplicaciones o divisiones por segundo para resolver un sistema lineal de 1000 ecuaciones con 1000 incógnitas utilizando el método de Gauss.

Solución:

La cantidad de productos y cocientes es del orden de $\frac{1}{3}n^3 = \frac{1}{3}(1000)^3 = 333\,000\,000$

El tiempo necesario será aproximadamente: $\frac{333\,000\,000}{50\,000} = 6600$ segundos, es decir, aproximadamente 2 horas.

Ejercicios

En los siguientes ejercicios, utilice un programa computacional del método de Gauss, preferiblemente confeccionado por usted. Si no cuenta con un programa adecuado, realice los cálculos mediante una calculadora trabajando con 5 cifras exactas.

- Resuelva los siguientes sistemas lineales utilizando el método de Gauss con estrategia parcial de pivote.

$2,6x - 2,7y + 5,2z = 7,25$ a) $3,1x + 0,5y - z = 4,33$ $2,9x - 5,2y - 4,3z = -5,6$ $\frac{2}{5}x - \frac{1}{3}y + \frac{2}{7}z = \frac{1}{9}$ c) $\frac{1}{2}x + \frac{4}{5}y = \frac{3}{11}z$ $\frac{2}{3}x - \frac{3}{4}y + \frac{4}{5}z = \frac{5}{6}$	$3x - 45 = u - v + z$ b) $z = x + u$ $4u + v = 18 - z$ $5 = x + z + u + v$ $e^x + 3y - \tan z = 22$ d) $-2y = 2e^x + 3 \tan z$ $\frac{1}{3}e^x = 5 - y$ $e^{2x}e^{3y}e^{-4z} - 3e^xe^{2y}e^z + 4e^{4x}e^ye^{-z} = -2,5$ e) $2e^xe^{2y}e^z - e^{2x}e^{3y}e^{-4z} - 5e^{4x}e^ye^{-z} = -1,2$ $e^{4x}e^ye^{-z} + 2e^{2x}e^{3y}e^{-4z} + 3e^xe^ze^{2y} = 13,8$
---	---

- Se sabe que el polinomio $p(x) = ax^3 + bx^2 + cx + d$ satisface las condiciones:

$$\begin{aligned}
 p(1,3) &= 2,8 \\
 p(2,1) &= 3,2 \\
 p(2,7) &= 6 \\
 p(3,1) &= 0
 \end{aligned}$$

Halle a, b, c y d .

3. Halle el punto de intersección de los planos: $2x + y - z = 7$; $2x + y + z = 8$; $x - 3y + z = -2$.

4. Halle el punto del espacio donde la recta:

$$\frac{x-2}{3} = \frac{y-1}{4} = z+2$$

se encuentra con el plano cuyos interceptos con los ejes coordenados son $x = 1$, $y = 2$, $z = 3$.

5. La función $f(x, y, z) = kx^A y^B z^C$ satisface las condiciones:

x	y	z	$f(x, y, z)$
2	3	1	1,38
1	2	2	1,25
3	1	1	1,05
1	4	3	4,3

Determine los valores de los parámetros k , A , B y C .

6. Halle la ecuación del plano vertical que pasa por el origen de coordenadas y por el punto donde se intersecan las rectas

$$\frac{x-2}{2} = \frac{y+1}{-3} = \frac{z-2}{4} \quad \text{y} \quad \frac{x-3}{1} = \frac{y+2}{-2} = \frac{z-1}{5}$$

7. Halle una curva paramétrica del tipo: $x = a_1 t^2 + b_1 t + c_1$; $y = a_2 t^2 + b_2 t + c_2$ que pase por los tres puntos $P_1(3, 4)$, $P_2(1, 2)$, $P_3(1, 3)$ en ese mismo orden.
8. Elabore un algoritmo en seudo código que permita hallar y dibujar curvas paramétricas del tipo anterior para tres puntos ordenados cualesquiera del plano.
9. Elabore un algoritmo en seudo código que permita hallar la ecuación de la circunferencia que determinen tres puntos no colineales cualesquiera del plano.
10. El esquema de la figura 1 representa un sistema de cañerías de dos entradas con gastos de $Q_1 = 200$ litros por minuto y $Q_2 = 300$ litros por minuto y una salida con gasto Q_3 litros por minuto. Determine el gasto que circula por cada tramo de tubería si se sabe que, debido al diámetro y longitud de los tubos, x_2 es tres veces menor que x_1 y x_3 es dos veces mayor que x_4 .

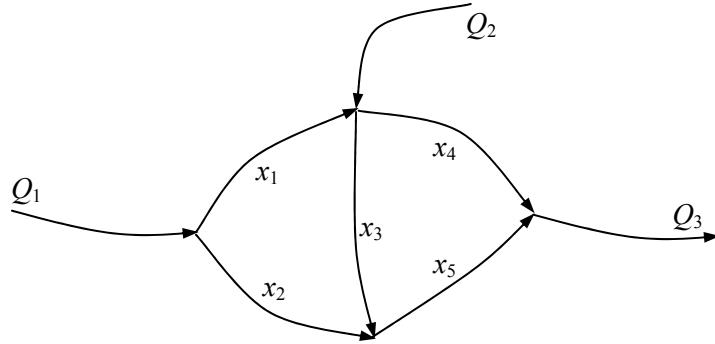


Figura 1

3.3 Consecuencias del método de Gauss

El método de Gauss puede ser adaptado para resolver otros problemas especiales. Aquí se analizarán tres de ellos: la solución de sistemas tridiagonales, el cálculo de determinantes y la inversión de matrices.

El método de Gauss para sistemas tridiagonales

En muchos problemas importantes aparecen sistemas de ecuaciones muy grandes con la característica especial de que la matriz del sistema es tridiagonal, esto es, solo la diagonal principal y las dos adyacentes, no son nulas. Para sistemas de este tipo se puede obtener un algoritmo de Gauss especializado sumamente eficiente. Ante todo, es obvio que en lugar de guardar todos los elementos de la matriz del sistema (la mayoría de los cuales son ceros) es preferible guardar solamente las tres diagonales no nulas y el vector de términos independientes. El sistema, por tanto se expresa matricialmente como:

$$\begin{bmatrix} a_1 & c_1 & & & \\ b_2 & a_2 & c_2 & & \\ b_3 & a_3 & c_3 & & \\ b_4 & \ddots & \ddots & & \\ & \ddots & a_{n-1} & c_{n-1} & \\ & & b_n & a_n & \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix}$$

donde se ha tomado el convenio de no escribir los términos (todos ellos ceros) fuera de las tres diagonales. Para aplicar el método de Gauss al sistema, primero debe construirse la matriz ampliada:

$$\begin{bmatrix} a_1 & c_1 & & & d_1 \\ b_2 & a_2 & c_2 & & d_2 \\ b_3 & a_3 & c_3 & & d_3 \\ b_4 & \ddots & \ddots & & \vdots \\ & \ddots & a_{n-1} & c_{n-1} & d_{n-1} \\ & & b_n & a_n & d_n \end{bmatrix} \quad (1)$$

En el desarrollo que sigue se supondrá que los términos que se utilizan como denominadores no son ceros. Posteriormente se darán condiciones que garantizan que ello se cumpla. Dividiendo por a_1 toda la primera fila queda:

$$\left[\begin{array}{ccc|c} 1 & p_1 & & q_1 \\ b_2 & a_2 & c_2 & d_2 \\ b_3 & a_3 & c_3 & d_3 \\ b_4 & \ddots & \ddots & \vdots \\ \vdots & a_{n-1} & c_{n-1} & d_{n-1} \\ b_n & a_n & d_n & \end{array} \right]$$

donde $p_1 = \frac{c_1}{a_1}$ y $q_1 = \frac{d_1}{a_1}$

Ahora se multiplicará la primera fila por $-b_2$ para sumársela a la segunda fila, de modo que se anule el segundo elemento de la columna 1 (que es único de esa columna por debajo de la diagonal que no era ya cero). La matriz quedará entonces:

$$\left[\begin{array}{ccc|c} 1 & p_1 & & q_1 \\ 0 & r_2 & c_2 & d_2 - b_2 q_1 \\ b_3 & a_3 & c_3 & d_3 \\ b_4 & \ddots & \ddots & \vdots \\ \vdots & a_{n-1} & c_{n-1} & d_{n-1} \\ b_n & a_n & d_n & \end{array} \right]$$

donde: $r_2 = a_2 - b_2 p_1$

Ahora se divide toda la segunda fila por r_2 con el objetivo de hacer 1 el elemento de la diagonal principal. La matriz ampliada queda en la forma:

$$\left[\begin{array}{ccc|c} 1 & p_1 & & q_1 \\ 0 & 1 & p_2 & q_2 \\ b_3 & a_3 & c_3 & d_3 \\ b_4 & \ddots & \ddots & \vdots \\ \vdots & a_{n-1} & c_{n-1} & d_{n-1} \\ b_n & a_n & d_n & \end{array} \right]$$

donde: $p_2 = \frac{c_2}{r_2}$ y $q_2 = \frac{d_2 - b_2 q_1}{r_2}$

Si ahora se usa la segunda fila como pivote para colocar un cero en lugar de b_3 y se definen r_3, p_3 y q_3 de manera análoga al paso anterior, la matriz se transforma en:

$$\begin{bmatrix} 1 & p_1 & & q_1 \\ 0 & 1 & p_2 & q_2 \\ 0 & 1 & p_3 & q_3 \\ b_4 & \ddots & \ddots & \vdots \\ & \ddots & a_{n-1} & c_{n-1} d_{n-1} \\ & & b_n & a_n d_n \end{bmatrix}$$

$$\text{donde: } r_3 = a_3 - b_3 p_2 \quad p_3 = \frac{c_3}{r_3} \quad \text{y} \quad q_3 = \frac{d_3 - b_3 q_2}{r_3}$$

Procediendo de esta manera, la matriz ampliada se transforma en una matriz equivalente con todos los elementos nulos por debajo de la diagonal:

$$\begin{bmatrix} 1 & p_1 & & q_1 \\ & 1 & p_2 & q_2 \\ & & 1 & p_3 & q_3 \\ & & \ddots & \ddots & \vdots \\ & & & 1 & p_{n-1} & q_{n-1} \\ & & & & 1 & q_n \end{bmatrix}$$

donde $p_1 = \frac{c_1}{a_1}$; $q_1 = \frac{d_1}{a_1}$ (2)

$$y \quad r_i = a_i - b_i p_{i-1} \quad p_i = \frac{c_i}{r_i} \quad y \quad q_i = \frac{d_i - b_i q_{i-1}}{r_i} \quad \text{para } i = 2, 3, \dots, n \quad (3)$$

El sistema de ecuaciones representado por esta matriz tiene la forma:

$$\left[\begin{array}{cc|c} 1 & p_1 & x_1 \\ 1 & p_2 & x_2 \\ 1 & p_3 & x_3 \\ \ddots & \ddots & \vdots \\ 1 & p_{n-1} & x_{n-1} \\ 1 & p_n & x_n \end{array} \right] = \left[\begin{array}{c} q_1 \\ q_2 \\ q_3 \\ \vdots \\ q_{n-1} \\ q_n \end{array} \right]$$

Esto es:

$$\left\{ \begin{array}{l} x_1 + p_1 x_2 = q_1 \\ x_2 + p_2 x_3 = q_2 \\ x_3 + p_3 x_4 = q_3 \\ \vdots \\ x_{n-1} + p_{n-1} x_n = q_{n-1} \\ x_n = q_n \end{array} \right.$$

El proceso inverso se simplifica notablemente por el hecho de que en cada ecuación solamente aparecen dos incógnitas, salvo en la última que permite encontrar x_n . Despejando sucesivamente las incógnitas de abajo hacia arriba, se obtiene:

$$\left\{ \begin{array}{l} x_n = q_n \\ x_{n-1} = q_{n-1} - p_{n-1} x_n \\ x_{n-2} = q_{n-2} - p_{n-2} x_{n-1} \\ \vdots \\ x_2 = q_2 - p_2 x_3 \\ x_1 = q_1 - p_1 x_2 \end{array} \right. \quad (4)$$

Obsérvese que para resolver un sistema tridiagonal se requiere de muy pocas operaciones aritméticas:

Ecuaciones (2) y (3)	6 operaciones n veces = $6n$ operaciones
Ecuaciones (4)	2 operaciones n veces = $2n$ operaciones

Total:	8n operaciones
--------	----------------

La cantidad de operaciones que requiere el método de Gauss especializado en sistemas tridiagonales para resolver un sistema de n ecuaciones con n incógnitas es del orden de

$$8n$$

Algoritmo en seudo código para sistemas tridiagonales

Se trata de resolver un sistema tridiagonal para el cual se conocen las diagonales:

- $\{a_1, a_2, \dots, a_n\}$ (diagonal principal)
- $\{b_2, b_3, \dots, b_n\}$ (diagonal debajo de la principal)
- $\{c_1, c_2, \dots, c_{n-1}\}$ (diagonal encima de la principal)

y los términos independientes $\{d_1, d_2, \dots, d_n\}$.

Puede demostrarse que si la diagonal principal es predominante, es decir si en cada fila i el valor absoluto de a_i es mayor que la suma de los valores absolutos de b_i y c_i , entonces, ninguno de los denominadores que aparecen en el algoritmo que sigue se hacen cero.

```

 $p_1 := \frac{c_1}{a_1}$ 
 $q_1 := \frac{d_1}{a_1}$ 
for  $i = 2$  to  $n$ 
     $r_i = a_i - b_i p_{i-1}$ 
     $p_i = \frac{c_i}{r_i}$ 
     $q_i = \frac{d_i - b_i q_{i-1}}{r_i}$ 
end
 $x_n := q_n$ 
 $i := n - 1$ 
repeat
     $x_i := q_i - p_i x_{i+1}$ 
     $i := i - 1$ 
until  $i = 0$ 
La solución es  $x_1, x_2, \dots, x_n$ 
Terminar

```

Ejemplo 1

Haga un estimado del tiempo que requerirá una computadora capaz de realizar 50 000 multiplicaciones o divisiones por segundo para resolver un sistema tridiagonal de 1000 ecuaciones con 1000 incógnitas utilizando el método de Gauss especializado para sistemas tridiagonales.

Solución:

El total de operaciones es del orden de $8n$, es decir, 8000. El tiempo requerido será de:

$$\frac{8000}{50000} = 0,16 \text{ segundos}$$

Compárese con el tiempo de unas dos horas que se hubiera requerido (ejemplo 2 de la sección anterior) para resolver un sistema de este tamaño mediante el algoritmo general de Gauss en esta misma máquina.

Cálculo de determinantes

Como se vio en la sección anterior, en el proceso directo de Gauss para llevar la matriz del sistema a una forma escalonada, se utilizan transformaciones elementales de tipo I y de tipo II. La

transformación de tipo I (intercambiar dos filas) produce el efecto colateral de cambiar el signo del determinante de la matriz \mathbf{A} del sistema; la transformación de tipo II no afecta el valor del determinante.

Por otra parte, es muy simple demostrar que el determinante de una matriz triangular es el producto de los elementos en la diagonal de la matriz. En efecto, utilizando el desarrollo por menores de un determinante de orden n respecto a su primera columna, que tiene un solo elemento no nulo, se tiene que:

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{vmatrix}$$

Desarrollando el determinante de orden $n - 1$ respecto a su primera columna, que solo posee un elemento no nulo, se tiene que:

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{vmatrix} = a_{11}a_{22} \begin{vmatrix} a_{33} & a_{34} & \cdots & a_{3n} \\ 0 & a_{44} & \cdots & a_{4n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{vmatrix}$$

Repetiendo la operación las veces necesarias se llega a:

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{vmatrix} = a_{11}a_{22}a_{33}\cdots a_{nn}$$

A partir de lo anterior, una forma eficiente de calcular el determinante de una matriz es realizar sobre ella transformaciones elementales de tipo I y II que la reduzcan a una matriz triangular (tal como hace el proceso directo del método de Gauss) tomando en cuenta que cada aplicación de una transformación de tipo I produce un cambio en el signo del determinante. Una vez que la matriz tiene forma triangular, el determinante se obtiene como el producto de los elementos de la diagonal principal.

El cálculo de un determinante de orden n por esta vía requiere una cantidad de operaciones que está determinada por el proceso directo de Gauss, que es mayor consumidor de tiempo de cómputo. Por tanto en número de operaciones es del orden de $\frac{1}{3}n^3$.

En el algoritmo que sigue, se utiliza una variable llamada *Signo* a la que inicialmente se le da valor 1, y que cambia de signo cada vez que se realiza un intercambio de filas. Al final el producto de los elementos de la diagonal es afectada por la variable *Signo* para obtener el valor del determinante. En el proceso directo de Gauss se utiliza la estrategia de pivote parcial.

Algoritmo para calcular determinantes

El algoritmo que sigue utiliza como datos el entero $n > 1$ y los elementos de la matriz cuadrada \mathbf{A} de orden n .

```

Signo := 1
for  $i = 1$  to  $n - 1$ 
    {En este sector se halla la fila pivote mediante la estrategia parcial}
    max :=  $|a_{ii}|$ 
    Fila de max :=  $i$ 
    for  $k = i + 1$  to  $n$ 
        if  $|a_{ki}| > max$  then
            max :=  $|a_{ki}|$ 
            Fila de max :=  $k$ 
        end
    end
    if max = 0 then
        El determinante es cero
        Terminar
    end
    if  $k > i$  then
        Intercambiar fila( $k$ ) con fila( $i$ )
        Signo := - Signo
    end

    {En el sector que sigue se anulan los elementos de la columna  $i$  que se hallan
    debajo de la diagonal}
    for  $k = i + 1$  to  $n$ 
         $m := \frac{a_{ki}}{a_{ii}}$ 
        fila( $k$ ) := fila( $k$ ) -  $m$  fila( $i$ )
    end
end
Determinante := Signo
for  $i = 1$  to  $n$ 
    Determinante :=  $(a_{ii})(Determinante)$ 
end
El valor del determinante es Determinante
Terminar

```

Inversión de matrices mediante el método de Gauss

El problema de hallar la inversa de la matriz cuadrada \mathbf{A} de orden n equivale a encontrar otra matriz \mathbf{X} del mismo orden tal que:

$$\mathbf{AX} = \mathbf{I} \quad (5)$$

donde \mathbf{I} es la matriz identidad de orden n . En forma desarrollada, la ecuación (5) es:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \quad (6)$$

Si las matrices \mathbf{X} e \mathbf{I} se dividen en bloques por columnas, la ecuación matricial (6) toma la forma:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \cdot \begin{pmatrix} \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} & \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} & \cdots & \begin{bmatrix} x_{1n} \\ x_{2n} \\ \vdots \\ x_{nn} \end{bmatrix} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} & \cdots & \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \end{pmatrix}$$

Esta ecuación puede ser separada en n ecuaciones más simples:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_{1n} \\ x_{2n} \\ \vdots \\ x_{nn} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

Cada una de estas n ecuaciones matriciales no es otra cosa que un sistema lineal de ecuaciones. Resolviendo cada uno de los sistemas mediante el método de Gauss, se tendrían las n columnas de la matriz inversa. Como los n sistemas poseen la misma matriz de coeficientes (lo que cambia es el vector de términos independientes) se puede hacer una adaptación del método de Gauss en la cual la matriz ampliada se forma añadiendo a la matriz \mathbf{A} las n columnas de términos independientes y entonces el proceso directo (que es el que más operaciones requiere) se realiza con esta matriz ampliada. Por supuesto, el proceso inverso del método de Gauss habrá que realizarlo n veces, una con cada una de las columnas con que se amplió la matriz \mathbf{A} . Puede

probarse que la cantidad de operaciones que requiere todo el algoritmo es del orden de $\frac{4}{3}n^3$.

Algoritmo para invertir una matriz mediante el método de Gauss

El algoritmo en seudo código que sigue utiliza la idea anterior para hallar la inversa de la matriz cuadrada \mathbf{A} de orden n . En el proceso directo de Gauss se utiliza la estrategia parcial de pivote. Se supone que \mathbf{A} es no singular. El algoritmo utiliza como datos el entero $n > 1$ y los elementos de la matriz \mathbf{A} .

```

C := [A|I] {La matriz C se obtiene ampliando A con los elementos de la idéntica}
for i = 1 to n - 1
    {En este sector se halla la fila pivote mediante la estrategia parcial}
     $max := |c_{ii}|$ 
    Fila de max := i
    for k = i + 1 to n
        if  $|c_{ki}| > max$  then
             $max := |c_{ki}|$ 
            Fila de max := k
        end
    end
    if  $k > i$  then
        Intercambiar fila(k) con fila(i)
    end

    {En el sector que sigue se anulan los elementos de la columna  $i$  que se hallan
     debajo de la diagonal}
    for k = i + 1 to n
         $m := \frac{c_{ki}}{c_{ii}}$ 
        fila(k) := fila(k) - m fila(i)
    end
end
{A partir de aquí se realiza el proceso inverso}

for j = 1 to n {El proceso inverso se repetirá  $n$  veces}
     $i := n$ 
    repeat {En este ciclo se halla el elemento  $i$  de la columna  $j$  de la matriz inversa}
         $x_{ij} := c_{i,n+j}$  {El término independiente está en la columna  $n + j$ }
        for k = i + 1 to n {Cuando  $i = n$  este lazo no se ejecuta}
             $x_{ij} := x_{ij} - c_{ik}x_{kj}$ 
        end
         $x_{ij} := \frac{x_{ij}}{c_{ii}}$ 
         $i := i - 1$ 
    until  $i = 0$ 
end
La matriz inversa es  $\mathbf{X} = [x_{ij}]$ 
Terminar

```

Ejercicios

En los siguientes ejercicios, utilice programas computacionales, preferiblemente confeccionados por usted. De no contar con un programa adecuado, use una calculadora y trabaje con cinco cifras significativas en los cálculos.

- Resuelva los siguientes sistemas de ecuaciones mediante el algoritmo de Gauss especializado en sistemas tridiagonales. Verifique previamente que en cada caso la diagonal es predominante.

$$\begin{array}{l}
 \begin{array}{lll}
 7x_1 + 3x_2 = 5 & 5x_1 + 2x_2 = 6 & -3x_1 + x_2 = 6 \\
 2x_1 + 6x_2 - x_3 = 6 & -x_1 + 7x_2 - 3x_3 = 4 & 3x_1 + 7x_2 + 3x_3 = 2 \\
 \text{a) } x_2 + 8x_3 + 3x_4 = 2 & \text{b) } 2x_2 + 9x_3 + 6x_4 = 10 & \text{c) } x_2 + 6x_3 + 2x_4 = 3 \\
 3x_3 + 8x_4 - x_5 = 2 & x_3 + 6x_4 - 2x_5 = 1 & 3x_3 + 7x_4 + 2x_5 = 5 \\
 x_4 + 3x_5 = 4 & 3x_4 + 5x_5 = 6 & 4x_4 + 6x_5 = 2
 \end{array}
 \end{array}$$

- Como se verá en el próximo capítulo, para hacer pasar un tipo de curva llamado spline cúbico natural por n puntos del plano, se requiere resolver un sistema de ecuaciones de $n-2$ ecuaciones lineales, el cual es tridiagonal y con la diagonal predominante. Suponga una computadora que es capaz de realizar 50000 multiplicaciones por segundo y estime qué tiempo requiere esa máquina para calcular un spline que pasa por 100 puntos de la pantalla. Considere dos posibilidades: a) Utilizar el método general de Gauss, b) Utilizar el método de Gauss especializado en sistemas tridiagonales.
- Para hallar un spline cúbico natural que pase por los $n+1$ puntos del plano: $P_0(x_0, y_0)$, $P_1(x_1, y_1)$, ..., $P_n(x_n, y_n)$ con $x_0 < x_1 < \dots < x_n$ se necesita resolver el sistema tridiagonal

$$\frac{h_{i-1}}{6}M_{i-1} + \frac{h_{i-1} + h_i}{3}M_i + \frac{h_i}{6}M_{i+1} = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}} \quad i = 1, 2, \dots, n-1$$

donde $h_i = x_{i+1} - x_i \quad i = 0, 1, \dots, n-1$
 $M_0 = M_n = 0$

Elabore un algoritmo en seudo código que utilice el método de Gauss especializado en sistemas tridiagonales, para calcular los coeficientes M_i ($i = 1, 2, \dots, n-1$). Verifique previamente si este sistema tiene la diagonal predominante.

- Calcule los siguientes determinantes mediante el algoritmo basado en el proceso directo de Gauss.

$$\begin{array}{lll}
 \text{a) } \left| \begin{array}{rrrr} 3 & -4 & 3 & 2 \\ 5 & 7 & 1 & -4 \\ 3 & 2 & 2 & 1 \\ 0 & 2 & 7 & 1 \end{array} \right| & \text{b) } \left| \begin{array}{rrrr} 2 & 3 & 1 & 1 \\ 0 & 4 & 2 & 6 \\ 1 & 4 & -2 & 4 \\ 2 & 1 & 3 & 5 \end{array} \right| & \text{c) } \left| \begin{array}{rrrr} 3 & -2 & 0 & 2 \\ 1 & 2 & 3 & 4 \\ 3 & 2 & 3 & 7 \\ 3 & 5 & 6 & 2 \end{array} \right|
 \end{array}$$

5. El siguiente algoritmo para calcular determinantes es una modificación del que se mostró en las páginas anteriores. Analice cómo funciona el mismo y detecte y rectifique un error que contiene.

```

Cambios := 0
for  $i = 1$  to  $n - 1$ 
    if  $a_{ii} = 0$  then
         $k := i$ 
        repeat
             $k := k + 1$ 
        until  $a_{ki} \neq 0$  or  $k > n$ 
        if  $k > n$  then
            El determinante es cero
            Terminar
        end
        if  $k < i$  then
            Intercambiar  $\text{fila}(k)$  con  $\text{fila}(i)$ 
             $\text{Cambios} := \text{Cambios} + 1$ 
        end
    end
    for  $k = i + 1$  to  $n$ 
         $m := \frac{a_{ki}}{a_{ii}}$ 
         $\text{fila}(k) := \text{fila}(k) - m \text{fila}(i)$ 
    end
end
Determinante :=  $\text{Cambios} \bmod 2$ 
for  $i = 1$  to  $n$ 
     $\text{Determinante} := (a_{ii})(\text{Determinante})$ 
end
El valor del determinante es Determinante
Terminar

```

6. Suponga que utilizando una calculadora de mano usted es capaz de multiplicar o sumar dos números cualesquiera en 5 segundos. Calcule que tiempo tardaría en calcular manualmente determinantes de orden 4, 6, 8 y 10 en dos variantes: a) Utilizando el método basado en el algoritmo de Gauss. b) Desarrollando el determinante por menores. En ambos casos suponga que los algoritmos requieren tantas sumas como multiplicaciones.
7. Halle la matriz inversa de cada una de las siguientes matrices mediante el método basado en el algoritmo de Gauss.

$$\begin{array}{l}
 \text{a)} \begin{bmatrix} 3 & 5 & 2 \\ 3 & -1 & 3 \\ 4 & 2 & 4 \end{bmatrix} \quad \text{b)} \begin{bmatrix} 1 & 2 & -1 & 3 \\ 3 & 0 & 1 & 2 \\ 2 & 3 & 2 & -2 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \text{c)} \begin{bmatrix} 1 & -1 & 0 & 2 \\ 2 & 1 & 3 & 1 \\ 1 & 1 & 2 & 2 \\ 3 & 0 & 1 & 1 \end{bmatrix}
 \end{array}$$

8. Suponga que usted necesita resolver un número grande p de sistemas lineales, todos con la misma matriz de coeficientes \mathbf{A} pero con diferentes vectores de términos independientes. Elabore un algoritmo basado en el método de Gauss que utilice una idea similar a la que se

utilizó para hallar la inversa de una matriz: ampliar la matriz de los coeficientes con todos los vectores independientes, realizar una vez el proceso inverso y p veces el proceso inverso.

9. El esquema de la figura 2 representa un sistema real con cuatro causas x_1, x_2, x_3, x_4 y cuatro efectos y_1, y_2, y_3, y_4 . Como el sistema es lineal, las relaciones entre causas y efectos se pueden representar con una matriz \mathbf{A} , de modo que $\mathbf{y} = \mathbf{Ax}$ donde \mathbf{y} es el vector de efectos y \mathbf{x} el vector de causas. Determine la relación entre efectos y causas, de modo que a partir de un vector de efectos se pueda hallar el vector de causas que lo originó. Se sabe que:

$$\begin{aligned}y_1 &= 3x_1 - 2x_2 + x_3 - 5x_4 \\y_2 &= 2x_1 + 3x_2 - x_3 + 2x_4 \\y_3 &= -x_1 + 2x_2 + 3x_3 + x_4 \\y_4 &= x_1 - 3x_2 + 2x_3 - 4x_4\end{aligned}$$

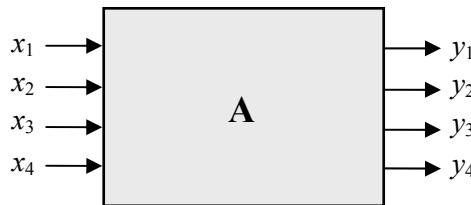


Figura 2

10. En el ejemplo 1 de la sección 3.1 se analizó un circuito de cuatro mallas por el método de la corrientes de malla. El sistema de ecuaciones obtenido se puede generalizar del siguiente modo:

$$\begin{aligned}R_{11}i_1 - R_{12}i_2 - \cdots - R_{1n}i_n &= V_1 \\- R_{21}i_1 + R_{22}i_2 - \cdots - R_{2n}i_n &= V_2 \\&\vdots \\- R_{n1}i_1 - R_{n2}i_2 - \cdots + R_{nn}i_n &= V_n\end{aligned}$$

- a) Determine el sentido físico de los diferentes parámetros que aparecen en el sistema
- b) Exprese el sistema matricialmente como $\mathbf{V} = \mathbf{RI}$
- c) ¿Qué características tiene la matriz \mathbf{R} ?
- d) ¿Cómo se puede expresar el vector \mathbf{I} de corrientes de malla en función del vector \mathbf{V} de voltajes de malla?

3.4 Sistemas mal condicionados

En la sección 1.7 fue tratado en general el problema de la estabilidad. Allí se dijo que un problema inestable es aquel en el cual pequeños cambios en los datos pueden producir grandes cambios en los resultados. Entre los sistemas lineales de ecuaciones se presentan a veces problemas inestables, que reciben más frecuentemente el nombre de mal condicionados (del inglés *ill-conditioned*). En el ejemplo 2 de aquella sección se mostró un pequeño sistema de dos

ecuaciones lineales mal condicionado, en el cual un pequeño cambio de 1% en uno de los datos causó un cambio del orden de 100% en la solución. En esta sección se verá un poco más profundamente el tema de los sistemas lineales mal condicionados, se estudiará la manera de medir el mal condicionamiento y se analizará la forma en que debe procederse en estos casos.

Ejemplo 1

Analice el mal condicionamiento de los sistemas lineales:

$$a) \begin{cases} 4x_1 + 3x_2 + 8x_3 - 7x_4 = 8 \\ 5x_1 + 6x_2 - 3x_3 + 5x_4 = 13 \\ 2x_1 + 7x_2 + 4x_3 - 4x_4 = 9 \\ 3x_1 - 2x_2 + 5x_3 + 8x_4 = 14 \end{cases}$$

$$b) \begin{cases} 5x_1 + 7x_2 + 6x_3 + 5x_4 = 23 \\ 7x_1 + 10x_2 + 8x_3 + 7x_4 = 32 \\ 6x_1 + 8x_2 + 10x_3 + 9x_4 = 33 \\ 5x_1 + 7x_2 + 9x_3 + 10x_4 = 31 \end{cases}$$

Solución:

Ambos sistemas tienen como solución: $x_1 = 1; x_2 = 1; x_3 = 1; x_4 = 1$.

a) Al cambiar ligeramente los términos independientes la solución cambia de la siguiente forma:

$$\text{Para } \mathbf{b} = \begin{bmatrix} 8,1 \\ 13 \\ 9 \\ 14 \end{bmatrix} \text{ se obtiene } \mathbf{x} = \begin{bmatrix} 1,0199 \\ 0,9901 \\ 0,9984 \\ 0,9910 \end{bmatrix}$$

$$\text{Para } \mathbf{b} = \begin{bmatrix} 8 \\ 13,1 \\ 9 \\ 14 \end{bmatrix} \text{ se obtiene } \mathbf{x} = \begin{bmatrix} 1,0157 \\ 0,9999 \\ 0,9916 \\ 0,9993 \end{bmatrix}$$

$$\text{Para } \mathbf{b} = \begin{bmatrix} 8 \\ 13 \\ 9,1 \\ 14 \end{bmatrix} \text{ se obtiene } \mathbf{x} = \begin{bmatrix} 0,9769 \\ 1,0188 \\ 1,0105 \\ 1,9968 \end{bmatrix}$$

$$\text{Para } \mathbf{b} = \begin{bmatrix} 8 \\ 13 \\ 9 \\ 14,1 \end{bmatrix} \text{ se obtiene } \mathbf{x} = \begin{bmatrix} 0,9961 \\ 1,0008 \\ 1,0091 \\ 1,0085 \end{bmatrix}$$

En todos los casos que se analizaron, al realizar cambios del orden de 1% en el vector \mathbf{b} se obtuvo una solución con cambios del orden de 1% del valor original.

$$b) \text{ Para } \mathbf{b} = \begin{bmatrix} 23,1 \\ 32 \\ 33 \\ 31 \end{bmatrix} \text{ se obtiene } \mathbf{x} = \begin{bmatrix} 7,8000 \\ -3,1000 \\ -0,7000 \\ 2,0000 \end{bmatrix}$$

Con un cambio en \mathbf{b} del orden de 0,5% se han producido en la solución cambios del orden de 700%.

Conclusión: Todo parece indicar que el sistema a) es bien condicionado. El sistema b) presenta mal condicionamiento.

El ejemplo 1 ilustra la forma más efectiva de probar que un sistema es mal condicionado: verificar que pequeños cambios en los datos provocan cambios mucho mayores en la solución. Esta prueba, en cambio no asegura el buen condicionamiento pues no es posible realizar toda la variedad de pequeñas alteraciones en los datos del problema; por eso en el caso a) la respuesta no es concluyente.

Una medida del mal condicionamiento

En lo que sigue se tratará de hallar una relación entre el cambio relativo del vector \mathbf{b} y el cambio relativo del vector \mathbf{x} en un sistema general

$$\mathbf{Ax} = \mathbf{b} \quad (1)$$

Como la norma de un vector mide el tamaño del vector, es natural definir el cambio sufrido por el vector \mathbf{b} al pasar a \mathbf{b}_1 como la norma del vector diferencia $\mathbf{b} - \mathbf{b}_1$. El cambio relativo se definirá entonces dividiendo el resultado anterior por la norma de \mathbf{b} . Esto es:

Cambio relativo al cambiar \mathbf{b} por \mathbf{b}_1 :
$$\frac{\|\mathbf{b} - \mathbf{b}_1\|}{\|\mathbf{b}\|}$$

Sea ahora \mathbf{x}_1 el vector solución del sistema lineal obtenido al sustituir \mathbf{b} por \mathbf{b}_1 , es decir:

$$\mathbf{Ax}_1 = \mathbf{b}_1 \quad (2)$$

El cambio relativo sufrido por la solución será:
$$\frac{\|\mathbf{x} - \mathbf{x}_1\|}{\|\mathbf{x}\|}$$

Para establecer una relación entre estos cambios relativos, se restan miembro a miembro las ecuaciones (1) y (2) y se extrae \mathbf{A} factor común:

$$\mathbf{A}(\mathbf{x} - \mathbf{x}_1) = \mathbf{b} - \mathbf{b}_1$$

Premultiplicando por \mathbf{A}^{-1} ambos miembro de esta ecuación:

$$\mathbf{x} - \mathbf{x}_1 = \mathbf{A}^{-1}(\mathbf{b} - \mathbf{b}_1)$$

Ahora bien, según el axioma 6 de las normas matriciales, la norma del producto de una matriz por un vector es menor o igual que el producto de sus respectivas normas, así que:

$$\|\mathbf{x} - \mathbf{x}_1\| \leq \|\mathbf{A}^{-1}\| \cdot \|\mathbf{b} - \mathbf{b}_1\|$$

Dividiendo en ambos miembros por la norma de \mathbf{x} se obtiene:

$$\frac{\|\mathbf{x} - \mathbf{x}_1\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\| \cdot \|\mathbf{b} - \mathbf{b}_1\|}{\|\mathbf{x}\|} = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \cdot \frac{\|\mathbf{b} - \mathbf{b}_1\|}{\|\mathbf{A}\| \cdot \|\mathbf{x}\|} \quad (3)$$

Pero como $\mathbf{Ax} = \mathbf{b}$, aplicando de nuevo el axioma 6, se tiene que $\|\mathbf{A}\| \cdot \|\mathbf{x}\| \geq \|\mathbf{b}\|$, así que cambiando el producto $\|\mathbf{A}\| \cdot \|\mathbf{x}\|$ en el denominador de (3) por la norma de \mathbf{b} la desigualdad se refuerza y se obtiene:

$$\frac{\|\mathbf{x} - \mathbf{x}_1\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \cdot \frac{\|\mathbf{b} - \mathbf{b}_1\|}{\|\mathbf{b}\|}$$

Como se ve, el término $\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \cdot \frac{\|\mathbf{b} - \mathbf{b}_1\|}{\|\mathbf{b}\|}$ constituye una cota para el cambio relativo que sufrirá la solución del sistema. Esta cota se obtiene multiplicando el cambio relativo que haya sufrido \mathbf{b} por el número

$$\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$$

Valores pequeños de este número garantizan que el cambio relativo de \mathbf{x} no será grande respecto al cambio relativo de \mathbf{b} .

Definición 1

El número de condición del sistema lineal $\mathbf{Ax} = \mathbf{b}$ se denota $\text{cond}(\mathbf{A})$ y se define como

$$\text{cond}(\mathbf{A}) = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \quad \blacksquare$$

El número de condición de la matriz \mathbf{A} permite medir el mal condicionamiento del sistema. Es fácil ver que:

$$\text{cond}(\mathbf{A}) = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \geq \|\mathbf{A}\mathbf{A}^{-1}\| = \|\mathbf{I}\| = 1$$

esto es, el número de condición siempre es mayor o igual que 1. Valores de $\text{cond}(\mathbf{A})$ próximos a 1 aseguran un buen condicionamiento del sistema lineal mientras valores mucho mayores que 1 indican un mal condicionamiento. Puede probarse que siempre se puede seleccionar un vector \mathbf{b} y uno \mathbf{b}_1 de modo que

$$\frac{\|\mathbf{x} - \mathbf{x}_1\|}{\|\mathbf{x}\|} = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \cdot \frac{\|\mathbf{b} - \mathbf{b}_1\|}{\|\mathbf{b}\|}$$

así que un valor grande de $\text{cond}(\mathbf{A})$ significa que para determinados vectores \mathbf{b}_1 pudiera haber grandes cambios relativos en \mathbf{x} . En la práctica, valores de $\text{cond}(\mathbf{A})$ menores que 10 se toman como problemas bien condicionados mientras que valores mayores que 10 ya comienzan a ser considerados como mal condicionados.

Ejemplo 2

Halle el número de condición para cada uno de los sistemas lineales del ejemplo 1.

Solución:

- a) En este caso el sistema es

$$\begin{cases} 4x_1 + 3x_2 + 8x_3 - 7x_4 = 8 \\ 5x_1 + 6x_2 - 3x_3 + 5x_4 = 13 \\ 2x_1 + 7x_2 + 4x_3 - 4x_4 = 9 \\ 3x_1 - 2x_2 + 5x_3 + 8x_4 = 14 \end{cases}$$

La matriz \mathbf{A} del sistema se obtiene directamente. Su inversa se obtuvo mediante un programa confeccionado a partir del algoritmo mostrado en la sección anterior:

$$\mathbf{A} = \begin{bmatrix} 4 & 3 & 8 & -7 \\ 5 & 6 & -3 & 5 \\ 2 & 7 & 4 & -4 \\ 3 & -2 & 5 & 8 \end{bmatrix} \quad \mathbf{A}^{-1} = \begin{bmatrix} 0,1990 & 0,1569 & -0,2310 & -0,0394 \\ -0,0991 & -0,0006 & 0,1880 & 0,0077 \\ -0,0157 & -0,0839 & 0,1047 & 0,0910 \\ -0,0896 & -0,0065 & 0,0682 & 0,0848 \end{bmatrix}$$

$$\|\mathbf{A}\| = \max\{22, 19, 17, 18\} = 22 \quad \|\mathbf{A}^{-1}\| = \max\{0,6263, 0,2954, 0,2953, 0,2491\} = 0,6263$$

$$\text{cond}(\mathbf{A}) = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| = (22)(0,6263) = 13,8$$

El sistema presenta solamente un ligero mal condicionamiento.

b) En este caso el sistema es

$$\begin{cases} 5x_1 + 7x_2 + 6x_3 + 5x_4 = 23 \\ 7x_1 + 10x_2 + 8x_3 + 7x_4 = 32 \\ 6x_1 + 8x_2 + 10x_3 + 9x_4 = 33 \\ 5x_1 + 7x_2 + 9x_3 + 10x_4 = 31 \end{cases}$$

La matriz \mathbf{A} del sistema y su inversa son:

$$\mathbf{A} = \begin{bmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix} \quad \mathbf{A}^{-1} = \begin{bmatrix} 68 & -41 & -17 & 10 \\ -41 & 25 & 10 & -6 \\ -17 & 10 & 5 & -3 \\ 10 & -6 & -3 & 2 \end{bmatrix}$$

$$\|\mathbf{A}\| = \max\{23, 32, 33, 31\} = 33 \quad \|\mathbf{A}^{-1}\| = \max\{136, 82, 35, 21\} = 136$$

$$\text{cond}(\mathbf{A}) = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| = (33)(136) = 4488$$

El sistema es muy mal condicionado. ■

¿Qué hacer?

Una vez que se tiene la certeza o la sospecha de que un sistema puede presentar mal condicionamiento, es importante tomar medidas para disminuir los errores, ya que ellos representan pequeños cambios en los datos del problema. Para ello hay dos cosas que deben hacerse. Primero utilizar una estrategia de pivote parcial o total (si está disponible), de modo que no se produzca una ampliación de los errores de redondeo. En segundo lugar, desminuir los errores de redondeo trabajando con un número mayor de cifras exactas; la mayoría de los lenguajes de programación brindan la posibilidad de trabajar en precisión doble o extendida, lo cual hace más lentos los programas pero utilizan entre 12 y 18 cifras exactas, lo cual es suficiente para que la respuesta obtenida sea correcta.

Existen además algunas técnicas iterativas para mejorar la solución, una de las cuales es la que sigue:

Sea el sistema a resolver

$$\mathbf{Ax} = \mathbf{b} \quad (4)$$

Sea \mathbf{x}_1 una solución obtenida por el método de Gauss (o por otro) la cual se sospecha que posee errores debido al mal condicionamiento del sistema. Si \mathbf{x}_1 fuera exacta, la diferencia vectorial

$$\mathbf{r}_1 = \mathbf{b} - \mathbf{Ax}_1 \quad (5)$$

sería el vector nulo, de modo que el vector \mathbf{r}_1 , llamado *residuo* es una manera de medir la discrepancia de una solución aproximada \mathbf{x}_1 con la verdadera solución \mathbf{x} . El residuo de cualquier solución aproximada se obtiene sin mucho esfuerzo computacional. El proceso que sigue tiene por objetivo disminuir el residuo hasta valores aceptables.

Sea $\text{error}(\mathbf{x}_1) = \mathbf{x} - \mathbf{x}_1$

Sustituyendo en la ecuación (5) \mathbf{b} por \mathbf{Ax} se obtiene:

$$\mathbf{r}_1 = \mathbf{Ax} - \mathbf{Ax}_1 = \mathbf{A}(\mathbf{x} - \mathbf{x}_1)$$

Pero esto significa que

$$\mathbf{A} \text{error}(\mathbf{x}_1) = \mathbf{r}_1 \quad (6)$$

Resolviendo la ecuación (6), que tiene la misma matriz de coeficientes que el sistema original, se obtiene $\text{error}(\mathbf{x}_1)$ en forma aproximada (debido al mal condicionamiento de \mathbf{A}), sin embargo, la solución aproximada:

$$\mathbf{x}_2 = \mathbf{x}_1 + \text{error}(\mathbf{x}_1)$$

está mucho más próxima a la solución exacta. Ahora puede hallarse el residuo de \mathbf{x}_2 :

$$\mathbf{r}_2 = \mathbf{b} - \mathbf{Ax}_2$$

y el error de \mathbf{x}_2 resolviendo el sistema: $\mathbf{A} \text{error}(\mathbf{x}_2) = \mathbf{r}_2$

y el proceso se continúa hasta que la diferencia entre dos aproximaciones sucesivas \mathbf{x}_n y \mathbf{x}_{n-1} sea suficientemente pequeña.

Ejercicios

1. Para cada uno de los sistemas lineales que siguen, halle el número de condición y analice su mal condicionamiento.

$$\begin{array}{ll}
 28x_1 + 17x_2 + 35x_3 + 29x_4 = 48 & 26x_1 + 15x_2 + 18x_3 + 28x_4 = 37 \\
 \text{a) } 23x_1 + 28x_2 + 32x_3 + 15x_4 = 31 & 45x_1 + 25x_2 + 34x_3 + 53x_4 = 62 \\
 43x_1 + 20x_2 + 29x_3 + 25x_4 = 36 & 42x_1 + 23x_2 + 27x_3 + 45x_4 = 19 \\
 31x_1 + 44x_2 + 18x_3 + 27x_4 = 55 & 28x_1 + 16x_2 + 19x_3 + 30x_4 = 44
 \end{array}$$

2. En los dos sistemas del ejercicio 1, realice pequeños cambios en el vector de términos independientes y analice los cambios que sufre la respuesta. Verifique las conclusiones a las que usted llegó en el ejercicio 1 acerca del mal condicionamiento de cada uno de ellos.
3. Suponga que usted ya tiene un algoritmo que, a partir de una matriz \mathbf{A} cuadrada de orden n y no singular y un vector \mathbf{b} le da como salida la solución del sistema $\mathbf{Ax} = \mathbf{b}$. Utilice este procedimiento para elaborar un algoritmo que calcule el vector de residuos y mejore iterativamente la respuesta hasta que la norma de la diferencia entre dos aproximaciones sucesivas no exceda de una cierta tolerancia ε .
4. La matriz \mathbf{H}_4 que se muestra a continuación se llama matriz de Hilbert de orden 4. Este tipo de matriz aparece en muchas cuestiones prácticas. La matriz $\hat{\mathbf{H}}_4$ es una aproximación de \mathbf{H}_4 obtenida redondeando sus elementos a tres cifras decimales exactas.

$$\mathbf{H}_4 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{bmatrix} \quad \hat{\mathbf{H}}_4 = \begin{bmatrix} 1 & 0,5 & 0,333 & 0,25 \\ 0,25 & 0,333 & 0,25 & 0,2 \\ 0,333 & 0,25 & 0,2 & 0,167 \\ 0,25 & 0,2 & 0,167 & 0,143 \end{bmatrix}$$

Al calcular las inversas de estas matrices se obtienen resultados muy diferentes:

$$(\mathbf{H}_4)^{-1} = \begin{bmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{bmatrix} \quad (\hat{\mathbf{H}}_4)^{-1} = \begin{bmatrix} 2,3 & 32,8 & -127,4 & 98,9 \\ 32,8 & -499,9 & 1381,8 & -971,8 \\ -127,4 & 1381,8 & -3314,1 & 2160,5 \\ 98,9 & -971,8 & 2160,5 & -1329,9 \end{bmatrix}$$

Explique a qué se debe este comportamiento.

3.5 Métodos iterativos para sistemas lineales

Del mismo modo que la solución de ecuaciones escalares puede obtenerse por métodos iterativos (bisección, Regula Falsi, Newton – Raphson, Secantes) los sistemas de ecuaciones lineales pueden ser resueltos también por procedimientos de este estilo, lo cual tiene a veces sus ventajas. Una de estas ventajas es la sencillez de estos algoritmos, lo cual significa programas computacionales más simples y seguros. Otra de sus ventajas es la no propagación de errores. En

efecto, como en los métodos iterativos se obtiene una sucesión de soluciones aproximadas que converge hacia la solución exacta, la aproximación número $n - 1$ solo sirve para encontrar una mejor solución en la aproximación número n , por tanto, los errores de la aproximación $n - 1$ no se propagan a la aproximación n .

A pesar de estos dos atractivos, la cuestión más importante al decidirse por uno u otro algoritmo es la eficiencia computacional, medida en tiempo de máquina, y, para sistemas muy voluminosos, la cantidad de memoria necesaria. En este sentido, los métodos iterativos no siempre compiten con el método de Gauss. Sobre este aspecto se volverá más adelante.

La diferencia entre los distintos métodos iterativos existentes está en la forma en que se genera la sucesión de soluciones aproximadas. De ellos se verá el método de Jacobi y el método de Seidel que son dos de los más empleados actualmente.

El método de Jacobi

Sea $\mathbf{Ax} = \mathbf{b}$ un sistema de n ecuaciones lineales. Se supone, como hasta ahora, que este sistema es cuadrado y de solución única. Escrito en forma desarrollada, el sistema es:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned} \quad (1)$$

Si todos los elementos a_{ii} ($i = 1, 2, \dots, n$) de la diagonal son no nulos, el sistema se puede escribir así:

$$\begin{aligned} x_1 &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2 - \cdots - \frac{a_{1n}}{a_{11}}x_n \\ x_2 &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1 - \frac{a_{23}}{a_{22}}x_3 - \cdots - \frac{a_{2n}}{a_{22}}x_n \\ &\vdots \\ x_n &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1 - \frac{a_{n2}}{a_{nn}}x_2 - \cdots - \frac{a_{n(n-1)}}{a_{nn}}x_{n-1} \end{aligned} \quad (2)$$

donde, como se aprecia, en la i -sima ecuación ($i = 1, 2, \dots, n$) se ha despejado x_i . Utilizando la forma matricial, el sistema puede escribirse:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & -\frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \vdots \\ \frac{b_n}{a_{nn}} \end{bmatrix}$$

Llamando \mathbf{M} y \mathbf{c} a las matrices:

$$\mathbf{M} = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots & -\frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \dots & 0 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \vdots \\ \frac{b_n}{a_{nn}} \end{bmatrix}$$

el sistema toma la forma

$$\mathbf{x} = \mathbf{Mx} + \mathbf{c} \quad (3)$$

A partir de la ecuación (3) se puede generar un proceso iterativo matricial, partiendo de un vector inicial $\mathbf{x}^{(0)}$ y generando una sucesión de vectores $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \dots$ mediante la ecuación recursiva:

$$\mathbf{x}^{(k)} = \mathbf{Mx}^{(k-1)} + \mathbf{c} \quad k = 1, 2, 3, \dots \quad (4)$$

La sucesión generada puede o no converger pero, cuando lo hace, su límite es la solución del sistema (3) o (2) o (1) que son todos equivalentes. Este algoritmo recibe el nombre de método de Jacobi.

Ejemplo 1

Escriba los siguientes sistemas lineales en la forma $\mathbf{x} = \mathbf{Mx} + \mathbf{c}$ y aplíquelo el algoritmo de Jacobi.

a) $10x_1 - x_2 + 2x_3 = 6$	b) $10x_1 + 8x_2 - 7x_3 = -3$	c) $4x_1 - 3x_2 + 5x_3 = 25$
$x_1 - 5x_2 + x_3 = -10$	$-3x_1 + 10x_2 + x_3 = -25$	$3x_1 + 2x_2 + 2x_3 = 11$
$2x_1 - x_2 + 8x_3 = -8$	$5x_1 - x_2 + 10x_3 = 22$	$4x_1 - 2x_2 + 3x_3 = -4$

Solución:

a) Despejando x_1 en la primera ecuación, x_2 en la segunda y x_3 en la tercera, se tiene:

$$\begin{aligned} x_1 &= 0x_1 + 0,1x_2 - 0,2x_3 + 0,6 \\ x_2 &= 0,2x_1 + 0x_2 + 0,2x_3 + 2 \\ x_3 &= -0,25x_1 + 0,125x_2 + 0x_3 - 1 \end{aligned}$$

que en forma matricial resulta:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 0,1 & -0,2 \\ 0,2 & 0 & 0,2 \\ -0,25 & 0,125 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0,6 \\ 2 \\ -1 \end{bmatrix}$$

Tomando $\mathbf{x}^{(0)}$ como el vector nulo y generando la sucesión de vectores $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \dots$ mediante la ecuación recursiva:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}^{(k)} = \begin{bmatrix} 0 & 0,1 & -0,2 \\ 0,2 & 0 & 0,2 \\ -0,25 & 0,125 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}^{(k-1)} + \begin{bmatrix} 0,6 \\ 2 \\ -1 \end{bmatrix} \quad k = 1, 2, 3, \dots$$

se obtienen los resultados que se muestran en la tabla 1. Resulta evidente que en este caso el método de Jacobi converge hacia la solución del sistema de ecuaciones, que es $x_1 = 1$, $x_2 = 2$ y $x_3 = -1$.

Iteración	x_1	x_2	x_3
0	0,0	0,0	0,0
1	0,6	2,0	-1,0
2	1,0	1,92	-0,9
3	0,972	2,02	-1,01
4	1,004	1,9924	-0,9905
5	0,99734	2,0027	-1,00195
6	1,00066	1,999078	-0,998997
7	0,999707	2,000333	-1,00028
8	1,000089	1,999885	-0,999885
9	0,999966	2,000041	-1,000037

Tabla 1

- b) De manera similar al anterior, este sistema queda expresado en la forma $\mathbf{x} = \mathbf{Mx} + \mathbf{c}$ como sigue:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & -0,8 & 0,7 \\ 0,3 & 0 & -0,1 \\ -0,5 & 0,1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} -0,3 \\ -2,5 \\ 2,2 \end{bmatrix}$$

Al ser aplicado el algoritmo de Jacobi se obtienen los resultados que muestra la tabla 2. Por razones de espacio no se muestran todas las filas de la tabla. Como se observará, el método en este caso también converge hacia la solución del sistema, que es $x_1 = 2$, $x_2 = -2$ y $x_3 = 1$, pero ahora la convergencia se produce más lentamente: en el sistema a) se había logrado cuatro cifras decimales exactas en la novena iteración, sin embargo, en este con 40 iteraciones aun no se ha logrado esa exactitud. Más adelante se comprenderá la causa de este comportamiento.

Iteración	x_1	x_2	x_3
0	0,0	0,0	0,0
1	-0,3	-2,5	2,2
2	3,24	-2,81	2,1
3	3,418	-1,738	0,299
4	1,2997	-1,5045	0,3172
5	1,12564	-2,14181	1,3997
:	:	:	:
10	2,121377	-2,111874	1,161146
:	:	:	:
20	1,994785	-1,995258	0,986102
:	:	:	:
30	2,000055	-2,000701	1,001144
:	:	:	:
40	2,000024	-1,999948	0,99991

Tabla 2

c) El sistema queda expresado en la forma $\mathbf{x} = \mathbf{Mx} + \mathbf{c}$ tal como se hizo en los casos anteriores:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 0,75 & -1,25 \\ -1,5 & 0 & -1 \\ -1,3333 & 0,6667 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 6,25 \\ 5,5 \\ -1,3333 \end{bmatrix}$$

Los resultados obtenidos al aplicar el algoritmo de Jacobi se muestran en la tabla 3. Obsérvese como aquí el proceso iterativo no converge hacia la solución del sistema, que es $x_1 = 3$, $x_2 = -1$ y $x_3 = 2$. En las páginas que sigue se hará claro el motivo de este comportamiento.

Iteración	x_1	x_2	x_3
0	0,0	0,0	0,0
1	6,25	5,5	-1,3333
2	12,0417	-2,5417	10,6667
3	-8,9896	-23,2292	13,0298
4	-27,4566	5,9566	-28,8056
5	46,7244	75,4905	-33,9711
6	105,3317	-30,6155	111,2928
7	-155,8277	-263,7907	118,6985

Tabla 3

Convergencia del método de Jacobi

Como se ha visto, no siempre el método de Jacobi da lugar a una sucesión convergente hacia la solución del sistema. En lo que sigue se establecerán algunas condiciones de uso práctico que garanticen la obtención de la solución.

Para estudiar la convergencia del método de Jacobi (y, más adelante, el de Seidel) es necesario definir la forma en que se medirá el error en la k -sima aproximación $\mathbf{x}^{(k)}$, respecto al vector solución \mathbf{x} .

Definición 1

Si $\mathbf{x}^{(k)}$ denota la k -sima aproximación de un proceso iterativo a la solución \mathbf{x} de un sistema lineal, entonces el error absoluto de $\mathbf{x}^{(k)}$ se denota por $E(\mathbf{x}^{(k)})$ y se define como:

$$E(\mathbf{x}^{(k)}) = \|\mathbf{x} - \mathbf{x}^{(k)}\|$$

■

Nótese que, según los axiomas de las normas matriciales, el error absoluto es cero cuando los vectores \mathbf{x} y $\mathbf{x}^{(k)}$ coinciden en todas sus componentes, es decir, cuando su diferencia es el vector nulo.

Ejemplo 2

En el sistema lineal del ejemplo 1 a) muestre \mathbf{x} , $\mathbf{x}^{(4)}$ y $E(\mathbf{x}^{(4)})$

Solución:

La solución \mathbf{x} del sistema es $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}$

La aproximación $\mathbf{x}^{(4)}$ obtenida en la cuarta iteración es $\mathbf{x}^{(4)} = \begin{bmatrix} 1.004 \\ 1.9924 \\ -0.9905 \end{bmatrix}$

El error absoluto de $\mathbf{x}^{(4)}$, de acuerdo con la definición que se acaba de dar es:

$$E(\mathbf{x}^{(4)}) = \|\mathbf{x} - \mathbf{x}^{(4)}\| = \begin{vmatrix} 1 - 1.004 \\ 2 - 1.9924 \\ -1 - (-0.9905) \end{vmatrix} = \begin{bmatrix} -0.004 \\ 0.0076 \\ -0.0095 \end{bmatrix} = \max\{0.004; 0.0076; 0.0095\}$$

$$E(\mathbf{x}^{(4)}) = 0.0095$$

■

Considérese ahora el sistema de ecuaciones cuya solución se desea hallar, escrito como en la ecuación (3):

$$\mathbf{x} = \mathbf{M}\mathbf{x} + \mathbf{c}$$

Si de ella se resta miembro a miembro la ecuación recursiva del método de Jacobi (4)

$$\mathbf{x}^{(k)} = \mathbf{M}\mathbf{x}^{(k-1)} + \mathbf{c} \quad k = 1, 2, 3, \dots$$

Se obtiene:

$$\mathbf{x} - \mathbf{x}^{(k)} = \mathbf{M}(\mathbf{x} - \mathbf{x}^{(k-1)}) \quad k = 1, 2, 3, \dots$$

Tomando norma en cada miembro y aplicando el axioma 6 de las normas matriciales, se obtiene:

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \|\mathbf{M}\| \cdot \|\mathbf{x} - \mathbf{x}^{(k-1)}\|$$

Si se tiene en cuenta la definición de error absoluto:

$$E(\mathbf{x}^{(k)}) \leq \|\mathbf{M}\| \cdot E(\mathbf{x}^{(k-1)}) \quad k = 1, 2, 3, \dots \quad (5)$$

La norma de \mathbf{M} juega un papel muy importante en la convergencia del algoritmo de Jacobi, por lo cual se le dará un nombre y una notación especiales.

Definición 2

Sea el sistema lineal $\mathbf{x} = \mathbf{Mx} + \mathbf{c}$. Se llama *factor de convergencia* del método de Jacobi para este sistema a la norma de \mathbf{M} y se denotará como α , es decir:

$$\alpha = \|\mathbf{M}\|$$

■

De acuerdo con esta definición, la ecuación (5) toma la forma más simple:

$$E(\mathbf{x}^{(k)}) \leq \alpha E(\mathbf{x}^{(k-1)}) \quad k = 1, 2, 3, \dots \quad (6)$$

De la ecuación (6) se obtiene una consecuencia muy importante relacionada con la convergencia del algoritmo de Jacobi.

Teorema 1

Una condición suficiente para que el método de Jacobi converja hacia la solución del sistema $\mathbf{x} = \mathbf{Mx} + \mathbf{c}$ independientemente de la aproximación inicial $\mathbf{x}^{(0)}$ es que el factor de convergencia α , sea menor que 1.

Demostración:

Aplicando k veces la desigualdad (6):

$$E(\mathbf{x}^{(k)}) \leq \alpha E(\mathbf{x}^{(k-1)}) \leq \alpha^2 E(\mathbf{x}^{(k-2)}) \leq \dots \leq \alpha^k E(\mathbf{x}^{(0)})$$

Como $E(\mathbf{x}^{(0)})$ no depende de k , tomando límites para k tendiendo hacia infinito, resulta:

$$\lim_{k \rightarrow \infty} E(\mathbf{x}^{(k)}) \leq E(\mathbf{x}^{(0)}) \lim_{k \rightarrow \infty} \alpha^k = 0$$

por ser $\alpha < 1$. Esto significa que $\lim_{k \rightarrow \infty} E(\mathbf{x}^{(k)}) = 0$ de donde resulta que

$$\lim_{k \rightarrow \infty} \mathbf{x}^{(k)} = \mathbf{x}$$

como se quería demostrar.

■

De la ecuación (6) también se puede ver con claridad que la rapidez de convergencia del método de Jacobi depende del valor del parámetro α de convergencia. En efecto, la ecuación (6) se puede escribir en términos de error absoluto máximo como:

$$E_m(\mathbf{x}^{(k)}) = \alpha E_m(\mathbf{x}^{(k-1)})$$

de manera que se trata de un método de convergencia lineal. Valores de α próximos a cero garantizan una alta rapidez en la convergencia, valores menores que 1 pero próximos a él, producen una convergencia lenta. Nótese que el teorema 1 establece una condición suficiente para la convergencia; esta condición no es necesaria y, de hecho, se pueden hallar sistemas lineales con $\alpha > 1$ para los cuales el método de Jacobi converge. No es difícil demostrar que una condición necesaria y suficiente para la convergencia es que el mayor valor propio de la matriz \mathbf{M} sea menor que 1, sin embargo esta condición no posee gran valor práctico dado que hallar el mayor valor propio de \mathbf{M} es un problema, en general, más complicado que resolver el sistema lineal.

Ejemplo 3

Halle el factor de convergencia α de cada uno de los sistemas del ejemplo 1 y explique la convergencia del método de Jacobi en cada uno de ellos.

Solución:

a) La matriz \mathbf{M} del sistema es

$$\mathbf{M} = \begin{bmatrix} 0 & 0,1 & -0,2 \\ 0,2 & 0 & 0,2 \\ -0,25 & 0,125 & 0 \end{bmatrix}$$

su factor de convergencia se obtiene como:

$$\alpha = \|\mathbf{M}\| = \max\{0,3; 0,4; 0,375\} = 0,4$$

Como se ve, es próximo a cero. En este caso, el error absoluto máximo en cada iteración será menos de la mitad del de la iteración anterior y se obtendrá una alta velocidad de convergencia.

b) En este caso

$$\mathbf{M} = \begin{bmatrix} 0 & -0,8 & 0,7 \\ 0,3 & 0 & -0,1 \\ -0,5 & 0,1 & 0 \end{bmatrix}$$

y el parámetro de convergencia: $\alpha = \|\mathbf{M}\| = \max\{1,5; 0,4; 0,6\} = 1,5$

Como $\alpha > 1$, la convergencia no se puede garantizar por el teorema 1. En este caso el método de Jacobi converge, aunque la convergencia, como ya se había visto, es bastante lenta.

c) Para este sistema:

$$\mathbf{M} = \begin{bmatrix} 0 & 0,75 & -1,25 \\ -1,5 & 0 & -1 \\ -1,3333 & 0,6667 & 0 \end{bmatrix}$$

El parámetro de convergencia es: $\alpha = \|\mathbf{M}\| = \max\{2; 2,5; 2\} = 2,5$

Como $\alpha > 1$, el teorema 1 no garantiza que exista convergencia. Como se recordará del ejemplo 1, en este caso el método de Jacobi diverge. ■

El teorema que sigue es una consecuencia inmediata del anterior. Su importancia radica en que permite analizar la convergencia del método de Jacobi para un sistema lineal cuando este aun se encuentra en la forma $\mathbf{Ax} = \mathbf{b}$.

Teorema 2

Sea el sistema lineal de n ecuaciones con n incógnitas $\mathbf{Ax} = \mathbf{b}$. Una condición suficiente para que el método de Jacobi aplicado a dicho sistema, sea convergente, es que \mathbf{A} tenga la diagonal predominante, esto es, que para cada fila $i = 1, 2, 3, \dots, n$, el elemento de la diagonal sea, en valor absoluto, mayor que la suma de los valores absolutos de los otros elementos. Es decir que:

$$|a_{ii}| > |a_{i1}| + |a_{i2}| + \dots + |a_{i,i-1}| + |a_{i,i+1}| + \dots + |a_{in}| \quad \text{para } i = 1, 2, 3, \dots, n \quad (7)$$

Demostración:

Como

$$\mathbf{M} = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots & -\frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \dots & 0 \end{bmatrix}$$

entonces, si i es una fila cualquiera de \mathbf{A} la desigualdad (7) implica, dividiendo en cada uno de sus términos por $|a_{ii}|$, que:

$$\left| \frac{a_{i1}}{a_{ii}} \right| + \left| \frac{a_{i2}}{a_{ii}} \right| + \dots + \left| \frac{a_{i,i-1}}{a_{ii}} \right| + \left| \frac{a_{i,i+1}}{a_{ii}} \right| + \dots + \left| \frac{a_{in}}{a_{ii}} \right| < 1$$

lo cual significa que la suma de los valores absolutos de los elementos de la i -sima fila de \mathbf{M} es menor que 1. Como esto sucede para todas las filas de \mathbf{M} entonces la norma de \mathbf{M} será necesariamente menor que 1 y el algoritmo de Jacobi converge, según el teorema 1. ■

El error en el método de Jacobi

Al igual que en todos los demás métodos iterativos, se necesita una forma de acotar el error en cada iteración del método de Jacobi para poder detener el proceso iterativo cuando el error sea suficientemente pequeño. Para ello recuérdese que se definió:

$$E(\mathbf{x}^{(k)}) = \|\mathbf{x} - \mathbf{x}^{(k)}\|$$

por tanto,

$$E(\mathbf{x}^{(k-1)}) = \|\mathbf{x} - \mathbf{x}^{(k-1)}\|$$

Sumando y restando $\mathbf{x}^{(k)}$ a la diferencia del segundo miembro, se obtiene:

$$E(\mathbf{x}^{(k-1)}) = \|\mathbf{x} - \mathbf{x}^{(k)} + \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

Pero, según el axioma 3 de las normas, la norma de la suma de dos vectores es menor o igual a la suma de las normas de esos vectores, así que:

$$E(\mathbf{x}^{(k-1)}) \leq \|\mathbf{x} - \mathbf{x}^{(k)}\| + \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

El primer sumando del segundo miembro no es más que $E(\mathbf{x}^{(k)})$. Esto conduce a:

$$E(\mathbf{x}^{(k-1)}) \leq E(\mathbf{x}^{(k)}) + \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \quad (8)$$

El primer miembro de esta inecuación se puede modificar teniendo en cuenta la desigualdad (6),

$$E(\mathbf{x}^{(k)}) \leq \alpha E(\mathbf{x}^{(k-1)}) \text{ que implica } \frac{E(\mathbf{x}^{(k)})}{\alpha} \leq E(\mathbf{x}^{(k-1)})$$

Entonces la expresión (8) queda:

$$\frac{E(\mathbf{x}^{(k)})}{\alpha} \leq E(\mathbf{x}^{(k)}) + \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

Despejando de esta inecuación el error absoluto de $\mathbf{x}^{(k)}$, para lo cual se supondrá que $\alpha < 1$:

$$E(\mathbf{x}^{(k)}) \left(\frac{1}{\alpha} - 1 \right) \leq \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

$$E(\mathbf{x}^{(k)}) \leq \left(\frac{\alpha}{1-\alpha} \right) \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

Esto significa que el error absoluto máximo puede tomarse como el miembro de la derecha de esta desigualdad, es decir, que:

$$E_m(\mathbf{x}^{(k)}) = \left(\frac{\alpha}{1-\alpha} \right) \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \quad (9)$$

La fórmula (9) permite establecer una condición práctica para la terminación del proceso iterativo de Jacobi.

Condición de terminación:

Si se desea obtener la solución de un sistema lineal con un error absoluto menor que ε , y el factor de convergencia del método de Jacobi es $\alpha < 1$, entonces el proceso iterativo de Jacobi se llevará a cabo hasta la aproximación $\mathbf{x}^{(k)}$ para la cual:

$$E_m(\mathbf{x}^{(k)}) = \left(\frac{\alpha}{1-\alpha} \right) \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \varepsilon$$

En los casos en que α es menor que 0,5, la condición de terminación se puede simplificar un poco. En efecto, si $\alpha < 0,5$:

$$\frac{\alpha}{1-\alpha} < \frac{0,5}{1-\alpha} < \frac{0,5}{1-0,5} = 1$$

y entonces:

$$\left(\frac{\alpha}{1-\alpha} \right) \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

y puede tomarse como error absoluto máximo:

$$E_m(\mathbf{x}^{(k)}) = \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

De aquí resulta:

Condición de terminación:

Si se desea obtener la solución de un sistema lineal con un error absoluto menor que ε , y el factor de convergencia del método de Jacobi es $\alpha < 0,5$, entonces el proceso iterativo de Jacobi se llevará a cabo hasta la aproximación $\mathbf{x}^{(k)}$ para la cual:

$$E_m(\mathbf{x}^{(k)}) = \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \varepsilon$$

Ejemplo 4

Complete la tabla 1 del ejemplo 1 a) con una columna que muestre el error absoluto máximo para cada iteración. Indique en qué iteración debe detenerse el proceso si se desea obtener la solución con 3 cifras decimales exactas.

Solución:

En el ejemplo 3 se determinó que el factor de convergencia para este sistema es $\alpha = 0,4$. Por tanto, para cada iteración, el error absoluto máximo se puede tomar como:

$$E_m(\mathbf{x}^{(k)}) = \left(\frac{\alpha}{1-\alpha} \right) \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| = \frac{0,4}{1-0,4} \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| = (0,67) \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

En cada iteración, $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$ se obtiene como el mayor valor absoluto de las componentes del vector diferencia entre la iteración actual y la anterior. En la tabla 4 se muestra el error absoluto máximo para cada iteración. En cada iteración se ha marcado con * la componente donde ocurre la máxima diferencia con la iteración anterior.

Iteración	x_1	x_2	x_3	$E_m(\mathbf{x}^{(k)})$
0	0,0	0,0	0,0	-----
1	0,6	2,0*	-1,0	1,34
2	1,0*	1,92	-0,9	0,27
3	0,972	2,02	-1,01*	0,074
4	1,004*	1,9924	-0,9905	0,021
5	0,99734	2,0027	-1,00195*	0,0077
6	1,00066	1,999078*	-0,998997	0,0024
7	0,999707	2,000333	-1,00028*	0,00086
8	1,000089	1,999885*	-0,999885	0,00030
9	0,999966	2,000041*	-1,000037	0,00010

Tabla 4

Para obtener 3 cifras decimales exactas el error absoluto debe ser menor que 0,0005, por tanto, habría que tomar como solución la iteración 8.

Algoritmo del método de Jacobi

Se desea hallar una solución del sistema $\mathbf{Ax} = \mathbf{b}$ con error absoluto menor que ε . Se supone que \mathbf{A} tiene diagonal predominante y se conoce el valor del factor de convergencia α . El algoritmo utiliza como datos las matrices \mathbf{A} y \mathbf{b} , el factor de convergencia α , la tolerancia ε y la aproximación inicial $\mathbf{x}^{(0)}$. De no conocerse $\mathbf{x}^{(0)}$ se puede tomar el vector nulo $\mathbf{0}$.

```

xv := x(0) {El vector xv representa la solución vieja, es decir, la aproximación
obtenida en la iteración anterior, el vector xa representa la solución
nueva, es decir, la aproximación obtenida en la iteración actual}
repeat
    Error := 0
    for i = 1 to n
        xai := bi {En esta sección se calcula la i-sima componente de xa}
        for j := 1 to n
            if j ≠ i then
                xai := xai - aijxvj
            end
            xai :=  $\frac{x{a}_i}{a_{ii}}$ 
        end
        if |xai - xvi| > Error then {En este lazo se va guardando la máxima
diferencia entre las dos ultimas iteraciones}
            Error := |xai - xvi|
        end
    end
    xv := xa {La solución obtenida se transfiere a la vieja}
    Error := Error ·  $\left(\frac{\alpha}{1-\alpha}\right)$ 
until Error <  $\epsilon$ 
La solución del sistema es xa con error absoluto menor que Error
Terminar

```

Ejemplo 5

Resuelva con cuatro cifras decimales exactas el siguiente sistema de ecuaciones mediante el método de Jacobi.

$$9x_1 - x_2 + 2x_3 = 9$$

$$x_1 + 8x_2 + 2x_3 = 19$$

$$x_1 - x_2 + 11x_3 = 10$$

Solución:

Como la diagonal es predominante, se puede asegurar la convergencia del método de Jacobi. Para obtener el factor de convergencia se requiere hallar la matriz **M**:

$$\mathbf{M} = \begin{bmatrix} 0 & \frac{1}{9} & -\frac{2}{9} \\ -\frac{1}{8} & 0 & -\frac{1}{4} \\ -\frac{1}{11} & \frac{1}{11} & 0 \end{bmatrix}$$

de donde:

$$\alpha = \max\left\{\frac{1}{3}, \frac{3}{8}, \frac{2}{11}\right\} = \frac{3}{8} = 0,375$$

Por ser $\alpha < 0,5$, se puede tomar el error absoluto máximo como la norma de $\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$. Esto es, el proceso iterativo será detenido cuando

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq 0,00005$$

La tabla 5 muestra las aproximaciones obtenidas y el error absoluto máximo de cada una. Aparecen marcadas con * las componentes de la solución donde ocurre la máxima diferencia con la iteración anterior.

Iteración	x_1	x_2	x_3	$E_m(\mathbf{x}^{(k)})$
0	0,0	0,0	0,0	-----
1	1,0	2,375*	0,909091	2,375
2	1,061869	2,022727*	1,034091	0,352273
3	0,994949*	1,983744	0,996442	0,066919
4	0,998984	2,001521*	0,998981	0,017777
5	1,000395*	2,000382	1,000231	0,001411
6	0,999991	1,999893*	0,999999	0,000489
7	0,999988	2,000001*	0,999991	0,000108
8	1.000002*	2.000004	1.000001	0.000014

Tabla 5

La solución aproximada, con cuatro cifras decimales exactas, es:

$$\begin{aligned}x_1 &= 1,000002 \\x_2 &= 2,000004 \\x_3 &= 1,000001\end{aligned}$$

Ejemplo 6

Se desea obtener la solución del siguiente sistema tridiagonal con 100 ecuaciones y 100 incógnitas:

$$\begin{aligned}8x_1 - x_2 &= 10 \\x_1 + 8x_2 - x_3 &= 11 \\x_2 + 8x_3 - x_4 &= 12 \\x_3 + 8x_4 - x_5 &= 13 \\\vdots \\x_{98} + 8x_{99} - x_{100} &= 108 \\x_{99} + 8x_{100} &= 109\end{aligned}$$

Realice un algoritmo en seudo código para hallar la solución mediante el método de Jacobi con 5 cifras decimales exactas. Este ejemplo ilustra una de las ventajas más importantes de los métodos iterativos: grandes sistemas de ecuaciones donde todas las ecuaciones poseen una misma estructura general.

Solución:

Se trata de un sistema con diagonal predominante. Es más, el factor de convergencia se puede calcular rápidamente como $\alpha = 0,25$. Salvo la primera y la ultima ecuación, todas presentan una estructura similar:

$$x_{i-1} + 8x_i - x_{i+1} = i + 9 \quad i = 2, 3, 4, \dots, 99$$

Despejando x_i en cada una de las ecuaciones, el sistema sería:

$$\begin{aligned} x_1 &= \frac{10+x_2}{8} \\ x_i &= \frac{i+9-x_{i-1}+x_{i+1}}{8} \quad \text{para } i = 2, 3, 4, \dots, 99 \\ x_{100} &= \frac{109-x_{99}}{8} \end{aligned}$$

Por ser $\alpha = 0,25$, se tomará el error absoluto máximo de cada aproximación como:

$$E_m(\mathbf{x}^{(k)}) = \left(\frac{0,25}{1-0,25} \right) \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| = (0,34) \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

Como la estructura del sistema de ecuaciones es especial, es ventajoso utilizar este hecho para no tener que guardar en memoria las matrices \mathbf{A} y \mathbf{b} . Por esta razón el algoritmo que se usará no es el mismo que el algoritmo general elaborado antes. El algoritmo en seudo código es:

```

xv := x(0)
repeat
     $xa_1 = \frac{10+xv_2}{8}$ 
     $Error := |xa_1 - xv_1|$ 
    for  $i = 2$  to 99
         $xa_i = \frac{i+9-xv_{i-1}+xv_{i+1}}{8}$ 
        if  $|xa_i - xv_i| > Error$  then
             $Error := |xa_i - xv_i|$ 
        end
    end
     $xa_{100} = \frac{109-xv_{99}}{8}$ 
    if  $|xa_{100} - xv_{100}| > Error$  then
         $Error := |xa_{100} - xv_{100}|$ 
    end
    xv := xa {La solución obtenida se transfiere a la vieja}
     $Error := (0,34)Error$ 
until  $Error < 0,000005$ 
La solución del sistema es xa con error absoluto menor que Error
Terminar

```

El método de Seidel

El método de Seidel, también llamado de Gauss – Seidel, es una variación del método de Jacobi que logra simplificar dicho algoritmo y mejorar la rapidez de la convergencia en la mayoría de los casos.

En el método de Jacobi, al calcular la variable x_i se utilizan los valores de las demás variables que se obtuvieron en la iteración anterior, sin embargo, en ese momento ya se han calculado los nuevos valores de x_1, x_2, \dots, x_{i-1} que son, por lo general, mejores aproximaciones que los obtenidos en la iteración anterior. En el método de Seidel, una vez que se obtiene el nuevo valor de una variable, éste se utiliza para actualizar los valores de las variables que siguen; de esta forma, no se necesita guardar los valores de la iteración anterior, lo cual simplifica el algoritmo y ahorra memoria y, por otra parte, la velocidad de la convergencia mejora sustancialmente.

En términos más precisos, sea $\mathbf{Ax} = \mathbf{b}$ un sistema de n ecuaciones lineales. Se supone, como hasta ahora, que este sistema es cuadrado y que posee diagonal predominante. Escrito en forma desarrollada, el sistema es:

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

Como los elementos a_{ii} ($i = 1, 2, \dots, n$) de la diagonal son no nulos, el sistema, igual que antes, se puede escribir así:

$$\begin{aligned} x_1 &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2 - \cdots - \frac{a_{1n}}{a_{11}}x_n \\ x_2 &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1 - \frac{a_{23}}{a_{22}}x_3 - \cdots - \frac{a_{2n}}{a_{22}}x_n \\ &\vdots \\ x_n &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1 - \frac{a_{n2}}{a_{nn}}x_2 - \cdots - \frac{a_{n,n-1}}{a_{nn}}x_{n-1} \end{aligned}$$

El proceso iterativo de Seidel queda entonces definido de la siguiente forma:

$$\begin{aligned} x_1^{(k)} &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2^{(k-1)} - \frac{a_{13}}{a_{11}}x_3^{(k-1)} - \cdots - \frac{a_{1n}}{a_{11}}x_n^{(k-1)} \\ x_2^{(k)} &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1^{(k)} - \frac{a_{23}}{a_{22}}x_3^{(k-1)} - \cdots - \frac{a_{2n}}{a_{22}}x_n^{(k-1)} \\ x_3^{(k)} &= \frac{b_3}{a_{33}} - \frac{a_{31}}{a_{33}}x_1^{(k)} - \frac{a_{32}}{a_{33}}x_2^{(k)} - \cdots - \frac{a_{3n}}{a_{33}}x_n^{(k-1)} \\ &\vdots \\ x_n^{(k)} &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1^{(k)} - \frac{a_{n2}}{a_{nn}}x_2^{(k)} - \cdots - \frac{a_{n,n-1}}{a_{nn}}x_{n-1}^{(k)} \end{aligned} \tag{10}$$

Aunque el método de Seidel se puede expresar en forma matricial, no se gana mucho en este caso, pues las matrices que aparecen ya no son tan simples.

Ejemplo 7

Los siguientes sistemas de ecuaciones fueron resueltos por el método de Jacobi en el ejemplo 1. Revuélvalos por el método de Seidel.

$$\begin{array}{l} \text{a) } \begin{aligned} 10x_1 - x_2 + 2x_3 &= 6 \\ x_1 - 5x_2 + x_3 &= -10 \\ 2x_1 - x_2 + 8x_3 &= -8 \end{aligned} \quad \begin{array}{l} \text{b) } \begin{aligned} 10x_1 + 8x_2 - 7x_3 &= -3 \\ -3x_1 + 10x_2 + x_3 &= -25 \\ 5x_1 - x_2 + 10x_3 &= 22 \end{aligned} \quad \begin{array}{l} \text{c) } \begin{aligned} 4x_1 - 3x_2 + 5x_3 &= 25 \\ 3x_1 + 2x_2 + 2x_3 &= 11 \\ 4x_1 - 2x_2 + 3x_3 &= -4 \end{aligned} \end{array} \end{array}$$

Solución:

Aplicando el algoritmo de Seidel se obtiene los resultados que muestran respectivamente las tablas 6, 7 y 8 para los sistemas a), b) y c). Al igual que sucedió en el ejemplo 1 para el algoritmo de Jacobi, el proceso iterativo converge en los sistemas a) y b) y diverge en el c).

Iteración	x_1	x_2	x_3
0	0,0	0,0	0,0
1	0,6	2,12	-0,885
2	0,989	2,0208	-0,99465
3	1,00101	2,001272	-1,000094
4	1,000146	2,000010	-1,000035
5	1,000008	1,999995	-1,000003

Tabla 6

Iteración	x_1	x_2	x_3
0	0,0	0,0	0,0
1	-0,3	-2,59	2,091
2	3,2357	-1,73839	0,408311
3	1,37653	-2,127872	1,298948
4	2,311561	-1,936426	0,850577
5	1,844545	-2,031694	1,074558
⋮	⋮	⋮	⋮
10	2,004801	-1,999021	0,997698
⋮	⋮	⋮	⋮
15	1,999852	-2,000074	1,000071
16	2,000074	-1,999985	0,999965
17	1,999963	-2,000008	1,000018

Tabla 7

Iteración	x_1	x_2	x_3
0	0,0	0,0	0,0
1	6,25	-3,875	4,4167
2	-2,1771	4,3490	-1,3368
3	11,1827	-9,9373	6,9521
4	-9,8931	13,3875	-5,5991
5	23,2895	-23,8352	13,8293
6	-28,9130	35,0402	-16,5238
7	53,1849	-57,7536	31,0775
8	-75,9121	88,2906	-43,6890

Tabla 8

Nótese, sin embargo, que el método de Seidel ha requerido menos iteraciones para llegar a la solución de a) y b) con una exactitud similar.

Convergencia del método de Seidel

Considérese la i -sima ecuación del algoritmo de Seidel:

$$x_i^{(k)} = \frac{b_i}{a_{ii}} - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(k)} - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} x_j^{(k-1)} \quad i = 1, 2, \dots, n \quad (11)$$

La i -sima ecuación del sistema se puede escribir, agrupando convenientemente, como:

$$x_i = \frac{b_i}{a_{ii}} - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} x_j \quad i = 1, 2, \dots, n \quad (12)$$

Restando miembro a miembro las ecuaciones (12) menos (11), se obtiene:

$$x_i - x_i^{(k)} = - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} (x_j - x_j^{(k)}) - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} (x_j - x_j^{(k-1)})$$

Tomando módulos en ambos miembros de esta igualdad y recordando que el módulo de una suma es menor o igual que la suma de los módulos, resulta:

$$|x_i - x_i^{(k)}| \leq \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \cdot |x_j - x_j^{(k)}| + \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \cdot |x_j - x_j^{(k-1)}| \quad (13)$$

Si en la ecuación (13) los términos $|x_j - x_j^{(k)}|$ y $|x_j - x_j^{(k-1)}|$ se sustituyen respectivamente por $\|\mathbf{x} - \mathbf{x}^{(k)}\| = E(\mathbf{x}^{(k)})$ y $\|\mathbf{x} - \mathbf{x}^{(k-1)}\| = E(\mathbf{x}^{(k-1)})$ la desigualdad se cumple con más razón pues las normas son los máximos valores que toman $|x_j - x_j^{(k)}|$ y $|x_j - x_j^{(k-1)}|$ para las diferentes j . Se obtiene:

$$|x_i - x_i^{(k)}| \leq \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \cdot E(\mathbf{x}^{(k)}) + \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \cdot E(\mathbf{x}^{(k-1)}) \quad (14)$$

Los errores absolutos, como son independientes de j se pueden extraer como factores comunes en las sumas respectivas, lo cual conduce a:

$$|x_i - x_i^{(k)}| \leq E(\mathbf{x}^{(k)}) \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| + E(\mathbf{x}^{(k-1)}) \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \quad (15)$$

$$\text{Conviene ahora definir, } p_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \quad \text{y} \quad q_i = \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \quad i = 1, 2, \dots, n \quad (16)$$

Obsérvese que p_i es la suma de los módulos de los elementos de la fila i de la matriz \mathbf{M} , sumando desde el primer elemento hasta el que antecede la diagonal, mientras que q_i es la suma desde el elemento después de la diagonal hasta el final de la fila. La desigualdad (15) queda ahora:

$$|x_i - x_i^{(k)}| \leq E(\mathbf{x}^{(k)}) p_i + E(\mathbf{x}^{(k-1)}) q_i \quad i = 1, 2, \dots, n \quad (17)$$

Como la desigualdad (17) es cierta para todas las ecuaciones $i = 1, 2, \dots, n$ del sistema, será también cierta, en particular, para aquella, llamémosla m , en la cual es máximo $|x_i^{(k)} - x_i|$.

Entonces para $i = m$ el primer miembro de la desigualdad es $E(\mathbf{x}^{(k)})$ y se tiene:

$$E(\mathbf{x}^{(k)}) \leq E(\mathbf{x}^{(k)}) p_m + E(\mathbf{x}^{(k-1)}) q_m$$

$$\text{Despejando } E(\mathbf{x}^{(k)}): \quad E(\mathbf{x}^{(k)}) \leq \left(\frac{q_m}{1 - p_m} \right) E(\mathbf{x}^{(k-1)}) \quad (18)$$

supuesto que $p_m < 1$, lo cual está garantizado, por ejemplo, si el sistema tiene diagonal predominante.

El valor de m puede variar de una iteración a otra. Sin embargo, si se define:

$$\beta = \max_i \frac{q_i}{1 - p_i}$$

la desigualdad (18) conduce a:

$$E(\mathbf{x}^{(k)}) \leq \beta \cdot E(\mathbf{x}^{(k-1)}) \quad (19)$$

Si se supone que la diagonal de \mathbf{A} es predominante, entonces

$$p_i + q_i < 1 \quad i = 1, 2, 3, \dots, n$$

esto es:

$$q_i < 1 - p_i$$

$$\text{y} \quad \frac{q_i}{1-p_i} < 1 \quad i = 1, 2, 3, \dots, n$$

en ese caso el parámetro $\beta = \max_i \frac{q_i}{1-p_i}$ será menor que 1, lo cual prueba que el proceso iterativo converge hacia la solución del sistema.

El parámetro β juega en el método de Seidel un papel análogo al α del método de Jacobi. Por su importancia, la siguiente definición formaliza este concepto.

Definición 3

Sea el sistema lineal $\mathbf{Ax} = \mathbf{b}$ con diagonal predominante. Se llama *factor de convergencia* del método de Seidel para este sistema al número β definido como:

$$\beta = \max_i \frac{q_i}{1-p_i}$$

donde $p_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right|$ y $q_i = \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \quad i = 1, 2, \dots, n$ ■

Como las fórmulas (19) para el método de Seidel y (6) para el método de Jacobi tienen idéntica estructura con el único cambio de α por β , no es necesario repetir para el método de Seidel todas las deducciones realizadas en el método de Jacobi. A continuación se resumen los resultados fundamentales:

Teorema 3

Una condición suficiente para que el método de Seidel converja hacia la solución del sistema $\mathbf{Ax} = \mathbf{b}$ independientemente de la aproximación inicial $\mathbf{x}^{(0)}$ es que el sistema tenga diagonal predominante.

El error absoluto máximo en la iteración k -sima del algoritmo de Seidel puede tomarse como:

$$E_m(\mathbf{x}^{(k)}) = \left(\frac{\beta}{1-\beta} \right) \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \quad (20)$$

y de aquí resulta la condición de terminación:

Condición de terminación:

Si se desea obtener la solución de un sistema lineal con un error absoluto menor que ε , y el factor de convergencia del método de Seidel es $\beta < 1$, entonces el proceso iterativo de Seidel se llevará a cabo hasta la aproximación $\mathbf{x}^{(k)}$ para la cual:

$$E_m(\mathbf{x}^{(k)}) = \left(\frac{\beta}{1-\beta} \right) \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \varepsilon$$

En los casos en que β es menor que 0,5, se tiene:

$$\frac{\beta}{1-\beta} < 1$$

y la fórmula (20) implica:

$$E_m(\mathbf{x}^{(k)}) = \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

De aquí resulta:

Condición de terminación:

Si se desea obtener la solución de un sistema lineal con un error absoluto menor que ε , y el factor de convergencia del método de Seidel es $\beta < 0,5$, entonces el proceso iterativo de Seidel se llevará a cabo hasta la aproximación $\mathbf{x}^{(k)}$ para la cual:

$$E_m(\mathbf{x}^{(k)}) = \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \varepsilon$$

La fórmula (19) puede ser escrita en términos del error absoluto máximo:

$$E_m(\mathbf{x}^{(k)}) = \beta \cdot E_m(\mathbf{x}^{(k-1)}) \quad (21)$$

La cual deja ver que el método de Seidel posee convergencia lineal y que valores pequeños de β (próximos a cero) aseguran una convergencia rápida hacia la solución, mientras que valores próximos y menores que 1, dan lugar a una convergencia más lenta. Cuando el factor de convergencia es mayor que 1, no se asegura la convergencia, aunque puede suceder.

Comparación entre la convergencia de los métodos de Jacobi y de Seidel

Se pueden encontrar sistemas en los cuales el método de Seidel converge y el de Jacobi no y también a la inversa, sistemas en los que el método de Seidel diverge y converge el de Jacobi. Sin

embargo, cuando la matriz \mathbf{A} tiene diagonal predominante, de modo que $\alpha < 1$, la convergencia del método de Jacobi nunca es más rápida que la de Seidel. En efecto, nótese que:

$$p_i + q_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| + \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| = \sum_{j=1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \quad \text{para } i = 1, 2, \dots, n$$

y como α es el máximo valor que alcanza la última suma, se tiene que:

$$p_i + q_i \leq \alpha \quad \text{para } i = 1, 2, \dots, n$$

es decir;

$$q_i \leq \alpha - p_i$$

$$\text{Entonces, } \frac{q_i}{1-p_i} \leq \frac{\alpha - p_i}{1-p_i} \leq \frac{\alpha - \alpha p_i}{1-p_i} = \frac{\alpha(1-p_i)}{1-p_i} = \alpha$$

y como β es el máximo valor del primer miembro, se tiene que:

$$\text{Si } \alpha < 1 \text{ entonces } \beta \leq \alpha \quad (22)$$

Lo cual prueba que en los casos de diagonal predominante, el método de Seidel tiene una convergencia más rápida o igual que el de Jacobi. En la práctica la diferencia es bastante apreciable y, en casi todos los casos de diagonal predominante, el método de Seidel necesita aproximadamente la mitad de las iteraciones que requiere el método de Jacobi.

Como el parámetro β puede ser más complicado de calcular que α , en ocasiones, cuando la diagonal es predominante, se suele usar α en lugar de β al trabajar con el método de Seidel y esto no conduce a errores precisamente porque en estos casos $\beta \leq \alpha$.

Ejemplo 8

Determine el parámetro β para el sistema lineal a) del ejemplo 7. Compare con el valor de α obtenido.

$$\begin{aligned} 10x_1 - x_2 + 2x_3 &= 6 \\ x_1 - 5x_2 + x_3 &= -10 \\ 2x_1 - x_2 + 8x_3 &= -8 \end{aligned}$$

Solución:

$$\text{La matriz } \mathbf{M} \text{ del sistema es} \quad \mathbf{M} = \begin{bmatrix} 0 & 0,1 & -0,2 \\ 0,2 & 0 & 0,2 \\ -0,25 & 0,125 & 0 \end{bmatrix}$$

$$p_1 = 0 \quad q_1 = 0,1 + 0,2 = 0,3 \quad \frac{q_1}{1-p_1} = \frac{0,3}{1-0} = 0,3$$

$$p_2 = 0,2 \quad q_2 = 0,2 \quad \frac{q_2}{1-p_2} = \frac{0,2}{1-0,2} = 0,25$$

$$p_3 = 0,25 + 0,125 = 0,375 \quad q_3 = 0 \quad \frac{q_3}{1 - p_3} = \frac{0}{1 - 0,375} = 0$$

y como

$$\beta = \max_i \frac{q_i}{1 - p_i}$$

se tiene

$$\beta = \max\{0,3; 0,25; 0\} = 0,3$$

En este mismo sistema, el parámetro de convergencia de Jacobi es $\alpha = 0,4$. ■

Algoritmo del método de Seidel

Se desea hallar una solución del sistema $\mathbf{Ax} = \mathbf{b}$ con error absoluto menor que ε . Se supone que \mathbf{A} tiene diagonal predominante y se conoce el valor del factor de convergencia β . El algoritmo utiliza como datos las matrices \mathbf{A} y \mathbf{b} , el factor de convergencia β , la tolerancia ε y la aproximación inicial $\mathbf{x}^{(0)}$. De no conocerse $\mathbf{x}^{(0)}$ se puede tomar el vector nulo $\mathbf{0}$.

```

x := x(0)
repeat
    Error := 0
    for i = 1 to n
        prov := bi {En esta sección se calcula la i-sima componente de x,
        que se almacena provisionalmente en prov, de modo que
        no se afecte xi para poder determinar la diferencia entre
        la iteración actual y la anterior}
        for j := 1 to n
            if j ≠ i then
                prov := prov - aijxj
            end
            prov := prov / aii
        end
        if |prov - xi| > Error then {En este lazo se va guardando la máxima
        diferencia entre las dos ultimas iteraciones}
            Error := |prov - xi|
        end
        xi := prov {El antiguo valor de xi se cambia por el nuevo}
    end

    Error := Error · (β / (1 - β))

until Error < ε
La solución del sistema es x con error absoluto menor que Error
Terminar

```

Ejemplo 9

Resuelva con cuatro cifras decimales exactas el siguiente sistema de ecuaciones mediante el método de Seidel. El sistema ya fue resuelto por el método de Jacobi en el ejemplo 5.

$$\begin{aligned} 9x_1 - x_2 + 2x_3 &= 9 \\ x_1 + 8x_2 + 2x_3 &= 19 \\ x_1 - x_2 + 11x_3 &= 10 \end{aligned}$$

Solución:

Como la diagonal es predominante, se puede asegurar la convergencia del método de Seidel. Para obtener el factor de convergencia se requiere hallar la matriz \mathbf{M} :

$$\mathbf{M} = \begin{bmatrix} 0 & \frac{1}{9} & -\frac{2}{9} \\ -\frac{1}{8} & 0 & -\frac{1}{4} \\ -\frac{1}{11} & \frac{1}{11} & 0 \end{bmatrix}$$

$$\begin{array}{lll} p_1 = 0 & q_1 = 0,33 & \frac{q_1}{1-p_1} = \frac{0,33}{1-0} = 0,33 \\ p_2 = 0,125 & q_2 = 0,25 & \frac{q_2}{1-p_2} = \frac{0,25}{1-0,125} = 0,29 \\ p_3 = 0,182 & q_3 = 0 & \frac{q_3}{1-p_3} = \frac{0}{1-0,182} = 0 \end{array}$$

de donde:

$$\beta = \max\{0,33; 0,29; 0\} = 0,33$$

Por ser $\beta < 0,5$, se puede tomar el error absoluto máximo como la norma de $\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$. Esto es, el proceso iterativo será detenido cuando

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq 0,00005$$

La tabla 9 muestra las aproximaciones obtenidas y el error absoluto máximo de cada una.

Aparecen marcadas con * las componentes de la solución donde ocurre la máxima diferencia con la iteración anterior.

Iteración	x_1	x_2	x_3	$E_m(\mathbf{x}^{(k)})$
0	0,0	0,0	0,0	-----
1	1,0	2,25*	1,022727	2,25
2	1,022727	1,991477*	0,997159	0,258523
3	0,999684*	2,000750	1,000097	0,023043
4	1,000062	1,999968*	0,999991	0,000782
5	0,999998*	2,000002	1,000000	0,000063
6	1,000000	2,000000*	1,000000	0,000002

Tabla 9

La solución aproximada, con cinco cifras decimales exactas, es:

$$\begin{aligned}x_1 &= 1,000000 \\x_2 &= 2,000000 \\x_3 &= 1,000000\end{aligned}$$

Aunque la solución se pidió con cuatro cifras decimales exactas, fueron obtenidas cinco, ya que en la quinta iteración todavía el error era mayor que 0,00005 pero en la sexta iteración está por debajo de 0,000005. Como en este problema se sabe que la solución exacta es $x_1 = 1$; $x_2 = 2$; $x_3 = 1$ puede apreciarse que ya desde la quinta iteración se tenían 5 cifras decimales exactas, pero la cota del error, que es lo que permite detener el proceso, siempre es mayor que el verdadero error absoluto.

Ejemplo 10

En el ejemplo 2 del comienzo del capítulo, se llegó al sistema de 16 ecuaciones con 16 incógnitas:

$$4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = 0$$

para $i = 1, 2, 3, 4$; $j = 1, 2, 3, 4$

donde u_{ij} representa la temperatura en el punto (i, j) de una red definida sobre una lámina metálica cuadrada sometida en sus bordes a temperaturas fijas. Como las temperaturas en el contorno son conocidas, se sabe que:

$$\begin{aligned}u_{01} &= u_{02} = u_{03} = u_{04} = 20 \\u_{51} &= u_{52} = u_{53} = u_{54} = 60 \\u_{10} &= u_{20} = u_{30} = u_{40} = 80 \\u_{15} &= u_{25} = u_{35} = u_{45} = 40\end{aligned}$$

Elabore un algoritmo en seudo código basado en el método de Seidel, que permita obtener la solución del sistema con tres cifras decimales exactas.

Solución:

Si en cada ecuación del sistema se coloca en la diagonal la variable con coeficiente 4, es claro que la matriz del sistema (de 16 por 16) tiene la diagonal casi predominante, pues la suma de los elementos (en valor absoluto) que no están en la diagonal es exactamente 4. Aunque en este caso, la convergencia no se puede garantizar, es muy probable que el algoritmo converja ya que el parámetro α es exactamente 1 en lugar de la condición suficiente que es $\alpha < 1$. En estos casos, el algoritmo se utiliza a riesgo y se toma como condición de terminación $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \varepsilon$ con una tolerancia mucho menor que la que se desea. Como en el problema se piden tres cifras decimales exactas, se tomará una tolerancia $\varepsilon = 0,00005$, diez veces menor que la necesaria.

En el algoritmo que sigue, se emplea la variable provisional *prov* para guardar el nuevo valor de cada incógnita sin borrar el de la iteración anterior, de manera que pueda ser hallada la diferencia entre ambas. La variable *Error* va actualizándose con la máxima diferencia entre las dos iteraciones hasta el momento en que se calcula la nueva u_{ij} .

```

for  $k = 1$  to 4           {Aquí se fijan las condiciones de frontera}
     $u_{0k} := 20$ 
     $u_{5k} := 60$ 
     $u_{k0} := 80$ 
     $u_{k5} := 60$ 
end
for  $i = 1$  to 4           {Todas las incógnitas se inicializan en cero}
    for  $j = 1$  to 4
         $u_{ij} := 0$ 
    end
end
repeat
     $Error := 0$ 
    for  $i = 1$  to 4
        for  $j = 1$  to 4
             $prov := (u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1})/4$ 
            if  $|prov - u_{ij}| > Error$  then
                 $Error := |prov - u_{ij}|$ 
            end
             $u_{ij} := prov$ 
        end
    end
until  $Error < 0,00005$ 
La solución es  $u_{ij}$  ( $i = 1, 2, 3, 4$ ;  $j = 1, 2, 3, 4$ )
Terminar

```

El algoritmo anterior converge en unas 30 iteraciones con una baja rapidez debido a que la diagonal no es predominante.

Comentarios finales sobre los métodos iterativos

A diferencia de los métodos directos, en que la cantidad de operaciones aritméticas requeridas está determinada únicamente por el orden del sistema, en los métodos iterativos hay que tener en cuenta la cantidad de iteraciones que se necesitará; esta depende de la exactitud deseada y de la

rapidez de convergencia del método que se emplee. En la demostración del teorema 1 sobre la convergencia del método de Jacobi se obtuvo una fórmula que puede utilizarse para decidir:

$$E(\mathbf{x}^{(k)}) \leq \alpha^k E(\mathbf{x}^{(0)})$$

que se puede escribir en términos de errores absolutos máximos como:

$$E_m(\mathbf{x}^{(k)}) = \alpha^k E_m(\mathbf{x}^{(0)}) \quad (23)$$

Para el método de Seidel, análogamente, se tiene:

$$E_m(\mathbf{x}^{(k)}) = \beta^k E_m(\mathbf{x}^{(0)}) \quad (24)$$

Estas fórmulas permiten predecir cuantas iteraciones se necesitará para obtener una exactitud deseada si se puede hacer una estimación del error absoluto máximo en la iteración inicial, lo cual requiere tener una idea aproximada de los valores que tomará la solución. Por ejemplo, si se desea la solución de un sistema con tolerancia 0,00005 (cuatro cifras decimales exactas), se toma la aproximación inicial como 0 y se sabe que la solución debe tener todas sus componentes por debajo de 5, entonces $E_m(\mathbf{x}^{(0)}) = 5$ y $E_m(\mathbf{x}^{(k)}) = 0,00005$. Si se selecciona el método de Seidel, entonces la cantidad de iteraciones que se requerirá se puede despejar la fórmula (24):

$$k = \frac{\ln \left[\frac{E_m(\mathbf{x}^{(k)})}{E_m(\mathbf{x}^{(0)})} \right]}{\ln \beta}$$

Si, en este mismo ejemplo $\beta = 0,3$ entonces harán falta:

$$k = \frac{\ln \left[\frac{0,00005}{5} \right]}{\ln 0,3} = 9,56$$

es decir, unas 10 iteraciones. Sin embargo, si $\beta = 0,9$ entonces el número de iteraciones se incrementa a:

$$k = \frac{\ln \left[\frac{0,00005}{5} \right]}{\ln 0,9} = 109,27$$

o sea 110 iteraciones. Tanto en el método de Jacobi como en el de Seidel, se requieren n^2 operaciones de multiplicar y dividir en cada iteración (n operaciones para cada ecuación y son n ecuaciones), así que multiplicando se obtiene la cantidad de iteraciones necesarias.

Sin embargo, el único factor para decidir si se aplica un método directo o uno iterativo no es la cantidad de operaciones. Si el sistema tiene muchas ecuaciones, todas con la misma estructura, como en los ejemplos 6 y 10, seguramente es preferible escribir un pequeño programa que implemente el método de Seidel o el Jacobi, que tener que escribir todos los coeficientes de la

matriz \mathbf{A} (que pueden ser decenas de miles) para utilizar un programa profesional de alguna biblioteca numérica con un método directo.

Como los métodos iterativos solo se utilizan para sistemas con diagonal predominante, muchas veces no hay nada que decidir, porque aunque teóricamente siempre es posible hacer transformaciones elementales en el sistema que hagan predominante la diagonal, esto puede resultar prácticamente imposible en sistemas de más de tres o cuatro ecuaciones.

En cuanto a la selección entre Seidel y Jacobi, el método de Seidel converge mucho más rápido, es más fácil de programar y requiere la mitad de la memoria. Solo en una cosa lleva ventaja el método de Jacobi: si se va a implementar el algoritmo en una máquina de cálculo paralelo, el algoritmo de Jacobi permite calcular en cada iteración todas las incógnitas (si la máquina posee esta posibilidad) simultáneamente, ya que cada una requiere los mismos datos y es independiente de la otra; el método de Seidel es, por naturaleza, un método secuencial, ya que para calcular una incógnita en una iteración hay que haber calculado todas las que le preceden.

Ejemplo 11

En un sistema lineal con diagonal predominante, se puede hallar una estimación inicial de la solución de tal modo que $E_m(\mathbf{x}^{(0)}) = 1$. Se necesita la solución con cuatro cifras decimales exactas y los parámetros de convergencia son $\alpha = 0,6$ y $\beta = 0,45$. Si el sistema tiene 70 ecuaciones con 70 incógnitas, calcule cuantas operaciones se necesitarán utilizando el método de Gauss, el de Jacobi y el de Seidel.

Solución:

El método de Gauss requeriría: $\frac{1}{3}n^3 = \frac{1}{3}70^3 = 114334$ operaciones

El método de Jacobi, necesita realizar

$$k = \frac{\ln \left[\frac{E_m(\mathbf{x}^{(k)})}{E_m(\mathbf{x}^{(0)})} \right]}{\ln \alpha} = \frac{\ln \left[\frac{0,00005}{1} \right]}{\ln 0,6} = 19,39 \text{ iteraciones}$$

es decir, 20. En cada iteración realizará $70^2 = 4900$ operaciones. Por tanto necesitará

$$20 \cdot 4900 = 98000 \text{ operaciones}$$

El método de Seidel, requiere:

$$k = \frac{\ln \left[\frac{E_m(\mathbf{x}^{(k)})}{E_m(\mathbf{x}^{(0)})} \right]}{\ln \beta} = \frac{\ln \left[\frac{0,00005}{1} \right]}{\ln 0,45} = 12,40$$

o sea, 13 iteraciones. A razón de 4900 operaciones por cada iteración, un total de

$$13 \cdot 4900 = 63700 \text{ operaciones}$$

En este ejemplo, el método de Seidel lleva la ventaja, aunque la diferencia no es realmente significativa.

Ejercicios

En los siguientes ejercicios, utilice programas computacionales, preferiblemente confeccionados por usted. De no contar con un programa adecuado, use una calculadora y trabaje con cinco cifras significativas en los cálculos.

- Para los siguientes sistemas de ecuaciones, determine si la diagonal es predominante, calcule el valor del factor α de convergencia del método de Jacobi y, si $\alpha < 1$, determine el valor del factor de convergencia β del método de Seidel. De cumplirse las condiciones suficientes de convergencia, establezca el criterio de terminación para obtener la solución con cuatro cifras decimales exactas si se utilizara cada uno de estos métodos.

$$\begin{array}{ll}
 \text{a)} & \begin{aligned} 10x_1 - 2x_2 + 3x_3 + 2x_4 &= 15 \\ 3x_1 - 10x_2 - 4x_3 + 2x_4 &= 4 \\ 5x_1 + 3x_2 + 10x_3 - x_4 &= 13 \\ 4x_1 - 3x_2 - 4x_3 + 10x_4 &= -2 \end{aligned} \\
 \text{b)} & \begin{aligned} 10x_1 + x_2 + 2x_3 - x_4 &= 3 \\ -2x_1 + 10x_2 + x_3 - x_4 &= 4 \\ 2x_1 + 3x_2 - 10x_3 - 2x_4 &= -7 \\ 3x_1 - 2x_2 + 4x_3 + 10x_4 &= 4 \end{aligned} \\
 \text{c)} & \begin{aligned} 10x_1 + 2x_2 - 3x_3 + 4x_4 &= 1 \\ -x_1 + 10x_2 + 3x_3 + 2x_4 &= -3 \\ 2x_1 - 3x_2 - 10x_3 + 2x_4 &= 2 \\ x_1 - 2x_2 - 4x_3 + 10x_4 &= -1 \end{aligned}
 \end{array}$$

- Escriba los siguientes sistemas lineales de manera que se pueda garantizar la convergencia de los métodos de Jacobi y de Seidel y calcule los valores de los parámetros α y β .

$$\begin{array}{lll}
 \text{a)} & \begin{aligned} x + 3y - 9z &= 3 \\ 5x - y + 2z &= 25 \\ 4x - 15y + z &= -7 \end{aligned} & \begin{aligned} 9z + 3y - x &= 4 \\ 8x - y + 2z &= 14 \\ 3x - 7y - z &= 5 \end{aligned} & \begin{aligned} -x + 4y - 2z &= 5 \\ 3x + 2y + 9z &= 25 \\ 5x + 2y + z &= 4 \end{aligned}
 \end{array}$$

- Intente resolver los tres sistemas del ejercicio 1 (a pesar de que el primero no tiene la diagonal predominante) por los métodos de Jacobi y de Seidel. Compare la rapidez con que se obtiene la solución con 5 cifras decimales exactas para cada sistema y para cada método y analice su relación con los respectivos factores de convergencia.
- Elabore un algoritmo en seudo código para resolver el siguiente sistema lineal mediante el método de Seidel con cuatro cifras decimales exactas.

$$8u_{ij} - u_{i-1,j} - 2u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = 0 \quad i = 1, 2, 3; \quad j = 1, 2, 3$$

Se sabe que:

$u_{01} = u_{02} = u_{03} = 25$
$u_{41} = u_{42} = u_{43} = -3$
$u_{10} = u_{20} = u_{30} = 18$
$u_{14} = u_{24} = u_{34} = 12$

- Las x_{ij} de la región de la figura 3 satisfacen la ecuación

$$4u_{ij} = u_{i-1,j} + u_{i+1,j} + \frac{u_{i,j-1} + u_{i,j+1}}{2}$$

siempre que (i, j) corresponda a un punto interior (no en la frontera) de la región triangular. Los valores de u_{ij} para puntos (i, j) de la frontera se muestran en la propia figura. Elabore un algoritmo en seudo código basado en el método de Seidel para obtener la solución con cuatro cifras decimales exactas.

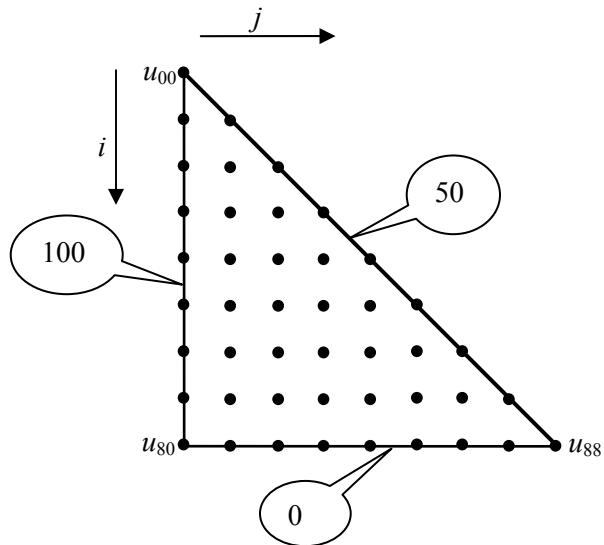


Figura 3

6. Se necesita resolver un sistema con diagonal predominante de n ecuaciones y n incógnitas con 4 cifras decimales exactas. Se puede hallar una aproximación inicial con error absoluto menor que 0,5. Determine la cantidad de operaciones que se necesitaría para hallar la solución mediante el método de Gauss y mediante el método iterativo de Seidel si el valor de n fuera 10, 100 ó 1000 y si el factor de convergencia fuera $\beta = 0,2$; $\beta = 0,5$ ó $\beta = 0,8$. Analice qué método sería más rápido en cada una de las nueve combinaciones de n y β .

3.6 Cálculo de valores y vectores propios

Definición 1

Si \mathbf{A} es una matriz cuadrada de orden n se dice que el número λ es un valor propio de \mathbf{A} si existe algún vector \mathbf{x} no nulo de R^n tal que:

$$\mathbf{Ax} = \lambda \mathbf{x} \quad (1)$$

Si λ es un valor propio de \mathbf{A} , a todos los vectores \mathbf{x} que satisfacen (1), incluido el vector nulo, se les llama vectores propios de \mathbf{A} asociados al valor propio λ . ■

En la definición anterior y en todo lo que sigue, se supone que los elementos de la matriz \mathbf{A} son números reales. No obstante, aun en ese caso, la ecuación (1) puede ser satisfecha para valores imaginarios de λ y de \mathbf{x} . Aunque el interés fundamental en este libro se centrará en el caso de valores y vectores propios reales, algunos resultados requieren que se consideren también los valores imaginarios; en esos casos se hará la advertencia explícitamente, de lo contrario, puede suponer que λ es un número real y que los elementos de \mathbf{x} son números reales.

En algunas cuestiones teóricas y prácticas aparece la necesidad de calcular los valores y vectores propios de una matriz. Por ejemplo, la estabilidad de algunos procesos numéricos iterativos depende del valor propio de mayor valor absoluto de alguna matriz, la frecuencia con que puede oscilar un edificio alto y el modo en que puede producirse la oscilación depende de los valores y vectores propios de la matriz de rigidez de la estructura. Incluso en algunos procesos naturales, como el que se muestra en el siguiente ejemplo, aparecen valores y vectores propios.

Ejemplo 1

Para estudiar cómo cambia la estructura de una población en cuanto a las edades, suponga que esta se ha dividido en 5 grupos de edad del siguiente modo:

grupo 1:	de 0 a 9 años
grupo 2:	de 10 a 14 años
grupo 3:	de 15 a 20 años
grupo 4:	de 21 a 45 años
grupo 5:	más de 45 años

Se supone, además, que en todos los grupos de edades la cantidad de individuos de cada sexo es idéntico, lo cual es muy próximo a la realidad. Se define un periodo de tiempo, digamos de un año, para cuantificar la cantidad de individuos en cada grupo de edad. Para el año n , la población en cada grupo de edad se define como $p_{1,n}$, $p_{2,n}$, $p_{3,n}$, $p_{4,n}$ y $p_{5,n}$ respectivamente. Supóngase que la cantidad de individuos en cada grupo de edad en el año siguiente ($n + 1$) depende de las cifras del año actual mediante las siguientes relaciones:

Los individuos del grupo 1 en el año $n + 1$ proceden o de ese grupo o nacieron en el transcurso del año, a partir de individuos de otros grupos de edades. Se supone que cada grupo de edad tiene su propia taza de procreación t_i ($i = 1, 2, 3, 4, 5$) que es 0 en el primer grupo, muy baja en los grupos 2 y 5 y mayor en los grupos 3 y 4. Así:

$$p_{1,n+1} = \beta_1 p_{1,n} + t_2 p_{2,n} + t_3 p_{3,n} + t_4 p_{4,n} + t_5 p_{5,n}$$

Para los demás grupos de edad, la cantidad de individuos en el año $n + 1$ viene dada por una parte (β) de los que estaban en ese grupo de edad y que todavía sigue estando, más una parte (α) de los del grupo más joven que arribaron durante el año al grupo mayor. Esto es:

$$\begin{aligned} p_{2,n+1} &= \alpha_2 p_{1,n} + \beta_2 p_{2,n} \\ p_{3,n+1} &= \alpha_3 p_{2,n} + \beta_3 p_{3,n} \\ p_{4,n+1} &= \alpha_4 p_{3,n} + \beta_4 p_{4,n} \\ p_{5,n+1} &= \alpha_5 p_{4,n} + \beta_5 p_{5,n} \end{aligned}$$

Si se define el vector población \mathbf{p}_n como un vector cuyas cinco componentes representan la población en año n de cada grupo poblacional, entonces las relaciones anteriores se pueden representar en forma matricial como:

$$\mathbf{p}_{n+1} = \mathbf{A} \mathbf{p}_n \quad (2)$$

donde: $\mathbf{p}_{n+1} = \begin{bmatrix} p_{1,n+1} \\ p_{2,n+1} \\ p_{3,n+1} \\ p_{4,n+1} \\ p_{5,n+1} \end{bmatrix}$ $\mathbf{A} = \begin{bmatrix} \beta_1 & t_2 & t_3 & t_4 & t_5 \\ \alpha_2 & \beta_2 & 0 & 0 & 0 \\ 0 & \alpha_3 & \beta_3 & 0 & 0 \\ 0 & 0 & \alpha_4 & \beta_4 & 0 \\ 0 & 0 & 0 & \alpha_5 & \beta_5 \end{bmatrix}$ $\mathbf{p}_n = \begin{bmatrix} p_{1,n} \\ p_{2,n} \\ p_{3,n} \\ p_{4,n} \\ p_{5,n} \end{bmatrix}$

Surge entonces el siguiente problema: ¿existirá alguna estructura de edades estable? es decir, tal que para algún número positivo λ se satisfaga:

$$\mathbf{p}_{n+1} = \lambda \mathbf{p}_n$$

Si esta estructura poblacional existe, teniendo en cuenta la ecuación (2), se deberá cumplir que:

$$\mathbf{A} \mathbf{p}_n = \lambda \mathbf{p}_n$$

Comparando esta ecuación con (1), se ve que el problema equivale a buscar los valores y vectores propios de la matriz \mathbf{A} .

Ejemplo 2

Compruebe que $\mathbf{x} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ es un vector propio de la matriz $\mathbf{A} = \begin{bmatrix} 5 & 3 & -1 \\ 1 & -2 & 2 \\ 4 & 3 & 3 \end{bmatrix}$ y determine a qué valor propio corresponde.

Solución:

En efecto, el producto \mathbf{Ax} da como resultado:

$$\begin{bmatrix} 5 & 3 & -1 \\ 1 & -2 & 2 \\ 4 & 3 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 4 \\ 12 \end{bmatrix} = 4 \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$$

Es decir, $\mathbf{Ax} = 4\mathbf{x}$, lo cual demuestra que \mathbf{x} es un vector propio de \mathbf{A} que corresponde al valor propio $\lambda = 4$. ■

Aunque se supone que el lector tiene los conocimientos básicos de un curso elemental de Álgebra Lineal, a continuación se tratan algunos de estos conceptos a modo de repaso.

Sub espacios propios

Es sencillo demostrar que todos los vectores propios de una matriz \mathbf{A} de orden n , que corresponden a un mismo valor propio λ , forman un sub espacio vectorial $E(\lambda)$ de R^n con dimensión 1 o mayor. Esto significa que cuando se conoce un vector propio de una matriz, entonces todos los múltiplos de ese vector son también vectores propios de la matriz.

Si se conoce λ , el sub espacio propio $E(\lambda)$ se puede determinar sin dificultad resolviendo el sistema lineal homogéneo:

$$\mathbf{Ax} = \lambda \mathbf{x}$$

el cual siempre es indeterminado, pues de lo contrario su única solución sería el vector nulo.

Polinomio característico

Teóricamente, los valores propios se hallan de la siguiente forma. Si la ecuación (1) se escribe:

$$\mathbf{Ax} = \lambda \mathbf{Ix}$$

donde \mathbf{I} es la matriz idéntica de orden n , y se traspone el segundo miembro al primero:

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0} \quad (3)$$

Para que el sistema homogéneo (3) pueda tener alguna solución no trivial es necesario y suficiente que el determinante de la matriz $(\mathbf{A} - \lambda \mathbf{I})$ sea nulo. Se demuestra que el determinante:

$$\det(\mathbf{A} - \lambda \mathbf{I})$$

es un polinomio de grado n en λ y se le llama *polinomio característico* de la matriz \mathbf{A} . Así pues, los valores propios de \mathbf{A} son las n raíces de la ecuación algebraica de grado n :

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

que recibe el nombre de *ecuación característica*.

Para matrices de orden pequeño, no es difícil hallar el polinomio característico, pero si n excede de 4 o 5 es bastante complicado, por cuanto no se trata de un problema aritmético sino que se requiere calcular un determinante con términos literales, lo cual requiere de operaciones algebraicas simbólicas. Ningún método numérico con valor práctico supone que se halle en algún momento el polinomio característico de la matriz.

Ejemplo 3

Halle el polinomio característico de la matriz \mathbf{A} del ejemplo anterior y calcule sus raíces para hallar los valores propios de \mathbf{A} .

Solución:

Como la matriz es

$$\mathbf{A} = \begin{bmatrix} 5 & 3 & -1 \\ 1 & -2 & 2 \\ 4 & 3 & 3 \end{bmatrix}$$

el polinomio característico es:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \begin{vmatrix} 5 - \lambda & 3 & -1 \\ 1 & -2 - \lambda & 2 \\ 4 & 3 & 3 - \lambda \end{vmatrix} = -\lambda^3 + 6\lambda^2 + 6\lambda - 56$$

La gráfica del polinomio característico se muestra en la figura 1. Es evidente la presencia de tres raíces reales en los intervalos $[-3,5; -2,5]$, $[3,5; 4,5]$ y $[4,5; 5,5]$. En el segundo de estos intervalos se encuentra el valor propio $\lambda = 4$ analizado en el ejemplo 2.

$$-\lambda^3 + 6\lambda^2 + 6\lambda - 56$$

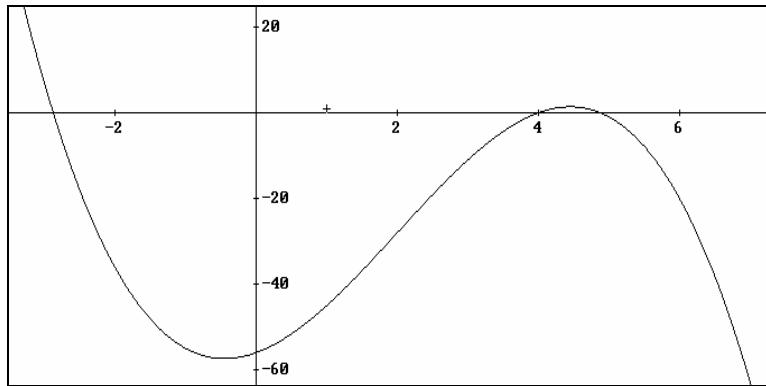


Figura 1

■

El hecho de que los valores propios son raíces de una ecuación algebraica de grado n explica por qué pueden aparecer valores propios imaginarios, aun cuando los elementos de la matriz \mathbf{A} sean números reales. Se demuestra que si el valor propio λ es una raíz de multiplicidad k del polinomio característico, entonces la dimensión del sub espacio propio $E(\lambda)$ no puede ser mayor que k .

El caso especial de las matrices simétricas

La matriz \mathbf{A} se llama simétrica si cada elemento a_{ij} coincide con su elemento simétrico a_{ji} . En muchas aplicaciones importantes aparecen matrices simétricas, para las cuales el problema de los valores y vectores propios se simplifica notablemente. En efecto, si \mathbf{A} es simétrica entonces todos los valores propios son reales (aunque no necesariamente diferentes) y los vectores propios son ortonormales (ortogonales y con norma 1). Como se verá posteriormente, estas características permiten hallar los vectores y valores propios de manera más eficiente y segura.

Localización de los valores propios

Aunque los valores propios son las raíces de una ecuación polinomial, en la práctica no se conocen los coeficientes del polinomio (salvo algunos de ellos). Por esta causa, las reglas de Descartes y de Lagrange no serán aquí de utilidad. En su lugar, existen algunas propiedades interesantes que permiten localizar a grosor modo los valores propios de la matriz.

Todos los valores propios λ (reales e imaginarios) de la matriz \mathbf{A} satisfacen la desigualdad:

$$|\lambda| \leq \|\mathbf{A}\|$$

Esto significa que si se traza un círculo con centro en el origen del plano coordenado y radio $\|\mathbf{A}\|$, entonces todos los valores propios de la matriz \mathbf{A} están dentro de este círculo.

Otros resultados importantes que permiten en algunos casos encontrar algunos de los valores propios de la matriz y que por lo menos, se pueden utilizar para comprobar los cálculos realizados, son los siguientes:

Si $\lambda_1, \lambda_2, \dots, \lambda_n$ son todos los valores propios (reales e imaginarios) de la matriz \mathbf{A} de orden n , se cumple que:

$$\lambda_1 + \lambda_2 + \dots + \lambda_n = \text{Traza}(\mathbf{A}) = a_{11} + a_{22} + \dots + a_{nn}$$

$$\lambda_1 \cdot \lambda_2 \cdot \dots \cdot \lambda_n = \det(\mathbf{A})$$

Ejemplo 4

Para la matriz \mathbf{A} de los ejemplos 2 y 3, ya se sabe que uno de sus valores propios es 4. Halle los otros dos a partir de la propiedad anterior.

Solución:

Como $\mathbf{A} = \begin{bmatrix} 5 & 3 & -1 \\ 1 & -2 & 2 \\ 4 & 3 & 3 \end{bmatrix}$ se tiene que $\text{Traza}(\mathbf{A}) = 5 - 2 + 3 = 6$ y $\det(\mathbf{A}) = -56$, de aquí se tiene

que:

$$\lambda_1 + \lambda_2 + \lambda_3 = 6$$

y

$$\lambda_1 \lambda_2 \lambda_3 = -56$$

Como uno de los valores propios (digamos λ_3) tiene valor 4, los otros dos cumplirán que:

$$\lambda_1 + \lambda_2 = 2 \quad (4)$$

y $\lambda_1 \lambda_2 = -14$

Despejando λ_2 en la primera ecuación y sustituyendo en la segunda:

$$\lambda_1(2 - \lambda_1) = -14$$

esto es: $\lambda_1^2 - 2\lambda_1 - 14 = 0$

de donde: $\lambda_1 = 1 \pm \sqrt{15}$

Tomando para λ_1 el valor negativo, se tiene: $\lambda_1 = 1 - \sqrt{15}$ y, utilizando la ecuación (4), $\lambda_2 = 1 + \sqrt{15}$. Es obvio que al tomar λ_1 como el valor positivo se obtendría para λ_2 el valor negativo. En resumen, los tres valores propios son en este caso:

$$\begin{aligned}\lambda_1 &= 1 - \sqrt{15} = -2,872983 \\ \lambda_2 &= 1 + \sqrt{15} = 4,872983 \\ \lambda_3 &= 4\end{aligned}$$

■

Gráfica del polinomio característico

Aunque hallar el polinomio característico es un problema simbólico engoroso cuando el orden de la matriz es mayor que 4, es relativamente simple, con un pequeño programa computacional, evaluar el polinomio característico para valores numéricos específicos de la variable λ , lo cual solamente implica calcular determinantes *numéricos* de orden n .

Ejemplo 5

Para la matriz \mathbf{M} de orden 5, determine una zona donde estén todos sus valores propios reales y haga la gráfica del polinomio característico evaluándolo para algunos valores de la variable λ .

$$\mathbf{M} = \begin{bmatrix} 2 & 1 & 0 & -1 & 3 \\ 1 & 0 & 3 & -2 & 1 \\ 3 & 1 & -1 & 2 & 1 \\ 2 & 3 & 0 & -1 & 2 \\ 2 & -1 & 1 & 3 & 0 \end{bmatrix}$$

Solución:

La norma de la matriz es $\|\mathbf{M}\| = \max\{7, 7, 8, 8, 7\} = 8$, así que todos los valores propios reales se encuentran en el intervalo $[-8, 8]$. Evaluando el determinante:

$$\det(\mathbf{M} - \lambda \mathbf{I})$$

para valores de λ desde -8 hasta 8 con paso $0,2$, se obtienen 81 puntos que aparecen en la gráfica de la figura 2.

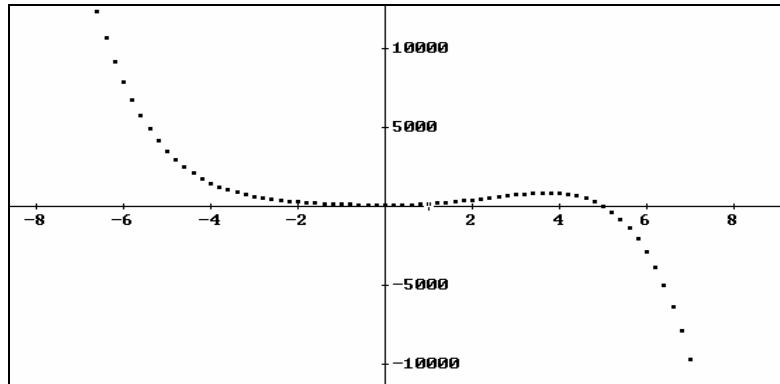


Figura 2

De la gráfica se aprecia que existe un valor propio cerca de 5. Para poder apreciar mejor qué sucede en el intervalo $[-2, 2]$, la figura 3 muestra la gráfica en ese intervalo a una escala mayor:

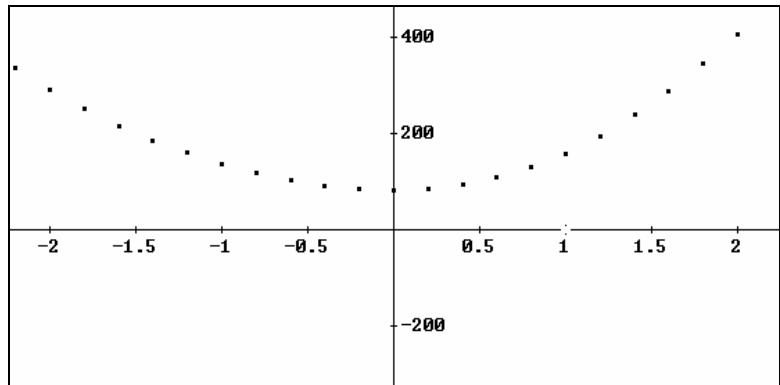


Figura 3

y ya de aquí resulta obvio que en ese intervalo no hay valores propios. En conclusión, la matriz **M** solo posee un valor propio real y se halla próximo a 5. Obsérvese que todo este trabajo gráfico se ha realizado sin utilizar el polinomio característico. ■

Para hallar uno o varios valores propios de una matriz existe una gran cantidad de técnicas, muchas de las cuales están más allá del carácter elemental de este texto. En la bibliografía recomendada al final de este capítulo pueden verse algunas referencias importantes. En lo que sigue se hace una descripción general de las estrategias de solución más empleadas.

Solución numérica de la ecuación característica

Como los valores propios de una matriz son las raíces de la ecuación característica, estos pueden hallarse por varios de los métodos estudiados en el capítulo 2. Si lo que se desea son los valores propios reales, los métodos de bisección, Regula Falsi y el método de las secantes pueden ser empleados. Como cada vez que se evalúa el polinomio característico, cuya expresión analítica se supone desconocida, se requiere calcular un determinante de orden n , aquí es importante utilizar un método eficiente, de manera que haya que evaluarlo pocas veces; por esta razón el método preferido en este caso es el de las secantes. En cuanto al método de Newton – Raphson, la necesidad de hallar la derivada del polinomio característico, lo hace inadecuado en este caso.

Para hallar los valores propios imaginarios no se aconseja utilizar el método de Newton – Bairstow, el cual requiere conocer los coeficientes del polinomio característico. Existe un método que no ha sido tratado en este texto, llamado método de Muller, que generaliza el método de las secantes. En el método de Muller en lugar de hallar la recta que pasa por dos puntos de la gráfica de la función, determinados por las dos aproximaciones anteriores, se hace pasar una parábola por tres puntos de la gráfica que corresponden a las tres últimas aproximaciones. Este procedimiento es sumamente eficiente, aunque el algoritmo es algo complicado, y permite hallar no solo los valores propios reales sino también los imaginarios. En la bibliografía recomendada del capítulo se dan referencias acerca del método.

Transformación de la matriz en una similar

Una gran cantidad de algoritmos están basados en la idea de realizar sobre la matriz **A**, cuyos valores propios se desea hallar, transformaciones que simplifiquen su estructura sin afectar los valores propios. Una vez simplificada la matriz, se hallan los valores propios de una manera más fácil. Las transformaciones que se realizan se llaman de similitud. El término matrices similares se refiere a matrices **A** y **B** relacionadas entre sí por la ecuación:

$$\mathbf{B} = \mathbf{P}^{-1} \mathbf{A} \mathbf{P}$$

donde **P** se llama matriz de transformación. Se puede demostrar sin mucha dificultad que dos matrices similares tienen los mismos valores propios.

Por ejemplo, para matrices tridiagonales y simétricas (ambas cosas a la vez) la evaluación del polinomio característico se realiza de modo muy eficiente y la ecuación característica puede ser resuelta numéricamente con muy pocas operaciones. Entonces, para hallar los valores propios de una matriz simétrica, se realizan sobre ella transformaciones de similitud que la transformen en una tridiagonal simétrica. Para ello se han usado diferentes matrices de transformación. Una de los algoritmos de este tipo más empleado es el que utiliza las matrices de transformación de Householder acerca del cual puede encontrar referencias al final del capítulo.

El método de la potencia

Este es un método muy simple basado en una ingeniosa idea, que permite hallar el valor propio (real) de mayor valor absoluto. Para comprender como el método funciona, suponga una matriz **A** de orden 2, que posee dos valores propios reales λ_1 y λ_2 tales que λ_1 es mayor que λ_2 en valor absoluto. Para concretar, se supondrá que ambos son positivos y que $\lambda_1 = k\lambda_2$ con $k > 1$. Si se

llama \mathbf{x}_1 y \mathbf{x}_2 a dos vectores propios unitarios (con norma 1) de la matriz \mathbf{A} correspondientes a los valores propios λ_1 y λ_2 , se tiene que:

$$\begin{aligned}\mathbf{Ax}_1 &= \lambda_1 \mathbf{x}_1 \\ \text{y} \quad \mathbf{Ax}_2 &= \lambda_2 \mathbf{x}_2\end{aligned}$$

Sea ahora un vector \mathbf{z} no nulo cualquiera de R^2 . Como \mathbf{x}_1 y \mathbf{x}_2 forman una base de R^2 , el vector \mathbf{z} se puede expresar como combinación lineal de ellos, es decir:

$$\mathbf{z} = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 \quad (5)$$

Si se multiplica la matriz \mathbf{A} por el vector \mathbf{z} , la imagen \mathbf{Az} obtenida será un vector de R^2 dado por:

$$\mathbf{Az} = \mathbf{A}(\alpha_1 \mathbf{x}_1) + \mathbf{A}(\alpha_2 \mathbf{x}_2) = \alpha_1 \mathbf{Ax}_1 + \alpha_2 \mathbf{Ax}_2 = \alpha_1 \lambda_1 \mathbf{x}_1 + \alpha_2 \lambda_2 \mathbf{x}_2$$

$$\mathbf{Az} = \lambda_1 (\alpha_1 \mathbf{x}_1) + \lambda_2 (\alpha_2 \mathbf{x}_2) \quad (6)$$

Si se compara la expansión del vector \mathbf{z} (5) con la de su imagen (6) se observa que las componentes sobre \mathbf{x}_1 y sobre \mathbf{x}_2 no se transformaron de igual modo, mientras la primera quedó multiplicada por λ_1 la segunda lo hizo por λ_2 . Bajo la suposición de que $\lambda_1 > \lambda_2$ la componente a lo largo de \mathbf{x}_1 ha adquirido una mayor preponderancia que la componente sobre \mathbf{x}_2 . En la figura 4 se ilustra geométricamente esta situación, en ella se ha tomado $\lambda_1 = 3$ y $\lambda_2 = 0,6$ es decir, $k = 5$.

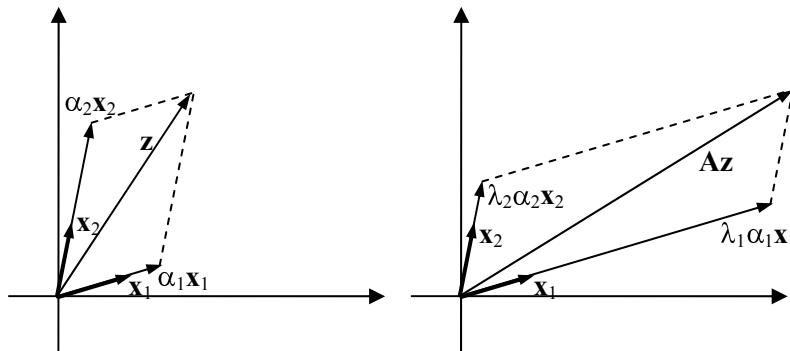


Figura 4

Como se ve, el vector \mathbf{Az} está más cerca del vector propio \mathbf{x}_1 de lo que lo estaba \mathbf{z} . Si esto proceso se repite, tomando ahora como vector de partida \mathbf{Az} se obtiene un nuevo vector $\mathbf{A}^2 \mathbf{z}$ que estará aun más próximo a la dirección de \mathbf{x}_1 . Es fácil comprender que el vector $\mathbf{A}^n \mathbf{z}$ converge hacia la dirección de \mathbf{x}_1 cuando n tiende hacia infinito con una rapidez que depende del valor de k . Para evitar que en este proceso el vector imagen crezca desmesuradamente (si $\lambda_1 > 1$) o decrezca mucho (si $\lambda_1 < 1$), en cada paso del proceso el vector se normaliza (se divide por su norma, para hacerlo unitario) antes de hallarle su imagen. Con esta idea, el algoritmo de la potencia, consta esencialmente de los siguientes pasos:

1. Tomar un vector \mathbf{z} cualquiera
2. Normalizarlo, para obtener \mathbf{z}_0
3. Hallar el vector imagen $\mathbf{w} = \mathbf{Az}_0$
4. Regresar a 2.

Para terminar el proceso iterativo, observe que, en la medida en que el vector \mathbf{z}_0 se va aproximando al vector propio \mathbf{x}_1 el vector \mathbf{w} se va aproximando a $\mathbf{Ax}_1 = \lambda_1 \mathbf{x}_1$, así que, en cada iteración, se puede calcular una aproximación a λ_1 hallando cuánto cambian las componentes de \mathbf{z}_0 al aplicarle la matriz \mathbf{A} para obtener \mathbf{w} . Si se llama λ_{aprox} a esta aproximación, el proceso se puede detener cuando su valor se va estabilizando, es decir, cuando el cambio sufrido entre dos iteraciones llegue a ser suficientemente pequeño.

Las ideas anteriores se generalizan y formalizan en el siguiente algoritmo en seudo código, en el cual se supone que la matriz \mathbf{A} posee un valor propio λ que es mayor en valor absoluto que todos los demás (nótese que esto descarta la posibilidad de que λ sea imaginario, pues en ese caso su conjugado sería también un valor propio y tendría el mismo valor absoluto). Se suponen conocidos la matriz cuadrada \mathbf{A} de orden n , una aproximación inicial \mathbf{x}_0 al vector propio que corresponde a λ y la tolerancia ε con que se desea hallar el valor propio λ . En caso de no conocerse una aproximación \mathbf{x}_0 adecuada, se puede tomar un vector con todas sus componentes iguales a 1.

```

z := x0
λanterior := 108 {Se toma un valor arbitrariamente grande para que se pueda evaluar
                      Error en la primera iteración del algoritmo}
repeat
    w := Az
    NormaW := 0
    imax := 0
    for i = 1 to n {En este lazo se hallan la norma de W así como el número imax
                      de la componente máxima de W}
        if |wi| > NormaW then
            NormaW := |wi|
            imax := i
        end
    end
    λaprox :=  $\frac{w_{imax}}{z_{imax}}$  {Se aproxima el valor propio por el cambio que ha tenido una de
                      las componentes del vector z al ser multiplicada por A}
    Error := |λaprox - λanterior|
    λanterior := λaprox
    z :=  $\frac{w}{NormaW}$ 
until Error < ε
El valor propio de mayor valor absoluto de A es λaprox y z es un vector propio asociado.
Terminar

```

Ejemplo 6

Halle con error menor que 0,001 el valor propio de mayor valor absoluto de la matriz \mathbf{A} y un vector propio correspondiente, mediante el método de la potencia.

$$\mathbf{A} = \begin{bmatrix} 6 & 3 & -2 \\ 5 & 1 & 0 \\ 2 & 3 & -4 \end{bmatrix}$$

Solución:

Aplicando a la matriz un programa basado en el algoritmo anterior, se obtienen los resultados que muestra la tabla 1.

Iter	x_1	x_2	x_3	λ	$E_m(\lambda)$
0	1,000000	1,000000	1,000000	-----	-----
1	1,000000	0,857143	0,142857	7,000000	-----
2	1,000000	0,706897	0,482759	8,285714	1,2857
3	1,000000	0,797590	0,306024	7,155172	1,1305
4	1,000000	0,745122	0,407247	7,780723	0,6256
5	1,000000	0,774184	0,351223	7,420873	0,3598
6	1,000000	0,757756	0,383890	7,620107	0,1992
7	1,000000	0,766935	0,365197	7,507489	0,1126
8	1,000000	0,761773	0,375147	7,570412	0,0629
9	1,000000	0,764665	0,369571	7,535025	0,0354
10	1,000000	0,763041	0,372702	7,554853	0,0198
11	1,000000	0,763952	0,370946	7,543720	0,0111
12	1,000000	0,763441	0,371932	7,549964	0,0062
13	1,000000	0,763728	0,371379	7,546460	0,0035
14	1,000000	0,763567	0,371689	7,548426	0,0020
15	1,000000	0,763657	0,371515	7,547322	0,0011
16	1,000000	0,763607	0,371613	7,547941	0,0006

Tabla 1

Entonces, con error menor que 0,001 el mayor valor propio de la matriz \mathbf{A} es $\lambda = 7,5479$ y un vector propio asociado a este valor propio es:

$$\mathbf{x} = \begin{bmatrix} 1,0000 \\ 0,7636 \\ 0,3716 \end{bmatrix}$$

■

Ejercicios

1. A continuación se muestran tres matrices \mathbf{A} , \mathbf{B} y \mathbf{C} y tres vectores \mathbf{x} , \mathbf{y} y \mathbf{z} . Determine de qué matriz o matrices es vector propio cada uno de los vectores. Determine en cada caso a qué valor propio corresponde.

$$\mathbf{A} = \begin{bmatrix} 129 & -26 & 92 \\ 93 & -20 & 66 \\ 150 & -30 & -107 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 182 & -28 & -136 \\ 78 & -8 & 60 \\ 228 & -36 & -170 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} -8 & -4 & 28 \\ -22 & 22 & 22 \\ -40 & 2 & 63 \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 5 \\ 3 \\ 6 \end{bmatrix} \quad \mathbf{z} = \begin{bmatrix} 4 \\ 2 \\ 5 \end{bmatrix}$$

2. En el ejercicio 1 de la sección anterior apareció el sistema

$$\begin{aligned} 10x_1 - 2x_2 + 3x_3 + 2x_4 &= 15 \\ 3x_1 - 10x_2 - 4x_3 + 2x_4 &= 4 \\ 5x_1 + 3x_2 + 10x_3 - x_4 &= 13 \\ 4x_1 - 3x_2 - 4x_3 + 10x_4 &= -2 \end{aligned}$$

para el cual el método de Jacobi converge, a pesar de que su diagonal no es predominante. Esto no es contradictorio porque la condición de diagonal predominante es suficiente pero no necesaria para la convergencia. Una condición necesaria y suficiente es que el valor propio de mayor valor absoluto de la matriz \mathbf{M} tenga módulo menor que 1. Compruebe que en este sistema esta condición se satisface.

3. Utilice un algoritmo para calcular determinantes basado en el método de Gauss, para trazar un número suficiente de puntos de la gráfica del polinomio característico de la matriz \mathbf{M} correspondiente al sistema del ejercicio anterior. Compruebe que esta matriz posee sus cuatro valores propios reales, dos positivos y dos negativos todos en el intervalo $[-1, 1]$.
4. Cada una de las cuatro matrices que siguen, tienen un valor propio (real) que es el de mayor valor absoluto. Utilice el método de la potencia para hallar este valor propio de máximo valor absoluto y determine en cada caso los otros dos a partir de la propiedad de que la suma de todos los valores propios es igual a la traza de la matriz y su producto igual al determinante de ella.

$$\text{a)} \begin{bmatrix} 3 & 4 & -3 \\ 3 & 5 & 3 \\ 4 & 2 & 8 \end{bmatrix} \quad \text{b)} \begin{bmatrix} -1 & 3 & 2 \\ 1 & -2 & 2 \\ -1 & 3 & 5 \end{bmatrix} \quad \text{c)} \begin{bmatrix} 3 & 1 & -2 \\ 1 & 2 & 3 \\ -2 & 3 & -3 \end{bmatrix} \quad \text{d)} \begin{bmatrix} 3 & 1 & -2 \\ 1 & 2 & -3 \\ 2 & 3 & -3 \end{bmatrix}$$

5. Elabore un algoritmo basado en el método de bisección que permita hallar con una exactitud dada, un valor propio real de una matriz dada que se encuentre comprendido en un intervalo dado. Establezca qué condiciones deben cumplirse para que se garantice el resultado del algoritmo.
6. Elabore un algoritmo basado en el método de las secantes que permita hallar con una exactitud dada, un valor propio real de una matriz dada conociendo dos aproximaciones iniciales del valor propio.

Otras lecturas recomendadas

Para cualquiera de los temas vinculados con los métodos numéricos de los sistemas lineales, es obligada la referencia la obra clásica “Computational Methods of Linear Algebra” de D. K. Fadeev y V. N. Fadeeva, que ha sido impreso en Cuba desde la década del 70.

Para un estudio más pormenorizado sobre normas matriciales es recomendable consultar el libro “Introduction to Numerical Analysis” de K. E. Atkinson, editorial John Wiley and Sons, 1989 y también la obra clásica de E. Isaacson y H. B. Keller, “Analysis of Numerical Methods”, reproducida en Cuba por Ediciones R.

Acerca de los temas relacionados con valores y vectores propios, la parte conceptual está muy bien expuesta en el libro “Multivariable Calculus, Linear Álgebra and Differential Equations” de S. I. Grossman. Este mismo autor posee otras obras sobre Álgebra Lineal de gran valor didáctico y conceptual. Sobre el acotamiento de los valores propios en el plano complejo existen resultados muy interesantes, como el teorema de Gershgorin, que no han sido tratados en el texto y que se encuentran explicados en el texto de Atkinson, antes citado, donde también aparece con detalle los métodos para calcular valores propios basados en las transformaciones con las matrices ortogonales de Householder. El método de Muller para hallar los valores propios reales e imaginarios de matrices reales está desarrollado, al menos en parte, en el libro “Elementary Numerical Analysis” de S. D. Conte, reproducido en Cuba a partir de una edición de McGraw-Hill.

Principales ideas del capítulo

- En la práctica suelen aparecer sistemas lineales de gran tamaño (cientos y miles de ecuaciones) para los cuales se necesita contar con algoritmos eficientes, pues de lo contrario, aun en una rápida computadora, el tiempo de solución podría ser muy largo.
- Los métodos para resolver sistemas de ecuaciones lineales son, en su casi totalidad, de dos tipos: métodos directos y métodos iterativos.
- El método de Gauss es un eficiente método directo en el cual la cantidad de operaciones para resolver un sistema lineal de n ecuaciones y n incógnitas es del orden de $\frac{1}{3}n^3$.
- En el método de Gauss existen tres estrategias de pivote: elemental, parcial y total, de las cuales la más empleada es la parcial porque es fácil de implementar y evita la propagación de los errores de redondeo.
- Para el caso especial de sistemas lineales tridiagonales con la diagonal predominante, el método de Gauss da lugar a un algoritmo muy compacto y eficiente. Este algoritmo solo requiere guardar en memoria los coeficientes no nulos y la cantidad de operaciones que realiza es del orden de $8n$.
- A partir del proceso directo del método de Gauss se elabora un algoritmo para calcular determinantes que resulta muy eficiente pues requiere una cantidad de operaciones de orden $\frac{1}{3}n^3$ para hallar un determinante de orden n .
- A partir de una modificación del método de Gauss se obtiene un algoritmo muy cómodo y eficiente para hallar la inversa de una matriz, que consiste, en esencia en resolver simultáneamente n sistemas lineales de n ecuaciones y n incógnitas. Este algoritmo requiere una cantidad de operaciones del orden de $\frac{4}{3}n^3$ para invertir una matriz cuadrada de orden n .

- En el mismo tiempo que se necesita para hallar la inversa de la matriz de un sistema, se pueden resolver cuatro sistemas de ecuaciones de ese mismo tamaño utilizando el método de Gauss.
- Sistemas lineados mal condicionados son aquellos para los cuales pequeños cambios en los coeficientes producen grandes cambios en la solución del sistema.
- La cuantía del mal condicionamiento se mide mediante el número de condición de la matriz del \mathbf{A} sistema que se define como $\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$.
- El número de condición es siempre mayor o igual que 1. Valores por debajo de 10 se consideran sistemas bien condicionados. Por encima de 10 ya se empiezan a considerar con mal condicionamiento.
- Una solución \mathbf{x}_1 de mala calidad de un sistema lineal se puede mejorar iterativamente resolviendo el sistema: $\mathbf{A} \text{ error}(\mathbf{x}_1) = \mathbf{r}_1$.
- Los métodos iterativos de Jacobi y de Seidel son una alternativa útil para resolver sistemas lineales con diagonal predominante.
- El factor de convergencia del método de Jacobi se llama α y el del método de Seidel es β . Ambos son menores que 1 cuando la diagonal del sistema es predominante y en ese caso siempre es $\beta \leq \alpha$, por lo cual la convergencia del método de Seidel suele ser más rápida que la de Jacobi.
- Además de una convergencia más rápida el método de Seidel posee un algoritmo más simple y requiere menos memoria. El método de Jacobi, sin embargo, se adapta muy fácilmente al trabajo con cálculo paralelo.
- Los métodos iterativos son muy apropiados para grandes sistemas de ecuaciones con muchos coeficientes nulos en los cuales todas las ecuaciones obedecen a una misma estructura general.
- Calcular todos los valores propios de una matriz cualquiera es un problema muy complejo. Cuando la matriz es simétrica o cuando solo se desea hallar el valor propio de mayor valor absoluto, existen métodos no muy complicados y eficientes para resolver el problema.
- El método de la potencia es un algoritmo iterativo muy fácil de programar capaz de hallar el valor propio de mayor valor absoluto de una matriz y un vector propio asociado a este valor propio.
- Si se tiene un programa eficiente para calcular el determinante de una matriz numérica, el polinomio característico se puede graficar trazando un suficiente número de puntos aislados de la gráfica y la ecuación característica se puede resolver numéricamente por métodos como biseción y secantes.

Auto examen

1. En la solución de sistemas de ecuaciones lineales, ¿qué significado poseen los términos “métodos directos” y “métodos iterativos” ?
2. Dado el siguiente sistema de ecuaciones:

$$\begin{aligned}x_2 + 2x_3 &= -3 \\3x_1 - 2x_2 + x_3 &= 6 \\2x_1 + 3x_2 - 2x_3 &= 5\end{aligned}$$

- a) Expréselo en la forma matricial $\mathbf{Ax} = \mathbf{b}$.
- b) Halle la norma de \mathbf{A} y la norma de \mathbf{b} .

- c) Resuelva el sistema mediante el método de Gauss utilizando la estrategia parcial de pivote.
- d) Compruebe si $\mathbf{Ax} = \mathbf{b}$.
- e) Halle la norma de \mathbf{x} , analice si $\|\mathbf{A}\| \cdot \|\mathbf{x}\| \geq \|\mathbf{b}\|$ y justifique su respuesta.
3. Dado el sistema de 80 ecuaciones:
- $$2x_{i-1} + 8x_i - x_{i+1} = b_i \quad i = 1, 2, \dots, 80$$
- donde $x_0 = 6$, $x_{81} = 10$ y los coeficientes b_i ($i = 1, 2, \dots, 80$) son datos,
- a) Determine qué cantidad de operaciones se requiere para resolver el sistema mediante el método de Gauss especializado en sistemas tridiagonales.
- b) Analice si es posible aplicar a este sistema los métodos iterativos de Jacobi y de Seidel y, de ser así, halle qué cantidad de operaciones se necesitará en cada uno para lograr 4 cifras decimales exactas si el error inicial es del orden de 1,5.
4. Halle la inversa de la matriz \mathbf{A} del sistema del ejercicio 2 mediante el método de Gauss y diga el valor del determinante de \mathbf{A} .
5. ¿Qué se entiende por un sistema mal condicionado y cómo se mide el mal condicionamiento?
6. Elabore un algoritmo en seudo código para obtener la solución del sistema del ejercicio 3 con cuatro cifras decimales exactas mediante el método de Seidel.
7. ¿A qué se llama valores propios y vectores propios de una matriz?
8. Sin hallar el polinomio característico de la matriz \mathbf{C} halle varios valores del mismo que le permitan graficarlo en el intervalo en el que están acotados todos los valores propios.

$$\mathbf{C} = \begin{bmatrix} 2 & 3 & -1 \\ 3 & 1 & 1 \\ -1 & 1 & 3 \end{bmatrix}$$

9. Halle el valor propio de mayor valor absoluto de la matriz \mathbf{C} .

CAPÍTULO 4

Matemática Numérica, 2da Edición

Manuel Álvarez, Alfredo Guerra, Rogelio Lau

APROXIMACIÓN DE FUNCIONES

Objetivos

Al finalizar el estudio y la ejercitación de este capítulo, el lector debe ser capaz de:

- Describir el problema general de la aproximación de funciones y, como casos particulares, los problemas de interpolación y de ajuste de curvas.
- Distinguir entre un problema que debe ser resuelto por interpolación y otro que corresponde al ajuste de curvas.
- Establecer la relación entre el grado de un polinomio interpolador y la cantidad de nodos de interpolación de manera que dicho polinomio exista y sea único.
- Explicar el concepto de error de interpolación y acotarlo haciendo uso de la fórmula correspondiente.
- Hallar valores interpolados o el polinomio interpolador utilizando cualquiera de los métodos: Lagrange y Diferencias Divididas de Newton.
- Utilizar la interpolación polinomial para resolver problemas concretos.
- Explicar el concepto de función spline.
- Explicar en qué consiste el carácter local de la interpolación polinomial y el carácter global de la interpolación mediante splines.
- Explicar la relación entre las diferencias divididas y las derivadas y emplear esta relación para determinar el carácter polinomial de un conjunto de datos.
- Explicar las características de los splines cúbicos natural, anclado y periódico.
- Plantear los sistemas de ecuaciones lineales necesarios para calcular splines cúbicos de interpolación de los tipos estudiados y para hallar curvas paramétricas abiertas y cerradas que pasen por puntos específicos del plano coordenado.
- Deducir el sistema normal de ecuaciones para ajustar un modelo lineal a un conjunto de datos.
- Ajustar modelos lineales a conjuntos de datos para resolver problemas aplicados.
- Ajustar modelos no lineales en casos en que el modelo sea linealizable o se pueda resolver numéricamente de manera simple.
- Describir mediante seudo código cualquiera de los algoritmos estudiados en este capítulo.

4.1 Introducción

El problema de la aproximación funcional

En este capítulo se estudiarán algunas técnicas para resolver el siguiente tipo de problema: hallar la expresión analítica de una función $g(x)$ que sirva para aproximar a otra función $f(x)$ para x en algún intervalo $[a, b]$.

Problemas como este aparecen en la teoría y en la práctica con gran frecuencia; a veces porque no se conoce una expresión analítica para la función $f(x)$, sino valores aislados $f(x_1), f(x_2), \dots, f(x_n)$ de la misma y se necesita disponer de una expresión analítica que permita, aunque sea de manera aproximada, poder evaluar la función en otros valores de x ; en otras ocasiones el algoritmo algebraico para calcular $f(x)$, aunque se conoce, resulta tan complicado que se prefiere hallar una función $g(x)$ de una clase más simple y utilizarla en lugar de $f(x)$, aun sabiendo que se está incurriendo en un error. En los ejemplos que siguen se ilustran varias situaciones típicas.

Ejemplo 1

La función Γ (gamma) se utiliza en el cálculo de algunas integrales impropias que se presentan, por ejemplo, al resolver ecuaciones diferenciales mediante la transformada de Laplace. Esta función se define mediante la expresión:

$$\Gamma(a) = \int_0^{\infty} e^{-x} x^{a-1} dx \quad \text{para } a > 0 \quad (1)$$

a	$\Gamma(a)$	a	$\Gamma(a)$	a	$\Gamma(a)$
1,00	1,00000	1,35	0,89115	1,70	0,90865
1,05	0,97350	1,40	0,88726	1,75	0,91906
1,10	0,95135	1,45	0,88565	1,80	0,93138
1,15	0,93304	1,50	0,88623	1,85	0,94561
1,20	0,91817	1,55	0,88887	1,90	0,96177
1,25	0,90640	1,60	0,89352	1,95	0,97988
1,30	0,89747	1,65	0,90012	2,00	1,00000

Tabla 1

Salvo para valores enteros de a , la integral (1) es muy difícil de evaluar. En muchos manuales se ofrecen valores de la función $\Gamma(a)$ para $a \in [1, 2]$ como se muestra en la tabla 1. Para valores de a fuera de este intervalo se aplica la fórmula:

$$\Gamma(a + 1) = a\Gamma(a)$$

Si se desea obtener $\Gamma(a)$ para algún a entre 1 y 2 que no aparezca en la tabla, es preferible, en lugar de utilizar la integral impropia (1) que define a la función, determinar una función simple g (por ejemplo, un polinomio entero) que en un entorno de a se aproxime a $\Gamma(a)$ y, entonces, hallar $g(a)$.

Ejemplo 2

Con el objetivo de encontrar una relación entre el peso y la talla de un determinado sector poblacional, se selecciona al azar una muestra de 100 individuos del grupo y se obtiene para cada persona i , su peso (p_i) y su estatura (t_i). Al representar estas mediciones en un sistema de ejes $p-t$,

se obtiene un diagrama de puntos como el que se muestra en la figura 1. Se desea hallar una fórmula que permita establecer una relación entre el peso p y la talla t de los individuos del grupo.

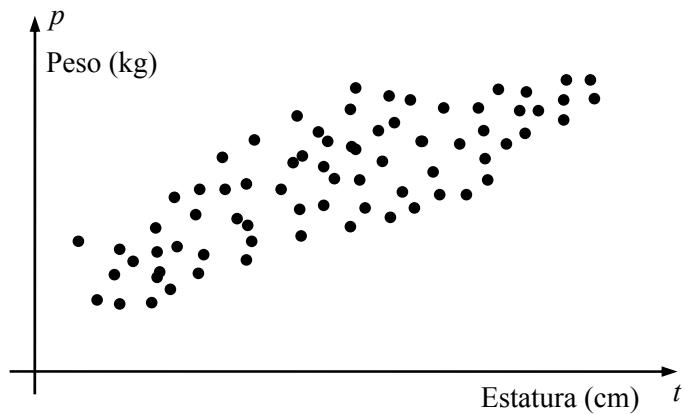


Figura 1

Ejemplo 3

Para producir en un torno de mando numérico una pieza con el perfil longitudinal que se muestra en la figura 2, es necesario obtener una función simple (de modo que pueda ser evaluada en un tiempo muy breve) que describa el contorno de la pieza. Esta función servirá para fijar la posición de la cuchilla del torno en cada instante. Algunas dimensiones, marcadas en la figura en milímetros, deberán ser respetadas y el perfil de la pieza debe ser una curva suave.

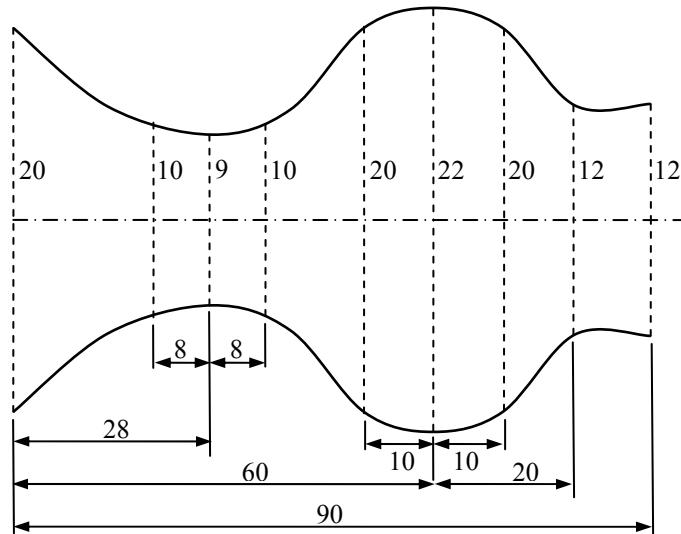


Figura 2

Ejemplo 4

En un laboratorio se está analizando la resistencia de las barras de acero corrugadas de 12 mm para la construcción. Para ello se someten 10 muestras de la misma longitud a grandes esfuerzos de tensión y se mide la deformación sufrida. Los resultados se muestran en la tabla 2 y en la figura 3 se han representado en un plano $T-s$.

T (tensión) kilonewton	5,1	7,7	10,8	13,2	15,6	18,1	22,2	23,9	26,3	27,5
s (deformación) milímetro	0,10	0,17	0,24	0,30	0,36	0,40	0,53	0,70	0,85	1,03

Tabla 2

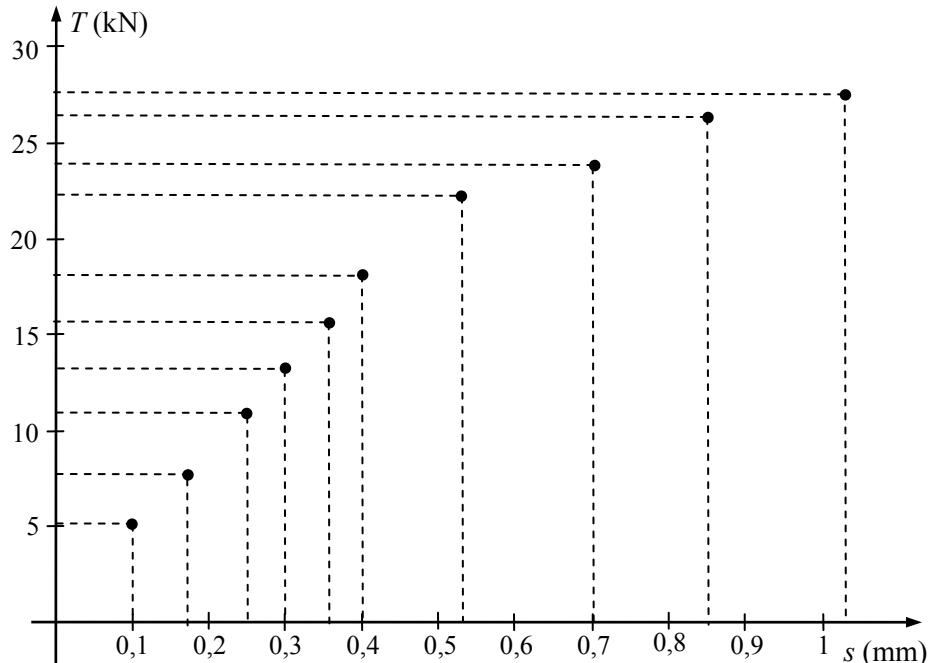


Figura 3

Se quiere hallar un modelo lineal

$$T = a_0 + a_1 s$$

y otro cuadrático:

$$T = b_0 + b_1 s + b_2 s^2$$

para representar el comportamiento de este material para tensiones en el rango de 0 a 30 kilonewton.

En los cuatro ejemplos mostrados se ha intentado cubrir la mayor variedad de situaciones. Así, en el ejemplo 1, existe una expresión analítica para determinar $\Gamma(a)$ para todo a positivo, mientras en

los demás ejemplos no existe una fórmula que relacione a las variables. En el ejemplo 2 la relación entre peso y talla será de un tipo estadístico, ya que el peso de un individuo no es una función únicamente de su estatura; de lo que se trata realmente es de hallar una relación matemática entre la estatura de una persona y el valor esperado o valor medio del peso de todas las personas que poseen esa talla. En el ejemplo 4 existe también una componente aleatoria significativa, ya que en la medición de T y de s estarán presentes errores inevitables y, además, porque es imposible lograr que todas las muestras tomadas sean idénticas. En el ejemplo 3, puede suponerse que los valores dados en el esquema son exactos, y en esto hay una gran semejanza con el ejemplo 1; en el ejemplo 3, sin embargo, la necesidad de que la expresión matemática que describa al perfil sea simple y su representación gráfica sea una curva suave, impone requisitos difíciles de satisfacer. Los ejemplos 1 y 3 son situaciones típicas de problemas de interpolación, mientras que el 2 y el 4 representan casos en que la solución se logra mediante técnicas de ajuste de funciones. Estos son los dos tipos de enfoque que se estudiarán en el resto del capítulo.

Conceptos básicos

Si se conocen los valores que toma la función $f(x)$ en los $n + 1$ puntos diferentes x_0, x_1, \dots, x_n , el problema de interpolación consiste en hallar una función $g(x)$ cuyos valores puedan ser calculados para cualquier x en un intervalo que contiene a x_0, x_1, \dots, x_n , de manera que:

$$\begin{aligned} g(x_0) &= f(x_0) \\ g(x_1) &= f(x_1) \\ &\vdots \\ g(x_n) &= f(x_n) \end{aligned}$$

Los números x_0, x_1, \dots, x_n suelen llamarse *puntos* o *nodos de interpolación*. Si x no es un nodo de interpolación, al número real $g(x)$ se le llama *valor interpolado*. Con frecuencia se utiliza la frase *valor extrapolado* para referirse a $g(x)$ cuando x es mayor que el mayor nodo de interpolación o menor que el menor de ellos. La función $g(x)$ se denomina función *interpoladora* y debe ser lo suficientemente simple como para que resulte fácil y rápido evaluarla en los puntos deseados; por esta razón, lo más usual es utilizar polinomios de grado pequeño con este fin.

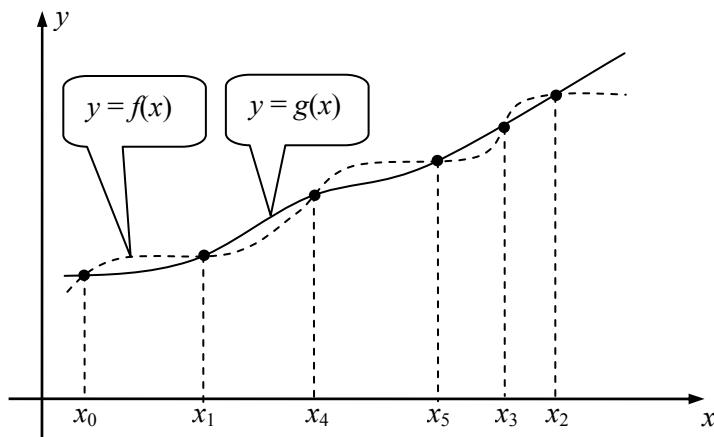


Figura 1

La figura 1 ilustra geométricamente los conceptos anteriores para el caso de una función $f(x)$ con seis nodos de interpolación que aparece representada en línea de puntos para indicar que sus valores son desconocidos salvo para los puntos x_0, x_1, x_2, x_3, x_4 y x_5 , que es la situación más frecuente. Nótese que la función interpoladora $g(x)$ coincide con la función interpolada $f(x)$ en los nodos de interpolación, lo cual significa que sus gráficas se cortan en los puntos $(x_i, f(x_i))$, $i = 0, 1, 2, 3, 4, 5$. Note que los nodos de interpolación tienen que ser diferentes pero no es obligatorio que estén ordenados.

A la diferencia entre la función interpolada y la interpoladora se le llama *error de interpolación* y se denota $R(x)$, es decir:

$$R(x) = \text{error}(g(x)) = f(x) - g(x)$$

El error de interpolación depende de x ; es cero si x es un nodo de interpolación y, por lo general, aumenta a medida que x está más distante de los nodos. En particular, el error de interpolación suele ser mucho mayor (en valor absoluto) en los casos de extrapolación que de interpolación.

4.2 Interpolación polinomial

Cuando la función interpoladora es un polinomio, la interpolación se llama polinomial. Si se supone conocido el conjunto $\{x_0, x_1, \dots, x_n\}$ de nodos de interpolación, para los cuales se conocen las imágenes de la función f :

$$y_i = f(x_i) \quad i = 0, 1, \dots, n$$

existen tres problemas fundamentales relacionados con la interpolación polinómica:

- a) Existencia: ¿Hay algún polinomio $p(x)$ tal que $p(x_i) = y_i$ para $i = 0, 1, \dots, n$?
- b) Unicidad: Si tal polinomio existe, ¿será único?
- c) Construcción: Si existe el polinomio interpolador y es único, ¿cómo hallarlo?

El teorema que sigue resuelve definitivamente los dos primeros problemas. El tercero tiene varias soluciones, pues, aunque se trata no solo de hallar el polinomio interpolador, sino de hacerlo de un modo eficiente, para esto existe una gran cantidad de enfoques de los cuales en este capítulo se verán algunos de los más importantes.

Existencia y unicidad del polinomio interpolador

Teorema 1

Sean $n + 1$ números reales diferentes x_0, x_1, \dots, x_n y f una función que toma valores $y_0 = f(x_0)$, $y_1 = f(x_1)$, ..., $y_n = f(x_n)$. Entonces existe uno y solo un polinomio p de grado menor o igual que n tal que:

$$\begin{aligned} p(x_0) &= y_0 \\ p(x_1) &= y_1 \\ &\vdots \\ p(x_n) &= y_n \end{aligned} \tag{1}$$

Demostración:

Cualquier polinomio de grado menor o igual que n se puede representar como:

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \quad (2)$$

Hay que probar que existe uno y solo un juego de coeficientes a_0, a_1, \dots, a_n que den lugar a un polinomio que satisface las condiciones (1). Esto es:

$$\begin{aligned} p(x_0) &= a_0 + a_1x_0 + a_2x_0^2 + \cdots + a_nx_0^n = y_0 \\ p(x_1) &= a_0 + a_1x_1 + a_2x_1^2 + \cdots + a_nx_1^n = y_1 \\ &\vdots \\ p(x_n) &= a_0 + a_1x_n + a_2x_n^2 + \cdots + a_nx_n^n = y_n \end{aligned} \quad (3)$$

Este sistema de ecuaciones se puede expresar en forma matricial como:

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

El determinante de la matriz de este sistema se llama determinante de Vandermonde y será designado por V_{n+1} . Es decir:

$$V_{n+1} = \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix}$$

Aplicando el principio de inducción completa, se prueba fácilmente que, si los números x_i ($i = 0, 1, 2, \dots, n$) son distintos, entonces V_{n+1} no es nulo. En efecto:

$$V_2 = \begin{vmatrix} 1 & x_0 \\ 1 & x_1 \end{vmatrix} = x_1 - x_0 \neq 0$$

Si se supone que V_n no es cero, se llega a que V_{n+1} tampoco lo es. Para ello, se realizan sobre el determinante V_{n+1} las siguientes operaciones (en ese orden), ninguna de las cuales altera el valor del determinante:

Restar a la columna $n+1$, la columna n multiplicada por x_0
 Restar a la columna n , la columna $n-1$ multiplicada por x_0
 \vdots
 Restar a la columna 1, la columna 2 multiplicada por x_0

Con esto se logra anular todos los elementos de la fila 1 excepto el primero. Se obtiene:

$$V_{n+1} = \begin{vmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & x_1 - x_0 & x_1^2 - x_0 x_1 & \cdots & x_1^n - x_0 x_1^{n-1} \\ 1 & x_2 - x_0 & x_2^2 - x_0 x_2 & \cdots & x_2^n - x_0 x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n - x_0 & x_n^2 - x_0 x_n & \cdots & x_n^n - x_0 x_n^{n-1} \end{vmatrix}$$

Desarrollando por menores a V_{n+1} respecto a la primera fila:

$$V_{n+1} = \begin{vmatrix} x_1 - x_0 & x_1^2 - x_0 x_1 & \cdots & x_1^n - x_0 x_1^{n-1} \\ x_2 - x_0 & x_2^2 - x_0 x_2 & \cdots & x_2^n - x_0 x_2^{n-1} \\ \vdots & \vdots & & \vdots \\ x_n - x_0 & x_n^2 - x_0 x_n & \cdots & x_n^n - x_0 x_n^{n-1} \end{vmatrix}$$

Extrayendo factores comunes en los elementos de las columnas 2, 3, ..., n:

$$V_{n+1} = \begin{vmatrix} x_1 - x_0 & x_1(x_1 - x_0) & \cdots & x_1^{n-1}(x_1 - x_0) \\ x_2 - x_0 & x_2(x_2 - x_0) & \cdots & x_2^{n-1}(x_2 - x_0) \\ \vdots & \vdots & & \vdots \\ x_n - x_0 & x_n(x_n - x_0) & \cdots & x_n^{n-1}(x_n - x_0) \end{vmatrix}$$

Ahora se puede extraer de la primera fila el factor $(x_1 - x_0)$, de la segunda fila $(x_2 - x_0)$, etcétera, para obtener:

$$V_{n+1} = (x_1 - x_0)(x_2 - x_0) \cdots (x_n - x_0) \begin{vmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^{n-1} \end{vmatrix}$$

Como todos los binomios que aparecen multiplicando al determinante son no nulos (los nodos de interpolación son diferentes) y el determinante de la derecha es V_n , que se ha supuesto no nulo, resulta que:

$$V_{n+1} \neq 0$$

El hecho de que $V_n \neq 0 \Rightarrow V_{n+1} \neq 0$ junto a la certeza de que $V_2 \neq 0$, prueban que para todo $n \geq 2$ es $V_n \neq 0$ siempre que los nodos sean diferentes. De aquí resulta que el sistema (3) admite solución única y esto prueba el teorema. ■

Es importante notar que, para garantizar la existencia y la unicidad del polinomio interpolador, el teorema exige una relación precisa entre el número de nodos y el grado del polinomio: si hay

$n + 1$ nodos entonces el grado del polinomio será n o menor, pues algunos de los coeficientes pudieran ser cero. Obsérvese que un polinomio de grado n posee $n + 1$ coeficientes: a_0, a_1, \dots, a_n . De no cumplirse esta relación, pueden fallar la existencia o la unicidad. Esto se demuestra en los ejemplos que siguen.

Ejemplo 1

Pruebe que para tres nodos de interpolación no puede asegurarse la existencia de un polinomio interpolador con grado menor o igual que 1.

Solución:

Basta hallar un ejemplo para probar la proposición. Sean:

$$\begin{aligned}x_0 &= 1 & y_0 &= f(x_0) = 1 \\x_1 &= 2 & y_1 &= f(x_1) = 3 \\x_2 &= 3 & y_2 &= f(x_2) = 1\end{aligned}$$

los nodos de interpolación y los respectivos valores de la función interpolada. Sea

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

un polinomio de grado menor o igual que uno. Como la gráfica de todo polinomio de este tipo es una línea recta y los tres puntos (x_0, y_0) , (x_1, y_1) y (x_2, y_2) no están alineados, es obvio que ningún polinomio de grado menor o igual que uno puede interpolar estos valores.

Ejemplo 2

Probar que para dos nodos de interpolación no puede asegurarse la unicidad del polinomio interpolador de grado menor o igual que dos.

Solución:

Es suficiente con dar un ejemplo para demostrar la proposición. Sean

$$\begin{aligned}x_0 &= 1 & y_0 &= f(x_0) = 3 \\x_1 &= 2 & y_1 &= f(x_1) = 4\end{aligned}$$

los dos nodos con sus correspondientes imágenes. Sea el polinomio interpolador

$$p(x) = a_0 + a_1 x + a_2 x^2$$

Hay que satisfacer las condiciones:

$$\begin{aligned}p(x_0) &= a_0 + a_1 x_0 + a_2 x_0^2 = y_0 \\p(x_1) &= a_0 + a_1 x_1 + a_2 x_1^2 = y_1\end{aligned}$$

esto es:

$$\begin{aligned}a_0 + a_1 + a_2 &= 3 \\a_0 + 2a_1 + 4a_2 &= 4\end{aligned}$$

que es un sistema de rango 2 con 3 incógnitas, que posee, por tanto, infinitas soluciones, cada una de las cuales representa a un polinomio de grado menor o igual a dos. Geométricamente, estos polinomios corresponden a las infinitas parábolas de eje vertical que pasan por los puntos (1, 3) y (2, 4). ■

A partir del teorema de existencia y unicidad y de los dos ejemplos mostrados, si se denota por P_m la familia de todos los polinomios con grado menor o igual que m , puede concluirse que:

Dados $n + 1$ nodos de interpolación diferentes: x_0, x_1, \dots, x_n de una función f , entonces:
Si $m < n$ No puede asegurarse que en P_m exista algún polinomio que interpole la función.
Si $m > n$ En P_m existen infinitos polinomios que interpolan la función
Si $m = n$ Existe en P_m uno y solo un polinomio que interpola la función.

En todo lo que sigue, siempre que se tengan $n + 1$ nodos de interpolación, se sobreentenderá como polinomio interpolador aquel de grado menor o igual que n .

Error del polinomio interpolador

Cuando la función $g(x)$ interpoladora es un polinomio, el error de interpolación

$$R(x) = f(x) - g(x)$$

puede expresarse mediante una fórmula muy interesante. Sea $f(x)$ la función interpolada, x_0, x_1, \dots, x_n los nodos de interpolación y $p(x)$ el polinomio interpolador. El error de interpolación es una función de x definida por:

$$R(x) = f(x) - p(x)$$

y como el polinomio interpolar coincide con $f(x)$ en los nodos de interpolación, es evidente que:

$$R(x_i) = 0 \quad \text{para } i = 0, 1, 2, \dots, n$$

Como la función $k(x - x_0)(x - x_1) \cdots (x - x_n)$

también se anula en estos mismos valores, independientemente del coeficiente k , resulta interesante estudiar la diferencia entre ambas funciones. Considérese entonces la función:

$$F(x) = R(x) - k(x - x_0)(x - x_1) \cdots (x - x_n) \quad (4)$$

Es obvio que $F(x_i) = 0 \quad \text{para } i = 0, 1, 2, \dots, n$

Supóngase que se quiere investigar el error de interpolación para un valor x^* que no es un nodo de interpolación. Evaluando para $x = x^*$ la función (4):

$$F(x^*) = R(x^*) - k(x^* - x_0)(x^* - x_1) \cdots (x^* - x_n) \quad (5)$$

Como ninguno de los factores $(x^* - x_i)$ es cero, el producto $(x^* - x_0)(x^* - x_1) \cdots (x^* - x_n)$ tampoco lo puede ser, así que es posible encontrar un valor de k que para el cual:

$$F(x^*) = 0$$

Por supuesto, este valor de k depende de x^* . Dando a k este valor, la función $F(x)$ posee $n + 2$ ceros: x_0, x_1, \dots, x_n y x^* .

Si se supone que f es derivable $n + 1$ veces en un intervalo I que incluya a todos los nodos de interpolación y al valor x^* , la función F también lo será ya que:

$$F(x) = f(x) - p(x) - k(x - x_0)(x - x_1) \cdots (x - x_n) \quad (6)$$

y, aparte de $f(x)$, el resto de las funciones que intervienen en su definición son polinomios, los cuales son infinitamente derivables.

Como se recordará del curso de Cálculo de una variable, el teorema de Rolle garantiza que si una función derivable en un intervalo cerrado toma en ambos extremos del intervalo el mismo valor, entonces su derivada se anula en algún punto de ese intervalo. Como $F(x)$ es derivable y toma el mismo valor (cero) en por lo menos $n + 2$ puntos de I , entonces su derivada se anulará al menos una vez entre cada dos de estos puntos; esto es:

$F'(x)$ posee al menos $n + 1$ ceros en I

Con razonamientos similares se llega sucesivamente a las siguientes conclusiones:

$F''(x)$ posee al menos n ceros en I

$F'''(x)$ posee al menos $n - 1$ ceros en I

⋮

$F^{(n+1)}(x)$ posee al menos 1 cero en I

Si a este cero se le asigna la letra c , se ha probado que existe al menos un número c en I tal que:

$$F^{(n+1)}(c) = 0$$

Por otra parte, derivando $n + 1$ veces en ambos miembros de la ecuación (6):

$$F^{(n+1)}(x) = f^{(n+1)}(x) - p^{(n+1)}(x) - kD^{n+1}[(x - x_0)(x - x_1) \cdots (x - x_n)]$$

Ahora bien, como $p(x)$ es un polinomio de grado menor o igual que n , su derivada de orden $n + 1$ es cero para todo x . La función $(x - x_0)(x - x_1) \cdots (x - x_n)$ es un polinomio de grado $n + 1$ cuyo

término de mayor grado es x^{n+1} , así que su derivada de orden $n+1$ es la constante $(n+1)!$. De todo esto resulta:

$$F^{(n+1)}(x) = f^{(n+1)}(x) - k(n+1)!$$

Evaluando para $x = c$:

$$F^{(n+1)}(c) = f^{(n+1)}(c) - k(n+1)! = 0$$

despejando k :
$$k = \frac{f^{(n+1)}(c)}{(n+1)!}$$

Resumiendo: si x^* no es un nodo de interpolación, puede asegurarse la existencia en I de un valor c tal que, de acuerdo con la ecuación (5):

$$F(x^*) = R(x^*) - \frac{f^{(n+1)}(c)}{(n+1)!} (x^* - x_0)(x^* - x_1) \cdots (x^* - x_n) = 0$$

esto es:
$$R(x^*) = \frac{f^{(n+1)}(c)}{(n+1)!} (x^* - x_0)(x^* - x_1) \cdots (x^* - x_n)$$

La deducción anterior es válida si x^* no es un nodo de interpolación. Sin embargo, la expresión obtenida también resulta cierta cuando x^* es alguno de los nodos, ya que entonces la fórmula se reduciría a:

$$R(x_i) = 0 \quad \text{para } i = 0, 1, 2, \dots, n$$

lo cual es un hecho conocido. Puede entonces permitirse que x^* sea cualquier número real. Por la importancia de esta conclusión será establecida como un teorema:

Teorema 2

Si $f(x)$ es derivable $n+1$ veces en un intervalo cerrado I que incluye a los nodos de interpolación x_0, x_1, \dots, x_n del polinomio interpolador $p(x)$ y al número x , entonces existe en I al menos un valor c tal que el error de interpolación en x es:

$$R(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x - x_0)(x - x_1) \cdots (x - x_n) \quad (7)$$

■

Es conveniente destacar acerca de esta importante fórmula algunos aspectos que pudieran pasar inadvertidos:

- Acerca del número c solo puede asegurarse su existencia en el intervalo I pero no su valor preciso.
- Si x es un nodo de interpolación, alguno de los binomios que contiene la expresión (7) se anula, sin importar que valor tome c y se obtiene $R(x) = 0$.
- El valor de c depende de la función $f(x)$, de la ubicación de los nodos de interpolación y del valor x en el cual se interpole.

- El factor $(x - x_0)(x - x_1) \cdots (x - x_n)$ se anula en los nodos de interpolación y tiende hacia infinito cuando x tiende hacia infinito. Esto explica por qué, cuando el punto de interpolación está cerca de un nodo, el error de interpolación suele ser pequeño y por qué en el caso de la extrapolación, el error se hace, por regla general, muy grande.
- Si se puede hallar una cota de la función $|f^{(n+1)}(x)|$ para $x \in I$, la fórmula (7) brinda una cota para el error absoluto de interpolación, pues:

$$\text{Si } |f^{(n+1)}(x)| \leq M \text{ para } x \in I, \text{ entonces: } |R(x)| \leq \frac{M}{(n+1)!} (x - x_0)(x - x_1) \cdots (x - x_n)$$

Ejemplo 3

Dada la función $f(x) = \sin x$, obtenga su polinomio de interpolación para los nodos $x_0 = 0$, $x_1 = \frac{\pi}{4}$ y $x_2 = \frac{\pi}{2}$ y dé una cota del error de interpolación para $0 \leq x \leq \frac{\pi}{2}$. Con estos resultados, estime el valor de $\sin(\frac{\pi}{6})$.

Solución:

$$\begin{aligned} y_0 &= f(x_0) = \sin 0 = 0 \\ y_1 &= f(x_1) = \sin \frac{\pi}{4} = \frac{\sqrt{2}}{2} \\ y_2 &= f(x_2) = \sin \frac{\pi}{2} = 1 \end{aligned}$$

Polinomio de interpolación: $p(x) = ax^2 + bx + c$

Luego: $p(x_0) = p(0) = c = 0$

$$\begin{aligned} p(x_1) &= p\left(\frac{\pi}{4}\right) = a\left(\frac{\pi}{4}\right)^2 + b\left(\frac{\pi}{4}\right) + c = \frac{\sqrt{2}}{2} \\ p(x_2) &= p\left(\frac{\pi}{2}\right) = a\left(\frac{\pi}{2}\right)^2 + b\left(\frac{\pi}{2}\right) + c = 1 \end{aligned}$$

de donde $c = 0$ y:

$$\begin{aligned} a\left(\frac{\pi}{4}\right)^2 + b\left(\frac{\pi}{4}\right) &= \frac{\sqrt{2}}{2} \\ a\left(\frac{\pi}{2}\right)^2 + b\left(\frac{\pi}{2}\right) &= 1 \end{aligned}$$

Resolviendo el sistema: $a = -0,335749$ y $b = 1,164013$

Resulta: $p(x) = -0,335749x^2 + 1,164013x$

Como la función $f(x) = \sin x$ es derivable indefinidamente y

$$f^{(3)}(x) = -\cos x$$

se puede acotar esta derivada: $|f^{(3)}(x)| \leq 1$

y con esto acotar también el error de interpolación

$$|R(x)| \leq \frac{M}{(n+1)!} (x - x_0)(x - x_1) \cdots (x - x_n) = \frac{1}{3!} (x - x_0)(x - x_1)(x - x_2)$$

$$|R(x)| \leq \frac{1}{6} x(x - \frac{\pi}{4})(x - \frac{\pi}{2})$$

en particular, para $x = \frac{\pi}{6}$: $p(\frac{\pi}{6}) = -0,335749(\frac{\pi}{6})^2 + 1,164013(\frac{\pi}{6}) = 0,5174$

$$|R(\frac{\pi}{6})| \leq \frac{1}{6} \left[\frac{\pi}{6} \left(\frac{\pi}{6} - \frac{\pi}{4} \right) \left(\frac{\pi}{6} - \frac{\pi}{2} \right) \right] = 0,0239$$

De aquí: $\sin(\frac{\pi}{6}) = 0,5174 \pm 0,0239$

Como se sabe, el valor exacto de $\sin(\frac{\pi}{6})$ es 0,5 y, por tanto, el verdadero error de interpolación es 0,0174, de manera que la cota 0,0239 resulta algo conservadora.

La figura 1 muestra las gráficas de $f(x) = \sin x$ y del polinomio $p(x) = -0,335749x^2 + 1,164013x$ en el intervalo $[0, \frac{\pi}{2}]$. Obsérvese la similitud de ambas gráficas así como su coincidencia en los nodos de interpolación. Nótese también la gran discrepancia de ambas funciones para valores de x fuera del intervalo de interpolación.

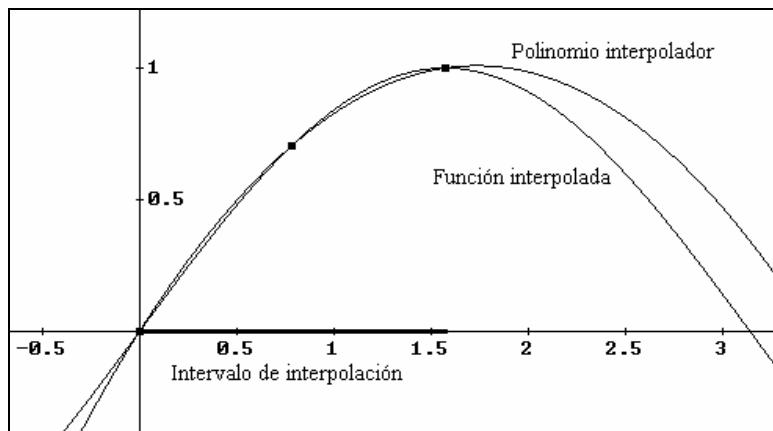


Figura 1

4.3 El método de Lagrange

El procedimiento seguido en el ejemplo anterior (ejemplo3 de la sección 4.2) para determinar el polinomio de interpolación requiere la resolución de un sistema lineal de ecuaciones de orden $n + 1$. Esta vía es poco eficiente para hallar el polinomio de interpolación pero lo es aun menos si lo que se desea es solamente hallar el valor interpolado para una x específica y no interesa obtener el polinomio $p(x)$.

El método de Lagrange brinda un algoritmo eficiente para hallar el polinomio interpolador pero, además, permite encontrar el valor interpolado para una x específica sin necesidad de hallar la expresión analítica del polinomio interpolador.

Sean x_0, x_1, \dots, x_n , $n + 1$ nodos de interpolación diferentes y $f(x)$ la función a interpolar. Sean

$$y_i = f(x_i) \quad i = 0, 1, 2, \dots, n$$

los valores de f en los nodos. El método de Lagrange consiste en encontrar $n + 1$ polinomios básicos de grado n :

$$L_0(x), L_1(x), \dots, L_n(x)$$

que satisfagan las siguientes condiciones:

$$\begin{array}{lllll} L_0(x_0) = 1 & L_1(x_0) = 0 & L_2(x_0) = 0 & \cdots & L_n(x_0) = 0 \\ L_0(x_1) = 0 & L_1(x_1) = 1 & L_2(x_1) = 0 & \cdots & L_n(x_1) = 0 \\ L_0(x_2) = 0 & L_1(x_2) = 0 & L_2(x_2) = 1 & \cdots & L_n(x_2) = 0 \\ \vdots & \vdots & \vdots & & \vdots \\ L_0(x_n) = 0 & L_1(x_n) = 0 & L_2(x_n) = 0 & \cdots & L_n(x_n) = 1 \end{array}$$

Como se verá enseguida, es muy fácil construir estos polinomios básicos; una vez obtenidos, el polinomio:

$$p(x) = y_0L_0(x) + y_1L_1(x) + \dots + y_nL_n(x) \quad (1)$$

cumple las condiciones:

$$\begin{aligned} p(x_0) &= y_0L_0(x_0) + y_1L_1(x_0) + \dots + y_nL_n(x_0) = y_0 \cdot 1 + y_1 \cdot 0 + \dots + y_n \cdot 0 = y_0 \\ p(x_1) &= y_0L_0(x_1) + y_1L_1(x_1) + \dots + y_nL_n(x_1) = y_0 \cdot 0 + y_1 \cdot 1 + \dots + y_n \cdot 0 = y_1 \\ &\vdots \\ p(x_n) &= y_0L_0(x_n) + y_1L_1(x_n) + \dots + y_nL_n(x_n) = y_0 \cdot 0 + y_1 \cdot 0 + \dots + y_n \cdot 1 = y_n \end{aligned}$$

Además, como $p(x)$ es una combinación lineal de polinomios de grado n su grado será menor o igual que n y es, por tanto, el polinomio de interpolación buscado.

Para hallar el polinomio básico $L_i(x)$, $i = 0, 1, 2, \dots, n$, nótese que el mismo posee n ceros:

$$x_0, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$$

o sea, cada nodo de interpolación excepto x_i es un cero de $L_i(x)$, por tanto:

$$L_i(x) = K(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n) \quad (2)$$

donde K es un coeficiente aún por determinar. Como se debe cumplir que $L_i(x_i) = 1$, el valor de K se hallará de modo que esta condición se satisfaga. De aquí resulta:

$$L_i(x_i) = K(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n) = 1$$

Despejando K :

$$K = \frac{1}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}$$

y sustituyendo esta expresión para K en la ecuación (2):

$$L_i(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} \quad i = 0, 1, 2, \dots, n$$

Ejemplo 1

De la función $f(x)$ se conoce que: $f(1) = 2$, $f(2) = 3$ y $f(4) = 1$.

- a) Halle el polinomio interpolador $p(x)$ correspondiente a estos nodos.
- b) Halle directamente $p(2,5)$ sin utilizar $p(x)$.

Solución:

- a) Como se tienen tres nodos de interpolación, el grado del polinomio interpolador será menor o igual que 2. Para hallar $p(x)$ primero es necesario calcular los tres polinomios de segundo grado $L_0(x)$, $L_1(x)$ y $L_2(x)$.

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x - 2)(x - 4)}{(1 - 2)(1 - 4)} = \frac{1}{3}(x - 2)(x - 4)$$

$$L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x - 1)(x - 4)}{(2 - 1)(2 - 4)} = -\frac{1}{2}(x - 1)(x - 4)$$

$$L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(x - 1)(x - 2)}{(4 - 1)(4 - 2)} = \frac{1}{6}(x - 1)(x - 2)$$

Es fácil comprobar que:

$$\begin{array}{lll} L_0(1) = 1 & L_1(1) = 0 & L_2(1) = 0 \\ L_0(2) = 0 & L_1(2) = 1 & L_2(2) = 0 \\ L_0(4) = 0 & L_1(4) = 0 & L_2(4) = 1 \end{array}$$

El polinomio interpolador es:

$$p(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x)$$

donde $y_0 = 2$, $y_1 = 3$, $y_2 = 1$. De esto resulta:

$$p(x) = 2 \cdot \frac{1}{3}(x-2)(x-4) + 3 \cdot (-\frac{1}{2})(x-1)(x-4) + 1 \cdot \frac{1}{6}(x-1)(x-2)$$

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

Véase que, efectivamente:

$$\begin{aligned} p(x_0) &= p(1) = -\frac{2}{3} + 3 - \frac{1}{3} = 2 = y_0 \\ p(x_1) &= p(2) = -\frac{2}{3} \cdot 4 + 3 \cdot 2 - \frac{1}{3} = 3 = y_1 \\ p(x_2) &= p(4) = -\frac{2}{3} \cdot 16 + 3 \cdot 4 - \frac{1}{3} = 1 = y_2 \end{aligned}$$

- b) Para hallar $p(2,5)$ bastaría con evaluar el polinomio del inciso a). Sin embargo puede calcularse $p(2,5)$ sin tener que encontrar previamente $p(x)$. Esta forma de proceder tiene la ventaja de que no es necesario realizar operaciones algebraicas, sino que todo el cálculo es puramente con números. Todo lo que se requiere es tomar $x = 2,5$ desde un principio.

$$L_0(2,5) = \frac{(2,5-x_1)(2,5-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(2,5-2)(2,5-4)}{(1-2)(1-4)} = -0,25$$

$$L_1(2,5) = \frac{(2,5-x_0)(2,5-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(2,5-1)(2,5-4)}{(2-1)(2-4)} = -1,125$$

$$L_2(2,5) = \frac{(2,5-x_0)(2,5-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(2,5-1)(2,5-2)}{(4-1)(4-2)} = 0,125$$

$$p(2,5) = y_0 L_0(2,5) + y_1 L_1(2,5) + y_2 L_2(2,5) = 2(-0,25) + 3(1,125) + 1(0,125) = 3$$

Algoritmo en seudo código

El algoritmo que sigue permite calcular $p(x)$ donde p es el polinomio interpolador y x es un valor numérico. El algoritmo funciona también si x es una variable pero en este caso será necesario realizar operaciones simbólicas (literales) de suma y multiplicación. Se suponen conocidos los nodos de interpolación: x_0, x_1, \dots, x_n y los valores correspondientes de la función a interpolar: y_0, y_1, \dots, y_n . El resultado será un número $p(x)$ si x es un número o el polinomio $p(x)$ si x es un literal.

```

Resultado := 0
for  $i = 0$  to  $n$ 
     $L := 1$ 
    for  $j = 0$  to  $n$ 
        if  $j \neq i$  then
             $L := L \cdot \frac{x - x_j}{x_i - x_j}$ 
        end
    end
    Resultado := Resultado +  $y_i \cdot L$ 
end
Terminar

```

Ejemplo 2

En la tabla 1 se muestran datos que corresponden a la función

$$f(x) = \frac{1}{1+25x^2}$$

evaluada para 13 valores equidistantes de x . En la figura 1 se muestra la gráfica de la función junto con los 13 puntos de la tabla. Suponga que la función $f(x)$ no se conoce, sino solamente los puntos de la tabla. Utilice interpolación polinomial para calcular aproximadamente $f(0,35)$ utilizando como nodos:

- a) $\{-0,6; -0,4; -0,2; 0; 0,2; 0,4; 0,6\}$
- b) $\{0,2; 0,3; 0,4; 0,5\}$

Compare los resultados obtenidos con el valor $f(0,35)$. Para explicar los resultados obtenidos, utilice las graficas de los polinomios interpoladores correspondientes a los nodos en a) y en b).

x	$f(x)$	x	$f(x)$	x	$f(x)$
-0.6	0.100000	-0.1	0.800000	0.4	0.200000
-0.5	0.137931	0	1.000000	0.5	0.137931
-0.4	0.200000	0.1	0.800000	0.6	0.100000
-0.3	0.307692	0.2	0.500000		
-0.2	0.500000	0.3	0.307692		

Tabla 1

$$f(x) = \frac{1}{1+25x^2}$$

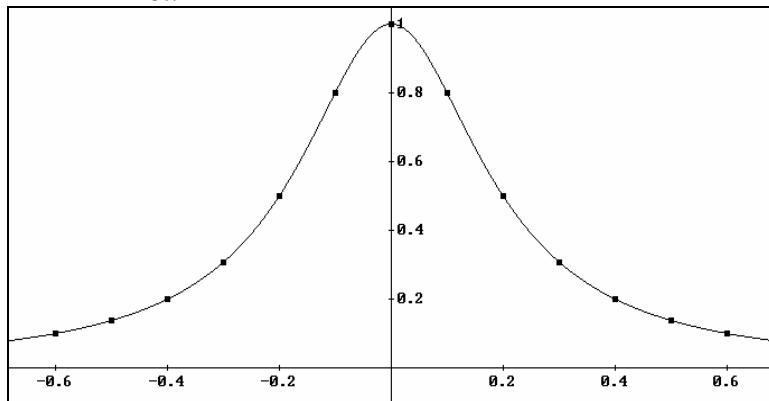


Figura 1

Solución:

- a) Utilizando los 7 nodos de interpolación $\{-0,6; -0,4; -0,2; 0; 0,2; 0,4; 0,6\}$, lo cual significa

interpolación de sexto grado, se obtiene como resultado: $p_1(0,35) = 0.159607$. El verdadero valor de la función para $x = 0,35$ es: $f(0,35) = 0.246154$. Se obtiene un error de interpolación tan grande que el valor interpolado no posee ni siquiera una cifra exacta.

- b) Utilizando los 4 nodos de interpolación $\{0,2; 0,3; 0,4; 0,5\}$, lo que quiere decir, interpolación de tercer grado, se obtiene $p_2(0,35) = 0.245307$, resultado cuyo error de interpolación es menor que 0,001.

La diferencia entre la forma de proceder en a) y en b) es que en el primer caso se tomaron muchos nodos y muy dispersos, algunos de ellos muy alejados del valor $x = 0,35$ donde se deseaba aproximar la función. En el inciso b), en cambio, se tomaron menos nodos, pero más concentrados en la proximidad de $x = 0,35$. En este caso, a pesar de que el polinomio interpolador es de menor grado, se obtiene un resultado mucho más exacto que en a).

Utilizando el método de Lagrange fueron obtenidos los polinomios de interpolación correspondientes a los incisos a) y b):

$$p_1(x) = -156,25x^6 + 93,75x^4 - 16x^2 + 1$$

$$p_2(x) = -6,498673x^3 + 10,079575x^2 - 5,728116x + 1,294030$$

En la figura 2 se muestran los 13 puntos de la tabla 1 y las graficas de los polinomios p_1 y p_2 cada una en el intervalo donde se encuentran sus nodos de interpolación.

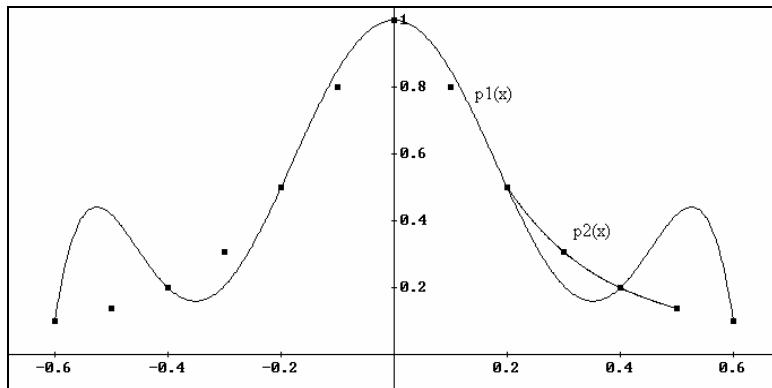


Figura 2

Aunque la función $f(x)$ fue seleccionada para este ejemplo porque en ella este efecto se aprecia con mucha claridad, la conclusión que se deriva tiene carácter general. La interpolación polinomial no debe usarse como una técnica global para representar a una función en un intervalo grande utilizando muchos nodos, todo lo contrario, se trata de un procedimiento para aproximar *localmente* una función mediante un polinomio. Por tanto, la estrategia debe ser seleccionar pocos nodos de interpolación concentrados en la región donde se desea aproximar la función. Si el intervalo en que se desea aproximar la función es muy extenso, en lugar de utilizar un polinomio interpolador de grado alto, siempre es preferible aproximar la función por varios polinomios cada uno en un pequeño tramo de la función. Esta estrategia, llamada interpolación por tramos será objeto de estudio en la sección 4.5.

Ejemplo 3

La tabla 2 muestra un conjunto de valores de la función exponencial $f(x) = e^x$ con incrementos 0,05. Halle aproximadamente $e^{1,07}$ utilizando interpolación de tercer grado. Halle una cota del error de interpolación y compárela con el verdadero error de interpolación.

x	$f(x)$	x	$f(x)$	x	$f(x)$
1,00	2,718282	1,10	3,004166	1,20	3,320117
1,05	2,857651	1,15	3,158193	1,25	3,490342

Tabla 2

Solución:

Para obtener interpolación de grado 3 se requieren 4 nodos. Como el punto en que se interpolará es $x = 1,07$ se tomarán los cuatro nodos más próximos: $\{1,00; 1,05; 1,10; 1,15\}$. El resultado se obtuvo con un programa confeccionado según el seudo código del método de Lagrange. Se obtuvo:

$$p(1,07) = 2.915379$$

Para hallar una cota del error, se emplea la fórmula (7) de la sección 4.2:

$$R(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x - x_0)(x - x_1)\cdots(x - x_n)$$

Como $n = 3$, se tiene que

$$R(x) = \frac{f^{(4)}(c)}{4!} (x - x_0)(x - x_1)(x - x_2)(x - x_3)$$

Tomando en cuenta que $f^{(4)}(x) = e^x$ y que $x = 1,07$, resulta:

$$R(1,07) = \frac{e^c}{24} (1,07 - 1,00)(1,07 - 1,05)(1,07 - 1,10)(1,07 - 1,15)$$

Acerca del número c , solo se sabe que se encuentra en el intervalo $[1,00; 1,15]$. Como e^x es una función creciente, se tiene que:

$$e^c \leq e^{1,15} = 3,158193 < 3,16$$

Por tanto, $|R(1,07)| < \frac{3,16}{24} |(1,07 - 1,00)(1,07 - 1,05)(1,07 - 1,10)(1,07 - 1,15)| = 0,00000044$

Lo cual significa que el resultado obtenido posee seis cifras decimales exactas. A este error debido a la interpolación, habría que añadirle el error debido al redondeo de los datos (que han venido dados con seis cifras exactas); debido a esto, el resultado puede haber perdido una cifra decimal de exactitud. Usando una calculadora científica, el verdadero valor de la función es:

$$f(1,07) = 2.9153795$$

Es decir, que el verdadero error del resultado es 0,0000005, que es la suma del error de redondeo más el de interpolación.

Ejercicios

1. Se conocen los siguientes pares de valores de una función $y = f(x)$:

x	0	1	3	4	5
y	1	-3	25	149	381

- a) Aproxime y para $x = 2$ mediante interpolación lineal.
- b) Aproxime y para $x = 2$ mediante interpolación cuadrática.
- 2. Se conocen valores de la función $f(x)$: $f(0) = 3$, $f(0,5) = 2$, $f(2) = 2$, $f(3) = 4$
 - a) Halle aproximadamente $f(1)$ mediante interpolación cúbica (sin hallar el polinomio interpolador).
 - b) Halle el polinomio cúbico de interpolación correspondiente a los nodos dados.
- 3. A partir de los valores de la función $\tan x$ en los ángulos notables (en radianes): $\frac{\pi}{6}$, $\frac{\pi}{4}$ y $\frac{\pi}{3}$ determine, aproximadamente, la tangente de 1 radián y dé un estimado del error de interpolación cometido.
- 4. Halle el polinomio interpolador de la función $y = \sqrt{x}$ tomando como nodos de interpolación: 0, 1, 4 y 9. Con el polinomio hallado encuentre una aproximación para $\sqrt{2}$. Repita ahora la operación pero utilizando solamente los nodos 1, 4 y 9. Compare la exactitud de las aproximaciones obtenidas. ¿A qué conclusión usted llega?
- 5. Halle un polinomio interpolador de segundo grado que aproxime a la función $y = \cosh x$ en el intervalo $[0, 1]$ y dé una cota para el error de interpolación.
- 6. Halle un polinomio interpolador de tercer grado que aproxime a la función $y = \ln x$ en el intervalo $[1; 1,5]$ y dé una cota para el error de interpolación.
- 7. Dada la función $f(x) = x - \cos x$ (x en radianes) se construye la tabla

x	0	0,5	1
y	-1	-0,37758	0,45970

Tome los valores de y como nodos de interpolación y halle aproximadamente el valor de x para $y = 0$ (a esta forma de proceder se le llama *interpolación inversa* y aquí se está empleando para resolver aproximadamente la ecuación $x - \cos x = 0$). ¿Cómo podría mejorar la solución hallada mediante este mismo procedimiento?

8. Se sabe que la suma de los cuadrados de los n primeros números naturales es un polinomio de tercer grado en n . Hállelo utilizando interpolación.

9. El algoritmo que sigue se supone que permite interpolar en un valor x conocidos los nodos de interpolación: x_0, x_1, \dots, x_n y los valores correspondientes de la función a interpolar: y_0, y_1, \dots, y_n . Sin embargo, el algoritmo contiene un grave error. Detéctelo y corríjalo.

```

Resultado := 0
for i = 0 to n
    L := 1
    for j = 0 to n
        L := L ·  $\frac{x - x_j}{x_i - x_j}$ 
    end
    Resultado := Resultado +  $y_i \cdot L$ 
end
Terminar

```

10. Aunque por lo general la aceleración de los cuerpos en caída libre se toma como una constante $g = 9,8 \text{ m/s}^2$, se sabe que su valor depende de varios factores, entre ellos la latitud geográfica y la altura sobre el nivel del mar. En la tabla que sigue se ofrecen los valores de g a nivel del mar para varias latitudes geográficas. Halle aproximadamente el valor de g en la ciudad de La Habana, que está a orillas del mar a 23° de latitud Norte.

Latitud	0°	10°	20°	30°	40°	50°
$g (\text{m/s}^2)$	9,7803	9,7819	9,7863	9,7932	9,8017	9,8107

11. La viscosidad de un fluido se mide en poises: 1 poise = $1 \frac{\text{g}\cdot\text{cm}}{\text{cm}\cdot\text{s}}$. La viscosidad varía muy marcadamente con la temperatura. En la tabla que sigue se da la viscosidad (multiplicada por 10^5) del agua para algunas temperaturas. Halle aproximadamente la viscosidad del agua a 27° grados centígrados.

Temperatura ($^\circ\text{C}$)	10	15	20	25	30	40
Viscosidad por 10^5 (poises)	1307	1140	1004	895	803	655

12. Cuando un alambre eléctrico de cobre (forrado) conduce una corriente eléctrica excesiva durante un tiempo prolongado, se corre el riesgo de que se dañe y puede, incluso, provocar un accidente. La siguiente tabla da los valores de corriente (en ampere) permisibles durante una operación prolongada para conductores de cobre de diferente sección transversal (en milímetros cuadrados). Halle aproximadamente, la corriente permisible para un alambre de cobre de 2 mm^2 de sección transversal.

Área de la sección transversal (mm^2)	1	1,5	2,5	4	6
Corriente permisible (ampere)	11	14	20	25	31

4.4 El método de Newton

La idea fundamental del método de Newton es realizar la interpolación en un punto de forma sucesiva: partiendo de dos nodos ir agregando los demás, uno por uno, en el orden que se desee, de tal manera que en cada paso solo se requiera agregar un nuevo término a los cálculos precedentes. El método permite, sin realizar ninguna operación adicional, ir obteniendo en cada paso del proceso una estimación del error de interpolación, de manera que el proceso iterativo se pueda detener si se alcanza un error suficientemente pequeño.

Sea $f(x)$ la función a interpolar, $p_{n-1}(x)$ el polinomio interpolador (de grado menor o igual que $n - 1$) correspondiente a los nodos $\{x_0, x_1, \dots, x_{n-1}\}$ y $p_n(x)$ al polinomio interpolador (de grado menor o igual que n) que corresponde a los nodos $\{x_0, x_1, \dots, x_{n-1}, x_n\}$. Se supone, igual que hasta ahora, que todos los nodos de interpolación son diferentes, aunque no tienen que estar ordenados de ninguna forma. Con el propósito de simplificar la notación, se llamará $y_i = f(x_i)$ para $i = 0, 1, 2, \dots$.

Como el grado de $p_n(x)$ es menor o igual que n y el de $p_{n-1}(x)$ es menor o igual que $n - 1$, se tiene que:

$$p_n(x) = p_{n-1}(x) + C_n(x) \quad (1)$$

donde $C_n(x)$ es algún polinomio de grado menor o igual que n . Por otra parte, como ambos polinomios interpolan a $f(x)$ se tiene que:

$$\begin{array}{ll} p_{n-1}(x_0) = y_0 & p_n(x_0) = y_0 \\ p_{n-1}(x_1) = y_1 & p_n(x_1) = y_1 \\ \vdots & \vdots \\ p_{n-1}(x_{n-1}) = y_{n-1} & p_n(x_{n-1}) = y_{n-1} \\ & p_n(x_n) = y_n \end{array}$$

Como, según (1), $C_n(x) = p_n(x) - p_{n-1}(x)$,

resulta que: $C_n(x_0) = C_n(x_1) = C_n(x_2) = \dots = C_n(x_{n-1}) = 0$

El hecho de que $C_n(x)$ es un polinomio de grado menor o igual que n asegura que no puede tener más que estos n ceros. Por esta razón será:

$$C_n(x) = a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

con lo cual la expresión (1) toma la forma:

$$p_n(x) = p_{n-1}(x) + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \quad (2)$$

El coeficiente a_n queda determinado por la condición:

$$p_n(x_n) = y_n$$

Con el objetivo de llegar a una formulación general es conveniente comenzar a partir de $n = 1$. En ese caso la expresión (2) es:

$$p_1(x) = p_0(x) + a_1(x - x_0) \quad (3)$$

donde $p_0(x)$ es un polinomio de grado cero (una constante) determinado por un solo nodo; esto es:

$$p_0(x) = y_0$$

y a_1 se halla a partir de: $p_1(x_1) = y_1$

Evaluando para $x = x_1$ en la fórmula (3):

$$y_1 = y_0 + a_1(x_1 - x_0)$$

de donde:

$$a_1 = \frac{y_1 - y_0}{x_1 - x_0}$$

El cociente anterior se denomina primera diferencia dividida de la función f para los puntos x_0 y x_1 y se denota $f[x_0, x_1]$, esto es:

$$f[x_0, x_1] = \frac{y_1 - y_0}{x_1 - x_0}$$

Con esta notación, la ecuación (3) se transforma en:

$$p_1(x) = p_0(x) + f[x_0, x_1](x - x_0) \quad (4)$$

Para obtener el polinomio $p_2(x)$ se llevará a cabo un procedimiento similar:

$$p_2(x) = p_1(x) + a_2(x - x_0)(x - x_1) \quad (5)$$

y como

$$p_1(x) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0)$$

$$\text{resulta: } p_2(x) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + a_2(x - x_0)(x - x_1)$$

$$\text{Evaluando para } x = x_2: \quad y_2 = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1)$$

de donde, restando y sumando y_1 , se obtiene:

$$a_2(x_2 - x_0)(x_2 - x_1) = y_2 - y_1 + y_1 - y_0 - \frac{y_1 - y_0}{x_1 - x_0}(x_2 - x_0)$$

Agrupando términos:

$$a_2(x_2 - x_0)(x_2 - x_1) = y_2 - y_1 + (y_1 - y_0) \left[1 - \frac{x_2 - x_0}{x_1 - x_0} \right]$$

$$\text{Efectuando: } a_2(x_2 - x_0)(x_2 - x_1) = y_2 - y_1 - (y_1 - y_0) \left[\frac{x_2 - x_1}{x_1 - x_0} \right]$$

$$\text{Dividiendo por } (x_2 - x_1): \quad a_2(x_2 - x_0) = \frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0} = f[x_1, x_2] - f[x_0, x_1]$$

y, finalmente:

$$a_2 = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

A esta diferencia dividida de dos primeras diferencias divididas de f , se le llama segunda diferencia dividida de f para los puntos x_0, x_1, x_2 y se denotará por $f[x_0, x_1, x_2]$; es decir:

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

De acuerdo con esto, la ecuación (5) se expresa ahora por:

$$p_2(x) = p_1(x) + f[x_0, x_1, x_2] (x - x_0)(x - x_1)$$

Repetiendo este proceso se llega, en general, a:

$$p_n(x) = p_{n-1}(x) + f[x_0, x_1, \dots, x_n] (x - x_0)(x - x_1) \cdots (x - x_{n-1}) \quad (6)$$

donde, las diferencias divididas de cualquier orden se definen recursivamente por:

$$f[x_0, x_1, \dots, x_k] = \frac{f[x_1, x_2, \dots, x_k] - f[x_0, x_1, \dots, x_{k-1}]}{x_k - x_0} \quad k = 1, 2, 3, \dots, n \quad (7)$$

La expresión $f[x_0, x_1, \dots, x_k]$ se lee “diferencia dividida de orden k de f respecto a los nodos x_0, x_1, \dots, x_k ”. Aplicando reiteradamente la fórmula (6), se obtiene para el polinomio interpolador una sugestiva expresión:

$$\begin{aligned} p_n(x) &= f(x_0) + f[x_0, x_1] (x - x_0) + f[x_0, x_1, x_2] (x - x_0)(x - x_1) + \cdots \\ &\quad \cdots + f[x_0, x_1, \dots, x_n] (x - x_0)(x - x_1) \cdots (x - x_{n-1}) \end{aligned} \quad (8)$$

que guarda una interesante analogía con el desarrollo de una función en polinomios de Taylor.

Cuando se trabaja manualmente resulta cómodo organizar los datos como muestra la tabla 1 para calcular las diferencias divididas necesarias.

x	$f(x)$	1 ^a dif.	2 ^a dif.	3 ^a dif.
x_0	$f(x_0)$	$f[x_0, x_1]$	$f[x_0, x_1, x_2]$	$f[x_0, x_1, x_2, x_3]$
x_1	$f(x_1)$	$f[x_1, x_2]$	$f[x_1, x_2, x_3]$	$f[x_1, x_2, x_3, x_4]$
x_2	$f(x_2)$	$f[x_2, x_3]$	$f[x_2, x_3, x_4]$	$f[x_2, x_3, x_4, x_5]$
x_3	$f(x_3)$	$f[x_3, x_4]$	$f[x_3, x_4, x_5]$	$f[x_3, x_4, x_5, x_6]$
x_4	$f(x_4)$	$f[x_4, x_5]$	$f[x_4, x_5, x_6]$	\vdots
x_5	$f(x_5)$	$f[x_5, x_6]$	\vdots	
x_6	$f(x_6)$	\vdots		
\vdots	\vdots			

Tabla 1

Nótese que las diferencias necesarias para la aplicación de las fórmula (6) u (8) del método de Newton, aparecen en la tabla 1 en una misma fila, como la que aparece señalada en negritas.

Estimación del error de interpolación

El error de interpolación cometido cuando se utiliza el método de Newton, es el mismo que cuando se usa cualquier otro método, ya que el polinomio interpolador es único para un conjunto de nodos y una función dados. Por tanto, se puede usar la fórmula (7) deducida en la sección 4.2:

$$R(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x - x_0)(x - x_1) \cdots (x - x_n)$$

Sin embargo, es más eficiente utilizar una expresión equivalente que puede calcularse directamente utilizando la tabla de diferencias divididas y no requiere el cálculo de derivadas, aunque se supone, igual que antes, que $f(x)$ posee derivadas hasta el orden $n + 1$ en un intervalo que incluye a todos los valores de x involucrados en la deducción que sigue.

Sea $p_n(x)$ el polinomio interpolador correspondiente a los nodos $\{x_0, x_1, \dots, x_n\}$. Se desea calcular o por lo menos estimar, el error de interpolación de $p_n(x)$ en el punto \hat{x} :

$$R_n(\hat{x}) = f(\hat{x}) - p_n(\hat{x})$$

Considérese ahora el polinomio interpolador de grado menor o igual que $n + 1$ que corresponde a los $n + 2$ nodos $\{x_0, x_1, \dots, x_n, \hat{x}\}$. De acuerdo con la fórmula (6):

$$p_{n+1}(x) = p_n(x) + f[x_0, x_1, \dots, x_n, \hat{x}] (x - x_0)(x - x_1) \cdots (x - x_n) \quad (9)$$

Como \hat{x} es uno de los nodos de interpolación del polinomio $p_{n+1}(x)$, se cumple que:

$$p_{n+1}(\hat{x}) = f(\hat{x}) \quad (10)$$

Por tanto, haciendo $x = \hat{x}$ en (9) y tomando en cuenta (10):

$$f(\hat{x}) = p_n(\hat{x}) + f[x_0, x_1, \dots, x_n, \hat{x}] (\hat{x} - x_0)(\hat{x} - x_1) \cdots (\hat{x} - x_n)$$

$$\text{Por tanto: } R_n(\hat{x}) = f(\hat{x}) - p_n(\hat{x}) = f[x_0, x_1, \dots, x_n, \hat{x}] (\hat{x} - x_0)(\hat{x} - x_1) \cdots (\hat{x} - x_n)$$

Como en esta expresión, \hat{x} representa un número arbitrario, no existe inconveniente por cambiarlo por x . Esto es:

$$R_n(x) = f[x_0, x_1, \dots, x_n, x] (x - x_0)(x - x_1) \cdots (x - x_n) \quad (11)$$

La fórmula (11) brinda el valor exacto del error de interpolación del polinomio $p_n(x)$ pero su valor no es calculable pues calcular la diferencia dividida $f[x_0, x_1, \dots, x_n, x]$ requeriría conocer $f(x)$ lo cual carece de sentido pues, precisamente se quiere calcular $p_n(x)$ porque no se dispone de $f(x)$. Si

en esta diferencia dividida se toma x_{n+1} en lugar de \hat{x} se obtiene una aproximación del error de interpolación. Por tanto, se define:

$$\tilde{R}_n(x) = f[x_0, x_1, \dots, x_n, x_{n+1}](x - x_0)(x - x_1) \cdots (x - x_n) \quad (12)$$

y puede esperarse que, si las diferencias divididas de orden $n + 1$ no son funciones que experimenten grandes cambios, se cumpla la aproximación:

$$\tilde{R}_n(x) \approx R_n(x)$$

Si la fórmula (6) se compara con la (12) entonces aquella se puede escribir como:

$$p_{n+1}(x) = p_n(x) + \tilde{R}_n(x) \quad (13)$$

Es decir, el error estimado para $p_n(x)$, sumado a $p_n(x)$ da la aproximación $p_{n+1}(x)$, lo cual hace sumamente cómodo y eficiente el modo de estimar el error de interpolación cuando se utiliza el método de Newton.

Debe tenerse cuidado al ordenar los nodos para formar la tabla de diferencias divididas si se desea que las sucesivas aproximaciones que se van obteniendo tengan la mejor calidad. Así, si se desea que la aproximación $p_1(x)$ sea una buena aproximación, los nodos x_0 y x_1 deben ser escogidos de modo que estén lo más cerca posible y que entre ellos se encuentre el valor x donde se ha de interpolar. El nodo x_2 se tomará de tal modo que el conjunto $\{x_0, x_1, x_2\}$ esté lo más cerca posible del valor x y similarmente con los demás nodos que se incluyan en la tabla. De no procederse así, puede suceder que algunas de las aproximaciones obtenidas en los pasos sucesivos del algoritmo correspondan a extrapolaciones en lugar de interpolaciones y se obtengan errores de interpolación muy exagerados.

Ejemplo 1

La tabla 2 reproduce a otra mostrada en la sección 4.1 y que contiene algunos valores de la función gamma. Utilizando esta tabla obtenga $\Gamma(1,17)$ con un error de interpolación menor que 0,00005 (cuatro cifras decimales exactas).

a	$\Gamma(a)$	a	$\Gamma(a)$	a	$\Gamma(a)$
1,00	1,00000	1,35	0,89115	1,70	0,90865
1,05	0,97350	1,40	0,88726	1,75	0,91906
1,10	0,95135	1,45	0,88565	1,80	0,93138
1,15	0,93304	1,50	0,88623	1,85	0,94561
1,20	0,91817	1,55	0,88887	1,90	0,96177
1,25	0,90640	1,60	0,89352	1,95	0,97988
1,30	0,89747	1,65	0,90012	2,00	1,00000

Tabla 2

Solución:

Como la interpolación nunca se hará con polinomios de grado alto, no es necesario incluir muchos nodos para formar la tabla de diferencias, se tomarán en este caso 6 nodos próximos a 1,17 y ordenados convenientemente: {1,15; 1,20; 1,10; 1,25; 1,05; 1,30}. La tabla 3 muestra las diferencias obtenidas en este caso. Realmente, no es necesario construir toda la tabla de una vez, es preferible ir calculando las diferencias en la medida en que van siendo necesarias. Aquí, por un problema de espacio, se muestran todas las diferencias, hasta la de quinto orden.

x	$f(x)$	1 ^a dif	2 ^a dif	3 ^a dif	4 ^a dif	5 ^a dif
1,15	0,93304	- 0,29740	0,68800	- 0,45333	0,40000	0,53333
1,20	0,91817	- 0,33180	0,64267	- 0,49333	0,48000	
1,10	0,95135	- 0,29967	0,71667	- 0,44533		
1,25	0,90640	- 0,33550	0,62760			
1,05	0,97350	- 0,30412				
1,30	0,89747					

Tabla 3

En la tabla aparecen en negrita las diferencias que se utilizarán en cada paso del algoritmo.

Aproximación con polinomio de grado 1:

$$p_1(x) = p_0(x) + f[x_0, x_1](x - x_0) = 0,93304 + (-0,29790)(1,17 - 1,15) = 0,92708$$

$$\tilde{R}_1(x) = f[x_0, x_1, x_2](x - x_0)(x - x_1) = (0,68800)(1,17 - 1,15)(1,17 - 1,20) = -0,00041$$

Como este error es muy grande, se calculará la siguiente aproximación:

Aproximación con polinomio de grado 2:

$$p_2(x) = p_1(x) + \tilde{R}_1(x) = 0,92708 - 0,00041 = 0,92667$$

$$\begin{aligned} \tilde{R}_2(x) &= f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2) = \\ &= (-0,45333)(1,17 - 1,15)(1,17 - 1,20)(1,17 - 1,10) = 0,00002 \end{aligned}$$

Como el error es suficientemente pequeño, se tomará $p_2(x)$ como la aproximación buscada. No obstante, puesto que ya se tiene calculada la estimación del error, puede utilizarse para tener un resultado aun mejor, solamente sumándolo a $p_2(x)$. Se toma pues:

$$\Gamma(1,17) = 0,92669$$

con cuatro cifras decimales exactas. El valor de $\Gamma(1,17)$ con 5 cifras decimales exactas que aparece en una tabla es 0,92670.

Relación entre diferencias y derivadas

Si se compara la fórmula (7) de la sección 4.2:

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x - x_0)(x - x_1) \cdots (x - x_n)$$

con la fórmula (11): $R_n(x) = f[x_0, x_1, \dots, x_n, x] (x - x_0)(x - x_1) \cdots (x - x_n)$

resulta claro que

$$\frac{f^{(n+1)}(c)}{(n+1)!} = f[x_0, x_1, \dots, x_n, x]$$

Si se considera que x es el nodo número $n + 1$ ($x = x_{n+1}$) y se toma $m = n + 1$, la ecuación anterior queda como:

$$\frac{f^{(m)}(c)}{m!} = f[x_0, x_1, \dots, x_m] \quad (14)$$

lo cual significa que existe algún número c en el intervalo I que incluya a todos los nodos $\{x_0, x_1, \dots, x_m\}$ donde la derivada de orden m es exactamente $m!$ veces la diferencia dividida que corresponde a esos nodos. Una consecuencia importante de esta relación entre diferencias divididas y derivadas, es el siguiente teorema:

Teorema 1

Las diferencias divididas de orden m de un polinomio de grado m son constantes (es decir, no dependen de los nodos seleccionados) y las de orden mayor que m son nulas.

Demostración:

En efecto, si p_m es un polinomio de grado m y $\{x_0, x_1, \dots, x_m\}$ son $m + 1$ nodos cualesquiera, entonces, de acuerdo con la igualdad (14), existe un punto c tal que:

$$\frac{p_m^{(m)}(c)}{m!} = p_m[x_0, x_1, \dots, x_m]$$

pero como la derivada de orden m de un polinomio de grado m es una constante K (no depende de x) entonces, no importa el valor de c :

$$p_m[x_0, x_1, \dots, x_m] = \frac{K}{m!}$$

independientemente de los nodos utilizados. Al ser iguales todas las diferencias divididas de orden m , las diferencias de orden $m + 1$ son todas cero y lo mismo sucede entonces con todas las de orden superior. ■

El resultado anterior se utiliza frecuentemente para detectar si un conjunto de pares (x, y) corresponden a una función polinómica; en efecto, si este es el caso, entonces alguna columna de la tabla de diferencias se hace constante y las siguientes se anulan. El orden de las diferencias

constantes indica el grado del polinomio de que se trata. Con un poco de imaginación, la idea se puede extender a funciones que no son polinómicas pero se *aproximan* a un polinomio de grado m ya que en ese caso la columna de las diferencias de grado m se hace *aproximadamente* constante.

Ejemplo 2

Investigue si los datos que muestra la tabla 4 se aproximarán adecuadamente con algún polinomio.

x	$f(x)$	x	$f(x)$	x	$f(x)$
2,0	3	4,0	107	5,5	336
2,5	14	4,5	165	6,0	451
3,0	34	5,0	240	6,5	591
3,5	64				

Tabla 4

Solución:

La tabla 5 muestra las diferencias divididas hasta el orden 5. Note que las diferencias de orden 1 y 2 muestran una clara tendencia creciente a medida que x crece. La columna correspondiente a las diferencias de tercer orden, en cambio oscila en un intervalo pequeño. Puede entonces afirmarse que la función se comporta aproximadamente como un polinomio de tercer grado y, por lo tanto sus valores serán bien aproximados utilizando interpolación polinómica de tercer grado.

x	$f(x)$	1 ^a dif	2 ^a dif	3 ^a dif	4 ^a dif	5 ^a dif
2,0	3	22	18	1,333	1,333	-0,800
2,5	14	40	20	4,000	-0,667	0,267
3,0	34	60	26	2,667	0,000	0,267
3,5	64	86	30	2,667	0,667	-0,888
4,0	107	116	34	4,000	-1,333	1,067
4,5	165	150	40	1,333	1,333	
5,0	240	190	42	4,000		
5,5	336	232	48			
6,0	451	280				
6,5	591					

Tabla 5

Algoritmo en seudo código

El algoritmo que sigue permite calcular $p(x)$ donde p es el polinomio interpolador y x es un valor numérico. Se suponen conocidos los nodos de interpolación: x_0, x_1, \dots, x_n , los valores correspondientes de la función a interpolar: y_0, y_1, \dots, y_n , el número x donde se desea interpolar y la tolerancia ϵ del error de interpolación la cual debe ser mucho mayor que el error por redondeo,

o por otras causas, que contengan los datos. El resultado será el número $p(x)$ y el error de interpolación estimado.

```

for  $i = 0$  to  $n$  {En este lazo, los valores de la función se transfieren al arreglo  $d_i$ 
 $d_i := f(x_i)$ 
end

for  $i = 1$  to  $n$  {A la salida de este lazo, cada variable  $d_i$  contiene la diferencia dividida
de orden  $i$  que corresponde a la primera fila de la tabla de diferencias}
 $j := n$ 
repeat
 $d_j := \frac{d_j - d_{j-1}}{x_j - x_{j-1}}$ 
 $j := j - 1$ 
until  $j < i$ 
end
 $p := d_0$ 
 $producto := (x - x_0)$ 
 $error := d_1 \cdot producto$ 
 $i := 1$ 
do while  $error > \varepsilon$  and  $i < n$  {El algoritmo termina cuando el error estimado es menor
que la tolerancia o cuando el grado de interpolación se
hace muy alto}
 $p := p + error$ 
 $producto := producto \cdot (x - x_i)$ 
 $i := i + 1$ 
 $error := d_i \cdot producto$ 
end
```

El resultado de la interpolación es p con error estimado de $error$
Terminar

Ejercicios

- Dada la siguiente tabla de valores de la función $f(x)$, obtenga $f(0,38)$ y $f(0,5)$ con tres cifras decimales exactas. Recuerde en cada caso ordenar la tabla adecuadamente.

x	0,30	0,35	0,40	0,48	0,55	0,60
$f(x)$	1,23114	1,27456	1,31951	1,39474	1,46409	1,51572

- A partir de la siguiente tabla de la función de Bessel de orden cero, obtenga una con incremento 1 en x en el intervalo $[1, 5]$ con dos cifras decimales exactas. Tenga presente en cada caso, la forma más conveniente de ordenar la tabla.

x	0,0	0,8	1,5	2,2	3,0	3,6	4,4	5,0
$f(x)$	1.0000	0,8463	0,5118	0,1104	-0,2601	-0,3918	-0,3423	-0,1776

3. A partir de los valores de la función $\cos x$ en los ángulos notables ($0^\circ, 30^\circ, 45^\circ, 60^\circ, 90^\circ$) construya una tabla de cosenos con incrementos de 10° con dos cifras decimales exactas. Tenga presente en cada caso, la forma más conveniente de ordenar la tabla.
4. Un programa para la inversión de matrices ha demorado los siguientes tiempos al aplicarse a matrices de diferentes órdenes:

Orden de la matriz	5	7	9	10	12	14	15	20
Tiempo (s)	14	44	97	133	232	368	452	1067

Se sabe que el tiempo de ejecución del algoritmo empleado varía polinomialmente con el orden de la matriz. Mediante una tabla de diferencias divididas halle de qué grado es este polinomio.

5. Las imágenes de la función $f(x) = (1,5)^x$ son muy fáciles de hallar si x es entero y positivo. Utilice este hecho y un proceso de interpolación para hallar aproximadamente $f(2,7)$ con dos cifras decimales exactas.
6. La probabilidad de que una variable aleatoria con distribución normal, media cero y varianza 1, sea, en valor absoluto, menor que x , viene dada por la integral:

$$\Phi(x) = \sqrt{\frac{2}{\pi}} \int_0^x e^{-\frac{1}{2}t^2} dt$$

Esta integral no puede ser calculada por métodos analíticos. En el capítulo 5 se estudiarán los métodos numéricos para calcularla. Por esa vía se calculan y tabulan los valores de probabilidad para valores de x con incremento de 0,1, tal como muestra el segmento de tabla que sigue. Halle $\Phi(0,323)$ con tres cifras decimales exactas.

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,30	0,2358	0,33	0,2586	0,36	0,2812
0,31	0,2434	0,34	0,2661	0,37	0,2886
0,32	0,2510	0,35	0,2737	0,38	0,2961

7. Para hallar la longitud de la elipse

$$\begin{aligned} x &= 3 \cos t \\ y &= 4 \sin t \quad 0 \leq t \leq 2\pi \end{aligned}$$

mediante la fórmula:

$$L = 4 \int_0^{\frac{\pi}{2}} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt$$

se llega a la expresión:

$$L = 16 \int_0^{\frac{\pi}{2}} \sqrt{1 - \left(\frac{7}{16}\right) \sin^2 t} dt$$

donde aparece una integral elíptica de segunda especie:

$$\int_0^{\frac{\pi}{2}} \sqrt{1 - k^2 \sin^2 t} dt \quad 0 < k < 1$$

las cuales vienen tabuladas para diferentes valores del parámetro $\alpha = \arcsen k$. A continuación se reproduce un fragmento de dicha tabla. Utilice interpolación para calcular la longitud de la elipse con tres cifras decimales exactas.

$\alpha = \arcsen k$	39°	40°	41°	42°	43°	44°
$\int_0^{\frac{\pi}{2}} \sqrt{1 - k^2 \sin^2 t} dt$	1,4013	1,3931	1,3849	1,3765	1,3680	1,3594

8. El calor específico de un material es la cantidad de calor (en calorías) que se necesita para elevar un grado centígrado la temperatura de un gramo del material. El calor específico de un material depende de la temperatura. En la tabla que sigue se da el calor específico c_p del agua a diferentes temperaturas. Calcule, con cuatro cifras decimales exactas el calor específico del agua a 14°C .

Temperatura ($^\circ\text{C}$)	0°	5°	10°	15°	20°	30°
c_p (cal/g·grado)	1,0104	1,0063	1,0033	1,0013	1,0000	0,9986

9. En el algoritmo en seudo código del método de Newton, se utilizó el segmento que sigue para calcular las diferencias divididas que este método requiere.

```

for  $i = 1$  to  $n$  {A la salida de este lazo, cada variable  $d_i$  contiene la diferencia dividida de orden  $i$  que corresponde a la primera fila de la tabla de diferencias}
     $j := n$ 
    repeat
         $d_j := \frac{d_j - d_{j-1}}{x_j - x_{j-1}}$ 
         $j := j - 1$ 
    until  $j < i$ 
end

```

Explique cómo funciona este segmento del algoritmo.

10. La tensión superficial de un líquido varía en general con la temperatura y viene tabulada en los manuales. La tabla que sigue muestra la tensión superficial del alcohol etílico (en dinas por centímetro) para diferentes temperaturas. Halle la tensión superficial del alcohol etílico a 34°C y dé un estimado del error debido a la interpolación.

Temperatura ($^\circ\text{C}$)	0°	30°	60°	90°	120°	150°
Tensión superficial (dina/cm)	24,4	21,9	19,2	16,4	13,4	10,1

11. El sonido se refleja en las superficies de diferente manera según el material de que están formadas. El coeficiente de absorción del sonido se define como el cociente de la energía absorbida entre la energía que incide sobre la superficie reflectante. El coeficiente de absorción varía con la frecuencia del sonido. En la tabla que sigue se ofrece el coeficiente de absorción de una pared de ladrillos para diferentes frecuencias. Determine el coeficiente de absorción de este material para un sonido de 440 ciclo/s (que corresponde a la nota La del centro del teclado del piano). Dé un estimado del error de interpolación.

Frecuencia (ciclo/s)	125	250	500	1000	2000	4000
Coeficiente de absorción	0,024	0,025	0,032	0,041	0,049	0,07

4.5 Interpolación mediante splines

El problema de la interpolación global

Ya en la sección 4.3 (vea el ejemplo 2 de esa sección) se llegó a la conclusión de que la interpolación polinomial no es una técnica adecuada para representar globalmente una función en un intervalo largo, todo lo contrario, se trata de un procedimiento para buscar un polinomio de grado reducido, que represente localmente a una función de forma adecuada dentro de un pequeño intervalo. Sin embargo, a veces se presenta la necesidad de aproximar globalmente una función, de la cual se conocen valores aislados, mediante una función sencilla. Ya esto fue visto en el ejemplo 3 de la sección 4.1 donde se deseaba aproximar el perfil de una pieza mediante una curva suave. En la construcción de gráficos por computadora, este problema se presenta con gran frecuencia.

Una solución muy elemental al problema de la interpolación global es descomponerla en varios problemas de interpolación local. Por ejemplo, si se desea hallar una función interpoladora de $f(x)$ para los nueve nodos ordenados $\{x_0, x_1, \dots, x_8\}$ el problema se puede descomponer en cuatro problemas sencillos: hallar cuatro polinomios interpoladores, cada uno de grado menor o igual que dos, para los conjuntos de nodos $\{x_0, x_1, x_2\}$, $\{x_2, x_3, x_4\}$, $\{x_4, x_5, x_6\}$, $\{x_6, x_7, x_8\}$. En este caso, la gráfica de la función interpoladora, como muestra la figura 1, estará formada por cuatro arcos

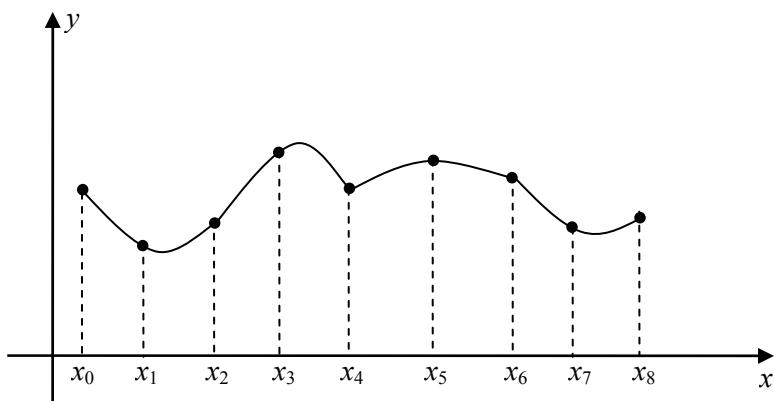


Figura 1

de paráolas de eje vertical; como dos arcos consecutivos comparten un nodo, los cuatro arcos formarán una curva continua que puede usarse para representar la función $f(x)$. Cada polinomio local se determina por cualquiera de los métodos vistos en las secciones anteriores, de modo que no se requiere ninguna teoría adicional para realizar este enfoque. Sin embargo, ya el lector habrá observado que la función que resulta no es suave, es decir, en los nodos que separan un polinomio local de otro, se produce un cambio de derivada, que se presenta en la gráfica como un punto anguloso. En algunos casos esto carece de importancia pero muchas veces se requiere obtener una curva que, además de continua, sea suave (derivada continua) e incluso que posea continuidad en la curvatura (segunda derivada continua). Si es así, este enfoque resulta insatisfactorio.

Funciones spline

En términos muy generales, una función spline es una función polinomial por tramos que es continua y posee derivadas continuas hasta un cierto orden. Además de las condiciones de continuidad y suavidad, el spline deberá satisfacer algunas otras condiciones adecuadas al problema que se desea resolver: pasar por un conjunto de puntos de la gráfica de $f(x)$ (spline interpolador), aproximarse a un conjunto de puntos experimentales (spline de mejor ajuste), cumplir ciertos requerimientos estéticos y además en cuanto al valor en algunos puntos de control (problemas de diseño gráfico), etc. Para lograr todas estas condiciones, el spline contiene un conjunto de parámetros cuyos valores se escogen de forma que se satisfagan todas las condiciones deseadas. Para precisar ideas, supóngase un conjunto de $n + 1$ números ordenados en forma creciente $\{x_0, x_1, \dots, x_n\}$ y que se utilicen polinomios de grado k , entonces el spline $s(x)$ es una función de la forma:

$$s(x) = \begin{cases} p_1(x) & \text{si } x_0 \leq x < x_1 \\ p_2(x) & \text{si } x_1 \leq x < x_2 \\ \vdots \\ p_n(x) & \text{si } x_{n-1} \leq x \leq x_n \end{cases}$$

donde $p_i(x)$ ($i = 1, 2, \dots, n$) representa un polinomio de grado k . Como un polinomio de grado k posee $k + 1$ coeficientes, el spline en su conjunto posee $n(k + 1)$ coeficientes y podrá satisfacer esa misma cantidad de condiciones siempre que las mismas no encierren contradicciones que las hagan incompatibles. El hecho de que $s(x)$ debe ser continua en todos los nodos interiores $\{x_1, x_2, \dots, x_{n-1}\}$ representan ya $n - 1$ condiciones. Para lograr que el spline posea además varias derivadas continuas es necesario tomar un grado k lo suficientemente elevado de manera que la cantidad de parámetros permita satisfacer todas las condiciones requeridas.

Dado el interés limitado de este texto, aquí solo se considerará el spline como función de interpolación y formado por polinomios de grado menor o igual que 3, es decir, el spline cúbico interpolador. En la bibliografía recomendada al final de este capítulo se incluyen referencias a otras fuentes donde aparecen tratados los splines con otros objetivos.

El spline cúbico de interpolación

Considérese que para cada uno de los $n + 1$ nodos ordenados en forma creciente $\{x_0, x_1, \dots, x_n\}$ se conoce el valor de una función $f(x)$. Sea

$$y_i = f(x_i) \quad i = 0, 1, 2, \dots, n$$

y se necesita que el spline satisfaga las condiciones de interpolación global:

$$s(x_i) = y_i \quad i = 0, 1, 2, \dots, n \quad (1)$$

Como se trata de un spline cúbico, su expresión analítica será:

$$s(x) = \begin{cases} a_1x^3 + b_1x^2 + c_1x + d_1 & \text{si } x_0 \leq x < x_1 \\ a_2x^3 + b_2x^2 + c_2x + d_2 & \text{si } x_1 \leq x < x_2 \\ \vdots \\ a_nx^3 + b_nx^2 + c_nx + d_n & \text{si } x_{n-1} \leq x \leq x_n \end{cases} \quad (2)$$

Aquí se está suponiendo que los nodos de interpolación coinciden con los puntos que limitan los tramos del spline; en un enfoque más general, esto no tendría que ser así. Sin embargo este es el caso más simple y más frecuente y a él se limitará este análisis. En la bibliografía recomendada puede hallar referencias para un tratamiento más amplio del problema de interpolación.

Como cada uno de los n polinomios de tercer grado que conforman el spline posee cuatro coeficientes, el spline posee $4n$ coeficientes los cuales permiten satisfacer, como se verá a continuación, las siguientes condiciones:

- Condiciones de interpolación: $s(x_i) = y_i \quad i = 0, 1, 2, \dots, n$
- Condiciones de continuidad: $s(x)$ es continua en $x_i \quad i = 1, 2, \dots, n-1$
- Condiciones de suavidad: $s'(x)$ es continua en $x_i \quad i = 1, 2, \dots, n-1$
 $s''(x)$ es continua en $x_i \quad i = 1, 2, \dots, n-1$

Estas condiciones suman en total $4n - 2$ lo que significa que aun se cuenta con la posibilidad de imponer otras dos condiciones al spline, lo cual se suele hacer de diversas maneras para lograr diferentes propósitos. Estas dos condiciones se impondrán más adelante.

Para encontrar las fórmulas que determinan a $s(x)$ se seguirá el procedimiento de ir imponiendo sucesivamente las condiciones de interpolación, continuidad y suavidad, aunque no en ese orden. Se utilizará la siguiente notación:

$$M_i = s''(x_i) \quad i = 0, 1, 2, \dots, n$$

$$h_i = x_{i+1} - x_i \quad i = 0, 1, 2, \dots, n-1$$

es decir, M_0, M_1, \dots, M_n , representará el valor de la segunda derivada del spline en los nodos de interpolación y h_0, h_1, \dots, h_{n-1} , las longitudes de los tramos en que están definidos los n polinomios del spline.

La deducción que sigue, se enmarca en el tramo número i donde

$$x_i \leq x \leq x_{i+1} \quad i = 0, 1, 2, \dots, n-1$$

Como en este tramo $s(x)$ es un polinomio cúbico, su segunda derivada es una función lineal que toma valores

$$s''(x_i) = M_i \quad \text{y} \quad s''(x_{i+1}) = M_{i+1} \quad (3)$$

es decir:

$$s''(x) = \frac{(x_{i+1} - x)M_i + (x - x_i)M_{i+1}}{h_i} \quad (4)$$

Compruebe el lector que esta es una función lineal y que satisface las condiciones (3). Nótese también que, como en cada tramo sucede algo similar, la función en el intervalo $[x_0, x_n]$ es una poligonal continua, como muestra la figura 2. Esto significa que ya ha quedado satisfecha la condición de continuidad de la segunda derivada del spline.

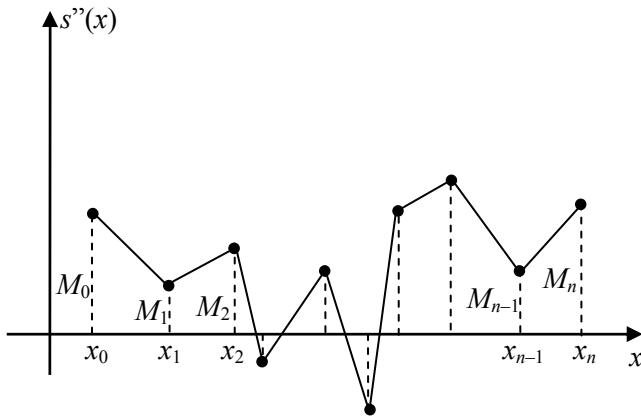


Figura 2

Si se integra indefinidamente cada miembro de la ecuación (4), puede obtenerse, salvo una constante arbitraria K_1 la función $s'(x_i)$ en el intervalo $x_i \leq x \leq x_{i+1}$

$$\begin{aligned} s'(x) &= \frac{1}{h_i} \int [(x_{i+1} - x)M_i + (x - x_i)M_{i+1}] dx \\ s'(x) &= -\frac{1}{2h_i}(x_{i+1} - x)^2 M_i + \frac{1}{2h_i}(x - x_i)^2 M_{i+1} + K_1 \end{aligned}$$

Integrando de nuevo se obtiene, siempre en el intervalo $x_i \leq x \leq x_{i+1}$:

$$\begin{aligned} s(x) &= -\frac{1}{2h_i} \int (x_{i+1} - x)^2 M_i dx + \frac{1}{2h_i} \int (x - x_i)^2 M_{i+1} dx + \int K_1 dx \\ s(x) &= \frac{1}{6h_i}(x_{i+1} - x)^3 M_i + \frac{1}{6h_i}(x - x_i)^3 M_{i+1} + K_1 x + K_2 \end{aligned} \quad (5)$$

Si en la ecuación (5) se imponen las condiciones

$$s(x_i) = y_i \quad \text{y} \quad s(x_{i+1}) = y_{i+1}$$

serán satisfechas las condiciones de interpolación y de continuidad de $s(x)$; de esta forma se hallan las constantes K_1 y K_2 . En efecto:

$$s(x_i) = \frac{1}{6h_i}(x_{i+1} - x_i)^3 M_i + \frac{1}{6h_i}(x_i - x_{i+1})^3 M_{i+1} + K_1 x_i + K_2$$

de donde: $y_i = \frac{1}{6}h_i^2 M_i + K_1 x_i + K_2$ (6)

$$s(x_{i+1}) = \frac{1}{6h_i}(x_{i+1} - x_{i+1})^3 M_i + \frac{1}{6h_i}(x_{i+1} - x_i)^3 M_{i+1} + K_1 x_{i+1} + K_2$$

es decir: $y_{i+1} = \frac{1}{6}h_i^2 M_{i+1} + K_1 x_{i+1} + K_2$ (7)

Las ecuaciones (6) y (7) forman un sistema lineal en K_1 y K_2 que se resuelve fácilmente. Se obtiene:

$$K_1 = \frac{y_{i+1} - y_i}{h_i} - \frac{h_i(M_{i+1} - M_i)}{6}$$

$$K_2 = \frac{y_i x_{i+1} - y_{i+1} x_i}{h_i} - \frac{h_i(M_{i+1} x_{i+1} - M_i x_i)}{6}$$

Sustituyendo estos valores de K_1 y K_2 en la expresión (5) y agrupando términos, se llega a:

$$s(x) = \frac{(x_{i+1} - x)^3 M_i + (x - x_i)^3 M_{i+1}}{6h_i} + \frac{(x_{i+1} - x)y_i + (x - x_i)y_{i+1}}{h_i} - \frac{h_i[(x_{i+1} - x)M_i + (x - x_i)M_{i+1}]}{6}$$

para $x_i \leq x \leq x_{i+1}$ (8)

A los efectos de la evaluación de $s(x)$ es preferible, para disminuir el número de operaciones, definir las variables u y v como:

$$u = x - x_i \quad y \quad v = x_{i+1} - x$$

y entonces, en forma más compacta:

$$s(x) = \frac{v^3 M_i + u^3 M_{i+1}}{6h_i} + \frac{vy_i + uy_{i+1}}{h_i} - \frac{h_i(vM_i + uM_{i+1})}{6} \quad (9)$$

Para satisfacer las condiciones de continuidad de $s'(x)$ se requiere que, en cada uno de los nodos interiores x_i ($i = 1, 2, \dots, n-1$) la derivada lateral por la izquierda y la derivada lateral por la derecha coincidan. Como la fórmula (8) da la ecuación de $s(x)$ en el intervalo $x_i \leq x \leq x_{i+1}$, derivando esta función y evaluando en x_i se obtiene la derivada lateral en x_i por la derecha, esto es:

$$s'(x) = \frac{-(x_{i+1} - x)^2 M_i + (x - x_i)^2 M_{i+1}}{2h_i} + \frac{y_{i+1} - y_i}{h_i} - \frac{h_i(M_{i+1} - M_i)}{6} \quad \text{para } x_i \leq x \leq x_{i+1} \quad (10)$$

$$s'(x_i^+) = \frac{-(x_{i+1} - x_i)^2 M_i + \frac{y_{i+1} - y_i}{h_i} - \frac{h_i(M_{i+1} - M_i)}{6}}{2h_i}$$

$$\text{y, como } h_i = x_{i+1} - x_i: \quad s'(x_i^+) = \frac{-h_i M_i}{2} + \frac{y_{i+1} - y_i}{h_i} - \frac{h_i(M_{i+1} - M_i)}{6} \quad (11)$$

Si en la ecuación (10) se cambia i por $i - 1$ se obtiene la expresión de $s'(x)$ en el intervalo $x_{i-1} \leq x \leq x_i$. Así que:

$$s'(x) = \frac{-(x_i - x)^2 M_{i-1} + (x - x_{i-1})^2 M_i + \frac{y_i - y_{i-1}}{h_{i-1}} - \frac{h_{i-1}(M_i - M_{i-1})}{6}}{2h_{i-1}} \quad \text{para } x_{i-1} \leq x \leq x_i$$

Al evaluar esta función en x_i se obtiene la derivada lateral en x_i por la izquierda, esto es:

$$s'(x_i^-) = \frac{(x_i - x_{i-1})^2 M_i + \frac{y_i - y_{i-1}}{h_{i-1}} - \frac{h_{i-1}(M_i - M_{i-1})}{6}}{2h_{i-1}}$$

$$\text{y, como } h_{i-1} = x_i - x_{i-1}: \quad s'(x_i^-) = \frac{h_{i-1} M_i}{2} + \frac{y_i - y_{i-1}}{h_{i-1}} - \frac{h_{i-1}(M_i - M_{i-1})}{6} \quad (12)$$

Igualando (11) y (12) se obtiene una importante relación entre M_{i-1} , M_i y M_{i+1} :

$$\frac{-h_i M_i}{2} + \frac{y_{i+1} - y_i}{h_i} - \frac{h_i(M_{i+1} - M_i)}{6} = \frac{h_{i-1} M_i}{2} + \frac{y_i - y_{i-1}}{h_{i-1}} - \frac{h_{i-1}(M_i - M_{i-1})}{6}$$

Agrupando términos y trasponiendo:

$$\frac{h_{i-1}}{6} M_{i-1} + \frac{h_{i-1} + h_i}{3} M_i + \frac{h_i}{6} M_{i+1} = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}} \quad \text{para } i = 1, 2, \dots, n-1 \quad (13)$$

La expresión (13) constituye un sistema lineal tridiagonal con la diagonal predominante que contiene $n - 1$ ecuaciones y $n + 1$ incógnitas: M_0, M_1, \dots, M_n . Agregando las dos condiciones que aun faltan, se convierte en un sistema lineal de $n + 1$ ecuaciones que se puede resolver muy eficientemente. Una vez resuelto este sistema, las ecuaciones (8), o mejor en su forma (9), permiten evaluar el spline $s(x)$ para cualquier x del intervalo $x_i \leq x \leq x_{i+1}$ ($i = 0, 1, \dots, n-1$).

El spline natural

Este es el spline interpolador más empleado. Se obtiene añadiendo las dos condiciones:

$$s''(x_0) = M_0 = 0 \quad \text{y} \quad s''(x_n) = M_n = 0 \quad (14)$$

Cuando se toman estas condiciones adicionales, puede demostrarse que el spline que resulta es la función que, pasando por los $n + 1$ puntos $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, hace mínima la integral:

$$\int_{x_0}^{x_n} [g''(x)]^2 dx$$

lo cual, como $g''(x)$ está relacionada con la curvatura de la gráfica de $g(x)$, significa, geométricamente, que minimiza la curvatura global de la función interpoladora. Desde un punto de vista físico, como la energía potencial de una varilla delgada, flexible y elástica, depende de la curvatura en cada punto, resulta que, si una varilla con tales propiedades, es obligada a pasar por los $n + 1$ puntos del plano: $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, ella toma la forma que minimiza su energía potencial elástica, que es, precisamente, la del spline cúbico natural que interpola a dichos puntos. De este hecho físico proviene la palabra “spline” que originalmente se usaba para designar una especie de curvígrafo empleado en dibujo, formado por una varilla que se hacía pasar por varios puntos de un plano.

Teniendo en cuenta las condiciones adicionales (14), el sistema de ecuaciones (13) toma para el spline natural, la forma:

$$\begin{aligned} \frac{h_0 + h_1}{3} M_1 + \frac{h_1}{6} M_2 &= \frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\ \frac{h_1}{6} M_1 + \frac{h_1 + h_2}{3} M_2 + \frac{h_2}{6} M_3 &= \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\ \frac{h_2}{6} M_2 + \frac{h_2 + h_3}{3} M_3 + \frac{h_3}{6} M_4 &= \frac{y_4 - y_3}{h_3} - \frac{y_3 - y_2}{h_2} \\ &\vdots \\ \frac{h_{n-2}}{6} M_{n-2} + \frac{h_{n-2} + h_{n-1}}{3} M_{n-1} &= \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \end{aligned}$$

Que, escrito en forma matricial resulta:

$$\mathbf{HM} = \mathbf{Y} \quad (15)$$

donde:

$$\mathbf{H} = \begin{bmatrix} \frac{h_0 + h_1}{3} & \frac{h_1}{6} & & & & \\ \frac{h_1}{6} & \frac{h_1 + h_2}{3} & \frac{h_2}{6} & & & \\ & \frac{h_2}{6} & \frac{h_2 + h_3}{3} & \ddots & & \\ & & \ddots & \ddots & \frac{h_{n-2}}{6} & \\ & & & & \frac{h_{n-2}}{6} & \frac{h_{n-2} + h_{n-1}}{3} \end{bmatrix} \quad \mathbf{M} = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ \vdots \\ M_{n-1} \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} \frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\ \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\ \frac{y_4 - y_3}{h_3} - \frac{y_3 - y_2}{h_2} \\ \vdots \\ \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \end{bmatrix}$$

Aquí se ha seguido el criterio, usual en las matrices tridiagonales, de no escribir los elementos nulos. El sistema que se obtiene es tridiagonal y con diagonal predominante, por tanto puede ser resuelto con gran eficiencia por el método de Gauss especializado en sistemas tridiagonales,

estudiado en la sección 3.3. También puede ser utilizado uno de los métodos iterativos estudiados en el capítulo 3 pues la convergencia es rápida; observe que el factor de convergencia del método de Jacobi es $\alpha = 0,5$ y el factor β , que determina la rapidez de convergencia del método de Seidel, es aun menor.

Algoritmo en seudo código para el spline cúbico natural

El siguiente algoritmo determina el valor interpolado $s(x)$ de un spline cúbico natural, para lo cual previamente se calculan los parámetros M_1, M_2, \dots, M_{n-1} . Con una ligera modificación, el algoritmo puede ser utilizado para calcular un conjunto suficientemente grande de puntos del spline como para poder trazar su gráfica. Se supone conocidos los nodos de interpolación $\{x_0, x_1, x_2, \dots, x_n\}$, ordenados en forma creciente, los valores correspondientes de la función a interpolar $\{y_0, y_1, y_2, \dots, y_n\}$ y el número x donde se desea interpolar, el cual debe estar comprendido en el intervalo $[x_0, x_n]$. El algoritmo da como salida el número $s(x)$.

```

for  $i = 0$  to  $n - 1$ 
     $h_i := x_{i+1} - x_i$ 
end
Formar las tres diagonales de la matriz H
Formar la matriz columna Y
Resolver el sistema tridiagonal HM = Y utilizando los métodos de Gauss o de Seidel
 $i := n$ 
do while  $x_i > x$  {En este lazo se determina el tramo del spline a que pertenece  $x$ }
     $i := i - 1$ 
end
 $u := x - x_i$ 
 $v := x_{i+1} - x$ 

$$s(x) := \frac{v^3 M_i + u^3 M_{i+1}}{6h_i} + \frac{vy_i + uy_{i+1}}{h_i} - \frac{h_i(vM_i + uM_{i+1})}{6}$$

Terminar

```

Ejemplo 1

(El enunciado de este ejemplo se analizó en la sección 4.1. Aquí se repite el enunciado y la figura correspondiente para comodidad del lector). Para producir en un torno de mando numérico una pieza con el perfil longitudinal que se muestra en la figura 3, es necesario obtener una función simple (de modo que pueda ser evaluada en un tiempo muy breve) que describa el contorno de la pieza. Esta función servirá para fijar la posición de la cuchilla del torno en cada instante. Algunas dimensiones, marcadas en la figura en milímetros, deberán ser respetadas y el perfil de la pieza debe ser una curva suave.

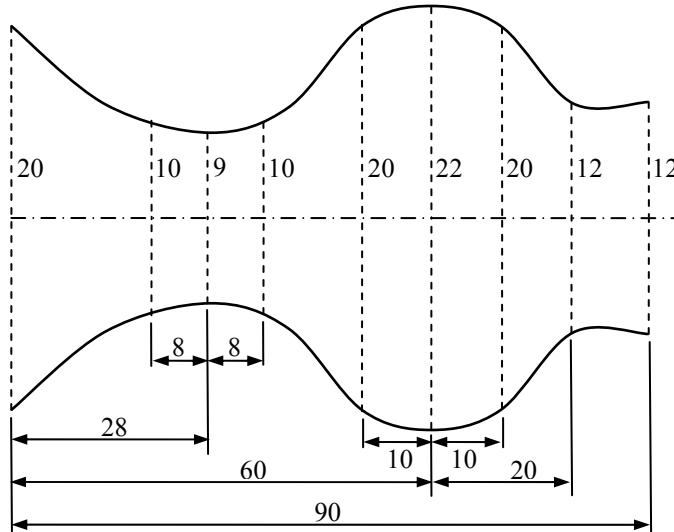


Figura 3

Solución:

Como se desea hallar una función simple que describa el contorno, es claro que no se debe intentar hallar un polinomio que satisfaga las 9 condiciones de interpolación, ya que sería un polinomio de grado 8 y esto implicaría complicaciones en su evaluación y, lo que es mucho peor, el peligro de que ese polinomio presente oscilaciones indeseadas. Para hallar un spline cúbico natural, se determinaron los siguientes nodos (a partir de la figura 3):

$$\begin{array}{ll}
 x_0 = 0 & y_0 = 20 \\
 x_1 = 20 & y_1 = 10 \\
 x_2 = 28 & y_2 = 9 \\
 x_3 = 36 & y_3 = 10 \\
 x_4 = 50 & y_4 = 20 \\
 x_5 = 60 & y_5 = 22 \\
 x_6 = 70 & y_6 = 20 \\
 x_7 = 80 & y_7 = 12 \\
 x_8 = 90 & y_8 = 12
 \end{array}$$

Con un algoritmo similar al del seudo código que se acaba de ver, se hallaron los coeficientes M_i ($i = 1, 2, 3, \dots, 7$). Recuérdese que, por ser un spline natural, M_0 y M_8 son nulos. Los resultados fueron:

$$\begin{aligned}
 M_0 &= 0 \\
 M_1 &= 0,03875 \\
 M_2 &= 0,00999 \\
 M_3 &= 0,10879 \\
 M_4 &= -0,09507 \\
 M_5 &= -0,00453 \\
 M_6 &= 0,12679 \\
 M_7 &= 0,15170 \\
 M_8 &= 0
 \end{aligned}$$

Con estos resultados basta para evaluar el spline en cualquier punto del intervalo $0 \leq x \leq 90$ o para trazarlo, si se desea. De hecho, las curvas que se muestran en la figura 3, son los splines cúbicos que corresponden a los resultados anteriores.

El spline cúbico anclado

En el spline interpolador anclado (“clamped spline”) las dos condiciones adicionales que se toman son:

$$s'(x_0^+) = m_0 \quad \text{y} \quad s'(x_n^-) = m_n$$

es decir, si fijan las pendientes en los extremos del spline a valores deseados m_0 y m_n , lo cual es importante en muchas aplicaciones. Las expresiones para las derivadas laterales fueron antes halladas en las ecuaciones (11) y (12):

$$s'(x_i^+) = \frac{-h_i M_i}{2} + \frac{y_{i+1} - y_i}{h_i} - \frac{h_i(M_{i+1} - M_i)}{6} \quad \text{y} \quad s'(x_i^-) = \frac{h_{i-1} M_i}{2} + \frac{y_i - y_{i-1}}{h_{i-1}} - \frac{h_{i-1}(M_i - M_{i-1})}{6}$$

De aquí resulta:

$$s'(x_0^+) = \frac{-h_0 M_0}{2} + \frac{y_1 - y_0}{h_0} - \frac{h_0(M_1 - M_0)}{6} = m_0 \quad (16)$$

$$s'(x_n^-) = \frac{h_{n-1} M_n}{2} + \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{h_{n-1}(M_n - M_{n-1})}{6} = m_n \quad (17)$$

Después de simplificar y ordenar, se obtiene:

$$\frac{h_0}{3} M_0 + \frac{h_0}{6} M_1 = \frac{y_1 - y_0}{h_0} - m_0 \quad \text{y} \quad \frac{h_{n-1}}{6} M_{n-1} + \frac{h_{n-1}}{3} M_n = m_n - \frac{y_n - y_{n-1}}{h_{n-1}}$$

Cuando estas dos ecuaciones se unen a las $n - 1$ ecuaciones (13):

$$\frac{h_{i-1}}{6} M_{i-1} + \frac{h_{i-1} + h_i}{3} M_i + \frac{h_i}{6} M_{i+1} = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}} \quad \text{para } i = 1, 2, \dots, n-1$$

se forma el sistema

$$\begin{aligned}
\frac{h_0}{3}M_0 + \frac{h_0}{6}M_1 &= \frac{y_1 - y_0}{h_0} - m_0 \\
\frac{h_0}{6}M_0 + \frac{h_0 + h_1}{3}M_1 + \frac{h_1}{6}M_2 &= \frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\
\frac{h_1}{6}M_1 + \frac{h_1 + h_2}{3}M_2 + \frac{h_2}{6}M_3 &= \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\
\frac{h_2}{6}M_2 + \frac{h_2 + h_3}{3}M_3 + \frac{h_3}{6}M_4 &= \frac{y_4 - y_3}{h_3} - \frac{y_3 - y_2}{h_2} \\
&\vdots \\
\frac{h_{n-2}}{6}M_{n-2} + \frac{h_{n-2} + h_{n-1}}{3}M_{n-1} + \frac{h_{n-1}}{6}M_n &= \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \\
\frac{h_{n-1}}{6}M_{n-1} + \frac{h_{n-1}}{3}M_n &= m_n - \frac{y_n - y_{n-1}}{h_{n-1}}
\end{aligned}$$

el cual, si se escribe en forma matricial, queda:

$$\mathbf{HM} = \mathbf{Y}$$

donde ahora:

$$\begin{aligned}
\mathbf{M} &= \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ \vdots \\ M_n \end{bmatrix} \\
\mathbf{H} &= \begin{bmatrix} \frac{h_0}{3} & \frac{h_0}{6} & & & & \\ \frac{h_0}{6} & \frac{h_0 + h_1}{3} & \frac{h_1}{6} & & & \\ & \frac{h_1}{6} & \frac{h_1 + h_2}{3} & \frac{h_2}{6} & & \\ & & \ddots & \ddots & & \\ & & & & \frac{h_{n-2}}{6} & \frac{h_{n-2} + h_{n-1}}{3} & \frac{h_{n-1}}{6} \\ & & & & & \frac{h_{n-1}}{6} & \frac{h_{n-1}}{3} \end{bmatrix} \\
\mathbf{Y} &= \begin{bmatrix} \frac{y_1 - y_0}{h_0} - m_0 \\ \frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\ \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\ \vdots \\ \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \\ m_n - \frac{y_n - y_{n-1}}{h_{n-1}} \end{bmatrix}
\end{aligned}$$

de manera que ahora el sistema es de orden $n + 1$, pero sigue siendo tridiagonal y con la diagonal principal fuertemente predominante ($\alpha = 0,5$).

Algoritmo en seudo código para el spline cúbico anclado

El siguiente algoritmo determina el valor interpolado $s(x)$ de un spline cúbico anclado, para lo cual previamente se calculan los parámetros $M_0, M_1, M_2, \dots, M_n$. Con una ligera modificación, el algoritmo puede ser utilizado para calcular un conjunto suficientemente grande de puntos del spline como para poder trazar su gráfica. Se supone conocidos los nodos de interpolación $\{x_0, x_1,$

$x_2, \dots, x_n\}$, ordenados en forma creciente, los valores correspondientes de la función a interpolar $\{y_0, y_1, y_2, \dots, y_n\}$ la pendiente deseada a la derecha del nodo x_0 (m_0) y a la izquierda del nodo x_n (m_n) y el número x donde se desea interpolar, el cual debe estar comprendido en el intervalo $[x_0, x_n]$. El algoritmo da como salida el número $s(x)$. El algoritmo que sigue es idéntico al anterior, pues solo difiere en las matrices \mathbf{H} , \mathbf{M} e \mathbf{Y} que se utilizan.

```

for  $i = 0$  to  $n - 1$ 
     $h_i := x_{i+1} - x_i$ 
end
Formar las tres diagonales de la matriz H
Formar la matriz columna Y
Resolver el sistema tridiagonal HM = Y utilizando los métodos de Gauss o de Seidel
 $i := n$ 
do while  $x_i > x$  {En este lazo se determina el tramo del spline a que pertenece  $x$ }
     $i := i - 1$ 
end
 $u := x - x_i$ 
 $v := x_{i+1} - x$ 

$$s(x) := \frac{v^3 M_i + u^3 M_{i+1}}{6h_i} + \frac{vy_i + uy_{i+1}}{h_i} - \frac{h_i(vM_i + uM_{i+1})}{6}$$

Terminar

```

El spline cúbico periódico

Si al spline interpolador se le imponen las condiciones:

$$y_0 = y_n \quad s'(x_0^+) = s'(x_n^-) \quad \text{y} \quad s''(x_0^+) = s''(x_n^-)$$

se logra una función cuya gráfica es tal que, si ella se repite una y otra vez, en los puntos de unión la función es continua y son continuas las derivadas de orden 1 y 2. La figura 4 muestra esta idea. El spline original (en línea más gruesa) se ha repetido a la izquierda y a la derecha. Como las ordenadas, las pendientes y las concavidades en ambos extremos coinciden, resulta una función periódica, continua y suave hasta la segunda derivada.

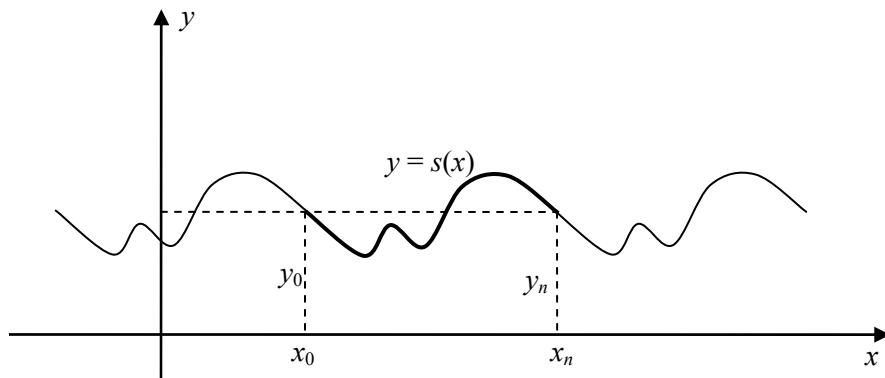


Figura 4

A este spline se le llama periódico aunque, en rigor, solo representa un período de la función periódica que se forma por la repetición de $s(x)$. Como los números $\{y_0, y_1, y_2, \dots, y_n\}$ son arbitrarios, tomar $y_0 = y_n$ no implica una condición adicional que haya que satisfacer, así que, de nuevo, se trata de agregar dos condiciones adicionales a las $n - 1$ que establecen las ecuaciones (13). El sistema (13) sin condición adicional alguna sería:

$$\begin{aligned} \frac{h_0}{6}M_0 + \frac{h_0 + h_1}{3}M_1 + \frac{h_1}{6}M_2 &= \frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\ \frac{h_1}{6}M_1 + \frac{h_1 + h_2}{3}M_2 + \frac{h_2}{6}M_3 &= \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\ \frac{h_2}{6}M_2 + \frac{h_2 + h_3}{3}M_3 + \frac{h_3}{6}M_4 &= \frac{y_4 - y_3}{h_3} - \frac{y_3 - y_2}{h_2} \\ &\vdots \\ \frac{h_{n-2}}{6}M_{n-2} + \frac{h_{n-2} + h_{n-1}}{3}M_{n-1} + \frac{h_{n-1}}{6}M_n &= \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \end{aligned} \quad (18)$$

Al hacer $y_0 = y_n$ y $s''(x_0^+) = s''(x_n^-)$, lo cual significa $M_0 = M_n$, la última ecuación se transforma en:

$$\frac{h_{n-2}}{6}M_{n-2} + \frac{h_{n-2} + h_{n-1}}{3}M_{n-1} + \frac{h_{n-1}}{6}M_0 = \frac{y_0 - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \quad (19)$$

Por otra parte, al igualar las derivadas en x_0 y x_n , obtenidas en las ecuaciones (16) y (17), resulta

$$s'(x_0^+) = \frac{-h_0 M_0}{2} + \frac{y_1 - y_0}{h_0} - \frac{h_0(M_1 - M_0)}{6} = \frac{h_{n-1} M_n}{2} + \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{h_{n-1}(M_n - M_{n-1})}{6} = s'(x_n^-)$$

Si en esta ecuación se sustituye y_n por y_0 y M_n por M_0 queda:

$$\frac{-h_0 M_0}{2} + \frac{y_1 - y_0}{h_0} - \frac{h_0(M_1 - M_0)}{6} = \frac{h_{n-1} M_0}{2} + \frac{y_0 - y_{n-1}}{h_{n-1}} - \frac{h_{n-1}(M_0 - M_{n-1})}{6}$$

Agrupando y ordenando convenientemente:

$$\frac{h_0 + h_{n-1}}{3}M_0 + \frac{h_0}{6}M_1 + \frac{h_{n-1}}{6}M_{n-1} = \frac{y_1 - y_0}{h_0} - \frac{y_0 - y_{n-1}}{h_{n-1}} \quad (20)$$

Al cambiar la última ecuación del sistema (18) por la ecuación (19) y agregarle la condición (20) se obtiene el siguiente sistema lineal de n ecuaciones con las n incógnitas $\{M_0, M_1, M_2, \dots, M_{n-1}\}$:

$$\begin{aligned}
\frac{h_0 + h_{n-1}}{3} M_0 + \frac{h_0}{6} M_1 + \frac{h_{n-1}}{6} M_{n-1} &= \frac{y_1 - y_0}{h_0} - \frac{y_0 - y_{n-1}}{h_{n-1}} \\
\frac{h_0}{6} M_0 + \frac{h_0 + h_1}{3} M_1 + \frac{h_1}{6} M_2 &= \frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\
\frac{h_1}{6} M_1 + \frac{h_1 + h_2}{3} M_2 + \frac{h_2}{6} M_3 &= \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\
&\vdots \\
\frac{h_{n-1}}{6} M_0 + \frac{h_{n-2}}{6} M_{n-2} + \frac{h_{n-2} + h_{n-1}}{3} M_{n-1} &= \frac{y_0 - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}}
\end{aligned}$$

La estructura de este sistema lineal se aprecia con más claridad si se escribe en la forma matricial:

$$\mathbf{HM} = \mathbf{Y}$$

donde ahora:

$$\mathbf{H} = \left[\begin{array}{cccccc}
\frac{h_{n-1} + h_0}{3} & \frac{h_0}{6} & 0 & 0 & \dots & \frac{h_{n-1}}{6} \\
\frac{h_0}{6} & \frac{h_0 + h_1}{3} & \frac{h_1}{6} & 0 & \dots & 0 \\
0 & \frac{h_1}{6} & \frac{h_1 + h_2}{3} & \frac{h_2}{6} & \dots & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & \dots & \frac{h_{n-3}}{6} & \frac{h_{n-3} + h_{n-2}}{3} & \frac{h_{n-2}}{6} \\
\frac{h_{n-1}}{6} & 0 & \dots & 0 & \frac{h_{n-2}}{6} & \frac{h_{n-2} + h_{n-1}}{3}
\end{array} \right]$$

$$\mathbf{Y} = \left[\begin{array}{c}
\frac{y_1 - y_0}{h_0} - \frac{y_0 - y_{n-1}}{h_{n-1}} \\
\frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\
\frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\
\vdots \\
\frac{y_{n-1} - y_{n-2}}{h_{n-2}} - \frac{y_{n-2} - y_{n-3}}{h_{n-3}} \\
\frac{y_0 - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}}
\end{array} \right] \quad \text{y} \quad \mathbf{M} = \left[\begin{array}{c}
M_0 \\
M_1 \\
M_2 \\
\vdots \\
M_{n-1}
\end{array} \right]$$

Como se observa, ahora \mathbf{H} no es tridiagonal, pero sigue siendo de diagonal predominante con factor de convergencia $\alpha = 0,5$, por lo cual los métodos iterativos permiten su solución con eficiencia.

Algoritmo en seudo código para el spline cúbico periódico

El siguiente algoritmo determina el valor interpolado $s(x)$ de un spline cúbico periódico, para lo cual previamente se calculan los parámetros $M_0, M_1, M_2, \dots, M_{n-1}$ (no se calcula M_n , ya que, $M_n = M_0$). Con una ligera modificación, el algoritmo puede ser utilizado para calcular un conjunto suficientemente grande de puntos del spline como para poder trazar su gráfica. Se supone conocidos los nodos de interpolación $\{x_0, x_1, x_2, \dots, x_n\}$, ordenados en forma creciente, los valores correspondientes de la función a interpolar $\{y_0, y_1, y_2, \dots, y_{n-1}\}$ (se supone que $y_n = y_0$) y el número x donde se desea interpolar, el cual debe estar comprendido en el intervalo $[x_0, x_n]$. El algoritmo da como salida el número $s(x)$.

```

for  $i = 0$  to  $n - 1$ 
     $h_i := x_{i+1} - x_i$ 
end
Formar la matriz H
Formar la matriz columna Y
Resolver el sistema tridiagonal HM = Y utilizando el método de Seidel
 $M_n := M_0$ 
 $i := n$ 
do while  $x_i > x$  {En este lazo se determina el tramo del spline a que pertenece  $x$ }
     $i := i - 1$ 
end
 $u := x - x_i$ 
 $v := x_{i+1} - x$ 

$$s(x) := \frac{v^3 M_i + u^3 M_{i+1}}{6h_i} + \frac{vy_i + uy_{i+1}}{h_i} - \frac{h_i(vM_i + uM_{i+1})}{6}$$

Terminar

```

Ejemplo 2

Halle las ecuaciones paramétricas de una curva cerrada, continua, con tangente en cada punto y curvatura continua que pase por los cuatro puntos: $P_0(1, 1), P_1(3, 1), P_2(2, 2), P_3(4, 2), P_0(1, 1)$ en ese mismo orden.

Solución:

Para que la curva paramétrica $x = x(t); y = y(t)$ pase por los puntos $P_0(x_0, y_0), P_1(x_1, y_1), P_2(x_2, y_2), P_3(x_3, y_3), P_0(x_0, y_0)$ basta con tomar cinco números $t_0 < t_1 < t_2 < t_3 < t_4$ y hacer que las funciones $x(t); y(t)$ satisfagan las condiciones de interpolación:

$$\begin{array}{ll}
 x(t_0) = x_0 & y(t_0) = y_0 \\
 x(t_1) = x_1 & y(t_1) = y_1 \\
 x(t_2) = x_2 & y(t_2) = y_2 \\
 x(t_3) = x_3 & y(t_3) = y_3 \\
 x(t_4) = x_0 & y(t_4) = y_0
 \end{array}$$

Las condiciones de continuidad y suavidad se obtienen si $x(t)$ e $y(t)$ son funciones continuas y con primera y segunda derivadas continuas. Como se ha tomado $x(t_0) = x(t_4)$ e $y(t_0) = y(t_4)$ la curva obtenida será cerrada. Si se utilizan para $x(t)$ e $y(t)$ splines cúbicos periódicos, se logrará la

continuidad y suavidad de la curva, inclusive en el punto P_0 donde la curva se cierra, ya que coinciden los valores iniciales y finales de $x'(t), x''(t), y'(t), y''(t)$.

Existen diferentes criterios para seleccionar los números t_0, t_1, t_2, t_3, t_4 , que determinan diferentes características geométricas de la curva obtenida. Aquí se utilizará la variante más simple (pero no siempre la mejor) de tomar valores equidistantes. Sean entonces: $t_0 = 1, t_1 = 2, t_2 = 3, t_3 = 4, t_4 = 5$. Con esto queda:

Nodos de $x(t)$: (1, 1), (2, 3), (3, 2), (4, 4), (5, 1)

Nodos de $y(t)$: (1, 1), (2, 1), (3, 2), (4, 2), (5, 1)

Al tomar para t_0, t_1, t_2, t_3, t_4 , valores equidistantes las ecuaciones se simplificarán ya que, en cada spline, será: $h_0 = h_1 = h_2 = h_3 = h_4 = 1$.

Spline $x(t)$:

En este caso es: $x_0 = 1, x_1 = 3, x_2 = 2, x_3 = 4$ y el sistema $\mathbf{HM} = \mathbf{X}$ resulta:

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & \frac{1}{6} \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \frac{1}{6} & 0 & \frac{1}{6} & \frac{2}{3} \end{bmatrix} \cdot \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \end{bmatrix} = \begin{bmatrix} x_1 - 2x_0 + x_3 \\ x_2 - 2x_1 + x_0 \\ x_3 - 2x_2 + x_1 \\ x_0 - 2x_3 + x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ -3 \\ 3 \\ -5 \end{bmatrix}$$

Como se trata de un sistema muy pequeño, cualquier método de solución puede servir. Se obtiene:

$$M_0 = 5, M_1 = -3, M_2 = 3, M_3 = 5$$

y como $M_4 = M_0$, se tiene, $M_4 = 5$.

A partir de la ecuación (8), el tramo del spline $x(t)$ correspondiente a $t_i < t < t_{i+1}$ será:

$$x(t) = \frac{(t_{i+1} - t)^3 M_i + (t - t_i)^3 M_{i+1}}{6h_i} + \frac{(t_{i+1} - t)x_i + (t - t_i)x_{i+1}}{h_i} - \frac{h_i[(t_{i+1} - t)M_i + (t - t_i)M_{i+1}]}{6}$$

para $i = 0, 1, 2, 3$. Al darle estos valores a i y tomando para las t , sus valores correspondientes, resulta:

$$x(t) = \begin{cases} \frac{(2-t)^3 M_0 + (t-1)^3 M_1}{6} + (2-t)x_0 + (t-1)x_1 - \frac{(2-t)M_0 + (t-1)M_1}{6} & \text{para } 1 \leq t \leq 2 \\ \frac{(3-t)^3 M_1 + (t-2)^3 M_2}{6} + (3-t)x_1 + (t-2)x_2 - \frac{(3-t)M_1 + (t-2)M_2}{6} & \text{para } 2 \leq t \leq 3 \\ \frac{(4-t)^3 M_2 + (t-3)^3 M_3}{6} + (4-t)x_2 + (t-3)x_3 - \frac{(4-t)M_2 + (t-3)M_3}{6} & \text{para } 3 \leq t \leq 4 \\ \frac{(5-t)^3 M_3 + (t-4)^3 M_4}{6} + (5-t)x_3 + (t-4)x_4 - \frac{(5-t)M_3 + (t-4)M_4}{6} & \text{para } 4 \leq t \leq 5 \end{cases}$$

Después de asignarle a las M y las x sus correspondientes valores y simplificando, se obtiene:

$$x(t) = \begin{cases} -\frac{1}{4}(16t^3 - 75t^2 + 105t - 50) & \text{para } 1 \leq t \leq 2 \\ \frac{1}{4}(14t^3 - 105t^2 + 255t - 190) & \text{para } 2 \leq t \leq 3 \\ -\frac{1}{4}(16t^3 - 165t^2 + 555t - 620) & \text{para } 3 \leq t \leq 4 \\ \frac{1}{4}(18t^3 - 243t^2 + 1077t - 1556) & \text{para } 4 \leq t \leq 5 \end{cases}$$

En la figura 5 se muestra la gráfica del spline periódico $x(t)$ y sus nodos.

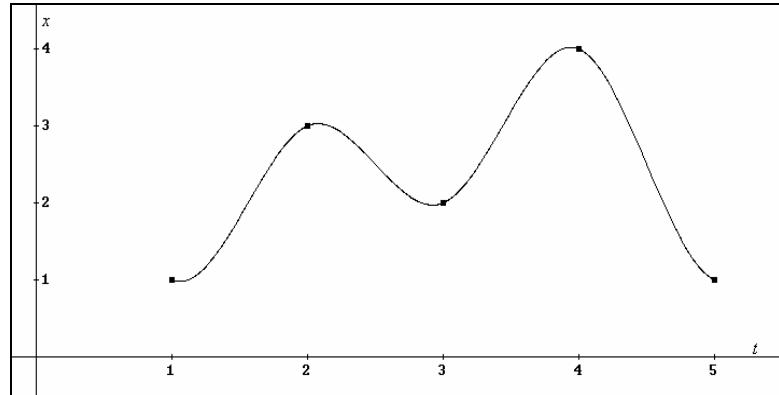


Figura 5

Spline $y(t)$:

En este caso es: $y_0 = 1, y_1 = 1, y_2 = 2, y_3 = 2$ y el sistema $\mathbf{HM} = \mathbf{Y}$ resulta:

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & \frac{1}{6} \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \frac{1}{6} & 0 & \frac{1}{6} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \end{bmatrix} = \begin{bmatrix} y_1 - 2y_0 + y_3 \\ y_2 - 2y_1 + y_0 \\ y_3 - 2y_2 + y_1 \\ y_0 - 2y_3 + y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

Al resolver el sistema se obtiene:

$$M_0 = 1,5, M_1 = 1,5, M_2 = -1,5, M_3 = -1,5$$

y como $M_4 = M_0$, se tiene, $M_4 = 1,5$.

Utilizando la ecuación (8), el tramo del spline $y(t)$ correspondiente a $t_i < t < t_{i+1}$ será:

$$y(t) = \frac{(t_{i+1} - t)^3 M_i + (t - t_i)^3 M_{i+1}}{6h_i} + \frac{(t_{i+1} - t)y_i + (t - t_i)y_{i+1}}{h_i} - \frac{h_i[(t_{i+1} - t)M_i + (t - t_i)M_{i+1}]}{6}$$

para $i = 0, 1, 2, 3$. Al darle estos valores a i y tomando para las t , sus valores correspondientes, resulta:

$$y(t) = \begin{cases} \frac{(2-t)^3 M_0 + (t-1)^3 M_1}{6} + (2-t)y_0 + (t-1)y_1 - \frac{(2-t)M_0 + (t-1)M_1}{6} & \text{para } 1 \leq t \leq 2 \\ \frac{(3-t)^3 M_1 + (t-2)^3 M_2}{6} + (3-t)y_1 + (t-2)y_2 - \frac{(3-t)M_1 + (t-2)M_2}{6} & \text{para } 2 \leq t \leq 3 \\ \frac{(4-t)^3 M_2 + (t-3)^3 M_3}{6} + (4-t)y_2 + (t-3)y_3 - \frac{(4-t)M_2 + (t-3)M_3}{6} & \text{para } 3 \leq t \leq 4 \\ \frac{(5-t)^3 M_3 + (t-4)^3 M_4}{6} + (5-t)y_3 + (t-4)y_4 - \frac{(5-t)M_3 + (t-4)M_4}{6} & \text{para } 4 \leq t \leq 5 \end{cases}$$

Asignando a las M y las y sus correspondientes valores y simplificando, se obtiene:

$$y(t) = \begin{cases} \frac{1}{4}(3t^2 - 9t + 10) & \text{para } 1 \leq t \leq 2 \\ -\frac{1}{4}(2t^3 - 15t^2 + 33t - 26) & \text{para } 2 \leq t \leq 3 \\ -\frac{1}{4}(3t^2 - 21t + 28) & \text{para } 3 \leq t \leq 4 \\ \frac{1}{4}(2t^3 - 27t^2 + 117t - 156) & \text{para } 4 \leq t \leq 5 \end{cases}$$

La figura 6 muestra la gráfica del spline periódico $y(t)$ y los nodos correspondientes. En la figura 7 se encuentra la curva con ecuaciones paramétricas $x = x(t)$; $y = y(t)$ y los cuatro puntos $P_0(1, 1)$, $P_1(3, 1)$, $P_2(2, 2)$, $P_3(4, 2)$. Observe que, en efecto, es una curva que reúne todos los requisitos exigidos.

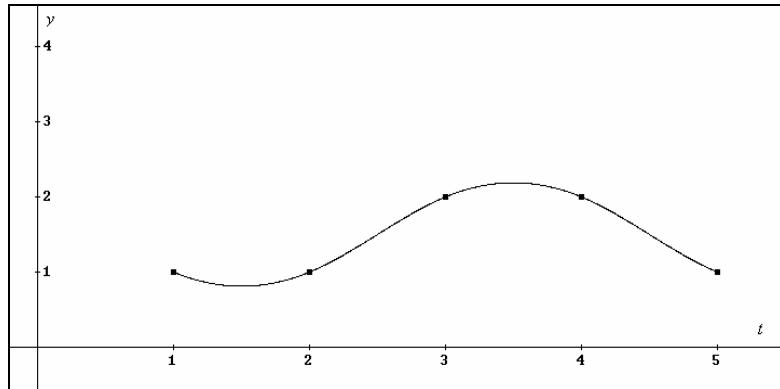


Figura 6

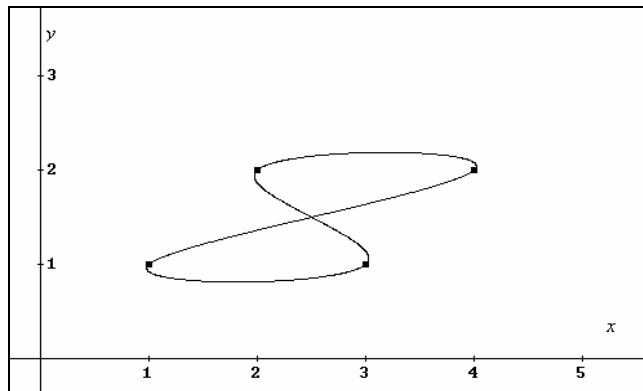


Figura 7

Ejercicios

1. Dados los siguientes puntos, halle la expresión analítica de una función formada por polinomios de interpolación de grado k determinados por $k + 1$ nodos consecutivos, para $k = 1, 2$ y 3 . Construya las gráficas de las funciones obtenidas.

i	0	1	2	3	4	5	6
x_i	3	3,5	4	4,5	5	5,5	6
y_i	2,2	4,3	4,8	4,1	3,9	5,3	7,9

2. Determine el spline cúbico natural correspondiente a los nodos del ejercicio anterior. Trace su gráfica y compárela con las de dicho ejercicio.
3. En la figura 8 se muestra un corte transversal de un canal para regadío, incluyendo las cotas (en metros) de varios puntos (equidistantes un metro) respecto a una cierta referencia. Emplee un spline cúbico natural para representar analíticamente este corte.

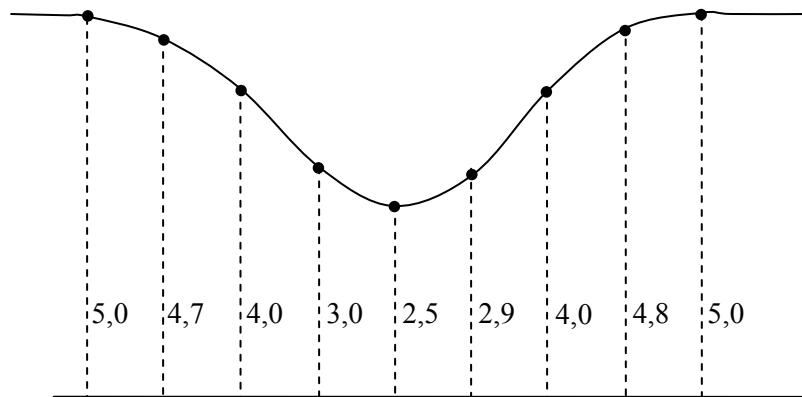


Figura 8

4. A continuación se muestran las coordenadas Lambert de varios puntos de la costa norte de la provincia de Pinar del Río. Determine un spline cúbico natural para representar el segmento de costa que corresponde a estos puntos.

i	1	2	3	4	5	6	7	8
Coordenadas del punto i	Longitud	10,0	10,2	11,0	12,2	13,5	14,0	14,5
	Latitud	7,3	6,7	6,5	7,0	6,0	6,4	7,0

5. En la figura 9 se muestra el perfil de una botella de cerveza y algunas de sus dimensiones en milímetros. Halle un spline cúbico natural que represente a la curva comprendida entre los puntos A y B .

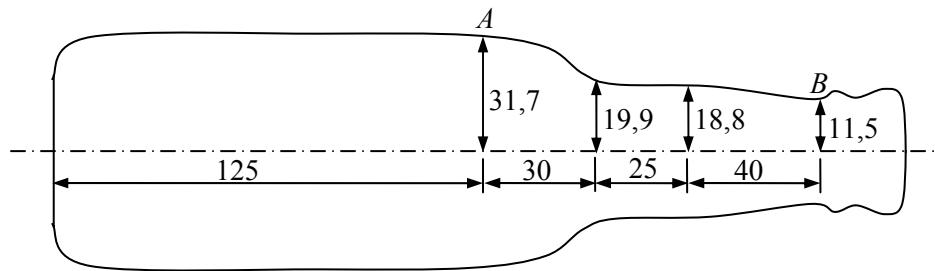


Figura 9

6. Para trazar el perfil de un tramo recto de 1000 metros de longitud de un terraplén (figura 10), se han determinado las cotas (en metros) que el mismo debe tener en puntos separados 100 m entre sí. Utilice un spline cúbico natural para obtener cotas cada 50 m.

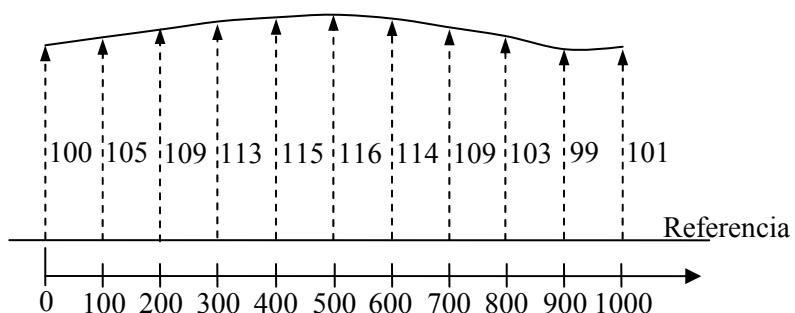


Figura 10

7. En un proceso térmico se ha logrado normar en intervalos de 10 minutos la temperatura que debe existir en el interior de un horno eléctrico. Estos datos se muestran en la tabla. Para diseñar un sistema de control computarizado se requiere conocer la norma cada cinco minutos. Utilice un spline cúbico natural para hallar dichas temperaturas.

Tiempo (min.)	0	10	20	30	40	50	60	70	80
Temperatura (°C)	30	40	60	80	95	100	110	100	80

8. En la tabla que se acompaña aparece la presión atmosférica (en hectopascales) en 35 nodos sobre una cuadrícula que cubre una región de 20 km por 30 km. Utilizando interpolación lineal halle las coordenadas de los puntos (marcados con círculos en blanco en el mapa de la figura 11) donde la presión es de 1005 hectopascales y después, mediante un spline cúbico natural, trace en el mapa una curva suave que pase por los puntos marcados (isobara de 1005 hectopascales).

$i \backslash j$	1	2	3	4	5	6	7
1	990	994	998	993	992	994	995
2	995	1000	1000	995	997	997	998
3	1000	1007	1007	999	1001	1000	1000
4	1010	1012	1013	1013	1009	1012	1003
5	1015	1016	1015	1014	1012	1014	1013

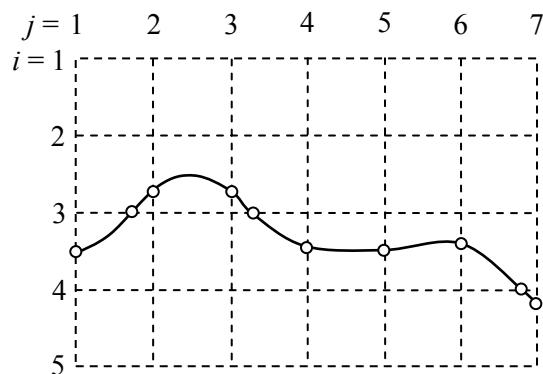


Figura 11

9. Plantee el sistema de ecuaciones necesario para determinar el spline cúbico anclado que pasa por los puntos del plano: $(2, 4)$, $(3, 3)$, $(5, 4)$, $(6, 3)$ y que satisface las condiciones $s'(2^+) = 1$ y $s'(6^-) = -1$.
10. Plantee el sistema de ecuaciones necesario para determinar un spline cúbico periódico que pase por los puntos $(2, 4)$, $(3, 3)$, $(5, 4)$, $(6, 3)$ y que permita formar una función periódica con periodo 5.
11. Plantee los sistemas de ecuaciones necesarios para determinar los dos splines cúbicos naturales que forman las ecuaciones paramétricas de una curva como la que se muestra en la figura 12, que pasa por los puntos A, B, C, D, E .

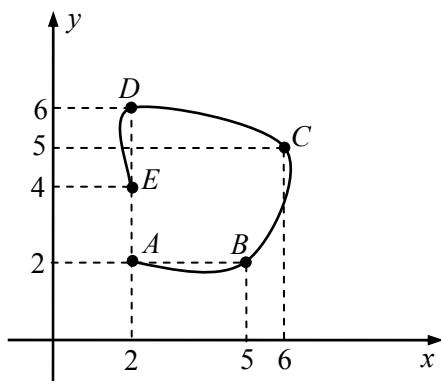


Figura 12

12. Plantee los sistemas de ecuaciones necesarios para determinar los dos splines cúbicos periódicos que forman las ecuaciones paramétricas de una curva cerrada como la que se muestra en la figura 12, que pasa por los puntos $(1, 1)$, $(7, 1)$ y $(4, 7)$, posee tangente en cada punto y curvatura que varía en forma continua.

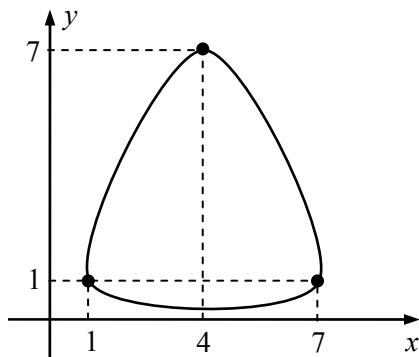


Figura 13

4.6 Ajuste de curvas

En los ejemplos introductorios al comienzo de este capítulo se vieron dos situaciones en las que el enfoque de la interpolación no resuelve el problema. En el ejemplo 2 se deseaba encontrar una relación funcional entre el peso y la talla de los individuos de un sector poblacional a partir de un conjunto de datos correspondientes a 100 de estas personas. Cuando estos datos se representan en un sistema t (estatura) contra p (peso) muestran un comportamiento como el de la figura 1 (que es una copia de la figura 1 de la sección 4.1). Es obvio que sería absurdo aplicar a este problema

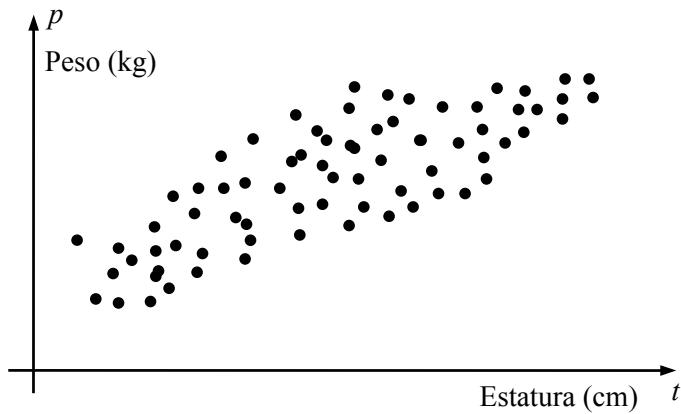


Figura 1

el enfoque de la interpolación, aun cuando fuera posible encontrar alguna función interpoladora capaz de pasar por los 100 puntos de la figura.

En el ejemplo 4 de esa misma sección se analizó otro caso en el que el uso de la interpolación tampoco es recomendable. En dicho ejemplo, se tienen los resultados de la deformación de diez muestras de barras de acero de 12 mm sometidas a esfuerzos de tracción conocidos. En este caso, las mediciones realizadas contienen los inevitables errores de observación y de los instrumentos y es imposible garantizar la identidad de las diez barras usadas para el experimento. Por tanto, no sería conveniente encontrar una función interpoladora de mucha complejidad, cuya mayor complicación se debería posiblemente a haber sido obligada a satisfacer un conjunto de datos que contienen errores.

En ambos ejemplos, lo que realmente conviene es encontrar una función f sencilla que satisfaga “lo mejor posible” las exigencias de cada caso. Este es el enfoque del ajuste de curvas, que se precisará a continuación.

El problema del ajuste de curvas

Sea $A = \{x_1, x_2, \dots, x_m\}$ un conjunto de valores de la variable x entre los cuales puede existir cualquier orden e incluso, puede haber repeticiones. Sea f una función definida en A , que toma valores $f_j = f(x_j)$ ($j = 1, 2, \dots, m$), los cuales usualmente se desconocen. Sea y_j el valor observado de f_j . A la diferencia entre el valor real y el observado se le denomina error de observación.

Por otra parte, sea G una familia de funciones aproximantes. Esta familia consta en general de infinitas funciones que se diferencian entre sí por los valores que toman ciertos parámetros cuyo número es una característica importante de la familia. Decidir cual es la familia de funciones aproximantes que se utilizará en un problema de ajuste de curvas es una cuestión importante y a veces muy difícil y que suele encontrar solución a partir de un conocimiento profundo del problema que se está resolviendo. Por el momento, no es necesario entrar en otros detalle acerca de G .

El problema del ajuste de curvas consiste, en encontrar aquella función g de la familia G que se aproxime “lo mejor posible” a los datos (x_j, y_j) ($j = 1, 2, \dots, m$). En términos más formales, se trata de encontrar la función $g \in G$ que haga mínima la desviación cuadrática:

$$D = \sum_{j=1}^m [g(x_j) - y_j]^2$$

La desviación cuadrática es una manera muy conveniente de medir la separación que existe entre el modelo $g(x)$ y los valores observados $\{y_1, y_2, \dots, y_m\}$. Por una parte, las desviaciones de los datos respecto al modelo aparecen elevadas al cuadrado, esto las hace positivas e impide que desviaciones positivas y negativas se compensen mutuamente. Por otra parte, al elevar las desviaciones al cuadrado las desviaciones mayores resultan incrementadas respecto a las más pequeñas, produciéndose un efecto de penalización que está muy acorde con el sentido que las personas otorgan a los errores, pues siempre se considera preferible muchas desviaciones pequeñas que pocas desviaciones grandes. Además, la suma de cuadrados es una función muy fácil de tratar matemáticamente.

Antes de estudiar la forma general en que el problema se resuelve, se incluye un ejemplo muy sencillo para aclarar los conceptos anteriores.

Ejemplo 1

Halle las ecuaciones que permiten determinar la recta no vertical que mejor se ajusta a un conjunto de datos (x_j, y_j) ($j = 1, 2, \dots, m$).

Solución:

Todas las rectas no verticales responden al modelo general:

$$g(x) = C_1x + C_2$$

donde C_1 y C_2 son parámetros reales. En este caso la familia G será:

$$G = \{g: g(x) = C_1x + C_2\}$$

La recta de mejor ajuste será aquella que minimice a:

$$D = \sum_{j=1}^m [C_1x_j + C_2 - y_j]^2$$

Como D es una función cuadrática de C_1 y C_2 ella posee un punto de mínimo que puede obtenerse igualando a cero las derivadas parciales respecto a C_1 y C_2 , es decir:

$$\frac{\partial D}{\partial C_1} = \sum_{j=1}^m 2[C_1x_j + C_2 - y_j]x_j = 0$$

$$\text{y } \frac{\partial D}{\partial C_2} = \sum_{j=1}^m 2[C_1x_j + C_2 - y_j] = 0$$

Separando cada suma en tres sumandos y extrayendo C_1 y C_2 factores comunes, se obtiene:

$$C_1 \sum_{j=1}^m x_j^2 + C_2 \sum_{j=1}^m x_j = \sum_{j=1}^m x_j y_j$$

$$C_1 \sum_{j=1}^m x_j + mC_2 = \sum_{j=1}^m y_j$$

Estas dos ecuaciones forman un sistema lineal con las incógnitas C_1 y C_2 ; la solución de este sistema determina la recta

$$g(x) = C_1x + C_2$$

de mejor ajuste a los datos (x_j, y_j) ($j = 1, 2, \dots, m$).

Modelos lineales

Cuando la familia G de funciones aproximantes está formada por funciones del tipo:

$$g(x) = C_1 g_1(x) + C_2 g_2(x) + \cdots + C_n g_n(x)$$

se dice que se está ajustando un modelo lineal. Realmente, sería más riguroso decir que el modelo es lineal respecto a los parámetros C_1, C_2, \dots, C_n pues, en general, no es lineal respecto a la variable x .

Nótese que, cuando se utiliza un modelo lineal, la familia G es un espacio vectorial generado por el conjunto de funciones $\{g_1, g_2, \dots, g_n\}$. Si este conjunto es linealmente independiente, es decir, si ninguna de las funciones $g_i(x)$ se puede expresar como combinación lineal de las demás, entonces el conjunto $\{g_1, g_2, \dots, g_n\}$ forma una base del espacio vectorial G . En todo lo que sigue, se supondrá que este es el caso.

Cuando se utiliza un modelo lineal, el problema de ajuste de curvas tiene una solución muy simple y elegante; en el caso de modelos no lineales el problema es, en general, mucho más complejo y será tratado más adelante. A continuación se ofrece una lista de ejemplos importantes de modelos lineales y de modelos no lineales.

Modelos lineales:

$$\begin{array}{ll} g(x) = C_1 x + C_2 & g_1(x) = x, \quad g_2(x) = 1 \\ g(x) = C_1 x^2 + C_2 x + C_3 & g_1(x) = x^2, \quad g_2(x) = x, \quad g_3(x) = 1 \\ g(x) = C_1 x + C_2 + C_3 \cos \frac{\pi}{6} x + C_3 \sin \frac{\pi}{6} x & g_1(x) = x, \quad g_2(x) = 1, \quad g_3(x) = \cos \frac{\pi}{6} x \\ & g_4(x) = \sin \frac{\pi}{6} x \end{array}$$

Modelos no lineales:

$$\begin{aligned} g(x) &= C_1 x^{C_2} \\ g(x) &= C_1 (C_2)^x \\ g(x) &= C_1 + C_2 e^{C_3 x} \end{aligned}$$

Para el caso del un modelo lineal

$$g(x) = C_1 g_1(x) + C_2 g_2(x) + \cdots + C_n g_n(x)$$

la desviación cuadrática toma la forma:

$$D = \sum_{j=1}^m [C_1 g_1(x_j) + C_2 g_2(x_j) + \cdots + C_n g_n(x_j) - y_j]^2$$

Como D es una función cuadrática de las variables C_1, C_2, \dots, C_n ella posee un punto de mínimo que se obtiene igualando a cero las derivadas de D respecto a C_1, C_2, \dots, C_n .

Derivando parcialmente con respecto a C_i ($i = 1, 2, \dots, n$) e igualando a cero:

$$\frac{\partial D}{\partial C_i} = \sum_{j=1}^m 2[C_1 g_1(x_j) + C_2 g_2(x_j) + \dots + C_n g_n(x_j) - y_j] g_i(x_j) = 0$$

que se puede escribir, separando en varias sumas y extrayendo factores comunes, como:

$$C_1 \sum_{j=1}^m g_1(x_j) g_i(x_j) + C_2 \sum_{j=1}^m g_2(x_j) g_i(x_j) + \dots + C_n \sum_{j=1}^m g_n(x_j) g_i(x_j) = \sum_{j=1}^m y_j g_i(x_j)$$

Dando a i los valores 1, 2, ..., n , se obtiene el sistema lineal de ecuaciones:

$$\begin{aligned} \left[\sum_{j=1}^m g_1^2(x_j) \right] C_1 + \left[\sum_{j=1}^m g_1(x_j) g_2(x_j) \right] C_2 + \dots + \left[\sum_{j=1}^m g_1(x_j) g_n(x_j) \right] C_n &= \sum_{j=1}^m g_1(x_j) y_j \\ \left[\sum_{j=1}^m g_2(x_j) g_1(x_j) \right] C_1 + \left[\sum_{j=1}^m g_2^2(x_j) \right] C_2 + \dots + \left[\sum_{j=1}^m g_2(x_j) g_n(x_j) \right] C_n &= \sum_{j=1}^m g_2(x_j) y_j \\ &\vdots \\ \left[\sum_{j=1}^m g_n(x_j) g_1(x_j) \right] C_1 + \left[\sum_{j=1}^m g_n(x_j) g_2(x_j) \right] C_2 + \dots + \left[\sum_{j=1}^m g_n^2(x_j) \right] C_n &= \sum_{j=1}^m g_n(x_j) y_j \end{aligned} \quad (1)$$

que se denomina *sistema normal*. La notación se simplifica un poco si se tiene en cuenta que en el espacio vectorial de las funciones definidas en el conjunto $\{x_1, x_2, \dots, x_m\}$ con $m > n$ se puede definir el producto escalar como:

$$f \cdot g = \sum_{j=1}^m f(x_j) g(x_j)$$

Con este convenio, todas las sumas que aparecen en el sistema normal se pueden expresar como productos escalares de funciones:

$$\begin{aligned} (g_1 \cdot g_1) C_1 + (g_1 \cdot g_2) C_2 + \dots + (g_1 \cdot g_n) C_n &= g_1 \cdot y \\ (g_2 \cdot g_1) C_1 + (g_2 \cdot g_2) C_2 + \dots + (g_2 \cdot g_n) C_n &= g_2 \cdot y \\ &\vdots \\ (g_n \cdot g_1) C_1 + (g_n \cdot g_2) C_2 + \dots + (g_n \cdot g_n) C_n &= g_n \cdot y \end{aligned} \quad (2)$$

El determinante del sistema (1) se llama determinante de Gram del conjunto $\{g_1, g_2, \dots, g_n\}$ y se demuestra que no es cero si el conjunto $\{g_1, g_2, \dots, g_n\}$ es linealmente independiente. Como este es el caso que se está suponiendo, puede asegurarse que el sistema posee solución única.

El sistema normal se puede representar también en una forma más compacta y más eficiente en cuanto al tiempo de ejecución, utilizando la matriz \mathbf{G} de orden n por m , la cual se define como:

$$\mathbf{G} = \begin{bmatrix} g_1(x_1) & g_1(x_2) & g_1(x_3) & \cdots & g_1(x_m) \\ g_2(x_1) & g_2(x_2) & g_2(x_3) & \cdots & g_2(x_m) \\ \vdots & \vdots & \vdots & & \vdots \\ g_n(x_1) & g_n(x_2) & g_n(x_3) & \cdots & g_n(x_m) \end{bmatrix}$$

Note que el producto de \mathbf{G} por su traspuesta \mathbf{G}' es:

$$\begin{aligned} \mathbf{GG}' &= \begin{bmatrix} g_1(x_1) & g_1(x_2) & g_1(x_3) & \cdots & g_1(x_m) \\ g_2(x_1) & g_2(x_2) & g_2(x_3) & \cdots & g_2(x_m) \\ \vdots & \vdots & \vdots & & \vdots \\ g_n(x_1) & g_n(x_2) & g_n(x_3) & \cdots & g_n(x_m) \end{bmatrix} \begin{bmatrix} g_1(x_1) & g_2(x_1) & \cdots & g_n(x_1) \\ g_1(x_2) & g_2(x_2) & \cdots & g_n(x_2) \\ g_1(x_3) & g_2(x_3) & \cdots & g_n(x_3) \\ \vdots & \vdots & & \vdots \\ g_1(x_m) & g_2(x_m) & \cdots & g_n(x_m) \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j=1}^m g_1^2(x_j) & \sum_{j=1}^m g_1(x_j)g_2(x_j) & \cdots & \sum_{j=1}^m g_1(x_j)g_n(x_j) \\ \sum_{j=1}^m g_2(x_j)g_1(x_j) & \sum_{j=1}^m g_2^2(x_j) & \cdots & \sum_{j=1}^m g_2(x_j)g_n(x_j) \\ \vdots & \vdots & & \vdots \\ \sum_{j=1}^m g_n(x_j)g_1(x_j) & \sum_{j=1}^m g_n(x_j)g_2(x_j) & \cdots & \sum_{j=1}^m g_n^2(x_j) \end{bmatrix} \end{aligned}$$

que es la matriz cuadrada de orden n de los coeficientes del sistema normal (1). Si se definen las matrices columna:

$$\mathbf{C} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{bmatrix} \quad \text{y} \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

el sistema normal se puede escribir como:

$$(\mathbf{GG}') \mathbf{C} = \mathbf{GY} \quad (3)$$

Desde el punto de vista del algoritmo computacional, lo más eficiente es calcular previamente y almacenar en memoria la matriz \mathbf{G} para obtener los coeficientes de la matriz \mathbf{GG}' y los términos independientes \mathbf{GY} .

Algoritmo en seudo código para ajustar modelos lineales

El siguiente algoritmo permite ajustar el modelo lineal $g(x) = C_1g_1(x) + C_2g_2(x) + \cdots + C_ng_n(x)$ a los datos (x_j, y_j) ($j = 1, 2, \dots, m$) donde se supone que el conjunto $\{g_1, g_2, \dots, g_n\}$ es linealmente independiente y que $m > n$. El algoritmo calcula los coeficientes del modelo que minimizan la desviación cuadrática:

$$D = \sum_{j=1}^m [g(x_j) - y_j]^2$$

Calcular la matriz $\mathbf{G} = [g(x_j)]$ ($i = 1, 2, \dots, n; j = 1, 2, \dots, m$)

Formar la matriz \mathbf{GG}' del sistema normal

Formar la matriz columna $\mathbf{Y} = [y_j]$

Formar la matriz de términos independientes \mathbf{GY}

Resolver el sistema lineal de n por n (\mathbf{GG}') $\mathbf{C} = \mathbf{GY}$ mediante el método de Gauss

La matriz columna \mathbf{C} contiene los coeficientes del modelo ajustado

Terminar

Para muchos modelos lineales importantes la matriz \mathbf{GG}' suele ser mal condicionada, por tanto, debe utilizarse doble precisión en los cálculos y trabajar con estrategia parcial de pivote o, mejor aun, estrategia total. Para evitar el mal condicionamiento puede también utilizarse la idea de hallar para el espacio G una base $\{g_1, g_2, \dots, g_n\}$ de funciones ortogonales. Esto introduce un paso adicional en el algoritmo, pero, a cambio de ello la matriz del sistema (2) es diagonal y su solución es inmediata sin necesidad de utilizar el método de Gauss. En la literatura recomendada al final del capítulo puede encontrar referencias acerca de las familias de funciones ortogonales.

Ejemplo 2

(El siguiente enunciado es el del ejemplo 4 del comienzo del capítulo). En un laboratorio se está analizando la resistencia de las barras de acero corrugadas de 12 mm para la construcción. Para ello se someten 10 muestras de la misma longitud a grandes esfuerzos de tensión y se mide la deformación sufrida. Los resultados se muestran en la tabla 1 y en la figura 2 se han representado en un plano $T-s$.

T (tensión) kilonewton	5,1	7,7	10,8	13,2	15,6	18,1	22,2	23,9	26,3	27,5
s (deformación) milímetro	0,10	0,17	0,24	0,30	0,36	0,40	0,53	0,70	0,85	1,03

Tabla 2

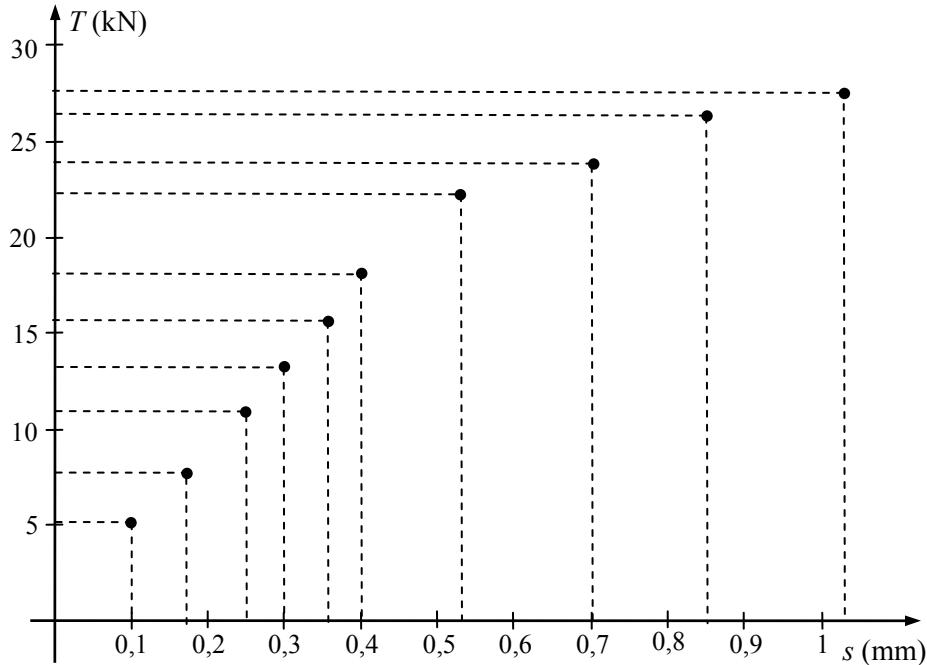


Figura 3

- a) Ajuste un modelo del tipo $T = a_0 + a_1 s$ y b) otro del tipo $T = b_0 + b_1 s + b_2 s^2$ para representar el comportamiento de este material para tensiones en el rango de 0 a 30 kilonewton.

Solución:

- a) Para ajustar a los datos el modelo lineal $T = a_0 + a_1 s$ se escogerán como funciones base:

$$g_0(s) = 1 \quad \text{y} \quad g_1(s) = s$$

de modo que $T(s) = a_0 g_0(s) + a_1 g_1(s) = a_0 + a_1 s$

Los coeficientes a_0 y a_1 se hallan resolviendo el sistema normal:

$$\begin{aligned} (g_0 \cdot g_0) a_0 + (g_0 \cdot g_1) a_1 &= g_0 \cdot T \\ (g_1 \cdot g_0) a_0 + (g_1 \cdot g_1) a_1 &= g_1 \cdot T \end{aligned}$$

$$\text{donde: } g_0 \cdot g_0 = \sum_{j=1}^{10} g_0^2(s_j) = \sum_{j=1}^{10} 1^2 = 10$$

$$g_0 \cdot g_1 = g_0 \cdot g_1 = \sum_{j=1}^{10} g_0(s_j) g_1(s_j) = \sum_{j=1}^{10} 1 s_j = \sum_{j=1}^{10} s_j = 4,68$$

$$g_1 \cdot g_1 = \sum_{j=1}^{10} g_1^2(s_j) = \sum_{j=1}^{10} s_j^2 = 3,0304$$

$$g_0 \cdot T = \sum_{j=1}^{10} g_0(s_j) T_j = \sum_{j=1}^{10} 1 T_j = \sum_{j=1}^{10} T_j = 170,4$$

$$g_1 \cdot T = \sum_{j=1}^{10} g_1(s_j) T_j = \sum_{j=1}^{10} s_j T_j = 100,403$$

Entonces, el sistema normal resulta:

$$\begin{aligned} 10 a_0 + 4,68 a_1 &= 170,4 \\ 4,68 a_0 + 3,0304 a_1 &= 100,403 \end{aligned}$$

cuya solución es: $a_0 = 5,53$
 $a_1 = 24,595$

por tanto, la recta de mejor ajuste será:

$$T = 5,53 + 24,595 s$$

b) En el caso del modelo cuadrático $T = b_0 + b_1 s + b_2 s^2$ se pueden tomar las funciones base

$$g_0(s) = 1; \quad g_1(s) = s \quad y \quad g_2(s) = s^2$$

y los coeficientes b_0, b_1 y b_2 se calculan mediante el sistema normal:

$$\begin{aligned} (g_0 \cdot g_0)b_0 + (g_0 \cdot g_1)b_1 + (g_0 \cdot g_2)b_2 &= g_0 \cdot T \\ (g_1 \cdot g_0)b_0 + (g_1 \cdot g_1)b_1 + (g_1 \cdot g_2)b_2 &= g_1 \cdot T \\ (g_2 \cdot g_0)b_0 + (g_2 \cdot g_1)b_1 + (g_2 \cdot g_2)b_2 &= g_2 \cdot T \end{aligned}$$

Esto es:

$$\begin{aligned} 10b_0 + \left(\sum_{j=1}^{10} s_j \right) b_1 + \left(\sum_{j=1}^{10} s_j^2 \right) b_2 &= \sum_{j=1}^{10} T_j \\ \left(\sum_{j=1}^{10} s_j \right) b_0 + \left(\sum_{j=1}^{10} s_j^2 \right) b_1 + \left(\sum_{j=1}^{10} s_j^3 \right) b_2 &= \sum_{j=1}^{10} s_j T_j \\ \left(\sum_{j=1}^{10} s_j^2 \right) b_0 + \left(\sum_{j=1}^{10} s_j^3 \right) b_1 + \left(\sum_{j=1}^{10} s_j^4 \right) b_2 &= \sum_{j=1}^{10} s_j^2 T_j \end{aligned}$$

Una vez calculadas las sumas, el sistema queda:

$$\begin{aligned} 10b_0 + 4,68b_1 + 3,0304b_2 &= 170,4 \\ 4,68b_0 + 3,0304b_1 + 2,35612b_2 &= 100,403 \\ 3,0304b_0 + 2,35612b_1 + 2,02127b_2 &= 73,1248 \end{aligned}$$

Cuya solución, mediante el método de Gauss, es:

$$\begin{aligned} b_0 &= -0,63 \\ b_1 &= 55,955 \\ b_2 &= -28,100 \end{aligned}$$

Luego, el modelo cuadrático de mejor ajuste será:

$$T = -0,63 + 55,955 s - 28,1 s^2$$

En la figura 3 se muestran los puntos experimentales y los dos modelos ajustados.

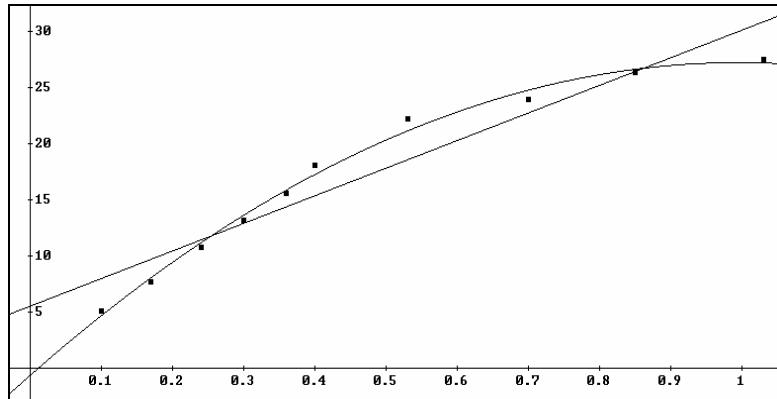


Figura 3

Ejemplo 3

Plantee el sistema normal de ecuaciones para ajustar el modelo $y = a_0 + a_1x + a_2e^{-x} + a_3 \operatorname{sen} x$ a un conjunto de datos $(x_j, y_j), j = 1, 2, \dots, 50$.

Solución:

Tomando como funciones base:

$$\begin{aligned} g_0(s) &= 1 \\ g_1(s) &= x \\ g_2(s) &= e^{-x} \\ g_3(s) &= \operatorname{sen} x \end{aligned}$$

se tiene el sistema normal:

$$\begin{aligned} (g_0 \cdot g_0)a_0 + (g_0 \cdot g_1)a_1 + (g_0 \cdot g_2)a_2 + (g_0 \cdot g_3)a_3 &= g_0 \cdot y \\ (g_1 \cdot g_0)a_0 + (g_1 \cdot g_1)a_1 + (g_1 \cdot g_2)a_2 + (g_1 \cdot g_3)a_3 &= g_1 \cdot y \\ (g_2 \cdot g_0)a_0 + (g_2 \cdot g_1)a_1 + (g_2 \cdot g_2)a_2 + (g_2 \cdot g_3)a_3 &= g_2 \cdot y \\ (g_3 \cdot g_0)a_0 + (g_3 \cdot g_1)a_1 + (g_3 \cdot g_2)a_2 + (g_3 \cdot g_3)a_3 &= g_3 \cdot y \end{aligned}$$

Esto es:

$$\begin{aligned}
& 50a_0 + \left(\sum_{j=1}^{50} x_j \right) a_1 + \left(\sum_{j=1}^{50} e^{-x_j} \right) a_2 + \left(\sum_{j=1}^{50} \operatorname{sen} x_j \right) a_3 = \sum_{j=1}^{50} x_j \\
& \left(\sum_{j=1}^{50} x_j \right) a_0 + \left(\sum_{j=1}^{50} x_j^2 \right) a_1 + \left(\sum_{j=1}^{50} x_j e^{-x_j} \right) a_2 + \left(\sum_{j=1}^{50} x_j \operatorname{sen} x_j \right) a_3 = \sum_{j=1}^{50} x_j y_j \\
& \left(\sum_{j=1}^{50} e^{-x_j} \right) a_0 + \left(\sum_{j=1}^{50} x_j e^{-x_j} \right) a_1 + \left(\sum_{j=1}^{50} e^{-2x_j} \right) a_2 + \left(\sum_{j=1}^{50} e^{-x_j} \operatorname{sen} x_j \right) a_3 = \sum_{j=1}^{50} y_j e^{-x_j} \\
& \left(\sum_{j=1}^{50} \operatorname{sen} x_j \right) a_0 + \left(\sum_{j=1}^{50} x_j \operatorname{sen} x_j \right) a_1 + \left(\sum_{j=1}^{50} e^{-x_j} \operatorname{sen} x_j \right) a_2 + \left(\sum_{j=1}^{50} \operatorname{sen}^2 x_j \right) a_3 = \sum_{j=1}^{50} y_j \operatorname{sen} x_j
\end{aligned}$$

Ajuste de modelos no lineales mediante cambios de variables

Cuando el modelo que se necesita ajustar no es lineal respecto a sus parámetros, el problema se hace mucho más complicado. No obstante, existen formas de abordarlo, algunas que brindan los valores exactos de los parámetros del modelo, otras que ofrecen una solución aproximada.

En algunos casos, el modelo no lineal se puede linealizar mediante cambios de variables y, entonces, se aplica al modelo linealizado el procedimiento de las ecuaciones normales. Aunque esta forma de proceder no ofrece los valores exactos de los parámetros que optimizan el sistema original, la solución que se obtiene suele ser bastante aproximada y, por lo general, siempre que esta variante es posible, se prefiere a otros métodos exactos pero más complicados. A continuación se muestran algunos modelos no lineales y los cambios de variables que los convierten en lineales.

a) $y = a(b)^x$

Aplicando logaritmos: $\ln y = \ln a + (\ln b)x$

Haciendo $Y = \ln y$, $K_1 = \ln a$, $K_2 = \ln b$ el sistema se reduce a:

$$Y = K_1 + K_2 x$$

que es un modelo lineal, el cual se ajusta a los datos modificados: (x_j, Y_j) donde $Y_j = \ln y_j$ para $j = 1, 2, \dots, m$. Una vez conocidos K_1 y K_2 , los parámetro a y b se hallan como:

$$a = e^{K_1} \quad y \quad b = e^{K_2}$$

b) $y = ax^b$

Aplicando logaritmos: $\ln y = \ln a + b(\ln x)$

Haciendo $Y = \ln y$, $X = \ln x$, $K_1 = \ln a$, $K_2 = b$ el sistema se convierte en:

$$Y = K_1 + K_2 X$$

que se trata de un modelo lineal en las variables X e Y , el cual se puede ajustar a los datos transformados: (X_j, Y_j) donde $X_j = \ln x_j$ y $Y_j = \ln y_j$ para $j = 1, 2, \dots, m$. Hallados K_1 y K_2 , los parámetro a y b se determinan sin dificultad.

c) $y = a + \frac{b}{x}$

Se convierte en un modelo lineal si se define la variable: $X = \frac{1}{x}$. Resulta el modelo lineal:

$$y = a + bX$$

el cual es ajustado a los datos transformados (X_j, y_j) donde $X_j = \frac{1}{x_j}$ para $j = 1, 2, \dots, m$.

d) $y = \frac{1}{ax + b}$

Se puede escribir como

$$\frac{1}{y} = ax + b$$

Llamando $Y = \frac{1}{y}$, queda: $Y = ax + b$

que es un modelo lineal, el cual se debe ajustar a los datos modificados (x_j, Y_j) donde

$$Y_j = \frac{1}{y_j} \text{ para } j = 1, 2, \dots, m.$$

e) $y = \frac{x}{ax + b}$

El cual se puede expresar como $\frac{1}{y} = \frac{ax + b}{x} = a + \frac{b}{x}$

Haciendo $X = \frac{1}{x}$ y $Y = \frac{1}{y}$ el modelo se reduce a la forma lineal:

$$Y = a + bX$$

que se ajusta a los datos (X_j, Y_j) donde $X_j = \frac{1}{x_j}$ y $Y_j = \frac{1}{y_j}$ para $j = 1, 2, \dots, m$.

Ejemplo 4

Cuando un condensador de C farad con una carga de q coulomb se descarga a través de una resistencia de R ohm, como se muestra en la figura 4, la corriente de descarga, en ampere, viene dada por

$$i = \frac{q}{RC} e^{-\frac{1}{RC}t}$$

donde t es el tiempo, medido en segundos, a partir del comienzo de la descarga. Determine los valores de q y la constante RC a partir de los datos de la corriente de descarga medida en cinco instantes y que muestra la tabla 3.

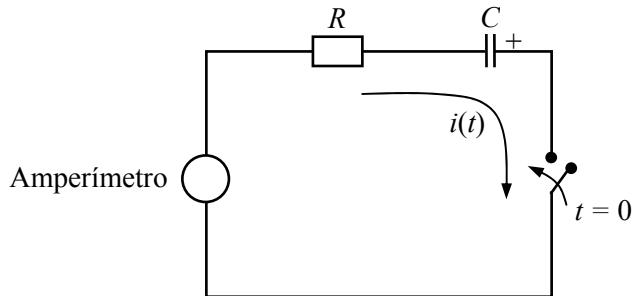


Figura 4

j	t_j	$i(t_j)$
1	0,1	1,228
2	0,2	1,005
3	0,3	0,823
4	0,4	0,674
5	0,5	0,552

Tabla 3

Solución:

Para simplificar, se llamará: $a = \frac{q}{RC}$ y $b = -\frac{1}{RC}$

con lo cual el modelo a ajustar es: $i = ae^{bt}$

Aplicando logaritmos: $\ln i = \ln a + bt$

Introduciendo las nuevas variables: $y = \ln i$ y $c = \ln a$

se obtiene el modelo lineal: $y = c + bt$

El cual será ajustado a los datos transformados (t_j, y_j) donde $y_j = \ln i(t_j)$ $j = 1, 2, 3, 4, 5$ (tabla 4).

Las funciones básicas serán: $g_0(t) = 1$ y $g_1(t) = t$

y el sistema normal de ecuaciones:

$$\begin{aligned} (g_0 \cdot g_0)c + (g_0 \cdot g_1)b &= g_0 \cdot y \\ (g_1 \cdot g_0)c + (g_1 \cdot g_1)b &= g_1 \cdot y \end{aligned}$$

Es decir:

$$\begin{aligned} 5c + \left(\sum_{j=1}^5 t_j \right) b &= \sum_{j=1}^5 y_j \\ \left(\sum_{j=1}^5 t_j \right) c + \left(\sum_{j=1}^5 t_j^2 \right) b &= \sum_{j=1}^5 t_j y_j \end{aligned}$$

j	t_j	$i(t_j)$	$y_j = \ln i(t_j)$
1	0,1	1,228	0,205387
2	0,2	1,005	0,004988
3	0,3	0,823	-0,194799
4	0,4	0,674	-0,394525
5	0,5	0,552	-0,693147

Tabla 4

Después de calcular las sumas indicadas, se obtiene:

$$\begin{aligned} 5c + 1,5b &= -0,973172 \\ 1,5c + 0,55b &= -0,491817 \end{aligned}$$

y de aquí: $c = 0,404979$ y $b = -1,998701$

a partir de c se halla $a = e^c = 1,499271$

El modelo ajustado para la corriente de descarga será:

$$i = 1,499271e^{-1,998701t}$$

Como $a = \frac{q}{RC}$ y $b = -\frac{1}{RC}$, resulta:

$$\frac{q}{RC} = 1,499271 \quad y \quad -\frac{1}{RC} = -1,998701$$

y de aquí: $RC = 0,500325$ segundos y $q = 0,750123$ coulomb

Ajuste numérico de modelos no lineales

El procedimiento de linealizar un modelo para calcular aproximaciones a los parámetros no siempre es posible. En estos casos hay otras dos formas de proceder. Una de ellas es minimizar numéricamente la función desviación cuadrática mediante alguno de los métodos que se estudiarán en el capítulo 6. Aunque este es el enfoque más recomendado, su estudio queda pospuesto para ese momento.

El otro enfoque es calcular las derivadas parciales de la función desviación cuadrática, igualarlas a cero y resolver el sistema de ecuaciones no lineales que resulta. Por esta vía suelen aparecer sistemas no lineales muy complejos, cuya solución numérica puede ser muy problemática. Sin embargo, en algunos casos simples la solución numérica es factible e, incluso, a veces se puede reducir el sistema no lineal a una sola ecuación no lineal, la cual se puede resolver por algunos de los métodos estudiados en el capítulo 2. Por esta vía, a modo de muestra, se resolverá el ejemplo 4 de las páginas anteriores.

Ejemplo 5

Ajuste el modelo $i = ae^{bt}$ (vea el ejemplo anterior) a los datos de la tabla 3, sin linealizarlo.

Solución:

La desviación cuadrática viene dada por la expresión:

$$D = \sum_{j=1}^5 (ae^{bt_j} - i_j)^2$$

Derivando parcialmente respecto a a y a b , e igualando a cero:

$$\frac{\partial D}{\partial a} = \sum_{j=1}^5 2(a e^{bt_j} - i_j) e^{bt_j} = 0$$

$$\frac{\partial D}{\partial b} = \sum_{j=1}^5 2(a e^{bt_j} - i_j) a t_j e^{bt_j} = 0$$

Dividiendo la primera ecuación por 2, la segunda por $2a$ y separando cada suma:

$$a \sum_{j=1}^5 e^{2bt_j} - \sum_{j=1}^5 i_j e^{bt_j} = 0 \quad (4)$$

$$a \sum_{j=1}^5 t_j e^{2bt_j} - \sum_{j=1}^5 i_j t_j e^{bt_j} = 0 \quad (5)$$

Si la ecuación (4) se multiplica por $\sum_{j=1}^5 t_j e^{2bt_j}$, la ecuación (5) por $\sum_{j=1}^5 e^{2bt_j}$ y se restan las ecuaciones resultantes, el parámetro a desaparece:

$$\sum_{j=1}^5 i_j e^{bt_j} \sum_{j=1}^5 t_j e^{2bt_j} - \sum_{j=1}^5 e^{2bt_j} \sum_{j=1}^5 i_j t_j e^{bt_j} = 0 \quad (6)$$

y el sistema queda reducido a una ecuación con una sola incógnita: b . Evaluando las sumas para varios valores de b se concluye que en el intervalo $[-2, 0]$ la ecuación tiene una raíz. Aplicando en este intervalo el método de bisección (esto requiere la confección de un pequeño programa) se determinó con seis cifras decimales exactas el valor de $b = -1,999357$. De cualquiera de las ecuaciones (4) o (5) se halla el parámetro $a = 1,499527$. Compárese estos valores con los obtenidos por el método aproximado del ejemplo 4 ($b = -1,998701$ y $a = 1,499271$) que difieren de estos en menos de una milésima.

Ejercicios

- Ajuste a los datos de la siguiente tabla, polinomios de grados 1, 2 y 3. Determine para cada uno la desviación cuadrática obtenida y verifique como la desviación cuadrática es mucho menor en el polinomio de grado 2 que en el de grado 1, pero que, sin embargo, no disminuye significativamente en el polinomio de grado 3. ¿Qué conclusión obtiene de este resultado?

i	1	2	3	4	5	6	7	8	9	10	11
x_i	0	0,5	2	3	3,5	4,5	5	6	7	7,7	8,5
y_i	27,8	22,5	12,9	9,3	6,8	8,1	8,6	11,2	17,9	21,6	31,3

2. Con vista a determinar la velocidad más económica de un tipo de automóvil de pasajeros, se mide, para diferentes velocidades de conducción, cuántos kilómetros recorre por litro de gasolina. Los resultados se muestran en la tabla que sigue. Ajuste un polinomio de segundo grado y halle después para qué velocidad se hace mínimo el consumo (o máximo los kilómetros por litro).

Velocidad (km/h)	30	40	50	60	70	80	90	100
km/litro de gasolina	9,2	9,7	9,9	10	9,8	9,4	8,8	8,1

3. Cuando un cuerpo sólido se mueve a través de un fluido viscoso se produce sobre aquel una fuerza f_r que se opone al movimiento (fuerza resistiva) y que depende de la velocidad del cuerpo. Con vista a esclarecer esta dependencia, se efectúa un experimento en el cual un cuerpo se hace mover a diferentes velocidades a través de un medio viscoso y se mide la fuerza de fricción. Los resultados obtenidos se muestran en la siguiente tabla. Halle el modelo del tipo $f_r = \mu v^k$ que mejor se ajusta a estos datos.

Velocidad (m/s)	1	2	3	4	5
Fricción (newton)	50	153	293	464	663

4. Se sabe que existe una relación entre el nivel de lluvias caído durante un año y la altura sobre el nivel del mar, correspondiendo, por regla general, mayores precipitaciones a los lugares más altos. Para cuantificar este fenómeno se cuenta con los datos de lluvia de la tabla 5, registrados por pluviómetros ubicados en lugares a diferentes altitudes en la región de la Sierra Maestra al oeste de Santiago de Cuba. Determine que relación lineal aproximada existe entre estas variables.

Zona	Altitud (m)	Precipitación anual (mm)	Zona	Altitud (m)	Precipitación anual (mm)
Batey Mabay	60	1290	Macuba	320	1660
El molino	80	1290	Los Lajiales	380	1625
Batey América	100	1310	Oro de Guisa	560	1780
El Alcázar	180	1310	Las Coloradas	420	1840
Maravilla	160	1330	Buena Vista	555	1910
Las Mantecas	160	1400	Duaba	600	1940
Las Pozas	325	1420	Caña Brava	700	2190
Cruce de los baños	160	1510			

Tabla 5

5. El tiempo que se necesita para detener un vehículo es mayor mientras mayor es la velocidad a la que circula y, por tanto, mayor será la longitud L que recorrerá desde que el conductor descubre un hecho peligroso hasta que el vehículo queda inmóvil. En un código de vialidad y tránsito se brinda la siguiente tabla. Ajuste un polinomio de segundo grado a estos datos y determine a qué velocidad circulaba un automóvil que recorrió 70 metros antes de detenerse.

Velocidad (km/h)	30	40	50	60	70	80	90	100	120
L (m)	11,6	18,3	26,1	34,9	47,3	60,1	75,0	91,4	130

6. La resistividad R de un material conductor de electricidad, se define como la resistencia eléctrica de un conductor de una unidad de longitud y una unidad de área en su sección transversal. En ingeniería eléctrica se suele medir en ohm·mm²/m. La resistividad de la mayoría de los metales aumenta con la temperatura. Esta dependencia se puede representar aproximadamente por la relación:

$$R_t = R_0(1 + \alpha\Delta t)$$

donde:
 R_t : Resistividad a la temperatura t
 R_0 : Resistividad a la temperatura t_0
 α : Coeficiente térmico de la resistividad
 Δt : $t - t_0$

Determine los coeficientes R_0 y α correspondientes al cobre a la temperatura $t_0 = 20$ °C a partir de las siguientes mediciones:

Temperatura (°C)	22	24	25	27	28	30
Resistividad (ohm·mm ² /m)	0,01764	0,01778	0,01784	0,01798	0,01805	0,01818

7. Los datos que muestra la tabla 6 corresponden a la entrada mensual de grupos turísticos a las instalaciones de una empresa hotelera durante los dos últimos años. Se desea realizar un pronóstico acerca de la afluencia de grupos turísticos en el año siguiente y para ello se utilizará el modelo:

$$y = a + bt + c \cos \frac{\pi}{6}t + d \sin \frac{\pi}{6}t$$

donde t : meses contados a partir del primer dato de la tabla 6
 y : Grupos turísticos que llegaron en el mes t

Ajuste el modelo a los datos, represente en un mismo sistema coordenado los datos y el modelo ajustado y pronostique cuantos grupos turísticos llegarán en primer trimestre del año próximo.

Mes	Primer año	Segundo año
Enero	954	968
Febrero	869	928
Marzo	832	873
Abril	634	665
Mayo	476	?
Junio	379	?
Julio	591	675
Agosto	571	642
Septiembre	472	578
Octubre	553	703
Noviembre	608	623
Diciembre	797	882

Tabla 6

8. Siguiendo un procedimiento similar al del ejemplo 5, elabore una algoritmo en seudo código para ajustar el modelo no lineal:

$$y = ax^b$$

a un conjunto de datos: $(x_j, y_j) j = 1, 2, \dots, m$.

9. Por análisis teóricos se sabe que el flujo de un líquido a través de una tubería depende de la pendiente de la misma, según un modelo del tipo:

$$Q = as^b$$

donde s es la pendiente de la tubería y a y b son parámetros a determinar. Mediante experimentos se han obtenido los siguientes datos:

Experimento	1	2	3	4	5	6	7	8
Pendiente	0,001	0,003	0,007	0,01	0,02	0,03	0,05	0,06
Flujo (m^3/s)	0,80	1,45	2,30	2,75	4,00	5,00	6,55	7,25

- a) Utilice el algoritmo elaborado en ejercicio 8 para ajustar el modelo.
 b) Aplique logaritmos para linealizar el modelo y, mediante cambios de variables, obtenga a y b .
 c) Compare los resultados obtenidos en a) y en b).
10. Cuando un cuerpo es calentado, su densidad cambia. La densidad a la temperatura t viene dada por la fórmula:

$$\delta_t = \frac{\delta_0}{1 + \beta(t - t_0)}$$

donde δ_0 : Densidad (g/cm^3) a la temperatura t_0 de referencia
 β : Coeficiente de expansión volumétrica ($1/\text{°C}$)

Con vistas a determinar las constantes δ_0 y β para el ácido nítrico a 20 °C, se midió la densidad a diferentes temperaturas y se obtuvo:

t (°C)	19	21	24	25	28	30	32	35
δ (g/cm ³)	1,5102	1,5098	1,5093	1,5091	1,5085	1,5081	1,5077	1,5072

Linealice el modelo y, mediante cambios de variables, determine aproximadamente, los parámetros δ_0 y β .

Otras lecturas recomendadas

El tema de aproximación funcional es sumamente amplio y en él existen multitud de aspectos útiles e interesantes, muchos de los cuales no han podido ser tratados en este texto, dado su carácter introductorio.

Solamente han sido estudiados dos criterios de aproximación: la interpolación y el ajuste de curvas; sin embargo existen otros muy importantes, sobre todo el conocido como criterio minimax y el criterio de Taylor. En el criterio minimax la distancia entre una función y su función aproximante se toma como la máxima diferencia entre las imágenes correspondientes para un cierto conjunto de definición y la aproximación optima será aquella que minimice esta distancia; para el caso en que las funciones aproximantes sean polinomios existen algoritmos aproximados que resuelven el problema con eficiencia y la exactitud deseada, pero el tratamiento requiere de mucho más espacio y tiempo del que se dispone en un curso introductorio. El lector interesado puede consultar “Computing Methods” de I. S. Berezin y N. P. Zhidkov o también “Introduction to Numerical Analysis” de K. E. Atkinson. En el criterio de aproximación de Taylor la proximidad entre dos funciones viene dada por la coincidencia en un punto de referencia de una mayor o menor cantidad de derivadas; se estudia a un nivel elemental en los cursos de Cálculo Infinitesimal empleando como funciones aproximantes los polinomios de Taylor; muchos de los métodos numéricos tratados en este libro están basados en esta forma de aproximación, la cual se supone conocida del lector. En cursos más avanzados se utilizan las funciones racionales de Padé que demuestran ser una vía más eficientes a la hora de evaluar una función aproximante. El estudio de las funciones de Padé para implementar el criterio de Taylor puede ser consultado en muchos libros de Análisis Numérico, entre ellos: “First Course in Numerical Analysis” de A. Ralston.

Los métodos de interpolación tratados en el libro pueden ser ampliados en varios sentidos. Tanto el método de Lagrange como el de Newton dan lugar a fórmulas mucho más simples cuando los nodos de interpolación están ordenados (en forma creciente o decreciente) y son equidistantes. En el caso del método de Newton, aparecen entonces los conceptos de diferencia finita hacia delante, diferencia finita hacia atrás y diferencia central y existen varias fórmulas para realizar la interpolación mediante estos tipos de diferencias. En los cursos de Matemática numérica de algunos años atrás, el tema de interpolación polinomial era tratado mucho más ampliamente debido a que la necesidad de realizar los cálculos manualmente obligaba a la búsqueda de muchos algoritmos que abordaban casos particulares, con el objetivo de hacer las operaciones aritméticas más sencillas (evitar las multiplicaciones y divisiones) y disminuir las equivocaciones. El tema se encuentra tratado ampliamente en el libro “Computational Mathematics” de B. P. Demidovich e I. A. Maron.

El problema de interpolación puede combinarse con el criterio de proximidad de Taylor exigiendo que en los nodos de interpolación la función original y la aproximante coincidan no solo en valor sino en una o varias derivadas; este es conocido como el criterio de aproximación de Hermite. El lector interesado en este enfoque puede consultar la obra de Atkinson antes citada o, con más amplitud, en el libro de Berezin y Zhidkov también citado más arriba.

Aquí se han utilizado como funciones interpoladoras los polinomios y los splines cúbicos, pero otras funciones podrían ser empleadas, en particular las trigonométricas y las exponenciales. La interpolación también puede extenderse a funciones de más de una variable. Aunque el problema puede muchas veces ser reducido a realizar varias interpolaciones unidimensionales, se han desarrollado algoritmos para el tratamiento general en funciones de n variables, los cuales, por su complejidad y uso menos frecuente, han sido excluidos de este libro. Ambos temas pueden ser consultados en la enciclopédica obra ya citada de Berezin y Zhidkov.

En muchas ocasiones los nodos de interpolación están predeterminados por el problema que se está resolviendo, pero en otras ocasiones, el analista puede decidir el valor exacto donde desea tomarlos. El problema de ubicar los nodos de la mejor manera posible (minimizando el máximo error de interpolación) está tratado en muchos textos, vinculado con el tema de los polinomios de Chebyshev. Los lectores motivados por este asunto encontrarán una buena referencia, a un nivel elemental en “Elementary Numerical Analysis” de S. D. Conte y, con más profundidad en el texto de Atkinson, ya citado.

Aunque en este libro el tema de las funciones spline para la interpolación está tratado con una relativa amplitud para un texto de carácter elemental, muchos aspectos no han sido tocados. En particular, el error de interpolación el cual puede ser consultado en muchos lugares, entre ellos en el libro, varias veces citado, de K. E. Atkinson. Un enfoque mucho más general y amplio de los splines, no solamente como funciones interpoladoras, sino para el ajuste y para el diseño gráfico, construido sobre el enfoque de las funciones spline básicas (B-splines), está contenido en el excelente informe técnico “DT_NURBS Spline Geometry Subprogram Library Theory Document”, Versión 3.6, preparado por Boeing Shared Services Group y editado por Naval Surface Warfare Center, Carderock División.

El tema del ajuste de modelos no lineales como un problema de optimización de varias variables, ha sido pospuesto para el capítulo 6, como una aplicación de las técnicas numéricas de optimización. En el caso de los modelos lineales, el uso de polinomios ortogonales como funciones básicas no se ha contemplado en el texto. La teoría y práctica de este tipo de técnica puede consultarse a un nivel básico en la obra de A. Ralston antes citada y, con más profundidad, en la obra clásica: “Analysis of Numerical Methods” de E. Isaacson y H. B. Keller.

Principales ideas del capítulo

- Los problemas de aproximación de funciones aparecen en la teoría y en la práctica con gran frecuencia; bien porque no se conoce una expresión analítica para la función $f(x)$, sino valores aislados $f(x_1), f(x_2), \dots, f(x_n)$ de la misma o porque su expresión exacta es muy complicada, y se necesita disponer de una expresión más simple que permita, aunque sea de manera aproximada, evaluar la función en valores de x necesarios.
- Si se conocen los valores que toma la función $f(x)$ en los $n + 1$ puntos diferentes x_0, x_1, \dots, x_n , (llamados nodos de interpolación), el problema de interpolación consiste en hallar una

función $g(x)$ cuyos valores puedan ser calculados para cualquier x en un intervalo que contiene a x_0, x_1, \dots, x_n , y de manera que $g(x_i) = f(x_i)$ para $i = 0, 1, 2, \dots, n$.

- A la diferencia entre la función interpolada y la interpoladora se le llama *error de interpolación* y se denota $R(x)$.
- El error de interpolación depende del punto x en que se interpole; es cero si x es un nodo de interpolación y, por lo general, aumenta a medida que x está más distante de los nodos. En particular, el error de interpolación suele ser mucho mayor (en valor absoluto) en los casos de extrapolación que de interpolación.
- Para $n + 1$ nodos de interpolación diferentes x_0, x_1, \dots, x_n y una función f que toma valores $y_0 = f(x_0), y_1 = f(x_1), \dots, y_n = f(x_n)$, existe uno y solo un polinomio interpolador de grado menor o igual que n .
- Si $f(x)$ es derivable $n + 1$ veces en un intervalo cerrado I que incluye a los nodos de interpolación x_0, x_1, \dots, x_n del polinomio interpolador $p(x)$ y al número x , entonces existe en I al menos un valor c tal que, el error de interpolación en x es:

$$R(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x - x_0)(x - x_1) \cdots (x - x_n)$$

- La fórmula de interpolación de Lagrange se basa en la expresión:

$$p(x) = y_0 L_0(x) + y_1 L_1(x) + \cdots + y_n L_n(x)$$

$$\text{donde: } L_i(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} \quad i = 0, 1, 2, \dots, n$$

- La interpolación polinomial no debe usarse como una técnica global para representar a una función en un intervalo grande utilizando muchos nodos, todo lo contrario, se trata de un procedimiento para aproximar *localmente* una función mediante un polinomio.
- Las diferencias divididas de una función $f(x)$ definida en los números diferentes x_0, x_1, \dots, x_n de cualquier orden k se definen recursivamente como:

$$f[x_0, x_1, \dots, x_k] = \frac{f[x_1, x_2, \dots, x_k] - f[x_0, x_1, \dots, x_{k-1}]}{x_k - x_0} \quad k = 1, 2, 3, \dots, n$$

donde la diferencia de orden cero es $f[x] = f(x)$

- El método de interpolación de Newton se basa en la fórmula:

$$p_{n+1}(x) = p_n(x) + \tilde{R}_n(x)$$

$$\text{donde: } \tilde{R}_n(x) = f[x_0, x_1, \dots, x_n, x_{n+1}] (x - x_0)(x - x_1) \cdots (x - x_n)$$

es una estimación del error de interpolación de $p_n(x)$.

- Las diferencias divididas de orden m de un polinomio de grado m son constantes (es decir, no dependen de los nodos seleccionados) y las de orden mayor que m son nulas.

- Una función spline es una función polinomial por tramos que es continua y posee derivadas continuas hasta un cierto orden. Además de las condiciones de continuidad y suavidad, el spline deberá satisfacer algunas otras condiciones adecuadas al problema que se desea resolver.
- El spline cúbico interpolador, para el tramo $x_i \leq x \leq x_{i+1}$, viene dado por

$$s(x) = \frac{v^3 M_i + u^3 M_{i+1}}{6h_i} + \frac{vy_i + uy_{i+1}}{h_i} - \frac{h_i(vM_i + uM_{i+1})}{6}$$

donde: $u = x - x_i$
 $v = x_{i+1} - x$

y los números M_i , $i = 0, 1, \dots, n$ se calculan resolviendo un sistema lineal de ecuaciones con diagonal predominante cuya estructura es diferente para el spline natural, anclado o periódico.

- El problema del ajuste de curvas consiste, en encontrar aquella función g de la familia G que se aproxime “lo mejor posible” a los datos (x_j, y_j) ($j = 1, 2, \dots, m$) lo cual significa hacer mínima la desviación cuadrática:

$$D = \sum_{j=1}^m [g(x_j) - y_j]^2$$

- Cuando el modelo a ajustar es lineal, es decir, del tipo:

$$g(x) = C_1 g_1(x) + C_2 g_2(x) + \dots + C_n g_n(x)$$

entonces los parámetros se hallan resolviendo el sistema normal de ecuaciones:

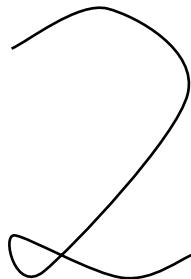
$$\begin{aligned} (g_1 \cdot g_1)C_1 + (g_1 \cdot g_2)C_2 + \dots + (g_1 \cdot g_n)C_n &= g_1 \cdot y \\ (g_2 \cdot g_1)C_1 + (g_2 \cdot g_2)C_2 + \dots + (g_2 \cdot g_n)C_n &= g_2 \cdot y \\ &\vdots \\ (g_n \cdot g_1)C_1 + (g_n \cdot g_2)C_2 + \dots + (g_n \cdot g_n)C_n &= g_n \cdot y \end{aligned}$$

donde $f \cdot g = \sum_{j=1}^m f(x_j)g(x_j)$

- Cuando el modelo a ajustar no es lineal el problema se puede resolver aproximadamente a veces linealizando el modelo, o bien resolviendo el sistema no lineal de ecuaciones que surge de igualar a cero las derivadas parciales de la desviación cuadrática, o bien resolviendo numéricamente el problema de minimizar la desviación cuadrática, como se verá en el capítulo 6.

Auto examen

1. ¿En qué consiste el problema general de la aproximación de funciones y, en particular, el problema de interpolación?
2. Ofrezca dos ejemplos prácticos de problemas de aproximación funcional: uno que deba ser resuelto por interpolación y otro por ajuste de curvas.
3. Se tienen los valores de una función $f(x)$ en cuatro nodos de interpolación. Diga cuales de las siguientes afirmaciones son verdaderas y cuales son falsas y porqué.
 - a) Existen infinitos polinomios de grado menor o igual que dos para dichos nodos.
 - b) Hay uno y solo un polinomio de grado tres para dichos nodos.
 - c) Pudiera haber un polinomio interpolador de grado menor o igual que 5, pero no se puede asegurar.
4. Halle un polinomio interpolador de grado 2 que represente a la función $\frac{\sin x}{x}$ en el intervalo $[-1, 1]$. Dé una cota del error de interpolación.
5. ¿A qué se llama un spline cúbico interpolador?
6. Plantee los sistemas de ecuaciones necesarios para determinar los splines cúbicos naturales que permitan obtener una curva paramétrica con la forma de la figura que sigue:



7. El consumo eléctrico en un domicilio depende de muchos factores pero fundamentalmente de la época del año. Proponga un modelo lineal que sea adecuado para ajustarse a los datos mensuales de consumo eléctrico (en kilowatt·hora). Ajuste el modelo propuesto a los datos de consumo eléctrico de su casa en los últimos 24 meses (los recibos que conserve) y utilice el modelo para pronosticar el consumo del mes próximo.
8. Ajuste a los datos que siguen los modelos no lineales biparamétricos $y = ax^b$ e $y = ab^x$. Determine cual de los dos se ajusta mejor a los datos.

x	1	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2
y	5,1	5,4	5,7	6,0	6,3	6,6	7,0	7,4	7,8	8,2	8,7

9. Explique tres formas de resolver el problema de ajustar un modelo no lineal a un conjunto de valores experimentales.

CAPÍTULO 5 **Matemática Numérica, 2da Edición**
 Manuel Álvarez, Alfredo Guerra, Rogelio Lau
INTEGRACIÓN NUMÉRICA

Objetivos

Al finalizar el estudio y la ejercitación de este capítulo, el lector debe ser capaz de:

- Describir las limitaciones de la fórmula de Barrow para el cálculo de integrales definidas.
- Describir geométricamente los métodos de trapecios y de Simpson para el cálculo aproximado de integrales definidas.
- Justificar las fórmulas del método de los trapecios y del método de Simpson y de las expresiones para estimar el error de truncamiento.
- Describir la idea fundamental del método de integración numérica de Gauss.
- Justificar algebraicamente la fórmula de Gauss para tres puntos y generalizarla a otros valores de m .
- Deducir la fórmula de extrapolación de Richardson y explicar, a través de ella, el método de integración de Romberg.
- Describir los algoritmos en seudo código de los métodos de trapecios, Simpson, Gauss y Romberg.
- Calcular integrales definidas mediante los algoritmos de trapecios, Simpson, Gauss y Romberg, estimando el error de truncamiento si fuera necesario.
- Modelar problemas simples que conducen a integrales definidas y resolverlos mediante métodos numéricos de integración.
- Explicar la forma en que una integral doble se puede calcular aproximadamente mediante los algoritmos numéricos para integrales unidimensionales, exemplificando para los casos de Simpson y Gauss.
- Calcular integrales dobles mediante los métodos de Gauss y de Simpson y obtener, en este caso, una estimación del error de truncamiento.
- Explicar el algoritmo en seudo código de los métodos de Gauss y de Simpson para el cálculo de integrales dobles.
- Modelar problemas simples que conduzcan a integrales dobles y calcularlas numéricamente por los métodos estudiados.

5.1 Introducción

En el curso de Cálculo Infinitesimal se ha estudiado el concepto de integral definida de una función $f(x)$ en un intervalo $[a, b]$:

$$\int_a^b f(x)dx$$

como el resultado de un proceso de límite de una suma finita. El lector seguramente recuerda las importantes aplicaciones que poseen las integrales en las diferentes ramas de la Geometría, la Física, la Química, las Ciencias Económicas y, prácticamente, en todas las ramas del saber.

Supuesto que $f(x)$ sea continua en $[a, b]$, la integral se puede calcular mediante la regla de Barrow (también llamada de Newton – Leibniz):

$$\int_a^b f(x)dx = F(b) - F(a)$$

donde $F(x)$ es cualquier primitiva de $f(x)$, es decir, una función cuya derivada sea $f(x)$.

El punto débil de este procedimiento analítico para evaluar una integral es la obtención de una función primitiva. Para muchas funciones sencillas se obtienen primitivas con mayor o menor dificultad pero, en muchos casos se presentan integrales para las cuales no existen primitivas que se puedan expresar en términos de funciones elementales. Lo peor es que, en muchas ocasiones, se trata de integrandos sencillos (es decir, formados por funciones elementales). Por supuesto que, si no se tiene una primitiva expresada en términos de funciones elementales, no es posible evaluarla en los límites de integración, y la regla de Barrow se hace inaplicable.

Ejemplo 1

Al calcular la longitud de una elipse aparece la integral

$$\int_0^{\frac{\pi}{2}} \sqrt{1 - k^2 \sin^2 t} dt$$

donde el parámetro k depende de la excentricidad de la elipse. En el caso más simple en que la excentricidad es 1 (la elipse es una circunferencia), el parámetro k toma el valor 1 y la integral se calcula fácilmente. En el caso más general en que $0 < k < 1$ se ha demostrado que no existen funciones elementales que puedan ser primitivas. Este tipo de integral se llama “elíptica”.

Ejemplo 2

Las variables aleatorias se caracterizan por su función de densidad probabilística. De todas estas funciones de densidad, la más importante es la distribución Normal, que aparece de una u otra manera en casi todos los fenómenos naturales de carácter aleatorio. La probabilidad de que la variable con distribución normal tome un valor entre 0 y x viene dada por la integral:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}t^2} dt$$

para la cual está demostrado que no existe una primitiva elemental.

Ejemplo 3

En el patrón de transmisión de algunas antenas aparece la función elemental

$$\frac{\sin \theta}{\theta}$$

donde θ representa una dirección en el espacio. Para calcular la potencia radiada por la antena en un sector, se requiere calcular la integral:

$$\int_{\theta_1}^{\theta_2} \frac{\sin \theta}{\theta} d\theta$$

y, también en este caso, es imposible hallar una primitiva elemental del integrando. ■

En otros casos la regla de Barrow no puede usarse porque no se conoce la expresión analítica del integrando, sino valores obtenidos a través de mediciones.

Ejemplo 4

El gasto (en m^3/s) de agua que circula por un río es el volumen de agua que atraviesa una sección cualquiera del río en una unidad de tiempo. Suponiendo, para simplificar el problema, que en todos los puntos de la sección seleccionada la velocidad es la misma (es decir, tomando una velocidad promedio v_m), el gasto se puede calcular como:

$$Q = v_m A$$

donde A es el área de la sección seleccionada. En la figura 1 se muestra una hipotética sección del río y se ha introducido un sistema de referencia con el eje vertical dirigido hacia abajo y el eje horizontal en la superficie del agua y a lo largo de la sección y con el origen en una de las orillas. Se ha representado por $h(x)$ la función que describe la profundidad del río para cada x . Es claro que el área de la sección será:

$$A = \int_0^b h(x) dx$$

pero la función $h(x)$ no tiene una expresión analítica conocida a la que se pueda encontrar una primitiva. La regla de Barrow, de nuevo, no se puede usar.

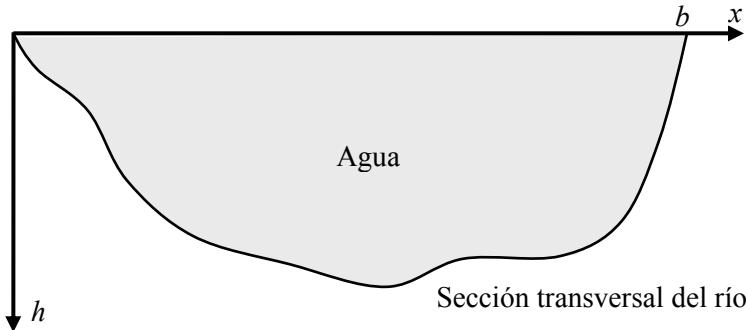


Figura 1

5.2 El método de los trapecios

Aunque el método más elemental para calcular la integral definida de una función consiste en utilizar la misma definición de la integral definida pero prescindiendo del proceso de límite, es decir:

$$\int_a^b f(x)dx \approx \sum_{i=1}^n f(c_i)\Delta x_i \quad (1)$$

el resultado que se obtiene es muy inexacto, a menos que se tome un número n de subintervalos muy grande. El método de los trapecios que se expone a continuación requiere una cantidad de cálculos semejante a la expresión (1) y produce resultados mucho mejores.

En esta sección y en todo lo que sigue de este capítulo, se supondrá que $f(x)$ es una función continua en $[a, b]$. Aunque algunas de las expresiones que se deducirán pueden usarse aun sin el requisito de la continuidad de $f(x)$, las fórmulas para el cálculo de los errores siempre requerirán la continuidad, e incluso la derivabilidad de $f(x)$, de modo que no vale la pena contemplar el caso de funciones discontinuas. Por otra parte, en la mayoría de los casos de interés práctico el integrando es continuo o la integral se puede descomponer en una suma de integrales con el integrando continuo en el intervalo de integración.

El método de los trapecios se basa en la idea de dividir el intervalo de integración en n subintervalos de amplitud h mediante un conjunto de puntos $\{a = x_0, x_1, x_2, \dots, x_n = b\}$ y descomponer la integral en n integrales, cada una de las cuales posee un intervalo de integración pequeño de longitud h , como se ilustra en la figura 1. El parámetro h se denomina *paso* y juega un papel importante en la exactitud del resultado obtenido.

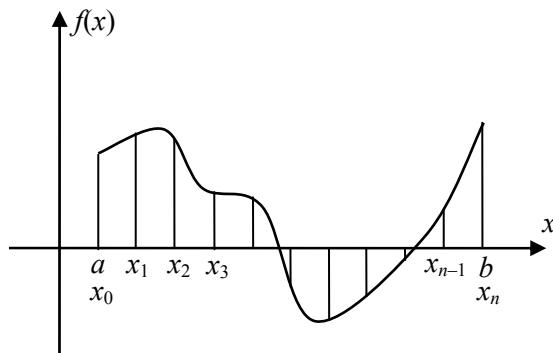


Figura 1

Es decir:

$$\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \dots + \int_{x_{n-1}}^{x_n} f(x)dx \quad (2)$$

donde: $x_i - x_{i-1} = h$ para $i = 1, 2, \dots, n$ y $h = \frac{b-a}{n}$

Ahora, cada una de las n integrales resultantes se aproximan sustituyendo el integrando $f(x)$ por un polinomio interpolador de grado 1, es decir, un segmento de recta, como se ilustra en la figura 2.

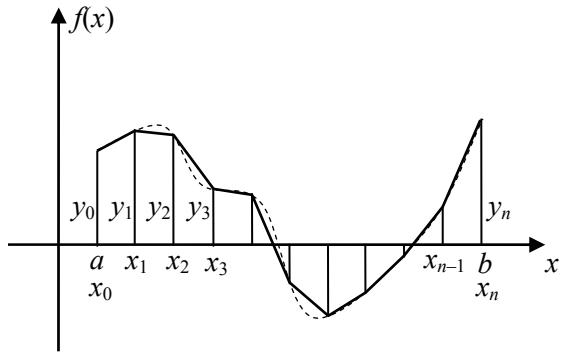


Figura 2

Nótese que la región entre la gráfica de $f(x)$ y el eje x se aproximará mediante un conjunto de n regiones, que tienen, en general, forma de trapecios, es decir, cuadriláteros con dos de sus lados paralelos, lo cual da nombre al método. Para simplificar la notación, se llamará: $y_i = f(x_i)$ para $i = 0, 1, 2, \dots, n$, como se muestra también en la figura 2.

Trabajando en el primero de los trapecios, que corresponde al sub-intervalo $[x_0, x_1]$, la ecuación de la recta de interpolación que le corresponde será:

$$y = y_0 + \frac{y_1 - y_0}{x_1 - x_0} (x - x_0)$$

Es decir:

$$y = y_0 + \frac{y_1 - y_0}{h} (x - x_0)$$

Sustituyendo el integrando $f(x)$ por el polinomio $p_1(x) = y_0 + \frac{y_1 - y_0}{h} (x - x_0)$, la primera de las n integrales queda aproximada como:

$$\begin{aligned} \int_{x_0}^{x_1} f(x) dx &\approx \int_{x_0}^{x_1} p_1(x) dx = \int_{x_0}^{x_1} \left(y_0 + \frac{y_1 - y_0}{h} (x - x_0) \right) dx \\ &= y_0 \int_{x_0}^{x_1} dx + \frac{y_1 - y_0}{h} \int_{x_0}^{x_1} (x - x_0) dx \\ &= y_0 x \Big|_{x_0}^{x_1} + \frac{y_1 - y_0}{2h} (x - x_0)^2 \Big|_{x_0}^{x_1} \\ &= y_0 (x_1 - x_0) + \frac{y_1 - y_0}{2h} (x_1 - x_0)^2 \\ &= y_0 h + \frac{y_1 - y_0}{2h} h^2 \end{aligned}$$

$$\int_{x_0}^{x_1} f(x)dx \approx \int_{x_0}^{x_1} p_1(x)dx = h \left(\frac{y_0 + y_1}{2} \right) \quad (3)$$

Nótese que el resultado obtenido coincide con la conocida fórmula para el área de un trapecio (semisuma de las bases por la altura) cuando y_0 y y_1 son positivas, pero la expresión (3) tiene un carácter más general, ya que es válida para cualesquiera valores (positivos, negativos o cero) que tomen y_0 y y_1 .

Es evidente que aplicando el mismo procedimiento para la segunda integral se obtendría:

$$\int_{x_1}^{x_2} f(x)dx \approx \int_{x_1}^{x_2} p_2(x)dx = h \left(\frac{y_1 + y_2}{2} \right)$$

y similarmente ocurriría para cada una de las n integrales de la expresión (2). Por tanto, de ella resulta:

$$\int_a^b f(x)dx \approx h \left(\frac{y_0 + y_1}{2} \right) + h \left(\frac{y_1 + y_2}{2} \right) + \dots + h \left(\frac{y_{n-1} + y_n}{2} \right)$$

Simplificando: $\int_a^b f(x)dx \approx h \left(\frac{1}{2} y_0 + y_1 + y_2 + \dots + y_{n-1} + \frac{1}{2} y_n \right) \quad (4)$

que es la forma usual de expresar el método de los trapecios para el cálculo numérico aproximado de integrales definidas.

Ejemplo 1

Calcule aproximadamente la integral

$$\int_0^\pi \sin x dx$$

mediante el método de los trapecios, utilizando 10, 20 y 40 sub-intervalos. Compare con el resultado exacto, que es 2.

Solución:

Tomando $n = 10$ sub-intervalos, resulta $h = \frac{\pi - 0}{10} = 0,314159$. La tabla 1 muestra los valores de x_i y $y_i = \sin x_i$ para $i = 0, 1, \dots, 10$. A partir de ahí se obtiene:

$$\int_0^\pi \sin x dx \approx h \left(\frac{1}{2} y_0 + y_1 + y_2 + \dots + y_9 + \frac{1}{2} y_{10} \right) = 1,983524$$

Similarmente, para $n = 20$, $h = 0,157080$ y se obtiene la aproximación:

$$\int_0^\pi \sin x dx \approx h \left(\frac{1}{2} y_0 + y_1 + y_2 + \dots + y_{19} + \frac{1}{2} y_{20} \right) = 1,995886$$

y, para $n = 40$, $h = 0,078540$ y resulta:

$$\int_0^{\pi} \sin x \, dx \approx h \left(\frac{1}{2} y_0 + y_1 + y_2 + \cdots + y_{39} + \frac{1}{2} y_{40} \right) = 1,998972$$

i	x_i	$y_i = \sin x_i$
0	0,000000	0,000000
1	0,314159	0,309017
2	0,628319	0,587785
3	0,942478	0,809017
4	1,256637	0,951057
5	1,570796	1,000000
6	1,884956	0,951057
7	2,199115	0,809017
8	2,513274	0,587785
9	2,827433	0,309017
10	3,141593	0,000000

Tabla 1

Como era de esperar, a medida que el paso h toma valores más pequeños se obtienen resultados más exactos, así los errores respectivos para $n = 10, 20$ y 40 fueron: $0,016476$, $0,004114$ y $0,001028$. Observe el hecho, que no es casual, de que en la medida en que el paso se reduce a la mitad, el error se reduce aproximadamente a la cuarta parte.

Algoritmo en seudo código

El algoritmo que sigue calcula aproximadamente la integral definida de la función continua $f(x)$ en el intervalo $[a, b]$ mediante el método de los trapecios tomando n sub intervalos. Se suponen conocidos la función $f(x)$, el intervalo $[a, b]$ y el número entero $n > 1$. El resultado es, aproximadamente,

$$\int_a^b f(x) dx$$

$$h := \frac{b-a}{n}$$

$$Suma := \frac{1}{2} [f(a) + f(b)]$$

for $i = 1$ **to** $n - 1$

$$x = a + ih$$

$$Suma := Suma + f(x)$$

end

$$Integral := h \cdot Suma$$

El resultado aproximado es $Integral$

Terminar

El error de truncamiento en el método de los trapecios

Por supuesto que la fórmula (4) de los trapecios no tiene gran valor si no se puede acotar el error que con ella se comete. A continuación se hallará una expresión que en muchos casos permite determinar un error absoluto máximo de truncamiento. Además del error de truncamiento, la fórmula (4) introduce un error de redondeo debido a los errores de este tipo que contienen los sumandos y_i , del cual se hablará posteriormente, pero que desde ahora puede adelantarse que suele ser insignificante.

El error de truncamiento total se hallará como la suma de los errores de truncamiento que aparecen en cada uno de los subintervalos en que se divide $[a, b]$. Considérese, en general, el subintervalo $[x_i, x_{i+1}]$ ($i = 0, 1, 2, \dots, n-1$) y sea $p(x)$ el polinomio de interpolación de primer grado mediante el cual se aproxima a $f(x)$ en dicho intervalo. Sea $r(x)$ el error de interpolación correspondiente, de modo que:

$$f(x) = p(x) + r(x) \quad \text{para } x_i \leq x \leq x_{i+1}$$

Integrando en cada miembro y trasponiendo, se obtiene:

$$\int_{x_i}^{x_{i+1}} f(x) dx - \int_{x_i}^{x_{i+1}} p(x) dx = \int_{x_i}^{x_{i+1}} r(x) dx$$

Es decir, que el error de truncamiento que se introduce en el intervalo $[x_i, x_{i+1}]$, puede obtenerse integrando el error de interpolación en ese intervalo. Se llamará:

$$R_i = \int_{x_i}^{x_{i+1}} r(x) dx \quad (5)$$

de manera que el error de truncamiento R del método de los trapecios pueda ser calculado como:

$$R = \sum_{i=0}^{n-1} R_i \quad (6)$$

El error de interpolación (vea la sección 4.2, fórmula (7)) para un polinomio de primer grado correspondiente a los nodos x_i y x_{i+1} viene dado por:

$$r(x) = \frac{f''(\alpha)}{2!} (x - x_i)(x - x_{i+1})$$

donde α es algún número del intervalo $[x_i, x_{i+1}]$ cuyo valor depende de x . De acuerdo con (5):

$$R_i = \frac{1}{2} \int_{x_i}^{x_{i+1}} f''(\alpha)(x - x_i)(x - x_{i+1}) dx \quad (7)$$

Considérese el caso en que $f''(x)$ es continua en $[a, b]$, entonces en el intervalo de integración $[x_i, x_{i+1}]$ ella tomará un valor mínimo m y uno máximo M , por lo tanto se puede asegurar que:

$$m \leq f''(\alpha) \leq M \quad (8)$$

Por otra parte, como en $[x_i, x_{i+1}]$ el factor $(x - x_i)$ es mayor o igual que cero, mientras que $(x - x_{i+1})$ es menor o igual que cero, el producto $(x - x_i)(x - x_{i+1})$ es negativo o cero, de modo que si cada miembro de (8) se multiplica por este producto, se obtiene:

$$M(x - x_i)(x - x_{i+1}) \leq f''(\alpha)(x - x_i)(x - x_{i+1}) \leq m(x - x_i)(x - x_{i+1}) \quad \text{para } x_i \leq x \leq x_{i+1}$$

De aquí resulta que:

$$M \int_{x_i}^{x_{i+1}} (x - x_i)(x - x_{i+1}) dx \leq \int_{x_i}^{x_{i+1}} f''(\alpha)(x - x_i)(x - x_{i+1}) dx \leq m \int_{x_i}^{x_{i+1}} (x - x_i)(x - x_{i+1}) dx$$

Esto significa que, para algún valor μ comprendido entre m y M se cumple que:

$$\int_{x_i}^{x_{i+1}} f''(\alpha)(x - x_i)(x - x_{i+1}) dx = \mu \int_{x_i}^{x_{i+1}} (x - x_i)(x - x_{i+1}) dx$$

Por ser $f''(x)$ continua en $[x_i, x_{i+1}]$, ella tomará en algún punto de ese intervalo el valor μ , esto es, existe un c_i en el intervalo $[x_i, x_{i+1}]$ para el cual:

$$f''(c_i) = \mu$$

y, por tanto:

$$\int_{x_i}^{x_{i+1}} f''(\alpha)(x - x_i)(x - x_{i+1}) dx = f''(c_i) \int_{x_i}^{x_{i+1}} (x - x_i)(x - x_{i+1}) dx$$

y la fórmula (7) queda:

$$R_i = \frac{1}{2} f''(c_i) \int_{x_i}^{x_{i+1}} (x - x_i)(x - x_{i+1}) dx$$

para algún c_i en el intervalo $[x_i, x_{i+1}]$. Calculando la integral y simplificando, tomando en cuenta que $x_{i+1} - x_i = h$, se obtiene:

$$R_i = \frac{1}{2} f''(c_i) \left(-\frac{h^3}{6} \right)$$

Esto es:

$$R_i = -\frac{h^3 f''(c_i)}{12}$$

Observe que el signo menos que aparece tiene un significado interesante: como era de esperar, cuando la segunda derivada es positiva en el intervalo $[x_i, x_{i+1}]$ (la gráfica de la función es cóncava hacia arriba) la recta de interpolación está por encima de $f(x)$ y el error es negativo, es decir por exceso, mientras que cuando la gráfica es cóncava hacia abajo (segunda derivada negativa) el error de integración es positivo, es decir, por defecto.

Volviendo a la expresión (6) se puede calcular ahora el error de truncamiento total como:

$$R = \sum_{i=0}^{n-1} R_i = \sum_{i=0}^{n-1} -\frac{h^3 f''(c_i)}{12}$$

Es decir:

$$R = -\frac{h^3}{12} \sum_{i=0}^{n-1} f''(c_i) \quad (9)$$

Con vista a obtener una fórmula más sencilla, conviene escribir esta ecuación como:

$$R = -\frac{nh^3}{12} \left[\frac{1}{n} \sum_{i=0}^{n-1} f''(c_i) \right]$$

pues ahora el término entre corchetes es el promedio de n valores de $f''(x)$, el cual tiene que estar comprendido entre los valores mínimo y máximo de $f''(x)$ en $[a, b]$ y, como se trata de una función continua, $f''(x)$ tomará este valor en algún punto del intervalo $[a, b]$. Llámese c a este número y se puede entonces decir que:

$$R = -\frac{nh^3}{12} f''(c)$$

Por último, como $n = \frac{b-a}{h}$, queda:

$$R = -\frac{b-a}{12} h^2 f''(c) \quad (10)$$

para algún c en el intervalo $[a, b]$.

En resumen:

Si en el intervalo $[a, b]$ $f(x)$ es continua y posee primera y segunda derivadas continuas, entonces existe en $[a, b]$ al menos un número c tal que, el error de truncamiento de la fórmula de los trapecios para la integral de $f(x)$ en $[a, b]$ viene dado por:

$$R = -\frac{b-a}{12} h^2 f''(c)$$

La fórmula anterior, además de su valor teórico, permite en muchos casos acotar el error cometido en el método de los trapecios. Para ello se necesita hallar una cota superior del valor absoluto de la segunda derivada en el intervalo de integración, lo cual a veces no es difícil de encontrar.

Ejemplo 2

Al calcular la integral

$$\int_0^\pi \sin x \, dx$$

por el método de los trapecios en el ejemplo 1, tomando $n = 10$ sub-intervalos, se obtuvo el resultado aproximado: 1,983524. Halle una cota del error que se ha cometido.

Solución:

La función $f(x) = \sin x$ satisface las condiciones de continuidad y derivabilidad exigidas para aplicar la fórmula (10). En este caso se tiene:

$$f''(x) = -\sin x$$

así que:

$$R = -\frac{b-a}{12} h^2 f''(c) = \frac{\pi-0}{12} \left(\frac{\pi}{10}\right)^2 \sin c$$

para algún c en el intervalo de integración $[0, \pi]$. Aunque no se conoce el número c el valor de $\sin c$ está en el intervalo $[0, 1]$, por tanto:

$$0 \leq R \leq \frac{\pi^3}{1200}$$

O sea:

$$0 \leq R \leq 0,0258$$

El error realmente fue de 0,016476 por lo cual, la cota es válida aunque un poco conservadora.

Una fórmula asintótica para el error de truncamiento

Es fácil obtener una fórmula para hallar una aproximación del error de truncamiento del método de los trapecios, que posee la ventaja de dar una estimación del error no tan conservadora como las cotas que se obtienen con la fórmula (10) y que, además, no requiere calcular la segunda derivada de $f(x)$. La fórmula se llama asintótica porque, a medida que h tiende hacia cero (o n hacia infinito) el resultado que ella brinda tiende hacia el verdadero error cometido, lo cual significa que el valor de R que ella estima es tanto mejor cuanto más pequeño sea h .

De acuerdo con la fórmula (9):

$$R = -\frac{h^3}{12} \sum_{i=0}^{n-1} f''(c_i)$$

la cual puede ser escrita como:

$$R = -\frac{h^2}{12} \sum_{i=0}^{n-1} f''(c_i)h$$

$$\text{estos es: } -\frac{12R}{h^2} = \sum_{i=0}^{n-1} f''(c_i)h$$

Tomando límites cuando h tiende hacia cero, se obtiene:

$$\lim_{h \rightarrow 0} \left(-\frac{12R}{h^2} \right) = \lim_{h \rightarrow 0} \sum_{i=0}^{n-1} f''(c_i)h$$

Suponiendo que $f''(x)$ es continua en el intervalo $[a, b]$ y, por lo tanto, integrable, el límite de la derecha existe y coincide con la integral definida de $f''(x)$ en el intervalo $[a, b]$. Es decir:

$$\lim_{h \rightarrow 0} \left(-\frac{12R}{h^2} \right) = \int_a^b f''(x) dx$$

y como se trata de un integrando continuo se puede utilizar la regla de Barrow. La ecuación queda:

$$\lim_{h \rightarrow 0} \left(-\frac{12R}{h^2} \right) = f'(b) - f'(a)$$

Si se prescinde del límite se obtiene la fórmula:

$$R \approx -\frac{f'(b) - f'(a)}{12} h^2 \quad (11)$$

donde el signo “ \approx ” se entiende en sentido asintótico, es decir, la fórmula se va haciendo exacta en la medida en que h tiende hacia cero.

Ejemplo 3

Halle un estimado del error de truncamiento cometido al calcular:

$$\int_0^\pi \sin x dx$$

por el método de los trapecios (ejemplo 1) con $n = 10$ sub-intervalos.

Solución:

Como $f(x) = \sin x$, se obtiene $f'(x) = \cos x$ y aplicando la fórmula (11):

$$R \approx -\frac{f'(b) - f'(a)}{12} h^2 = -\frac{\cos \pi - \cos 0}{12} \left(\frac{\pi}{10} \right)^2$$

$$R \approx \frac{\pi^2}{600} = 0,01645$$

Compárese con el valor exacto del error de truncamiento que fue de 0,016476.

Estimación del error de truncamiento por doble cálculo

La fórmula (11) permite apreciar claramente un hecho que ya se había notado en el ejemplo 1. Como

$$R \approx -\frac{f'(b) - f'(a)}{12} h^2$$

el error de truncamiento es aproximadamente una función cuadrática del paso h , lo cual significa, por ejemplo, que al disminuir h a la mitad, el error de truncamiento disminuirá aproximadamente a la cuarta parte. Este sencillo hecho, permitirá obtener una estimación de R sin necesidad de obtener la derivada de $f(x)$, a cambio de utilizar dos veces el método de los trapecios.

Como este razonamiento se puede generalizar fácilmente, se supondrá que el error de truncamiento al calcular la integral con paso h viene dado por:

$$R_h = Ch^p \quad (12)$$

donde, C es una constante (no depende de h) y la constante p vale 2 para el método de los trapecios. Observe que (12) es una fórmula exacta mientras que (11) es solo una aproximación; esto quiere decir que, aunque se utilizará el signo igual, solo se aspira a obtener fórmulas aproximadas.

Suponga que utilizando el método de los trapecios, se calcula una integral definida dos veces, una de ellas con un paso h y la otra con paso $2h$. Sea:

I_h : Resultado obtenido al calcular la integral con paso h

I_{2h} : Resultado obtenido al calcular la integral con paso $2h$

El error de truncamiento de I_h viene dado por $R_h = Ch^p$.

El error de truncamiento de I_{2h} será $R_{2h} = C(2h)^p$.

De aquí resulta que: $R_{2h} = C(2h)^p = 2^p Ch^p = 2^p R_h$ (13)

Por otra parte, $I_h + R_h = I_{2h} + R_{2h}$

ya que ambas expresiones dan el valor exacto de la integral. Sustituyendo en esta igualdad el resultado (13), queda:

$$I_h + R_h = I_{2h} + 2^p R_h$$

y puede despejarse R_h : $R_h \approx \frac{I_h - I_{2h}}{2^p - 1}$ (14)

donde se ha escrito el signo “ \approx ” tomando en cuenta que la fórmula (12) solo se cumple en sentido asintótico, es decir, cuando h tiende hacia cero. Como en el método de los trapecios $p = 2$, se tiene, como caso particular:

$$R_h \approx \frac{I_h - I_{2h}}{3} \quad (15)$$

Ejemplo 4

En el ejemplo 1 se calculó la integral

$$\int_0^\pi \sin x \, dx$$

tomando $n = 10, 20$ y 40 . Se obtuvo los resultados aproximados: 1,983524, 1,995886 y 1,998972 respectivamente. Utilice estos resultados para estimar el error de truncamiento de los dos últimos.

Solución:

Para $n = 20$, se obtuvo: $I_h = 1,995886$

Para $n = 10$, se obtuvo: $I_{2h} = 1,983524$.

El error de truncamiento de I_h viene dado aproximadamente por (15):

$$R_h \approx \frac{I_h - I_{2h}}{3} = \frac{1,995886 - 1,983524}{3} = 0,0041$$

Es decir, al tomar 20 sub-intervalos el error de truncamiento es aproximadamente de 0,0041. Nótese que el verdadero error de este resultado es de 0,004114.

Para $n = 40$, se obtuvo: $I_h = 1,998972$

Para $n = 20$, se obtuvo: $I_{2h} = 1,995886$

El error de truncamiento de I_h es aproximadamente:

$$R_h \approx \frac{I_h - I_{2h}}{3} = \frac{1,998972 - 1,995886}{3} = 0,0010$$

Esto es, tomando $n = 40$ sub-intervalos el error de truncamiento resulta aproximadamente de 0,0010. El verdadero error de truncamiento es en este caso de 0,001028.

Ejemplo 5

Calcule la integral $\int_1^2 e^{x^2} dx$ con cuatro cifras decimales exactas.

Solución:

En los cálculos que siguen se ha utilizado un programa basado en el seudo código del método de los trapecios, para efectuar cada una de las integrales. El error de truncamiento se ha estimado mediante doble cálculo utilizando la fórmula (15). El proceso se comenzó arbitrariamente con $n = 10$ y a partir de ahí se estimaron los errores comparando con el resultado anterior. La tabla 2 muestra todos los resultados. El proceso iterativo fue detenido tan pronto como el error estimado fue menor que 0,00005. Con cuatro cifras decimales exactas, el valor de la integral es 14.990019.

n	Integral	Error estimado
10	15.166784	
20	15.034301	0.044161
40	15.001065	0.011079
80	14.992749	0.002772
160	14.990669	0.000693
320	14.990149	0.000173
640	14.990019	0.000043

Tabla 2

El error de redondeo en la fórmula de los trapecios

En la fórmula de los trapecios

$$\int_a^b f(x)dx \approx h\left(\frac{1}{2}y_0 + y_1 + y_2 + \dots + y_{n-1} + \frac{1}{2}y_n\right)$$

se necesita calcular la suma de una gran cantidad de datos que contienen errores de redondeo. Puede entonces pensarse que se introducirá un error resultante muy grande cuando se sumen todos estos errores de redondeo. Realmente no es así. Para simplificar, supóngase que cada uno de los datos posee un mismo error absoluto máximo δ (que se supone causado por el redondeo, pero que podría ser motivado por cualquier otra factor, por ejemplo, por la medición u observación). Sea:

$$S = h \left(\frac{1}{2} y_0 + y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2} y_n \right)$$

Se trata de hallar $E_m(S)$. Para ello es mejor escribir h como $\frac{b-a}{n}$, es decir:

$$S = \frac{b-a}{n} \left(\frac{1}{2} y_0 + y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2} y_n \right)$$

Como, a , b y n se suponen exentos de error, resulta:

$$E_m(S) = \frac{b-a}{n} E_m \left(\frac{1}{2} y_0 + y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2} y_n \right)$$

$$= \frac{b-a}{n} \left(\frac{1}{2} \delta + \delta + \delta + \cdots + \delta + \frac{1}{2} \delta \right)$$

$$= \frac{b-a}{n} (n\delta)$$

$$E_m(S) = (b-a)\delta$$

Es decir, el error total en el resultado causado por errores en los datos, es solamente el error absoluto máximo de cada dato individual multiplicado por la amplitud del intervalo de integración y es independiente de la cantidad de términos que se tomen en la fórmula. Como se trata de un error insignificante en comparación con el error de truncamiento, este tipo de error no suele tomarse en consideración en el cálculo numérico de integrales.

En los demás métodos de integración numérica que serán estudiados en el resto de este capítulo sucede algo similar, así que en ellos se prescindirá del estudio del error por redondeo.

5.3 El método de Simpson

La fórmula de los trapecios se obtiene aproximando el integrando por un conjunto de polinomios de interpolación de primer grado. Esta idea puede ser mejorada si la aproximación se realiza con polinomios interpoladores de grado dos. Considérese entonces la integral:

$$\int_a^b f(x) dx$$

donde se supondrá que $f(x)$ es continua en $[a, b]$. Sea el intervalo de integración dividido en un número par de intervalos de igual amplitud h mediante los puntos $\{a = x_0, x_1, x_2, \dots, x_n = b\}$. Las figuras 1 y 2 muestran la idea geométrica del método de Simpson: en la figura 1 se aprecia que la región comprendida entre el eje horizontal y la gráfica de $f(x)$ se ha dividido en franjas verticales mediante rectas que determinan los puntos de la partición de $[a, b]$. En la figura 2 se ha sustituido la función $f(x)$ por un conjunto de polinomios de grado 2, es decir, parábolas de eje vertical. Posteriormente, se calculará la suma de las integrales de estas funciones cuadráticas.

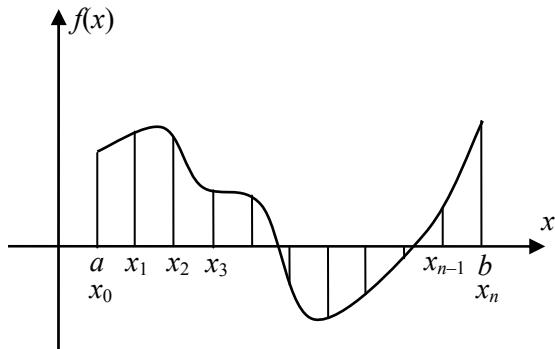


Figura 1

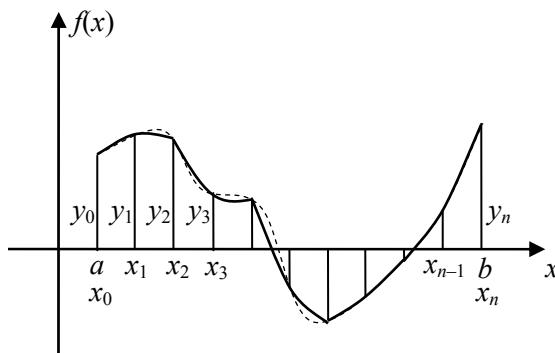


Figura 2

Al igual que en el método de los trapecios, la notación se simplificará llamando $y_i = f(x_i)$ para $i = 0, 1, 2, \dots, n$.

La obtención de la fórmula de Simpson requiere conocer el valor de la integral de una función cuadrática determinada por tres puntos $(x_0, y_0), (x_1, y_1)$ y (x_2, y_2) donde $x_1 - x_0 = h$ y $x_2 - x_1 = h$ como muestra la figura 3. El resultado, naturalmente, no cambiará si toda la figura se traslada de modo que $x_1 = 0$, como muestra la figura 4, y de esta manera el trabajo algebraico se simplifica.

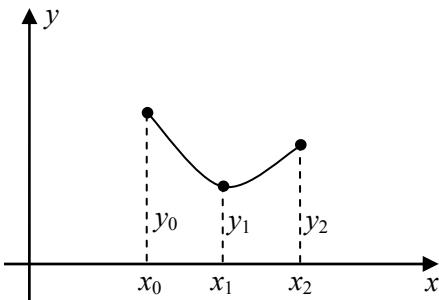


Figura 3

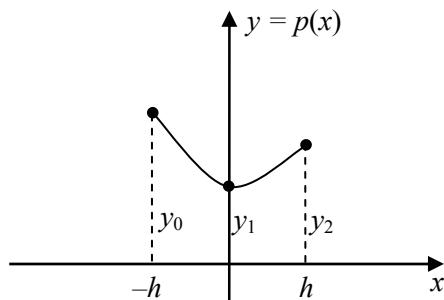


Figura 4

Sea entonces $p(x)$ el polinomio de segundo grado cuya gráfica aparece en la figura 4, determinado por los tres puntos $(-h, y_0)$, $(0, y_1)$ y (h, y_2) , cuya ecuación será de la forma:

$$p(x) = ax^2 + bx + c \quad (1)$$

Su integral vendrá dada por:

$$\begin{aligned} \int_{-h}^h p(x) dx &= \int_{-h}^h (ax^2 + bx + c) dx = \left(\frac{1}{3}ax^3 + \frac{1}{2}bx^2 + cx \right) \Big|_{-h}^h = \\ &= \frac{2}{3}ah^3 + 2ch \end{aligned} \quad (2)$$

Ahora bien, de acuerdo con (1), se cumple que:

$$p(-h) = ah^2 - bh + c = y_0 \quad (3)$$

$$p(0) = c = y_1 \quad (4)$$

$$y \quad p(h) = ah^2 + bh + c = y_2 \quad (5)$$

$$\text{Al sumar (5) con (3) queda: } 2ah^2 + 2c = y_0 + y_2$$

$$\text{y como, según (4), } c = y_1: \quad 2ah^2 + 2y_1 = y_0 + y_2$$

$$\text{Esto es: } a = \frac{y_0 - 2y_1 + y_2}{2h^2}$$

$$\text{Sustituyendo en (2): } \int_{-h}^h p(x) dx = \frac{2}{3} \left(\frac{y_0 - 2y_1 + y_2}{2h^2} \right) h^3 + 2y_1 h$$

Simplificando:

$$\int_{-h}^h p(x)dx = \frac{1}{3}h(y_0 + 4y_1 + y_2) \quad (6)$$

Aplicando ahora a la integral de $f(x)$ en $[a, b]$ la propiedad de aditividad respecto al intervalo de integración resulta:

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + \cdots + \int_{x_{n-2}}^{x_n} f(x)dx$$

Sustituyendo en cada integral la función $f(x)$ por el polinomio interpolador de segundo grado que corresponde a esos nodos, la integral resultante se puede calcular mediante la fórmula (6). Se obtiene la aproximación:

$$\int_a^b f(x)dx \approx \frac{1}{3}h(y_0 + 4y_1 + y_2) + \frac{1}{3}h(y_2 + 4y_3 + y_4) + \cdots + \frac{1}{3}h(y_{n-2} + 4y_{n-1} + y_n)$$

Que se simplifica como:

$$\int_a^b f(x)dx \approx \frac{1}{3}h(y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \cdots + 2y_{n-2} + 4y_{n-1} + y_n)$$

o, más brevemente:

$$\int_a^b f(x)dx \approx \frac{1}{3}h(E + 4I + 2P)$$

donde:

$E = y_0 + y_n$ (Suma de las ordenadas en los extremos)

$I = y_1 + y_3 + \dots + y_{n-1}$ (Suma de las ordenadas de índice impar)

$P = y_2 + y_4 + \dots + y_{n-2}$ (Suma de las ordenadas de índice par)

Ejemplo 1

Calcule aproximadamente la integral

$$\int_0^\pi \sin x dx$$

mediante el método de Simpson, utilizando 10, 20 y 40 sub-intervalos. Compare con el resultado exacto, que es 2 y con los resultados obtenidos mediante el método de los trapecios en el ejemplo 1 de la sección 5.2.

Solución:

Tomando $n = 10$ sub-intervalos, resulta $h = \frac{\pi - 0}{10} = 0,314159$. La tabla 1 muestra los valores de x_i y $y_i = \sin x_i$ para $i = 0, 1, \dots, 10$. Para facilitar las sumas cuando se trabaja manualmente, se suelen colocar los valores de la función en tres columnas.

i	x_i	$y_i = \sin x_i$
0	0,000000	0,000000
1	0,314159	0,309017
2	0,628319	0,587785
3	0,942478	0,809017
4	1,256637	0,951057
5	1,570796	1,000000
6	1,884956	0,951057
7	2,199115	0,809017
8	2,513274	0,587785
9	2,827433	0,309017
10	3,141593	0,000000

$$E = 0,000000 \quad I = 3,236068 \quad P = 3,077684$$

Tabla 1

El resultado será:

$$\int_a^b f(x)dx \approx \frac{1}{3}h(E + 4I + 2P) = \frac{1}{3}0,314159[0 + 4(3.236068) + 2(3,077684)] =$$

$$\int_a^b f(x)dx \approx 2,000110$$

Nótese que el error de truncamiento es solo de 0,00011. Con este mismo paso h , el resultado del método de los trapecios contenía un error de 0,016476, unas 150 veces mayor.

Con $n = 20$ se obtiene, de forma similar:

$$\int_a^b f(x)dx \approx 2,00000678$$

cuyo error de truncamiento: 0,00000678 es 600 veces menor que el de 0,004114 logrado con el método de los trapecios.

Para $n = 40$ el resultado del método de Simpson es:

$$\int_a^b f(x)dx \approx 2,00000042$$

con un error de truncamiento de solo 0,00000042 unas 2400 veces menor que el de 0,001028 que se logra con el método de los trapecios para este mismo paso.

Algoritmo en seudo código

El algoritmo que sigue calcula aproximadamente la integral definida de la función continua $f(x)$ en el intervalo $[a, b]$ mediante el método de Simpson tomando un número par n de sub intervalos. Se supone conocidos la función $f(x)$, el intervalo $[a, b]$ y el número entero par $n \geq 2$. El resultado es, aproximadamente,

$$\int_a^b f(x)dx$$

```


$$h := \frac{b-a}{n}$$


$$E := f(a) + f(b) \quad \{ \text{Se calcula la suma de las ordenadas extremas} \}$$


$$I := 0 \quad \{ \text{Se calcula la suma de las ordenadas de índice impar} \}$$


$$i := 1$$

do while  $i < n$ 
     $x = a + ih$ 
     $I := I + f(x)$ 
     $i := i + 2$ 
end
 $P := 0 \quad \{ \text{Se calcula la suma de las ordenadas de índice par, sin incluir las extremas} \}$ 
 $i := 2$ 
do while  $i < n$ 
     $x = a + ih$ 
     $P := P + f(x)$ 
     $i := i + 2$ 
end

$$\text{Integral} := \frac{1}{3} h \cdot (E + 4I + 2P)$$

El resultado aproximado es Integral
Terminar

```

Error de truncamiento en el método de Simpson

De forma semejante a como se procedió en el método de los trapecios, aunque con dificultades adicionales, se puede demostrar (en las lecturas recomendadas se indica donde encontrar esta demostración) que el error de truncamiento R_i que se produce en el método de Simpson en el subintervalo $[x_{2i}, x_{2i+2}]$ ($i = 0, 1, 2, \dots, \frac{1}{2}n - 1$), viene dado por:

$$R_i = -\frac{h^5 f^{(4)}(c_i)}{90}$$

donde c_i es algún número del intervalo $[x_{2i}, x_{2i+2}]$. Por tanto en todo el intervalo $[a, b]$ de integración, el error de truncamiento es:

$$R = -\frac{h^5}{90} \sum_{i=0}^{\frac{n}{2}-1} f^{(4)}(c_i) \tag{7}$$

Que puede ser escrito como:

$$R = -\frac{nh^5}{180} \left[\frac{2}{n} \sum_{i=0}^{\frac{n}{2}-1} f^{(4)}(c_i) \right]$$

Como la cantidad entre corchetes es un promedio de valores de la cuarta derivada, el mismo se encuentra entre el mínimo y el máximo valor que toma esa función en $[a, b]$. Suponiendo que $f^{(4)}(x)$ es continua en $[a, b]$ se puede afirmar que, al menos en un punto c de ese intervalo, la cuarta derivada toma este valor promedio. Resulta entonces que:

$$R = -\frac{nh^5}{180} f^{(4)}(c)$$

pero como $n = \frac{b-a}{h}$, se obtiene: $R = -\frac{b-a}{180} h^4 f^{(4)}(c)$

Resumiendo:

Si en el intervalo $[a, b]$ $f(x)$ es continua y posee hasta la cuarta derivada continuas, entonces existe en $[a, b]$ al menos un número c tal que, el error de truncamiento de la fórmula de Simpson para la integral de $f(x)$ en $[a, b]$ viene dado por:

$$R = -\frac{b-a}{180} h^4 f^{(4)}(c)$$

Esta fórmula, debida a Peano, requiere conocer una cota superior de la cuarta derivada para poder acotar el error de truncamiento, por lo cual su valor práctico es algo limitado. Su valor teórico fundamental reside en que ella pone claramente de manifiesto la dependencia del error con la potencia h^4 , lo cual indica que, a medida que h tome valores más pequeños, el error de truncamiento tenderá a cero con rapidez. Nótese que la presencia de la cuarta derivada en la expresión de R , indica que el método de Simpson es exacto si $f(x)$ es un polinomio de grado 2 (lo cual era de esperar, puesto que se aproxima con parábolas) y también para polinomios de grado 3, lo cual resulta realmente inesperado.

A los efectos prácticos, puede utilizarse la fórmula de Mansion: que permite acotar el error sin necesidad de hallar derivadas:

$$|R| \leq \frac{2}{3} h |I - P - \frac{1}{2} E|$$

y cuya demostración se indica en otras lecturas recomendadas.

Ejemplo 2

En el ejemplo 1 se calculó con $n = 10$, mediante el método de Simpson, la integral:

$$\int_0^\pi \sin x \, dx$$

Se obtuvo:

$$\int_0^\pi \sin x \, dx \approx 2,000110$$

Utilice la fórmula de Peano y la de Mansion para hallar el error absoluto máximo de este resultado.

Solución:

Como el integrando es $f(x) = \sin x$, la cuarta derivada será $f^{(4)}(x) = \sin x$, la cual está acotada entre 0 y 1 en el intervalo de integración $[0, \pi]$. Tomando 1 como la cota superior, la fórmula de Peano conduce a:

$$-\frac{\pi - 0}{180} \left(\frac{\pi}{10} \right)^4 \leq R \leq 0$$

$$-0,00017 \leq R \leq 0$$

Es decir un error negativo (error por exceso) y menor en valor absoluto que 0,00017. En realidad, el error de truncamiento es de -0,00011.

La fórmula de Mansion utiliza los valores, calculados en el ejemplo 1,

$$\begin{aligned} E &= 0,000000 \\ I &= 3,236068 \\ P &= 3,077684 \end{aligned}$$

$$\text{de donde: } |R| \leq \frac{2}{3} h |I - P - \frac{1}{2} E| = 0,033$$

que es una cota muy conservadora para el error de truncamiento real.

Una fórmula asintótica para el error de truncamiento del método de Simpson

Es fácil obtener una fórmula para hallar una aproximación del error de truncamiento en el método de Simpson, que posee la ventaja de dar una estimación del error menos conservadora que las cotas que se obtienen con las fórmulas de Peano y Mansion. Su deducción es muy similar a la que se empleó en el método de los trapecios. A partir de la fórmula (7):

$$R = -\frac{h^5}{90} \sum_{i=0}^{\frac{n}{2}-1} f^{(4)}(c_i)$$

se puede escribir:

$$R = -\frac{h^4}{180} \sum_{i=0}^{\frac{n}{2}-1} f^{(4)}(c_i) 2h$$

Entonces, trasponiendo y pasando al límite:

$$-\lim_{h \rightarrow 0} \frac{180R}{h^4} = \lim_{h \rightarrow 0} \sum_{i=0}^{\frac{n}{2}-1} f^{(4)}(c_i) 2h$$

y, como $f^{(4)}(x)$ es continua y, por lo tanto, integrable en $[a, b]$, el límite de la derecha existe y es la integral definida de esta función. Es decir:

$$-\lim_{h \rightarrow 0} \frac{180R}{h^4} = \int_a^b f^{(4)}(x) dx$$

Siendo $f^{(4)}(x)$ continua, puede utilizarse la regla de Barrow y resulta:

$$-\lim_{h \rightarrow 0} \frac{180R}{h^4} = f^{(3)}(b) - f^{(3)}(a)$$

Así que, para valores pequeños de h , se tiene:

$$R \approx -\frac{f^{(3)}(b) - f^{(3)}(a)}{180} h^4 \quad (8)$$

donde el signo “ \approx ” se está utilizando con el sentido asintótico: Cuando h tiende hacia cero, ambos miembros son infinitésimos equivalentes.

Estimación del error de truncamiento por doble cálculo

La fórmula (8) no es tan atractiva para la estimación del error de truncamiento como su similar del método de los trapecios, ya que aquí se requiere hallar hasta la tercera derivada del integrando. Sin embargo, ella deja ver claramente que, para valores pequeños del paso h , el error de truncamiento es de la forma:

$$R = Ch^4 \quad (9)$$

lo cual se puede utilizar de inmediato para hallar estimaciones de R a partir de dos aplicaciones del método de Simpson.

Sea I_h : Resultado obtenido al calcular la integral por el método de Simpson con paso h
 I_{2h} : Resultado obtenido al calcular la integral por el método de Simpson con paso $2h$
 R_h : Error de truncamiento de I_h

Como ya se demostró en la sección 5.2 (fórmula 14):

$$R_h \approx \frac{I_h - I_{2h}}{2^p - 1}$$

Como, de acuerdo con (9), para el método de Simpson es $p = 4$, resulta:

$$R_h \approx \frac{I_h - I_{2h}}{15} \quad (10)$$

Ejemplo 3

En el ejemplo 1 la integral

$$\int_0^\pi \sin x \, dx$$

fue calculada con el método de Simpson utilizando $n = 10, 20$ y 40 sub intervalos. Utilice estos resultados para estimar el error de truncamiento obtenido para $n = 20$ y $n = 40$.

Solución:

Para $n = 20$, se obtuvo: $I_h = 2,00000678$

Para $n = 10$, se obtuvo: $I_{2h} = 2,00011$.

El error de truncamiento de I_h viene dado aproximadamente por:

$$R_h \approx \frac{I_h - I_{2h}}{15} = \frac{2,00000678 - 2,00011}{15} = -0,00000688$$

Es decir, al tomar 20 sub-intervalos el error de truncamiento es aproximadamente de $-0,00000688$. Nótese que el verdadero error de este resultado es de $-0,00000678$.

Para $n = 40$, se obtuvo: $I_h = 2,00000042$

Para $n = 20$, se obtuvo: $I_{2h} = 2,00000678$

El error de truncamiento de I_h es aproximadamente:

$$R_h \approx \frac{I_h - I_{2h}}{3} = \frac{2,00000042 - 2,00000678}{15} = -0,00000424$$

Esto es, tomando $n = 40$ sub-intervalos el error de truncamiento resulta aproximadamente de $-0,00000424$. El verdadero error de truncamiento es en este caso de $-0,0000042$.

Ejemplo 4

Calcule la integral $\int_1^2 e^{x^2} dx$ con cuatro cifras decimales exactas utilizando el método de Simpson.

Solución:

Esta integral fue calculada mediante el método de los trapecios en la sección 5.2. Con ese procedimiento se requirió emplear hasta 640 sub intervalos. En los cálculos que siguen se ha utilizado un programa basado en el seudo código del método de Simpson, para efectuar cada una de las integrales. El error de truncamiento se ha estimado mediante doble cálculo utilizando la fórmula (10). El proceso se comenzó arbitrariamente con $n = 10$ y a partir de ahí se estimaron los errores comparando con el resultado anterior. La tabla 2 muestra todos los resultados. El proceso iterativo fue detenido tan pronto como el error estimado fue menor que 0,00005. Con cuatro cifras decimales exactas, el valor de la integral es 14.989986. Observe que se requirió calcular la integral solamente con 40 sub intervalos.

<i>n</i>	Integral	Error estimado
10	14.992527	
20	14.990140	0.000159
40	14.989986	0.000010

Tabla 2

Fórmulas de Newton – Cotes

La idea del método de los trapecios, posteriormente extendida al método de Simpson, puede generalizarse. En esencia, se trata de aproximar el integrando por un conjunto de polinomios de grado m (en el método de los trapecios $m = 1$ y en el método de Simpson $m = 2$) tomando para ello $m + 1$ nodos sucesivos del conjunto $\{a = x_0, x_1, x_2, \dots, x_n = b\}$. Después se integran estos polinomios y se obtienen fórmulas del tipo:

$$\int_a^b f(x) dx \approx \sum_{i=0}^n A_i y_i$$

donde $y_i = f(x_i)$ para $i = 0, 1, 2, \dots, n$ y las constantes A_i dependen del paso h y del grado m de los polinomios utilizados. Se supone, naturalmente, que n es un múltiplo de m .

Estas fórmulas reciben el nombre de Newton – Cotes. Por ejemplo, para $m = 3$, se obtiene la fórmula de Newton – Cotes de orden 3 (conocida también como fórmula de los tres octavos):

$$\int_a^b f(x)dx \approx \frac{3}{8}h[(y_0 + 3y_1 + 3y_2 + y_3) + (y_3 + 3y_4 + 3y_5 + y_6) + \dots + (y_{n-3} + 3y_{n-2} + 3y_{n-1} + y_n)]$$

donde se supone que n es múltiplo de $m = 3$. Los términos de sub índice múltiplo de tres pueden agruparse y resulta:

$$\int_a^b f(x)dx \approx \frac{3}{8}h(E + 3S_1 + 3S_2 + 2S_3)$$

donde:

$E = y_0 + y_n$	(Suma de las ordenadas en los extremos)
$S_1 = y_1 + y_4 + \dots + y_{n-2}$	(Suma de las ordenadas de índice de tipo $3i + 1$)
$S_2 = y_2 + y_5 + \dots + y_{n-1}$	(Suma de las ordenadas de índice de tipo $3i + 2$)
$S_3 = y_3 + y_6 + \dots + y_{n-3}$	(Suma de las ordenadas de índice de tipo $3i$)

Sin embargo, el error de truncamiento de esta fórmula es del mismo orden h^4 que el método de Simpson, así que no tienen mucho sentido el uso de esta expresión más complicada.

El lector interesado en fórmulas de Newton – Cotes de orden aun mayor puede consultar la bibliografía que se recomienda al final de este capítulo. Más adelante se verá que el enfoque de Romberg es una vía mucho más simple para obtener resultados mejores que los de las fórmulas de Newton – Cotes de orden superior.

Ejercicios

Algunos de los ejercicios que siguen son muy laboriosos para realizar los cálculos a mano. Se supone que posea programas computacionales, preferiblemente confeccionados por usted, para ayudarle en el trabajo. De no ser así, en los ejercicios que requieran mucho trabajo manual, determine los resultados con menos cifras exactas que las exigidas.

1. Dada la integral $\int_1^2 \frac{2}{x^3} dx = 0,75$
 - a) Mediante el método de los trapecios, calcúlela aproximadamente con 4, 8, 16, 32 y 64 sub intervalos, determine el error en cada caso y verifique que en la medida en que el paso se duplica el error disminuye aproximadamente a la cuarta parte. ¿Por qué sucede esto?
 - b) Repita el cálculo de la integral, pero empleando el método de Simpson y determine cómo va decreciendo el error en la medida en que la cantidad de sub intervalos se va duplicando. Justifique por qué sucede así.
2. Resulta difícil medir la carga eléctrica, sin embargo la corriente puede determinarse con suficiente exactitud mediante un amperímetro. Este hecho puede aprovecharse para calcular la carga que ha ido almacenando una batería de automóvil mediante:

$$Q = \int_0^t idt$$

Utilice el método de Simpson para determinar aproximadamente, a partir de los datos de corriente medidos en ampere cada 10 minutos, la carga adquirida por la batería en una hora (utilice hora como unidad de tiempo y obtenga el resultado en ampere-hora).

t (minuto)	0	10	20	30	40	50	60
I (ampere)	3,7	3,0	2,5	2,0	1,7	1,4	1,1

3. Calcule con cuatro cifras decimales exactas, el volumen del sólido engendrado por la rotación alrededor del eje x de la región comprendida entre la gráfica de $y = e^x$, el eje x , la recta $x = 0$ y la recta $x = 1$.
4. A partir de la igualdad

$$\int_0^1 \frac{dx}{1+x^2} = \frac{\pi}{4}$$

utilice el método de Simpson para calcular π con cinco cifras decimales exactas.

5. En el ejemplo 4 de la sección 5.1 se analizó el problema de determinar el área de la sección transversal de un río con vista a calcular su caudal. En la figura 5 se muestran los resultados de un conjunto de mediciones de la profundidad del río a lo largo de su sección transversal. Calcule el área de la sección transversal y halle un estimado del error cometido.

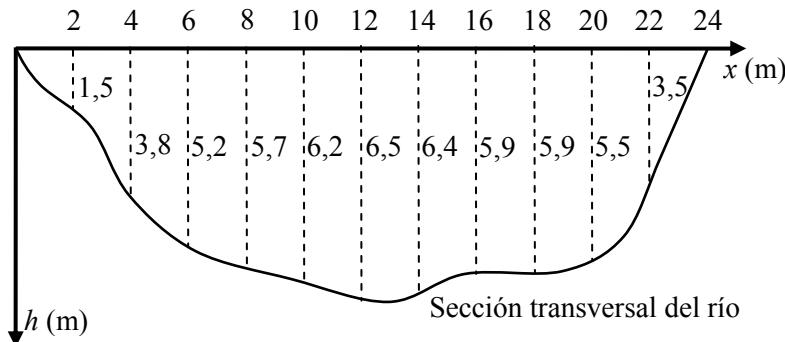


Figura 5

6. Dada la integral definida:

$$\int_1^3 (x^3 + 4x^2 - 5x + 4) dx$$

- a) Calcule su valor exacto por la regla de Barrow.
- b) Calcule su valor por el método de los trapecios con $n = 8$.
- c) Calcule su valor mediante el método de Simpson con $n = 8$.
- d) Explique por qué el resultado de Simpson es exacto y el de los trapecios no.

7. Calcule, con cuatro cifras decimales exactas, la integral:

$$\int_0^{1,5} \frac{\sin \theta}{\theta} d\theta$$

8. Cuando un cable o una cadena cuelga sujeto por los extremos y sometido solamente a su propio peso, adopta la forma de una catenaria, que es la gráfica del coseno hiperbólico. El cable que se muestra en la figura 6 tiene como ecuación:

$$y = \cosh 0,1x$$

respecto a un sistema de ejes, como muestra la figura, colocado de modo que el origen se halla un metro debajo del punto más bajo del cable y donde tanto x como y se miden en metros. Halle la longitud del cable con un error menor que un milímetro.

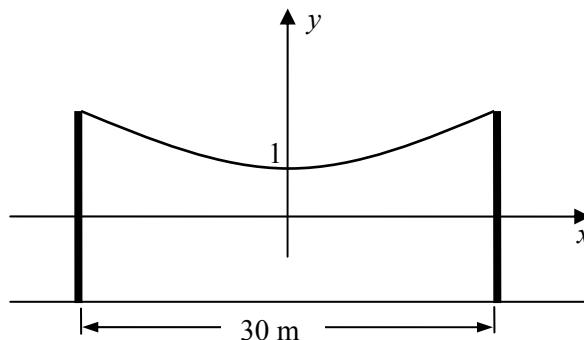


Figura 6

9. Al intentar calcular la longitud de una elipse, aparece un tipo de integral, llamado elíptica, la cual no puede ser evaluada por métodos analíticos. Estas integrales aparecen tabuladas y, en el capítulo 4 se vio cómo el problema se puede resolver interpolando en una de estas tablas. Por supuesto, también se puede calcular la integral elíptica numéricamente. Halle la longitud de una elipse de semiejes 3 y 4 con un error menor que 0,00005.
10. Se necesita calcular la superficie de una teja acanalada que se muestra en la figura 7. Suponga que el perfil de la teja es sinusoidal y calcule la superficie con un error menor que 5 cm².

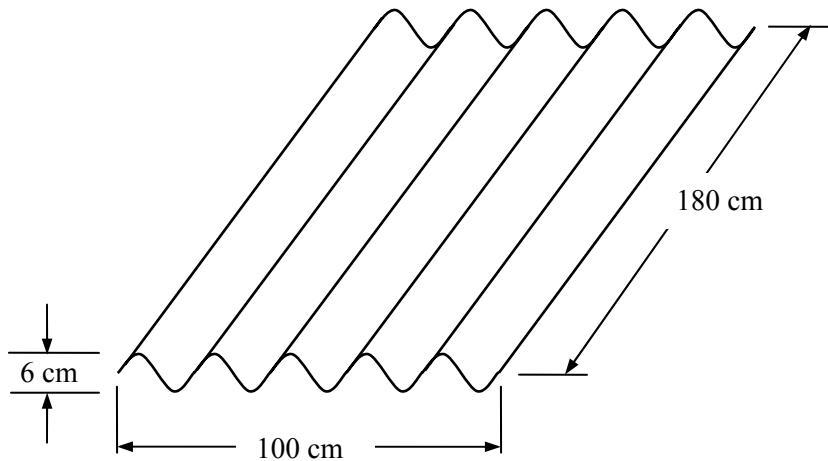


Figura 7

11. Uno de los generadores de una central termo eléctrica ha estado entregando durante un periodo de 4 horas los niveles de potencia que se indican en la tabla. Calcule la energía total entregada en ese periodo, en megawatt·hora y haga un estimado del error cometido.

t (hora)	0	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0
P (megawatt)	45	47	48	47	45	46	48	46	45

12. Elabore un algoritmo en seudo código que calcule una integral mediante el método de los trapecios utilizando como datos: el integrando, el intervalo de integración y la cantidad de subintervalos. El algoritmo debe dar como resultado no solo el valor aproximado de la integral sino, además, un estimado del error de truncamiento.
13. Elabore un algoritmo en seudo código que calcule una integral mediante el método de Simpson con un error menor que una cierta tolerancia. El algoritmo debe utilizar como datos: el integrando, el intervalo de integración y la tolerancia del error.

5.4 El método de Gauss

Los métodos considerados hasta aquí se han basado en dividir el intervalo de integración mediante un conjunto de puntos equidistantes en los cuales el integrando se evalúa. El método de integración numérica de Gauss, por el contrario, supone desconocidos los puntos del intervalo en que se evaluará la función y procede entonces a determinarlos de la manera más conveniente, es decir, de modo que la aproximación obtenida sea, en un cierto sentido, lo mejor posible.

Primero se considerará una integral en el intervalo particular $[-1, 1]$. Más adelante se verá que, mediante un sencillo cambio de variable, cualquier integral definida puede reducirse a una integral en ese intervalo. Sea la integral a calcular:

$$\int_{-1}^1 f(t) dt$$

Se trata, entonces, de encontrar una fórmula del tipo:

$$\int_{-1}^1 f(t)dt \approx \sum_{i=1}^m A_i f(t_i) \quad (1)$$

donde los parámetros A_1, A_2, \dots, A_m y t_1, t_2, \dots, t_m se hallarán siguiendo el criterio de que la fórmula (1) sea exacta para polinomios del mayor grado posible.

Como el conjunto P_n de todos los polinomios de grado menor o igual que n es un espacio vectorial de dimensión $n + 1$, es suficiente que la fórmula (1) sea exacta para los elementos de una base de P_n y entonces lo será para todos los elementos del espacio. En efecto, si $\{g_0(t), g_1(t), \dots, g_n(t)\}$ es una base de P_n , cualquier polinomio de grado menor o igual que n se puede escribir como combinación lineal de esta base, es decir:

$$p(t) = \lambda_0 g_0(t) + \lambda_1 g_1(t) + \dots + \lambda_n g_n(t)$$

y, debido a la linealidad de la integral:

$$\int_{-1}^1 p(t)dt = \lambda_0 \int_{-1}^1 g_0(t)dt + \lambda_1 \int_{-1}^1 g_1(t)dt + \dots + \lambda_n \int_{-1}^1 g_n(t)dt$$

Como se supone que la fórmula es exacta para las funciones que forman la base:

$$\int_{-1}^1 p(t)dt = \lambda_0 \sum_{i=1}^m A_i g_0(t_i) + \lambda_1 \sum_{i=1}^m A_i g_1(t_i) + \dots + \lambda_n \sum_{i=1}^m A_i g_n(t_i)$$

O sea:
$$\int_{-1}^1 p(t)dt = \sum_{i=1}^m A_i [\lambda_0 g_0(t_i) + \lambda_1 g_1(t_i) + \dots + \lambda_n g_n(t_i)]$$

Así que:
$$\int_{-1}^1 p(t)dt = \sum_{i=1}^m A_i p(t_i)$$

Es decir, que la aproximación (1) se convierte en exacta para cualquier polinomio de grado menor o igual que n .

Como en la fórmula (1) aparecen $2m$ coeficientes por determinar, es de esperar que esto permita hacerla exacta para $2m$ funciones independientes. Por tanto, debe ser posible lograr su exactitud para polinomios de grado menor o igual que $2m - 1$, cuyas bases son conjuntos de $2m$ funciones.

Con vistas a simplificar las notaciones, el análisis que sigue se limitará al caso concreto en que $m = 3$ que es suficientemente grande para que se comprenda la importancia de seleccionar adecuadamente la base de P_n con la que se operará. Como se verá, la generalización a otros valores de m será evidente.

La fórmula (1) para este caso particular tiene la forma:

$$\int_{-1}^1 f(t) dt \approx A_1 f(t_1) + A_2 f(t_2) + A_3 f(t_3) \quad (2)$$

y los seis parámetros $A_1, A_2, A_3, t_1, t_2, t_3$, serán calculados de manera que la fórmula (2) sea una igualdad para una base del espacio vectorial P_5 de los polinomios de grado menor o igual que cinco.

Suponga que se selecciona como base de P_5 el conjunto de funciones

$$\{1, t, t^2, t^3, t^4, t^5\}$$

(lo cual es una pésima selección), entonces sustituyendo en (2) a $f(t)$ por cada una de estas seis funciones y cambiando el signo “ \approx ” por un “ $=$ ”, se obtienen las seis ecuaciones que siguen:

$$\text{Para } f(t) = 1: \quad A_1 + A_2 + A_3 = \int_{-1}^1 dt = 2$$

$$\text{Para } f(t) = t: \quad A_1 t_1 + A_2 t_2 + A_3 t_3 = \int_{-1}^1 t dt = 0$$

$$\text{Para } f(t) = t^2: \quad A_1 t_1^2 + A_2 t_2^2 + A_3 t_3^2 = \int_{-1}^1 t^2 dt = \frac{2}{3}$$

$$\text{Para } f(t) = t^3: \quad A_1 t_1^3 + A_2 t_2^3 + A_3 t_3^3 = \int_{-1}^1 t^3 dt = 0$$

$$\text{Para } f(t) = t^4: \quad A_1 t_1^4 + A_2 t_2^4 + A_3 t_3^4 = \int_{-1}^1 t^4 dt = \frac{2}{5}$$

$$\text{Para } f(t) = t^5: \quad A_1 t_1^5 + A_2 t_2^5 + A_3 t_3^5 = \int_{-1}^1 t^5 dt = 0$$

Como se ve, se trata de un sistema no lineal de seis ecuaciones con seis incógnitas cuya solución es sumamente compleja y eso, a pesar de que se ha tomado un valor de m muy pequeño.

Los polinomios de Legendre

La base seleccionada por Gauss utiliza a una notable familia de polinomios llamada Polinomios de Legendre. El polinomio de Legendre de grado n ($n = 0, 1, 2, \dots$), se define como:

$$p_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n} [(t^2 - 1)^n]$$

de modo que:

$$p_0(t) = \frac{1}{2^0 0!} \frac{d^0}{dt^0} [(t^2 - 1)^0] = 1$$

$$p_1(t) = \frac{1}{2^1 1!} \frac{d}{dt} [(t^2 - 1)] = t$$

$$p_2(t) = \frac{1}{2^2 2!} \frac{d^2}{dt^2} [(t^2 - 1)^2] = \frac{1}{8} \frac{d^2}{dt^2} [t^4 - 2t^2 + 1] = \frac{1}{2} (3t^2 - 1)$$

$$p_3(t) = \frac{1}{2^3 3!} \frac{d^3}{dt^3} [(t^2 - 1)^3] = \frac{1}{24} (5t^3 - 3t)$$

$$p_4(t) = \frac{1}{2^4 4!} \frac{d^4}{dt^4} [(t^2 - 1)^4] = \frac{1}{192} (35t^4 - 30t^2 + 3)$$

⋮

Los polinomios de Legendre gozan de muchas interesantes propiedades, de las cuales aquí se citarán las dos que se necesitarán más adelante:

Propiedad 1: Los n ceros del polinomio de Legendre de grado n son reales y diferentes y se encuentran en el intervalo $[-1, 1]$

Propiedad 2: Si q es cualquier entero no negativo menor que n , se cumple que

$$\int_{-1}^1 t^q p_n(t) dt = 0$$

Una base para el espacio P_5

Una forma muy conveniente de seleccionar una base para el espacio de los polinomios de grado menor o igual que cinco es la siguiente:

$$\{1, t, t^2, p_3(t), t \cdot p_3(t), t^2 \cdot p_3(t)\}$$

Observe que estas funciones forman un conjunto linealmente independiente ya que sus grados son respectivamente 0, 1, 2, 3, 4 y 5. Para esta nueva base de P_5 se repetirá ahora el proceso de sustituir cada una de sus funciones en la expresión (2), haciéndola exacta, es decir, cambiando el signo “≈” por un “=”. Se obtienen las seis ecuaciones:

$$\text{Para } f(t) = 1: \quad A_1 + A_2 + A_3 = \int_{-1}^1 dt = 2$$

$$\text{Para } f(t) = t: \quad A_1 t_1 + A_2 t_2 + A_3 t_3 = \int_{-1}^1 t dt = 0$$

$$\text{Para } f(t) = t^2: \quad A_1 t_1^2 + A_2 t_2^2 + A_3 t_3^2 = \int_{-1}^1 t^2 dt = \frac{2}{3}$$

$$\text{Para } f(t) = p_3(t): \quad A_1 p_3(t_1) + A_2 p_3(t_2) + A_3 p_3(t_3) = \int_{-1}^1 p_3(t) dt = 0$$

$$\text{Para } f(t) = t \cdot p_3(t): \quad A_1 t_1 p_3(t_1) + A_2 t_2 p_3(t_2) + A_3 t_3 p_3(t_3) = \int_{-1}^1 t p_3(t) dt = 0$$

$$\text{Para } f(t) = t^2 \cdot p_3(t): \quad A_1 t_1^2 p_3(t_1) + A_2 t_2^2 p_3(t_2) + A_3 t_3^2 p_3(t_3) = \int_{-1}^1 t^2 p_3(t) dt = 0$$

Obsérvese que las tres últimas integrales no ha sido necesario calcularlas pues, por la propiedad 2 de los polinomios de Legendre, ellas valen cero.

Aunque, aparentemente, este sistema de ecuaciones es mucho más difícil que el anterior, su solución es realmente muy sencilla: Tomando los parámetros t_1, t_2 y t_3 como los tres ceros de $p_3(t)$ (recuérdese que son reales y distintos y están en el intervalo $[-1, 1]$), las tres últimas ecuaciones del sistema quedan satisfechas automáticamente (independientemente de los valores de A_1, A_2 y A_3), mientras que, conocidos t_1, t_2 y t_3 , los valores de A_1, A_2 y A_3 se hallan sin dificultad a partir de las tres primeras ecuaciones del sistema, que forman un sistema lineal. Es decir:

t_1, t_2 y t_3 son los ceros de $p_3(t) = \frac{1}{2}(5t^3 - 3t)$, que son:

$$t_1 = -\sqrt{\frac{3}{5}} = -0,774596669 \quad t_2 = 0 \quad t_3 = \sqrt{\frac{3}{5}} = 0,774596669$$

y A_1, A_2 y A_3 forman la solución del sistema:

$$\begin{aligned} A_1 + A_2 + A_3 &= 2 \\ A_1 t_1 + A_2 t_2 + A_3 t_3 &= 0 \\ A_1 t_1^2 + A_2 t_2^2 + A_3 t_3^2 &= \frac{2}{3} \end{aligned}$$

Sustituyendo t_1, t_2 y t_3 por sus valores, el sistema queda:

$$\begin{aligned} A_1 + A_2 + A_3 &= 2 \\ -\sqrt{\frac{3}{5}} A_1 + \sqrt{\frac{3}{5}} A_3 &= 0 \\ \frac{3}{5} A_1 + \frac{3}{5} A_3 &= \frac{2}{3} \end{aligned}$$

cuyas sus soluciones son:

$$A_1 = \frac{5}{9} = 0,55555555... \quad A_2 = \frac{8}{9} = 0,88888888... \quad A_3 = \frac{5}{9} = 0,55555555...$$

A modo de resumen, se puede plantear que:

La fórmula de Gauss para tres puntos:

$$\int_{-1}^1 f(t) dt \approx A_1 f(t_1) + A_2 f(t_2) + A_3 f(t_3)$$

$$\begin{aligned} \text{donde } t_1 &= -0,774596669 & A_1 &= 0,555555556 \\ t_2 &= 0 & A_2 &= 0,888888889 \\ t_3 &= 0,774596669 & A_3 &= 0,555555556 \end{aligned}$$

es exacta hasta para polinomios de grado menor o igual que 5.

Generalización para cualquier m

No ofrece ninguna dificultad generalizar para cualquier entero positivo m el procedimiento seguido para el caso particular de tres puntos.

La fórmula de Gauss (también conocida, justamente, como de Gauss y Legendre) para m puntos viene dada por:

$$\int_{-1}^1 f(t)dt \approx \sum_{i=1}^m A_i f(t_i) \quad (3)$$

donde: t_1, t_2, \dots, t_m son los ceros del polinomio de Legendre de grado m

y A_1, A_2, \dots, A_m son la solución del sistema lineal:

$$t_1^i A_1 + t_2^i A_2 + \dots + t_m^i A_m = \begin{cases} \frac{2}{n+1} & \text{si } i \text{ es par} \\ 0 & \text{si } i \text{ es impar} \end{cases} \quad i = 0, 1, 2, \dots, m-1$$

La fórmula (3) da resultados exactos si el integrando es cualquier polinomio de grado menor o igual que $2m-1$. Si $f(t)$ es una función no polinómica los resultados ya no serán exactos pero su error será tanto menor mientras mayor sea el valor de m .

Generalización para cualquier intervalo

Cuando el intervalo de integración es $[a, b]$, un cambio de variable lo puede siempre reducir al caso $[-1, 1]$. En efecto, considere la integral:

$$\int_a^b f(x)dx$$

Sea

$$x = \frac{a+b}{2} + \frac{b-a}{2}t \quad (4)$$

La ecuación (4) define a x como una función de primer grado de t , de tal modo que:

Para $t = -1$ resulta $x = a$

Para $t = 1$ resulta $x = b$

y por otra parte: $dx = \frac{b-a}{2} dt$

Entonces, la integral se transforma en:

$$\begin{aligned} \int_a^b f(x)dx &= \int_{-1}^1 f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right) \frac{b-a}{2} dt = \\ &= \frac{b-a}{2} \int_{-1}^1 f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right) dt \end{aligned}$$

A la última integral se le puede aplicar la fórmula (3), pues su intervalo de integración es $[-1, 1]$. De ahí resulta:

$$\frac{b-a}{2} \int_{-1}^1 f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right) dt \approx \frac{b-a}{2} \sum_{i=1}^m A_i f\left(\frac{a+b}{2} + \frac{b-a}{2}t_i\right)$$

Para que la fórmula resulte más corta, se acostumbra llamar:

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} t_i \quad \text{para } i = 1, 2, \dots, m$$

y se obtiene:

$$\int_a^b f(x) dx \approx \frac{b-a}{2} \sum_{i=1}^m A_i f(x_i)$$

en resumen:

La fórmula de Gauss para m puntos:

$$\int_a^b f(x) dx \approx \frac{b-a}{2} \sum_{i=1}^m A_i f(x_i)$$

donde

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} t_i \quad \text{para } i = 1, 2, \dots, m$$

da resultados exactos si $f(x)$ es cualquier polinomio de grado menor o igual que $2m - 1$.

Aunque los parámetros t_i y A_i pueden ser calculados sin dificultad, los mismos se encuentran tabulados con gran exactitud para su uso computacional. La tabla 1 ofrece sus valores hasta $m = 8$.

	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$
t_1	-0,57735027	-0,77459667	-0,86113631	-0,90617985	-0,93246951	-0,94910791	-0,96028986
t_2	0,57735027	0	-0,33998104	-0,53846931	-0,66120939	-0,74153119	-0,79666648
t_3		0,77459667	0,33998104	0	-0,23861919	-0,40584515	-0,52553242
t_4			0,86113631	0,53846931	0,23861919	0	-0,18343464
t_5				0,90617985	0,66120939	0,40584515	0,18343464
t_6					0,93246951	0,74153119	0,52553242
t_7						0,94910791	0,79666648
t_8							0,96028986
A_1	1	0,55555556	0,34785484	0,23692688	0,17132450	0,12948496	0,10122854
A_2	1	0,88888889	0,65214516	0,47862868	0,36076158	0,27970540	0,22238104
A_3		0,55555556	0,65214516	0,56888889	0,46791394	0,38183006	0,31370664
A_4			0,34785484	0,47862868	0,46791394	0,41795918	0,36268378
A_5				0,23692688	0,36076158	0,38183006	0,36268378
A_6					0,17132450	0,27970540	0,31370664
A_7						0,12948496	0,22238104
A_8							0,10122854

Tabla 1

Ejemplo 1

Calcular la integral:

$$\int_0^{\pi} \sin x \, dx$$

- a) Mediante el método de Gauss de tres puntos. b) Mediante el método de Gauss de 5 puntos.

Solución:

- a) A partir de la tabla 1 se tienen los valores de los parámetros t_i y A_i . Mediante los valores de t_i se obtiene:

$$x_i = \frac{0 + \pi}{2} + \frac{\pi - 0}{2} t_i = \frac{\pi}{2} + \frac{\pi}{2} t_i$$

De ahí:
 $x_1 = 0,35406272$
 $x_2 = 1,57079633$
 $x_3 = 2,78752993$

$$\begin{aligned} \int_0^{\pi} \sin x \, dx &\approx \frac{\pi - 0}{2} \sum_{i=1}^m A_i \sin x_i \\ &= 1,57079633[0,55555556 \sin(0,35406272) + 0,88888889 \sin(1,57079633) + \\ &\quad + 0,55555556 \sin(2,78752993)] \\ \int_0^{\pi} \sin x \, dx &\approx 2,00138892 \end{aligned}$$

El error de truncamiento es 0,0014, un resultado verdaderamente asombroso si se tiene en cuenta que el integrando solamente fue evaluado en tres puntos.

- b) De forma similar, para $m = 5$ resulta:

$$\int_0^{\pi} \sin x \, dx \approx 2,00000011$$

Como se observa, con solo cinco evaluaciones del integrando, se han logrado seis cifras decimales exactas. Un resultado similar requiere, con el método de Simpson, evaluar el integrando en 41 puntos (vea el ejemplo 1 de la sección 5.3).

Algoritmo en seudo código

El siguiente algoritmo calcula el valor aproximado de la integral de $f(x)$ en el intervalo $[a, b]$ mediante el método de Gauss de m puntos. Se supone que $f(x)$ es continua en $[a, b]$. Se utilizan como datos: la función $f(x)$, el intervalo de integración, el número entero $m > 1$ y una tabla con los valores de los parámetros t_i y A_i para $i = 1, 2, 3, \dots, m$.

```
Suma := 0
for i = 1 to m
```

```


$$x_i = \frac{a+b}{2} + \frac{b-a}{2} t_i$$

Suma := Suma + A_i f(x_i)
end
Integral :=  $\frac{b-a}{2}$  Suma
El resultado es Integral
Terminar

```

El error en el método de Gauss

La obtención de la fórmula del error de truncamiento en el método de Gauss es una tarea muy difícil. Puede demostrarse (ver otras lecturas recomendadas) que, para la fórmula de m puntos de Gauss en el intervalo $[a, b]$, el error de truncamiento viene dado por:

$$R = \frac{(m!)^4 f^{(2m)}(c)}{[(2m)!]^3 (2m+1)} (b-a)^{2m+1}$$

para algún número c en $[a, b]$.

Sin embargo, la necesidad de hallar derivadas de un orden muy alto hace poco útil esta fórmula. Nótese, por ejemplo que para acotar el error en la fórmula de Gauss de 5 puntos se requiere analizar la décima derivada de $f(x)$.

Bajo la suposición de que $f^{(2m)}(x)$ no sufra variaciones muy grandes en el intervalo $[a, b]$, se puede suponer que el error de truncamiento es, aproximadamente:

$$R \approx C(b-a)^{2m+1}$$

donde la constante C depende de $f(x)$ y de m pero no de a y de b . Considere entonces que la integral se calcula mediante la fórmula de Gauss de m puntos y se obtiene un resultado I_1 con error de truncamiento:

$$R_1 \approx C(b-a)^{2m+1}$$

Por otra parte, sea I_2 el resultado obtenido cuando se calcula la integral como:

$$I_2 = \int_a^\omega f(x) dx + \int_\omega^b f(x) dx$$

donde ω es el punto medio del intervalo $[a, b]$, es decir:

$$\omega = \frac{a+b}{2}$$

empleando en cada una de las integrales la misma fórmula de Gauss de m puntos. Como en cada una de las integrales el intervalo se ha reducido a la mitad, sus errores de truncamiento serán ambos aproximadamente:

$$C \left(\frac{b-a}{2} \right)^{2m+1}$$

por tanto el resultado I_2 tendrá un error de truncamiento aproximado de

$$R_2 = 2C\left(\frac{b-a}{2}\right)^{2m+1} = \frac{1}{2^{2m}} C(b-a)^{2m+1} = \frac{1}{4^m} R_1$$

Es decir: $R_1 = 4^m R_2$

Por otra parte, como $I_1 + R_1 = I_2 + R_2$

Resulta: $I_1 + 4^m R_2 = I_2 + R_2$

y, despejando R_2 se tiene: $R_2 \approx \frac{I_2 - I_1}{4^m - 1}$ (5)

La fórmula (5) puede utilizarse en forma iterativa, dividiendo de nuevo los intervalos de integración de las dos integrales parciales si el error de truncamiento fuera mayor que lo requerido o, si se desea, repitiendo los cálculos con un valor mayor de m .

Ejemplo 2

Mediante la fórmula de Gauss de 4 puntos calcule la integral

$$\int_1^2 e^{x^2} dx$$

con cuatro cifras decimales exactas.

Solución:

Para el intervalo de integración $[1, 2]$: $I_1 = 14,98899597$

Para el intervalo de integración $[1; 1,5]$: $2,60046192$

Para el intervalo de integración $[1,5; 2]$: $12,38950636$

$$I_2 = 14,98996828$$

$$R_2 \approx \frac{I_2 - I_1}{4^4 - 1} = \frac{14,98996828 - 14,98899597}{255} = 0,00000038$$

Por tanto, con cinco cifras decimales exactas, $\int_1^2 e^{x^2} dx = 14,989968$.

Nótese que, para llegar a este resultado el integrando fue evaluado 12 veces, 4 para la integral I_1 y 8 para obtener I_2 . En el método de Simpson fueron necesarias 73 evaluaciones, 11 para la primera aproximación, 21 para la segunda y 41 para la tercera y final.

Ejercicios

Algunos de los ejercicios que siguen son muy laboriosos para realizar los cálculos a mano. Se supone que posea un programa computacional, preferiblemente confeccionado por usted, para ayudarle en el trabajo. De no ser así, en los ejercicios que requieran mucho trabajo manual, determine los resultados con menos cifras exactas que las exigidas.

1. Calcule el área sombreada de la figura 1, limitada por la gráfica de la función $f(x) = \frac{\sin x}{x}$, mediante el método de Gauss de 4 puntos. (Ayuda: aproveche la simetría de la región).

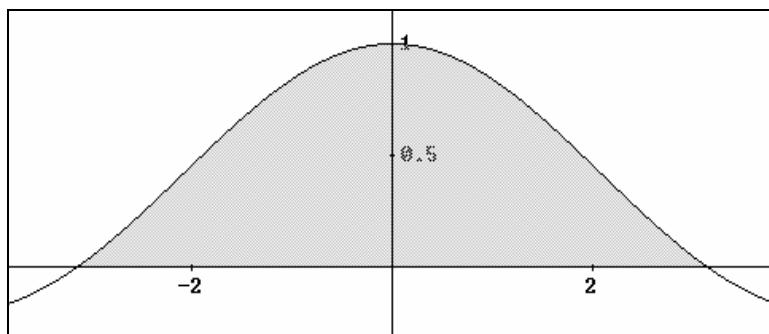


Figura 1

2. Se necesita calcular la integral:

$$\int_{1,75}^{3,17} (3x^7 - 5x^6 + 18x^4 - 17x^3 + 1,6x + 11,5) dx$$

¿Qué método de Gauss usted utilizaría para obtener el resultado exacto de la integral?

3. El método de Gauss y otros similares (por ejemplo, el de Chebyshev) debido a que requieren evaluar el integrando en muy pocos puntos, resultan ventajosos cuando se necesita calcular áreas de regiones limitadas por elementos constructivos curvos. Por ejemplo, en el diseño del casco de un barco, se necesita calcular el área de algunas secciones transversales con vistas a determinar el volumen de agua que se desplazará. En la figura 2 se muestra una de estas secciones. Determine en qué puntos sería necesario medir distancias verticales para poder calcular aproximadamente el área de la mitad de la sección mediante el método de Gauss de 5 puntos. Dé una fórmula para calcular el área total.

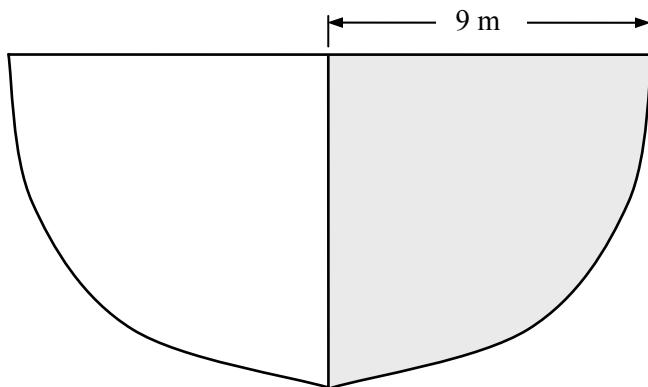


Figura 2

4. Halle aproximadamente el volumen del sólido de revolución que se genera cuando la región de la figura 3 gira alrededor del eje x. Utilice el método de Gauss de 4 puntos. Nota: para

hallar los puntos en que las curvas se intersecan utilice alguno de los métodos numéricos estudiados en el capítulo 2.

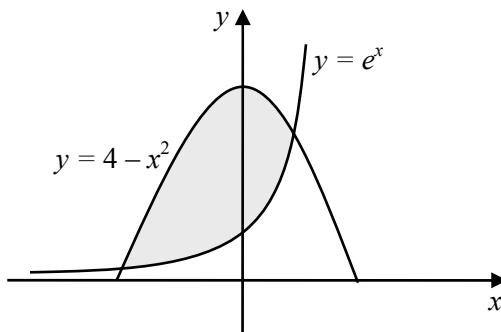


Figura 3

5. Halle la longitud de un periodo de la cosinusoide $y = \cos x$ con cuatro cifras decimales exactas, mediante el método de Gauss de cuatro puntos. Utilice la simetría de la curva.

6. Calcule la integral:

$$\int_0^2 \frac{dx}{\sqrt{1+x^3}}$$

con error menor que 0,00005 mediante algún método de Gauss.

7. Se quiere calcular el volumen que tendrá una copa (figura 4) que se encuentra en proceso de diseño. Determine a qué profundidades habría que tomar la longitud de tres diámetros con vista a calcular el volumen mediante el método de Gauss de 3 puntos. Dé una fórmula para calcular aproximadamente el volumen.

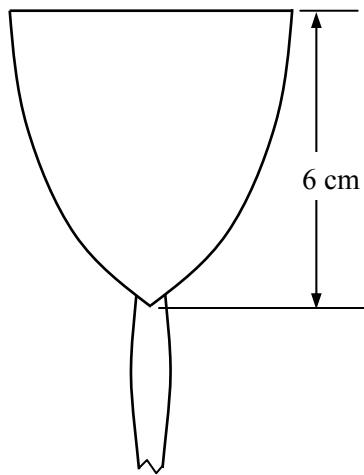


Figura 4

8. Calcule la integral que sigue con 5 cifras decimales exactas mediante el método de Gauss de 4 puntos.

$$\int_1^2 \frac{e^x}{x} dx$$

9. A partir de la igualdad

$$\int_a^b \frac{dx}{x} = \ln b - \ln a$$

para a y b positivos, calcule $\ln 2$ con cinco cifras decimales exactas mediante algún método de Gauss.

10. Elabore un algoritmo en seudo código que calcule el logaritmo neperiano de cualquier número mayor que 1 utilizando la idea del ejercicio 9. El algoritmo debe utilizar el método de Gauss de 4 puntos con doble cálculo para dar no solamente el valor del logaritmo sino un estimado del error de truncamiento.

5.5 El método de Romberg

El método de Romberg proporciona un procedimiento de cálculo numérico de integrales por aproximaciones sucesivas utilizando la idea de estimar el error mediante el método de doble cálculo. En la sección 5.2 se demostró que si la integral

$$I = \int_a^b f(x) dx$$

se halla con un método numérico cuyo error de truncamiento tiene la forma:

$$R_h \approx Ch^p$$

donde h es el paso de integración: $h = \frac{b-a}{n}$, entonces el error R_h puede ser estimado mediante:

$$R_h \approx \frac{I_h - I_{2h}}{2^p - 1}$$

donde I_h e I_{2h} son los resultados obtenidos al calcular la integral con pasos h y $2h$ respectivamente.

Es de suponer que, si se corrige el valor I_h con su error de truncamiento, se obtendrá una aproximación mejor del resultado exacto I . Es más, si R_h se calculara exactamente, la suma de I_h más R_h daría el resultado exacto de la integral. Puede probarse que, suponiendo que $f(x)$ es suficientemente derivable, la aproximación:

$$I_h + \frac{I_h - I_{2h}}{2^p - 1} = \frac{2^p I_h - I_{2h}}{2^p - 1} \quad (2)$$

posee un error de truncamiento de orden $p + 2$.

La fórmula (2), conocida como fórmula de extrapolación de Richardson es la base del método de Romberg. Para que este resulte más claro, conviene realizar algunos cambios en la notación.

- Sea:
- I_0^0 el resultado obtenido mediante el método de los trapecios usando un número pequeño n de subintervalos.
 - I_1^0 el resultado obtenido mediante el método de los trapecios con $2n$ subintervalos.
 - I_2^0 el resultado obtenido mediante el método de los trapecios con $4n$ subintervalos.

y, en general:

$$I_k^0 \quad \text{el resultado obtenido con } (2^k) \cdot n \text{ subintervalos, para } k = 0, 1, 2, \dots$$

Con estos valores iniciales, se pueden calcular mejores aproximaciones mediante la fórmula de Richardson, para los cuales se utilizará la notación: I_k^1

Con la nueva notación, la fórmula de Richardson (2) toma el aspecto:

$$I_k^1 = \frac{2^p I_k^0 - I_{k-1}^0}{2^p - 1} \quad k = 1, 2, 3, \dots \quad (3)$$

como, para el método de los trapecios es $p = 2$:

$$I_k^1 = \frac{4I_k^0 - I_{k-1}^0}{3} \quad k = 1, 2, 3, \dots \quad (4)$$

La fórmula (4) permite, con una reducida cantidad de operaciones aritméticas, mejorar las aproximaciones iniciales realizadas por el método de los trapecios. Los resultados de la generación 1: I_k^1 poseen un error de truncamiento de orden 4 y, de hecho, coinciden con los que se obtienen utilizando el método de Simpson.

El proceso puede continuarse. A partir de la generación 1, se pueden obtener los de generación 2, para los cuales en la fórmula (3) basta tomar $p+2$, en lugar de p :

$$I_k^2 = \frac{2^{p+2} I_k^1 - I_{k-1}^1}{2^{p+2} - 1} \quad k = 2, 3, 4, \dots$$

En general, la generación m se obtiene a partir de la $m-1$, mediante:

$$I_k^m = \frac{2^{p+2(m-1)} I_k^{m-1} - I_{k-1}^{m-1}}{2^{p+2(m-1)} - 1}$$

donde, el orden del error de truncamiento es $p + 2(m-1)$, ya que, con cada nueva generación el orden del error se incrementa en 2. Como $p = 2$, la fórmula anterior queda:

$$I_k^m = \frac{4^m I_k^{m-1} - I_{k-1}^{m-1}}{4^m - 1} \quad m = 1, 2, 3, \dots \quad k = m, m+1, m+2, \dots \quad (5)$$

La figura 1 muestra gráficamente la relación entre las diversas aproximaciones obtenidas en el proceso iterativo. La tabla que se muestra en la figura se va construyendo por filas: el primer elemento de la fila se calcula mediante el método de los trapecios y los restantes por la fórmula de Richardson (5). En cada fila el error se acota mediante la diferencia (en valor absoluto) entre el último elemento de la fila y el último elemento de la fila precedente, lo cual en la práctica, es una cota bastante conservadora.

k	Número de intervalos	Método de los trapecios	$I_k^1 = \frac{4I_k^0 - I_{k-1}^0}{3}$	$I_k^2 = \frac{16I_k^1 - I_{k-1}^1}{15}$	$I_k^3 = \frac{64I_k^2 - I_{k-1}^2}{63}$
0	n	I_0^0			
1	$2n$	I_1^0	I_1^1		
2	$4n$	I_2^0	I_2^1	I_2^2	
3	$8n$	I_3^0	I_3^1	I_3^2	I_3^3
4	$16n$	I_4^0	I_4^1	I_4^2	I_4^3
5	$32n$	I_5^0	I_5^1	I_5^2	I_5^3

Figura 1

Ejemplo 1

Calcule con cuatro cifras decimales exactas mediante el método de Romberg, la integral

$$\int_0^\pi \sin x \, dx$$

Solución:

Aplicando el método de los trapecios con $n = 4$ y $n = 8$ se obtiene, respectivamente:

$$I_0^0 = 1,896119$$

$$I_1^0 = 1,974232$$

y, con ellos: $I_1^1 = \frac{4I_1^0 - I_0^0}{3} = 2,000270$

con error absoluto menor que $|I_1^1 - I_0^0| = 0,104151$

Para comenzar la tercera fila ($k = 2$), se aplica el método de los trapecios con $n = 16$:

$$I_2^0 = 1,993570$$

y, mediante la fórmula de Richardson:

$$I_2^1 = \frac{4I_2^0 - I_1^0}{3} = 2,000016 \quad I_2^2 = \frac{16I_2^1 - I_1^1}{15} = 1,999999$$

El resultado I_2^2 tiene un error absoluto máximo de $|I_2^2 - I_1^1| = 0,000271$

Como esta cota es todavía alta, se pasa a la cuarta fila ($k = 3$):

Utilizando el método de los trapecios con 32 subintervalos:

$$I_3^0 = 1,998393$$

Y se aplica ahora tres veces la fórmula de Richardson, para obtener:

$$I_3^1 = \frac{4I_3^0 - I_2^0}{3} = 2,000001 \quad I_3^2 = \frac{16I_3^1 - I_2^1}{15} = 2,000000 \quad I_3^3 = \frac{64I_3^2 - I_2^2}{63} = 2,000000$$

El error absoluto máximo de este último resultado viene dado por: $|I_3^3 - I_2^2| = 0,000001$

que satisface el requerimiento de cuatro cifras decimales exactas (y también el de cinco decimales exactos). Así, la respuesta es: 2,000000 con cinco cifras decimales exactas.

Si no se posee un programa que haga todo el proceso automáticamente, pero sí uno con el método de los trapecios, los resultados se ordenan de forma similar a como se hizo en la figura 1, tal como muestra la tabla 1.

k	Intervalos	I_k^0	I_k^1	I_k^2	I_k^3	Error
0	4	1,896119				
1	8	1,974232	2,000270			0,104151
2	16	1,993570	2,000016	1,999999		0,000271
3	32	1,998393	2,000001	2,000000	2,000000	0,000001

Tabla 1

En la tabla 2 se muestra un resumen de los resultados que se han obtenido en el cálculo de la integral definida de $\sin x$ en el intervalo $[0, \pi]$ con los métodos de trapecios, Simpson y Romberg, para diferentes cantidades de subintervalos. Observe como el resultado de Romberg, que se ha comenzado con 4 sub intervalos, coincide con el resultado de los trapecios para esa misma cantidad, pero para 8 intervalos coincide con el resultado de Simpson y, a partir de ahí, logra valores con menor error que Simpson.

Intervalos	Trapecios	Simpson	Romberg
4	1,89611908	2,00455974	1,89611908
8	1,97423165	2,00026917	2,00026917
16	1,99357035	2,00001659	1,99999975
32	1,99839336	2,00000103	2,00000000
64	1,99959839	2,00000006	2,00000000

Tabla 2

Algoritmo en seudo código

El algoritmo que sigue calcula la integral definida de $f(x)$ en $[a, b]$ con error menor que ε mediante el método de Romberg. Se supone que $f(x)$ posee todas las derivadas requeridas en el intervalo $[a, b]$ y que los cálculos se realizan con las cifras decimales necesarias para no tomar en cuenta los errores por redondeo. El algoritmo utiliza como datos la función $f(x)$, el intervalo de integración $[a, b]$, la tolerancia ε del error, el número n de intervalos con que se desea comenzar y supone que existe un algoritmo auxiliar *Trapecios* que permite calcular integrales mediante el método de los trapecios.

```

 $I_0^0 := \text{Trapecios}(f(x), a, b, n)$  {El algoritmo Trapecios calcula la integral de  $f(x)$ 
en  $[a, b]$  utilizando  $n$  sub intervalos}
 $k := 0$ 
repeat
     $n := 2n$ 
     $k := k + 1$  {A partir de aquí se calcula la fila número  $k$ }
     $I_k^0 := \text{Trapecios}(f(x), a, b, n)$  {Se calcula el primer elemento de la fila}
    for  $m = 1$  to  $k$  {y en este lazo, los demás elementos de la fila}
         $I_k^m := \frac{4^m I_k^{m-1} - I_{k-1}^{m-1}}{4^m - 1}$ 
    end
     $Error := |I_k^k - I_{k-1}^{k-1}|$ 
until  $Error < \varepsilon$ 
El resultado es  $I_k^k$  con error absoluto máximo  $Error$ 

```

Ejemplo 2

Calcule la integral $\int_1^2 e^{x^2} dx$ con cuatro cifras decimales exactas utilizando el método de Romberg.

Solución:

La tabla 3 muestra los resultados obtenidos por un programa que calcula, a partir de $n = 4$, las filas de Romberg. En la tabla solo se muestra el elemento final de cada fila y la cota del error, que es el módulo de la diferencia entre dos resultados sucesivos.

k	Intervalos	Resultado	Error
0	4	16,07440693	
1	8	14,99608964	1.07831729
2	16	14,98999225	0.00609739
3	32	14.98997603	0.00001622
4	64	14.98997602	0.00000001

Tabla 3

Observe que, con 32 intervalos se ha logrado la exactitud requerida de cuatro cifras decimales exactas. Sin embargo, este resultado es mucho más exacto, ya que la cota del error que se utiliza en el método de Romberg es muy conservadora; la diferencia entre el resultado para $n = 32$ solo difiere en 10^{-8} del obtenido para $n = 64$, el cual posee siete cifras decimales exactas.

Ejercicios

Algunos de los ejercicios que siguen son muy laboriosos para realizar los cálculos a mano. Se supone que posea un programa computacional, preferiblemente confeccionado por usted, para ayudarle en el trabajo. De no ser así, en los ejercicios que requieran mucho trabajo manual, determine los resultados con menos cifras exactas que las exigidas.

1. Calcule la siguiente integral mediante el método de Romberg con error menor que 10^{-5} .

$$\int_0^2 \tan \sqrt{x} dx$$

2. Halle la longitud de la curva exponencial $y = e^x$ entre los puntos $(0, 1)$ y $(1, e)$ con cuatro cifras decimales exactas mediante el método de Romberg.
3. Mediante el método de Romberg calcule con cuatro cifras decimales exactas el área de la región sombreada de la figura 2, limitada por las gráficas de las curvas $y = \frac{\sin x}{x}$ y $y = \sin x$.

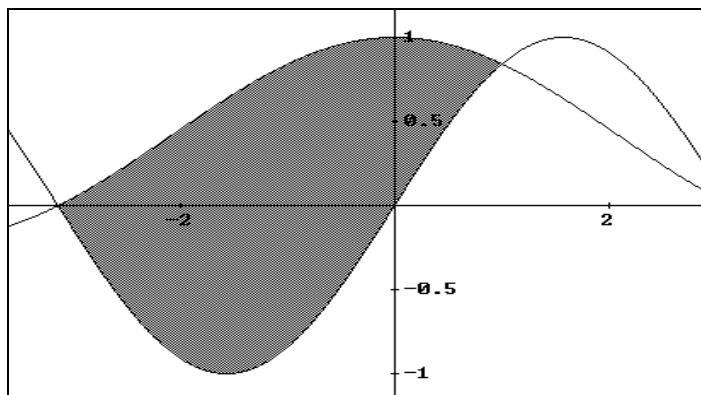


Figura 2

4. Utilice el método de Romberg para calcular, con cuatro cifras decimales exactas, la integral:

$$\int_0^{\pi} 3^{\sin x} dx$$

5. Si x es una variable aleatoria continua con distribución de probabilidad normal con media 0 y varianza 1, entonces su función de densidad probabilística viene dada por:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

que se denomina “campana de Gauss”. La probabilidad de que x tome algún valor entre 0 y 1 viene dada por la integral:

$$\int_0^1 \Phi(x) dx$$

Calcule esta integral mediante el método de Romberg con cinco cifras decimales exactas.

6. Puede probarse que

$$\int_0^{\infty} \Phi(x) dx = \frac{1}{2}$$

Determine con 3 cifras decimales exactas el número a tal que:

$$\int_0^a \Phi(x) dx = \frac{1}{4}$$

(Ayuda: Utilice algunos de los métodos del capítulo 2 para resolver ecuaciones numéricamente, por ejemplo, el método de bisección o el de la secante).

7. La figura 3 muestra la región sombreada R limitada por las curvas $y = \sqrt{\ln x}$ e $y = -(x - 2)(x - 5)$. Calcule el área de R con cuatro cifras decimales exactas mediante el método de Romberg.

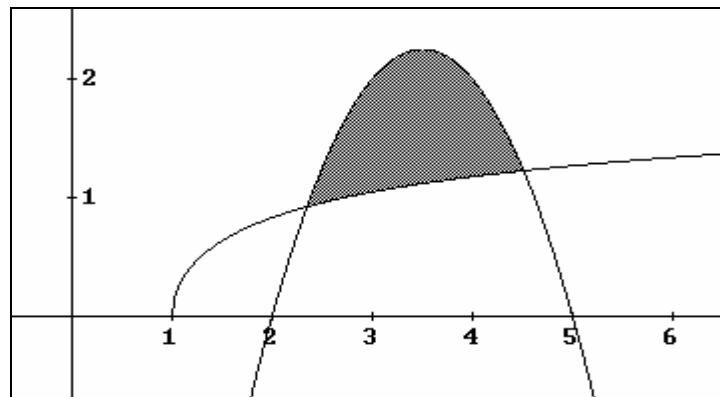


Figura 3

8. Un cable que está sometido a una carga uniforme (además de su propio peso), como ocurre en los puentes colgantes, adquiere la forma de una parábola de eje vertical. Halle la longitud

del cable que sostiene al puente colgante de la figura 4. Utilice el método de Romberg y halle el resultado con un error menor que 1 cm.

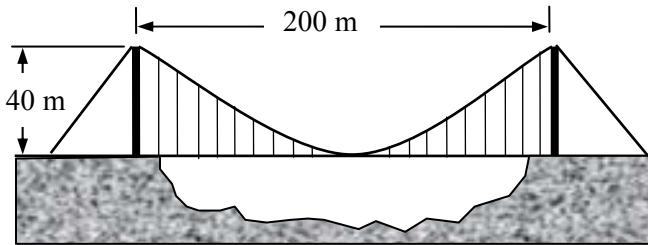


Figura 4

9. En el algoritmo en seudo código del método de Romberg se trabaja con el arreglo bidimensional I_k^m aunque el algoritmo realmente no requiere almacenar todas las aproximaciones que contiene este arreglo, sino solamente su última fila. Modifique el algoritmo de manera que solo se empleen dos arreglos unidimensionales, llamados *FilaActual* y *FilaAnterior*.

5.6 Cálculo numéricico de integrales dobles

Varios de los métodos numéricos estudiados antes pueden ser adaptados con facilidad al problema de calcular una integral doble. Sea la integral:

$$I = \int_a^b \int_{g(x)}^{h(x)} f(x, y) dy dx \quad (1)$$

cuya región de integración, limitada por las gráficas de $y = g(x)$, $y = h(x)$ y las rectas $x = a$ y $x = b$ se muestra en la figura 1.

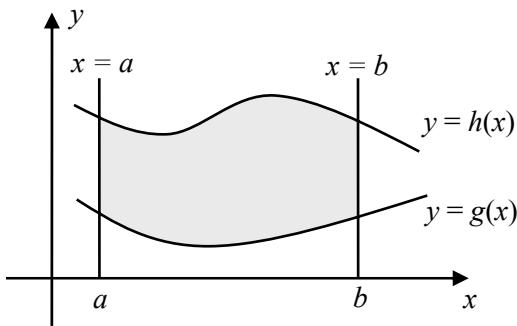


Figura 1

Realmente, el símbolo empleado en la fórmula (1), representa una integral iterada, es decir, una integral respecto a x , de una integral respecto a y . Así es como conviene enfocar este problema. Para ello la integral (1), se escribe:

$$I = \int_a^b F(x) dx \quad (2)$$

donde

$$F(x) = \int_{g(x)}^{h(x)} f(x, y) dy \quad (3)$$

y queda claro que el cálculo de I se reduce al cálculo de integrales unidimensionales.

Los métodos de los trapecios, de Simpson o de Gauss, permiten calcular aproximadamente la integral (2) mediante fórmulas del tipo:

$$\int_a^b F(x) dx \approx \sum_{i=1}^n B_i F(x_i) \quad (4)$$

El cálculo exacto de $F(x_i)$ requiere de la integral:

$$F(x_i) = \int_{g(x_i)}^{h(x_i)} f(x_i, y) dy \quad i = 1, 2, 3, \dots, n \quad (5)$$

pero su valor, para cada i , puede aproximarse por un método numérico (que no tiene que ser el mismo que el empleado en la expresión (4))

$$F(x_i) \approx \sum_{j=1}^m C_{ij} f(x_i, y_j) \quad i = 1, 2, 3, \dots, n \quad (6)$$

Observe que, como en cada x_i el intervalo de integración de (5) cambia, los coeficientes de (6) no son los mismos para cada i , sino que dependen de i y de j .

Como la cantidad de evaluaciones de $f(x, y)$ será de $m \cdot n$, se requiere utilizar métodos eficientes que no necesiten que m y n tomen valores altos. En lo que sigue se utilizarán, como casos particulares, el método de Gauss y el método de Simpson.

Cálculo de integrales dobles por el método de Gauss

Como el algoritmo de Gauss es muy simple de programar y requiere de muy pocos cálculos, es uno de los preferidos para evaluar integrales dobles. Por razones de simplicidad, se supondrá que todas las integrales unidimensionales se calculan con el mismo método de Gauss (realmente, esto no es obligatorio) de m puntos. Se supone que las funciones f , g y h son continuas y poseen las derivadas necesarias para aplicar este algoritmo. La integral que se desea calcular es:

$$I = \int_a^b \int_{g(x)}^{h(x)} f(x, y) dy dx$$

y su región de integración se muestra en la figura 1. Llamando, como hasta ahora:

$$F(x) = \int_{g(x)}^{h(x)} f(x, y) dy$$

la integral I se aproxima mediante la fórmula de Gauss:

$$I = \int_a^b F(x) dx \approx \frac{b-a}{2} \sum_{i=1}^m A_i F(x_i)$$

$$\text{donde } x_i = \frac{a+b}{2} + \frac{b-a}{2} t_i \quad \text{para } i = 1, 2, \dots, m \quad (7)$$

A_i y t_i son parámetros de la fórmula, previamente calculados y tabulados.

En cuanto al término $F(x_i)$, se aproxima también mediante la fórmula de Gauss para m puntos, es decir:

$$F(x_i) = \int_{g(x_i)}^{h(x_i)} f(x_i, y) dy \approx \frac{h(x_i) - g(x_i)}{2} \sum_{j=1}^m A_j f(x_i, y_j) \quad \text{para } i = 1, 2, \dots, m$$

$$\text{donde } y_j = \frac{g(x_i) + h(x_i)}{2} + \frac{h(x_i) - g(x_i)}{2} t_j \quad \text{para } j = 1, 2, \dots, m \quad (8)$$

A_j y t_j son, de nuevo, los parámetros de la fórmula de Gauss para m puntos.

El algoritmo se puede resumir en una sola fórmula:

$$\int_a^b \int_{g(x)}^{h(x)} f(x, y) dy dx \approx \frac{b-a}{2} \sum_{i=1}^m A_i \left[\frac{h(x_i) - g(x_i)}{2} \sum_{j=1}^m A_j f(x_i, y_j) \right]$$

con los x_i dados por (7) y, para cada i , los y_j calculados mediante (8).

La estimación del error de truncamiento es complicada y no será tratada.

Algoritmo en seudo código del método de Gauss para integrales dobles

El siguiente algoritmo calcula aproximadamente, mediante el método de Gauss de m puntos, la integral doble de $f(x, y)$ en la región

$$\{(x, y): a \leq x \leq b, g(x) \leq y \leq h(x)\}$$

como en la figura 1. Tanto el integrando como las funciones que limitan la región de integración se suponen continuas y con la cantidad de derivadas necesarias para que el algoritmo de Gauss que se emplee dé resultados satisfactorios. Se suponen conocidos las funciones f , g y h , el intervalo $[a, b]$, el orden m del método de Gauss que se empleará y los parámetros A_i y t_i ($i = 1, 2, \dots, m$) de la fórmula de Gauss.

```
Integral := 0
for i = 1 to m
```

```

 $x := \frac{a+b}{2} + \frac{b-a}{2} t_i$ 
 $c := g(x)$ 
 $d := h(x)$ 
 $F = 0$ 
for  $j = 1$  to  $m$ 
     $y := \frac{c+d}{2} + \frac{d-c}{2} t_j$ 
     $F := F + A_j f(x, y)$ 
end
 $F := \left( \frac{d-c}{2} \right) F$ 
 $Integral := Integral + A_i F$ 
end
 $Integral := \left( \frac{b-a}{2} \right) Integral$ 

```

El resultado es $Integral$

Terminar

Ejemplo 1

Calcule mediante el método de Gauss de 4 puntos la integral doble:

$$\int_2^4 \int_{\ln x}^{\sqrt{x}} \frac{dy dx}{x^2 + y^2}$$

Solución:

Los cálculos se realizarán paso a paso, siguiendo el algoritmo en seudo código, para ayudar a su comprensión.

Se toma $f(x, y) = \frac{1}{x^2 + y^2}$, $g(x) = \ln(x)$, $h(x) = \sqrt{x}$, $a = 2$, $b = 4$

y los valores de $t_1, t_2, t_3, t_4, A_1, A_2, A_3, A_4$ de la tabla 1 de la sección 5.4.

Para $i = 1$

$$x_1 = \frac{a+b}{2} + \frac{b-a}{2} t_1 = 2,138863690$$

Límites de integración: $c_1 = \ln(x_1) = 0,7602747020$ $d_1 = \sqrt{x_1} = 1,462485449$

Valores de y : $y_1 = 0,8090304898$

$$y_2 = 0,9920109056$$

$$y_3 = 1,230749245$$

$$y_4 = 1,413729661$$

$$F_1 = \frac{d_1 - c_1}{2} [A_1 f(x_1, y_1) + A_2 f(x_1, y_2) + A_3 f(x_1, y_3) + A_4 f(x_1, y_4)] = 0,1207277155$$

Para $i = 2$

$$x_2 = \frac{a+b}{2} + \frac{b-a}{2} t_2 = 2,660018959$$

$$\text{Límites de integración: } c_2 = \ln(x_2) = 0,9783332505 \quad d_2 = \sqrt{x_2} = 1,630956455$$

Valores de y :

$y_1 = 1,023646083$
$y_2 = 1,193705095$
$y_3 = 1,415584611$
$y_4 = 1,585643622$

$$F_2 = \frac{d_2 - c_2}{2} [A_1 f(x_2, y_1) + A_2 f(x_2, y_2) + A_3 f(x_2, y_3) + A_4 f(x_2, y_4)] = 0,07428028231$$

Para $i = 3$

$$x_3 = \frac{a+b}{2} + \frac{b-a}{2} t_3 = 3,339981040$$

$$\text{Límites de integración: } c_3 = \ln(x_3) = 1,205965130 \quad d_3 = \sqrt{x_3} = 1,827561501$$

Valores de y :

$y_1 = 1,249123713$
$y_2 = 1,411097825$
$y_3 = 1,622428805$
$y_4 = 1,784402918$

$$F_3 = \frac{d_3 - c_3}{2} [A_1 f(x_3, y_1) + A_2 f(x_3, y_2) + A_3 f(x_3, y_3) + A_4 f(x_3, y_4)] = 0,04615936537$$

Para $i = 4$

$$x_4 = \frac{a+b}{2} + \frac{b-a}{2} t_4 = 3,861136309$$

$$\text{Límites de integración: } c_4 = \ln(x_4) = 1,350961520 \quad d_4 = \sqrt{x_4} = 1,964977432$$

Valores de y :

$y_1 = 1,393593778$
$y_2 = 1,553592592$
$y_3 = 1,762346360$
$y_4 = 1,922345174$

$$F_4 = \frac{d_4 - c_4}{2} [A_1 f(x_4, y_1) + A_2 f(x_4, y_2) + A_3 f(x_4, y_3) + A_4 f(x_4, y_4)] = 0,03475074104$$

Cálculo final:

$$\text{Integral} = \frac{b-a}{2} (A_1 F_1 + A_2 F_2 + A_3 F_3 + A_4 F_4) = 0,1326280669$$

Luego el resultado es de 0,1326280669. La integral no es calculable por métodos analíticos. El resultado, hallado con 10 cifras exactas por vía numérica es: 0,1326309143, lo que significa que se obtuvo un error inferior a 0,000003.

Cálculo de integrales dobles por el método de Simpson

El método de Simpson para calcular integrales dobles, aunque requiere evaluar el integrando en muchos más puntos que el método de Gauss, tiene la ventaja, como se verá más adelante, de que el error de truncamiento se puede estimar por el proceso de doble cálculo. La deducción del algoritmo es muy similar a la del método de Gauss.

Tal como hasta aquí, la integral que se desea calcular es:

$$I = \int_a^b \int_{g(x)}^{h(x)} f(x, y) dy dx$$

con la misma región de integración mostrada en la figura 1. Llamando,

$$F(x) = \int_{g(x)}^{h(x)} f(x, y) dy$$

la integral I se puede aproximar por el método de Simpson para n subintervalos, como :

$$I = \int_a^b F(x) dx \approx \frac{h}{3} \sum_{i=0}^n A_i F(x_i) \quad (9)$$

$$\text{donde } h = \frac{b-a}{n}, \quad x_i = a + hi \text{ para } i = 0, 1, 2, \dots, n \quad (10)$$

y los coeficientes A_i son:

$$\begin{aligned} A_0 &= A_n = 1 \\ A_1 &= A_3 = A_5 = \dots = A_{n-1} = 4 \\ A_2 &= A_4 = A_6 = \dots = A_{n-2} = 2 \end{aligned} \quad (11)$$

En cuanto a $F(x_i)$, cuyo valor exacto sería:

$$F(x_i) = \int_{g(x_i)}^{h(x_i)} f(x_i, y) dy$$

Su valor se aproximará utilizando también el método de Simpson con la misma cantidad de sub intervalos (esto, realmente, no es necesario, pero de esta forma se obtienen expresiones más sencillas y el acotamiento del error es más simple). De esta manera resulta:

$$F(x_i) \approx \frac{h_i}{3} \sum_{j=0}^n A_j f(x_i, y_j) \quad (12)$$

$$\text{donde } h_i = \frac{h(x_i) - g(x_i)}{n}; \quad y_j = g(x_i) + jh_i \text{ para } j = 0, 1, 2, \dots, n \quad (13)$$

y los coeficiente A_j están dados por (11).

Combinando las ecuaciones (9) y (12) se obtiene una fórmula más compacta para el cálculo aproximado de la integral doble:

$$\int_a^b \int_{g(x)}^{h(x)} f(x, y) dy dx \approx \frac{h}{3} \sum_{i=0}^n A_i \left[\frac{h_i}{3} \sum_{j=0}^n A_j f(x_i, y_j) \right] \quad (14)$$

en la cual, según (10) y (13): $h = \frac{b-a}{n}$, $x_i = a + hi$ para $i = 0, 1, 2, \dots, n$

$$h_i = \frac{h(x_i) - g(x_i)}{n}; \quad y_j = g(x_i) + jh_i \text{ para } j = 0, 1, 2, \dots, n$$

y los coeficientes A_i y A_j se muestran en (11).

Algoritmo en seudo código del método de Simpson para integrales dobles

El siguiente algoritmo calcula aproximadamente, mediante el método de Simpson con n subintervalos, la integral doble de $f(x, y)$ en la región

$$\{(x, y) : a \leq x \leq b, g(x) \leq y \leq h(x)\}$$

como en la figura 1. Tanto el integrando como las funciones que limitan la región de integración se suponen continuas y con la cantidad de derivadas necesarias para que el algoritmo de Simpson ofrezca resultados satisfactorios. Se suponen conocidos las funciones f , g y h , el intervalo $[a, b]$, y el número n de subintervalos que se utilizará, el cual obligatoriamente debe ser par.

```

 $A_0 := 1$ 
 $A_n := 1$ 
for  $i = 1$  to  $n - 1$ 
     $A_i := 3 - (-1)^i$  {El resultado es 2 si  $i$  es par y 4 si  $i$  es impar}
end
 $Integral := 0$ 
 $hh := \frac{b-a}{n}$  {Este es el paso horizontal, para la variable  $x$ }
for  $i = 0$  to  $n$ 
     $x := a + hh \cdot i$ 
     $c := g(x)$ 
     $d := h(x)$ 
     $F := 0$ 
     $hv := \frac{d-c}{n}$  {Este es el paso vertical, para la variable  $y$ }
    for  $j = 0$  to  $n$ 
         $y := c + hv \cdot j$ 
         $F := F + A_j f(x, y)$ 
    end
     $F := \left( \frac{hv}{3} \right) F$ 
 $Integral := Integral + A_i F$ 

```

end

$$\text{Integral} := \left(\frac{hh}{3} \right) \text{Integral}$$

El resultado es *Integral*

Terminar

Ejemplo 2

Calcule mediante el método de Simpson con $n = 8$ subintervalos la integral doble:

$$\int_{2 \ln x}^{4 \sqrt{x}} \int_{x^2 + y^2} dy dx$$

Se trata de la misma integral que fue calculada por el método de Gauss en el ejemplo 1.

Solución:

Como $a = 2$ y $b = 4$, el paso de la variable x será: $h = \frac{4-2}{8} = 0,25$

La tabla 1 muestra, para cada uno de los 9 valores de x_i , el resultado $F_i \approx F(x_i)$ de la integral interior, calculado con el método de Simpson de 8 sub intervalos. Se han omitido los detalles del cálculo de los F_i por un problema de espacio. Todos los cálculos fueron ejecutados con un programa confeccionado a partir del algoritmo en seudo código.

i	x_i	F_i
0	2,00	0,1409302783
1	2,25	0,1075915751
2	2,50	0,0850521566
3	2,75	0,0691444759
4	3,00	0,0575205962
5	3,25	0,0487795461
6	3,50	0,0420453693
7	3,75	0,0367486993
8	4,00	0,0325071312

Tabla 1

El valor obtenido fue 0,1326442366, que comparado con el valor exacto de 0,1326309143, muestra que se lograron 4 cifras decimales exactas, menos que el método de Gauss de 4 puntos con el cual se obtuvo 5 cifras decimales exactas.

Estimación del error

Cuando la integral doble se calcula mediante el método de Simpson, la integral

$$I = \int_a^b \int_{g(x)}^{h(x)} f(x, y) dy dx$$

se aproxima como:

$$I = \int_a^b F(x) dx \approx \frac{h}{3} \sum_{i=0}^n A_i F_i \quad (15)$$

donde:

$$F_i = \frac{h_i}{3} \sum_{j=0}^n A_j f(x_i, y_j) \quad (16)$$

el resultado obtenido contiene errores de dos fuentes diferentes. Por una parte en la ecuación (15) se introduce un error de truncamiento propio del método aproximado de Simpson. Este error se llamará R_1 y, como se recordará, es asintóticamente proporcional a h^4 ; el mismo existiría aunque los valores de F_i coincidieran con $F(x_i)$. Por otra parte en (15) está presente el error debido a que los valores de F_i son solamente aproximaciones de $F(x_i)$, a este error se le llamará R_2 . Nótese que esta es la misma situación que se analizó en la sección 5.2 respecto al error introducido en el cálculo numérico de las integrales debido a errores en los datos. Si se supone que en cada uno de los F_i está presente el mismo error absoluto δ , entonces el error R_2 vendría dado por:

$$R_2 = \frac{h}{3} \sum_{i=0}^n A_i \delta = \frac{1}{3} \delta \frac{b-a}{n} \sum_{i=0}^n A_i$$

Ahora bien, es fácil calcular la suma indicada:

$$\sum_{i=0}^n A_i = 1 + 4 + 2 + 4 + \dots + 2 + 4 + 1$$

y, como aparece $\frac{n}{2}$ veces el sumando 4 y $(\frac{n}{2}-1)$ veces el sumando 2:

$$\sum_{i=0}^n A_i = 1 + 1 + 4(\frac{n}{2}) + 2(\frac{n}{2}-1) = 3n$$

De manera que:

$$R_2 = \frac{1}{3} \delta \frac{b-a}{n} (3n) = \delta(b-a) \quad (17)$$

En la realidad, todos los F_i no contendrán el mismo error. Cada uno de estos errores procede del error de truncamiento introducido en la ecuación (16) y, por tanto, será proporcional a h_i^4 . Con vista a obtener una cota superior del error, puede calcularse δ tomando el mayor de los h_i el cual toma el valor:

$$\frac{1}{n} \max_i [h(x_i) - g(x_i)]$$

Esto es:

$$\delta = K_3 \left(\frac{1}{n} \max_i [h(x_i) - g(x_i)] \right)^4$$

y, agrupando todas las constantes (independientes de n) en una sola:

$$\delta = K_2 \left(\frac{1}{n} \right)^4$$

Con esto, y tomando en cuenta la fórmula (17), el error R_2 puede expresarse como:

$$R_2 = K_2 (b-a) \left(\frac{1}{n} \right)^4 \quad (18)$$

Por su parte, el error R_1 viene dado por:

$$R_1 = Ch^4 = C\left(\frac{b-a}{n}\right)^4$$

Que, después de reunir las constantes, resulta:

$$R_1 = K_1\left(\frac{1}{n}\right)^4 \quad (19)$$

El error absoluto total de (15) puede entonces estimarse como la suma de $R_1 + R_2$

Esto es: $R = R_1 + R_2 = K_1\left(\frac{1}{n}\right)^4 + K_2(b-a)\left(\frac{1}{n}\right)^4$

que, con una nueva agrupación de constantes, resulta en:

$$R = K\left(\frac{1}{n}\right)^4$$

Esto prueba que el método de Simpson para integrales dobles continua siendo un método de orden 4, al igual que para integrales unidimensionales. Lo cual es suficiente para extender el procedimiento de doble cálculo utilizado para integrales simples al caso de las integrales dobles. Esto puede resumirse:

Si al calcular la integral doble:

$$\int_a^b \int_{g(x)}^{h(x)} f(x, y) dy dx$$

mediante el método de Simpson, se obtienen resultados I_h e I_{2h} al utilizar respectivamente pasos h y $2h$, entonces, el error en I_h puede estimarse como:

$$R_h \approx \frac{I_h - I_{2h}}{15}$$

Ejemplo 3

Calcule la integral doble:

$$\int_2^4 \int_{\ln x}^{\sqrt{x}} \frac{dy dx}{x^2 + y^2}$$

con cinco cifras decimales exactas utilizando el método de Simpson.

Solución:

Al calcular con $n = 4$, se obtiene el resultado:

$$I_{2h} = 0,1328112882$$

Utilizando $n = 8$:

$$I_h = 0,1326442366$$

El error de I_h se estima como:

$$R_h \approx \frac{I_h - I_{2h}}{15} = \frac{0,1326442366 - 0,1328112882}{15}$$

$$R_h \approx -0,000011$$

por lo cual, no es de esperar que I_h tenga cinco cifras decimales exactas. Se repite entonces el proceso, tomando ahora:

Para $n = 8$:

$$I_{2h} = 0,1326442366$$

Para $n = 16$:

$$I_h = 0,1326317902$$

El error contenido en este nuevo resultado puede estimarse como:

$$R_h \approx \frac{I_h - I_{2h}}{15} = \frac{0,1326317902 - 0,1326442366}{15}$$

$$R_h \approx -0,00000083$$

Por tanto, con cinco cifras decimales exactas, el resultado es: 0,1326318.

Observe que la estimación del error es bastante exacta, pues, teniendo en cuenta que el resultado exacto de la integral es 0,1326309143, el verdadero error resulta de: -0,000000876.

Ejercicios

1. Calcule la integral doble:

$$\iint_R xy^2 dxdy$$

donde R es la región que muestra la figura 2. Utilice el método de Gauss de 4 puntos.

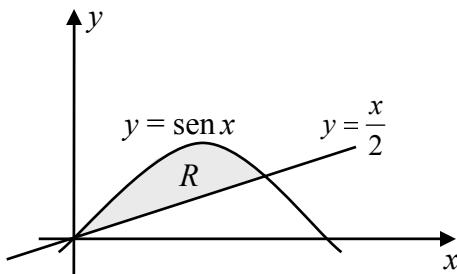


Figura 2

2. Calcule la integral del ejercicio anterior con un error menor que 0,00005 mediante el método de Simpson.

3. Calcule, con cinco cifras decimales exactas, la integral doble:

$$\int_1^3 \int_1^x \frac{\sin y}{y} dy dx$$

4. Determine cuál de los métodos de Gauss (es decir, con cuantos puntos) se requiere para calcular la integral del ejercicio 3 con una exactitud similar.
5. Mediante el método de Gauss de 4 puntos calcule la integral doble:

$$\int_{-2}^2 \int_0^{4-x^2} e^{xy} dy dx$$

6. Calcule la integral del ejercicio anterior mediante el método de Simpson con 16 sub intervalos y determine aproximadamente el error de truncamiento.
7. Calcule la integral doble

$$\iint_R \frac{dxdy}{x^2 + y^2 + 1}$$

donde R es la región limitada por una cardioide como se muestra en la figura 3. Utilice el método de Gauss de 4 puntos.

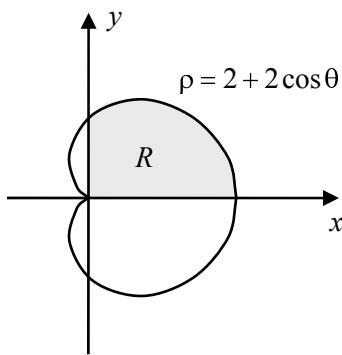


Figura 3

8. Calcule, con cuatro cifras decimales exactas, la integral doble:

$$\iint_R \arctan\left(\frac{y}{x}\right) dxdy$$

donde R es la región del primer cuadrante interior a la lemniscata $\rho^2 = 4 \cos 2\theta$.

9. En el caso particular en que la integral doble está definida sobre una región rectangular:

$$\int_a^b \int_c^d f(x, y) dy dx$$

los algoritmos numéricicos se simplifican considerablemente. Elabore el seudo código del método de Simpson para calcular este tipo especial de integral doble.

Otras lecturas recomendadas

En la deducción del error de truncamiento para el método de Simpson, ha sido omitida la deducción para el error cometido en un intervalo; en “Análisis Matemático” de R. Pastor et al. puede verse una demostración de la fórmula con todos los detalles. En esta misma obra se halla la deducción de la fórmula de Mansion para acotar el error del algoritmo de Simpson.

Acerca de las fórmulas de Newton – Cotes de orden superior puede consultarse el libro “Computing Methods” de Berezin y Zhidkov el cual contiene las deducciones generales y los coeficientes tabulados hasta para el caso de la fórmula de orden 11.

El método de Gauss, como se ha visto, descansa en las interesantes propiedades de los polinomios de Legendre, de las cuales solo se han mencionado las dos que eran necesarias para justificar el método. El lector interesado en profundizar acerca de los polinomios de Legendre, puede consultar “A First Course in Numerical Analysis” de A. Ralston. En cuanto a la deducción de la fórmula del error de truncamiento en el método de Gauss, puede consultarse la obra de Berezin y Zhidkov antes citada.

El método de Romberg se basa completamente en la fórmula de Richardson la cual fue demostrada en el texto. Sin embargo, no se probó que, cuando esta fórmula se aplica a un método cuyo error de truncamiento tiene orden p , ella da resultados con error de truncamiento $p + 2$. Esta demostración puede encontrarse en “Elementary Numerical Analysis” de S. D. Conte.

La interpolación mediante splines puede ser utilizada para construir fórmulas de integración numérica; este interesante tema está tratado en “Computer Methods for Mathematical Computations” de G. E. Forsythe et al.

Otro enfoque al problema de la integración numérica es mediante los métodos de Montecarlo, los cuales, aunque poseen una baja eficiencia numérica, constituyen una elegante forma de calcular integrales, sobre todo las dobles, triples y de mayor dimensión. El libro “Computacional Mathematics” de Demidovich y Maron incluye todo un capítulo al tratamiento de esta técnica.

Por razones de espacio, el cálculo numérico de las integrales impropias de primera y de segunda especie no se ha abordado. El lector interesado en el tema encontrará una buena referencia en “An Introduction to Numerical Analysis” de K. E. Atkinson.

Principales ideas del capítulo

- El método analítico para calcular integrales definidas está limitado a funciones que posean primitiva en términos de funciones elementales, lo cual en muchos casos importantes, no sucede. El procedimiento analítico tampoco es aplicable en los casos en que el integrando solo puede evaluarse de modo experimental.
- El método de los trapecios se basa en la idea de dividir el intervalo de integración en n subintervalos de amplitud h mediante un conjunto de puntos $\{a = x_0, x_1, x_2, \dots, x_n = b\}$ y descomponer la integral en n integrales, cada una de las cuales posee un intervalo de integración pequeño de longitud h , y se aproxima sustituyendo el integrando, para ese intervalo, por el polinomio interpolador de primer grado que determinan los puntos extremos del intervalo.
- El método de los trapecios consiste en la fórmula aproximada:

$$\int_a^b f(x)dx \approx h \left(\frac{1}{2}y_0 + y_1 + y_2 + \dots + y_{n-1} + \frac{1}{2}y_n \right)$$

- Si en el intervalo $[a, b]$ $f(x)$ es continua y posee primera y segunda derivadas continuas, entonces existe en $[a, b]$ al menos un número c tal que, el error de truncamiento de la fórmula de los trapecios para la integral de $f(x)$ en $[a, b]$ viene dado por:

$$R = -\frac{b-a}{12} h^2 f''(c)$$

- Para valores pequeños de h el error de truncamiento en el método de los trapecios se puede aproximar por la fórmula asintótica:

$$R \approx -\frac{f'(b) - f'(a)}{12} h^2$$

- Si el error de truncamiento de un método numérico de integración es, asintóticamente:

$$R_h = Ch^p$$

entonces se puede estimar, para valores pequeños de h , mediante:

$$R_h \approx \frac{I_h - I_{2h}}{2^p - 1}$$

donde I_h e I_{2h} son los resultados obtenidos al calcular la integral mediante el método numérico con pasos h y $2h$ respectivamente.

- Al calcular numéricamente una integral, se introduce, además del error de truncamiento, un error debido a los errores de los datos. Este error está acotado por el producto del error en cada dato por la longitud del intervalo de integración y es, en general, muy pequeño en comparación con el error de truncamiento.
- El método de Simpson, o método de las paráolas, se basa en la idea de dividir el intervalo de integración en n (número par) subintervalos de amplitud h mediante un conjunto de puntos $\{a = x_0, x_1, x_2, \dots, x_n = b\}$ y descomponer la integral en $n/2$ integrales, cada una de las cuales posee un intervalo de integración pequeño de longitud $2h$, y se approxima sustituyendo el integrando, para ese intervalo, por el polinomio interpolador de segundo grado que determinan los puntos extremos y el punto central del intervalo.
- La fórmula aproximada del método de Simpson es:

$$\int_a^b f(x)dx \approx \frac{1}{3}h(E + 4I + 2P)$$

donde:

$$E = y_0 + y_n \quad (\text{Suma de las ordenadas en los extremos})$$

$$I = y_1 + y_3 + \dots + y_{n-1} \quad (\text{Suma de las ordenadas de índice impar})$$

$$P = y_2 + y_4 + \dots + y_{n-2} \quad (\text{Suma de las ordenadas de índice par})$$

y su error de truncamiento viene dado por: $R = -\frac{b-a}{180} h^4 f^{(4)}(c)$

Como el error de truncamiento es de orden h^4 se puede estimar como:

$$R_h \approx \frac{I_h - I_{2h}}{15}$$

- La fórmula de Gauss para m puntos viene dada por:

$$\int_a^b f(x)dx \approx \frac{b-a}{2} \sum_{i=1}^m A_i f(x_i)$$

donde: $x_i = \frac{a+b}{2} + \frac{b-a}{2} t_i$ para $i = 1, 2, \dots, m$

t_1, t_2, \dots, t_m son los ceros del polinomio de Legendre de grado m . Los valores de t_1, t_2, \dots, t_m y A_1, A_2, \dots, A_m vienen tabulados y son tales que la fórmula de Gauss da resultados exactos para polinomios hasta de grado $2m - 1$.

- Si I_1 es el resultado obtenido a calcular la integral de $f(x)$ en $[a, b]$ mediante el método de Gauss con m puntos e I_2 es la suma los resultados obtenidos al aplicar el mismo método en los intervalos $[a, \omega]$ y $[\omega, b]$ (ω es el punto medio del intervalo $[a, b]$), entonces el error de truncamiento de I_2 puede estimarse como:

$$R_2 \approx \frac{I_2 - I_1}{4^m - 1}$$

- El método de Romberg es un algoritmo iterativo para calcular integrales. Se basa en la idea de calcular por filas una tabla de aproximaciones. La fila número k se calcula a partir de la fila $k - 1$, mediante la fórmula de Richardson:

$$I_k^m = \frac{4^m I_k^{m-1} - I_{k-1}^{m-1}}{4^m - 1} \quad m = 1, 2, \dots, k$$

excepto el primer elemento de la fila: I_k^0 que se calcula mediante el método de los trapecios con $2^k n$ subintervalos (n es un valor inicial que se selecciona como un número entero pequeño, por ejemplo 4). Cada fila posee un elemento más que la anterior. El error de truncamiento del último elemento de la fila está acotado por la diferencia (en valor absoluto) con el último elemento de la fila anterior.

- Algunos de los métodos numéricos de integración (Trapecios, Simpson, Gauss) aproximan una integral definida mediante una suma y, por tanto, como una integral doble se calcula como una integral simple iterada, los métodos anteriores permiten calcular integrales dobles reduciéndolas a una suma de sumas.
- El método de Gauss de m puntos para el cálculo de integrales dobles se puede resumir en la fórmula:

$$\int_a^b \int_g^h f(x, y) dy dx \approx \frac{b-a}{2} \sum_{i=1}^m A_i \left[\frac{h(x_i) - g(x_i)}{2} \sum_{j=1}^m A_j f(x_i, y_j) \right]$$

donde:

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} t_i \quad \text{para } i = 1, 2, \dots, m$$

y, para cada x_i , $y_j = \frac{g(x_i) + h(x_i)}{2} + \frac{h(x_i) - g(x_i)}{2} t_j$ para $j = 1, 2, \dots, m$

- El método de Simpson con n sub intervalos para integrales dobles se resume como:

$$\int_a^b \int_{g(x)}^{h(x)} f(x, y) dy dx \approx \frac{h}{3} \sum_{i=0}^n A_i \left[\frac{h_i}{3} \sum_{j=0}^n A_j f(x_i, y_j) \right] \quad (14)$$

donde: $h = \frac{b-a}{n}$, $x_i = a + hi$ para $i = 0, 1, 2, \dots, n$

y, para cada x_i : $h_i = \frac{h(x_i) - g(x_i)}{n}$; $y_j = g(x_i) + jh_i$ para $j = 0, 1, 2, \dots, n$

En todos los casos $\{A_i\} = \{1, 4, 2, 4, 2, \dots, 2, 4, 1\}$

- Al aplicar el método de Simpson al cálculo de integrales dobles, el error de truncamiento se puede estimar por el procedimiento de doble cálculo, ya que el error es del orden de h^4 , mediante la fórmula:

$$R_h \approx \frac{I_h - I_{2h}}{15}$$

Auto examen

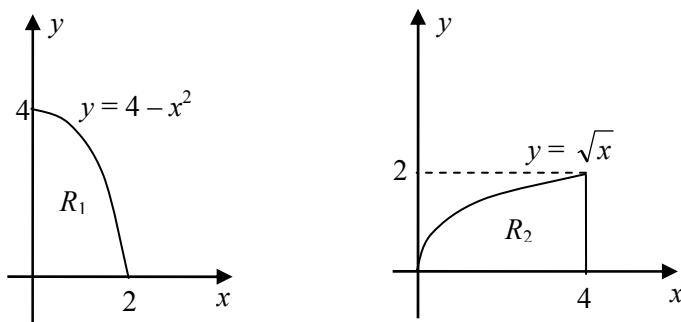
1. Cite dos razones que limitan el método analítico de calcular integrales definidas.
2. Explique la idea geométrica del método de los trapecios y diga por qué recibe este nombre.
3. ¿Qué significa la afirmación de que el error de truncamiento del método de los trapecios es de orden 2 y el de Simpson es de orden 4?
4. ¿Por qué en el método de Gauss el integrando se evalúa precisamente en los ceros de un polinomio de Legendre?
5. Pruebe la fórmula de extrapolación de Richardson y explique qué relación guarda la misma con el método de integración de Romberg.
6. La figura que sigue muestra dos regiones idénticas en cuanto a su forma. La diferencia solo está en cuanto a su posición respecto a los ejes. Si se utiliza la región R_1 el área viene dada por la integral

$$\int_0^2 (4 - x^2) dx$$

que, al ser calculada por el método de Simpson con cuatro sub intervalos da como resultado: 5,33333333, es decir, $16/3$, el valor exacto de la integral. ¿Por qué en este caso el método de Simpson obtiene un resultado exacto? Si se utiliza la región R_2 , hay que calcular la integral:

$$\int_0^4 \sqrt{x} dx$$

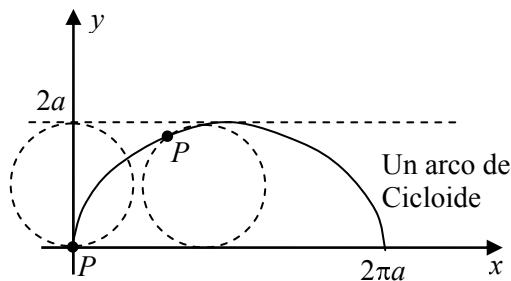
Cuando esta integral se calcula mediante el método de Simpson con 64 sub intervalos se obtiene el resultado 5,33206482 que solo posee dos cifras decimales exactas. ¿Por qué en este caso el método de Simpson conduce a un resultado tan inexacto? ¿Puede usted establecer alguna recomendación de carácter general?



7. Cuando una circunferencia rota (sin deslizamiento) a lo largo de una recta, cualquiera de los puntos de la circunferencia describe una curva periódica, llamada "cicloide". Si la circunferencia tiene radio a , rota sobre el eje x y el punto P inicialmente se halla en el origen, como en la figura que sigue, el punto P describe la cicloide de ecuación paramétrica:

$$x = a(t - \sin t); \quad y = a(1 - \cos t)$$

Si un auto viaja desde La Habana hasta Matanzas (100 km) ¿Qué longitud recorre una pequeña piedra que se encuentra en la superficie de uno de los neumáticos? Dé la respuesta con un error menor que un metro.



8. Calcule mediante el método de Simpson, con cuatro cifras decimales exactas, la integral doble:

$$\iint_R e^{-y} dy dx$$

donde $R = \{(x, y): \frac{\pi}{2} \leq x \leq \pi; \cos x \leq y \leq \sin x\}$

9. Respecto al algoritmo en seudo código que se muestra a continuación diga qué método de integración se está empleando y complete los espacios entre signos de interrogación.

```

 $h := \frac{b-a}{n}$ 
Suma := ?
for i = 1 to n - 1
    x := a + ih
    if i mod 2 = 1 then
        Suma := Suma + ?
    else

```

Suma := *Suma* + ζ ?
end

Integral := ζ ?

El resultado aproximado es *Integral*

Terminar

CAPÍTULO 6

Matemática Numérica, 2da Edición

Manuel Álvarez, Alfredo Guerra, Rogelio Lau

OPTIMIZACIÓN NUMÉRICA

Objetivos

Al finalizar el estudio y la ejercitación de este capítulo, el lector debe ser capaz de:

- Describir qué se entiende por optimizar una función.
- Decidir cuándo conviene utilizar un método analítico y cuándo uno numérico para realizar el proceso de optimización.
- Describir los conceptos de función unimodal de una variable y función linealmente unimodal de varias variables.
- Describir los métodos de búsqueda simultánea y búsqueda secuencial, uniforme y acelerada.
- Describir los métodos de búsqueda unidimensional en un intervalo: método de bisección y método de la sección áurea.
- Utilizar en forma combinada los métodos de búsqueda abierta y en un intervalo para optimizar funciones de una variable.
- Describir los métodos de optimización para funciones de varias variables: Búsqueda por coordenadas, gradiente, Powell y simplex secuencial, estableciendo las ventajas y desventajas de cada uno de ellos.
- Explicar el algoritmo en seudo código de cada uno de los métodos de optimización estudiados e ilustrarlos con ejemplos manuales.
- Utilizar los métodos de optimización multidimensional para hallar máximos y mínimos de funciones de varias variables independientes.
- Modelar problemas que conducen a problemas de optimización de una variable con o sin restricciones y problemas de varias variables sin restricciones.

6.1 Introducción

Problemas de optimización

La palabra “optimización” se utiliza en la Matemática para indicar aquellos procedimientos que permiten encontrar valores de las variables independientes para los cuales una cierta función toma su mayor (o menor) valor. En ocasiones el conjunto donde están definidas las variables independientes está restringido en una cierta región del dominio de definición.

Una enorme cantidad de problemas prácticos son problemas de optimización o pueden ser expresados como un problema de optimización, de los cuales existe una gran variedad de acuerdo con las siguientes posibilidades:

- Las características de la función a optimizar, por ejemplo, si es una función lineal o si es cuadrática o si es unimodal o si, por el contrario, no exhibe ninguna de estas propiedades.
- Si se conoce la expresión analítica de la función o si solamente puede ser evaluada de modo experimental.
- El número de variables independientes, que va desde una hasta cientos de variables, sobre todo en problemas relacionados con grandes sistemas económicos.

- Si existen restricciones en los valores de las variables independientes o, por el contrario, se trata de un problema de variables libres.
- En caso de que existan restricciones, el tipo de ellas resulta del mayor interés. Por ejemplo, si todas las restricciones son inecuaciones lineales, existen técnicas especiales para el problema.
- Si las variables solo pueden tomar valores enteros, las técnicas de solución son muy diferentes de las que se emplean cuando se trata de variables reales.

En cursos anteriores se estudiaron métodos analíticos para hallar los extremos de funciones de una y de más variables. Todos estos métodos partían de las siguientes hipótesis:

- Se conoce la expresión analítica de la función que se desea optimizar.
- La función a optimizar es continua y derivable una o más veces.
- Las ecuaciones o sistemas de ecuaciones (en general, no lineales) que se obtiene al igualar las derivadas a cero pueden ser resueltas.

A diferencia de esos casos, aquí se estudiarán técnicas numéricas, las cuales descansarán, fundamentalmente, en la operación de evaluar la función a optimizar.

Primero se analizará el caso más simple de las funciones que dependen de una sola variable independiente. Los procedimientos más generales, para funciones de n variables, casi siempre consisten en resolver secuencias de problemas unidimensionales; de aquí la importancia del caso unidimensional. Por otra parte, algunos problemas prácticos son a veces unidimensionales. A continuación se muestran varios ejemplos.

Ejemplo 1

En una empresa de producción agrícola desean estimar la cantidad óptima de fertilizante por hectárea que debe emplearse para cierto cultivo. Si la eficiencia en dicho cultivo se mide por P : relación entre el valor de la producción obtenida y el valor de los recursos gastados, es bastante evidente que P es una función de la cantidad x de fertilizante por hectárea. Se supone que otros importantes parámetros no han de cambiar.

Como se ve, se trata de un problema de optimización donde la función $P(x)$ depende de una sola variable real $x \geq 0$. La expresión analítica de $P(x)$ se desconoce.

Ejemplo 2

La función Gamma se define para $x > 0$ mediante la integral de Euler:

$$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1} dt$$

En el intervalo $[1; 2]$ esta función posee un punto de mínimo pero hallarlo mediante un método analítico no es posible por la dificultad de hallar la derivada respecto a x .

Ejemplo 3

Se necesita resolver el siguiente sistema de ecuaciones para el diseño de un circuito:

$$\begin{cases} a_1 V_1(T_1) - a_2 V_2(T_1) = V_0 & (1) \\ a_1 V_1(T_2) - a_2 V_2(T_2) = V_0 & (2) \\ a_1 V_1(T_3) - a_2 V_2(T_3) = V_0 & (3) \end{cases}$$

donde V_0 es un voltaje conocido y V_1 y V_2 son ciertos voltajes que dependen de la temperatura T . A las temperaturas T_1 , T_2 , T_3 el voltaje V_2 se conoce, pero no V_1 . Tampoco se conocen los coeficientes a_1 y a_2 . Se sabe además que el voltaje V_1 depende implícitamente de la temperatura según la ecuación:

$$\alpha T \left(1 + \frac{V_1(T)}{V_{ar}} \right) = KT^\eta \exp\left(\frac{V_1(T) - V_{go}}{T^{\frac{k}{q}}} \right) \quad (4)$$

donde K , V_{ar} , V_{go} , η , k y q son constantes conocidas. Sin embargo la constante α se desconoce. Se trata pues de un sistema de seis ecuaciones con seis incógnitas.

Ecuaciones: (1), (2), (3) y (4) para T_1 , T_2 y T_3 .

Incógnitas: $V_1(T_1)$, $V_1(T_2)$, $V_1(T_3)$, a_1 , a_2 y α .

Aunque es un sistema no lineal de ecuaciones, puede plantearse como un problema de optimización unidimensional. Para ello, se define la función $f(\alpha)$ de la siguiente forma:

Dado un valor de α :

1. Se hace $T = T_1$ en la ecuación (4) y se halla numéricamente su única incógnita $V_1(T_1)$ (se puede hacer, por ejemplo, mediante el método iterativo simple, despejando la incógnita $V_1(T_1)$ que aparece en el argumento de la exponencial)
2. Lo mismo para $T = T_2$. Se halla $V_1(T_2)$.
3. Lo mismo para $T = T_3$. Se halla $V_1(T_3)$.
4. Se resuelve el sistema formado por las ecuaciones (2) y (3) para hallar a_1 y a_2 (una vez que se han determinado $V_1(T_2)$ y $V_1(T_3)$, este sistema es lineal en a_1 y a_2).
5. Se define la función: $f(\alpha) = [a_1 V_1(T_1) - a_2 V_2(T_1) - V_0]^2$

El valor mínimo de $f(\alpha)$ es cero y se alcanza para un α que satisface la ecuación (1). O sea, el valor de α que minimiza a f conduce, mediante los pasos 1, 2, 3, 4 y 5, a la obtención de la solución del sistema de ecuaciones. Obsérvese, sin embargo, que a pesar de que existen expresiones analíticas que definen sin ambigüedades la función $f(\alpha)$, el tratamiento analítico de dicha función sería muy complicado.

Funciones unimodales de una variable

De manera no muy formal, puede decirse que una función unimodal de una variable es aquella que posee en su conjunto de definición un solo punto de extremo relativo. En la figura 1 se observan dos funciones a) y b) unimodales y una c) que no lo es.

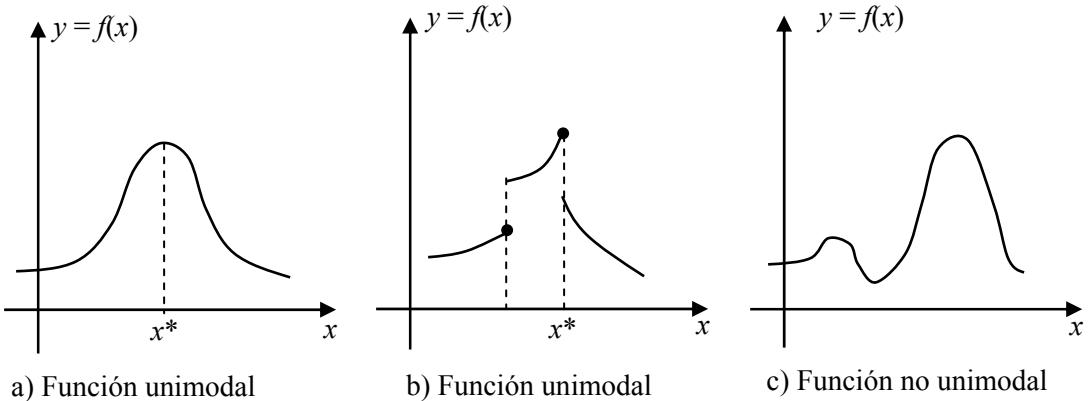


Figura 1

Una definición un poco más precisa es la siguiente:

Definición:

La función $f(x)$ (vea la figura 2) es unimodal con máximo si existe en su dominio un x^* tal que, si x_1 y x_2 pertenecen al dominio se cumple que:

$$\text{Si } x_1 < x_2 < x^* \text{ entonces } f(x_1) < f(x_2)$$

$$\text{Si } x^* < x_1 < x_2 \text{ entonces } f(x_1) > f(x_2)$$

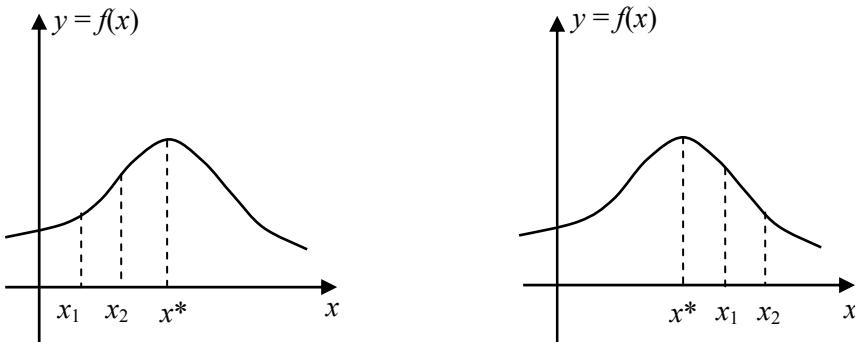


Figura 2

Nótese que la definición anterior simplemente establece que f es creciente a la izquierda de x^* y decreciente a la derecha de x^* . De manera evidente se puede modificar la definición para el caso de una función unimodal con mínimo. Sin embargo, no es necesario estudiar por separado los casos de máximo y de mínimo porque todo problema de minimización se puede reducir a uno de maximización (y viceversa). Por ejemplo, basta tener en cuenta que si $f(x)$ es unimodal con punto de máximo en x^* entonces $-f(x)$ es unimodal con punto de mínimo en x^* .

En todos los métodos que se estudiarán se aplica el siguiente resultado el cual, por ser tan elemental, no se ha querido llamarlo teorema.

Propiedad básica de la optimización unidimensional

Sea $f(x)$ una función unimodal con máximo en x^* y sean x_1 y x_2 dos valores de su dominio. Sean $y_1 = f(x_1)$ y $y_2 = f(x_2)$. Entonces (vea la figura 3):

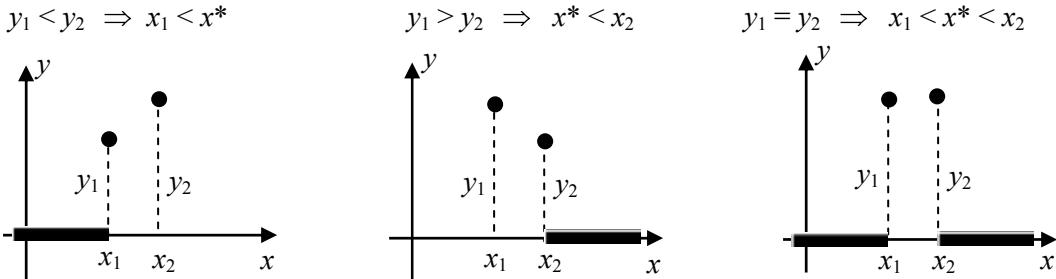


Figura 3

En la figura 3 se ha marcado con líneas gruesas la porción del dominio de $f(x)$ que puede ser desechara, ya que con seguridad el punto de óptimo no se encuentra en ella.

La justificación es bastante simple. Por ejemplo, en el primer caso, si se admitiera que el punto de máximo x^* está a la izquierda de x_1 entonces en el intervalo $[x_1, x_2]$ $f(x)$ tendría que ser decreciente y en ese caso debería ser $y_1 > y_2$; como esto contradice la hipótesis, entonces x^* no está a la izquierda sino a la derecha de x_1 . De modo similar se prueba para los otros dos casos. Se deja al lector como ejercicio que enuncie y demuestre la propiedad básica para una función unimodal con mínimo.

Clasificación de los métodos de búsqueda unidimensional

En la optimización numérica se acostumbra llamar experimento al proceso mediante el cual se conoce el valor de la función f para un determinado valor de x . Está claro que en muchas ocasiones la evaluación no necesita realizarse experimentalmente, sin embargo, esta forma de hablar es clásica. No obstante, aun cuando los experimentos se realicen evaluando una expresión algebraica en un programa de computadora o mediante un largo proceso práctico, el costo del algoritmo de optimización está directamente relacionado a la cantidad de experimentos que sea necesario realizar pues, en todo caso, los experimentos son la parte del algoritmo en que más recursos se consume.

Una primera clasificación entre los métodos de optimización numérica para funciones de una variable se basa en la simultaneidad o no de los experimentos. Aquellos algoritmos en los que los experimentos se realizan todos a un tiempo, se llaman de *búsqueda simultánea* mientras que los que siguen la estrategia de realizar un experimento después que han sido analizados los resultados de los experimentos anteriores, se llaman *métodos secuenciales*. Los algoritmos secuenciales requieren menos experimentos que los simultáneos, ya que en estos se realizan evaluaciones sin analizar resultados anteriores que podrían haber descartado la necesidad de muchos de estos experimentos. Sin embargo, en muchas ocasiones la naturaleza del problema investigado obliga a utilizar los métodos simultáneos. Por ejemplo, para determinar el punto de un territorio donde la temperatura fue mínima anoche, no queda más alternativa que medir la temperatura simultáneamente en una buena muestra de puntos; para determinar la cantidad óptima de fertilizante a utilizar en un cultivo sería demasiado lento el procedimiento secuencial, pues en este caso cada experimento requiere meses.

En la búsqueda secuencial hay dos situaciones bastante diferentes que requieren también procedimientos diferentes: los casos en que el punto de máximo x^* se halla dentro de un intervalo conocido (*búsqueda con restricciones*) y los casos en que no se sabe nada acerca de x^* (*búsqueda sin restricciones*). Es usual que ambos tipos de problemas se combinen y que se comience utilizando un método de búsqueda sin restricciones que termina dando un intervalo donde se halla x^* y, a continuación, aplicar procedimientos de búsqueda con restricciones para hallar x^* con una precisión aceptable.

En general, la búsqueda simultánea no tiene mucho que discutir y se le dedicará muy poca atención. La búsqueda secuencial es más rica en resultados y se estudiará con algún detalle posteriormente.

6.2 Optimización unidimensional sin restricciones

Búsqueda simultánea

Como ya se dijo, a menos que sea imprescindible, debe evitarse utilizar la búsqueda simultánea. En general los experimentos se sitúan uniformemente espaciados aunque el conocimiento de las características de la función a optimizar podría aconsejar concentrar más experimentos en determinadas regiones donde es más probable la existencia del punto x^* .

Si los experimentos se realizan en los valores $x_1 < x_2 < \dots < x_n$ con resultados y_1, y_2, \dots, y_n y se obtiene $y_k = \max \{y_i\}$ con $1 < k < n$, entonces (vea la figura 1), de acuerdo con la propiedad básica de las funciones unimodales, el punto de máximo se encuentra entre x_{k-1} y x_{k+1} , o sea:

$$x_{k-1} < x^* < x_{k+1}$$

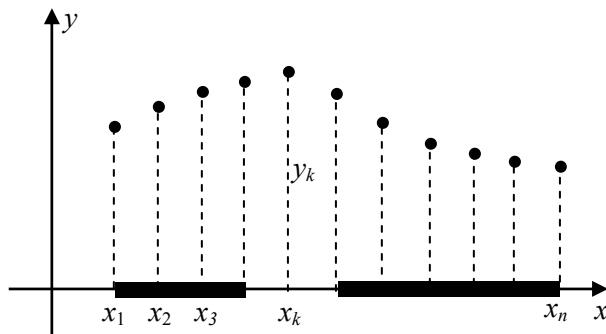


Figura 1

La cantidad de experimentos simultáneos depende del costo de cada experimento y de la exactitud con que se necesite determinar x^* .

Búsqueda secuencial uniforme

Sea $y = f(x)$ una función unimodal con un punto de máximo en x^* , de la cual solo se supone que está definida en un intervalo que incluye a x^* y que puede ser evaluada en cualquier punto de dicho intervalo. El método más simple para hallar x^* es la búsqueda secuencial uniforme.

Sea x_0 el punto donde se comenzará la búsqueda. La selección de x_0 requiere, por lo general, algún conocimiento previo de la función que se desea optimizar y a veces algún tipo de “sospecha” basada en análisis realizados a la misma o en la naturaleza del fenómeno con que se está tratando. En cualquier caso, como es natural, se seleccionará x_0 lo más próximo posible de x^* .

Sea s un número real distinto de cero (puede ser positivo o negativo) al que se llamará *paso*. El método de búsqueda secuencial uniforme consiste, simplemente, en generar la sucesión de valores:

$$x_i = x_0 + is \quad i = 0, 1, 2, 3, \dots$$

y obtener para cada uno de ellos la imagen de f , esto es:

$$y_i = f(x_i)$$

El proceso se detiene tan pronto se obtiene un y_k tal que $y_{k-1} > y_k$, es decir:

$$y_0 < y_1 < y_2 < \dots < y_{k-1} > y_k$$

y puede entonces asegurarse, se acuerda con la propiedad básica, que x^* se encuentra entre x_{k-2} y x_k . Teniendo en cuenta el signo de s , se puede escribir:

$$\text{si } s > 0, \quad x_{k-2} < x^* < x_k$$

$$\text{si } s < 0, \quad x_k < x^* < x_{k-2}$$

En la figura 2 se muestra geométricamente la situación en el caso en que $s > 0$:

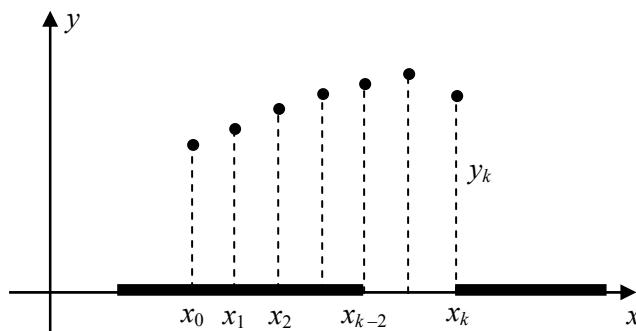


Figura 2

Algoritmo en seudo código

El algoritmo que sigue permite determinar un intervalo en el que se encuentra el punto x^* de máximo de una función $f(x)$ unimodal. Se supone que la función está definida en todo el intervalo en que se realiza la búsqueda. El algoritmo utiliza como datos: la función $f(x)$, el número x_0 a partir del cual se comienza la búsqueda y el paso s , que puede ser positivo o negativo.

```

 $k := 0$ 
 $y_0 := f(x_0)$ 
repeat
     $k := k + 1$ 
     $x_k := x_{k-1} + s$ 
     $y_k := f(x_k)$ 
until  $y_k < y_{k-1}$ 
 $x^*$  se encuentra entre  $x_{k-2}$  y  $x_k$ 
Terminar

```

Es de notar que, en cualquier caso, la región final en que queda acotado x^* tiene como amplitud $2|s|$; para obtener x^* con una exactitud adecuada se debe tomar un paso s pequeño pero, por otra parte, si s es pequeño se requiere de muchos experimentos para cubrir la distancia $|x^* - x_0|$. Hay varias formas de resolver esta contradicción, que se analizan a continuación.

- Busqueda secuencial por etapas

Esta estrategia consiste en tomar un paso s inicial grande (del orden de magnitud de la distancia $|x^* - x_0|$); una vez determinado el intervalo $x_{k-2} < x^* < x_k$ (aquí y en lo que sigue se está suponiendo $s > 0$; en el caso $s < 0$, los cambios a realizar son muy simples) tomar $x_0 = x_{k-2}$ y comenzar una nueva búsqueda con un paso s menor. Puede probarse que el valor esperado del número total de experimentos para encontrar el óptimo con una cierta precisión dada, se hace mínimo cuando, en cada etapa, el paso se reduce en e veces ($e = 2.71828\dots$). Se le propone al lector aventajado que lo demuestre.

- Búsqueda en un intervalo

En este caso, una vez determinado el intervalo, se aplican los métodos de búsqueda con restricciones, que se verán más adelante y que son más eficientes.

- Búsqueda secuencial acelerada

Esta estrategia resulta apropiada cuando no se tiene una idea clara del tamaño del problema, o sea, de la distancia $|x^* - x_0|$. Consiste en lo siguiente: En cada paso del algoritmo mientras no se cumpla la condición de parada, el paso se duplica en valor; de esta manera, aun cuando inicialmente s fuera demasiado pequeño, pronto toma valores suficientemente grandes para alcanzar a x^* en no muchas iteraciones. A este algoritmo se le llama *búsqueda secuencial acelerada* (en la figura 3 se ve geométricamente la idea). A continuación se muestra el algoritmo, que solo posee una ligera modificación respecto al anterior:

```

 $k := 0$ 
 $y_0 := f(x_0)$ 
repeat
     $k := k + 1$ 
     $x_k := x_{k-1} + s$ 
     $y_k := f(x_k)$ 
     $s := 2s$ 
until  $y_k < y_{k-1}$ 
 $x^*$  se encuentra entre  $x_{k-2}$  y  $x_k$ 
Terminar

```

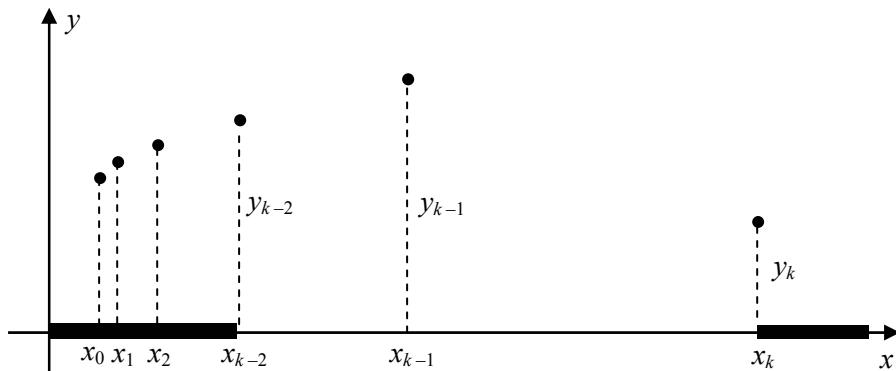


Figura 3

También se aprecia cómo, por lo general, el intervalo final en que queda encerrado x^* es apreciablemente grande; esto, sin embargo, puede resolverse posteriormente utilizando alguno de los métodos de búsqueda en un intervalo, los cuales son más eficientes.

Ejemplo 1

Halle el punto de máximo de la función $f(x) = \sin x$ tomando $x_0 = 0$ y $s = 0,1$ mediante búsqueda secuencial a) uniforme, b) acelerada.

Solución:

- a) En la tabla 1 se muestran parcialmente los experimentos efectuados. Se obtiene

$$1,5 < x^* < 1,7$$

(el valor de x^* es $\pi/2 = 1,5708\dots$) y se requieren 17 iteraciones.

i	x_i	$f(x_i)$
0	0,0	0,0000
1	0,1	0,0998
2	0,2	0,1987
:	:	:
14	1,4	0,9854
15	1,5	0,9975
16	1,6	0,9996
17	1,7	0,9917

Tabla 1

- b) Utilizando búsqueda secuencial acelerada, en 5 iteraciones se obtiene el resultado:

$$0,7 < x^* < 3,1$$

Los resultados se encuentran en la tabla 2.

i	x_i	$f(x_i)$
0	0,0	0,0000
1	0,1	0,0998
2	0,3	0,2955
3	0,7	0,6442
4	1,5	0,9975
5	3,1	0,0416

Tabla 2

Es de destacar el hecho de que el método de búsqueda acelerado llega más rápidamente a la solución pero da como resultado un intervalo final mayor.

Ejemplo 2

En este ejemplo se ilustra cómo el tamaño inicial del paso s_0 influye en el número de experimentos y en la amplitud del intervalo final $x_{k-2} < x^* < x_k$. Considere la función

$$f(x) = 100 - \frac{(x - 1000)^2}{100}$$

que es unimodal y alcanza su valor máximo en $x^* = 1000$. A partir de $x_0 = 0$ utilice el método de búsqueda secuencial acelerada para encontrar el punto de máximo con pasos iniciales de 1, 5, 10, 50, 100 y 900. Determine en cada caso el número de experimentos y la amplitud de la región donde queda acotado x^* .

Solución:

En la tabla 3 se muestran los resultados obtenidos procediendo de la misma forma que en el ejemplo anterior:

s_0	Iteraciones	Intervalo final	Amplitud
1	11	[511; 2047]	1536
5	9	[635; 2555]	1920
10	8	[630; 2550]	1920
50	5	[350; 1150]	800
100	4	[300; 1500]	1200
900	2	[0; 2700]	2700

Tabla 3

Como se observa, el hecho de comenzar con un x_0 muy alejado de x^* , trae como consecuencia que el intervalo final en que x^* queda encerrado, posee una gran amplitud; sin embargo, la selección del paso inicial no altera en forma significativa esta amplitud. Ahora bien, la cantidad de iteraciones necesarias sí depende fuertemente del valor inicial de s . Puede probarse que este comportamiento no es casual y que, en general, el valor de s que minimiza la cantidad de

iteraciones en el algoritmo de búsqueda acelerada es, aproximadamente, igual al *tamaño del problema*, que es la distancia entre x_0 y x^* .

Selección del sentido de la búsqueda

En algunos casos aparece una dificultad adicional, cuando se desconoce si x^* es mayor o menor que x_0 . Considérese que este es el caso para la función unimodal con máximo $f(x)$ y que se toma un paso $s > 0$. Si $f(x_1) > f(x_0)$, es obvio que puede descartarse la región $(-\infty, x_0)$, así que ya se sabe que $x^* > x_0$. Sin embargo, si $f(x_1) < f(x_0)$ la región descartable es (x_1, ∞) , o sea que $-\infty < x^* < x_1$, por tanto, pudiera ser $x^* > x_0$ o $x^* < x_0$ (ver figura 4). Observe que el resultado $f(x_1) < f(x_0)$ puede deberse a que se está buscando x^* en el sentido equivocado (a) o a que se ha tomado un paso tan grande que se ha sobrepasado x^* en la primera iteración (b). Hay dos modos de proceder: buscar de nuevo en el mismo sentido con un paso más pequeño a partir de x_0 o comenzar una nueva búsqueda a partir de x_1 pero en el sentido opuesto.

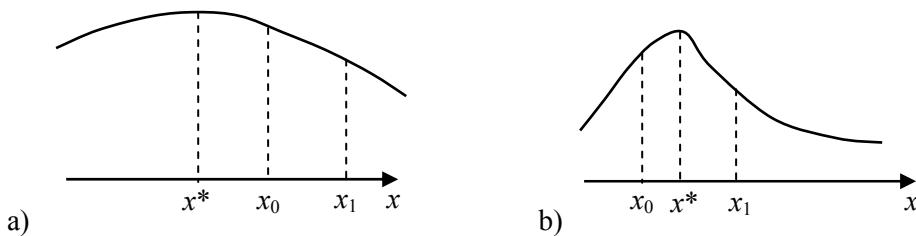


Figura 4

Ejemplo 3

Halle el punto de máximo de la función $f(x) = x \operatorname{sen} x$ que se encuentra entre 0 y π . Tome $x_0 = 2$ y utilice búsqueda secuencia uniforme con paso inicial 0,1. Reduzca el paso a la cuarta parte en cada etapa hasta acotar x^* en una región de amplitud 0,001.

Solución:

$$\text{Etapa 1: } x_0 = 2 \quad s = 0,1$$

x	$x \operatorname{sen} x$	
2,0	1,81859	
2,1	1,81273	Descartado $(2,1, \infty)$

$$\text{Etapa 2: } x_0 = 2,1 \quad s = -0,025$$

x	$x \operatorname{sen} x$	
2,1	1,81273	
2,075	1,81678	
2,050	1,81909	
2,025	1,81968	
2,000	1,81959	$2,000 < x^* < 2,050$

Etapa 3: $x_0 = 2$ $s = 0,00625$

x	$x \operatorname{sen} x$
2,00000	1,81859
2,00625	1,81902
2,01250	1,81935
2,01875	1,81957
2,02500	1,81968
2,03125	1,819697
2,03750	1,819602 $2,025 < x^* < 2,0375$

Etapa 4: $x_0 = 2,025$ $s = 0,0015625$

x	$x \operatorname{sen} x$
2,025	1,819687
2,0265625	1,819699
2,028125	1,819705
2,0296875	1,819704 $2,02656 < x^* < 2,02969$

Etapa 5: $x_0 = 2,0265$ $s = 0,0004$

x	$x \operatorname{sen} x$
2,0265	1,8196988
2,0269	1,8197011
2,0273	1,8197029
2,0277	1,8197042
2,0281	1,819705156
2,0285	1,819705651
2,0289	1,819705714
2,0293	1,819705344 $2,0285 < x^* < 2,0293$

Ejercicios

Algunos de los ejercicios que siguen son muy laboriosos para realizar los cálculos a mano. Se supone que posea programas computacionales, preferiblemente confeccionados por usted, para ayudarle en el trabajo. De no ser así, en los ejercicios que requieran mucho trabajo manual, determine los resultados con menos cifras exactas que las exigidas.

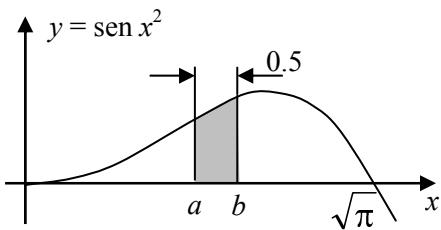
1. Halle el punto de máximo de la función $f(x) = 2x^2 e^{-3x}$ con un error menor que 0,001. Compruebe su respuesta hallando los ceros de la derivada de $f(x)$ mediante algún método numérico.

2. Halle con tres cifras decimales exactas el punto de mínimo de la función $f(x) = 3xe^x + e^{-x}$. Compruebe su respuesta hallando los ceros de la derivada de $f(x)$ mediante algún método numérico.
3. Utilice un asistente matemático, por ejemplo, Derive, para evaluar la función Gamma:

$$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1} dt \quad x > 0$$

y halle, con error menor que 0,001 el punto de mínimo que posee esta función.

4. En el ejemplo 4 de la sección 2.1 fue planteado el problema de hallar las dimensiones de un recipiente cilíndrico de 1000 cm^3 de capacidad utilizando la mínima cantidad de material y teniendo en cuenta que las piezas que lo conforman deben tener un sobrante de 0,25 cm para poder realizar las soldaduras (vea la figura 4 de ese capítulo). Allí fue resuelto hallando los ceros de la función derivada. Resuélvalo ahora mediante las técnicas estudiadas de optimización numérica con error menor que 0,1 mm.
5. Determine los valores de a y de b , ambos entre 0 y $\sqrt{\pi}$, de manera que el área sombreada de la figura sea máxima. Obtenga la respuesta con dos cifras decimales exactas.



6. Elabore un algoritmo en seudo código que permita resolver el problema anterior.

6.3 Optimización en un intervalo

Si en un problema de optimización unidimensional se conoce un intervalo $[a, b]$ donde se encuentra el punto de máximo x^* de la función unimodal $f(x)$, el problema puede resolverse de modo más efectivo que la búsqueda secuencial. El intervalo $[a, b]$ puede estar dado desde un inicio a partir de consideraciones físicas, económicas, etc., o puede haberse llegado a él a partir de una etapa previa de búsqueda secuencial. En cualquier caso pueden aplicarse los métodos que siguen.

El método de bisección

Sea f una función unimodal definida en $[a, b]$ y con un punto de máximo x^* en ese intervalo. Tómense dos puntos experimentales x_1 y x_2 muy próximos entre sí y a ambos lados del centro del intervalo $[a, b]$. Para concretar, sea δ la distancia entre x_1 y x_2 , entonces, como se aprecia en la figura 1:

$$x_1 = \frac{a+b}{2} - \frac{\delta}{2} \quad \text{y} \quad x_2 = \frac{a+b}{2} + \frac{\delta}{2}$$

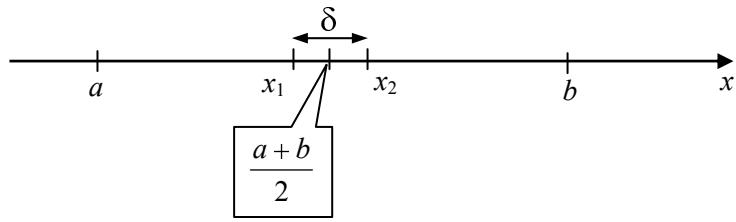


Figura 1

Como se verá de inmediato, debe procurarse que δ sea lo menor posible, sin embargo, debe ser razonablemente grande de modo que, al hallar $f(x_1)$ y $f(x_2)$ estos valores sean distinguibles uno del otro; esto es aún más importante si $f(x)$ se evalúa en forma experimental, pues entonces hay que tomar en cuenta los inevitables errores de observación, de medición, etc.

Sean

$$y_1 = f(x_1) \quad y \quad y_2 = f(x_2)$$

Como consecuencia inmediata de la propiedad básica de la optimización unidimensional, como muestra la figura 2, se tiene que:

$$y_1 < y_2 \Rightarrow x_1 \leq x^* \leq b$$

$$y_1 > y_2 \Rightarrow a \leq x^* \leq x_2$$

$$y_1 = y_2 \Rightarrow x_1 \leq x^* \leq x_2$$

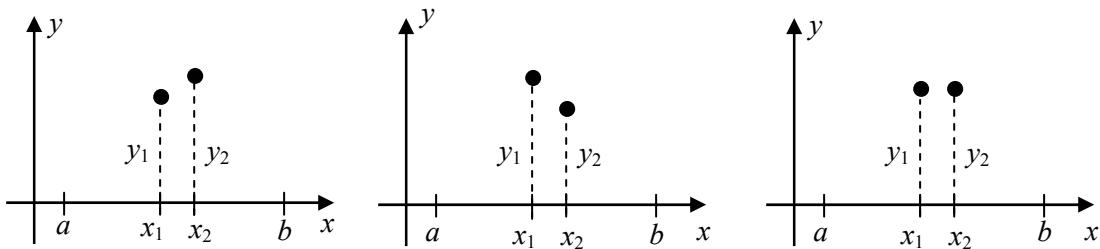


Figura 2

La convergencia del método de bisección es muy fácil de analizar. Si consideramos que δ es un número muy pequeño en relación con la amplitud del intervalo, puede suponerse que en cada iteración (que requiere dos evaluaciones de f) la longitud del intervalo de búsqueda se reduce a la mitad. Más formalmente, llamando:

L_n : Amplitud del intervalo de búsqueda después de n evaluaciones (n par)

$$L_0 = b - a$$

entonces:

$$L_n = \frac{L_0}{2^{n/2}}$$

lo cual prueba que L_n converge hacia 0 cuando n tiende hacia infinito. En la práctica, sin embargo, L_n no puede reducirse a valores menores que δ .

La relación L_n/L_0 se utiliza con frecuencia para medir la eficiencia de los métodos de optimización en un intervalo, pues indica cuánto se puede reducir la región de incertidumbre mediante n experimentos. En el método de bisección esta relación es:

$$\frac{L_n}{L_0} = \frac{1}{2^{n/2}}$$

Algoritmo en seudo código

El algoritmo que sigue describe el método de optimización de bisección del intervalo. Se supone que la función $f(x)$ es unimodal con máximo en el intervalo $[a, b]$ y que está definida en todos los puntos de ese intervalo. El algoritmo da como resultado un intervalo cerrado de amplitud menor que un número especificado ε y que contiene al punto de máximo. Los datos que se requiere son: la función $f(x)$, el intervalo $[a, b]$ en que se encuentra inicialmente el punto de máximo de la función, la distancia δ que separará a los puntos experimentales y la tolerancia ε , que determina la amplitud del intervalo final.

```

repeat
     $x_1 := \frac{a+b}{2} - \frac{\delta}{2}$ 
     $x_2 := \frac{a+b}{2} + \frac{\delta}{2}$ 
     $y_1 := f(x_1)$ 
     $y_2 := f(x_2)$ 
    if  $y_1 < y_2$  then  $a := x_1$  else  $b := x_2$ 
     $L := b - a$ 
until  $L < \varepsilon$ 
El punto de máximo  $x^*$  se halla en  $[a, b]$ 
Terminar

```

Ejemplo 1

La función $f(x) = x \operatorname{sen} x$ posee un punto x^* de máximo en el intervalo $[2; 2,1]$.

- ¿Cuántos experimentos se necesitará para determinar x^* en una región de amplitud 0,001 mediante el método de bisección?
- Hállelo.

Solución:

a) En este caso se tiene $L_0 = 0,1$

y se desea obtener un $L_n = 0,001$
 Como $L_n = \frac{L_0}{2^{n/2}}$

se tiene: $2^{n/2} = \frac{L_0}{L_n} = \frac{0,1}{0,001} = 100$

y, aplicando logaritmos:

$$\frac{n}{2} = \frac{\ln 100}{\ln 2}$$

o sea:

$$n = 2 \frac{\ln 100}{\ln 2} = 13.288$$

Luego, se requiere de $n = 14$ experimentos o, lo que es igual, 7 iteraciones del método. Compare con el ejemplo 3 de la sección anterior, y notará que en las etapas 2, 3, 4 y 5, para lograr la misma reducción mediante búsqueda secuencial uniforme, se requirió de 18 experimentos diferentes.

- b) Como se aspira a obtener un intervalo final de amplitud 0,001, se utilizará una distancia de separación $\delta = 0,0001$ (diez veces menor que el intervalo final). En la tabla 1 se muestran todos los resultados.

iteración	a	b	x_1	x_2	$f(x_1)$	$f(x_2)$
0	2,00000	2,10000	2,04995	2,05005	1,8190957	> 1,8190900
1	2,00000	2,05005	2,02498	2,02508	1,8196865	< 1,8196875
2	2,02508	2,05005	2,03751	2,03761	1,8196020	> 1,8195996
3	2,02508	2,03761	2,03129	2,03139	1,8196971	> 1,8196964
4	2,02508	2,03139	2,02819	2,02829	1,8197053	< 1,8197054
5	2,02819	2,03139	2,02974	2,02984	1,8197044	> 1,8197042
6	2,02819	2,02984	2,02897	2,02907	1,8197057	> 1,8197056
7	2,02819	2,02907				

Tabla 1

Completadas siete iteraciones (14 experimentos) se ha llegado a un intervalo con amplitud 0,00088 menor que la tolerancia 0,001 requerida. Por tanto, la respuesta buscada es:

$$2,02819 < x^* < 2,02907$$

Nótese que, si se desea resolver un problema de optimización unidimensional (y esto es también válido en el caso multidimensional) con una precisión ε , entonces la distancia δ debe ser mucho menor que ε pues en las iteraciones finales la amplitud de la región de búsqueda se va aproximando a ε . En la práctica puede tomarse $\delta = 0.1\varepsilon$, lo cual significa que en problemas de optimización se debe tener cuidado de no exigir una precisión ε exageradamente pequeña, ya que esto puede conducir a valores de δ más pequeños que lo razonable.

Ejemplo 2

Suponga que los especialistas en agua subterránea han determinado que, en una región bajo estudio, el nivel del manto freático solo varía significativamente cuando se mide en puntos alejados entre sí por lo menos 100 metros. ¿Puede pedirse entonces que se determine el punto en que el manto freático tiene su mayor altura con una precisión de 10 metros?

Solución:

Obviamente, si los experimentos no pueden hacerse a menos de 100 metros, no tiene mucho sentido pedir una precisión menor de 200 metros.

El método de Fibonacci

Es necesario aclarar que este método no es debido a Fibonacci. Se le llama así porque hace uso de los números de Fibonacci, matemático del siglo XIV que los definió para resolver un problema que nada tuvo que ver con la optimización.

Si se analiza con cuidado el método de optimización mediante bisección, se verá (observe la figura 3) que, en cada iteración, el intervalo de búsqueda que se selecciona contiene un punto experimental que, por estar muy próximo a uno de los extremos, no puede ser aprovechado en la próxima iteración:

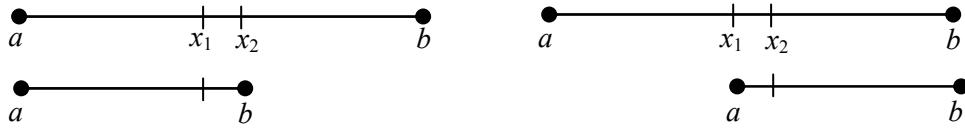


Figura 3

En el método de Fibonacci se trata de tomar los experimentos x_1 y x_2 más alejados entre sí, de modo que en la próxima iteración, el punto que quede dentro del nuevo intervalo de búsqueda, quede suficientemente cercano al centro como para que pueda ser utilizado como un punto experimental. De esta manera, cada iteración solo requerirá de un nuevo experimento, aunque a cambio de ello el intervalo no se reducirá a la mitad en cada paso sino a una proporción mayor.

En la figura 4 se muestra esta idea gráficamente:

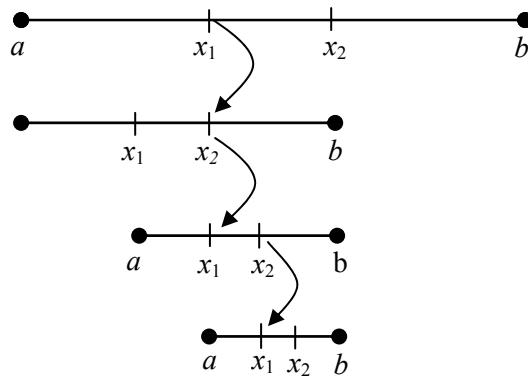


Figura 4

Para que el algoritmo sea sencillo, hay que procurar que las posiciones de x_1 y x_2 en cada iteración se puedan hallar aplicando una regla simple.

Una posible solución se basa en los números de Fibonacci, definidos como:

$$\begin{aligned}
 F_0 &= 1 \\
 F_1 &= 1 \\
 F_n &= F_{n-1} + F_{n-2} \quad \text{para } n = 2, 3, 4, \dots
 \end{aligned}$$

de donde se pueden fácilmente generar sus valores:

n	0	1	2	3	4	5	6	7	8	...
F_n	1	1	2	3	5	8	13	21	34	...

Si se supone que la amplitud $b - a$ del intervalo inicial coincide con el número de Fibonacci F_N se tiene:

$$L = F_N$$

Como se muestra en la figura 5, el intervalo $[a, b]$ se puede descomponer en dos subintervalos de longitudes F_{N-1} y F_{N-2} :

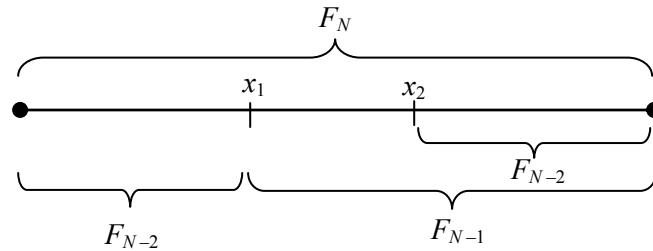


Figura 5

En la próxima iteración, como se observa en la figura 6, al aplicar la propiedad básica, el nuevo intervalo tendrá amplitud F_{N-1} y poseerá un punto experimental a una distancia F_{N-2} de uno de sus extremos:

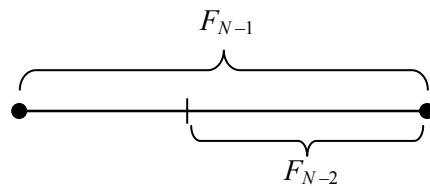


Figura 6

Obviamente (vea la figura 7), la distancia al otro extremo es F_{N-3} , pues $F_{N-1} = F_{N-2} + F_{N-3}$,

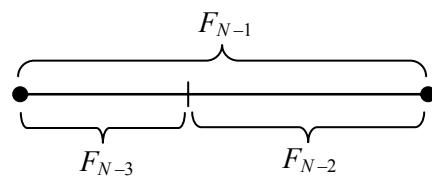


Figura 7

Falta solamente situar otro punto experimental simétricamente, como en la figura 8:

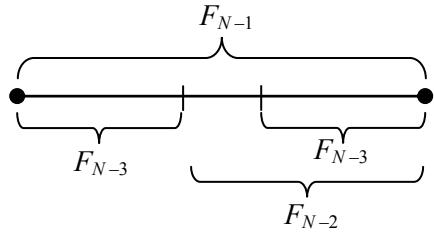


Figura 8

y el procedimiento puede repetirse hasta llegar a un intervalo de longitud $F_2 = 2$, como se ve en la figura 9:

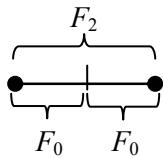


Figura 9

en el cual x_1 y x_2 coinciden. En la práctica, el procedimiento antes descrito tiene el inconveniente de que, desde que se comienza el proceso iterativo hay que seleccionar el número concreto de Fibonacci, F_N con el que se va a comenzar, del cual depende la reducción que habrá de conseguirse en el proceso, que será exactamente F_0/F_N . Esto puede evitarse con la siguiente modificación que, además, hace innecesario generar la sucesión de números de Fibonacci. Esta modificación se denomina, por razones que se verá enseguida, método de la *sección de oro*.

El método de la sección de oro

Considérense los números D_n definidos como una generalización de los de Fibonacci pero donde la condición $F_0 = F_1 = 1$ se sustituye por el requerimiento de que la sucesión formada sea geométrica. Esto es:

$$D_n = D_{n-1} + D_{n-2} \quad n = 2, 3, 4, \dots \quad (1)$$

$$D_n = r D_{n-1} \quad n = 1, 2, 3, \dots \quad (2)$$

El valor de r puede encontrarse fácilmente. Dividiendo en la ecuación (1) por D_{n-1} :

$$\frac{D_n}{D_{n-1}} = 1 + \frac{D_{n-2}}{D_{n-1}}$$

y como, según (2): $\frac{D_n}{D_{n-1}} = r$ y $\frac{D_{n-2}}{D_{n-1}} = \frac{1}{r}$ resulta:

$$r = 1 + \frac{1}{r} \quad (3)$$

de donde

$$r^2 - r - 1 = 0$$

cuya raíz positiva es:

$$r = \frac{1 + \sqrt{5}}{2} = 1,618034$$

$$G = \frac{1}{r} = r - 1 = 0,618034$$

es la famosa *sección de oro* que era conocida desde la antigua Grecia y que posee propiedades sumamente interesantes.

Como los números D_n satisfacen la condición de que $D_n = r D_{n-1}$, sus valores quedan determinados al asignarle a uno cualquiera de ellos un valor deseado. Puesto que los números D_n poseen la misma propiedad que caracteriza a los de Fibonacci:

$$D_n = D_{n-1} + D_{n-2}$$

ellos pueden usarse en el algoritmo anterior. Por otra parte poseen la ventaja de que cada D_n se obtiene directamente a partir de su antecedente:

$$D_n = r D_{n-1}$$

o de su sucesor:

$$D_{n-1} = G D_n$$

En la figura 10 se muestra la manera en que se obtienen los puntos experimentales en el algoritmo de la *sección de oro*

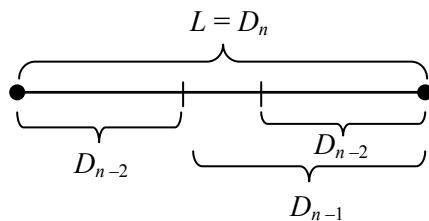


Figura 10

Pero como

$$D_{n-1} = G D_n = G L = 0,618034 L$$

y

$$D_{n-2} = L - D_{n-1} = L - G L = (1 - G)L = 0,381966 L$$

todo lo que se requiere en cada iteración es calcular el valor

$$(b - a)(1 - G)$$

para ubicar el nuevo punto experimental. Entre los números de Fibonacci y los D_n existe una interesante relación. En efecto, como

$$F_n = F_{n-1} + F_{n-2}$$

se tiene:

$$\frac{F_n}{F_{n-1}} = 1 + \frac{F_{n-2}}{F_{n-1}}$$

y, si n tiende hacia infinito

$$\lim_{n \rightarrow \infty} \frac{F_n}{F_{n-1}} = 1 + \lim_{n \rightarrow \infty} \frac{F_{n-2}}{F_{n-1}}$$

si se llama F al límite: $\lim_{n \rightarrow \infty} \frac{F_n}{F_{n-1}}$ entonces se tiene:

$$F = 1 + \frac{1}{F}$$

Esta ecuación es idéntica a la (3) y, por tanto, $F = r$. Esto es, cuando n tiende hacia infinito, los números de Fibonacci se comportan como los D_n .

Como el método de Fibonacci y el de la sección de oro poseen idénticas características, en general se prefiere esta segunda variante, más fácil de implementar. Para concretar, se da a continuación el algoritmo de la sección de oro.

Algoritmo en seudo código del método de la sección de oro

El algoritmo que sigue describe el método de optimización de la sección de oro. Se supone que la función $f(x)$ es unimodal con máximo en el intervalo $[a, b]$ y que está definida en todos los puntos de ese intervalo. El algoritmo da como resultado un intervalo cerrado de amplitud menor que un número especificado ε y que contiene al punto de máximo. Los datos que se requiere son: la función $f(x)$, el intervalo $[a, b]$ en que se encuentra inicialmente el punto de máximo de la función y la tolerancia ε , que determina la amplitud del intervalo final.

```
L := b - a
Factor := 1 - G = 0,381966
x1 := a + Factor · L
x2 := b - Factor · L
y1 := f(x1)
y2 := f(x2)
repeat
    if y1 < y2 then
        a := x1
        x1 := x2
        y1 := y2
        L := b - a
        x2 := b - Factor · L
        y2 := f(x2)
    else
        b := x2
        x2 := x1
        y2 := y1
        L := b - a
        x1 := a + Factor · L
        y1 := f(x1)
    end
until L < ε
```

El punto de máximo x^* se halla en $[a, b]$
 Terminar

Es fácil determinar la eficiencia del método de la sección áurea. Llamando, como hasta ahora:

Amplitud del intervalo original: $L_0 = b - a$:

Amplitud del intervalo después de n experimentos: L_n

se tiene: $L_n = G L_{n-1} = G^2 L_{n-2} = \dots = G^{n-2} L_2$

ya que, después de los dos primeros experimentos, con cada experimento nuevo la región de búsqueda queda multiplicada por G . Como en la primera iteración se requiere de dos experimentos nuevos,

$$L_2 = G L_0$$

de ahí que:

$$L_n = G^{n-1} L_0$$

Esto es:

$$\frac{L_n}{L_0} = G^{n-1}$$

En la tabla 2 se muestra la relación L_n/L_0 para diferentes valores de n en los métodos de bisección y de la sección áurea. Nótese que con la misma cantidad de experimentos, el método de la sección áurea logra una reducción mucho mayor de la región de búsqueda. La ventaja se ve más clara si se aprecia que en el método de bisección, cada dos nuevos experimentos reducen la región de incertidumbre a la mitad:

$$L_n = 0.5 L_{n-2}$$

pero en la sección áurea, después de la primera iteración

$$L_n = G^2 L_{n-2} = 0.382 L_{n-2}$$

n	L_n/L_0 en bisección	L_n/L_0 la sección áurea
10	0.03125	0.01316
12	0.01562	0.00502
14	0.00781	0.00192
16	0.00390	0.00073
18	0.00195	0.00028
20	0.00098	0.00011
22	0.00049	0.00004
24	0.00024	0.00002
26	0.00012	0.00001

Tabla 2

Ejemplo 3

Se sabe que la función $f(x) = x \operatorname{sen} x$ es unimodal y tiene un punto de máximo x^* en el intervalo $[2; 2,1]$. ¿Cuántos experimentos se necesitará para determinar x^* en una región de amplitud 0,001 usando el método de la sección de oro?

Solución:

$$\begin{aligned}L_0 &= 0,1 \\ \varepsilon &= 0,001\end{aligned}$$

entonces $L_n = G^{n-1}L_0 = G^{n-1}0,1 \leq \varepsilon = 0,001$

de donde: $G^{n-1} \leq \frac{0,001}{0,1} = 0,01$

y, aplicando logaritmos: $(n-1)\ln G \leq \ln 0,01$

Ahora, al dividir por $\ln G$, como es negativo, la desigualdad cambia de sentido:

$$\begin{aligned}n-1 &\geq \frac{\ln(0,01)}{\ln G} \\ n-1 &\geq 9,57\end{aligned}$$

$$n \geq 10,57$$

O sea:

$$n = 11$$

Compare con la solución del ejemplo 1, en el cual se calculó que el método de bisección requeriría 14 experimentos para este mismo problema.

Refinamiento del resultado mediante interpolación

En general, todos los algoritmos tratados hasta aquí concluyen su trabajo dando un intervalo o región de incertidumbre suficientemente pequeña donde se encuentra el punto de máximo. En muchos casos, sin embargo, es conveniente seleccionar un punto de dicho intervalo que se pueda tomar como una aproximación razonable de x^* , sobre todo cuando se usa la optimización unidimensional como procedimiento interno de un algoritmo mucho más amplio de optimización multidimensional.

Una solución que a veces se emplea es tomar el punto medio de la región de incertidumbre. Otra posibilidad usada con frecuencia es la que se verá a continuación.

Sea $f(x)$ una función unimodal con punto de máximo x^* y se suponen conocidos tres valores de x : $x_1 < x_2 < x_3$ tales que $x_1 < x^* < x_3$. Sean $y_1 = f(x_1)$, $y_2 = f(x_2)$ y $y_3 = f(x_3)$. Se pretende hallar el número \bar{x} , abscisa del vértice (vea la figura 11) de la parábola que determinan los tres puntos (x_1, y_1) , (x_2, y_2) y (x_3, y_3) . Sea $y = p(x)$ la ecuación de dicha parábola:

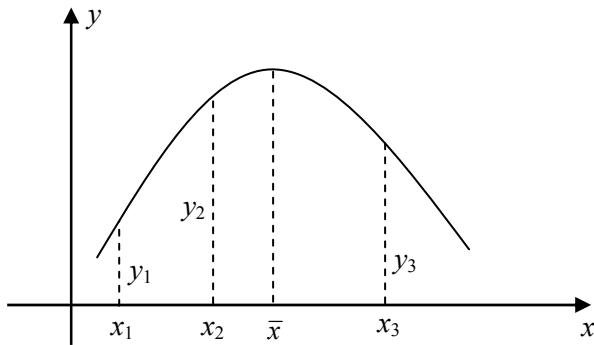


Figura 11

$$p(x) = a x^2 + bx + c$$

El valor \bar{x} anula a $p'(x)$ o sea: $2a\bar{x} + b = 0$

de donde: $\bar{x} = -\frac{b}{2a}$

Los valores de a y b se pueden obtener de diferentes formas, una de ellas es la que sigue. Como la parábola pasa por los tres puntos, se tiene:

$$\begin{cases} ax_1^2 + bx_1 + c = y_1 \\ ax_2^2 + bx_2 + c = y_2 \\ ax_3^2 + bx_3 + c = y_3 \end{cases}$$

Es decir:
$$\begin{cases} a(x_2^2 - x_1^2) + b(x_2 - x_1) = y_2 - y_1 \\ a(x_3^2 - x_2^2) + b(x_3 - x_2) = y_3 - y_2 \end{cases}$$

Si se divide la primera ecuación por $(x_2 - x_1)$, la segunda por $(x_3 - x_2)$ y se le llama

$$f_{12} = \frac{y_2 - y_1}{x_2 - x_1} \quad y \quad f_{23} = \frac{y_3 - y_2}{x_3 - x_2}$$

se obtiene:

$$\begin{cases} (x_1 + x_2)a + b = f_{12} \\ (x_2 + x_3)a + b = f_{23} \end{cases}$$

Ahora, utilizando la notación: $x_{12} = \frac{x_1 + x_2}{2}$ y $x_{23} = \frac{x_2 + x_3}{2}$ y resolviendo para a y b :

$$\bar{x} = \frac{x_{23}f_{12} - x_{12}f_{23}}{f_{12} - f_{23}}$$

$$\text{donde: } x_{12} = \frac{x_1 + x_2}{2} \quad f_{12} = \frac{y_2 - y_1}{x_2 - x_1} \quad x_{23} = \frac{x_2 + x_3}{2} \quad f_{23} = \frac{y_3 - y_2}{x_3 - x_2}$$

Ejemplo 4

Se sabe que la función $f(x) = x \operatorname{sen} x$ es unimodal y posee un punto de máximo en el intervalo $[2; 2,1]$. Tome $x_1 = 2$, $x_2 = 2,05$, $x_3 = 2,1$, determine \bar{x} y compare con $x^* = 2,0287578$.

Solución:

$$\begin{aligned} y_1 &= f(x_1) = 2 \operatorname{sen} 2 = 1,818595 \\ y_2 &= f(x_2) = 2,05 \operatorname{sen} 2,05 = 1,819093 \\ y_3 &= f(x_3) = 2,1 \operatorname{sen} 2,1 = 1,812740 \end{aligned}$$

$$\begin{aligned} x_{12} &= \frac{x_1 + x_2}{2} = 2,025 & f_{12} &= \frac{y_2 - y_1}{x_2 - x_1} = 0,009960 \\ x_{23} &= \frac{x_2 + x_3}{2} = 2,075 & f_{23} &= \frac{y_3 - y_2}{x_3 - x_2} = -1,127064 \end{aligned}$$

$$\text{De donde: } \bar{x} = \frac{x_{23}f_{12} - x_{12}f_{23}}{f_{12} - f_{23}} = 2,028634$$

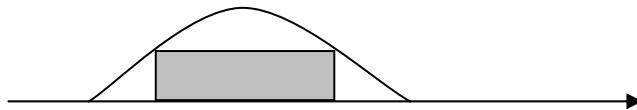
que solamente difiere de x^* en 0,000124. Como se ha visto en los ejemplos anteriores, para lograr un error unas 10 veces mayor que este, los métodos de bisección y de la sección áurea hubieran requerido respectivamente 14 y 11 evaluaciones de la función $f(x)$.

Ejercicios

Algunos de los ejercicios que siguen son muy laboriosos para realizar los cálculos a mano. Se supone que posea programas computacionales, preferiblemente confeccionados por usted, para ayudarle en el trabajo. De no ser así, en los ejercicios que requieran mucho trabajo manual, determine los resultados con menos cifras exactas que las exigidas.

1. Halle el menor punto de máximo positivo de la función $f(x)$ con error menor que 0,001. Primero determine, mediante búsqueda secuencial, un intervalo que contiene al punto buscado y, después, aplique el método de bisección o el de Fibonacci. Compruebe su respuesta hallando los ceros de la derivada de $f(x)$ mediante alguno de los métodos numéricos para resolver ecuaciones.
 - a) $f(x) = e^{-x} \operatorname{sen} x$
 - b) $f(x) = \operatorname{sen} x \cosh x$
 - c) $f(x) = \ln x - e^x$
 - d) $f(x) = \operatorname{sen}(\ln x)$
2. Halle la mínima distancia desde el punto $(3, 2)$ hasta la gráfica de la sinusode $y = \operatorname{sen} x$. Obtenga la respuesta con tres cifras decimales exactas.

3. De todos los triángulos rectángulos de área 5, halle las dimensiones del que posee menor perímetro. De la solución con cuatro cifras decimales exactas.
4. Halle las dimensiones (con tres cifras decimales exactas) del rectángulo de mayor área que se puede inscribir en un arco de sinusoida, como se muestra en la figura:



5. Dados los puntos $P_1 = (1; 1,2)$, $P_2 = (2; 2,1)$ y $P_3 = (3; 3)$, halle la recta $y = mx$ que minimiza la desviación cuadrática:

$$\sum_{i=1}^3 (y_i - mx_i)^2$$

Obtenga la solución con tres cifras decimales exactas.

6. Para hallar el punto de máximo de una función $f(x)$ en un intervalo $[a, b]$ existe un algoritmo que consiste en situar en el intervalo tres puntos distribuidos de manera que dividan al intervalo en cuatro partes iguales; evaluando $f(x)$ en los tres puntos se puede obtener un nuevo intervalo que contiene al punto de máximo y cuya longitud es la mitad de la del intervalo original. Analice la eficiencia de este método en comparación con el de bisección y elabore un algoritmo en seudo código para representarlo.

6.4 Conceptos básicos para la optimización multidimensional

La mayor parte de los problemas de optimización que surgen en la práctica corresponden a funciones que dependen de más de una variable. El problema general de la optimización multidimensional es mucho más complejo que el unidimensional, de manera que en este texto el análisis estará limitado a algunas técnicas numéricas fundamentales para resolver problemas sin restricciones. Aun así, será necesario dedicar esta sección a introducir algunos conceptos matemáticos imprescindibles para la comprensión de los algoritmos que se estudiarán posteriormente. Antes de acometer esta tarea resulta conveniente ver algunos ejemplos de problemas típicos de optimización multidimensional.

Ejemplo 1

Se desea buscar la mejor ubicación, dentro de una amplia región del país, para una estación de bombeo que alimentará a un acueducto. El criterio que se va a usar para determinar la localización óptima, es que la profundidad del manto freático sea mínima, de manera que las bombas subterráneas que se van a instalar requieran motores de menor potencia. Obviamente, la función $f(x, y)$ es la profundidad del manto freático y depende de dos variables: las coordenadas del punto donde se mide la profundidad.

Ejemplo 2

La eficiencia de una planta industrial depende de una enorme cantidad de parámetros; podrían citarse: los parámetros de funcionamiento (velocidad, presión, temperatura, etc.) de cada uno de

los equipos de la planta, años de experiencia de cada operario, nivel de producción de cada artículo que produce la planta, etc. Un complejo problema de optimización consiste en determinar los valores de todas estas variables que permiten alcanzar una máxima eficiencia.

Ejemplo 3

El campo de ajuste de curvas es una gran fuente de problemas de optimización. Si se desea ajustar un modelo lineal en sus parámetros:

$$g(x) = C_1 g_1(x) + C_2 g_2(x) + \dots + C_n g_n(x)$$

donde g_1, g_2, \dots, g_n son funciones conocidas de la variable real x , a un conjunto $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$ de datos, de manera que se minimice la desviación cuadrática:

$$D = \sum_{i=1}^m [g(x_i) - y_i]^2$$

entonces, como ya se sabe, el problema posee una solución analítica que se obtienen sin mucha dificultad resolviendo el sistema lineal de ecuaciones obtenido de hacer cero las derivadas parciales

$$\frac{\partial D}{\partial C_j} \quad j = 1, 2, \dots, n$$

Si el modelo es no lineal, o sea, si es del tipo general:

$$g(x) = F(x, C_1, C_2, \dots, C_n)$$

entonces la función a minimizar es muy complicada y sus derivadas parciales igualadas a cero no conducen a un sistema lineal. El problema de minimizar

$$\sum_{i=1}^m [F(x_i, C_1, C_2, \dots, C_n) - y_i]^2$$

se resuelve usualmente de forma numérica.

Notación y representación gráfica

Para representar más fácilmente una función real de n variables reales:

$$z = f(u_1, u_2, \dots, u_n)$$

se representará por \mathbf{X} el vector:

$$\mathbf{x} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

con lo cual, la función se escribe:

$$z = f(\mathbf{x})$$

Muchas de las técnicas de optimización multidimensional se comprenden mejor en el caso particular $n = 2$, o sea, para funciones de dos variables. La razón es obvia: en este caso la geometría puede ayudar a comprender la estrategia de búsqueda e, incluso, sugerir nuevas ideas y demostraciones. Cuando $n = 2$, se tiene:

$$\mathbf{x} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

y la función $z = f(\mathbf{X})$ puede visualizarse mediante sus curvas de nivel. A modo de ejemplo, la figura 1 muestra la presencia de dos puntos de máximo relativo y un punto de ensilladura en una región del plano (u_1, u_2) .

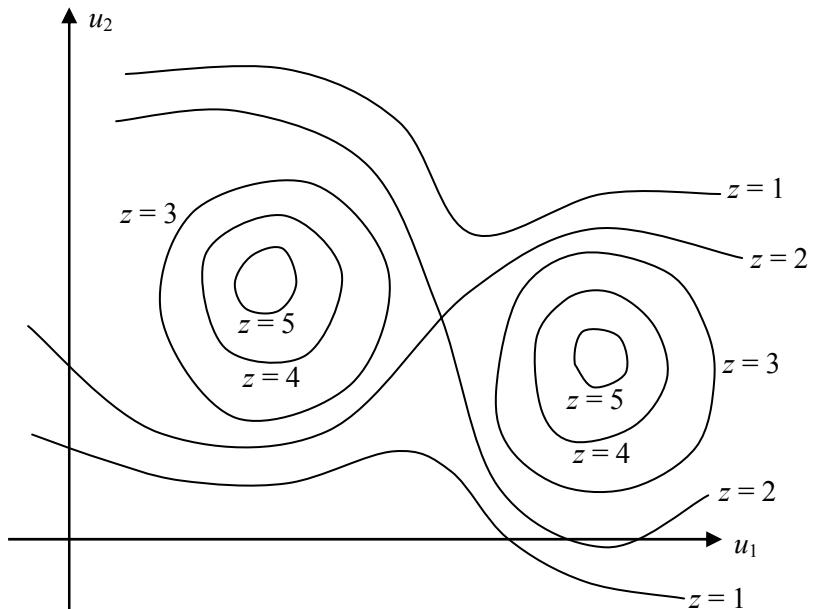


Figura 1

En lo que sigue se usarán con frecuencia funciones de dos variables y su representación mediante curvas de nivel como recurso didáctico para que se comprenda mejor propiedades que resultan necesariamente abstractas para $n > 2$.

Trayectoria lineal en R^n

El concepto de unimodalidad, que tan simple resulta en el caso unidimensional, puede hacerse muy complicado para el caso multidimensional. Por esa razón el análisis se restringirá a un tipo sencillo de unimodalidad, la unimodalidad lineal. Serán necesarias algunas definiciones previas.

Sea \mathbf{x}_0 un vector que representa un punto en R^n y sea \mathbf{v} otro vector de R^n con el que se determina una cierta dirección en ese espacio. Se llamará “recta en R^n por el punto \mathbf{x}_0 y con dirección \mathbf{v} ”, al conjunto de todos los \mathbf{X} que se obtienen de la expresión:

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{v}\lambda$$

al dar a λ todos los valores reales.

La figura 2 ilustra el concepto en el caso $n = 2$. Se dice también que los puntos \mathbf{x} forman una trayectoria lineal. Nótese que, aunque R^n posee dimensión n , la trayectoria lineal es un objeto unidimensional, en el cual cada punto \mathbf{x} está determinado por un solo parámetro real λ ; por esto, a veces se escribe:

$$\mathbf{x}(\lambda) = \mathbf{x}_0 + \lambda \mathbf{v}$$

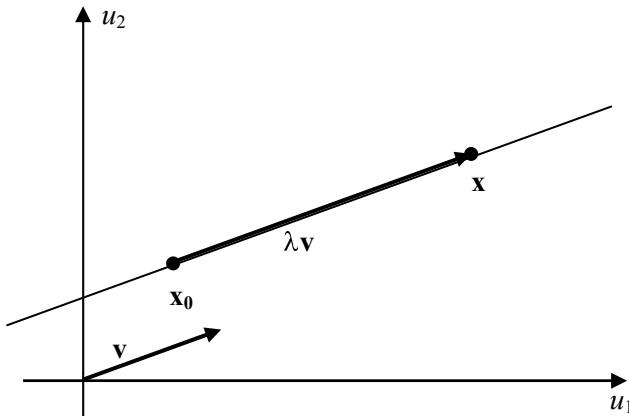


Figura 2

Ejemplo 4

Halle la ecuación de la recta de R^4 determinada por:

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 2 \end{bmatrix} \quad \mathbf{v} = \begin{bmatrix} 2 \\ 1 \\ 1 \\ -2 \end{bmatrix}$$

Solución:

$$\mathbf{x} = \mathbf{x}_0 + \lambda \mathbf{v} = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 2 \end{bmatrix} + \lambda \begin{bmatrix} 2 \\ 1 \\ 1 \\ -2 \end{bmatrix}$$

que se puede expresar como: $\mathbf{x} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 1+2\lambda \\ -1+\lambda \\ \lambda \\ 2-2\lambda \end{bmatrix}$

Función linealmente unimodal

Sea $z = f(\mathbf{x})$ una función real de n variables reales, o sea, $z \in R$, $\mathbf{x} \in R^n$. Si \mathbf{x} se restringe a que tome valores en una trayectoria lineal

$$\mathbf{x} = \mathbf{x}_0 + \lambda \mathbf{v}$$

entonces

$$z = f(\mathbf{x}_0 + \lambda \mathbf{v})$$

y como la única variable es λ , podemos decir que

$$z = F(\lambda)$$

es decir, una función unidimensional, en la cual el concepto de unimodalidad es muy sencillo.

Definición:

Se dice que $z = f(\mathbf{x})$ es linealmente unimodal con máximo en la región R , si ella es unimodal con máximo sobre cualquier trayectoria recta contenida en R .

■

En la figura 3 se muestran las curvas de nivel de una función de dos variables y una trayectoria recta y en la figura 4, el perfil de la función z a lo largo de la trayectoria recta. Si este comportamiento es así para todas las trayectorias rectas, entonces $z = f(\mathbf{x})$ es linealmente unimodal con máximo.

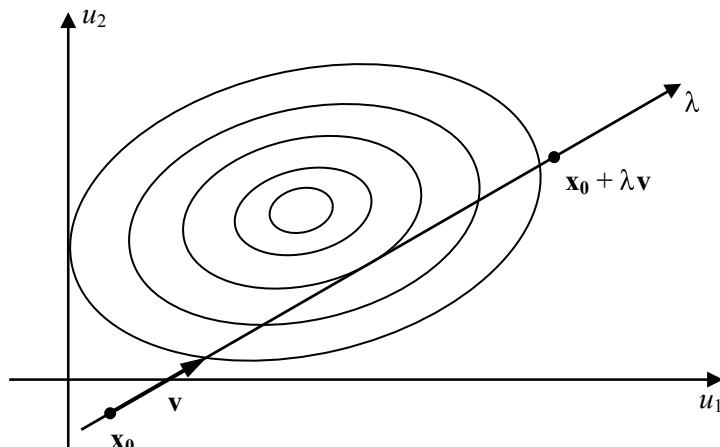


Figura 3

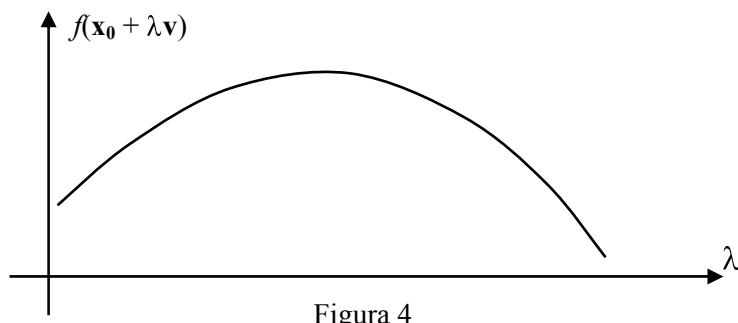


Figura 4

Funciones cuadráticas de n variables

Una clase importante de funciones linealmente unimodales son las funciones cuadráticas con punto de máximo.

Definición:

Se dice que la función $f(\mathbf{x})$ es cuadrática si la misma se puede expresar como:

$$\begin{aligned} f(\mathbf{x}) = & \frac{1}{2} \left[a_{11}u_1^2 + a_{12}u_1u_2 + \dots + a_{1n}u_1u_n + \right. \\ & + a_{21}u_2u_1 + a_{22}u_2^2 + \dots + a_{2n}u_2u_n + \dots \\ & \dots + a_{n1}u_nu_1 + a_{n2}u_nu_2 + \dots + a_{nn}u_n^2 \left. \right] \\ & - [b_1u_1 + b_2u_2 + \dots + b_nu_n] + c \end{aligned}$$

El factor $\frac{1}{2}$ común a todos los términos cuadráticos, así como el signo menos común a todos los términos lineales, persiguen el objetivo de obtener posteriormente expresiones más sencillas. Nótese que todos los términos rectangulares, esto es, del tipo u_iu_j ($i \neq j$) aparecen repetidos pues

$$a_{ij}u_iu_j \text{ y } a_{ji}u_ju_i$$

son semejantes y, de hecho, se podrían haber escrito como

$$(a_{ij} + a_{ji}) u_iu_j = k_{ij}u_iu_j$$

Sin embargo, a los efectos de introducir posteriormente la notación matricial, es preferible mantener la duplicidad, así que se conservarán los términos semejantes $a_{ij}u_iu_j$ y $a_{ji}u_ju_i$

Como, a los efectos de la optimización de $f(\mathbf{x})$, el término c carece de importancia, ya que

$$f(\mathbf{x}) \text{ y } f(\mathbf{x}) + c$$

poseen idénticos puntos de extremo, en todo lo que sigue, se supondrá que $c = 0$.

Las funciones cuadráticas pueden ser representadas matricialmente en forma muy cómoda. Para ello se introduce la matriz \mathbf{A} : coeficientes de los términos cuadráticos; y \mathbf{b} : coeficientes de los términos lineales:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

El producto $\frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x}$ al ser desarrollado resulta:

$$\frac{1}{2} \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\begin{aligned} &= \frac{1}{2} \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} a_{11}u_1 + a_{12}u_2 + \dots + a_{1n}u_n \\ a_{21}u_1 + a_{22}u_2 + \dots + a_{2n}u_n \\ \vdots \\ a_{n1}u_1 + a_{n2}u_2 + \dots + a_{nn}u_n \end{bmatrix} \\ &= \frac{1}{2} [a_{11}u_1^2 + a_{12}u_1u_2 + \dots + a_{1n}u_1u_n + \\ &\quad + a_{21}u_2u_1 + a_{22}u_2^2 + \dots + a_{2n}u_2u_n + \dots \\ &\quad \dots + a_{n1}u_nu_1 + a_{n2}u_nu_2 + \dots + a_{nn}u_n^2] \end{aligned}$$

que son los términos cuadráticos de $f(\mathbf{x})$. Por otra parte,

$$\mathbf{b}^T \mathbf{x} = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = b_1u_1 + b_2u_2 + \dots + b_nu_n$$

son los términos lineales de $f(\mathbf{x})$. Así que:

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$$

Como se recordará, los términos cuadráticos de tipo rectangular aparecen cada uno dos veces: $a_{ij}u_iu_j$ y $a_{ji}u_ju_i$. Puesto que lo importante es que la suma

$$a_{ij} + a_{ji}$$

tenga un valor determinado, no hay pérdida de generalidad si se supone que

$$a_{ij} = a_{ji} \quad \text{para } i \neq j$$

de modo que

$$a_{ij} + a_{ji} = 2a_{ij} = 2a_{ji}$$

El tomar $a_{ij} = a_{ji}$ hace que la matriz \mathbf{A} sea simétrica.

A modo de resumen se establecerá como teorema la propiedad que se acaba de probar.

Teorema:

Una función cuadrática cualquiera f de n variables independientes se puede expresar como

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$$

donde \mathbf{A} es una matriz $n \times n$ simétrica
y \mathbf{b} es una matriz $n \times 1$

■

Matrices positivas definidas y negativas definidas

Dentro de las matrices simétricas, existe un tipo especial de matriz, llamadas *definidas*, que juegan un papel importante en el análisis de las funciones cuadráticas.

Definición:

Sea $\mathbf{A}_{n \times n}$ una matriz simétrica. Si para todo vector no nulo

$$\mathbf{x} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

se cumple que $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$, entonces la matriz \mathbf{A} se llama positiva definida. Si para todo \mathbf{x} no nulo se tiene $\mathbf{x}^T \mathbf{A} \mathbf{x} < 0$, la matriz \mathbf{A} se llama negativa definida.

■

Ejemplo 5

Halle qué condiciones deben cumplir a , b y c para que la matriz simétrica de orden dos:

$$\mathbf{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

sea positiva definida.

Solución:

Sea $\mathbf{x} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ entonces:

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = [u_1 \ u_2] \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} > 0$$

$$= au_1^2 + 2bu_1u_2 + cu_2^2 > 0$$

De aquí resulta que tiene que ser $a > 0$ pues de lo contrario, bastaría tomar $u_2 = 0$ y $u_1 \neq 0$ para obtener un resultado negativo. Por razones similares es obvio que tiene también que ser $c > 0$. Pero esto no basta; completando cuadrados con los dos primeros términos de la expresión anterior:

$$= au_1^2 + 2bu_1u_2 + \frac{b^2u_2^2}{a} - \frac{b^2u_2^2}{a} + cu_2^2 > 0$$

$$= \left(\sqrt{au_1} + \frac{bu_2}{\sqrt{a}} \right)^2 - \frac{b^2u_2^2}{a} + cu_2^2 > 0$$

$$= \left(\sqrt{au_1} + \frac{bu_2}{\sqrt{a}} \right)^2 + \left(-\frac{b^2}{a} + c \right) u_2^2 > 0$$

$$= \left(\sqrt{au_1} + \frac{bu_2}{\sqrt{a}} \right)^2 + \left(\frac{ac - b^2}{a} \right) u_2^2 > 0$$

y para que esto sea cierto es necesario y suficiente que, además de la condiciones $a > 0$ y $c > 0$, sea

$$ac - b^2 > 0$$

■

Las matrices positivas definidas y negativas definidas poseen una gran importancia a la hora de caracterizar a una función cuadrática. Para ello es fundamental el siguiente teorema.

Teorema:

Si \mathbf{A} es positiva (negativa) definida la función cuadrática

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$$

posee un punto \mathbf{x}^* de mínimo (máximo) y se cumple que $\mathbf{A}\mathbf{x}^* = \mathbf{b}$.

Demostración:

Solo se probará el caso en que \mathbf{A} es positiva definida; el otro se demuestra de modo similar.

Sea \mathbf{x}^* tal que

$$\mathbf{A}\mathbf{x}^* = \mathbf{b}$$

(Puede probarse que toda matriz positiva definida es no singular, de ahí que la condición $\mathbf{A}\mathbf{x}^* = \mathbf{b}$ se satisface para uno y solo un \mathbf{x}^*).

Sea $\mathbf{x} \neq \mathbf{x}^*$

$$f(\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x} - \frac{1}{2} \mathbf{x}^{*T} \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}^*$$

Pero como $\mathbf{b}^T = (\mathbf{A}\mathbf{x}^*)^T = \mathbf{x}^{*T} \mathbf{A}^T = \mathbf{x}^{*T} \mathbf{A}$

$$f(\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{x}^{*T} \mathbf{A} \mathbf{x} - \frac{1}{2} \mathbf{x}^{*T} \mathbf{A} \mathbf{x}^* + \mathbf{x}^{*T} \mathbf{A} \mathbf{x}^*$$

$$= \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{x}^{*T} \mathbf{A} \mathbf{x} + \frac{1}{2} \mathbf{x}^{*T} \mathbf{A} \mathbf{x}^*$$

$$\begin{aligned}
&= \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \frac{1}{2} \mathbf{x}^{*T} \mathbf{A} \mathbf{x} - \frac{1}{2} \mathbf{x}^{*T} \mathbf{A} \mathbf{x} + \frac{1}{2} \mathbf{x}^{*T} \mathbf{A} \mathbf{x}^* \\
&= \frac{1}{2} (\mathbf{x}^T - \mathbf{x}^{*T}) \mathbf{A} \mathbf{x} - \frac{1}{2} \mathbf{x}^{*T} \mathbf{A} (\mathbf{x} - \mathbf{x}^*) \\
&= \frac{1}{2} (\mathbf{x} - \mathbf{x}^*)^T \mathbf{A} \mathbf{x} - \frac{1}{2} \mathbf{x}^{*T} \mathbf{A} (\mathbf{x} - \mathbf{x}^*)
\end{aligned}$$

Ambos sumandos de esta expresión son escalares (matrices de 1x1) así que el primer sumando es igual que su traspuesta:

$$\begin{aligned}
f(\mathbf{x}) - f(\mathbf{x}^*) &= \frac{1}{2} (\mathbf{A} \mathbf{x})^T (\mathbf{x} - \mathbf{x}^*) - \frac{1}{2} \mathbf{x}^{*T} \mathbf{A} (\mathbf{x} - \mathbf{x}^*) \\
&= \frac{1}{2} \mathbf{x}^T \mathbf{A} (\mathbf{x} - \mathbf{x}^*) - \frac{1}{2} \mathbf{x}^{*T} \mathbf{A} (\mathbf{x} - \mathbf{x}^*) \\
&= \frac{1}{2} (\mathbf{x}^T - \mathbf{x}^{*T}) \mathbf{A} (\mathbf{x} - \mathbf{x}^*) \\
&= \frac{1}{2} (\mathbf{x} - \mathbf{x}^*)^T \mathbf{A} (\mathbf{x} - \mathbf{x}^*) > 0
\end{aligned}$$

por ser \mathbf{A} positiva definida y $\mathbf{x} - \mathbf{x}^*$ no nulo. Como se ha probado que

$$\mathbf{x} \neq \mathbf{x}^* \Rightarrow f(\mathbf{x}) - f(\mathbf{x}^*) > 0$$

queda demostrado que \mathbf{x}^* es el único punto de mínimo de $f(\mathbf{x})$

■

Las funciones cuadráticas con mínimo o máximo son un importante ejemplo de funciones linealmente unimodales. Esto se establecerá como un teorema:

Teorema:

Si \mathbf{A} es positiva definida la función cuadrática con mínimo

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$$

es linealmente unimodal.

Demostración:

Sea $\mathbf{x} = \mathbf{x}_0 + \lambda \mathbf{v}$

$$\text{Entonces } f(\mathbf{x}) = \frac{1}{2} (\mathbf{x}_0 + \lambda \mathbf{v})^T \mathbf{A} (\mathbf{x}_0 + \lambda \mathbf{v}) - \mathbf{b}^T (\mathbf{x}_0 + \lambda \mathbf{v})$$

$$= \frac{1}{2} (\mathbf{x}_0^T + \lambda \mathbf{v}^T) \mathbf{A} (\mathbf{x}_0 + \lambda \mathbf{v}) - \mathbf{b}^T (\mathbf{x}_0 + \lambda \mathbf{v})$$

$$= \frac{1}{2} \mathbf{x}_0^T \mathbf{A} \mathbf{x}_0 + \frac{1}{2} \lambda \mathbf{v}^T \mathbf{A} \mathbf{x}_0 + \frac{1}{2} \lambda \mathbf{x}_0^T \mathbf{A} \mathbf{v} + \frac{1}{2} \lambda^2 \mathbf{v}^T \mathbf{A} \mathbf{v} - \mathbf{b}^T \mathbf{x}_0 - \lambda \mathbf{b}^T \mathbf{v}$$

$$= \frac{1}{2} (\mathbf{v}^T \mathbf{A} \mathbf{v}) \lambda^2 + (\frac{1}{2} \mathbf{v}^T \mathbf{A} \mathbf{x}_0 + \frac{1}{2} \mathbf{x}_0^T \mathbf{A} \mathbf{v} - \mathbf{b}^T \mathbf{v}) \lambda + (\frac{1}{2} \mathbf{x}_0^T \mathbf{A} \mathbf{x}_0 - \mathbf{b}^T \mathbf{x}_0)$$

Es decir, $f(\mathbf{x})$ evaluada sobre la recta $\mathbf{x} = \mathbf{x}_0 + \lambda \mathbf{v}$ es una función cuadrática de λ :

$$F(\lambda) = a\lambda^2 + b\lambda + c$$

donde $a = \frac{1}{2} (\mathbf{v}^T \mathbf{A} \mathbf{v})$ es positivo, pues el vector de dirección \mathbf{v} no puede ser nulo y \mathbf{A} es positiva definida. La función $F(\lambda) = a\lambda^2 + b\lambda + c$ con $a > 0$ es unimodal con mínimo independientemente de b y c , así que $f(\mathbf{x})$ es linealmente unimodal. ■

Ejemplo 6

Considere la función cuadrática de dos variables:

$$f(\mathbf{x}) = 100 - 3u_1^2 - 4u_2^2 + 5u_1u_2 + 2u_1$$

- a) Exprese $f(\mathbf{x})$ en forma matricial.
- b) Pruebe que \mathbf{A} es definida negativa.
- c) Halle el punto de máximo de $f(\mathbf{x})$.

Solución:

$$\text{a) } f(\mathbf{x}) = -3u_1^2 + \frac{5}{2}u_1u_2 + \frac{5}{2}u_1u_2 - 4u_2^2 + 2u_1 + 100$$

$$= \frac{1}{2}(-6u_1^2 + 5u_1u_2 + 5u_1u_2 - 8u_2^2) - (-2u_1) + 100$$

$$\begin{aligned} &= \frac{1}{2} \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} -6 & 5 \\ 5 & -8 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} - \begin{bmatrix} -2 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + 100 \\ &= \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x} + c \end{aligned}$$

$$f(\mathbf{X}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x} + c$$

$$\text{donde } \mathbf{A} = \begin{bmatrix} -6 & 5 \\ 5 & -8 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} -2 \\ 0 \end{bmatrix} \quad c = 100$$

- b) Para probar que \mathbf{A} es definida negativa basta probar que $-\mathbf{A}$ es definida positiva. En efecto:

$$-\mathbf{A} = \begin{bmatrix} 6 & -5 \\ -5 & 8 \end{bmatrix}$$

cumple las condiciones deducidas en el ejemplo 5 para una matriz

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

ya que posee a y c positivos y $ac - b^2 = 48 - 25 > 0$

- c) El punto de máximo \mathbf{x}^* satisface que $\mathbf{Ax}^* = \mathbf{b}$. Esto es:

$$\begin{bmatrix} 6 & -5 \\ -5 & 8 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \end{bmatrix}$$

$$\begin{cases} -6u_1 + 5u_2 = -2 \\ 5u_1 - 8u_2 = 0 \end{cases}$$

cuya solución es: $u_1^* = \frac{16}{23}$ $u_2^* = \frac{10}{23}$

Algoritmo para hallar el óptimo en una dirección

En varios de los métodos de optimización multidimensional que se estudiarán a partir de esta sección, se necesita hallar el punto de extremo de una función $f(\mathbf{X})$ a lo largo de una trayectoria recta. Esto requiere elaborar un algoritmo numérico adecuado a este fin.

Evidentemente el algoritmo necesita como datos:

- $f(\mathbf{x})$: función de u_1, u_2, \dots, u_n
- \mathbf{x}_0 : vector de R^n a partir del cual se buscará un punto de máximo
- \mathbf{v} : vector unitario de R^n en cuya dirección se hará la búsqueda
- ε : tolerancia con que se hallará el punto de máximo

El algoritmo será llamado MaxUnidim y será una función que posee parámetros de entrada $f(\mathbf{x})$, \mathbf{x}_0 , \mathbf{v} y ε y da como resultado un número real

$\text{MaxUnidim}(f, \mathbf{x}_0, \mathbf{v}, \varepsilon)$

Este número real se encuentra de la siguiente forma:

```

 $s := 1$ 
do while  $f(\mathbf{x}_0 + s\mathbf{v}) < f(\mathbf{x}_0)$  and  $s > \varepsilon/10$ 
     $s := s/2$ 
end
if  $s < \varepsilon/10$  then
    No hay máximo en la dirección y sentido de  $\mathbf{v}$ 
    Búsqueda sin éxito
else
    Realizar búsqueda secuencial uniforme con paso  $s$  para la función
     $F(\lambda) = f(\mathbf{x}_0 + \lambda\mathbf{v})$  hasta obtener un intervalo  $[\lambda_{\inf}, \lambda_{\sup}]$  o llegar a 1000 pasos
    (Búsqueda sin éxito).
    if (Búsqueda con éxito) then
        En el intervalo  $[\lambda_{\inf}, \lambda_{\sup}]$  realizar búsqueda por biseción con  $\delta = \varepsilon/20$ 
        hasta que  $\lambda_{\sup} - \lambda_{\inf} < \varepsilon$ 

```

```

    end
end
if (búsqueda con éxito) then
    MaxUnidim =  $\frac{1}{2}(\lambda_{\text{inf}} + \lambda_{\text{sup}})$ 
else
    MaxUnidim = -1
end
Terminar

```

Como se aprecia, no se han tomado las técnicas más eficientes sino las más robustas, teniendo en cuenta que no será posible observar directamente el funcionamiento del algoritmo.

6.5 El método de búsqueda por coordenadas

Este es posiblemente el método más elemental para optimizar funciones multidimensionales. En general, es un algoritmo poco eficiente pero, en algunos casos especiales, ofrece buenos resultados. Se basa en la idea de realizar búsquedas unidimensionales en las direcciones de las variables de la función f . Esto es:

1. Hallar el punto \mathbf{x}_1 de máximo de $f(\mathbf{x})$ a partir de \mathbf{x}_0 tomando u_1 variable y las demás constantes.
2. Hallar el punto \mathbf{x}_2 de máximo de $f(\mathbf{x})$ a partir de \mathbf{x}_1 tomando u_2 variable y las demás constantes.
3. Hallar el punto \mathbf{x}_3 de máximo de $f(\mathbf{x})$ a partir de \mathbf{x}_2 tomando u_3 variable y las demás constantes.

Hasta:

- n.* Hallar el punto \mathbf{x}_n de máximo de $f(\mathbf{x})$ a partir de \mathbf{x}_{n-1} tomando u_n variable y las demás constantes.

Hacer $\mathbf{x}_0 = \mathbf{x}_n$ y comenzar por el paso 1 de nuevo.

El algoritmo se detiene cuando los puntos $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_n$ difieren muy poco entre sí.

En la figura 1 se aprecia gráficamente como funciona el algoritmo para una función de dos variables u_1 y u_2 cuyas curvas de nivel se muestran, donde $z_1 < z_2 < z_3$ y \mathbf{x}^* es el punto de máximo. En la figura cada flecha representa una búsqueda en una dirección paralela a un eje coordenado.

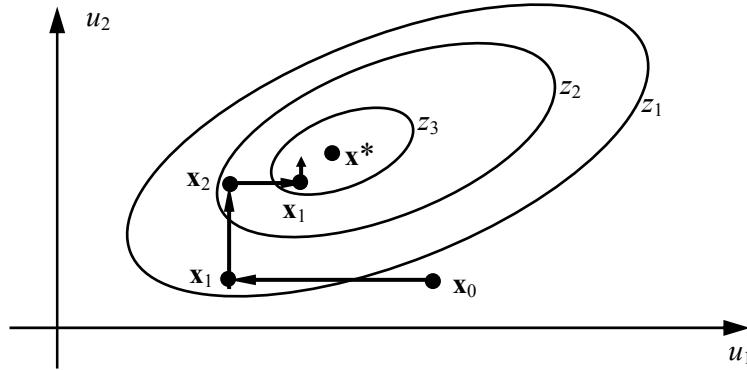


Figura 1

El algoritmo de búsqueda por coordenadas requiere que la función $f(\mathbf{x})$ sea linealmente unimodal, de modo que cada problema de búsqueda unidimensional tenga una solución.

Algoritmo en seudo código

El siguiente algoritmo halla el punto de máximo de una función $f(\mathbf{x})$ linealmente unimodal mediante el método de búsqueda por coordenadas. Utiliza como datos el número n de variables independientes, la función $f(\mathbf{x})$, el vector n dimensional \mathbf{x}_0 por donde comenzará la búsqueda y la tolerancia ε que se permitirá en el algoritmo. Se supone que se cuenta con un algoritmo como MaxUnidim que permite realizar búsquedas unidimensionales en cualquier dirección que se especifique mediante un vector n dimensional \mathbf{v} (este algoritmo fue desarrollado en la sección 6.4).

```

repeat
     $\lambda_{\max} := 0$  {Esta variable almacenará la mayor de las distancias recorridas en la secuencia de búsquedas unidimensionales}
    for i = 1 to n
         $\mathbf{v} := \mathbf{e}_i$  {i-simo vector canónico: un 1 en la  $i$ -sima coordenada y 0 en las otras}
         $\lambda := \text{MaxUnidim}(f, \mathbf{x}_{i-1}, \mathbf{v}, \varepsilon)$ 
        if  $\lambda < 0$  then {Esto solo sucede si la búsqueda en dirección  $\mathbf{e}_i$  no tuvo éxito}
             $\mathbf{v} := -\mathbf{e}_i$  {Se realiza una búsqueda en el sentido opuesto de  $\mathbf{e}_i$ }
             $\lambda := \text{MaxUnidim}(f, \mathbf{x}_{i-1}, \mathbf{v}, \varepsilon)$ 
            if  $\lambda < 0$  then  $\lambda = 0$  {Se supone que, si en ambos sentidos la búsqueda no tuvo éxito, entonces el punto de máximo se encuentra en el punto inicial}
        end
         $\mathbf{x}_i = \mathbf{x}_{i-1} + \lambda \mathbf{v}$ 
        if  $\lambda > \lambda_{\max}$  then  $\lambda_{\max} := \lambda$ 
    end
     $\mathbf{x}_0 := \mathbf{x}_n$ 
until  $\lambda_{\max} < \varepsilon$ 
El punto de máximo es  $\mathbf{x}_n$ 
Terminar

```

Observe que los vectores \mathbf{e}_i cuya i -sima componente es uno y las demás son ceros, son vectores unitarios. Como se ve, el algoritmo prevé la posibilidad de que en una iteración cualquiera el punto de máximo pueda estar en la dirección \mathbf{e}_i o $-\mathbf{e}_i$ (esto es, en la dirección del vector \mathbf{e}_i pero en su mismo sentido o en el sentido opuesto). Note también cómo, en el caso en que no hay punto de máximo ($\lambda < 0$) en ambos sentidos, se supone $\lambda = 0$, lo cual significa que el punto \mathbf{X}_{i-1} era, casualmente, el punto de máximo en la dirección de \mathbf{e}_i .

Ejemplo 1

En el ejemplo 6 de la sección 6.4 se demostró que la función cuadrática

$$f(\mathbf{x}) = 100 - 3u_1^2 - 4u_2^2 + 5u_1u_2 + 2u_1$$

posee un punto de máximo, el cual se calculó analíticamente:

$$\mathbf{x}^* = \begin{bmatrix} \frac{16}{23} \\ \frac{10}{23} \end{bmatrix} = \begin{bmatrix} 0.69565\dots \\ 0.43478\dots \end{bmatrix}$$

Determine \mathbf{x}^* por el método de búsqueda por coordenadas con precisión 0,0005 para:

$$\text{a) } \mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{b) } \mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{c) } \mathbf{x}_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{d) } \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Halle en cada caso la cantidad de búsquedas unidimensionales necesitadas.

Solución:

Nótese que la función $f(\mathbf{X})$, por ser cuadrática con máximo, es linealmente unimodal, por lo cual el algoritmo de búsqueda por coordenadas es aplicable. Los cálculos se realizaron con un programa elaborado utilizando el algoritmo anterior. Los resultados obtenidos fueron:

$$\text{a) } \mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \mathbf{x}^* = \begin{bmatrix} 0,6953 \\ 0,4345 \end{bmatrix} \quad \text{en 24 iteraciones.}$$

$$\text{b) } \mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathbf{x}^* = \begin{bmatrix} 0,6950 \\ 0,4343 \end{bmatrix} \quad \text{en 22 iteraciones.}$$

$$\text{c) } \mathbf{x}_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \mathbf{x}^* = \begin{bmatrix} 0,6961 \\ 0,4351 \end{bmatrix} \quad \text{en 24 iteraciones.}$$

$$\text{d) } \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \mathbf{x}^* = \begin{bmatrix} 0,6961 \\ 0,4351 \end{bmatrix} \quad \text{en 24 iteraciones.}$$

■

Como se observa en el ejemplo anterior, el método de búsqueda por coordenadas requiere una buena cantidad de búsquedas unidimensionales para su convergencia. Esto ocurre porque las curvas de nivel tienen forma de elipses inclinadas respecto a los ejes, lo cual sucede en presencia de términos rectangulares (del tipo de $5u_1u_2$ que posee este ejemplo). Si hay poca interacción entre las variables, lo que corresponde geométricamente con curvas de nivel de ejes paralelos a los coordenados, la misma interpretación geométrica permite asegurar que la convergencia ha de ser muy rápida.

Note, sin embargo, que la rapidez de convergencia no está influida significativamente por la selección de \mathbf{x}_0 siempre que se tome a una distancia adecuada del punto de óptimo.

Ejemplo 2

La función cuadrática: $f(\mathbf{x}) = 100 - 4u_1^2 - 5u_2^2 + 2u_1 + 3u_2$

posee un punto de máximo cuyo valor exacto es:

$$\mathbf{x}^* = \begin{bmatrix} 0,25 \\ 0,3 \end{bmatrix}$$

Utilice el método de búsqueda por coordenadas para hallarlo, con precisión de 0.0005. Tome punto de partida:

$$\text{a) } \mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{b) } \mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{c) } \mathbf{x}_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{d) } \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

y determine el número de búsquedas unidimensionales en cada caso.

Solución:

Utilizando el mismo programa del ejemplo anterior, los resultados fueron:

$$\text{a) } \mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \mathbf{x}^* = \begin{bmatrix} 0,25 \\ 0,3 \end{bmatrix} \quad \text{en 3 búsquedas unidimensionales.}$$

$$\text{b) } \mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathbf{x}^* = \begin{bmatrix} 0,25 \\ 0,3 \end{bmatrix} \quad \text{en 3 búsquedas unidimensionales.}$$

$$\text{c) } \mathbf{x}_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \mathbf{x}^* = \begin{bmatrix} 0,25 \\ 0,3 \end{bmatrix} \quad \text{en 3 búsquedas unidimensionales.}$$

$$\text{d) } \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \mathbf{x}^* = \begin{bmatrix} 0,25 \\ 0,3 \end{bmatrix} \quad \text{en 3 búsquedas unidimensionales.}$$

Nótese cómo, al no existir interacción entre u_1 y u_2 (curvas de nivel con ejes no inclinados) el método converge rápidamente.

Ejemplo 3

La función de Rosembrook ha sido propuesta para poner a prueba los métodos de optimización multidimensional. Es una función muy difícil, pues sus curvas de nivel son, como se muestra en la figura 2, “óvalos” muy alargados y con su eje mayor curvado en forma de parábola. Esta función posee un punto de mínimo en $(1, 1)$ y se recomienda comenzar la búsqueda en $(-1, 1)$ para someter a prueba a un algoritmo. Intente usar búsqueda por coordenadas para la función de Rosembrook:

$$f(\mathbf{x}) = 100(u_2 - u_1^2)^2 + (1 - u_1)^2$$

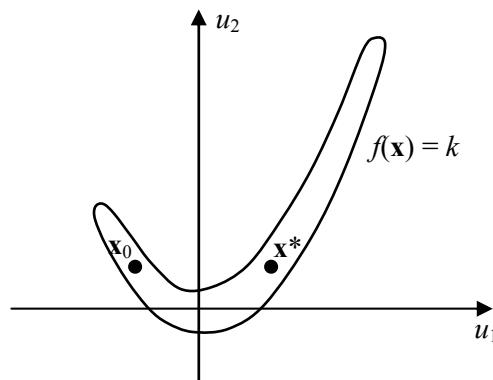


Figura 2

Solución:

La convergencia es sumamente lenta. Después de 1000 búsquedas unidimensionales se tiene:

$$\mathbf{x}^* \approx \begin{bmatrix} 0,922 \\ 0,850 \end{bmatrix}$$

después de 1500 búsquedas unidimensionales:

$$\mathbf{x}^* \approx \begin{bmatrix} 0,961 \\ 0,924 \end{bmatrix}$$

y después de 2000:

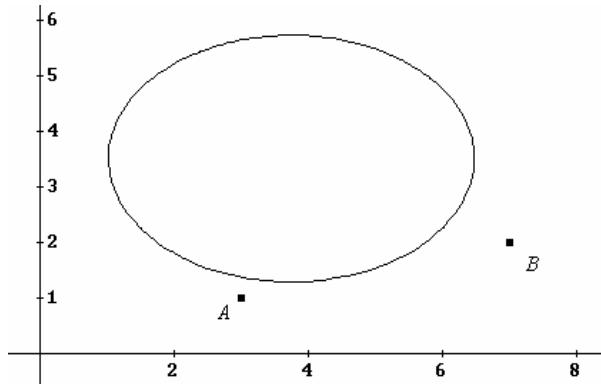
$$\mathbf{x}^* \approx \begin{bmatrix} 0,980 \\ 0,960 \end{bmatrix}$$

Ejercicios

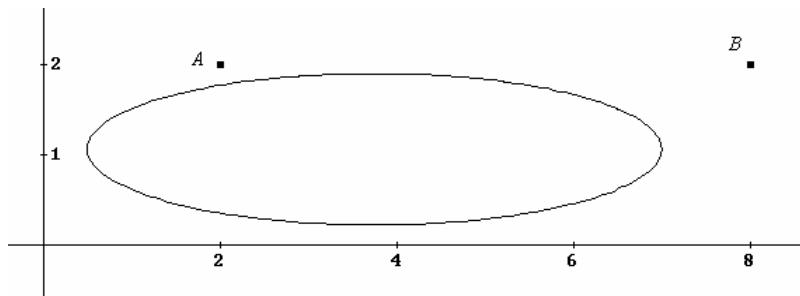
Algunos de los ejercicios que siguen son muy laboriosos para realizar los cálculos a mano. Se supone que posea programas computacionales, preferiblemente confeccionados por usted, para ayudarle en el trabajo. De no ser así, en los ejercicios que requieran mucho trabajo manual, determine los resultados con menos cifras exactas que las exigidas.

1. En cada una de las funciones cuadráticas que siguen se muestra una curva de nivel, cuya forma puede servir para predecir como se comportará el método de búsqueda por coordenadas al determinar su punto de extremo. En cada caso, exprese la función de forma matricial y aplique el método de búsqueda por coordenadas para hallar el punto de extremo utilizando como punto de partida los puntos A y B que se dan en la propia figura. Obtenga la solución con tres cifras decimales exactas. ¿Puede dar alguna conclusión general como resultado de este ejercicio?

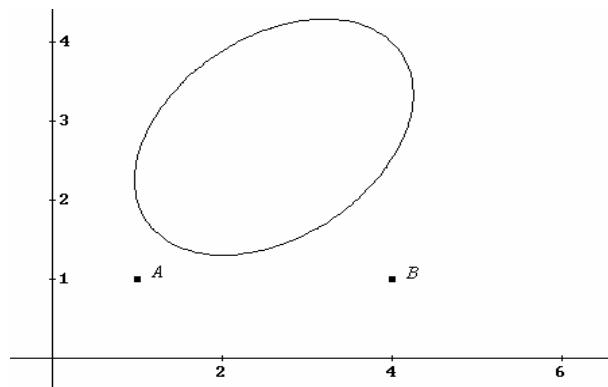
a) $f(x, y) = 2x^2 - 15x + 3y^2 - 21y; \quad A = (3, 1); \quad B = (7, 2)$



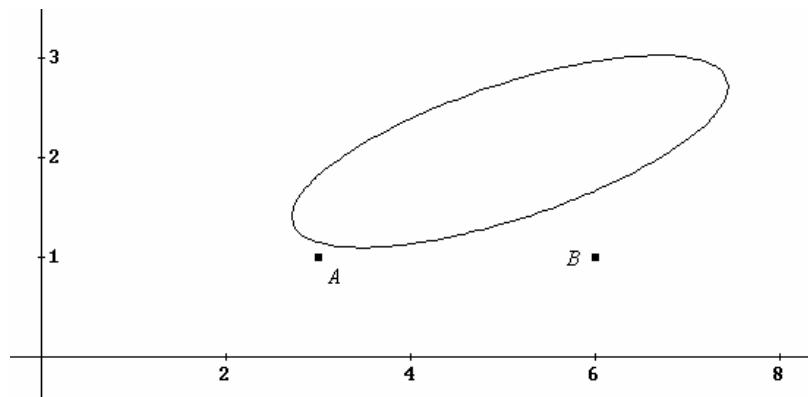
b) $f(x, y) = 2x^2 - 15x + 30y^2 - 63y; \quad A = (2, 2); \quad B = (8, 2)$



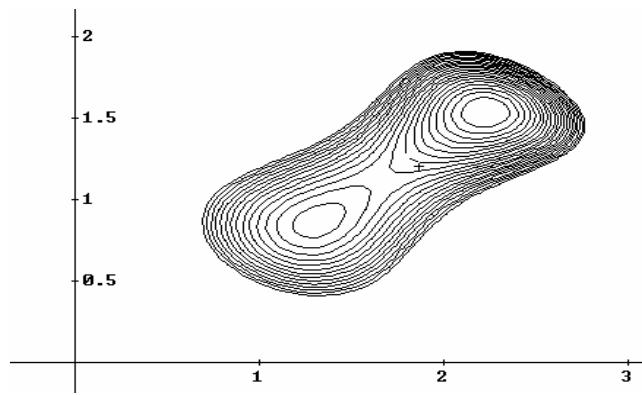
c) $f(x, y) = 5x^2 - 4xy + 6y^2 - 15x - 23y; \quad A = (1, 1); \quad B = (4, 1)$



d) $f(x, y) = 3x^2 - 10xy + 18y^2 - 10x - 23y; \quad A = (3, 1); \quad B = (6, 1)$



2. La función $f(x, y) = x^2 + \operatorname{sen}(xy) + 2y^2 - 3x - 4y$ no es unimodal. En la figura se observan algunas curvas de nivel que muestran que en esta región ella presenta dos puntos de mínimo. Seleccione adecuadamente el punto inicial \mathbf{x}_0 para determinar ambos puntos de mínimo con tres cifras decimales exactas.



3. Halle la recta de mejor ajuste a los cuatro puntos $(2, 3), (3, 4), (4, 6)$ y $(5, 5)$ como un problema de optimización. Obtenga los coeficientes con tres cifras decimales exactas. Compruebe sus resultados utilizando el sistema normal de ecuaciones estudiado en el capítulo 4.

4 Calcule la distancia mínima entre las curvas $y = \operatorname{sen} x$ y $y = x^2 - 6x + 10$. Obtenga el resultado con tres cifras decimales exactas.

5. Halle el plano

$$\frac{x}{a} + \frac{y}{b} + \frac{z}{c} = 1 \quad (a, b \text{ y } c \text{ son positivos})$$

que contiene al punto $(2,1; 0,9; 1,1)$ y que determina con los planos coordenados una pirámide de base triangular del menor volumen posible. Obtenga los resultados con error menor que 0,001.

6. A continuación se muestra un algoritmo para hallar un punto de máximo de la función f mediante el método de búsqueda por coordenadas. Este algoritmo posee un serio problema por el cual no funciona. Analice cual es este problema y repare el algoritmo.

```

repeat
     $\lambda_{\max} := 0$  {Esta variable almacenará la mayor de las distancias recorridas en la secuencia de búsquedas unidimensionales}
    for  $i = 1$  to  $n$ 
         $v := e_i$  { $i$ -simo vector canónico: un 1 en la  $i$ -sima coordenada y 0 en las otras}
         $\lambda := \operatorname{MaxUnidim}(f, X_{i-1}, V, \varepsilon)$ 
         $x_i = x_{i-1} + \lambda v$ 
        if  $\lambda > \lambda_{\max}$  then  $\lambda_{\max} := \lambda$ 
    end
     $x_0 = x_n$ 
until  $\lambda_{\max} < \varepsilon$ 
El punto de máximo es  $x_n$ 
Terminar

```

6.6 El método del gradiente

Cuando se conocen las derivadas parciales de $f(\mathbf{x})$ respecto a las variables u_1, u_2, \dots, u_n es posible seleccionar la dirección en el espacio R^n en que $f(\mathbf{x})$ aumenta con mayor rapidez. Como se conoce de los cursos de Cálculo, esta dirección viene dada en cada punto \mathbf{x} por el vector gradiente que corresponde a dicho punto:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial u_1} \\ \frac{\partial f}{\partial u_2} \\ \vdots \\ \frac{\partial f}{\partial u_n} \end{bmatrix}$$

En la figura 1 se muestran las curvas de nivel de una función $f(\mathbf{x})$ en R^2 y vectores gradiente en varios puntos. Recuérdese que el vector gradiente en cualquier punto es normal a la curva de nivel que pasa por dicho punto.

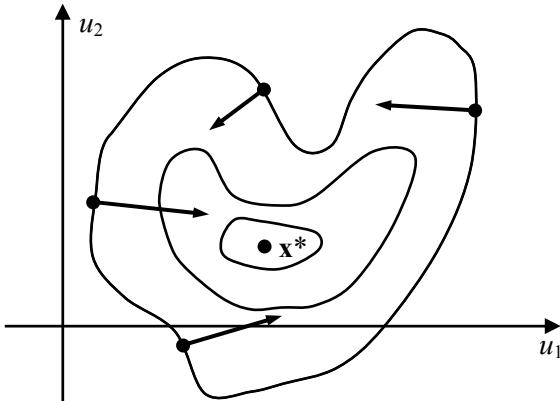


Figura 1

Es evidente que, a menos que las curvas de nivel sean circunferencias concéntricas, la dirección del gradiente en un punto no señala, en general, al punto \mathbf{x}^* de óptimo.

El método del gradiente, también llamado “steepest ascent” (o “steepest descent”, si se utiliza para minimizar) fue introducido por Cauchy en 1847 y se basa en:

Realizar una búsqueda unidimensional en la dirección de $\nabla f|_{\mathbf{x}_0}$ hasta encontrar el punto de máximo \mathbf{x}_1 en esa trayectoria.

Realizar una búsqueda unidimensional en la dirección de $\nabla f|_{\mathbf{x}_1}$ hasta encontrar el punto de máximo \mathbf{x}_2 en esa trayectoria.

Y así sucesivamente hasta que los puntos hallados difieran muy poco entre sí, o el gradiente tome valores muy pequeños o alguna otra condición que indique la proximidad de \mathbf{x}^* . Obsérvese que se ha supuesto que $f(\mathbf{x})$ es linealmente unimodal, de manera que en cada trayectoria lineal el problema unidimensional resultante tenga solución. El método puede resumirse en el algoritmo que sigue.

Algoritmo en seudo código

El algoritmo que sigue permite hallar el punto de máximo de la función linealmente unimodal $f(\mathbf{x})$ mediante el método del gradiente. Se supone que f es diferenciable, de modo que posee gradiente en cada punto. El algoritmo utiliza como datos: el número n de variables independientes de la función a maximizar, la función $f(\mathbf{x})$, las n derivadas parciales de $f(\mathbf{x})$, el vector \mathbf{x}_0 donde se comenzará el proceso de búsqueda y la tolerancia ε con que se desea hallar el punto de máximo. El algoritmo ofrece como resultado una aproximación del punto de máximo \mathbf{x}^* calculado con una tolerancia ε .

```

 $i := 0$ 
repeat
   $\mathbf{v} := \nabla f|_{\mathbf{x}_i}$ 
  Normalizar  $\mathbf{v}$ 
   $\lambda := \text{MaxUnidim}(f, \mathbf{x}_i, \mathbf{v}, \varepsilon)$ 
   $\mathbf{x}_{i+1} := \mathbf{x}_i + \lambda \mathbf{v}$ 
   $i := i + 1$ 
until  $\lambda < \varepsilon$ 

```

■

El método del gradiente posee algunos inconvenientes que deben señalarse. El primero, es la necesidad de conocer las derivadas parciales de $f(\mathbf{x})$, lo cual lo limita al caso en que se conoce la expresión analítica de f y sus derivadas pueden ser calculadas. El método es muy sensible a la geometría de las superficies de nivel. Como se aprecia de las figuras 2 y 3, cuando las curvas de nivel son elipses alargadas, el método posee una lenta convergencia; en cambio, con curvas de nivel circulares y concéntricas, basta una iteración para encontrar el punto de óptimo \mathbf{x}^* .

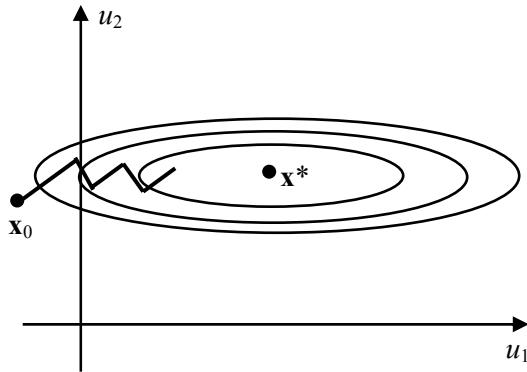


Figura 2

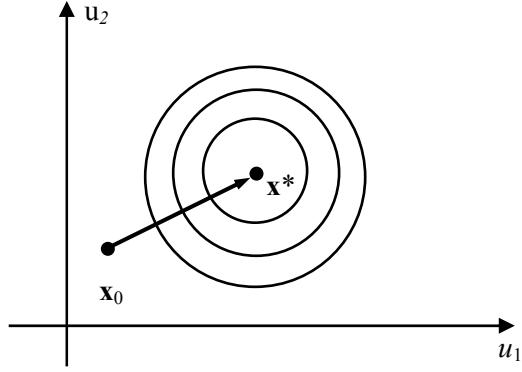


Figura 3

Esta característica trae como consecuencia que un simple cambio de escala, el cual puede ocurrir por utilizar diferentes unidades de medida en una de las variables, puede cambiar sustancialmente la rapidez de convergencia del método.

A diferencia de lo que ocurre en la búsqueda por coordenadas, en el método del gradiente resulta determinante la selección del punto inicial \mathbf{x}_0 . Véase en la figura 4 cómo la rapidez de convergencia puede diferir para distintas selecciones de \mathbf{x}_0 .

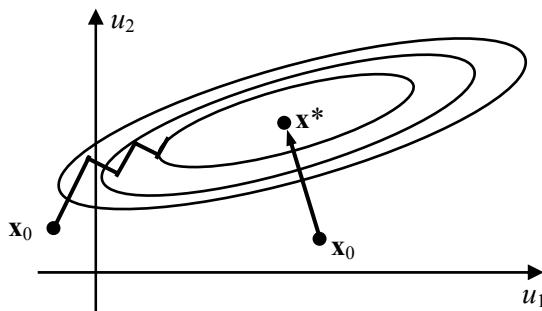


Figura 4

Ejemplo 1

Como ya se dijo en el ejemplo 6 de la sección 6.4, la función cuadrática

$$f(\mathbf{x}) = 100 - 3u_1^2 - 4u_2^2 + 5u_1u_2 + 2u_1$$

posee un punto de máximo

$$\mathbf{x}^* = \begin{bmatrix} \frac{16}{23} \\ \frac{10}{23} \end{bmatrix} = \begin{bmatrix} 0,69565\dots \\ 0,43478\dots \end{bmatrix}$$

Determine \mathbf{X}^* por el método gradiente con precisión 0,0005 para:

- a) $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ b) $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ c) $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ d) $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

y señale en cada caso la cantidad de búsquedas unidimensionales requeridas.

Solución:

Mediante un programa confeccionado según el seudo código anterior se obtuvo la solución de cada uno de los incisos:

a) $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\mathbf{x}^* = \begin{bmatrix} 0,6951 \\ 0,4342 \end{bmatrix}$ en 21 búsquedas unidimensionales.

b) $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\mathbf{x}^* = \begin{bmatrix} 0,6956 \\ 0,4347 \end{bmatrix}$ en 3 búsquedas unidimensionales.

c) $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $\mathbf{x}^* = \begin{bmatrix} 0,6955 \\ 0,4346 \end{bmatrix}$ en 5 búsquedas unidimensionales.

d) $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\mathbf{x}^* = \begin{bmatrix} 0,6961 \\ 0,4351 \end{bmatrix}$ en 13 búsquedas unidimensionales.

■

Nótese cómo varía la rapidez de convergencia de acuerdo con la posición de \mathbf{x}_0 , siendo en muchos casos inferior a la búsqueda por coordenadas. Como se observa en el ejemplo que sigue, este comportamiento persiste aun en casos en que no hay interacción entre las variables (ausencia de términos rectangulares) en los cuales la búsqueda por coordenadas tiene una magnífica eficiencia.

Ejemplo 2

La función cuadrática: $f(\mathbf{x}) = 100 - 4u_1^2 - 25u_2^2 + 2u_1 + 3u_2$

tiene curvas de nivel que son elipses con sus ejes horizontal y vertical respectivamente. Su punto de máximo es:

$$\mathbf{x}^* = \begin{bmatrix} 0,25 \\ 0,06 \end{bmatrix}$$

Utilice el método del gradiente para hallarlo con precisión de 0,0005 usando diferentes aproximaciones iniciales:

a) $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0,4 \end{bmatrix}$ b) $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ c) $\mathbf{x}_0 = \begin{bmatrix} 0,3 \\ 0 \end{bmatrix}$

y determine el número de iteraciones en cada caso.

Solución:

Mediante el mismo programa del ejemplo anterior fueron obtenidos los siguientes resultados:

a) $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0,4 \end{bmatrix}$ $\mathbf{x}^* = \begin{bmatrix} 0,2499 \\ 0,0600 \end{bmatrix}$ en 4 búsquedas unidimensionales.

b) $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\mathbf{x}^* = \begin{bmatrix} 0,2494 \\ 0,0601 \end{bmatrix}$ en 17 búsquedas unidimensionales.

c) $\mathbf{x}_0 = \begin{bmatrix} 0,3 \\ 0 \end{bmatrix}$ $\mathbf{x}^* = \begin{bmatrix} 0,2502 \\ 0,0600 \end{bmatrix}$ en 5 búsquedas unidimensionales.

■

Ejemplo 3

Halle el punto de mínimo de la función de Rosembrook (vea el ejemplo 3 de la sección 6.5)

$$f(\mathbf{x}) = 100(u_2 - u_1^2)^2 + (1 - u_1)^2$$

Compare con el número de iteraciones obtenido con el algoritmo de búsqueda por coordenadas.

Solución:

Tomando $\mathbf{x}_0 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ se obtiene $\mathbf{x}^* = \begin{bmatrix} 0.9202 \\ 0.8468 \end{bmatrix}$ después de 1000 búsquedas unidimensionales.

■

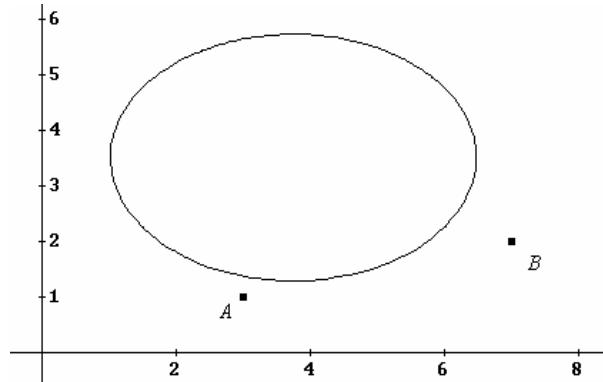
Aunque el resultado es algo menos lento que la búsqueda por coordenadas, el método se comporta con una convergencia sumamente lenta en esta función.

Ejercicios

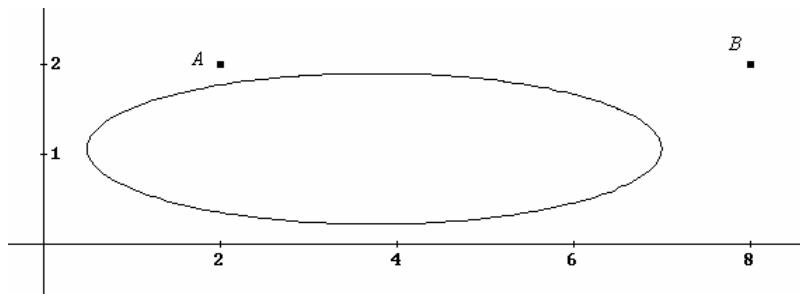
Algunos de los ejercicios que siguen son muy laboriosos para realizar los cálculos a mano. Se supone que posea programas computacionales, preferiblemente confeccionados por usted, para ayudarle en el trabajo. De no ser así, en los ejercicios que requieran mucho trabajo manual, determine los resultados con menos cifras exactas que las exigidas.

1. En cada una de las funciones cuadráticas que siguen se muestra una curva de nivel, cuya forma puede servir para predecir como se comportará el método del gradiente al determinar su punto de extremo. Estos problemas aparecieron también en los ejercicios de la sección anterior. En cada caso, aplique el método del gradiente para hallar el punto de extremo utilizando como punto de partida los puntos A y B que se dan en la propia figura. Obtenga la solución con tres cifras decimales exactas. Compare con la solución obtenida utilizando el método de búsqueda por coordenadas ¿Puede dar alguna conclusión general como resultado de este ejercicio?

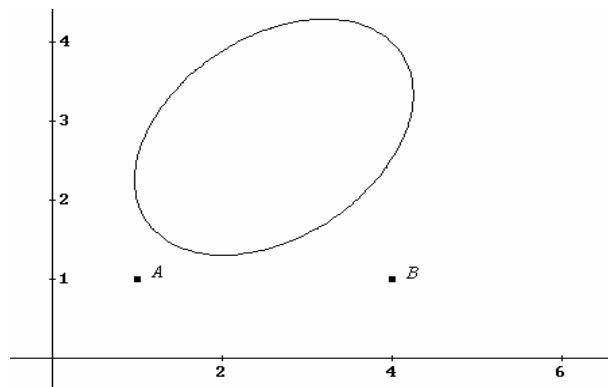
a) $f(x, y) = 2x^2 - 15x + 3y^2 - 21y; \quad A = (3, 1); \quad B = (7, 2)$



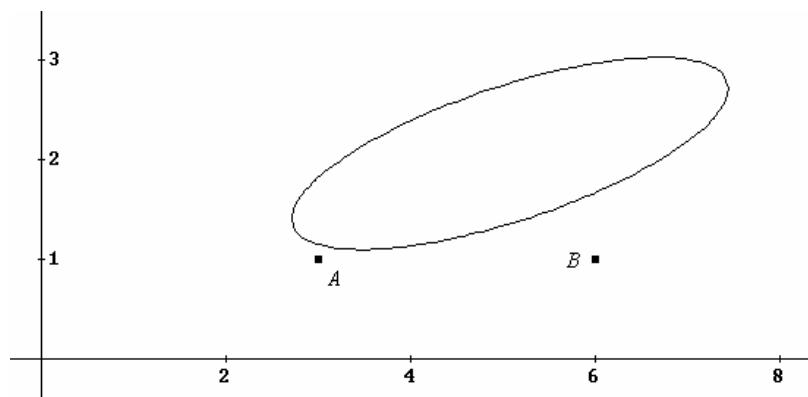
b) $f(x, y) = 2x^2 - 15x + 30y^2 - 63y; \quad A = (2, 2); \quad B = (8, 2)$



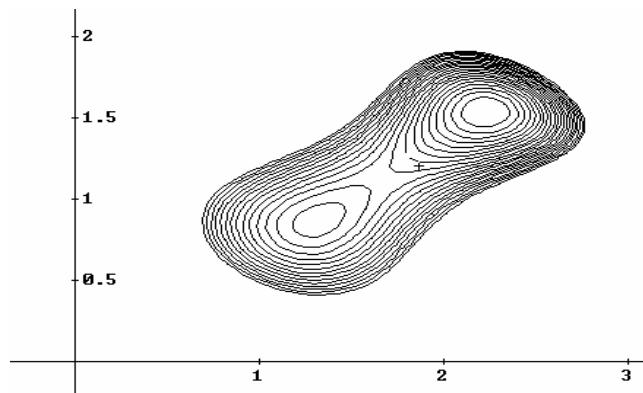
c) $f(x, y) = 5x^2 - 4xy + 6y^2 - 15x - 23y; \quad A = (1, 1); \quad B = (4, 1)$



d) $f(x, y) = 3x^2 - 10xy + 18y^2 - 10x - 23y; \quad A = (3, 1); \quad B = (6, 1)$



2. La función $f(x, y) = x^2 + \operatorname{sen}(xy) + 2y^2 - 3x - 4y$ no es unimodal. En la figura se observan algunas curvas de nivel que muestran que en esta región ella presenta dos puntos de mínimo. Seleccione adecuadamente el punto inicial \mathbf{x}_0 para determinar ambos puntos de mínimo mediante el método del gradiente. Obtenga los resultados con tres cifras decimales exactas.

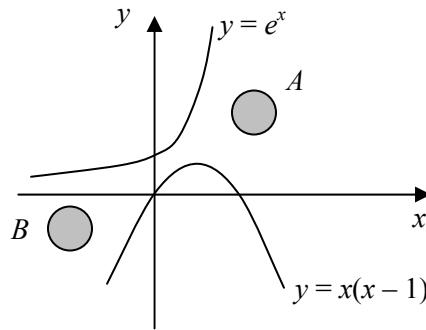


3. Ajuste un modelo no lineal del tipo $g(x) = ae^{bx}$ a los datos mostrados en la siguiente tabla:

x	1	2	3	4	5
y	1,18	1,26	1,36	1,46	1,57

obtenga la solución con tres cifras decimales exactas.

- 4 Halle el diámetro del mayor círculo que puede pasar de la posición A hasta la posición B entre las dos gráficas sin cortar a ninguna de las dos (solamente rozarlas). Obtenga el resultado con tres cifras decimales exactas.



5. En el método del gradiente se puede detectar la proximidad de un punto de extremo por que la norma del vector gradiente se aproxima a cero. Elabore un algoritmo en seudo código que utilice este criterio para detener el proceso iterativo.

6.7 El método de Powell

Direcciones conjugadas en una función cuadrática

Algunos algoritmos de optimización multidimensional están basados en la idea de obtener técnicas que permitan hallar eficientemente el óptimo de una función cuadrática, con la esperanza de que, en el caso de funciones no cuadráticas, el método mantenga su eficiencia si se parte de un punto cercano al óptimo. Varios de los métodos que han logrado mayor éxito utilizan el concepto de direcciones conjugadas, que es una generalización del concepto de direcciones perpendiculares.

Definición:

Sea \mathbf{A} una matriz cuadrada de orden n positiva (negativa) definida. Sean \mathbf{d}_i y \mathbf{d}_j vectores de R^n que señalan dos direcciones en ese espacio. Las direcciones \mathbf{d}_i y \mathbf{d}_j se llaman conjugadas respecto a \mathbf{A} , o también \mathbf{A} -ortogonales, si

$$\mathbf{d}_i^T \mathbf{A} \mathbf{d}_j = 0$$

Ejemplo 1

Sea la matriz

$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$$

la cual es positiva definida (compruébese). Sea $\mathbf{d}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Halle un vector \mathbf{d}_2 conjugado de \mathbf{d}_1 respecto a \mathbf{A} .

Solución:

Sea un vector

$$\mathbf{d}_2 = \begin{bmatrix} a \\ b \end{bmatrix}$$

Se debe cumplir que

$$\mathbf{d}_1^T \mathbf{A} \mathbf{d}_2 = 0$$

es decir:

$$[1 \quad 1] \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = 0$$

$$[5 \quad 6] \begin{bmatrix} a \\ b \end{bmatrix} = 0$$

$$5a + 6b = 0$$

Tomando una solución arbitraria:

$$a = 6; \quad b = -5$$

Por tanto,

$$\mathbf{d}_2 = \begin{bmatrix} 6 \\ -5 \end{bmatrix}$$

es conjugado de \mathbf{d}_1

Interpretación geométrica

Considérese el caso $n = 2$, de modo que la función:

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$$

con \mathbf{A} positiva definida representa un paraboloide elíptico con vértice (punto de mínimo) en $\mathbf{x} = \mathbf{0}$. Las curvas de nivel

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = k$$

con $k > 0$ son elipses con centro en $\mathbf{x} = \mathbf{0}$, como se muestra en la figura 1.

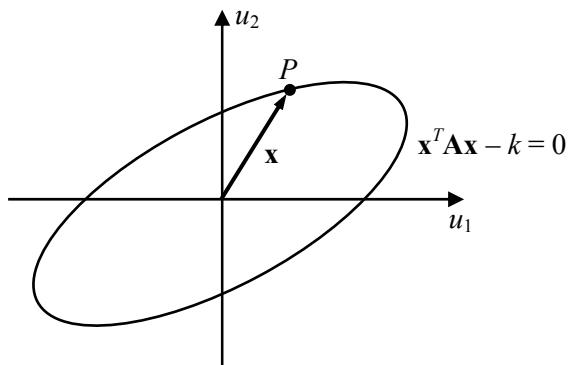


Figura 1

Sea \mathbf{x} la dirección correspondiente a un punto P (vector que une el centro de la elipse con el punto P). Hallemos el gradiente de $\mathbf{x}^T \mathbf{A} \mathbf{x} - k$ que, como se sabe, es un vector normal a la elipse en P :

$$\nabla(\mathbf{x}^T \mathbf{A} \mathbf{x} - k) = \nabla(a_{11}u_1^2 + a_{12}u_1u_2 + a_{21}u_2u_1 + a_{22}u_2^2 - k)$$

$$= \begin{bmatrix} 2a_{11}u_1 + 2a_{12}u_2 \\ 2a_{21}u_1 + 2a_{22}u_2 \end{bmatrix} = 2 \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 2\mathbf{A}\mathbf{x}$$

Es decir, el vector $\mathbf{A}\mathbf{x}$ es normal a la elipse en el punto P . Sea ahora \mathbf{y} un vector en una dirección conjugada a \mathbf{x} , esto es:

$$\mathbf{y}^T \mathbf{A} \mathbf{x} = 0$$

lo cual significa que

$$\mathbf{y} \cdot (\mathbf{A}\mathbf{x}) = 0$$

esto es, que \mathbf{y} es ortogonal a $\mathbf{A}\mathbf{x}$ y, por tanto, tangente a la elipse en P . Esto prueba que, si \mathbf{x} es el vector radial del punto P de la elipse

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = k$$

entonces su dirección conjugada \mathbf{y} es tangente a la elipse en el mismo punto P , tal como se muestra en la figura 2.

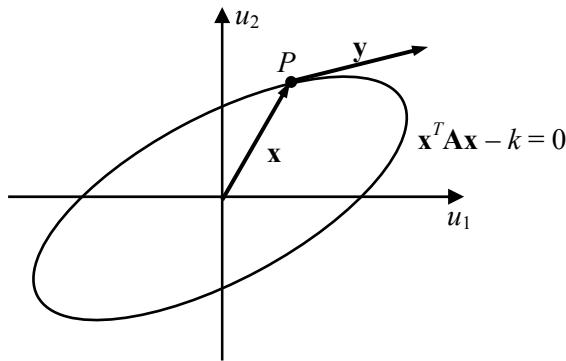


Figura 2

Propiedades fundamentales de las direcciones conjugadas

Propiedad 1:

Si $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_k$ son vectores de R^n no nulos conjugados entre sí respecto a la matriz \mathbf{A} positiva (negativa) definida, entonces el conjunto $\{\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_k\}$ es linealmente independiente.

Demostración:

Supóngase que

$$\alpha_0 \mathbf{d}_0 + \alpha_1 \mathbf{d}_1 + \dots + \alpha_k \mathbf{d}_k = \mathbf{0}$$

Premultiplicando en ambos miembros por \mathbf{A} :

$$\alpha_0 \mathbf{A} \mathbf{d}_0 + \alpha_1 \mathbf{A} \mathbf{d}_1 + \dots + \alpha_k \mathbf{A} \mathbf{d}_k = \mathbf{0}$$

Si ahora se premultiplica por el transpuesto de cualquiera de los vectores \mathbf{d}_i :

$$\alpha_0 \mathbf{d}_i^T \mathbf{A} \mathbf{d}_0 + \alpha_1 \mathbf{d}_i^T \mathbf{A} \mathbf{d}_1 + \dots + \alpha_k \mathbf{d}_i^T \mathbf{A} \mathbf{d}_k = 0$$

y, como \mathbf{d}_i y \mathbf{d}_j son conjugados para $i \neq j$, resulta:

$$\alpha_i \mathbf{d}_i^T \mathbf{A} \mathbf{d}_i = 0$$

Sin embargo, por ser \mathbf{A} positiva o negativa definida,

$$\mathbf{d}_i^T \mathbf{A} \mathbf{d}_i \neq 0$$

así que $\alpha_i = 0$. Como \mathbf{d}_i es cualquiera de los vectores del conjunto, se ha probado que

$$\alpha_0 = \alpha_1 = \dots = \alpha_k = 0$$

y, por tanto, el conjunto $\{\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_k\}$ es linealmente independiente.

Propiedad 2:

Sea la función cuadrática de n variables con punto de mínimo \mathbf{x}^*

$$f(\mathbf{X}) = \frac{1}{2} \mathbf{X}^T \mathbf{A} \mathbf{X} - \mathbf{B}^T \mathbf{X}$$

y sean $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{n-1}$ direcciones conjugadas entre sí respecto a la matriz \mathbf{A} . Como el conjunto $\{\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{n-1}\}$ es linealmente independiente y posee n vectores, es una base de R^n , así que \mathbf{x}^* puede expresarse como combinación de la base conjugada:

$$\mathbf{x}^* = \alpha_0 \mathbf{d}_0 + \alpha_1 \mathbf{d}_1 + \dots + \alpha_{n-1} \mathbf{d}_{n-1}$$

Los α_i pueden determinarse fácilmente pues:

$$\mathbf{A} \mathbf{x}^* = \alpha_0 \mathbf{A} \mathbf{d}_0 + \alpha_1 \mathbf{A} \mathbf{d}_1 + \dots + \alpha_{n-1} \mathbf{A} \mathbf{d}_{n-1}$$

pero como $\mathbf{A} \mathbf{x}^* = \mathbf{b}$:
$$\mathbf{b} = \alpha_0 \mathbf{A} \mathbf{d}_0 + \alpha_1 \mathbf{A} \mathbf{d}_1 + \dots + \alpha_{n-1} \mathbf{A} \mathbf{d}_{n-1}$$

$$\mathbf{d}_i^T \mathbf{b} = \alpha_0 \mathbf{d}_i^T \mathbf{A} \mathbf{d}_0 + \alpha_1 \mathbf{d}_i^T \mathbf{A} \mathbf{d}_1 + \dots + \alpha_{n-1} \mathbf{d}_i^T \mathbf{A} \mathbf{d}_{n-1}$$

$$\mathbf{d}_i^T \mathbf{b} = \alpha_i \mathbf{d}_i^T \mathbf{A} \mathbf{d}_i$$

de donde:

$$\alpha_i = \frac{\mathbf{d}_i^T \mathbf{b}}{\mathbf{d}_i^T \mathbf{A} \mathbf{d}_i} \quad i = 0, 1, 2, \dots, n-1$$

■

Lo notable de la fórmula anterior es que ella permite hallar las coordenadas $\alpha_0, \alpha_1, \dots, \alpha_{n-1}$ de \mathbf{x}^* respecto a una base sin tener que resolver un sistema lineal como $\mathbf{A} \mathbf{x}^* = \mathbf{b}$. Sin embargo, como establece la propiedad 3, estas coordenadas α_i se pueden obtener por otra vía mucho más interesante.

Propiedad 3:

Sea

$$f(\mathbf{X}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$$

una función cuadrática con punto de mínimo \mathbf{x}^* y $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{n-1}$ direcciones conjugadas entre sí respecto a la matriz \mathbf{A} . Sea \mathbf{x}_0 un punto arbitrario y:

$$\mathbf{x}^* - \mathbf{x}_0 = \alpha_0 \mathbf{d}_0 + \alpha_1 \mathbf{d}_1 + \dots + \alpha_{n-1} \mathbf{d}_{n-1}$$

de modo que:

$$\mathbf{x}^* = \mathbf{x}_0 + \alpha_0 \mathbf{d}_0 + \alpha_1 \mathbf{d}_1 + \dots + \alpha_{n-1} \mathbf{d}_{n-1}$$

Se define ahora el proceso iterativo:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \quad k = 0, 1, 2, \dots, n-1$$

esto es:

$$\mathbf{x}_1 = \mathbf{x}_0 + \alpha_0 \mathbf{d}_0$$

$$\mathbf{x}_2 = \mathbf{x}_1 + \alpha_1 \mathbf{d}_1 = \mathbf{X}_0 + \alpha_0 \mathbf{d}_0 + \alpha_1 \mathbf{d}_1$$

⋮

$$\mathbf{x}_n = \mathbf{x}_{n-1} + \alpha_{n-1} \mathbf{d}_{n-1} = \mathbf{X}_0 + \alpha_0 \mathbf{d}_0 + \alpha_1 \mathbf{d}_1 + \dots + \alpha_{n-1} \mathbf{d}_{n-1} = \mathbf{x}^*$$

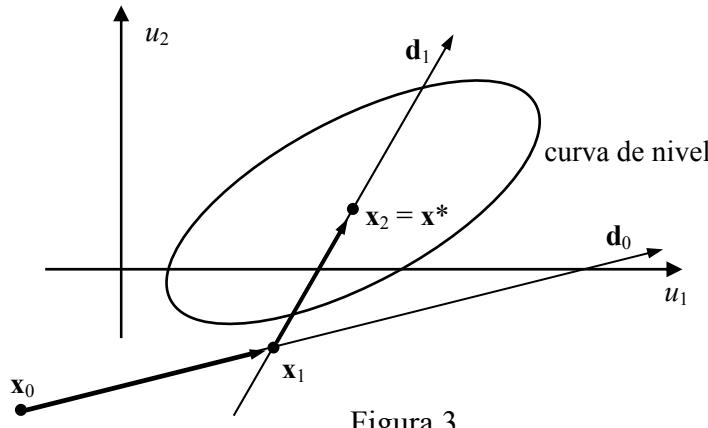
Entonces, puede demostrarse que, para $k = 0, 1, 2, \dots, n-1$:

α_k es el valor de λ para el cual la función unidimensional $f(\mathbf{x}_k + \lambda \mathbf{d}_k)$ toma su mínimo valor.

■

Esta propiedad permite ir buscando las coordenadas $\alpha_0, \alpha_1, \dots, \alpha_{n-1}$ mediante sendos problemas de optimización unidimensional. En este proceso se van obteniendo los puntos $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$.

Como se ve, el punto \mathbf{x}_n es ya el óptimo. Observe en la figura 3 cómo funciona el algoritmo para un caso donde $n = 2$.



Se puede entonces resumir que:

- Si $f(\mathbf{x})$ es una función cuadrática con punto de extremo \mathbf{x}^*
- Si $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{n-1}$ son direcciones conjugadas de su matriz \mathbf{A}
- Si \mathbf{x}_0 es un punto cualquiera

Entonces el punto \mathbf{x}^* se halla mediante n búsquedas unidimensionales sucesivas a partir de \mathbf{x}_0 en las direcciones conjugadas.

■

Podría pensarse que hallar un conjunto de n direcciones conjugadas es muy complicado pero, como se verá a continuación, hay formas muy eficientes de hacerlo.

Método de las tangentes paralelas

La idea geométrica del método puede verse muy claramente si $f(\mathbf{x})$ es una función cuadrática de dos variables, como en la figura 4. En ese caso, las curvas de nivel son elipses concéntricas (el centro es \mathbf{x}^*) y con la misma excentricidad.

Suponga que, partiendo de puntos diferentes P_1 y P_2 se trazan dos rectas paralelas, ambas en la dirección \mathbf{d}_1 . Sean L_1 y L_2 esas rectas. Obviamente, el mínimo de $f(\mathbf{x})$ a lo largo de L_1 se obtiene en el punto \mathbf{x}_1 donde L_1 es tangente a la curva de nivel que pasa por \mathbf{x}_1 y el mínimo sobre L_2 estará en un punto \mathbf{x}_2 donde L_2 es tangente a la curva de nivel que pasa por \mathbf{x}_2 .

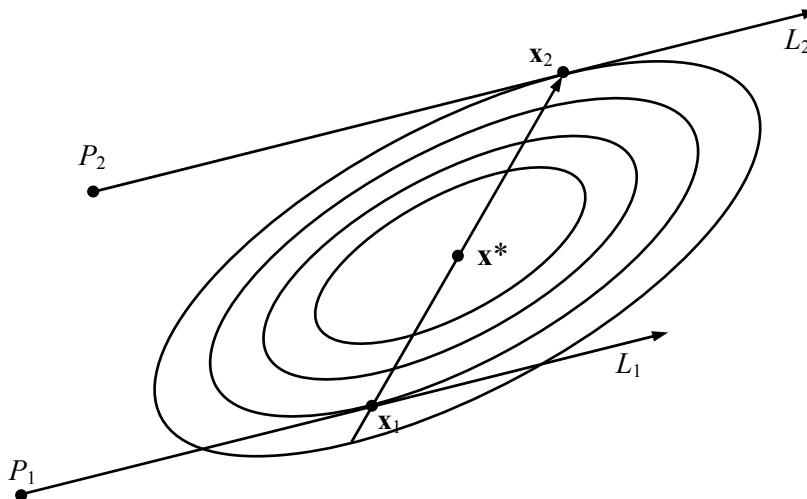


Figura 4

El segmento que une a \mathbf{x}_1 y \mathbf{x}_2 pasa por \mathbf{x}^* , o sea, la dirección del vector

$$\mathbf{d}_2 = \mathbf{x}_2 - \mathbf{x}_1$$

es radial. De acuerdo con la interpretación geométrica analizada, resulta que \mathbf{d}_2 es una dirección conjugada de \mathbf{d}_1 .

Lo anterior se puede establecer analíticamente.

Sea

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$$

una función cuadrática de n variables con punto de mínimo. Sea \mathbf{d} una dirección cualquiera en R^n y P_1 y P_2 dos puntos diferentes de ese espacio. Sean:

- \mathbf{x}_1 : Punto de mínimo de f sobre la recta L_1 que pasa por P_1 con dirección \mathbf{d}
 \mathbf{x}_2 : Punto de mínimo de f sobre la recta L_2 que pasa por P_2 con dirección \mathbf{d}

entonces el vector $\mathbf{x}_2 - \mathbf{x}_1$ es conjugado de \mathbf{d} .

Demostración:

La ecuación de L_1 es: $\mathbf{x} = \mathbf{x}_1 + \lambda \mathbf{d}$

La ecuación de L_2 es: $\mathbf{x} = \mathbf{x}_2 + \lambda \mathbf{d}$

Sobre L_1 : $f(\mathbf{x}) = \frac{1}{2} (\mathbf{x}_1 + \lambda \mathbf{d})^T \mathbf{A} (\mathbf{x}_1 + \lambda \mathbf{d}) - \mathbf{b}^T (\mathbf{x}_1 + \lambda \mathbf{d})$

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}_1^T \mathbf{A} \mathbf{x}_1 + \lambda \mathbf{d}^T \mathbf{A} \mathbf{x}_1 + \frac{1}{2} \lambda^2 \mathbf{d}^T \mathbf{A} \mathbf{d} - \mathbf{b}^T \mathbf{x}_1 - \lambda \mathbf{b}^T \mathbf{d}$$

Como esta función toma su mínimo en \mathbf{x}_1 (o sea, cuando $\lambda = 0$), debe cumplirse que su derivada respecto a λ evaluada en $\lambda = 0$ se anule, es decir:

$$\mathbf{d}^T \mathbf{A} \mathbf{x}_1 + \lambda \mathbf{d}^T \mathbf{A} \mathbf{d} - \mathbf{b}^T \mathbf{d} \Big|_{\lambda=0} = 0$$

esto es:

$$\mathbf{d}^T \mathbf{A} \mathbf{x}_1 - \mathbf{b}^T \mathbf{d} = 0$$

Realizando el mismo razonamiento para la recta L_2 se llega a:

$$\mathbf{d}^T \mathbf{A} \mathbf{x}_2 - \mathbf{b}^T \mathbf{d} = 0$$

Restando las dos últimas ecuaciones:

$$\mathbf{d}^T \mathbf{A} (\mathbf{x}_2 - \mathbf{x}_1) = 0$$

y esto significa que $\mathbf{x}_2 - \mathbf{x}_1$ y \mathbf{d} son direcciones conjugadas, como se deseaba probar.

El método de Powell

En las páginas anteriores se han establecido dos importantes propiedades de las direcciones conjugadas de una función cuadrática de n variables:

- Si se cuenta con un conjunto de n direcciones conjugadas de la función, entonces realizando n búsquedas unidimensionales sucesivas en dichas direcciones, se llega al óptimo de la función cuadrática.
- Si se realizan dos búsquedas unidimensionales en una misma dirección a partir de puntos distintos, el vector que une los óptimos alcanzados determina una dirección conjugada de la anterior.

En el método de Powell se combinan estos dos resultados del siguiente modo. Partiendo de un conjunto de n direcciones de búsqueda $\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n$ (usualmente, las direcciones de los ejes coordenados) y de un punto inicial \mathbf{x}_0 , se realizan n procesos de búsqueda unidimensional, obteniéndose sucesivamente:

- \mathbf{x}_1 : optimizando f en la dirección \mathbf{d}_1 desde \mathbf{x}_0
- \mathbf{x}_2 : optimizando f en la dirección \mathbf{d}_2 desde \mathbf{x}_1
- \vdots
- \mathbf{x}_n : optimizando f en la dirección \mathbf{d}_n desde \mathbf{x}_{n-1}

Al llegar al punto \mathbf{x}_n se realiza una nueva búsqueda unidimensional en la dirección del vector $\mathbf{x}_n - \mathbf{x}_0$. La norma de este vector sirve como indicador de la convergencia del proceso. Esta búsqueda suele llamarse un “paso de aceleración” por cuanto se realiza en una dirección que es el resumen de n búsquedas previas y recoge la “experiencia” obtenida en esos procesos. El vector alcanzado en este paso de aceleración se toma como un nuevo \mathbf{x}_0 para comenzar otra vez todo el proceso pero antes, se hace un cambio en el conjunto de direcciones $\{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n\}$. La dirección \mathbf{d}_1 se elimina, tomándose:

$$\begin{aligned}\mathbf{d}_1 &:= \mathbf{d}_2 \\ \mathbf{d}_2 &:= \mathbf{d}_3 \\ &\vdots \\ \mathbf{d}_{n-1} &:= \mathbf{d}_n\end{aligned}$$

y se introduce una nueva dirección \mathbf{d}_n según el vector $\mathbf{x}_n - \mathbf{x}_0$, previamente normalizado. Nótese que el nuevo \mathbf{x}_0 se obtuvo buscando en la dirección de este \mathbf{d}_n a partir de \mathbf{x}_n en el paso de aceleración. En la nueva iteración, el nuevo \mathbf{x}_n se obtendrá buscando a partir de \mathbf{x}_{n-1} en la dirección de \mathbf{d}_n . Como \mathbf{x}_0 y \mathbf{x}_n se han obtenido con búsquedas según rectas paralelas, el vector $\mathbf{x}_n - \mathbf{x}_0$ de la segunda iteración será conjugado de \mathbf{d}_n .

De este modo, el conjunto de direcciones $\{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n\}$ se va convirtiendo paulatinamente en un conjunto de direcciones conjugadas, eliminando en cada iteración el primer vector de ese conjunto y añadiendo uno nuevo por el final que es conjugado del último. Después de n iteraciones, el conjunto está formado por n direcciones conjugadas entre sí y, por tanto, en la próxima iteración se obtiene el punto de óptimo.

Claro que todo esto es así si f es una función cuadrática. Si no lo es, ya el concepto mismo de dirección conjugada carece de sentido pero, como se supone que en la región próxima a \mathbf{X}^* la función f se comporta similarmente a una función cuadrática, las direcciones que se van obteniendo, van formando una adecuada colección que, además, se va actualizando continuamente.

La figura 5 ilustra la forma en que funciona el algoritmo para el caso de una función cuadrática de dos variables. En este caso cada iteración utiliza tres búsquedas unidimensionales: dos en las direcciones \mathbf{d}_1 y \mathbf{d}_2 y el paso de aceleración, que sirve además para introducir una nueva dirección de búsqueda. Después de la segunda iteración, ya las direcciones \mathbf{d}_1 y \mathbf{d}_2 son conjugadas, de modo que en la segunda iteración (es decir, en la sexta búsqueda unidimensional) se alcanza el punto de máximo \mathbf{x}^* .

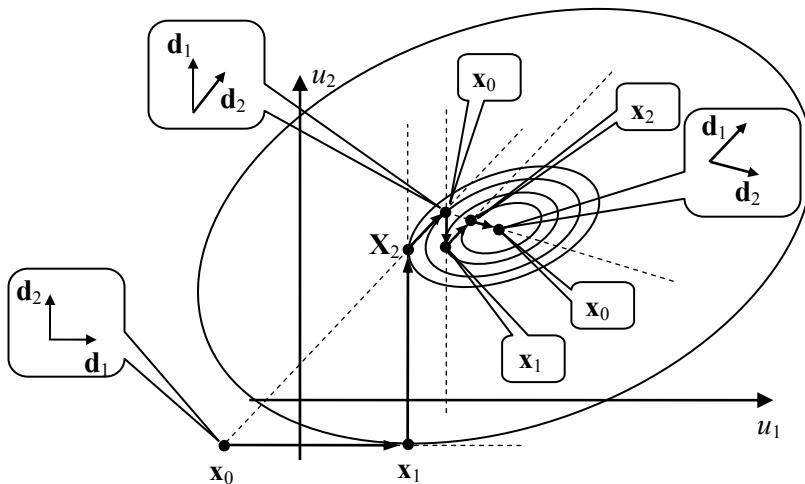


Figura 5

Algoritmo en seudo código

En el siguiente algoritmo se describe en detalle el método de Powell para hallar el punto de máximo de una función $f(\mathbf{x})$ de n variables independientes. Se supone que f es linealmente unimodal. Para que el algoritmo resulte más simple, el procedimiento $\text{MaxBidirec}(f, \mathbf{x}, \mathbf{v})$ empleado aquí realiza la búsqueda a partir de \mathbf{x} en la dirección del vector \mathbf{v} (en ambos sentidos) y devuelve como resultado el punto de óptimo obtenido. El algoritmo utiliza como datos, el número n de variables de la función a optimizar, la función f , el vector \mathbf{x}_0 , que indica el punto donde se comenzará todo el proceso, y la tolerancia ε con que se obtendrá el punto de máximo \mathbf{x}^* .

```

for  $i = 1$  to  $n$ 
     $\mathbf{d}_i := \mathbf{e}_i$  {Las direcciones de búsqueda se toman inicialmente en la dirección de los vectores canónicos}
end
repeat
    for  $i = 1$  to  $n$ 
         $\mathbf{x}_i := \text{MaxBidirec}(f, \mathbf{x}_{i-1}, \mathbf{d}_i)$ 
    end
    for  $i = 1$  to  $n - 1$ 
         $\mathbf{d}_i := \mathbf{d}_{i+1}$ 
    end
     $\mathbf{d}_n := \mathbf{x}_n - \mathbf{x}_0$ 
     $dist :=$  norma de  $\mathbf{d}_n$ 
     $\mathbf{d}_n := \frac{\mathbf{d}_n}{dist}$ 
     $\mathbf{x}_0 := \text{MaxBidirec}(f, \mathbf{x}_n, \mathbf{d}_n)$ 
until  $dist < \varepsilon$ 

```

Ejemplo 2

Halle el punto de máximo de la función cuadrática

$$f(\mathbf{x}) = 100 - 3u_1^2 - 4u_2^2 + 5u_1u_2 + 2u_1$$

por el método de Powell con precisión de tres cifras decimales exactas tomando

$$\mathbf{x}_0 = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$$

Solución:

La solución se llevará a cabo con detalle para ayudar a comprender el algoritmo.

Iteración 1:

$$\mathbf{x}_0 = \begin{bmatrix} 5 \\ 5 \end{bmatrix} \quad \mathbf{d}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathbf{d}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Búsqueda en dirección \mathbf{d}_1 : $\mathbf{x}_1 = \begin{bmatrix} 4,500040 \\ 5,000000 \end{bmatrix}$

Búsqueda en dirección \mathbf{d}_2 : $\mathbf{x}_2 = \begin{bmatrix} 4,500040 \\ 2,812548 \end{bmatrix}$

Dirección $\mathbf{x}_2 - \mathbf{x}_0$ (normalizada) = $\begin{bmatrix} -0,222813 \\ -0,974861 \end{bmatrix}$

Búsqueda en dirección $\mathbf{x}_2 - \mathbf{x}_0$: $\mathbf{x}_0 = \begin{bmatrix} 4,405260 \\ 2,397859 \end{bmatrix} \quad \lambda = 0,425383$

Iteración 2: $\mathbf{x}_0 = \begin{bmatrix} 4,405260 \\ 2,397859 \end{bmatrix} \quad \mathbf{d}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \mathbf{d}_2 = \begin{bmatrix} -0,222813 \\ -0,974861 \end{bmatrix}$

Búsqueda en dirección \mathbf{d}_1 : $\mathbf{x}_1 = \begin{bmatrix} 4,405260 \\ 2,753245 \end{bmatrix}$

Búsqueda en dirección \mathbf{d}_2 : $\mathbf{x}_2 = \begin{bmatrix} 4,312821 \\ 2,348980 \end{bmatrix}$

Dirección $\mathbf{x}_2 - \mathbf{x}_0$ (normalizada) = $\begin{bmatrix} -0,884022 \\ -0,467445 \end{bmatrix}$

Búsqueda en dirección $\mathbf{x}_2 - \mathbf{x}_0$: $\mathbf{x}_0 = \begin{bmatrix} 0,695992 \\ 0,435660 \end{bmatrix} \quad \lambda = 4,091764$

$$\text{Iteración 3: } \mathbf{x}_0 = \begin{bmatrix} 0,695992 \\ 0,435660 \end{bmatrix} \quad \mathbf{d}_1 = \begin{bmatrix} -0,222813 \\ -0,974861 \end{bmatrix} \quad \mathbf{d}_2 = \begin{bmatrix} -0,884022 \\ -0,467445 \end{bmatrix}$$

$$\text{Búsqueda en dirección } \mathbf{d}_1: \quad \mathbf{x}_1 = \begin{bmatrix} 0,695824 \\ 0,434922 \end{bmatrix}$$

$$\text{Búsqueda en dirección } \mathbf{d}_2: \quad \mathbf{x}_2 = \begin{bmatrix} 0,695761 \\ 0,434889 \end{bmatrix}$$

$$\text{Dirección } \mathbf{x}_2 - \mathbf{x}_0 \text{ (normalizada)} = \begin{bmatrix} -0,287006 \\ -0,957929 \end{bmatrix}$$

$$\text{Búsqueda en dirección } \mathbf{x}_2 - \mathbf{x}_0: \quad \mathbf{x}_0 = \begin{bmatrix} 0,695754 \\ 0,434865 \end{bmatrix} \quad \lambda = 0.000025 < \varepsilon$$

$$\text{Respuesta: Con cuatro cifras decimales exactas } \mathbf{x}^* = \begin{bmatrix} 0,695754 \\ 0,434865 \end{bmatrix}$$

Nótese como, sorpresivamente, en la sexta búsqueda (final de la segunda iteración) ya se alcanzó \mathbf{x}^* . La razón es obvia; como esta es una función cuadrática de dos variables, después de dos iteraciones ya las direcciones de búsqueda forman una base conjugada y el punto de óptimo se encuentra con toda precisión. El método requiere algunas iteraciones más solamente para verificar la satisfacción de algún criterio de parada.

Ejemplo 3

Pruebe el método de Powell con la función de Rosembrook. Compare el resultado con los obtenidos usando los métodos de búsqueda por coordenadas y del gradiente.

Solución:

$$\text{Tomando } \mathbf{x}_0 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \text{ y } f(\mathbf{x}) = 100(u_2 - u_1^2)^2 + (1 - u_1)^2$$

se obtiene \mathbf{x}^* con tres cifras decimales exactas:

$$\mathbf{x}^* = \begin{bmatrix} 1.000002 \\ 1.000004 \end{bmatrix}$$

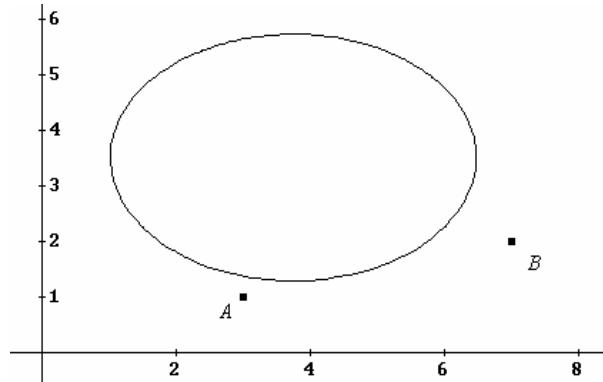
después de 29 búsquedas unidimensionales. En el método de búsqueda por coordenadas, después de 2000 búsquedas unidimensionales, aún no se había obtenido una milésima de precisión. En el método del gradiente, después de 1000 búsquedas unidimensionales, todavía el error es mayor de 0,15.

Ejercicios

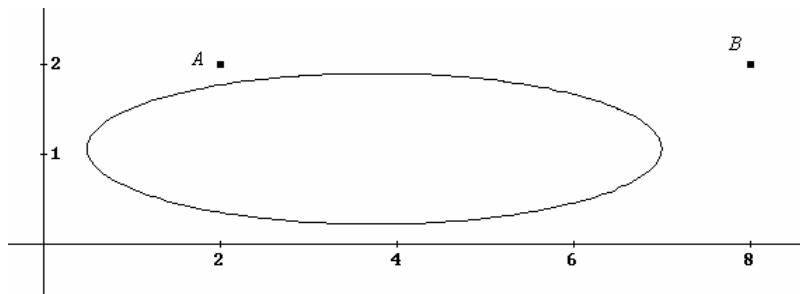
Algunos de los ejercicios que siguen son muy laboriosos para realizar los cálculos a mano. Se supone que posea programas computacionales, preferiblemente confeccionados por usted, para ayudarle en el trabajo. De no ser así, en los ejercicios que requieran mucho trabajo manual, determine los resultados con menos cifras exactas que las exigidas.

1. En cada una de las funciones cuadráticas que siguen se muestra una curva de nivel, cuya forma puede servir para predecir como se comportará el método de Powell al determinar su punto de extremo. Estos problemas aparecieron también en los ejercicios de las dos secciones anteriores. En cada caso, aplique el método de Powell para hallar el punto de extremo utilizando como punto de partida los puntos A y B que se dan en la propia figura. Obtenga la solución con tres cifras decimales exactas. Compare con la solución obtenida utilizando los métodos de búsqueda por coordenadas y del gradiente ¿Puede dar alguna conclusión general como resultado de este ejercicio?

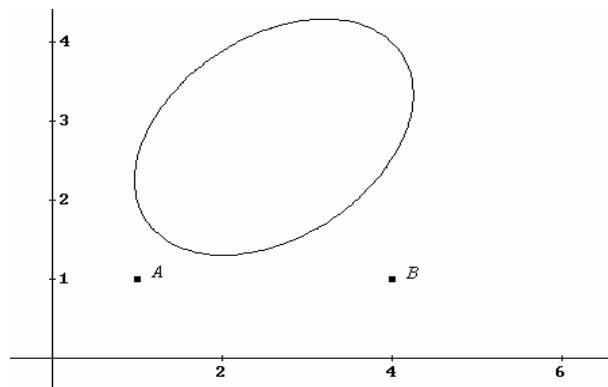
a) $f(x, y) = 2x^2 - 15x + 3y^2 - 21y; \quad A = (3, 1); \quad B = (7, 2)$



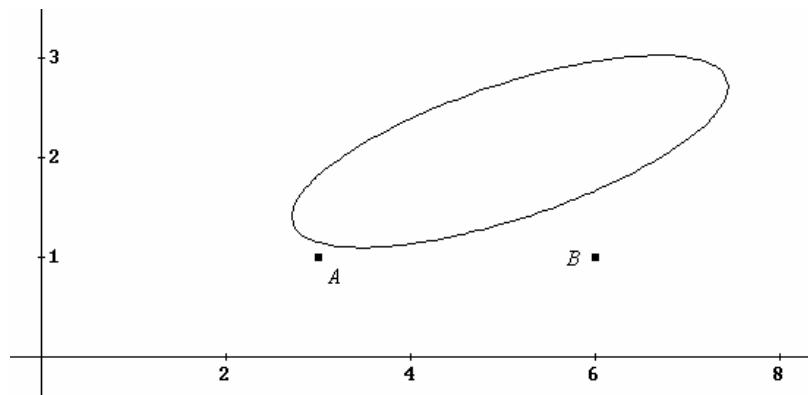
b) $f(x, y) = 2x^2 - 15x + 30y^2 - 63y; \quad A = (2, 2); \quad B = (8, 2)$



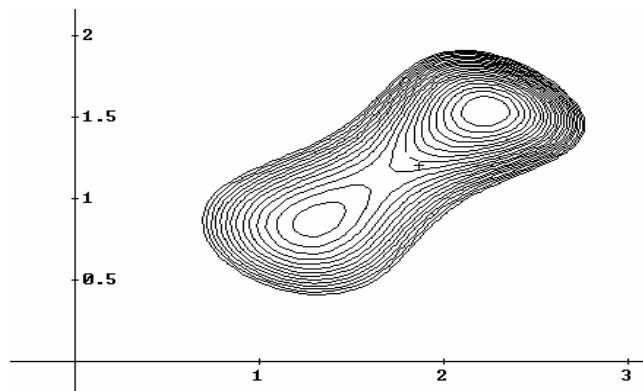
c) $f(x, y) = 5x^2 - 4xy + 6y^2 - 15x - 23y; \quad A = (1, 1); \quad B = (4, 1)$



d) $f(x, y) = 3x^2 - 10xy + 18y^2 - 10x - 23y; \quad A = (3, 1); \quad B = (6, 1)$



2. La función $f(x, y) = x^2 + \operatorname{sen}(xy) + 2y^2 - 3x - 4y$ no es unimodal. En la figura se observan algunas curvas de nivel que muestran que en esta región ella presenta dos puntos de mínimo. Seleccione adecuadamente el punto inicial \mathbf{x}_0 para determinar ambos puntos de mínimo mediante el método de Powell. Obtenga los resultados con tres cifras decimales exactas. Explique por qué en este caso la convergencia no es tan rápida como en el ejercicio anterior.



3. Halle la mínima distancia entre el punto $(2, 3, 1)$ y el parabolóide elíptico $z = x^2 + 2y^2$. Obtenga el resultado con tres cifras decimales exactas.

4. Dados los puntos $(1,2; 2)$, $(3,1; 2,5)$, $(3,9; 1,6)$, halle las coordenadas de un punto P del plano tal que la suma de sus distancias a los tres puntos dados sea lo menor posible. Obtenga la solución con error menor que 0,001.
5. Cuando un rayo de luz parte del punto A , se refleja en una superficie y después pasa por un punto B , lo hace de modo que su trayectoria es mínima. Si un rayo de luz parte del punto $(1, 1, 6)$, se refleja en la superficie $2x^2 + 3y^2 + 4z^2 = 12$ y después pasa por el punto $(1,5; 1,2; 10)$, ¿En qué punto de la superficie se reflejó? Determine el resultado con tres cifras decimales exactas.
6. Explique por qué en el algoritmo de Powell el paso de aceleración es fundamental.

6.8 El método del simplex secuencial

El método del simplex secuencial es una técnica de búsqueda multidimensional que no se basa en la realización de búsquedas unidimensionales como los otros tres procedimientos estudiados hasta aquí. Además de su eficiencia relativa, se trata de un algoritmo con una lógica muy simple y, por tanto, fácil de programar. Como su fundamento es esencialmente geométrico, es preferible estudiarlo primero en el caso de funciones de dos variables y posteriormente generalizarlo a funciones de n variables independientes, en las cuales ya no será posible una visualización. Como hasta ahora, se supone que todas las funciones involucradas en el análisis son linealmente unimodales con máximo.

El simplex

En el contexto de este método, se llamará *simplex* de un espacio n dimensional, al elemento con la misma dimensión del espacio que posea la estructura geométrica más simple posible. Así:

En el espacio unidimensional (la recta), el simplex es un intervalo cerrado, el cual está limitado por dos puntos.

En el espacio bidimensional (el plano), el simplex es un triángulo equilátero, el cual está limitado por tres segmentos iguales y posee tres puntos (vértices) que lo determinan.

En el espacio tridimensional, el simplex es un tetraedro (pirámide de base triangular) regular, que está limitado por cuatro triángulos equiláteros y posee cuatro puntos (vértices) que lo determinan.

En el espacio de dimensión 4, el simplex (que ya no posee un nombre especial y, naturalmente, no se puede representar) sería una “figura” regular delimitada por cinco puntos del espacio, que lo determinan. Esta idea se puede extender a espacios de cualquier dimensión: para un espacio de dimensión n , el simplex es la figura regular que está limitada por $n + 1$ puntos de ese espacio que equidistan entre sí. En la figura 1 se muestra el simplex del espacio bidimensional y el del espacio tridimensional.

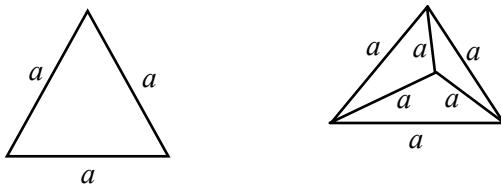


Figura 1

El método del simplex para funciones de dos variables

Sea f una función de las variables u_1 y u_2 y sean tres pares ordenados \mathbf{x}_1 , \mathbf{x}_2 y \mathbf{x}_3 que forman un triángulo equilátero, es decir, un simplex. La idea del algoritmo consiste en ir generando una sucesión de simplex adyacentes uno a otro, que conduzca al punto de máximo de la función f . La sucesión de simplex se genera siguiendo tres sencillas reglas, de las cuales la primera es muy obvia:

Regla 1: Si f_1 , f_2 y f_3 son los valores que toma la función $f(\mathbf{x})$ en los tres vértices del simplex, debe formarse un nuevo simplex en el cual se conserven los dos vértices donde la función toma los dos mayores valores y se sustituya el vértice donde la función toma el menor valor.

Esta es la regla que más frecuentemente se aplica, pero no siempre. En la figura 2, se muestran las curvas de nivel de una función de dos variables con punto de máximo en \mathbf{x}^* y se aprecian varios simplex, numerados 1, 2, 3, ..., 10 que han sido generados en ese orden aplicando la regla 1. En la figura se ha seguido el convenio de señalar con un punto negro el vértice donde la función toma el mayor valor, con un punto gris el vértice donde el valor de la función es intermedio y con un punto blanco el vértice donde f toma el valor más pequeño. Obsérvese que, en todos los casos, el vértice señalado con un punto blanco es el que desaparece para formar el simplex que sigue.

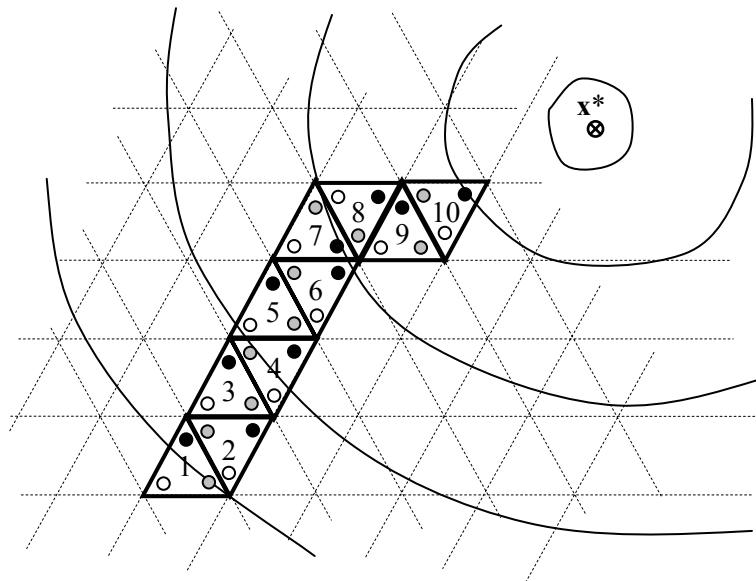


Figura 2

Al generar un nuevo simplex, puede suceder que en el vértice nuevo la función f tome el valor más pequeño. Si solo existiera la regla 1, el simplex siguiente coincidiría con el anterior y el algoritmo quedaría estancado. En la figura 3 se ilustra una situación en que esto sucede. Nótese

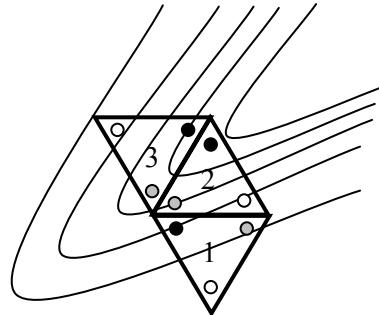


Figura 3

que, según la regla 1, después del simplex 3 el nuevo simplex coincidiría con el 2 y esto detendría el movimiento del simplex. Observe que las curvas de nivel presentan ángulos agudos muy marcados. Esta situación se evita con la regla 2.

Regla 2: El vértice más reciente del simplex no puede eliminarse. Si la función toma el menor valor en el vértice más reciente del simplex, entonces el vértice que se elimina es aquel donde la función toma el siguiente valor en orden creciente.

En la figura 4 se muestra como continuaría el proceso de la figura 3, con la introducción de esta nueva regla. Se muestran en gris los simplex (3 y 4) en que se necesita aplicar la regla 2.

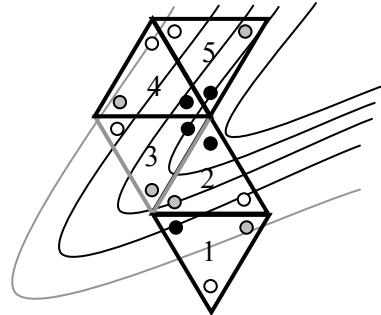


Figura 4

La tercera y última regla es necesaria cuando el simplex se encuentra en las proximidades del punto de máximo. En la figura 5 se aprecia lo que sucede. Los simplex del 4 al 9 van rodeando

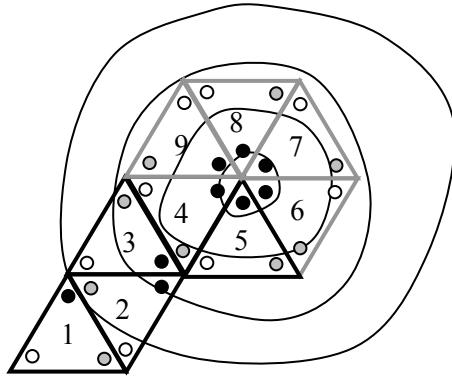


Figura 5

al punto donde ocurre el máximo, de manera que existe un vértice (el que se halla más cerca de \mathbf{X}^* , que permanece formando parte de los mismos. Evidentemente, el simplex número 10 coincidirá con el 4 y se produce un ciclo: 4, 5, 6, 7, 8, 9, 4,... Cuando se llega a una situación así, que se detecta porque hay un vértice que permanece establemente formando parte del simplex, ya no es posible mejorar la aproximación lograda con un simplex de este tamaño y se procede a tomar una de dos decisiones: Terminar el proceso y aceptar el vértice que se repite como la mejor aproximación a \mathbf{x}^* o reducir el tamaño del simplex y comenzar de nuevo el proceso a partir de la aproximación que se acaba de obtener. Esta es la regla número 3.

- Regla 3** Cuando un vértice permanece como el de más alto valor en más de M simplex consecutivos de arista a , debe reducirse la arista del simplex y comenzar de nuevo a partir del vértice en cuestión, o terminar el proceso y aceptar a dicho vértice como la solución del problema con error absoluto máximo a . El valor de M que sugieren los autores del método, depende del número n de variables independientes y viene dado por:

$$M = 1,65 n + 0,05 n^2$$

En la tabla 1 se muestra el valor de M para varias dimensiones n :

n	M
2	3,5
3	5,4
4	7,4
5	9,5
6	11,7
7	14,0
8	16,4
9	18,9
10	21,5

Tabla 1

así, por ejemplo, para el caso de funciones de dos variables, se sugiere que cuando el vértice óptimo se repita en cuatro simplex consecutivos, debe aplicarse la regla número 3, lo cual significa que, en el caso mostrado en la figura 5, después del simplex número 7 se debió aplicar ya la regla 3.

Cálculo del simplex inicial

Si el procedimiento secuencial se comenzará a partir del punto \mathbf{x}_0 , es necesario calcular las coordenadas de los vértices del simplex inicial sabiendo que uno de los vértices posee coordenadas \mathbf{x}_0 . Todos los restantes vértices se hallan a una distancia a de \mathbf{x}_0 y hay una distancia a entre dos cualesquiera de ellos. Se pueden definir muchos simplex que cumplan estas restricciones y, cuál de ellos se elija, no tiene mucha importancia. Para simplificar los cálculos se supone el simplex colocado en una posición que guarde simetría con todas las variables. En la figura 6 se muestra el simplex inicial para un caso de dos variables y para un caso de tres variables, suponiendo que el vértice inicial se hace corresponder con el origen de un sistema coordenado. En el caso bidimensional, las coordenadas de los tres vértices del simplex inicial son de la forma:

	u_1	u_2
Vértice 1:	0	0
Vértice 2:	p	q
Vértice 3:	q	p

donde p y q son parámetros a determinar. En el caso tridimensional, las coordenadas serían:

	u_1	u_2	u_3
Vértice 1:	0	0	0
Vértice 2:	p	q	q
Vértice 3:	q	p	q
Vértice 4:	q	q	p

donde, de nuevo, los valores de p y q deben ser determinados.

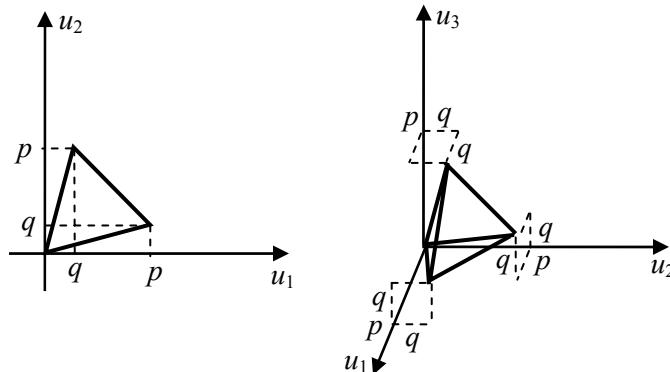


Figura 6

En general, para un caso en el espacio n dimensional, las coordenadas de los vértices, referidas a un sistema con el origen en el punto \mathbf{x}_0 , serían:

	u_1	u_2	u_3	...	u_n
Vértice 1:	0	0	0	...	0
Vértice 2:	p	q	q	...	q
Vértice 3:	q	p	q	...	q
Vértice 4:	q	q	p	...	q
⋮	⋮	⋮	⋮		⋮
Vértice $n + 1$:	q	q	q	...	p

Para hallar los valores de p y q basta imponer dos condiciones:

La distancia del vértice 1 a cualquiera de los otros debe ser a , es decir:

$$p^2 + (n-1)q^2 = a^2 \quad (1)$$

y la distancia entre dos cualesquier de los vértices $2, 3, \dots, n + 1$ debe ser también igual a a , esto es:

$$2(p-q)^2 = a^2 \quad (2)$$

De la ecuación (2) resulta:

$$p = \frac{a}{\sqrt{2}} + q \quad (3)$$

Sustituyendo en (1):

$$\left(\frac{a}{\sqrt{2}} + q \right)^2 + (n-1)q^2 = a^2$$

De donde

$$\frac{a^2}{2} + \sqrt{2}aq + q^2 + nq^2 - q^2 = a^2$$

Multiplicando por 2 y ordenando: $2nq^2 + 2\sqrt{2}aq - a^2 = 0$

$$\text{y, de aquí: } q = \frac{-2\sqrt{2}a + \sqrt{8a^2 + 8na^2}}{4n} = \frac{-2\sqrt{2}a + 2\sqrt{2}a\sqrt{n+1}}{4n}$$

Esto es:

$$q = \frac{a(\sqrt{n+1} - 1)}{\sqrt{2n}} \quad (4)$$

Sustituyendo en (3) se halla p : $p = \frac{a}{\sqrt{2}} + \frac{a(\sqrt{n+1} - 1)}{\sqrt{2n}}$

$$p = \frac{a}{\sqrt{2n}} \left(n + \sqrt{n+1} - 1 \right) \quad (5)$$

Nótese que, el hecho de que el sistema formado por las ecuaciones (1) y (2), posea solución, dada por las expresiones (4) y (5), prueba que la forma propuesta para los vértices del simplex es válida para cualquier dimensión n .

Ejemplo 1

Determine las coordenadas de los vértices de un simplex inicial para un problema de dimensión $n = 3$ si el punto donde comenzará el proceso es $\mathbf{x}_0 = (2,734; 1,281; -1,342)$ y la arista del simplex se ha seleccionado con longitud $a = 0,3$.

Solución:

A partir de las ecuaciones (4) y (5) se obtienen los valores de p y q :

$$p = \frac{a}{\sqrt{2n}}(n + \sqrt{n+1} - 1) = \frac{0,3}{3\sqrt{2}}(3 + \sqrt{3+1} - 1) = 0,2828$$

$$q = \frac{a(\sqrt{n+1} - 1)}{\sqrt{2n}} = \frac{0,3(\sqrt{3+1} - 1)}{3\sqrt{2}} = 0,0707$$

Con estos valores de p y q se pueden escribir las coordenadas de los cuatro vértices relativas al punto inicial \mathbf{X}_0 :

Vértice 1: $(0, 0, 0)$

Vértice 2: $(p, q, q) = (0,2828; 0,0707; 0,0707)$

Vértice 3: $(q, p, q) = (0,0707; 0,2828; 0,0707)$

Vértice 4: $(q, q, p) = (0,0707; 0,0707; 0,2828)$

Para tener las coordenadas absolutas de los cuatro vértices, basta sumar las coordenadas relativas con las del punto inicial \mathbf{X}_0 :

Coordenadas de los vértices:

Vértice 1: $(0, 0, 0) + (2,734; 1,281; -1,342) = (2,734; 1,281; -1,342)$

Vértice 2: $(0,2828; 0,0707; 0,0707) + (2,734; 1,281; -1,342) = (3,0168; 1,3517; -1,2713)$

Vértice 3: $(0,0707; 0,2828; 0,0707) + (2,734; 1,281; -1,342) = (2,8047; 1,5638; -1,2713)$

Vértice 4: $(0,0707; 0,0707; 0,2828) + (2,734; 1,281; -1,342) = (2,8047; 1,3517; -1,0592)$

Determinación de los vértices de un nuevo simplex

Para completar el algoritmo del simplex secuencial, se necesitan expresiones que permitan calcular analíticamente las coordenadas de los vértices de un nuevo simplex a partir de las del simplex anterior. Evidentemente, solo se necesita calcular las coordenadas de un vértice, ya que dos simplex consecutivos comparten n de sus $n + 1$ vértices. Para simplificar la deducción, el análisis que sigue se basa en la geometría del caso de dimensión 3, aunque las fórmulas se obtienen para el caso general de dimensión n .

La figura 7 muestra dos simplex consecutivos. Se ha llamado $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, a los vértices que comparten ambos simplex, \mathbf{S} al vértice que sale, es decir que forma parte del simplex anterior pero no del siguiente, y \mathbf{E} al vértice que entra, o sea, que forma parte del simplex nuevo pero no del anterior.

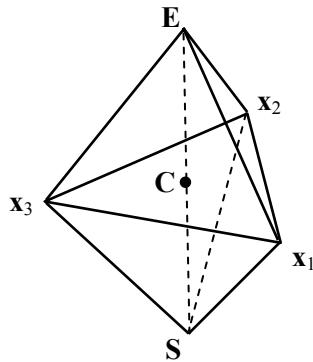


Figura 7

En la figura se muestra, además, el segmento que une a los vértices **E** y **S** y el punto **C** en que este segmento interseca al plano formado por los vértices que comparten ambos simplex. El punto **C** es el centro del segmento **ES** y es también el centroide del conjunto de vértices $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ que comparten ambos simplex. Por ser **C** el centro del segmento **ES** se cumple que:

$$\mathbf{C} = \frac{1}{2}(\mathbf{E} + \mathbf{S})$$

y por ser el centroide del conjunto de vértices $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, satisface la ecuación:

$$\mathbf{C} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

Igualando los miembros de la derecha de ambas igualdades se obtiene:

$$\frac{1}{2}(\mathbf{E} + \mathbf{S}) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

de donde:

$$\mathbf{E} = \frac{2}{n} \sum_{i=1}^n \mathbf{x}_i - \mathbf{S} \quad (6)$$

La ecuación (6) permite determinar las coordenadas del nuevo vértice **E** a partir de las coordenadas de los vértices del simplex actual y de la decisión de cual de ellos será el que dejará de formar parte del simplex, lo cual se decide a partir de las reglas 1 y 2.

Ejemplo 2

En el ejemplo 1 fueron halladas las coordenadas de los cuatro vértices de un simplex en el espacio tridimensional. Suponga que el vértice 2 será eliminado del simplex y halle las coordenadas del nuevo vértice **E** que formará parte del siguiente simplex.

Solución:

Aplicando la fórmula (6) resulta: $\mathbf{E} = \frac{2}{3}(\text{Vértice 1} + \text{Vértice 2} + \text{Vértice 4}) - \text{Vértice 2}$

ya que ya que los vértices 1, 3 y 4 son los que comparten ambos simplex y el vértice 2 es el que sale. Se obtiene:

$$\mathbf{E} = \frac{2}{3} [(2,734; 1,281; -1,342) + (2,8047; 1,5638; -1,2713) + (2,8047; 1,3517; -1,0592)] - \\ (3,0168; 1,3517; -1,2713) \\ \mathbf{E} = (2,5455; 1,4460; -1,1770)$$

Algoritmo en seudo código

El siguiente algoritmo permite hallar el punto de máximo de una función de n variables independientes linealmente unimodal mediante el método del simplex secuencial. Se utiliza la notación u_{ij} para denotar a la j -sima ($j = 1, 2, \dots, n$) componente del vector \mathbf{x}_i ($i = 0, 1, \dots, n$). El algoritmo utiliza como datos la función $f(\mathbf{x})$, el número n de variables independientes de dicha función, el punto $\mathbf{X}_0 = (u_{01}, u_{02}, \dots, u_{0n})$ donde comenzará el proceso y la longitud a de la arista del simplex y ofrece como resultado una aproximación al punto de máximo \mathbf{X}^* con error absoluto menor que a . Si este error no es lo suficientemente pequeño, todo el algoritmo se repite para un valor menor de a a partir del resultado anterior.

```

 $p := \frac{a}{\sqrt{2n}}(n + \sqrt{n+1} - 1)$ 
 $q := \frac{a(\sqrt{n+1} - 1)}{\sqrt{2n}}$ 
for  $i = 1$  to  $n$ 
    for  $j = 1$  to  $n$  {En este ciclo se forma el simplex inicial, cuyo vértice  $\mathbf{x}_0$  es dato}
        if  $j = i$  then  $u_{ij} := u_{0j} + p$  else  $u_{ij} := u_{0j} + q$ 
    end
end
for  $i = 1$  to  $n$ 
     $f_i := f(\mathbf{x}_i)$ 
     $R_i := 1$  {La variable  $R_i$  cuenta la cantidad de veces que el vértice  $i$  ha permanecido formando parte del simplex. Inicialmente todos valen 1}
end
 $mejor := i$  tal que  $f_i = \max \{f_0, f_1, \dots, f_n\}$ 
 $peor := i$  tal que  $f_i = \min \{f_0, f_1, \dots, f_n\}$ 
 $casipeor := i$  tal que  $f_i = \min[\{f_0, f_1, \dots, f_n\} - \{f_{peor}\}]$ 
 $último := mejor$  {último es el sub índice del último vértice que entró a formar parte del simplex. Al inicio se hace coincidir con el mejor}
 $M := 1,65 n + 0,05 n^2$ 
repeat
    if  $peor \neq último$  then  $k := peor$  else  $k := casipeor$  { $k$  indica el sub-índice del vértice que será cambiado}
     $último := k$ 
     $\mathbf{x}_k := \frac{2}{n} \sum_{i \neq k} \mathbf{x}_i - \mathbf{x}_k$  {Se hallan las coordenadas del nuevo vértice}
    for  $i := 0$  to  $n$  {Se actualizan los valores de  $R_i$ }
        if  $i = k$  then  $R_i := 1$  else  $R_i := R_i + 1$ 
    end
     $f_k := f(\mathbf{x}_k)$ 
     $mejor := i$  tal que  $f_i = \max \{f_0, f_1, \dots, f_n\}$ 
     $peor := i$  tal que  $f_i = \min \{f_0, f_1, \dots, f_n\}$ 
     $casipeor := i$  tal que  $f_i = \min[\{f_0, f_1, \dots, f_n\} - \{f_{peor}\}]$ 
until  $R_{mejor} > M$ 

```

El punto de máximo es \mathbf{x}_{mejor} con error absoluto menor que a .
Terminar

Ejemplo 3

Halle el punto de máximo de la función cuadrática $f(\mathbf{x}) = 100 - 3u_1^2 - 4u_2^2 + 5u_1u_2 + 2u_1$ por el método del simplex secuencial con precisión de tres cifras decimales exactas tomando $\mathbf{x}_0 = (5, 5)$.

Solución:

El problema se resolverá en varias etapas de búsqueda en las cuales se disminuye paulatinamente el valor de a hasta llegar a 0,0005 para obtener el resultado con tres cifras decimales exactas.

Etapa 1

$$\mathbf{x}_0 = (5, 5) \quad a = 1$$

La tabla 2 muestra los resultados de esta etapa. En cada simplex se muestra, junto al valor de la función en cada vértice, entre paréntesis las veces que ese vértice ha estado formando parte del simplex y un signo + ó – que señalará el vértice donde la función toma el mayor y el menor valor en este simplex. Por razones de espacio, las entradas de la tabla solo muestran las cifras decimales necesarias para la aplicación de las reglas correspondientes. Del mismo modo, en las coordenadas del mejor vértice (el señalado con +) solo se han incluido las cifras decimales más exactas, de acuerdo con el valor de a que se esté usando. En cada renglón de la tabla se señala la regla que será aplicada de inmediato. El proceso se detiene cuando el vértice (+) se mantiene en cuatro simplex sucesivos, ya que para el caso de $n = 2$ el valor de M es 3,5.

Simplex	f_0	f_1	f_2	Regla	Coordenadas de (+)
1	60,0(1)+	51,4(1)	42,1(1)-	1	5,00 5,00
2	60,0(2)	51,4(2)-	62,5(1)+	1	5,71 4,29
3	60,0(3)-	72,6(1)+	62,5(2)	1	4,74 4,03
4	68,2(1)	72,6(2)+	62,5(3)-	1	4,74 4,03
5	68,2(2)-	72,6(3)	79,8(1)+	1	4,48 3,07
⋮	⋮	⋮	⋮	⋮	⋮
13	99,1(1)+	96,9(3)	92,8(2)-	1	1,33 1,33
14	99,1(2)+	96,9(4)	96,4(1)-	2	1,33 1,33
15	99,1(3)	100,1(1)+	96,4(2)-	1	1,07 0,36
16	99,1(4)	100,1(2)+	97,7(1)-	2	1,07 0,36
17	100,18(1)+	100,12(3)	97,70(2)-	1	0,10 0,10
18	100,18(2)+	100,12(4)	95,74(1)-	2	0,10 0,10
19	100,18(3)+	97,30(1)	95,74(2)-	1	0,10 0,10
20	100,18(4)+	97,30(2)	96,61(2)-	3	0,10 0,10

Tabla 2

El proceso se detuvo en el punto $u_1 = 0,101021; u_2 = 0,101021$

Etapa 2

$$\mathbf{x}_0 = (0,101021; 0,101021); \alpha = 0,1$$

La tabla 3 muestra los resultados. Como los valores de la función en los tres vértices son muy similares, solo se muestra la cantidad que excede de 100 y se brindan tres o cuatro cifras decimales, según la necesidad. El proceso se detiene cuando el mejor valor obtenido se mantiene en 4 simplex consecutivos. Nótese como el final de la etapa se caracteriza por el empleo frecuente de la regla 2.

El mejor resultado de esta etapa es el punto $u_1 = 0,732340; u_2 = 0,449498$

Simplex	f_0	f_1	f_2	Regla	Coordenadas de (+)
	100,...	100,...	100,...		
1	182(1)	339(1)+	175(1)-	1	0,198 0,127
2	182(2)-	339(2)+	277(1)	1	0,198 0,127
3	383(1)+	339(3)	277(2)-	1	0,268 0,056
4	383(2)	339(4)-	460(1)+	1	0,294 0,153
5	383(2)-	453(4)	460(1)+	1	0,294 0,153
:	:	:	:	:	:
13	6463(4)	6733(2)+	6457(1)-	2	0,610 0,327
14	6877(1)+	6733(3)	6457(2)-	1	0,636 0,424
15	6877(2)+	6733(4)	6641(1)-	2	0,636 0,424
16	6877(3)	6934(1)+	6641(2)-	1	0,732 0,449
17	6877(4)	6934(2)+	6485(1)-	2	0,732 0,449
18	6692(1)	6934(3)+	6485(2)-	1	0,732 0,449
19	6692(2)	6934(4)+	6628(1)-	3	0,732 0,449

Tabla 3

Etapa 3

$$\mathbf{x}_0 = (0,732340; 0,449498), \alpha = 0,05$$

Los resultados se muestran en la tabla 4 en la cual se aprecia que fue necesaria una cantidad de simplex mucho menor, debido a que la reducción de α fue también menor que en otras etapas. Se obtuvo, como resultado el punto: $(0,684044; 0,436557)$.

Simplex	f_0	f_1	f_2	Regla	Coordenadas de (+)
	100,6...	100,6...	100,6...		
1	934(1)+	827(1)–	880(1)	1	0,7323 0,4495
2	934(2)+	860(1)–	880(2)	2	0,7323 0,4495
3	934(3)	860(2)–	951(1)+	1	0,6840 0,4366
4	934(4)	855(1)–	951(2)+	2	0,6840 0,4366
5	909(1)	855(2)–	951(3)+	1	0,6840 0,4366
6	909(2)	877(1)–	951(4)+	3	0,6840 0,4366

Tabla 4

Etapa 4

$$\mathbf{x}_0 = (0,684044; 0,436557), \alpha = 0,005$$

Los resultados se muestran en la tabla 5. Recuérdese que, para abreviar, los valores de la función que se muestran en cada fila son las cifras que siguen a 100,69. Se obtuvo como resultado el punto: (0,697239; 0,435610) que pasa a ser el punto de partida de la siguiente y última etapa.

Simplex	f_0	f_1	f_2	Regla	Coordenadas de (+)
	100,69...	100,69...	100,69...		
1	51(1)	54(1)+	48(1)–	1	0,6889 0,4379
2	5132(2)–	5373(2)	5515(1)+	1	0,6876 0,4330
3	5627(1)+	5373(3)–	5515(2)	1	0,6924 0,4343
4	5627(2)+	5598(1)	5515(3)–	1	0,6924 0,4343
5	5627(3)+	5598(2)	5582(1)–	2	0,6924 0,4343
6	5627(4)	5648(1)+	5582(2)–	1	0,6972 0,4356
7	5627(5)	5648(2)+	5522(1)–	2	0,6972 0,4356
8	5581(1)	5648(3)+	5522(2)–	1	0,6972 0,4356
9	5581(2)	5648(4)+	5579(1)–	3	0,6972 0,4356

Tabla 5

Etapa 5

$$\mathbf{x}_0 = (0,697239; 0,435610), \alpha = 0,0005$$

Los resultados se muestran en la tabla 6. Recuérdese que, para abreviar, los valores de la función que se muestran en cada fila son las cifras que siguen a 100,6956.

Se obtuvo, como resultado el punto: (0,695661; 0,434739) que es el punto de máximo buscado con tres cifras decimales exactas. Como se recordará, el valor exacto de \mathbf{X}^* es (0,6956521; 0,4347826) por lo cual, realmente, la respuesta hallada tiene 4 cifras decimales exactas.

Simplex	f_0	f_1	f_2	Regla	Coordenadas de (+)
	100,6956...	100,6956...	100,6956...		
1	4844(1)+	4556(1)-	4771(1)	1	0,6972 0,4356
2	4844(2)	4932(1)+	4771(2)-	1	0,6969 0,4360
3	4844(3)-	4932(2)	5042(1)+	1	0,6968 0,4355
4	5001(1)	4932(3)-	5042(2)+	1	0,6968 0,4355
5	5001(2)-	5149(1)+	5042(3)	1	0,6963 0,4354
6	5019(1)-	5149(2)+	5042(4)	2	0,6963 0,4354
7	5019(2)-	5149(3)	5163(1)+	1	0,6961 0,4349
8	5165(1)+	5149(4)-	5163(2)	1	0,6958 0,4352
9	5165(2)	5256(1)+	5163(3)-	1	0,6957 0,4347
10	5165(3)	5256(2)+	5090(1)-	2	0,6957 0,4347
11	5179(1)	5256(3)+	5090(2)-	1	0,6957 0,4347
12	5179(2)	5256(4)+	5134(1)-	3	0,6957 0,4347

Tabla 6

Ejemplo 4

Halle el punto de mínimo de la función de Rosembrook (vea el ejemplo 3 de la sección 6.5)

$$f(\mathbf{x}) = 100(u_2 - u_1^2)^2 + (1 - u_1)^2$$

mediante el método del simplex secuencial y compare su efectividad con los métodos anteriormente estudiados.

Solución:

Como se recordará, la función de Rosembrook está especialmente diseñada para que su tratamiento sea sumamente difícil. La presencia de “valles estrechos” hace que todos los métodos confronten serios problemas en la búsqueda del punto de mínimo, el cual se encuentra en

$$\mathbf{x}^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

El método simplex mal interpreta los puntos en el fondo del valle estrecho de esta función como si fuera el punto de mínimo y es necesario utilizar valores de a sumamente pequeños. A continuación se muestran los resultados finales de las 10 etapas utilizadas para hallar el punto de mínimo con error menor de 0,002. La primera etapa de búsqueda comienza en $(-1; 1)$. Las demás lo hacen en el resultado de la etapa precedente.

Etapa 1: $a = 0,01$	Evaluaciones: 511	Resultado: (0,55296; 0,30436)
Etapa 2: $a = 0,005$	Evaluaciones: 155	Resultado: (0,709335; 0,503162)
Etapa 3: $a = 0,002$	Evaluaciones: 452	Resultado: (0,867905; 0,753650)
Etapa 4: $a = 0,001$	Evaluaciones: 288	Resultado: (0,917658; 0,842300)
Etapa 5: $a = 0,0005$	Evaluaciones: 545	Resultado: (0,960028; 0,921793)
Etapa 6: $a = 0,0002$	Evaluaciones: 775	Resultado: (0,983935; 0,968186)

Etapa 7: $a = 0,0001$	Evaluaciones: 563	Resultado: (0,992300; 0,984683)
Etapa 8: $a = 0,00005$	Evaluaciones: 531	Resultado: (0,996134; 0,992300)
Etapa 9: $a = 0,00002$	Evaluaciones: 685	Resultado: (0,998207; 0,996424)
Etapa 10: $a = 0,00001$	Evaluaciones: 607	Resultado: (0,999111; 0,998226)

Cantidad de evaluaciones: 5112

La comparación con los métodos anteriormente estudiados resulta difícil, por cuanto todos ellos estaban basados en búsquedas unidimensionales. Si se supone, lo cual no es exagerado, que en cada búsqueda unidimensional se necesite evaluar la función en 20 puntos diferentes, entonces, los métodos anteriores habrían necesitado las siguientes cantidades de evaluaciones:

Búsqueda por coordenadas: 2000 búsquedas, unas 40000 evaluaciones, para llegar a

$$\mathbf{X}^* \approx \begin{bmatrix} 0,980 \\ 0,960 \end{bmatrix}$$

Método del gradiente: 1000 búsquedas unidimensionales, unas 20000 evaluaciones, para obtener:

$$\mathbf{X}^* \approx \begin{bmatrix} 0.9202 \\ 0.8468 \end{bmatrix}$$

Método de Powell: 29 búsquedas unidimensionales, unas 600 evaluaciones, para lograr:

$$\mathbf{X}^* \approx \begin{bmatrix} 1.000002 \\ 1.000004 \end{bmatrix}$$

Método del simplex secuencial: 5112 evaluaciones, para obtener:

$$\mathbf{X}^* \approx \begin{bmatrix} 0,999111 \\ 0,998226 \end{bmatrix}$$

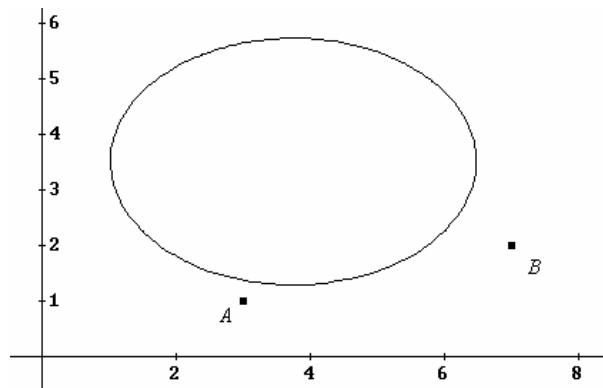
Como se observa, el método de Powell se comporta más eficientemente en el trabajo con la función de Rosembrook. Sin embargo, el método del simplex secuencial aventaja ampliamente a los métodos del gradiente y de búsqueda por coordenadas. Si se tiene en cuenta la dificultad de los algoritmos, el método del simplex secuencial posee la ventaja de no necesitar de algoritmos auxiliares de búsqueda unidimensional, cuya elaboración puede ser bastante compleja. Por otra parte, el método del simplex es, por la simplicidad de su lógica, un algoritmo sumamente robusto y confiable.

Ejercicios

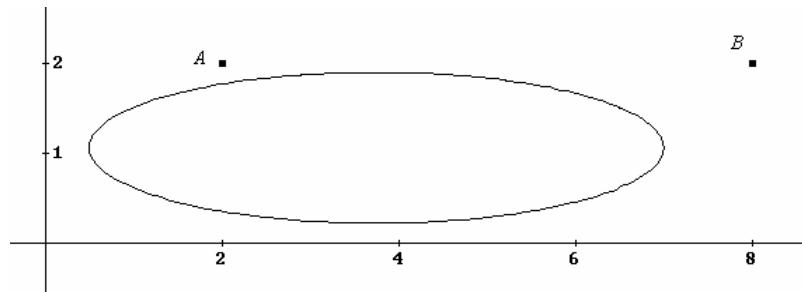
Algunos de los ejercicios que siguen son muy laboriosos para realizar los cálculos a mano. Se supone que posea programas computacionales, preferiblemente confeccionados por usted, para ayudarle en el trabajo. De no ser así, en los ejercicios que requieran mucho trabajo manual, determine los resultados con menos cifras exactas que las exigidas.

1. Cada una de las funciones cuadráticas que siguen aparecieron también en los ejercicios de las tres secciones anteriores. En cada caso aplique el método del simplex secuencial, utilizando como punto de partida los puntos A y B que se dan en la propia figura. Obtenga la solución con tres cifras decimales exactas. Compare con la solución obtenida utilizando los métodos anteriores. Observe la curva de nivel que acompaña a cada función y diga si puede sacar alguna conclusión general como resultado de este ejercicio.

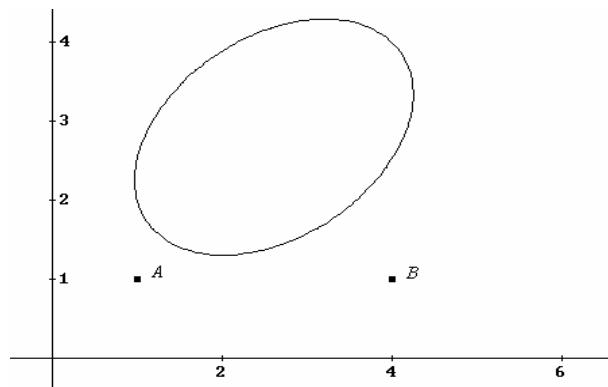
a) $f(x, y) = 2x^2 - 15x + 3y^2 - 21y; \quad A = (3, 1); \quad B = (7, 2)$



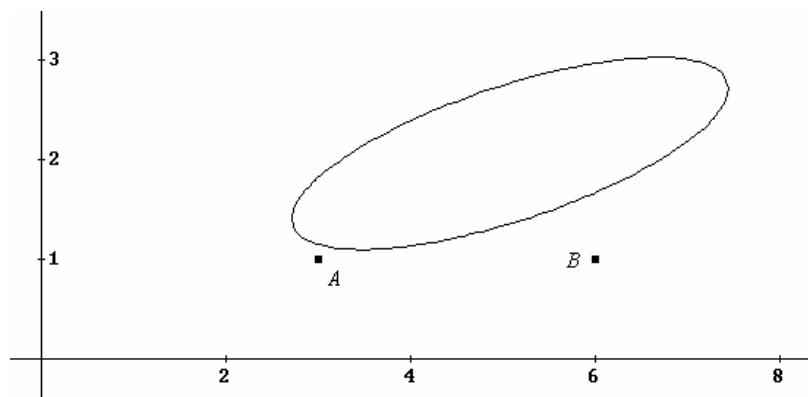
b) $f(x, y) = 2x^2 - 15x + 30y^2 - 63y; \quad A = (2, 2); \quad B = (8, 2)$



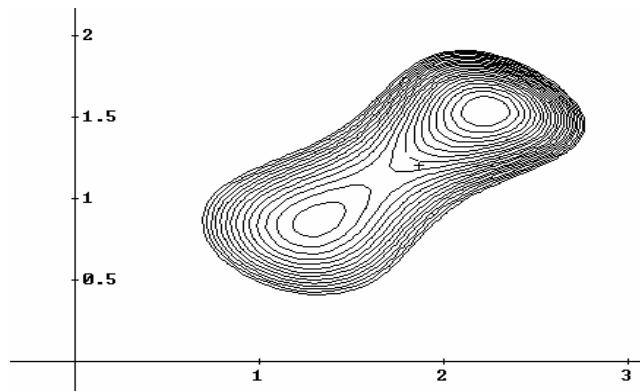
c) $f(x, y) = 5x^2 - 4xy + 6y^2 - 15x - 23y; \quad A = (1, 1); \quad B = (4, 1)$



d) $f(x, y) = 3x^2 - 10xy + 18y^2 - 10x - 23y; \quad A = (3, 1); \quad B = (6, 1)$



2. La función $f(x, y) = x^2 + \operatorname{sen}(xy) + 2y^2 - 3x - 4y$ no es unimodal. En la figura se observan algunas curvas de nivel que muestran que en esta región ella presenta dos puntos de mínimo. Seleccione adecuadamente el punto inicial \mathbf{x}_0 para determinar ambos puntos de mínimo mediante el método del simplex secuencial. Obtenga los resultados con tres cifras decimales exactas.



3. Ajuste el modelo no lineal $g(x) = a + be^{px}$ a los puntos que se muestran en la tabla. Obtenga el resultado con error menor que 0,001.

x	1	2	3	4	5
y	2,10	2,60	2,82	2,92	2,96

4. Halle la distancia más corta entre las superficies $z = 4 + 2x^2 + 3y^2$ y $y = 2 + 3x^2 + z^2$. Obtenga la solución con tres cifras decimales exactas.
5. En el algoritmo del simplex secuencial aparecen los pasos:

$$\begin{aligned} \text{mejor} &:= i \text{ tal que } f_i = \max \{f_0, f_1, \dots, f_n\} \\ \text{peor} &:= i \text{ tal que } f_i = \min \{f_0, f_1, \dots, f_n\} \end{aligned}$$

Elabore un algoritmo en seudo código que determine *mejor* y *peor* en un solo lazo.

Otras lecturas recomendadas

El tema de optimización numérica es sumamente amplio y en el texto solo han sido tocados los métodos más generalmente reconocidos, por lo cual existe una gran cantidad de algoritmos que ni siquiera han sido mencionados. En particular, se ha obviado el problema de la optimización multidimensional con restricciones. Al lector interesado en ampliar sus conocimientos con otros métodos o estudiar otros enfoques, puede consultar la enciclopédica obra “Optimization, Theory and Practice” de Bevridge y Schechter, el cual ha sido reproducido en Cuba. Con un enfoque más moderno, pero de lectura un poco más difícil, puede consultarse el libro “Optimization” de Luemberger, que también ha sido editado en Cuba.

Principales ideas del capítulo

- La función $f(x)$ es unimodal con máximo si existe en su dominio un x^* tal que, si x_1 y x_2 pertenecen al dominio se cumple que: Si $x_1 < x_2 < x^*$ entonces $f(x_1) < f(x_2)$ y si $x^* < x_1 < x_2$ entonces $f(x_1) > f(x_2)$.
- La propiedad básica de la optimización unidimensional establece que: Si $f(x)$ una función unimodal con máximo en x^* y x_1 y x_2 son dos valores de su dominio y se denomina $y_1 = f(x_1)$ y $y_2 = f(x_2)$. Entonces $y_1 < y_2 \Rightarrow x_1 < x^*$; $y_1 > y_2 \Rightarrow x^* < x_2$ y $y_1 = y_2 \Rightarrow x_1 < x^* < x_2$
- Aquellos algoritmos en los que los experimentos se realizan todos a un tiempo, se llaman de *búsqueda simultánea* mientras que los que siguen la estrategia de realizar un experimento después que han sido analizados los resultados de los experimentos anteriores, se llaman *métodos secuenciales*.
- Si se realizan n experimentos simultáneos en los valores $x_1 < x_2 < \dots < x_n$ con resultados y_1, y_2, \dots, y_n y se obtiene $y_k = \max \{y_i\}$ con $1 < k < n$, entonces el punto de máximo se encuentra entre x_{k-1} y x_{k+1} .
- El método de búsqueda secuencial uniforme consiste en generar la sucesión de valores: $x_i = x_0 + is$ ($i = 0, 1, 2, 3, \dots$) y obtener para cada uno de ellos la imagen $y_i = f(x_i)$. El proceso se detiene tan pronto se obtiene un y_k tal que $y_{k-1} > y_k$, y puede entonces asegurarse, se acuerdo con el principio básico, que x^* se encuentra entre x_{k-2} y x_k .
- La búsqueda secuencial acelerada consiste en lo siguiente: en cada paso del algoritmo mientras no se cumpla la condición de parada, el paso se duplica en valor; de esta manera, aun cuando inicialmente s fuera demasiado pequeño, pronto toma valores suficientemente grandes para alcanzar a x^* en no muchas iteraciones.

- En el método de bisección se toman dos puntos experimentales x_1 y x_2 muy próximos entre sí y a ambos lados del centro del intervalo $[a, b]$. Si se llama $y_1 = f(x_1)$ y $y_2 = f(x_2)$, entonces se tiene que $y_1 < y_2 \Rightarrow x_1 \leq x^* \leq b$; $y_1 > y_2 \Rightarrow a \leq x^* \leq x_2$ y $y_1 = y_2 \Rightarrow x_1 \leq x^* \leq x_2$.
- En el método de bisección se cumple que: $\frac{L_n}{L_0} = \frac{1}{2^{n/2}}$, donde L_n representa la longitud del intervalo de búsqueda después de haber realizado n experimentos.
- El método de la sección áurea utiliza los números D_n definidos mediante las condiciones: $D_n = D_{n-1} + D_{n-2}$ ($n = 2, 3, 4, \dots$) y $D_n = r D_{n-1}$ ($n = 1, 2, 3, \dots$). Se demuestra que: $r = 1,618034\dots$ y $G = \frac{1}{r} = r - 1 = 0,618034$. En cada paso del método la longitud del intervalo de búsqueda coincide con uno de los números D_n y los experimentos se realizan a distancia D_{n-2} de los extremos. De esta forma se logra que en cada iteración se requiere solamente un nuevo experimento.
- En el método de la sección áurea se cumple que $\frac{L_n}{L_0} = G^{n-1}$ y resulta mucho menor que para el método de la bisección.
- La función $z = f(\mathbf{x})$ se denomina linealmente unimodal con máximo en la región R , si ella es unimodal con máximo sobre cualquier trayectoria recta contenida en R .
- Una función cuadrática cualquiera f de n variables independientes se puede expresar como $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$ donde \mathbf{A} es una matriz $n \times n$ simétrica y \mathbf{b} es una matriz $n \times 1$.
- Una matriz simétrica $\mathbf{A}_{n \times n}$ se llama positiva definida si para todo vector no nulo \mathbf{x} se cumple que $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$. Si para todo \mathbf{X} no nulo se tiene $\mathbf{x}^T \mathbf{A} \mathbf{x} < 0$, la matriz \mathbf{A} se llama negativa definida.
- Si \mathbf{A} es positiva (negativa) definida la función cuadrática $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$ posee un punto \mathbf{x}^* de mínimo (máximo) y se cumple que $\mathbf{A}\mathbf{x}^* = \mathbf{b}$.
- Si \mathbf{A} es positiva (negativa) definida la función cuadrática con mínimo (máximo) $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$ es linealmente unimodal.
- El método de búsqueda por coordenadas consiste en realizar búsquedas sucesivas unidimensionales en las direcciones de los ejes de las variables independientes, es decir, haciendo variar una sola variable en cada búsqueda. Se trata de un método solamente recomendable cuando existe poca interacción entre las variables.
- El método del gradiente consiste en realizar sucesivas búsquedas unidimensionales en la dirección del vector gradiente. Presenta una buena rapidez de convergencia cuando las curvas de nivel son aproximadamente circulares pero se hace muy lento en presencia de valles estrechos.
- Si \mathbf{A} es una matriz cuadrada de orden n positiva (negativa) definida, las direcciones \mathbf{d}_i y \mathbf{d}_j se llaman conjugadas respecto a \mathbf{A} , o también \mathbf{A} -ortogonales, si $\mathbf{d}_i^T \mathbf{A} \mathbf{d}_j = 0$
- Si $f(\mathbf{x})$ es una función cuadrática con punto de extremo \mathbf{x}^* y $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{n-1}$ son direcciones conjugadas de su matriz \mathbf{A} , entonces el punto \mathbf{x}^* se halla mediante n búsquedas unidimensionales sucesivas en las direcciones conjugadas.
- El método de Powell consiste en realizar búsquedas unidimensionales en ciertas direcciones $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{n-1}$, que, aunque inicialmente pueden ser arbitrarias, en el transcurso del proceso se van convirtiendo en un conjunto de direcciones conjugadas, si la función que se optimiza es cuadrática. Se trata de un método muy eficiente aunque algo difícil de programar.
- Se llama *simplex* de un espacio n dimensional, al elemento con la misma dimensión del espacio que posea la estructura geométrica más simple posible. Así, en el espacio bidimensional (el plano), el simplex es un triángulo equilátero, el cual está limitado por tres

- segmentos iguales y posee tres puntos (vértices) que lo determinan y en el espacio tridimensional, el simplex es un tetraedro (pirámide de base triangular) regular, que está limitado por cuatro triángulos equiláteros y posee cuatro puntos (vértices) que lo determinan.
- El método del simplex secuencial se basa en evaluar la función a optimizar en los vértices de un simplex y, de acuerdo con los valores de la misma, formar otro simplex, sustituyendo uno de los vértices por un nuevo vértice, siguiendo dos reglas: 1) el vértice donde la función alcanza su peor valor es sustituido por uno nuevo, a menos que, 2) el peor vértice sea precisamente el último que entró a formar parte del simplex, en cuyo caso se elimina el vértice en el que la función alcanza es segundo peor valor. El método se detiene cuando el vértice donde la función alcanza el óptimo se mantiene muchas veces formando parte del simplex.

Auto examen

1. ¿Qué es un problema de optimización?
2. Cite algunos factores que limitan los procedimientos analíticos para optimizar funciones.
3. ¿Qué se entiende por función unimodal de una variable?
4. Explique cómo se utilizan en el método de la sección áurea las dos propiedades fundamentales de los números D_n : $D_n = D_{n-1} + D_{n-2}$ ($n = 2, 3, 4, \dots$) y $D_n = r D_{n-1}$ ($n = 1, 2, 3, \dots$).
5. Halle el valor de $x > 0$ más pequeño donde la función $x^{\sin x}$ posee un máximo.
6. Dos de los métodos más eficientes para optimizar funciones de varias variables son el método de Powell y el del simplex secuencial. Explique el fundamento de cada uno y analice las ventajas y desventajas relativas entre ellos.
7. Halle la mínima distancia entre la recta $x + y = 4$ y la cardiode $\rho = 3 - 3\cos\theta$. Obtenga el resultado con 3 cifras decimales exactas.

CAPÍTULO 7

Matemática Numérica, 2da Edición

Manuel Álvarez, Alfredo Guerra, Rogelio Lau

ECUACIONES DIFERENCIALES ORDINARIAS

Objetivos

Al finalizar el estudio y la ejercitación de este capítulo, el lector debe ser capaz de:

- Utilizar adecuadamente la terminología propia del tema de ecuaciones diferenciales ordinarias.
- Identificar problemas de condiciones iniciales y de condiciones de frontera.
- Comparar los métodos analíticos y los métodos numéricos que se utilizan para resolver ecuaciones diferenciales.
- Describir el concepto de campo de direcciones de una ecuación diferencial de primer orden.
- Analizar las características más sobresalientes de las soluciones de una ecuación diferencial de primer orden a partir de la observación del campo de direcciones.
- Describir el concepto de ecuación diferencial estable y de método numérico estable para resolver una ecuación diferencial.
- Interpretar geométricamente los métodos de Euler y de Runge – Kutta.
- Describir los métodos de Euler, Taylor y Runge – Kutta de orden 2 y 4 y Adams, su error de truncamiento y su estabilidad.
- Describir los conceptos de método de paso simple y de método de paso múltiple.
- Explicar la deducción de las fórmulas de Adams – Bashforth y de Adams – Moulton.
- Explicar el concepto de método predictor – corrector y exemplificarlo mediante los métodos predictor – corrector de Adams.
- Aplicar los métodos de Euler, Runge – Kutta y Adams a la solución de ecuaciones diferenciales de primer orden.
- Describir el concepto de problema de Cauchy de orden m .
- Describir los algoritmos de Runge – Kutta adaptados a la solución de problemas de Cauchy de orden m .
- Resolver ecuaciones diferenciales de orden superior con condiciones iniciales, mediante su transformación en problemas de Cauchy.
- Describir los pasos generales para resolver ecuaciones diferenciales ordinarias con condiciones de frontera mediante el método de los disparos.
- Analizar los algoritmos en seudo código correspondientes a todos los métodos estudiados en el capítulo.
- Modelar problemas sencillos que conducen a ecuaciones diferenciales ordinarias con condiciones iniciales o de frontera.

7.1 Introducción

Conceptos iniciales

Como se recordará de cursos previos, una ecuación diferencial es aquella en la que aparecen derivadas. Cuando hay una sola variable independiente respecto a la cual se plantean todas las derivadas, entonces las derivadas se llaman ordinarias y también la ecuación diferencial. Una gran cantidad de leyes físicas, biológicas, económicas, geométricas, etc., se plantean en forma concisa mediante una ecuación diferencial ordinaria o un conjunto de ellas. Algunos ejemplos sencillos de ecuaciones diferenciales ordinarias son:

$$\frac{dy}{dx} = x^3$$

$$\frac{dy}{dx} = 2y$$

$$\frac{dy}{dt} + 4ty = \sin t$$

$$\frac{d^2y}{dx^2} + x \frac{dy}{dx} + x^2 y = e^t$$

$$\begin{cases} \frac{dx}{dt} + 3y - \frac{dy}{dt} = t^2 \\ \frac{dy}{dt} + x + y + 2 \frac{dx}{dt} = t + 1 \end{cases}$$

Por supuesto, pueden diferir los nombres de las variables, el orden (dados por la derivada de mayor orden), el número de ecuaciones simultáneas, etc. En cualquier caso, el problema más importante que se presenta, una vez que la ecuación ha sido planteada, es hallar su solución. Se entiende por *solución*, una función (o un conjunto de funciones si se trata de un sistema de ecuaciones diferenciales ordinarias) que exprese la relación que debe existir entre las variables, de modo que la ecuación diferencial se satisfaga.

Ejemplo 1

Compruebe que $y = \frac{1}{4}x^4 + C_1$ e $y = C_2e^{2x}$, donde C_1 y C_2 son constantes reales, son soluciones, respectivamente, de las ecuaciones $\frac{dy}{dx} = x^3$ y $\frac{dy}{dx} = 2y$

Solución:

Si $y = \frac{1}{4}x^4 + C_1$ basta derivar para obtener $\frac{dy}{dx} = x^3$ independientemente del valor de C_1 . Así que, para cualquier constante C_1 , $y = \frac{1}{4}x^4 + C_1$ es solución de la ecuación diferencial ordinaria $\frac{dy}{dx} = x^3$.

En el segundo caso, al derivar $y = C_2e^{2x}$ se obtiene:

$$\frac{dy}{dx} = 2C_2e^{2x} = 2(C_2e^{2x}) = 2y$$

Luego, $y = C_2e^{2x}$ es solución de la ecuación $\frac{dy}{dx} = 2y$ para todo valor de C_2 . ■

Tipos de solución

Usualmente, una ecuación diferencial ordinaria posee infinitas soluciones que se expresan en una sola igualdad mediante el empleo de constantes arbitrarias. Así, las funciones que mostró el ejemplo 1 son las soluciones generales respectivas de las ecuaciones diferenciales que allí se mostraron.

En los cursos sobre ecuaciones diferenciales se prueba que el número de constantes arbitrarias esenciales que aparece en la solución general de una ecuación diferencial ordinaria coincide con el orden de la ecuación. De esto resulta que, si se desea precisar una única solución para una ecuación diferencial ordinaria de orden n , hay que suministrar n condiciones particulares que deba satisfacer la solución, de modo que no exista ambigüedad en la solución particular que se busca.

Ejemplo 2

Trace las gráficas de varias soluciones particulares de la ecuación diferencial ordinaria

$$\frac{dy}{dx} = 2y$$

y, en especial, aquella que satisface $y(1) = 2$.

Solución:

Como la solución general viene dada por $y = Ce^{2x}$, basta asignar a C valores arbitrarios. Tomando, por ejemplo, $C = 1, C = 2, C = 0, C = -\frac{1}{2}$, se tienen las soluciones particulares $y = e^{2x}, y = 2e^{2x}, y = 0, y = -\frac{1}{2}e^{2x}$, respectivamente. Si se exige que $y(1) = 2$, el valor de C debe satisfacer que:

$$2 = Ce^2$$

o sea, $C = 2e^{-2} = 0,27067$, y se tiene la solución:

$$y = 0,27067e^{2x}$$

En la figura 6.1 se muestran, aproximadamente, las cinco soluciones particulares halladas.

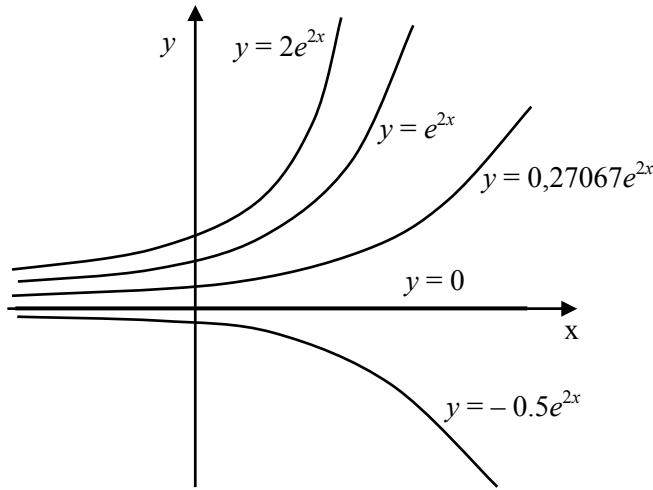


Figura 1

Condiciones iniciales y de frontera

Cuando se requiere más de una condición particular para determinar la solución que se desea, pueden darse dos situaciones muy diferentes:

- Que se especifiquen todas las condiciones particulares para un mismo valor de la variable independiente (*condiciones iniciales*).
- Que se incluyan condiciones particulares que deberá satisfacer la solución para dos o más valores de la variable independiente (*condiciones de frontera*).

Así:

$$\frac{d^2y}{dx^2} + 5 \frac{dy}{dx} + 6y = x^2 + \sin x$$
$$y(1) = 1$$
$$y'(1) = 3$$

es un problema de condiciones iniciales, mientras que

$$\frac{d^2y}{dx^2} + 5 \frac{dy}{dx} + 6y = x^2 + \sin x$$
$$y(1) = 2$$
$$y(2) = 5$$

es un problema de condiciones de frontera.

La diferencia entre ambos tipos de problema no es trivial. La forma de resolverlos es muy diferente y, usualmente, los problemas con condiciones de frontera son mucho más complicados. La mayor parte de este capítulo se dedicará a problemas con condiciones iniciales y solo al final, se estudiará un método para resolver ecuaciones diferenciales ordinarias con condiciones de frontera.

Limitaciones de los métodos analíticos

Ya en cursos anteriores se han tratado métodos *analíticos* para resolver ecuaciones diferenciales ordinarias. Acerca de ellos puede afirmarse que:

- Utilizan operaciones algebraicas, incluyendo la derivación y la integración, para obtener la solución general a partir de la ecuación diferencial.
- La solución particular deseada se halla a partir de la solución general, buscando valores adecuados para las constantes arbitrarias.
- Cada método analítico se ocupa de un tipo especial de ecuación diferencial ordinaria y es inaplicable en otros casos.
- A pesar de la diversidad de métodos analíticos, la mayoría de las ecuaciones diferenciales ordinarias no puede resolverse por esta vía.

Una gran cantidad de problemas se pueden modelar mediante ecuaciones diferenciales ordinarias que pueden resolverse analíticamente; en otros casos es necesario simplificar e incluso ignorar aspectos importantes, con tal de poder aplicar posteriormente un método analítico. En otras no pocas situaciones, las ecuaciones obtenidas no se pueden resolver analíticamente. A continuación se muestran algunos ejemplos clásicos.

Ejemplo 3: El péndulo simple

Considérese un cuerpo de masa m que cuelga de un soporte fijo mediante una cuerda de longitud L . Se suponen condiciones perfectamente ideales:

- La cuerda es perfectamente flexible.
- La cuerda no tiene masa.
- El aire no ejerce ninguna resistencia al movimiento del péndulo o la cuerda.

En estas condiciones, el péndulo se separa un ángulo θ_0 de la posición de reposo y se suelta en el instante $t = 0$. Se quiere conocer cómo cambia el ángulo θ de deflexión a lo largo del tiempo.

Solución:

En un instante cualquiera $t \geq 0$ el péndulo ocupará una posición θ (figura 2 a) y sobre él estarán actuando dos fuerzas externas: la atracción gravitacional mg y la fuerza de tensión T de la cuerda (figura 2 b).

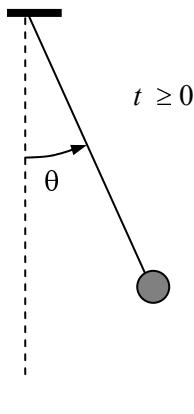


Figura 2 a

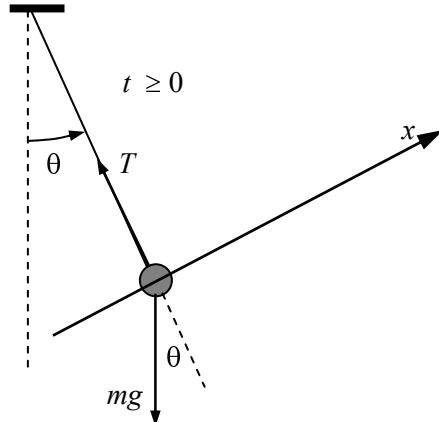


Figura 2 b

Tomando un eje x en el péndulo, tangente a su trayectoria (perpendicular a la cuerda), la segunda ley de Newton conduce a:

$$-mg \operatorname{sen} \theta = ma_x$$

donde a_x denota la aceleración tangencial del péndulo. Si se tiene en cuenta las leyes del movimiento circular:

$$a_x = L \frac{d^2\theta}{dt^2}$$

ya que L es el radio de la circunferencia que el péndulo describe. Queda entonces la ecuación:

$$-g \operatorname{sen} \theta = L \frac{d^2\theta}{dt^2}$$

que es una ecuación diferencial ordinaria de segundo orden.

A pesar de la apariencia sencilla de esta ecuación, su solución no se puede expresar en términos de funciones elementales y, por tanto, no es posible obtenerla por métodos analíticos. Si el análisis se limita a ángulos de deflexión muy pequeños, en los cuales la aproximación

$$\operatorname{sen} \theta \approx \theta$$

introduzca un error permisible, la ecuación se simplifica considerablemente y se puede encontrar una solución por la teoría de las ecuaciones diferenciales lineales con coeficientes constantes; pero esto solo permitiría considerar casos en los que el péndulo no se aleje más de cinco o seis grados de la posición de reposo.

Para que el problema del péndulo simple quede completo hay que añadir dos condiciones iniciales (la ecuación es de segundo orden) que están establecidas en el enunciado del ejemplo. El problema quedaría así:

Hallar $\theta(t)$ para $t \geq 0$ tal que:

$$-g \operatorname{sen} \theta = L \frac{d^2\theta}{dt^2}$$

$$\theta(0) = \theta_0$$

$$\theta'(0) = 0$$

donde L y g son constantes conocidas. Este problema será resuelto al final del capítulo.

Ejemplo 4: Depredadores y presas

A principios del siglo XX, el matemático italiano Vito Volterra (e, independientemente el norteamericano de origen austriaco, A. J. Lotka) estudió la dinámica de un sistema biológico muy simple en el cual interactúan dos especies, que se llamarán *depredadores y presas*. Se supone las siguientes condiciones ideales:

- Ambas especies conviven en una región cerrada (un bosque, un lago, una isla) de modo que no llegan individuos nuevos, ni salen. El crecimiento y decrecimiento solo es causado por nacimientos y muertes.
- Las presas viven del medio ambiente, que se supone puede sostener una población ilimitada de presas. El único factor que limita el crecimiento de la población de presas es su depredador. En ausencia de depredadores, la población de presas crecería sin límite.
- Los depredadores solo se alimentan de las presas. En ausencia de presas, la población de depredadores desaparece.

Considérense las funciones del tiempo:

$x(t)$: número de presas en el instante t

$y(t)$: número de depredadores en el instante t

En ausencia de depredadores ($y = 0$) las presas aumentan según la ecuación:

$$\frac{dx}{dt} = ax$$

donde $a > 0$ representa la tasa de crecimiento natural de la especie, la cual tiene en cuenta las tasas de natalidad y de mortalidad.

En ausencia de presas ($x = 0$) la población de depredadores decrecerá con una rapidez proporcional al número de individuos:

$$\frac{dy}{dt} = -cy$$

donde $c > 0$ es la taza de decrecimiento. Si $x > 0$ e $y > 0$, la probabilidad de que un depredador y una presa se encuentren será proporcional al producto xy ; como esa probabilidad constituye un factor de crecimiento para la población de depredadores y de decrecimiento para las presas, es razonable suponer que

$$\begin{cases} \frac{dx}{dt} = ax - bxy \\ \frac{dy}{dt} = -cy + dxy \end{cases}$$

que se conocen con el nombre de *ecuaciones de Lotka - Volterra*. Aunque su apariencia es sencilla, la presencia del término xy hace que el sistema sea no lineal y su solución analítica no es posible. Sin embargo, conocidos los parámetros a, b, c, d y las poblaciones x e y en un instante inicial, se puede determinar numéricamente como variará cada una de las poblaciones. Se pospone la solución del problema para la sección 7.5.

Ejemplo 5: La órbita de un planeta

Una de los descubrimientos más notables de Isaac Newton fue la ley de gravitación universal. Esta establece que dos cuerpos de masas m_1 y m_2 , cuyos centros de masa están separados una distancia d , se atraen con una fuerza F que es proporcional al producto de las masas e inversamente proporcional al cuadrado de la distancia. Simbólicamente:

$$F = \frac{km_1m_2}{d^2}$$

donde k es la constante de gravitación universal: $k = 6.67 \cdot 10^{-8} \text{ (cm}^3/\text{g})\text{s}^2$

Puesto que esta es la única ley que rige el movimiento planetario (dentro de los límites de la mecánica Newtoniana) ella permite conocer cómo variará la posición de un planeta respecto al sol, suponiendo conocidas su posición y su velocidad en un instante inicial.

Solución:

Se tomará al sol como punto de referencia y se situará en él un sistema de dos ejes coordenados coplanar con la órbita del planeta y de modo que el eje x pase por la posición del mismo en el instante $t = 0$.

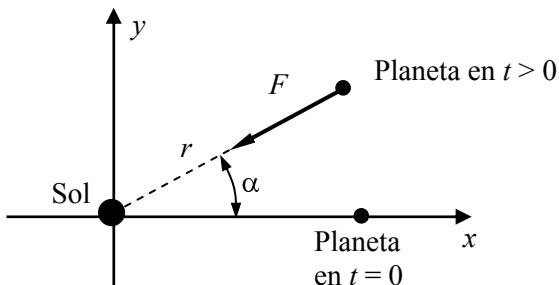


Figura 3

Considérese ahora el planeta en un instante $t > 0$ y sea r la distancia que lo separa del sol y α el ángulo que forma el radio vector con el eje x (figura 3). Para simplificar el problema, no se tendrá en cuenta ninguna otra influencia sobre el planeta, que no sea la fuerza F de atracción del sol.

Aplicando la segunda ley de Newton sobre ambos ejes coordenados, resulta:

$$\begin{cases} -F \cos \alpha = m_p a_x \\ -F \sin \alpha = m_p a_y \end{cases}$$

donde m_p es la masa del planeta y a_x y a_y son las componentes de la aceleración sobre los ejes x e y respectivamente. Según la ley de gravitación universal, las ecuaciones se transforman en:

$$\begin{cases} -\frac{km_s m_p}{r^2} \cos \alpha = m_p \frac{d^2 x}{dt^2} \\ -\frac{km_s m_p}{r^2} \sin \alpha = m_p \frac{d^2 y}{dt^2} \end{cases}$$

donde (x, y) son las coordenadas del planeta en el instante t , por lo cual:

$$\cos \alpha = \frac{x}{r} \quad \sin \alpha = \frac{y}{r} \quad r = \sqrt{x^2 + y^2}$$

Esto permite expresar las ecuaciones en términos de x e y :

$$\begin{cases} -\frac{km_s}{x^2 + y^2} \frac{x}{\sqrt{x^2 + y^2}} = \frac{d^2 x}{dt^2} \\ -\frac{km_s}{x^2 + y^2} \frac{y}{\sqrt{x^2 + y^2}} = \frac{d^2 y}{dt^2} \end{cases}$$

O sea:

$$\begin{cases} -\frac{km_s x}{(x^2 + y^2)^{3/2}} = \frac{d^2 x}{dt^2} \\ -\frac{km_s y}{(x^2 + y^2)^{3/2}} = \frac{d^2 y}{dt^2} \end{cases}$$

que forman un sistema de dos ecuaciones diferenciales no lineales de segundo orden. Las condiciones iniciales del problema son la posición y la velocidad del planeta (respecto al sol) en el instante $t = 0$; es decir:

$$x(0) = r_0$$

$$y(0) = 0$$

$$v_x(0) = v_{0x}$$

$$v_y(0) = v_{0y}$$

Al final del capítulo se estará en condiciones de resolver numéricamente este complicado problema.

Ejercicios

En las siguientes ecuaciones diferenciales diga su orden y determine si usted podría hallar su solución por alguno de los métodos analíticos que conoce. El objetivo de este ejercicio no es que usted repase todos los métodos analíticos que estudió en cursos pasados para resolver ecuaciones diferenciales, sino que se convenza de la gran variedad de tales procedimientos y del reducido número de ecuaciones que se pueden resolver con ellos; de ese modo apreciará mejor la enorme generalidad de los métodos numéricos que estudiará.

$$1. \quad (x^2 - xy)dx + (y^2 + x^2)dy = 0$$

$$2. \quad (x^2 - y)dx + (y^2 + x^2)dy = 0$$

$$3. \quad \frac{dy}{dx} + y = \frac{1}{x}$$

$$4. \quad \frac{dy}{dx} = x^2 + y^2$$

$$5. \quad \frac{dy}{dx} + \frac{y}{x} = xy^2$$

$$6. \quad (x + 2y + 1)dx + (x - 3y + 2)dy = 0$$

$$7. \quad \frac{dy}{dx} = -\frac{4}{x^2} - \frac{1}{x}y + y^2$$

$$8. \quad \frac{d^2y}{dx^2} + 4\frac{dy}{dx} = \tan x$$

$$9. \quad x^2 \frac{d^2y}{dx^2} - x \frac{dy}{dx} + y = \ln x$$

$$10. \quad (1 - x^2)y'' - 2xy' + n(n - 1)y = 0$$

$$11. \quad y'' - (\cos x)y' + y^2 = 0$$

$$12. \quad \frac{d^2y}{dx^2} = y^3$$

En cada uno de los siguientes ejercicios, determine si se trata de un problema de condiciones iniciales o de un problema con condiciones de frontera.

$$13. \quad \frac{d^2y}{dt^2} + t \frac{dy}{dt} - e^{-t}y = 4t; \quad y(0) = 0; \quad y(1) = 3$$

14. $\frac{d^3x}{dt^3} + \frac{dx}{dt} - x^2 = 0; \quad x(1) = 3; \quad x'(1) = 2; \quad x''(1) = -1$

15. $\frac{d^2y}{dx^2} + 3\frac{dy}{dx} - (\operatorname{sen} x)y = x + y; \quad y(2) = 1; \quad y'(2) = 3$

16. $\frac{d^2y}{dx^2} - 2xy = e^{-x}; \quad y(2) = 1; \quad y'(3) = 2$

17. $\frac{dy}{dx} = 4x + 3y^2; \quad y(2) = -3$

18. $\frac{dx}{dt} + 4xt = 3t^2; \quad y(0) = 1$

19.
$$\begin{cases} \frac{dx}{dt} - 3\frac{dy}{dt} = x \\ \frac{dy}{dt} + \frac{dx}{dt} = y + 1 \end{cases} \quad x(0) = 2 \quad y(0) = 3$$

20.
$$\begin{cases} \frac{dx}{dt} = y \\ \frac{dy}{dt} = z \\ \frac{dz}{dt} = -x + z - y \end{cases} \quad x(1) = 2 \quad y(1) = 1 \quad z(0) = 3$$

7.2 Ecuaciones diferenciales de primer orden

Como ya se sabe de cursos anteriores, una ecuación diferencial de primer orden es aquella en que sólo aparece la primera derivada de la función incógnita respecto a su variable independiente. En todos los casos de interés práctico, esta derivada puede expresarse en forma explícita en términos de las variables dependiente e independiente, así que en todo lo que sigue, se supondrá que la ecuación puede expresarse como

$$\frac{dy}{dx} = f(x, y)$$

Como los métodos numéricos sólo son capaces de hallar soluciones particulares de una ecuación diferencial, también se supone conocido el valor de la solución para algún valor de la variable independiente. Es decir, se trata de:

Hallar la solución del problema de Cauchy

$$\begin{aligned}\frac{dy}{dx} &= f(x, y) \\ y(x_0) &= y_0\end{aligned}\tag{1}$$

Ejemplo 1

Exprese de la forma (1) el problema:

$$\begin{aligned}(4x + y)dx + (2x + y^2 - 3)dy &= 0 \\ y(-1) &= 4\end{aligned}$$

Solución:

Despejando el cociente de diferenciales:

$$\begin{aligned}\frac{dy}{dx} &= -\frac{4x + y}{2x + y^2 - 3} \\ y(-1) &= 4\end{aligned}$$

Ejemplo 2

Proponga una ecuación diferencial de primer orden que no pueda ser expresada en la forma (1)

Solución:

Por ejemplo:

$$\frac{dy}{dx} + \operatorname{sen}\left(\frac{dy}{dx}\right) = x + y$$

ya que en una ecuación del tipo $u + \operatorname{sen} u = v$ no es posible despejar la variable u .

Las ecuaciones diferenciales de primer orden se presentan en multitud de problemas de la matemática aplicada pero no solo por esto son importantes. Resulta que cualquier ecuación diferencial ordinaria (o sistema de ecuaciones diferenciales ordinarias) con condiciones iniciales puede ser reducida a un conjunto de ecuaciones diferenciales de primer orden. Esto significa que, desarrollando métodos numéricos capaces de resolver ecuaciones diferenciales de primer orden, se podrá resolver cualquier problema de condiciones iniciales. Más adelante, en la sección 7.5, se verá cómo.

El campo de direcciones

La ecuación

$$\frac{dy}{dx} = f(x, y)$$

puede interpretarse como una relación que a cada par ordenado (x, y) le hace corresponder un valor de la derivada. Como un par ordenado (x, y) representa un punto en el plano coordenado y un valor de una derivada se puede interpretar como una pendiente en un punto, resulta que la ecuación le hace corresponder a cada punto P del plano xy , donde la función f esté definida, una pendiente. Es muy interesante expresar esta relación gráficamente.

Sea R una región del plano xy donde está definida la función $f(x, y)$. Considérese un conjunto discreto de puntos de $\{P_1, P_2, \dots, P_m\}$ (marcados en la figura 1 como puntos dentro de la región R)

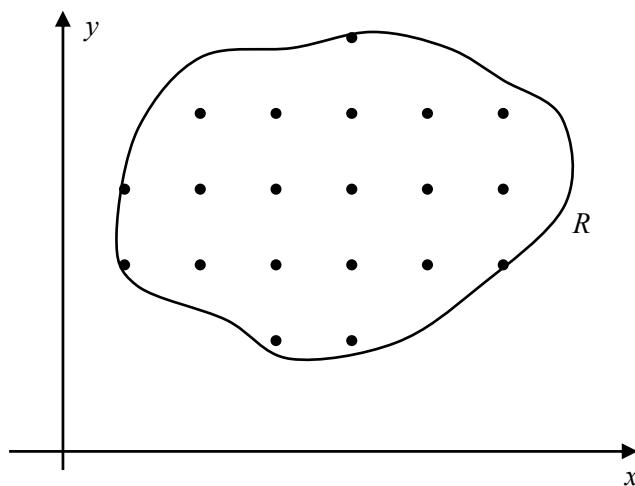


Figura 1

Sea f_i el valor de la función f para el punto P_i , es decir:

$$f_i = f(P_i) \quad i = 1, 2, \dots, m$$

Trazando un pequeño segmento de recta con centro en P_i y cuya pendiente sea f_i para cada uno de los puntos marcados en la región R , se obtiene una gráfica como la que muestra la figura 2. Al conjunto de segmentos obtenidos se le llama *campo de direcciones* de la ecuación diferencial.

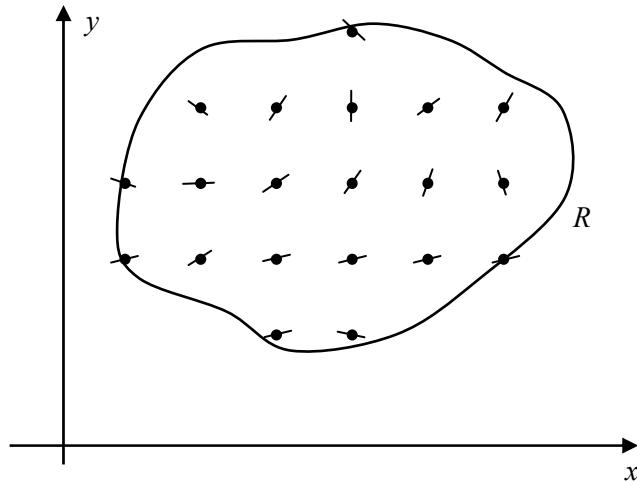


Figura 2

La mayoría de los programas de computadora que trabajan con ecuaciones diferenciales permiten trazar sus campos de direcciones para regiones rectangulares del plano xy tomando conjuntos de puntos distribuidos en forma regular.

Ejemplo 3

Trace el campo de direcciones de la ecuación diferencial $\frac{dy}{dx} = x^2 - y^2$ en la región rectangular definida por $-2 \leq x \leq 2$; $-2 \leq y \leq 2$

Solución:

En la figura 3 se muestra el campo de direcciones que ofrece un asistente matemático en la región definida por el rectángulo $-2 \leq x \leq 2$; $-2 \leq y \leq 2$, tomando 20 sub intervalos en ambas direcciones.

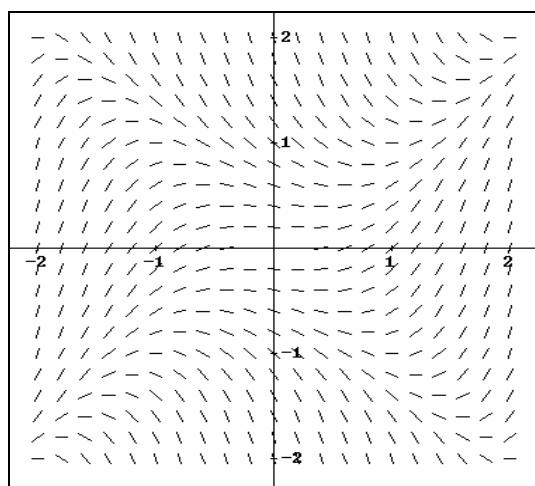


Figura 3

El campo de direcciones brinda una idea cualitativa muy abarcadora acerca de cómo se comportarán las soluciones de una ecuación diferencial y su importancia teórica y práctica no puede ser mayor. Si $y = y(x)$ es una solución particular de la ecuación diferencial

$$\frac{dy}{dx} = f(x, y)$$

entonces la gráfica de la función tendrá en cada punto (x, y) del plano una pendiente dada por $f(x, y)$; esto significa que, cuando la curva pase por un punto del campo de direcciones, lo hará en la dirección del segmento que corresponde a ese punto.

En la figura 4 se muestran algunas soluciones particulares de la ecuación (ejemplo 3):

$$\frac{dy}{dx} = x^2 - y^2$$

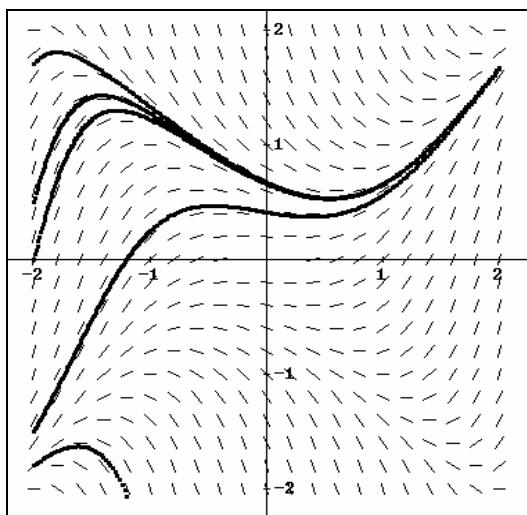


Figura 4

superpuestas a su campo de direcciones. El campo de direcciones se construye con gran facilidad y una simple observación del mismo da una idea acerca del comportamiento de las soluciones de la ecuación diferencial, tales como los intervalos en que crecen y decrecen, la ubicación aproximada de sus puntos de extremos, la existencia de asíntotas e incluso, la estabilidad de la solución.

Ejemplo 4

La ecuación

$$\frac{dy}{dx} = 2x^2 + y^2$$

no puede ser resuelta analíticamente. Trace su campo de direcciones en la región $R = \{(x, y): -1 \leq x \leq 1, -2 \leq y \leq 2\}$ y analice el comportamiento que poseen sus soluciones dentro de esa región.

Solución:

El campo de direcciones se muestra en la figura 5 y del mismo puede concluirse que:

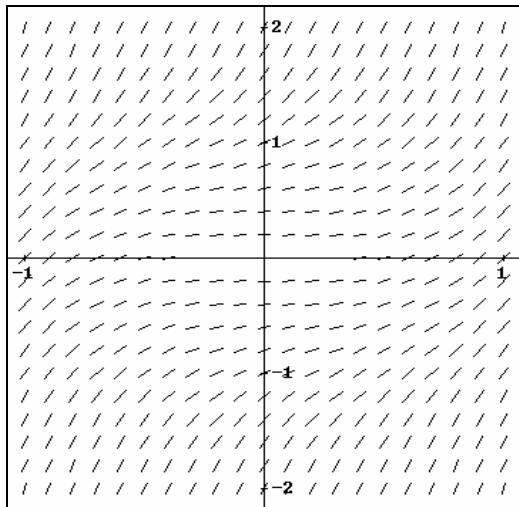


Figura 5

- La pendiente del campo de direcciones es positiva en todos los puntos, lo cual significa que todas las soluciones son funciones crecientes.
- En el origen el campo de direcciones es horizontal; en los demás puntos la pendiente es mayor a medida que aumenta la distancia al origen.
- Las soluciones que pasan cerca del origen tienen un punto de inflexión cuya tangente es casi horizontal; las soluciones que pasan más lejos del origen deben poseer una forma de S invertida con su punto de inflexión muy inclinado.
- Es obvio que ninguna solución de esta ecuación diferencial puede poseer puntos de extremo ni mas de un punto de inflexión.
- Es también bastante clara la ausencia de asíntotas de cualquier tipo.

En la figura 6 se han trazado sobre el campo de direcciones algunas soluciones verdaderas calculadas por métodos numéricos que se estudiarán más adelante, de modo que pueda verificar lo acertado del análisis cualitativo realizado.

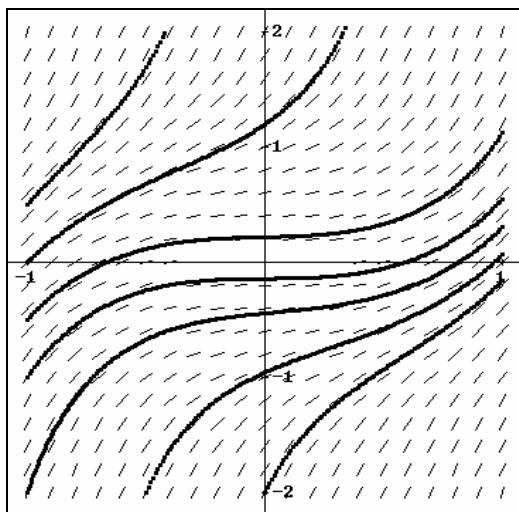


Figura 6

Isoclinas

En las ecuaciones diferenciales de primer orden tienen especial interés unas ciertas curvas llamadas *isoclinas* (igual inclinación) y que tienen la propiedad de que, en cada punto de una de estas curvas, el campo de direcciones posee la misma pendiente. Si la ecuación diferencial es

$$\frac{dy}{dx} = f(x, y)$$

las isoclinas tienen ecuaciones del tipo $f(x, y) = k$.

La determinación de algunas isoclinas importantes puede ser una fuente de información valiosa en el análisis cualitativo de una ecuación diferencial.

Ejemplo 5

Analice cualitativamente la ecuación diferencial

$$\frac{dy}{dx} = x + y$$

teniendo en cuenta su campo de direcciones en la región $R = \{(x, y): -4 \leq x \leq 4, -4 \leq y \leq 4\}$ y algunas isoclinas importantes.

Solución:

El campo de direcciones en la región deseada, se muestra en la figura 7

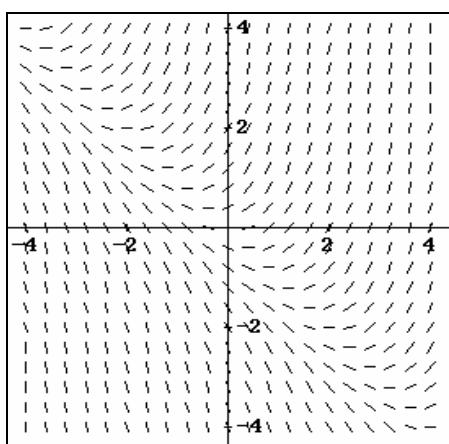


Figura 7

Las isoclinas son de la forma $x + y = k$, es decir, rectas con pendiente -1 . De particular interés es la isocrina que corresponde a $k = 0$:

$$x + y = 0$$

que contiene a los puntos estacionarios de las soluciones (puntos de tangente horizontal). La recta divide al plano en dos regiones:

$$R_1 = \{(x, y): x + y > 0\} \quad y \quad R_2 = \{(x, y): x + y < 0\}$$

En R_1 las soluciones son crecientes y en R_2 todas son decrecientes. De ahí resulta que todas las soluciones que cruzan la recta $x + y = 0$ poseen un punto de mínimo sobre ella (vea la figura 8)

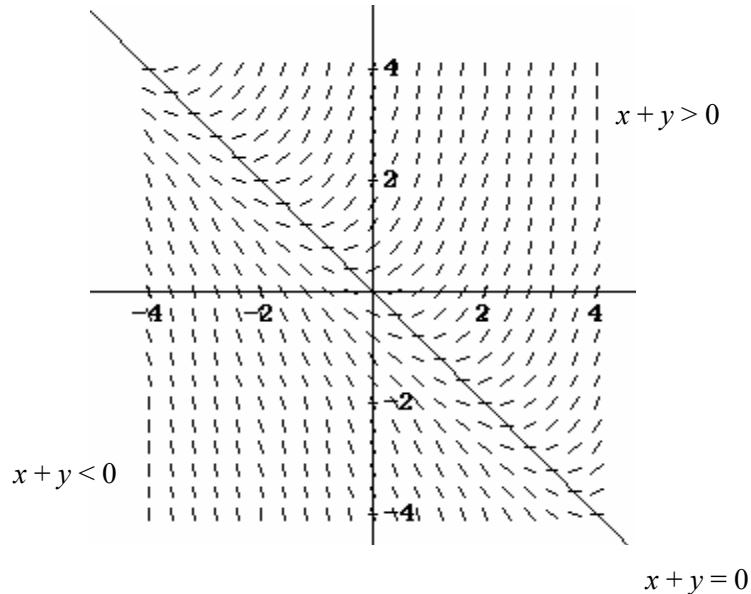


Figura 8

La isoclina que corresponde a $k = -1$ es sumamente interesante. Su ecuación es $x + y = -1$. Su pendiente es -1 , como en las demás isoclinas, pero en esta recta (figura 9) su pendiente en

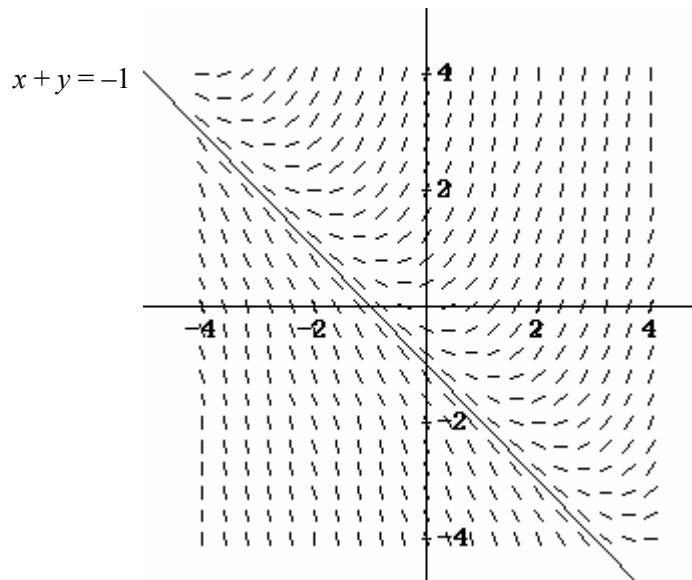


Figura 9

cada punto coincide con la del campo de direcciones. Nótese que esta recta es una solución de la ecuación diferencial porque ella es tangente al campo de direcciones en cada uno de sus puntos; es

la única isoclina que también es solución de la ecuación diferencial. Si usted analiza con cuidado la figura 9, verá que ninguna solución puede atravesar esta recta, la cual es una asíntota oblicua para todas las demás soluciones de la ecuación. Las soluciones son como se muestran en la figura 10. Nótese que si se dan condiciones iniciales (x_0, y_0) próximas a la recta $x + y = -1$ variaciones mínimas de x_0 ó y_0 causarán un importante cambio en el comportamiento de la solución para otros valores de x . Este fenómeno se llama inestabilidad y será objeto de estudio a continuación.

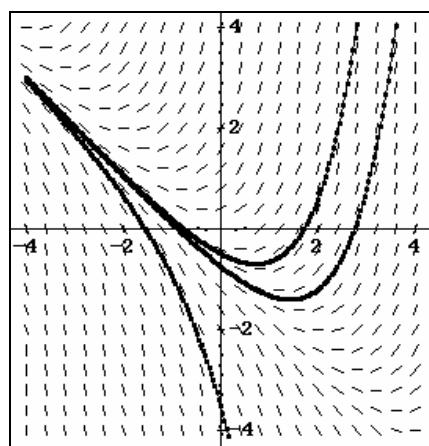


Figura 10

La estabilidad de las ecuaciones diferenciales

Recuérdese que en el primer capítulo fue caracterizado un problema inestable como aquel en el cual, pequeñas variaciones en los datos provocan grandes cambios en la respuesta. Es muy importante conocer si el problema que se va a resolver es estable o inestable pues de ello depende qué tan grandes son los errores que pueden permitirse en los datos iniciales o en el transcurso de la solución. En el campo de las ecuaciones diferenciales, se considerará que el problema de Cauchy:

$$\frac{dy}{dx} = f(x, y) \quad y(x_0) = y_0$$

es inestable si pequeños cambios en y_0 producen grandes cambios en la solución de la ecuación para valores de x alejados de x_0 . Este concepto puede precisarse más, pero para el objetivo que aquí se persigue, es suficiente con esta idea general.

Una ecuación diferencial será inestable en una cierta región del plano xy si sus soluciones presentan un comportamiento divergente a medida que x crece, como en la figura 11 a, mientras que será estable cuando sus soluciones tienden a agruparse a medida que crece x como en la figura 11 b.

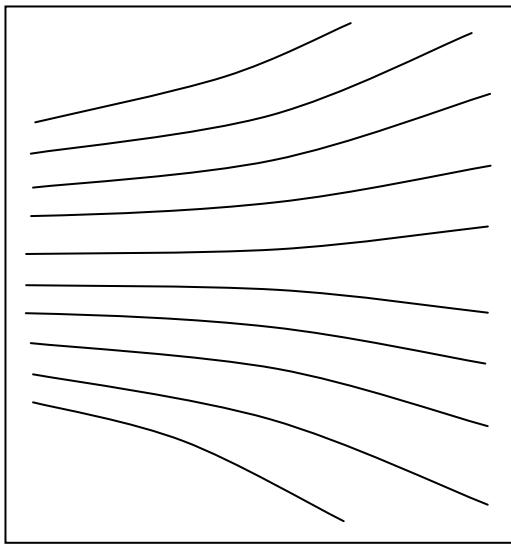


Figura 11 a

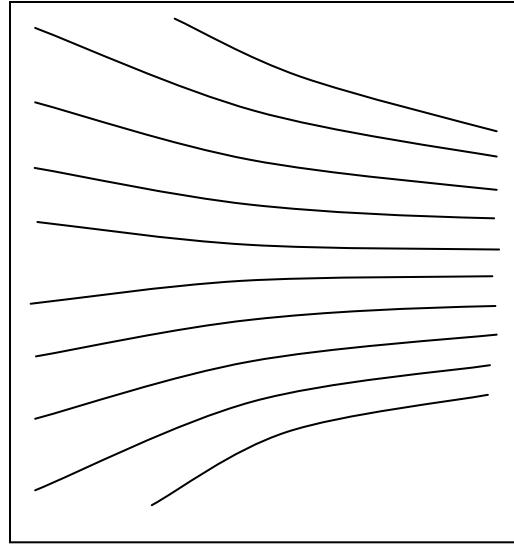


Figura 11 b

Ejemplo 6

Analice la estabilidad de la ecuación diferencial $\frac{dy}{dx} = x + y$

Solución:

Observe la figura 10. Las curvas solución poseen un patrón divergente en cualquier región del plano xy . Se trata de un problema inestable para cualquier condición inicial que se imponga. Sin embargo, es obvio que si (x_0, y_0) está próximo a la recta $x + y = -1$, el problema será más inestable.

Ejemplo 7

Analice la estabilidad de la ecuación diferencial $\frac{dy}{dx} = x^2 - y^2$

Solución:

En la figura 4 se mostró el campo de direcciones en la región $R = \{(x, y): -2 \leq x \leq 2, -2 \leq y \leq 2\}$ y las gráficas de varias soluciones. Resulta obvio que, de acuerdo con las condiciones iniciales que se tome, la ecuación diferencial tendrá un comportamiento estable o inestable. Mirando con cuidado el campo de direcciones se podrá apreciar que en la zona donde $y > 0$ las curvas tienden a unirse, mientras que donde $y < 0$ estas tienden a separarse; sin embargo, puede darse el caso que soluciones que inicialmente divergían, al entrar en la zona $y > 0$ converjan rápidamente. ■

La convergencia o divergencia de las soluciones de la ecuación diferencial

$$\frac{dy}{dx} = f(x, y)$$

y, por tanto, su estabilidad o inestabilidad, dependen de la forma en que varía $f(x, y)$ en la región considerada del plano xy . Si, para una x fija, $f(x, y)$ aumenta con y , ello significa que en la dirección

vertical el campo de direcciones aumenta su pendiente al crecer y , lo cual indica que las soluciones divergen en esa región (observe la figura 12 a); en este caso será positiva la derivada parcial de f respecto a y . Por el contrario, si esa derivada parcial es negativa, la pendiente del campo de direcciones decrece cuando y aumenta, y eso corresponde con una ecuación diferencial estable, con curvas convergentes, como se aprecia en la figura 12 b.

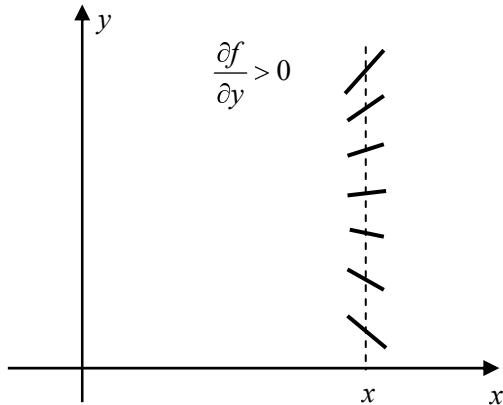


Figura 12 a

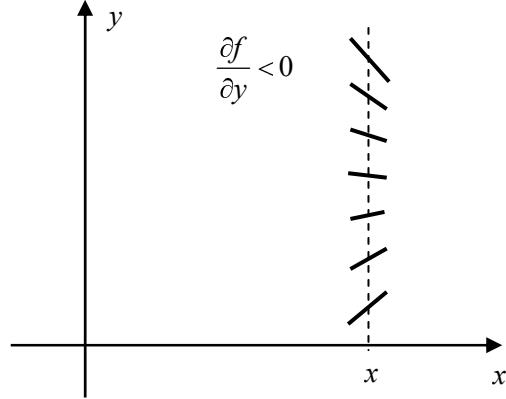


Figura 12 b

El análisis es bastante sencillo cuando la derivada parcial $\frac{\partial f}{\partial y}$ mantiene un solo signo en toda la región del plano xy donde puede desarrollarse la solución, pero el asunto se complica cuando dicha derivada posee diferentes signos en la región de interés.

Un ejemplo sumamente importante para el futuro, es el que se realiza a continuación.

Ejemplo 8

Analice la estabilidad de la ecuación diferencial:

$$\frac{dy}{dx} = Ay$$

de acuerdo con el signo de la constante A .

Solución:

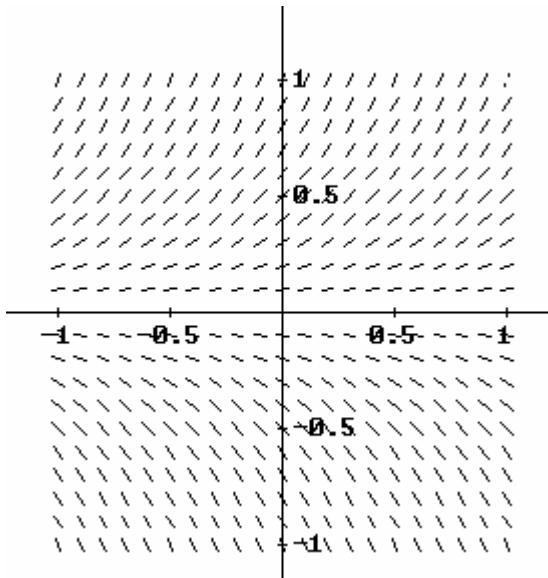
En este ejemplo $f(x, y) = Ay$, así que $\frac{\partial f}{\partial y} = A$ en todo el plano xy . Puede entonces afirmarse que:

Si $A < 0$ la ecuación diferencial es estable

Si $A > 0$ la ecuación diferencial es inestable

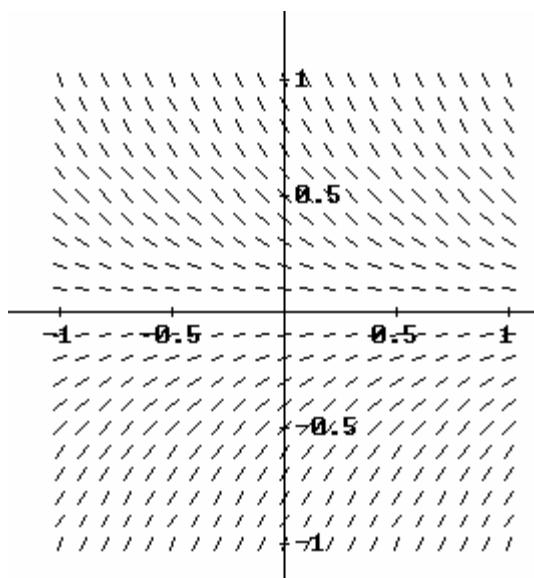
En la figura 13 se muestra el campo de direcciones de dos ecuaciones diferenciales de este tipo, una con $A = 2$ y otra con $A = -2$, en ambos casos en la región:

$$R = \{(x, y) : -1 \leq x \leq 1, -1 \leq y \leq 1\}$$



$$\frac{dy}{dx} = 2y$$

Figura 13 a



$$\frac{dy}{dx} = -2y$$

Figura 13 b

Ecuación estable modelo

La ecuación diferencial $\frac{dy}{dx} = -Ay$ con $A > 0$ se llamará *ecuación estable modelo*.

La ecuación diferencial modelo es sumamente sencilla, posee solución general en forma analítica:

$$y = Ce^{-Ax}$$

y servirá en lo adelante para analizar el comportamiento de los diferentes métodos numéricos que serán desarrollados frente a una ecuación diferencial estable.

Ejercicios

Exprese las siguientes ecuaciones diferenciales en la forma $\frac{dy}{dx} = f(x, y)$ o diga si no es posible.

$$1. \quad \frac{dy}{x+y} + \frac{dx}{x-y} = 0$$

$$2. \quad x \frac{dy}{dx} + \frac{3}{x} y = \cos x$$

$$3. \quad \sqrt{\frac{dy}{dx} + y} = 3x + 2$$

$$4. \quad \frac{dy}{dx} + \frac{dx}{dy} = x$$

En las siguientes ecuaciones diferenciales construya el campo de direcciones, utilizando un programa adecuado, en la región que se indica, haga un bosquejo de cómo serán las soluciones dentro de esa región y dibuje algunas isoclinas.

$$5. \quad \frac{dy}{dx} = -xy \quad R = \{(x, y): -1 \leq x \leq 1, -1 \leq y \leq 1\}$$

$$6. \quad \frac{dy}{dx} = -\frac{x}{y} \quad R = \{(x, y): -1 \leq x \leq 1, -1 \leq y \leq 1\}$$

$$7. \quad \frac{dy}{dx} = \frac{y-x}{x+y} \quad R = \{(x, y): -1 \leq x \leq 1, -1 \leq y \leq 1\}$$

$$8. \quad \frac{dy}{dx} = -2x \quad R = \{(x, y): -1 \leq x \leq 1, -1 \leq y \leq 1\}$$

$$9. \quad \frac{dy}{dx} = x - y \quad R = \{(x, y): -2 \leq x \leq 2, -2 \leq y \leq 2\}$$

$$10. \quad \frac{dy}{dx} = y^2 - 1 \quad R = \{(x, y): -2 \leq x \leq 2, -2 \leq y \leq 2\}$$

7.3 Métodos de paso simple

Dos tipos de métodos

Los métodos numéricos para resolver el problema de Cauchy:

$$\begin{aligned}\frac{dy}{dx} &= f(x, y) \\ y(x_0) &= y_0\end{aligned}\tag{1}$$

se basan en la idea de tomar un conjunto discreto de valores de x :

$$\{x_0, x_1, x_2, \dots\}$$

casi siempre uniformemente espaciados, y hallar valores:

$$\{y_0, y_1, y_2, \dots\}$$

que se aproximen a los valores de la solución exacta de (1):

$$\{y(x_0), y(x_1), y(x_2), \dots\}$$

Los algoritmos para hallar la solución aproximada $\{y_0, y_1, y_2, \dots\}$ son siempre iterativos y pueden simbolizarse mediante una ecuación del tipo:

$$y_{n+1} = G(y_n, y_{n-1}, y_{n-2}, \dots, y_{n-k}) \quad n \geq k$$

En el caso más sencillo, cuando $k = 0$:

$$y_{n+1} = G(y_n) \quad n \geq 0$$

los métodos se llaman *de paso simple*, y son los que se estudia en esta sección. En la sección 7.4 se analizarán métodos *de paso múltiple*, es decir, de los tipos:

paso doble: $y_{n+1} = G(y_n, y_{n-1},) \quad n \geq 1$

paso triple: $y_{n+1} = G(y_n, y_{n-1}, y_{n-2}) \quad n \geq 2$

etcétera.

El método de Euler

El método de Euler es el más elemental de los métodos de paso simple. Como se conoce el primer elemento (x_0, y_0) de la solución, se puede determinar el campo de direcciones en ese punto: $f(x_0, y_0)$. Para hallar el segundo punto de la solución aproximada (x_1, y_1) se sigue una trayectoria rectilínea en esa dirección hasta alcanzar la abscisa x_1 (ver figura 1).

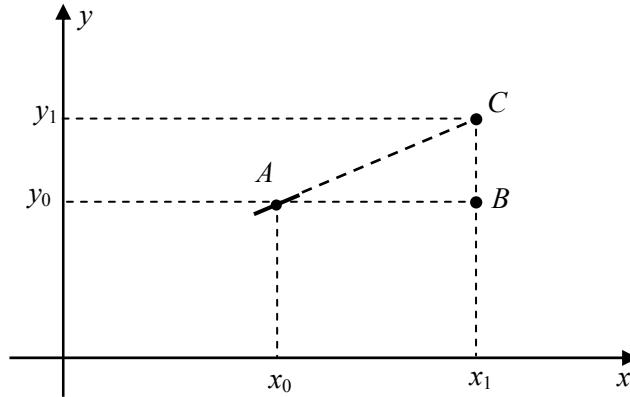


Figura 1

Como la pendiente del segmento AC es $f(x_0, y_0)$, entonces:

$$y_1 = y_0 + (x_1 - x_0)f(x_0, y_0)$$

Llamando $h = x_1 - x_0$, esta ecuación queda:

$$y_1 = y_0 + hf(x_0, y_0)$$

Una vez conocido (x_1, y_1) , se aplica la misma idea para determinar (x_2, y_2) . Suponiendo en la figura 2 que $x_2 - x_1 = h$ resulta:

$$y_2 = y_1 + hf(x_1, y_1)$$

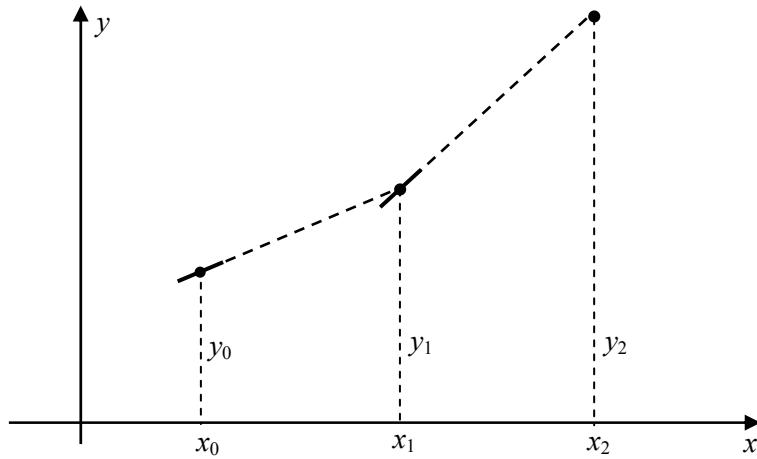


Figura 2

En general, si se toma un paso h como incremento de la variable independiente, resulta:

Para $n = 0, 1, 2, \dots$

$$\begin{aligned} x_{n+1} &= x_n + h \\ y_{n+1} &= y_n + hf(x_n, y_n) \end{aligned} \tag{2}$$

Obsérvese que nada impide que h tome valores diferentes en cada aplicación de la fórmula (2) y que, aunque no es lo usual, su valor pueda ser negativo.

Algoritmo en seudo código

El algoritmo que sigue obtiene la solución aproximada del problema de Cauchy

$$\frac{dy}{dx} = f(x, y)$$

$$y(x_0) = y_0$$

en un intervalo $x_0 \leq x \leq x_F$ con paso $h > 0$. Se suponen conocidos: la función $f(x, y)$, las condiciones iniciales (x_0, y_0) el valor final x_F y el paso h .

```

n := 0
do while  $x_n < x_F$ 
    ( $x_n, y_n$ ) es un punto de la solución aproximada
     $x_{n+1} := x_n + h$ 
     $y_{n+1} := y_n + hf(x_n, y_n)$ 
    n := n + 1
end
Terminar

```

Ejemplo 1

Resuelva el problema de Cauchy

$$\frac{dy}{dx} = \frac{1}{2}y \quad y(0) = 0,5$$

en el intervalo $0 \leq x \leq 3$ mediante el método de Euler con pasos $h = 0,4$; $h = 0,2$; $h = 0,1$ y $h = 0,05$. Analice cómo se comporta el error con el incremento de x y con la disminución de h , comparando las soluciones aproximadas obtenidas con la solución exacta:

$$y = \frac{1}{2} e^{\frac{x}{2}}$$

Solución:

Se calculará con detalle algunos valores de la solución aproximada para $h = 0,4$:

$$x_0 = 0 \quad y_0 = 0,5$$

$$x_1 = 0,4 \quad y_1 = y_0 + hf(x_0, y_0) = y_0 + h \frac{y_0}{2} = 0,5 + 0,4 \frac{0,5}{2} = 0,6$$

$$x_2 = 0,8 \quad y_2 = y_1 + hf(x_1, y_1) = y_1 + h \frac{y_1}{2} = 0,6 + 0,4 \frac{0,6}{2} = 0,72$$

$$x_3 = 1,2 \quad y_3 = y_2 + h f(x_2, y_2) = y_2 + h \frac{y_2}{2} = 0,72 + 0,4 \frac{0,72}{2} = 0,864$$

Etcétera.

En la tabla 1 se aprecia el resultado total, la solución exacta y el error para cada x_n

n	x_n	y_n	$y(x_n)$	$error(y_n) = y(x_n) - y_n$
0	0,0	0,5000	0,5000	0,0000
1	0,4	0,6000	0,6107	0,0107
2	0,8	0,7200	0,7459	0,0259
3	1,2	0,8640	0,9111	0,0471
4	1,6	1,0368	1,1128	0,0760
5	2,0	1,2442	1,3591	0,1149
6	2,4	1,4930	1,6601	0,1671
7	2,8	1,7916	2,0276	0,2360
8	3,2	2,1499	2,4765	0,3266
9	3,6	2,5799	3,0248	0,4449
10	4,0	3,0959	3,6945	0,5986

Tabla 1

Nótese como el error de la solución aproximada se va incrementando a medida que la solución progresá. Como se verá más adelante, este comportamiento no es así en todos los casos y está relacionado con el hecho de que esta ecuación diferencial es inestable, ya que $\frac{\partial f}{\partial y} = \frac{1}{2} > 0$. Una situación similar ocurre para valores menores del paso y se obtiene el mayor error al final del intervalo $[x_0, x_F]$. En la tabla 2 se muestra la solución obtenida para varios valores de x con pasos 0,4; 0,2; 0,1 y 0,05, así como los errores correspondientes.

Nótese cómo el error para una x fija disminuye a medida que se utilizan pasos menores; es más, se observa que, al disminuir h a la mitad, el error se hace también aproximadamente la mitad. Este fenómeno es general y tendrá su explicación más adelante.

x	Solución aproximada con $h = 0,4 \quad h = 0,2 \quad h = 0,1 \quad h = 0,05$				Error en la solución aproximada con $h = 0,4 \quad h = 0,2 \quad h = 0,1 \quad h = 0,05$			
	0,0	0,5000	0,5000	0,5000	0,5000	0,0000	0,0000	0,0000
0,8	0,7220	0,7320	0,7387	0,7423	0,0259	0,0139	0,0072	0,0036
1,6	1,0368	1,0718	1,0914	1,1019	0,0760	0,0410	0,0214	0,0109
2,4	1,4930	1,5692	1,6126	1,6357	0,0167	0,0909	0,0475	0,0244
3,2	2,1499	2,2975	2,3825	2,4283	0,3266	0,1790	0,0940	0,04820
4,0	3,0659	3,3638	3,5200	3,6048	0,5986	0,3407	0,1745	0,0897

Tabla 2

Error en el método de Euler

Considérese el problema de Cauchy

$$\frac{dy}{dx} = f(x, y) \quad y(x_0) = y_0$$

Sea $y(x)$: la solución exacta del problema

$\{y_0, y_1, y_2, \dots, y_n\}$: la solución obtenida por el método de Euler con paso h

$e_n = \text{error}(y_n) = y(x_n) - y_n$: el error que afecta a y_n

Se supondrá que $y(x)$ posee segunda derivada continua, de modo que puede usarse el polinomio de Taylor de primer grado con resto de Lagrange, alrededor de x_n :

$$y(x) = y(x_n) + y'(x_n)(x - x_n) + \frac{1}{2} y''(\xi)(x - x_n)^2$$

donde ξ es algún valor entre x_n y x . Evaluando para $x = x_{n+1}$ y considerando que $x_{n+1} - x_n = h$ resulta:

$$y(x_{n+1}) = y(x_n) + y'(x_n)h + \frac{1}{2} y''(c)h^2 \quad (3)$$

donde $x_n < c < x_{n+1}$

Por otra parte, según la fórmula (2) de Euler:

$$y_{n+1} = y_n + hf(x_n, y_n)$$

y, restando de (3):

$$y(x_{n+1}) - y_{n+1} = [y(x_n) - y_n] + h[y'(x_n) - f(x_n, y_n)] + \frac{1}{2} y''(c)h^2 \quad (4)$$

Ahora bien, $e_{n+1} = y(x_{n+1}) - y_{n+1}$. Además, $y(x_n)$ es la solución exacta de la ecuación diferencial, luego:

$$y'(x_n) = f[x_n, y(x_n)]$$

Sustituyendo en la ecuación (4):

$$e_{n+1} = e_n + h[f(x_n, y(x_n)) - f(x_n, y_n)] + \frac{1}{2} y''(c)h^2$$

Utilizando el teorema del valor medio del Cálculo Diferencial, la diferencia entre corchetes puede escribirse:

$$f(x_n, y(x_n)) - f(x_n, y_n) = f_y(x_n, \bar{y})[y(x_n) - y_n]$$

donde f_y : representa la derivada parcial de f respecto a y , que se supone continua
 \bar{y} : es un número entre $y(x_n)$ y y_n

Ahora queda:

$$e_{n+1} = e_n + hf_y(x_n, \bar{y})[y(x_n) - y_n] + \frac{1}{2}y''(c)h^2$$

es decir: $e_{n+1} = e_n [1 + hf_y(x_n, \bar{y})] + \frac{1}{2}y''(c)h^2 \quad (5)$

Esta fórmula es muy instructiva pues explica la forma en que el error de y_n da lugar al error de y_{n+1} . Primeramente observe que, aun en el caso en que y_n fuera exacto ($e_n = 0$), el error e_{n+1} tomaría el valor

$$\frac{1}{2}y''(c)h^2$$

Así que este término es el error nuevo que se incorpora en el paso $n+1$; por esta razón suele llamarse *error local*:

$$\text{Error local: } \frac{1}{2}y''(c)h^2$$

El otro sumando de la fórmula (5) se conoce como error propagado y representa la parte del error e_{n+1} que se debe al error que ya existía en el paso anterior:

$$\text{Error propagado: } [1 + hf_y(x_n, \bar{y})]e_n$$

Como usualmente h es pequeño, puede suponerse que

$$|hf_y(x_n, \bar{y})| < 1$$

Note como el signo de $f_y(x, y)$ es fundamental en el comportamiento del error propagado, ya que:

$$f_y(x_n, \bar{y}) > 0 \Rightarrow 1 + hf_y(x_n, \bar{y}) > 1$$

$$f_y(x_n, \bar{y}) < 0 \Rightarrow 0 < 1 + hf_y(x_n, \bar{y}) < 1$$

En el primer caso, el error propagado es mayor que e_n (en valor absoluto) y es de esperar que el error crezca rápidamente a medida que la solución avanza. En el segundo caso, el error propagado es menor (en valor absoluto) que e_n y, por lo general, sucederá que el error total se mantiene reducido e incluso puede tender hacia cero cuando la solución avanza.

La ecuación (5) es una ecuación en diferencias finitas (también, ecuación recursiva) de primer orden, que puede ser resuelta, de modo que su variable e_n quede como una función explícita de n . Los detalles de la obtención de la solución se omiten. Resulta:

$$|e_n| \leq \frac{hY}{2F} [e^{(x_n - x_0)F} - 1] \quad (6)$$

donde F : Cota superior de $|f_y(x, y)|$ en toda la región del plano xy donde se desarrolla el problema

Y : Cota superior de $|y''(x)|$ para $x_0 \leq x \leq x_n$

La fórmula (6) ofrece una cota del error total del método de Euler, pero su uso práctico es muy limitado por la presencia de los parámetros F e Y que usualmente no se conocen; lo más importante que aporta es el hecho de que

$$|e_n| \leq kh \quad (7)$$

donde $k = \frac{Y}{2F} [e^{(x_n - x_0)F} - 1]$

De la fórmula (7) se llega a dos conclusiones de gran importancia teórica y práctica:

1. Para una x_n fija, si se toman valores de h cada vez menores, se cumple que $\lim_{h \rightarrow 0} e_n = 0$
2. El error total del método de Euler depende de h de tal modo que esta dependencia no sobrepasa una cierta función lineal. Simbólicamente, $e_n = O(h)$, lo cual se expresa con la oración: El método de Euler es de orden uno.

Nótese ahora por qué en el ejemplo anterior el error aumenta rápidamente a medida que x se aleja de x_0 ; es que $f_y(x, y) = \frac{1}{2} > 0$ para todo el plano xy . Por otra parte, al ser el método de Euler de primer orden, el error total es proporcional (aproximadamente) a h , y esto explica por qué, para cada x fijo, cuando h se reduce a la mitad, el error también lo hace (aproximadamente).

Estimación del error por doble cálculo

Cuando se conoce el orden del error de un método aproximado, se puede hacer un estimado bastante bueno del error cometido hallando la solución dos veces. Ya esta técnica fue empleada en el capítulo 5 para estimar el error en la integración numérica. Salvo algunas notaciones que aquí son distintas, el análisis que sigue es completamente similar al que se realizó allí.

Sea $\frac{dy}{dx} = f(x, y) \quad y(x_0) = y_0$

la ecuación diferencial cuya solución se busca. Sea:

y_h :	solución aproximada obtenida para un cierto x usando paso h
y_{2h} :	solución aproximada obtenida para el mismo x usando paso $2h$
e_h :	error total de y_h
e_{2h} :	error total de y_{2h}

Si la solución aproximada se ha obtenido con un método numérico de orden p (por el momento, como el único método estudiado es el de Euler, puede suponerse que $p = 1$; más adelante p tomará otros valores), resulta que:

$$e_n = O(h^p)$$

Considérese aproximadamente: $e_n = kh^p$

(lo cual sólo es exacto en sentido asintótico, es decir, para h tendiendo hacia cero; en otras palabras, e_h y kh^p son infinitésimos equivalentes)

Entonces: $e_{2h} = k(2h)^p = 2^p kh^p = 2^p e_h \quad (8)$

Por otra parte: $y_h + e_h = y_{2h} + e_{2h}$

pues ambos miembros de esta igualdad son el valor de la solución $y(x)$. Teniendo en cuenta (8):

$$y_h + e_h = y_{2h} + 2^p e_h$$

y, despejando e_h

$$e_h = \frac{y_h - y_{2h}}{2^p - 1} \quad (9)$$

Como en el método de Euler el valor de p es 1, resulta:

Para el método de Euler

$$e_h \approx y_h - y_{2h} \quad (10)$$

Ejemplo 2

Resuelva el problema

$$\frac{dy}{dx} = x - y \quad y(1) = 3$$

en el intervalo $1 \leq x \leq 5$ mediante el método de Euler con dos cifras decimales exactas.

Solución:

Nótese que por ser $f_y(x, y) = -1$ se trata de una ecuación diferencial estable y es de esperar que el error no se propague demasiado cuando x se aleje de x_0 . Comenzando con $h = 0,1$, la tabla 3 muestra la solución que se obtiene con diferentes pasos para los valores de x : 1, 2, 3, 4 y 5, así como un estimado del error basado en la fórmula (10). Naturalmente, para la solución obtenida con $h = 0,1$ no se cuenta con la estimación del error.

$h = 0,1$		$h = 0,05$		$h = 0,025$		$h = 0,0125$		$h = 0,00625$	
x	y	y	Error	y	Error	y	Error	y	Error
1	3,0000	3,0000	0,0000	3,0000	0,0000	3,0000	0,0000	3,0000	0,0000
2	2,0460	2,0755	0,0295	2,0897	0,0142	2,0967	0,0070	2,1002	0,0035
3	2,3647	2,3855	0,0208	2,3958	0,0103	2,4009	0,0051	2,4035	0,0026
4	3,1272	3,1382	0,0110	3,1438	0,0056	3,1466	0,0028	3,1480	0,0014
5	4,0443	4,0495	0,0052	4,0522	0,0027	4,0536	0,0014	4,0543	0,0007

Tabla 3

Nótese como el error, después de un crecimiento local al inicio, disminuye a medida que x se aleja de x_0 . Este comportamiento se explica por el valor negativo de $f_y(x, y)$. También se observa en la tabla cómo el método de Euler converge linealmente a la solución exacta a medida que h tiende hacia cero. En la figura 3 se muestra el campo de direcciones y la gráfica de la solución en el intervalo $1 \leq x \leq 5$ obtenida con $h = 0,00625$.

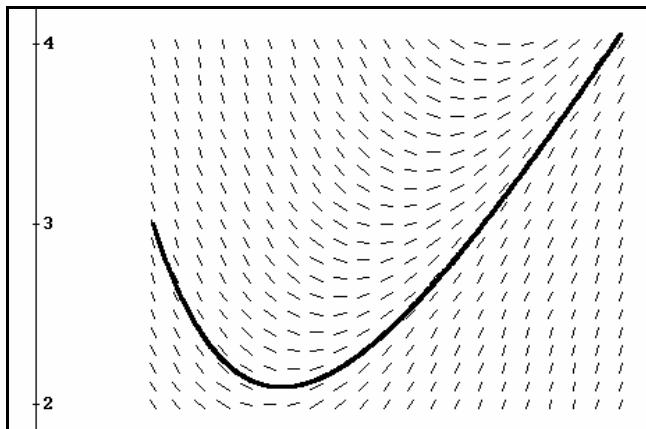


Figura 3

La solución analítica de esta ecuación diferencial puede hallarse sin mucha dificultad (es lineal de primer orden con coeficientes constantes):

$$y(x) = x - 1 + 3e^{1-x}$$

En la tabla 4 se muestran los valores de la solución exacta y de la solución hallada por el método de Euler, así como el verdadero error y el que se estimó por doble cálculo. Aprecie la similitud de ambos.

x	Solución exacta	Solución de Euler con $h = 0,00625$	Error exacto	Error estimado
1	3,000000	3,0000	0,000000	0,0000
2	2,103638	2,1002	0,003506	0,0035
3	2,406006	2,4035	0,002506	0,0026
4	3,149361	3,1480	0,001361	0,0014
5	4,054947	4,0543	0,000647	0,0007

Tabla 4

Estabilidad del método de Euler

El método de Euler, al igual que los demás métodos que se estudiarán en este capítulo, permite encontrar una solución aproximada del problema de Cauchy:

$$\text{Hallar } y(x) \text{ tal que: } \frac{dy}{dx} = f(x, y) \quad y(x_0) = y_0$$

resolviendo otro problema, en este caso:

$$\text{Hallar } \{y_0, y_1, y_2, \dots\} \text{ tal que } y_{n+1} = y_n + hf(x_n, y_n); \quad x_{n+1} = x_n + h$$

donde y_0 es conocido.

Ya en las páginas anteriores se vio que la solución del segundo problema converge hacia la solución del primero cuando h tiende hacia cero. Esto no significa, sin embargo, que sean problemas idénticos, así que tiene sentido analizar la estabilidad del segundo problema.

Es importante comenzar dejando claro un aspecto. Como la solución del método de Euler converge hacia la solución de la ecuación diferencial, cuando ésta es inestable, el método de Euler también tiene que serlo. Lo que interesa investigar es si, para una ecuación diferencial estable, pudiera suceder que el método de Euler fuera inestable; solo en ese caso se habla de inestabilidad de un método aproximado. Más formalmente:

Definición:

Se dice que un método aproximado para resolver el problema de Cauchy es inestable si presenta inestabilidad para una ecuación diferencial estable.

Para simplificar, el análisis se limitará a una ecuación diferencial estable modelo:

$$\frac{dy}{dx} = -Ay \quad A > 0$$

y se estudiará si, para esta ecuación diferencial, el método de Euler se comporta en forma estable. Para dicha ecuación, la solución de Euler será:

$$y_{n+1} = y_n + hf(x_n, y_n) = y_n + h(-Ay_n)$$

$$y_{n+1} = y_n(1 - hA)$$

Esta es una ecuación en diferencias de primer orden cuya solución general es

$$y_n = C(1 - hA)^n \quad n = 0, 1, 2, \dots$$

lo cual se verifica muy fácilmente, por sustitución.

Como y_0 es conocido: $y_0 = C(1 - hA)^0 = C$

así que la solución que brinda el método de Euler será:

$$y_n = y_0(1 - hA)^n \quad n = 0, 1, 2, \dots$$

Para analizar si este es un problema estable, considérese un pequeño cambio ε introducido en y_0 . El cambio que tiene lugar en y_n sería:

$$\begin{aligned} \Delta y_n &= (y_0 + \varepsilon)(1 - hA)^n - y_0(1 - hA)^n \\ &= \varepsilon(1 - hA)^n \end{aligned}$$

Si el producto hA es pequeño, de modo que

$$-1 < 1 - hA < 1$$

entonces, a medida que la solución avanza (n crece), el factor $(1 - hA)^n$ tiende hacia cero y por tanto, $\Delta y_n \rightarrow 0$, lo cual significa que el cambio ε en y_0 no provoca grandes variaciones en y_n .

Ahora bien, si hA es tal que $1 - hA < -1$

para lo cual basta con que $hA > 2$

entonces $|(1 - hA)^n|$ tiende hacia infinito al crecer n , es decir, que aun un pequeño cambio ε , provoca un Δy_n que crece en valor absoluto sin límite.

La conclusión es clara:

El método de Euler es estable si se toma un paso h tal que $hA < 2$. En casos así, en que la estabilidad se produce bajo ciertos requisitos, se dice que el método es condicionalmente estable.

La condición $hA < 2$ no es una exigencia importante porque, como ya se ha visto, el método de Euler requiere normalmente utilizar valores muy pequeños de h para lograr una aproximación aceptable a la solución de la ecuación diferencial.

Los métodos de Taylor

Considérese de nuevo el problema de Cauchy:

$$\frac{dy}{dx} = f(x, y) \quad y(x_0) = y_0$$

y que para $x = x_n$ se conoce una buena aproximación y_n de la solución exacta $y(x_n)$. Se desea buscar una aproximación y_{n+1} para $y(x_{n+1})$ mejor que la que ofrece la fórmula de Euler:

$$y_{n+1} = y_n + hf(x_n, y_n)$$

Esta fórmula consiste en hallar y_{n+1} approximando la solución $y(x)$ mediante una recta tangente en (x_n, y_n) (figura 4)

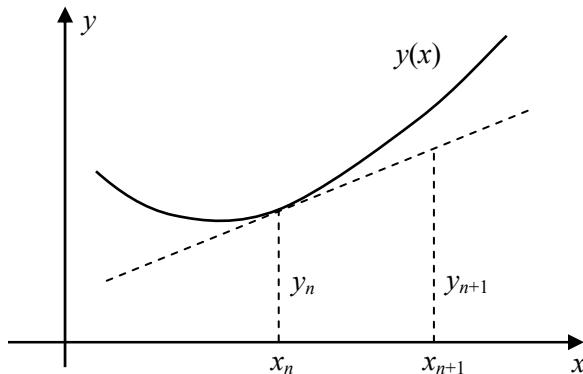


Figura 4

Pero existe la posibilidad de aproximar a $y(x)$ no mediante una línea recta sino mediante un polinomio de grado mayor. Para ello, se supone que la solución $y(x)$ posee una serie de Taylor alrededor de x_n , o sea, que:

$$y(x) = y(x_n) + y'(x_n)(x - x_n) + \frac{1}{2} y''(x_n)(x - x_n)^2 + \dots \quad (11)$$

Al tomar de esta serie solo dos términos, se obtiene:

$$y(x) \approx y(x_n) + y'(x_n)(x - x_n)$$

de donde:

$$y(x_{n+1}) \approx y(x_n) + y'(x_n)h$$

que da lugar a la fórmula de Euler:

$$y_{n+1} = y_n + hf(x_n, y_n)$$

con error local $O(h^2)$.

Si se toman más de dos términos en la serie (11), se pueden obtener fórmulas aproximadas de mayor calidad que la de Euler. A los métodos así obtenidos se les conoce como métodos de Taylor.

A modo de ejemplo, se verá el método de Taylor de orden 2, que surge de tomar tres términos de la serie (11):

$$y(x) \approx y(x_n) + y'(x_n)(x - x_n) + \frac{1}{2} y''(x_n)(x - x_n)^2$$

de donde:

$$y(x_{n+1}) \approx y(x_n) + y'(x_n)h + \frac{1}{2} y''(x_n)h^2 \quad (12)$$

Para obtener de aquí una fórmula recursiva que permita hallar y_{n+1} a partir de y_n hay que procurar una expresión para aproximar $y''(x_n)$, lo cual es un poco más complicado que la aproximación:

$$y'(x_n) = f(x_n, y(x_n)) \approx f(x_n, y_n) \quad (13)$$

Como

$$y'(x) = f(x, y(x))$$

aplicando la regla de la cadena resulta:

$$y''(x) = \frac{d}{dx}(y'(x)) = f_x(x, y(x)) + f_y(x, y(x))y'(x)$$

y de aquí se obtiene la aproximación:

$$y''(x_n) \approx f_x(x_n, y_n) + f_y(x_n, y_n)f(x_n, y_n) \quad (14)$$

Teniendo en cuenta (13) y (14), la aproximación (12) da lugar a la fórmula:

$$y_{n+1} = y_n + hf(x_n, y_n) + \frac{1}{2} h^2 [f_x(x_n, y_n) + f_y(x_n, y_n)f(x_n, y_n)] \quad (15)$$

que es la fórmula del método de Taylor de orden 2 con error local $O(h^3)$.

Similarmente se pueden obtener otras fórmulas de mayor orden pero, obviamente, la necesidad de aproximar derivadas de mayor orden, complica considerablemente las expresiones y, lo que es peor, introduce la necesidad de calcular derivadas parciales de la función $f(x, y)$.

Ejemplo 3

Halle las fórmulas necesarias para resolver la ecuación diferencial

$$\frac{dy}{dx} = x^2 + y^2 \quad y(x_0) = y_0$$

mediante el método de Taylor a) de orden dos; b) de orden tres.

Solución:

En lugar de utilizar la expresión (15), es preferible proceder directamente de la serie (11) y plantear:

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{1}{2}h^2y''(x_n) + \frac{1}{6}h^3y'''(x_n) + \dots$$

donde $y'(x) = f(x, y) = x^2 + y^2$

Derivando respecto a x y teniendo presente que y depende de x :

$$y''(x) = 2x + 2y \frac{dy}{dx} = 2x + 2y(x^2 + y^2)$$

$$y''(x) = 2x + 2x^2y + 2y^3$$

Derivando de nuevo respecto a x :

$$y'''(x) = 2 + 4xy + 2x^2 \frac{dy}{dx} + 6y^2 \frac{dy}{dx}$$

$$y'''(x) = 2 + 4xy + 2x^2(x^2 + y^2) + 6y^2(x^2 + y^2)$$

$$y'''(x) = 2 + 4xy + 2x^4 + 8x^2y^2 + 6y^4$$

Así que las fórmulas correspondientes son:

Taylor orden dos:

$$y_{n+1} = y_n + h(x^2 + y^2)_n + \frac{1}{2}h^2(2x + 2x^2y + 2y^3)_n$$

Taylor orden tres:

$$y_{n+1} = y_n + h(x^2 + y^2)_n + \frac{1}{2}h^2(2x + 2x^2y + 2y^3)_n + \frac{1}{6}h^3(2 + 4xy + 2x^4 + 8x^2y^2 + 6y^4)_n$$

donde la notación $(\)_n$ indica evaluar la expresión entre paréntesis para $x = x_n$ e $y = y_n$.

Error en el método de Taylor

Como las fórmulas de los diferentes métodos de Taylor se basan en truncar la serie de Taylor de $y(x)$, es claro que el error local de la fórmula de orden p es de orden h^{p+1} . Puede probarse que, si $f(x, y)$ tiene el número necesario de derivadas, el error total será de orden h^p . O sea:

El método de Taylor de orden p tiene: Error local $O(h^{p+1})$

Error total $O(h^p)$

Si se tiene en cuenta que el método de Euler posee error total $O(h)$, se comprenderá que los métodos de Taylor con $p > 1$ tienen una convergencia mucho más rápida hacia la solución de la ecuación diferencial. Sin embargo, la necesidad de calcular derivadas parciales es una gran desventaja, dada la complejidad de automatizar este proceso de derivación.

Para resolver este problema, a principios del siglo XX, dos matemáticos alemanes lograron obtener fórmulas mediante las cuales el problema de hallar derivadas parciales de $f(x, y)$ se transforma en el de evaluar la función $f(x, y)$ en varios puntos. De este modo obtuvieron métodos completamente equivalentes a los de Taylor pero sin el requerimiento de calcular derivadas. Estas fórmulas llevan los nombres de sus creadores: Carl Runge y Wilhelm Kutta. A continuación se hará la deducción para el caso de segundo orden, que es el más simple, y posteriormente se verán las fórmulas del método de orden cuatro que es uno de los más populares.

Método de Runge - Kutta de orden dos

El método de Taylor se basa en la aplicación reiterada de la ecuación (15):

$$y_{n+1} = y_n + hf(x_n, y_n) + \frac{1}{2}h^2[f_x(x_n, y_n) + f_y(x_n, y_n)f(x_n, y_n)]$$

la cual puede ser escrita:

$$y_{n+1} = y_n + hf(x_n, y_n) + \frac{1}{2}h[hf_x(x_n, y_n) + hf(x_n, y_n)f_y(x_n, y_n)]$$

Llamando: $K_1 = hf(x_n, y_n)$.

la ecuación queda:

$$y_{n+1} = y_n + K_1 + \frac{1}{2}h[hf_x(x_n, y_n) + K_1f_y(x_n, y_n)] \quad (16)$$

Si se supone que $f(x, y)$ es una función diferenciable, su diferencial viene dado por:

$$df(x, y) = f_x(x, y)\Delta x + f_y(x, y)\Delta y$$

de modo que el término entre corchetes de la ecuación (16) es el diferencial de f en (x_n, y_n) para

$$\Delta x = h \quad \text{y} \quad \Delta y = K_1$$

Para valores pequeños de h (y, por tanto, de K_1) este diferencial puede aproximarse mediante el incremento:

$$\Delta f(x_n, y_n) = f(x_n + h, y_n + K_1) - f(x_n, y_n)$$

con lo cual la ecuación (16) queda:

$$y_{n+1} = y_n + K_1 + \frac{1}{2}h[f(x_n + h, y_n + K_1) - f(x_n, y_n)] \quad (17)$$

La fórmula (17) es equivalente a la (16) en cuanto a exactitud, pero posee la enorme ventaja de no utilizar las derivadas de f . En su lugar, será necesario evaluar f dos veces: en (x_n, y_n) y en $(x_n + h, y_n + K_1)$. Para simplificar la notación, se suele llamar:

$$K_2 = hf(x_n + h, y_n + K_1)$$

y la ecuación (17) se reduce entonces a:

$$y_{n+1} = y_n + K_1 + \frac{1}{2}[K_2 - K_1] = y_n + \frac{1}{2}[K_1 + K_2]$$

Resumiendo lo anterior, las fórmulas de Runge - Kutta de orden dos (RK2) son:

$$\text{RK2 : } \begin{cases} K_1 = hf(x_n, y_n) \\ K_2 = hf(x_n + h, y_n + K_1) \\ y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2) \end{cases} \quad n = 0, 1, 2, 3, \dots$$

Interpretación geométrica de RK-2

Los términos $f(x_n, y_n)$ y $f(x_n + h, y_n + K_1)$ corresponden con las pendientes del campo de direcciones en los puntos (x_n, y_n) y $(x_n + h, y_n + K_1)$ (ver figura 5). Estas pendientes, multiplicadas por el incremento horizontal h , corresponden con incrementos verticales K_1 y K_2 . Entonces, geométricamente, el método RK2 consiste en determinar el valor K_1 utilizando la pendiente del campo de direcciones en (x_n, y_n) . Una vez conocido el punto $(x_n + h, y_n + K_1)$, tomar en él la pendiente del campo de direcciones y determinar el valor K_2 . El promedio de K_1 y K_2 permite calcular y_{n+1} al agregarlo a y_n . Obsérvese que promediar K_1 y K_2 es equivalente a haber promediado las pendientes del campo de direcciones en (x_n, y_n) y $(x_n + h, y_n + K_1)$ y calcular

$$y_{n+1} = y_n + h \cdot (\text{pendiente promedio})$$

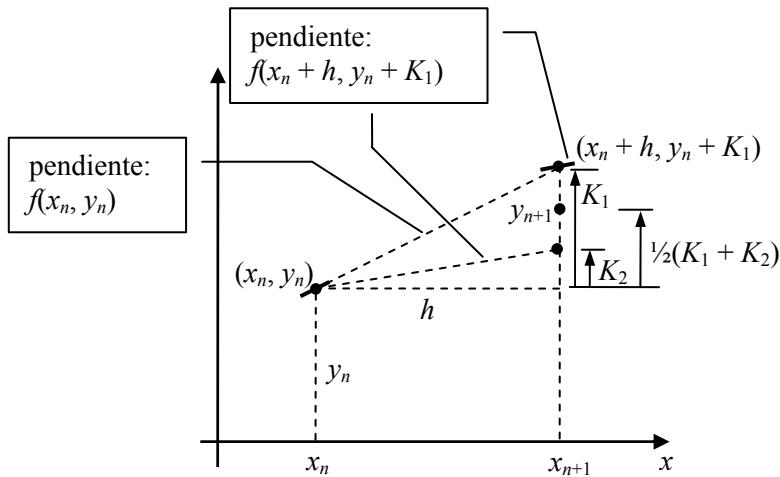


Figura 5

Algoritmo en seudo código para RK2

El algoritmo que sigue obtiene la solución aproximada del problema de Cauchy

$$\frac{dy}{dx} = f(x, y)$$

$$y(x_0) = y_0$$

en un intervalo $x_0 \leq x \leq x_F$ con paso $h > 0$. Se suponen conocidos: la función $f(x, y)$, las condiciones iniciales (x_0, y_0) el valor final x_F y el paso h .

```

n := 0
do while  $x_n < x_F$ 
    ( $x_n, y_n$ ) es un punto de la solución aproximada
     $K_1 := h f(x_n, y_n)$ 
     $K_2 := h f(x_n + h, y_n + K_1)$ 
     $y_{n+1} := y_n + \frac{1}{2}(K_1 + K_2)$ 
     $x_{n+1} := x_n + h$ 
    n := n + 1
end
Terminar

```

Ejemplo 4

Resuelva el problema de Cauchy

$$\frac{dy}{dx} = \frac{1}{2}y \quad y(0) = 0,5$$

en el intervalo $0 \leq x \leq 4$ mediante RK2 con pasos $h = 0,4; 0,2; 0,1; 0,05$. Analice cómo se comporta el error con la variación de x y con la disminución del paso h , comparando la solución obtenida con la solución exacta

$$y = \frac{1}{2} e^{\frac{x}{2}}$$

Solución:

Este mismo ejemplo fue resuelto por el método de Euler al principio de esta sección (Ejemplo 1). Resulta instructivo notar la diferencia en cuanto a los resultados en la solución por ambos métodos. Para el caso $h = 0,4$, se realizarán en detalle los primeros pasos de la solución.

$$x_0 = 0; \quad y_0 = 0,5$$

$$K_1 = h f(x_0, y_0) = 0,4 f(0; 0,5) = 0,4 \cdot \frac{1}{2}(0,5) = 0,1$$

$$K_2 = h f(x_0 + h, y_0 + K_1) = 0,4 f(0,4; 0,6) = 0,4 \cdot \frac{1}{2}(0,6) = 0,12$$

$$y_1 = y_0 + \frac{1}{2}(K_1 + K_2) = 0,5 + \frac{1}{2}(0,1 + 0,12) = 0,61$$

$$x_1 = 0,4; \quad y_1 = 0,61$$

$$K_1 = h f(x_1, y_1) = 0,4 f(0,4; 0,61) = 0,4 \cdot \frac{1}{2}(0,61) = 0,122$$

$$K_2 = h f(x_1 + h, y_1 + K_1) = 0,4 f(0,8; 0,732) = 0,4 \cdot \frac{1}{2}(0,732) = 0,1464$$

$$y_2 = y_1 + \frac{1}{2}(K_1 + K_2) = 0,61 + \frac{1}{2}(0,122 + 0,1464) = 0,7442$$

$$x_2 = 0,8; \quad y_2 = 0,7442$$

En la tabla 5 se muestran los resultados obtenidos para $x = 0,8; 1,6; 2,4; 3,2$ y 4 y los errores correspondientes, calculados por comparación con la solución exacta.

x	Solución aproximada con $h = 0,4 \quad h = 0,2 \quad h = 0,1 \quad h = 0,05$				Error en la solución aproximada con $h = 0,4 \quad h = 0,2 \quad h = 0,1 \quad h = 0,05$			
	$h = 0,4$	$h = 0,2$	$h = 0,1$	$h = 0,05$	$h = 0,4$	$h = 0,2$	$h = 0,1$	$h = 0,05$
0,0	0,5000	0,5000	0,5000	0,5000	0,0000	0,0000	0,0000	0,0000
0,8	0,7442	0,7455	0,7458	0,7459	0,0017	0,0004	0,0001	0,0000
1,6	1,1077	1,1114	1,1124	1,1127	0,0051	0,0014	0,0003	0,0001
2,4	1,6487	1,6570	1,6593	1,6599	0,0114	0,0031	0,0008	0,0002
3,2	2,4539	2,4704	2,4749	2,4761	0,0226	0,0061	0,0016	0,0004
4,0	3,6523	3,6831	3,9616	3,6938	0,0422	0,0114	0,0029	0,0007

Tabla 5

Debido a que $f_y(x, y) = 0,5 > 0$, se ve un incremento del error a medida que x se aleja de x_0 . Para cada x fija, el error disminuye con h . Nótese que, en este caso, el error disminuye unas cuatro veces cuando h se reduce a la mitad. Esto se debe a que RK2 posee un error total $O(h^2)$ al igual que el método de Taylor de orden 2.

Estimación del error en RK2

Como RK2 es una adaptación del método de Taylor de orden 2, pose errores local y total del mismo orden que aquel. Esto es:

$$\begin{aligned} \text{Error local: } & O(h^3) \\ \text{Error total: } & O(h^2) \end{aligned}$$

El error total se puede estimar por doble cálculo mediante la fórmula (9):

$$e_h = \frac{y_h - y_{2h}}{2^p - 1}$$

donde: y_h : Solución obtenida para un cierto x mediante RK2 con paso h

y_{2h} : Solución obtenida para un cierto x mediante RK2 con paso $2h$

e_h : Error total en y_h

El parámetro p , como se recordará, es el orden del error total y toma en este caso el valor 2. Resulta entonces:

$$e_h = \frac{y_h - y_{2h}}{3}$$

Ejemplo 5

Resuelva el problema

$$\frac{dy}{dx} = x - y \quad y(1) = 3$$

en el intervalo $1 \leq x \leq 5$ mediante RK2 con tres cifras decimales exactas en los resultados.

Solución:

Este ejemplo coincide con el ejemplo 2 de esta sección, en el que se utilizó el método de Euler. Es conveniente que compare los resultados de ambos. En la tabla 6 se muestran los resultados para $x = 1, 2, 3, 4$ y 5 con pasos $h = 0,1; 0,05$ y $0,025$ así como los errores estimados mediante doble cálculo (excepto para $h = 0,1$ donde no existe la información requerida)

x	$h = 0,1$		$h = 0,05$		$h = 0,025$	
	y	y	Error	y	Error	
1	3,0000	3,0000	0,0000	3,0000	0,0000	
2	2,1056	2,1041	0,0005	2,1038	0,0001	
3	2,4075	2,4064	0,0004	2,4061	0,0001	
4	3,1502	3,1496	0,0002	3,1494	0,0001	
5	4,0553	4,0550	0,0001	4,0550	0,0000	

Tabla 6

De la tabla se aprecia que el error tiende a disminuir a medida que x se aleja de x_0 . Esto se debe a que $f_y(x, y) = -1 < 0$, lo cual determina el decrecimiento del error propagado. Para cada x

fija, el error va reduciéndose a la cuarta parte en cada disminución de h a la mitad, como corresponde a un método con error total de orden dos.

Estabilidad de RK2

Al igual que en el método de Euler, la estabilidad será analizada para la ecuación modelo estable

$$\frac{dy}{dx} = -Ay \quad A > 0$$

El método RK2 aplicado a esta ecuación se reduce a

$$K_1 = hf(x_n, y_n) = h(-Ay_n) = -hAy_n$$

$$K_2 = hf(x_n + h, y_n + K_1) = h[-A(y_n + K_1)] = -hA(y_n + K_1) = -hA(y_n - hAy_n) =$$

$$= -hA(1 - hA)y_n = (-hA + h^2A^2)y_n$$

$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2) = y_n + \frac{1}{2}[-hAy_n + (-hA + h^2A^2)y_n] = y_n \left[1 + \frac{1}{2}(-hA - hA + h^2A^2)\right] =$$

$$y_{n+1} = y_n \left[1 - hA + \frac{1}{2}h^2A^2\right]$$

Esta ecuación en diferencias finitas es estable si

$$\left|1 - hA + \frac{1}{2}h^2A^2\right| < 1 \quad (18)$$

La forma cuadrática $\frac{1}{2}h^2A^2 - hA + 1$ posee discriminante negativo:

$$D = b^2 - 4ac = 1 - 4(\frac{1}{2})(1) = -1$$

así que no se anula para ningún hA real y, como vale 1 para $hA = 0$, es positiva para todo hA real. Luego, la inecuación (18) equivale a

$$1 - hA + \frac{1}{2}h^2A^2 < 1$$

esto es:

$$hA(\frac{1}{2}hA - 1) < 0$$

y, como $hA > 0$, equivale a

$$\frac{1}{2}hA - 1 < 0$$

es decir,

$$hA < 2$$

Se puede concluir entonces que el método RK2 es condicionalmente estable; se requiere tomar

$$h < \frac{2}{A}$$

En general, para valores de h que son suficientemente pequeños para lograr un error aceptable, se satisface también la condición de estabilidad.

Ejemplo 6

Resuelva la ecuación

$$\frac{dy}{dx} = -7y \quad y(0) = 5$$

en el intervalo $0 \leq x \leq 5$ con paso $h = 0,4$ y $h = 0,2$ utilizando RK2. Analice la estabilidad del método.

Solución:

La condición de estabilidad requiere en este caso que

$$h < \frac{2}{7} = 0,2857$$

lo cual no se cumple si se toma $h = 0,4$. En la tabla 7 se muestra la solución analítica:

$$y = 5e^{-7x}$$

y las obtenidas con RK2 para $h = 0,4$ y $h = 0,2$. Nótese la inestabilidad en los resultados obtenidos con paso $h = 0,4$.

x	Solución exacta	RK2 con $h = 0,4$	RK2 con $h = 0,2$
0,0	5,0000000	5,0000	5,000000
0,4	0,3040503	10,600	1,682000
0,8	0,0184893	22,472	0,565825
1,2	0,0011243	47,641	0,190343
1,6	0,0000684	101,00	0,064032
2,0	0,0000042	214,11	0,021540
2,4	0,0000003	453,93	0,007246
2,8	0,0000000	962,32	0,002436
3,2	0,0000000	2040,1	0,000820
3,6	0,0000000	4325,1	0,000276
4,0	0,0000000	9169,1	0,000093
4,4	0,0000000	19438	0,000031
4,8	0,0000000	41210	0,000011

Tabla 7

Método de Runge - Kutta de orden cuatro

A partir del método de Taylor de orden cuatro, se pueden deducir varios esquemas de Runge – Kutta orden cuatro, de modo similar (naturalmente, mucho más laborioso) al usado para obtener RK2. De estos esquemas, el más popular es el siguiente:

$$\text{RK4 : } \begin{cases} K_1 = hf(x_n, y_n) \\ K_2 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_1) \\ K_3 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_2) \\ K_4 = hf(x_n + h, y_n + K_3) \\ y_{n+1} = y_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4) \end{cases} \quad n = 0, 1, 2, 3, \dots$$

cuya interpretación geométrica se deja al lector. El método, por ser una consecuencia de Taylor orden con cuatro, posee:

$$\begin{aligned} \text{Error local: } & O(h^5) \\ \text{Error total: } & O(h^4) \end{aligned}$$

por tanto, el error total se puede estimar mediante doble cálculo como:

$$e_h = \frac{y_h - y_{2h}}{2^p - 1}$$

con $p = 4$, es decir:

$$e_h = \frac{y_h - y_{2h}}{15} \quad (19)$$

De forma similar a la utilizada para RK2, se demuestra que el método es estable para la ecuación modelo:

$$\frac{dy}{dx} = -Ay \quad A > 0$$

si se toma $hA < 2,785$, lo cual se satisface para los valores de h utilizados normalmente.

Algoritmo en seudo código para RK4

El algoritmo que sigue obtiene la solución aproximada del problema de Cauchy

$$\begin{aligned} \frac{dy}{dx} &= f(x, y) \\ y(x_0) &= y_0 \end{aligned}$$

en un intervalo $x_0 \leq x \leq x_F$ con paso $h > 0$. Se suponen conocidos: la función $f(x, y)$, las condiciones iniciales (x_0, y_0) el valor final x_F y el paso h .

```

n := 0
do while  $x_n < x_F$ 
    ( $x_n, y_n$ ) es un punto de la solución aproximada
     $K_1 := hf(x_n, y_n)$ 
     $K_2 := hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_1)$ 

```

```

 $K_3 := hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_2)$ 
 $K_4 := hf(x_n + h, y_n + K_3)$ 
 $y_{n+1} := y_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4)$ 
 $x_{n+1} := x_n + h$ 
 $n := n + 1$ 
end
Terminar

```

Ejemplo 7

Resuelva el problema

$$\frac{dy}{dx} = x - y \quad y(1) = 3$$

en el intervalo $1 \leq x \leq 5$ mediante RK4 con cuatro cifras decimales exactas.

Solución:

Se comenzara tomando $h = 0,2$. El primer paso del método se muestra a continuación en detalle.

$$x_0 = 1 \quad y_0 = 3$$

$$K_1 = hf(x_0, y_0) = 0,2f(1; 3) = 0,2(1 - 3) = -0,4$$

$$K_2 = hf(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}K_1) = 0,2f(1,1; 2,8) = 0,2(1,1 - 2,8) = -0,34$$

$$K_3 = hf(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}K_2) = 0,2f(1,1; 2,83) = 0,2(1,1 - 2,83) = -0,346$$

$$K_4 = hf(x_0 + h, y_0 + K_3) = 0,2f(1,2; 2,654) = 0,2(1,2 - 2,654) = -0,2908$$

$$y_1 = y_0 + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4) = 3 + \frac{1}{6}(-0,4 - 0,68 - 0,692 - 0,2908) = 2,6562$$

El resto de la solución se muestra en la tabla 8. También se incluyen los cálculos para $h = 0,1$ y los errores correspondientes obtenidos mediante la fórmula (19) para $x = 1, 2, 3, 4$ y 5 . En este caso no ha sido necesario tomar valores de h más pequeños pues, como se observa, con $h = 0,1$ se obtiene un error suficientemente pequeño.

x	$h = 0,2$		$h = 0,1$
	y	y	Error
1	3,000000	3,000000	0,000000
2	2,103656	2,103639	0,000001
3	2,406019	2,406007	0,000001
4	3,149368	3,149362	0,000000
5	4,054950	4,054947	0,000000

Tabla 8

Obsérvese que, para obtener resultados con solo tres cifras decimales exactas para este mismo problema (ejemplo 5) utilizando RK2 fue necesario emplear $h = 0,025$ mientras que con RK4 se han obtenido cinco cifras decimales exactas con paso $h = 0,1$. Por supuesto, debe tomarse en cuenta que cada paso de RK4 requiere evaluar la función f cuatro veces mientras que RK2 solo requiere dos evaluaciones, pero aun así, el método de cuarto orden es mucho más eficiente. En este problema, por ejemplo, utilizando RK2 con $h = 0,025$ se necesitan 160 pasos para cubrir el intervalo $1 \leq x \leq 5$, lo que representa 320 evaluaciones de la función f ; utilizando RK4 con $h = 0,1$ se emplean 40 pasos para un total de 160 evaluaciones de f , es decir, en la mitad del tiempo de cálculo se obtienen resultados con dos cifras decimales exactas adicionales.

Ejercicios

1. Considere el problema de Cauchy

$$\frac{dy}{dx} = \frac{x-y}{\ln x} \quad y(2) = 1$$

- a) Analice la estabilidad de la ecuación en el intervalo $2 \leq x \leq 6$ mediante el comportamiento de $f_y(x, y)$.
 - b) Analice su campo de direcciones (utilice un programa computacional adecuado) y verifique su estabilidad en el intervalo dado.
 - c) Halle su solución con tres cifras decimales exactas en el intervalo $2 \leq x \leq 6$ utilizando los métodos de Euler, RK2 y RK4. Compare la cantidad de veces que fue necesario evaluar la función f en cada caso.
2. Sea la ecuación diferencial
- $$(y^2 - x)dx + (y + 3x)dy = 0 \quad y(0) = 4$$
- a) Analice la estabilidad de este problema mediante la construcción (con un programa computacional) y la observación del campo de direcciones.
 - b) Halle la solución de la ecuación en el intervalo $0 \leq x \leq 5$ con cuatro cifras decimales exactas mediante RK4.
 - c) Repita el inciso b) cambiando la condición inicial por $y(0) = 4,001$, compare con el resultado obtenido en b) y explique.
3. Resuelva la siguiente ecuación diferencial con cuatro cifras decimales exactas

$$\frac{dy}{dx} = y - x$$

en el intervalo $0 \leq x \leq 5$ con las condiciones iniciales:

- a) $y(0) = 0,999$
 - b) $y(0) = 1,001$
- mediante RK4, con cuatro cifras decimales exactas.
- c) Compare los resultados obtenidos en a) y b) y explique, observando el campo de direcciones de la ecuación.
4. En las páginas anteriores se han dado algoritmos para los métodos de Euler, RK2 y RK4 que suponen $h > 0$. Introduzca en dichos algoritmos las modificaciones necesarias de manera que

admitan lo mismo valores positivos que negativos de h . Tenga en cuenta que los valores de x_0 y x_F deben ser consecuentes con el signo de h .

5. No es difícil elaborar algoritmos para resolver ecuaciones diferenciales que seleccionen automáticamente el tamaño h del paso. Sea ϵ la tolerancia relativa que se desea obtener en la solución (ϵ es el mayor error que se permitirá acumular por cada unidad que avance la solución; puede obtenerse dividiendo el máximo error absoluto permitido entre la amplitud del intervalo en que se desea obtener la solución). Construya un algoritmo basado en RK4 que utilice el método de doble cálculo para estimar el error en cada paso y que reduzca el paso a la mitad si se excede la tolerancia relativa y lo duplique si se está obteniendo un error 40 veces menor que lo que permite la tolerancia relativa.
6. Una curva pasa por el punto $(1, 1)$ y posee la siguiente propiedad: La recta tangente a la curva en un punto (x, y) (x positivo) corta al eje y a una distancia x del origen de coordenadas. Obtenga la ordenada del punto de la curva cuya abscisa es $1,5$. Dé el resultado con cuatro cifras decimales exactas.
7. La suposición de que una población crece con una tasa constante conduce a la sencilla ecuación:

$$\frac{dP}{dt} = kP$$

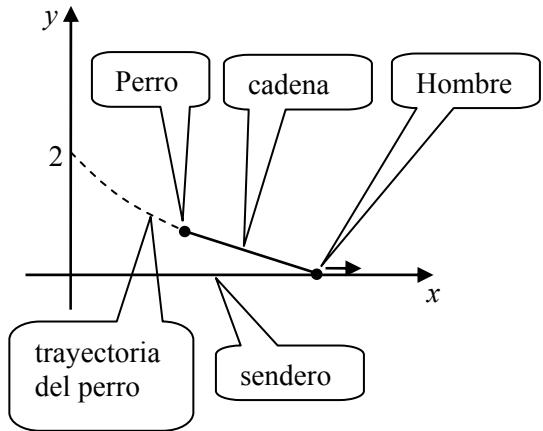
donde P : Población en el instante t .
 k : Tasa de crecimiento.

En la mayoría de los casos este es un modelo demasiado simple para un problema real, pues sucede que, al aumentar P (que lo haría exponencialmente, si se toma este modelo) aparecen factores ambientales y limitaciones de recursos materiales que actúan en contra de este aumento y provocan el efecto de disminuir la tasa de crecimiento. Un modelo más satisfactorio es:

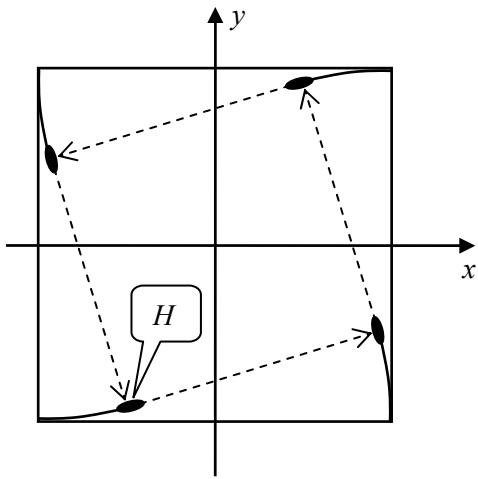
$$\frac{dP}{dt} = (a - bP)P$$

con a y b positivas, propuesto en 1840 por el matemático y biólogo belga P. F. Verhulst, y conocido como *ecuación logística*.

- a) Interprete el modelo logístico comparándolo con el modelo simple $\frac{dP}{dt} = kP$.
- b) Considere el caso $\frac{dP}{dt} = (0,9 - 0,01 \cdot P)P$; $P(0) = 10$ y determine la población en $t = 10$ con todas las cifras enteras exactas.
8. Un hombre saca a pasear a su perro. El hombre camina por un sendero en línea recta y el perro no quiere moverse, pero el hombre lo va halando con una cadena de tres metros de largo, como se muestra en la figura. Considere como instante inicial, el momento en que el perro se hallaba a dos metros del sendero. Defina un sistema coordenado xy donde el eje x coincide con el sendero y el eje y es perpendicular al x y pasa por la posición inicial del perro. Halle la trayectoria del perro para $0 \leq x \leq 6$ con error menor que 1 mm.



9. Sobre una mesa cuadrada de dos metros de lado, hay cuatro hormigas, una en cada esquina de la mesa. En el instante $t = 0$ comienzan a caminar las cuatro hormigas, todas a la misma velocidad, y lo hacen de manera que cada una se mueve en la dirección en que se encuentra en ese instante la hormiga a su derecha, como se muestra en la figura. Tome un sistema de referencia con origen en el centro de la mesa y ejes paralelos a los bordes y halle, con error menor que 1 mm la trayectoria de la hormiga H desde $x = -1$ hasta $x = 0$.



7.4 Métodos de paso múltiple

Todos los métodos estudiados en la sección anterior son de paso simple. En ellos el valor de y_{n+1} se calcula teniendo en cuenta solamente y_n mediante algún algoritmo que puede simbolizarse por

$$y_{n+1} = G(y_n)$$

Los métodos de paso múltiple, por el contrario, utilizan varios valores anteriores a y_n para calcular y_{n+1} . Así, si se usa los últimos $p + 1$ valores de la solución: $y_n, y_{n-1}, \dots, y_{n-p}$, para hallar y_{n+1} , se dice que se trata de un método de paso $p + 1$, que se simboliza:

$$y_{n+1} = G(y_n, y_{n-1}, \dots, y_{n-p})$$

Por ejemplo, un método de paso cuádruple sería del tipo:

$$y_{n+1} = G(y_n, y_{n-1}, y_{n-2}, y_{n-3})$$

Comparación entre los métodos de paso simple y paso múltiple

En términos generales, existen varias importantes diferencias entre estos dos tipos de métodos:

- Como en los métodos de paso múltiple se utiliza una mayor cantidad de información acerca de la solución ya calculada, se logra una mayor eficiencia computacional, en el sentido de una menor cantidad de operaciones para obtener una exactitud similar.
- Un método de paso $p + 1$ no puede funcionar mientras no se conozcan $p + 1$ valores de la solución: $y_0, y_1, y_2, \dots, y_p$. Como en el problema de Cauchy solo se conoce y_0 , es necesario utilizar algún método de paso simple para calcular los primeros p valores de la solución: y_1, y_2, \dots, y_p . Por esto se dice con frecuencia en el lenguaje técnico que los métodos de paso múltiple no son capaces de “arrancar”.
- Como los métodos de paso simple solamente utilizan el valor y_n en el cálculo de y_{n+1} , ellos se basan en aproximar la solución de la ecuación diferencial mediante un polinomio de Taylor en el punto x_n . Recuérdese que los métodos de Euler y Runge – Kutta fueron deducidos de esta forma. Los métodos de paso múltiple, por otra parte, puesto que usan varios valores de la solución, suelen estar basados en aproximar la solución de la ecuación diferencial mediante alguna función interpoladora, usualmente un polinomio.

Los métodos de Adams – Bashforth

Los métodos de Adams – Bashforth forman una familia de métodos de paso múltiple típica. Uno de los más usados es el de paso cuádruple que se va a deducir y estudiar a continuación con algún detalle. Después se extenderán los resultados obtenidos a otros métodos de Adams – Bashforth de diferente orden.

Sea, como hasta ahora, el problema de Cauchy:

$$\frac{dy}{dx} = f(x, y) \quad y(x_0) = y_0$$

y suponga que se han obtenido valores

$$y_0, y_1, y_2, \dots, y_n$$

los cuales son aproximaciones de la solución exacta:

$$y(x_0), y(x_1), y(x_2), \dots, y(x_n)$$

para valores de x tomados con paso h .

Se desea hallar y_{n+1} como aproximación de $y(x_{n+1})$.

Como $y(x)$ es la solución del problema de Cauchy, se cumple que:

$$\frac{dy(x)}{dx} = f(x, y(x))$$

Esto es,

$$dy(x) = f(x, y(x))dx$$

Integrando:

$$\int_{x_n}^{x_{n+1}} dy(x) = \int_{x_n}^{x_{n+1}} f(x, y(x))dx$$

Es decir:

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x))dx \quad (1)$$

Para calcular aproximadamente la integral que aparece en (1), se sustituirá la función $f(x, y(x))$ por el polinomio interpolador de tercer grado $p(x)$ correspondiente a los cuatro nodos:

$$(x_{n-3}, f_{n-3}), (x_{n-2}, f_{n-2}), (x_{n-1}, f_{n-1}), (x_n, f_n)$$

donde se utiliza la notación: $f_i = f(x_i, y_i)$ para $i = n-3, n-2, n-1, n$

De este modo, la fórmula exacta (1) da lugar a la ecuación:

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p(x)dx \quad (2)$$

Si se supone que los valores y_i son buenas aproximaciones de $y(x_i)$, entonces f_i será una adecuada aproximación de $f(x_i, y(x_i))$ y el polinomio $p(x)$ no diferirá mucho de la función $f(x, y(x))$ en el intervalo $[x_{n-3}, x_n]$. Nótese, sin embargo, que aun cuando todas estas aproximaciones fueran satisfactorias, se está empleando un polinomio interpolador calculado con nodos en el intervalo $[x_{n-3}, x_n]$ para calcular una integral en el intervalo $[x_n, x_{n+1}]$; este hecho tendrá una importante incidencia en la exactitud de la fórmula obtenida.

El polinomio interpolador puede expresarse mediante la fórmula de Lagrange como:

$$p(x) = L_n(x)f_n + L_{n-1}(x)f_{n-1} + L_{n-2}(x)f_{n-2} + L_{n-3}(x)f_{n-3}$$

así que la ecuación (2) pasa a ser:

$$y_{n+1} = y_n + f_n \int_{x_n}^{x_{n+1}} L_n(x) dx + f_{n-1} \int_{x_n}^{x_{n+1}} L_{n-1}(x) dx + f_{n-2} \int_{x_n}^{x_{n+1}} L_{n-2}(x) dx + f_{n-3} \int_{x_n}^{x_{n+1}} L_{n-3}(x) dx \quad (3)$$

donde: $L_n(x) = \frac{(x - x_{n-1})(x - x_{n-2})(x - x_{n-3})}{(x_n - x_{n-1})(x_n - x_{n-2})(x_n - x_{n-3})}$

$$L_{n-1}(x) = \frac{(x - x_n)(x - x_{n-2})(x - x_{n-3})}{(x_{n-1} - x_n)(x_{n-1} - x_{n-2})(x_{n-1} - x_{n-3})}$$

$$L_{n-2}(x) = \frac{(x - x_n)(x - x_{n-1})(x - x_{n-3})}{(x_{n-2} - x_n)(x_{n-2} - x_{n-1})(x_{n-2} - x_{n-3})}$$

$$L_{n-3}(x) = \frac{(x - x_n)(x - x_{n-1})(x - x_{n-2})}{(x_{n-3} - x_n)(x_{n-3} - x_{n-1})(x_{n-3} - x_{n-2})}$$

Las integrales de la ecuación (3) no son difíciles de calcular. A modo de ilustración se calculará la primera en detalle y se escribirá el resultado obtenido para las otras.

$$\int_{x_n}^{x_{n+1}} L_n(x) dx = \int_{x_n}^{x_{n+1}} \frac{(x - x_{n-1})(x - x_{n-2})(x - x_{n-3})}{(x_n - x_{n-1})(x_n - x_{n-2})(x_n - x_{n-3})} dx \quad (4)$$

Para simplificar los cálculos es útil hacer el cambio de variables:

$$s = x - x_n$$

En la figura 1 se muestra el significado geométrico de s y de los demás términos que aparecen en la integral.

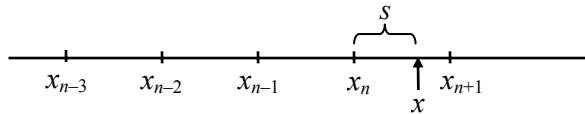


Figura 1

De la figura es obvio que:

$$\begin{aligned} x - x_{n-1} &= h + s \\ x - x_{n-2} &= 2h + s \\ x - x_{n-3} &= 3h + s \\ x_n - x_{n-1} &= h \\ x_n - x_{n-2} &= 2h \\ x_n - x_{n-3} &= 3h \end{aligned}$$

También se observa que:

$$\begin{aligned} x = x_n &\Leftrightarrow s = 0 \\ x = x_{n+1} &\Leftrightarrow s = h \end{aligned}$$

Como, además, $ds = dx$, la ecuación (4) se convierte en:

$$\begin{aligned}
 \int_{x_n}^{x_{n+1}} L_n(x) dx &= \int_0^h \frac{(s+h)(s+2h)(s+3h)}{(h)(2h)(3h)} ds \\
 &= \frac{1}{6h^3} \int_0^h (s^3 + 6hs^2 + 11h^2s + 6h^3) ds \\
 &= \frac{1}{6h^3} \left[\frac{1}{4}s^4 + 2hs^3 + \frac{11}{2}h^2s^2 + 6h^3s \right]_0^h \\
 &= \frac{1}{6h^3} \left[\frac{1}{4} + 2 + \frac{11}{2} + 6 \right] h^4 \\
 &= \frac{55}{24} h
 \end{aligned}$$

En forma similar:

$$\int_{x_n}^{x_{n+1}} L_{n-1}(x) dx = -\frac{59}{24} h \quad \int_{x_n}^{x_{n+1}} L_{n-2}(x) dx = \frac{37}{24} h \quad \int_{x_n}^{x_{n+1}} L_{n-3}(x) dx = -\frac{9}{24} h$$

Sustituyendo las integrales calculadas, la ecuación (3) se transforma en:

$$y_{n+1} = y_n + \frac{h}{24} (55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \quad (5)$$

que es la fórmula de Adams – Bashforth de orden 4 (AB4). Integrando en $[x_n, x_{n+1}]$ el error de interpolación de Lagrange:

$$R(x) = \frac{\frac{d^4}{dx^4} f(x, y(x)) \Big|_{x=\xi}}{4!} (x - x_n)(x - x_{n-1})(x - x_{n-2})(x - x_{n-3})$$

se obtiene una expresión para el error local del método:

$$\text{Error local: } \frac{251}{720} y^{(5)}(c) h^5 \quad (6)$$

El error total será del orden de h^4 , o sea:

$$\text{Error total: } O(h^4)$$

Como se ve, el método de Adams – Bashforth de paso cuádruple, es un método de cuarto orden (error del orden de h^4), al igual que RK4.

Los métodos de Adams – Bashforth de otros órdenes se obtienen de forma similar. A continuación se resumen las fórmulas correspondientes a los más usados, así como los errores locales correspondientes.

Método	Fórmula	Error local	Error total
AB2	$y_{n+1} = y_n + \frac{h}{2}(3f_n - f_{n-1})$	$\frac{5}{12}y^{(3)}(c)h^3$	$O(h^2)$
AB3	$y_{n+1} = y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2})$	$\frac{3}{8}y^{(4)}(c)h^4$	$O(h^3)$
AB4	$y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$	$\frac{251}{720}y^{(5)}(c)h^5$	$O(h^4)$

Algoritmo en seudo código para AB4

El algoritmo que sigue obtiene la solución aproximada del problema de Cauchy

$$\begin{aligned}\frac{dy}{dx} &= f(x, y) \\ y(x_0) &= y_0\end{aligned}$$

en un intervalo $x_0 \leq x \leq x_F$ con paso $h > 0$. Se suponen conocidos: la función $f(x, y)$, las condiciones iniciales (x_0, y_0) el valor final x_F y el paso h .

```

Calcular  $y_1, y_2, y_3$  mediante RK4
for  $i = 0$  to  $3$ 
     $f_i := f(x_i, y_i)$ 
end
 $n := 3$ 
 $x_n := x_0 + nh$ 
do while  $x_n < x_F$ 
     $(x_n, y_n)$  es un punto de la solución aproximada
     $y_{n+1} := y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$ 
     $x_{n+1} := x_n + h$ 
     $f_{n+1} := f(x_{n+1}, y_{n+1})$ 
     $n := n + 1$ 
end
Terminar

```

Ejemplo 1

Resuelva el problema de Cauchy

$$\frac{dy}{dx} = \frac{1}{2}y \quad y(0) = 0,5$$

en el intervalo $0 \leq x \leq 5$ mediante RK4 y AB4 con pasos $h = 0,2$ y $0,1$. Compare ambos resultados con la solución exacta

$$y = \frac{1}{2}e^{\frac{x}{2}}$$

Solución:

Los valores exactos de la solución son:

$$y(1) = 0,8243606; \quad y(2) = 1,3591409; \quad y(3) = 2,2408445; \quad y(4) = 3,6945280; \quad y(5) = 6,0912469.$$

En la tabla 1 se muestran los resultados obtenidos. Observe que, a pesar de que ambos métodos son de orden 4, los resultados de RK4 son más exactos que los de AB4; esto se debe a que el coeficiente de h^4 es menor en el primero que en el segundo. Tenga presente que AB4 requiere cuatro veces menos evaluaciones de $f(x, y)$ que RK4 (AB4: una evaluación por paso; RK4: cuatro evaluaciones por paso).

x	$h = 0,2$ RK4		$h = 0,2$ AB4		$h = 0,1$ RK4		$h = 0,2$ AB4	
	Solución	Error	Solución	Error	Solución	Error	Solución	Error
0	0,500000	0,000000	0,500000	0,000000	0,500000	0,000000	0,500000	0,000000
1	0,824360	0,000001	0,824355	0,000006	0,824361	0,000000	0,824360	0,000001
2	1,359140	0,000001	1,359112	0,000029	1,359141	0,000000	1,359139	0,000002
3	2,240842	0,000002	2,240764	0,000080	2,240844	0,000000	2,240838	0,000006
4	3,694522	0,000006	3,694340	0,000188	3,694528	0,000000	3,694514	0,000014
5	6,091235	0,000012	6,090845	0,000402	6,091246	0,000001	6,091218	0,000029

Tabla 1

Comparando los resultados con paso $h = 0,2$ y $h = 0,1$ se aprecia que, en ambos casos el error disminuye unas 16 veces al disminuir el paso a la mitad, lo cual se debe a que se trata de métodos de cuarto orden.

Los métodos de Adams – Moulton

Como ya se ha señalado, la causa principal de la menor exactitud de los métodos de Adams – Bashforth radica en el hecho de que utilizan un polinomio interpolador en el intervalo $[x_{n-p}, x_n]$ para aproximar a la función $f(x, y(x))$ en el intervalo $[x_n, x_{n+1}]$, y ya se sabe lo inconveniente de la extrapolación polinomial.

Los métodos de Adams – Moulton resuelven este problema tomando los nodos de interpolación en $[x_{n-p}, x_{n+1}]$, lo cual aumenta considerablemente la exactitud pero, como se verá, introduce nuevas dificultades. Al igual que antes, se deducirá solamente la fórmula para el método de cuarto orden (que, en este caso, es de paso triple).

Sea el problema de Cauchy:

$$\frac{dy}{dx} = f(x, y) \quad y(x_0) = y_0$$

y suponga que se han obtenido valores aproximados de la solución:

$$y_0, y_1, y_2, \dots, y_n$$

para los valores $x_0, x_1, x_2, \dots, x_n$ de x tomados con paso h . Considérese de nuevo la ecuación (1):

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x)) dx$$

Ahora la integral se calculará aproximando la función $f(x, y(x))$ mediante el polinomio interpolador $p(x)$ de tercer grado correspondiente a los cuatro nodos:

$$(x_{n-2}, f_{n-2}), (x_{n-1}, f_{n-1}), (x_n, f_n), (x_{n+1}, f_{n+1})$$

Nótese que, al proceder así, se está suponiendo conocido $f_{n+1} = f(x_{n+1}, y_{n+1})$, que depende precisamente del valor y_{n+1} que se está procurando calcular. Sin embargo, no es nada extraño en Álgebra, plantear una ecuación que contenga a la incógnita en ambos miembros. El polinomio interpolador, utilizando de nuevo la fórmula de Lagrange, sería:

$$p(x) = L_{n+1}(x)f_{n+1} + L_n(x)f_n + L_{n-1}(x)f_{n-1} + L_{n-2}(x)f_{n-2}$$

Sustituyendo $p(x)$ en lugar de $f(x, y(x))$ en el integrando de (1), resulta:

$$y_{n+1} = y_n + f_{n+1} \int_{x_n}^{x_{n+1}} L_{n+1}(x) dx + f_n \int_{x_n}^{x_{n+1}} L_n(x) dx + f_{n-1} \int_{x_n}^{x_{n+1}} L_{n-1}(x) dx + f_{n-2} \int_{x_n}^{x_{n+1}} L_{n-2}(x) dx \quad (7)$$

$$\text{donde: } L_{n+1}(x) = \frac{(x - x_n)(x - x_{n-1})(x - x_{n-2})}{(x_{n+1} - x_n)(x_{n+1} - x_{n-1})(x_{n+1} - x_{n-2})}$$

$$L_n(x) = \frac{(x - x_{n+1})(x - x_{n-1})(x - x_{n-2})}{(x_n - x_{n+1})(x_n - x_{n-1})(x_n - x_{n-2})}$$

$$L_{n-1}(x) = \frac{(x - x_{n+1})(x - x_n)(x - x_{n-2})}{(x_{n-1} - x_{n+1})(x_{n-1} - x_n)(x_{n-1} - x_{n-2})}$$

$$L_{n-2}(x) = \frac{(x - x_{n+1})(x - x_n)(x - x_{n-1})}{(x_{n-2} - x_{n+1})(x_{n-2} - x_n)(x_{n-2} - x_{n-1})}$$

A modo de muestra, se calculará en detalle la última de estas integrales.

$$\int_{x_n}^{x_{n+1}} L_{n-2}(x) dx = \int_{x_n}^{x_{n+1}} \frac{(x - x_{n+1})(x - x_n)(x - x_{n-1})}{(x_{n-2} - x_{n+1})(x_{n-2} - x_n)(x_{n-2} - x_{n-1})} dx$$

Haciendo de nuevo el cambio de variables:

$$s = x - x_n$$

$$\text{resulta: } x - x_{n+1} = x - x_n + x_n - x_{n+1} = (x - x_n) - (x_{n+1} - x_n) = s - h$$

$$x - x_n = s$$

$$x - x_{n-1} = s + h$$

$$x_{n-2} - x_{n+1} = -3h$$

$$\begin{aligned}
x_{n-2} - x_n &= -2h \\
x_{n-2} - x_{n-1} &= -h \\
x = x_n &\Leftrightarrow s = 0 \\
x = x_{n+1} &\Leftrightarrow s = h \\
dx &= ds
\end{aligned}$$

La integral queda:

$$\begin{aligned}
\int_{x_n}^{x_{n+1}} L_{n-2}(x) dx &= \int_0^h \frac{(s-h)s(s+h)}{(-3h)(-2h)(-h)} ds \\
&= -\frac{1}{6h^3} \int_0^h (s^3 - h^2 s) ds \\
&= -\frac{1}{6h^3} \left[\frac{1}{4}s^4 - \frac{1}{2}h^2 s^2 \right]_0^h \\
&= -\frac{1}{6h^3} \left[\frac{1}{4}h^4 - \frac{1}{2}h^4 \right] \\
&= \frac{1}{24}h
\end{aligned}$$

De forma similar:

$$\int_{x_n}^{x_{n+1}} L_{n+1}(x) dx = \frac{9}{24}h \quad \int_{x_n}^{x_{n+1}} L_n(x) dx = \frac{19}{24}h \quad \int_{x_n}^{x_{n+1}} L_{n-1}(x) dx = -\frac{5}{24}h$$

Sustituyendo en (7):

$$y_{n+1} = y_n + \frac{h}{24} (9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}) \quad (8)$$

Que es la fórmula de Adams – Moulton de orden 4 (AM4). Puede probarse, integrando el error de interpolación en el intervalo $[x_n, x_{n+1}]$, que el error local tiene la forma:

$$\text{Error local: } -\frac{19}{720} y^{(5)}(\tau)h^5 \quad (9)$$

y que el error total es $O(h^4)$; por esta razón se dice que es un método de orden 4. Nótese, sin embargo, que para determinar y_{n+1} se necesitan y_n , y_{n-1} y y_{n-2} , por lo que se trata de un método de paso triple.

Si se comparan los errores locales de los métodos AB4 y AM4, ecuaciones (9) y (6), se aprecia que el de AB4 es unas 13 veces mayor que AM4. Esto, como se había previsto, se debe a que en el método de Adams – Moulton no se ha empleado la extrapolación de $p(x)$.

Sin embargo, la fórmula AM4, al igual que los demás métodos de Adams – Moulton, presentan el inconveniente de que el valor y_{n+1} buscado no puede ser despejado por lo general. Por esta razón se dice que es un método implícito. La ecuación (8) suele resolverse por el método iterativo simple, estudiado en el capítulo 2. Como se recordará, la ecuación

$$f(x) = 0$$

se escribe en la forma

$$x = g(x)$$

y se realiza el proceso iterativo

$$x_{n+1} = g(x_n) \quad x_0 \in R \quad (10)$$

Para que el proceso (10) converja, es suficiente que $|g'(x)| \leq K < 1$ en un entorno de la solución que incluya a x_0 . En el caso de la ecuación (8), la incógnita es y_{n+1} , así que la solución ya está escrita en la forma adecuada para iterar:

$$y_{n+1}^{(k+1)} = y_n + \frac{h}{24} [9f(x_{n+1}, y_{n+1}^{(k)}) + 19f_n - 5f_{n-1} + f_{n-2}] \quad k = 0, 1, 2, \dots$$

Para que el proceso iterativo converja, es suficiente que:

$$\left| \frac{\partial}{\partial y_{n+1}} \left[y_n + \frac{h}{24} [9f(x_{n+1}, y_{n+1}) + 19f_n - 5f_{n-1} + f_{n-2}] \right] \right| \leq K < 1$$

o sea:

$$\left| \frac{9h}{24} \frac{\partial f}{\partial y} \right| \leq K < 1$$

Como se observa, ya que h es un valor reducido, la condición de convergencia se satisface sin dificultad; es más, la cantidad

$$\left| \frac{9h}{24} \frac{\partial f}{\partial y} \right|$$

es, usualmente, próxima a cero, lo cual garantiza una rápida convergencia del proceso iterativo. Nótese, sin embargo, que este proceso iterativo converge a la solución de la ecuación (8), no a la solución exacta $y(x_{n+1})$ de la ecuación diferencial.

De forma análoga al caso de orden cuatro, se pueden deducir otras fórmulas de Adams – Moulton. Las más usadas se resumen a continuación:

Método	Fórmula	Error local	Error total
AM2	$y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n)$	$-\frac{1}{12} y^{(3)}(c)h^3$	$O(h^2)$
AM3	$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1})$	$-\frac{1}{24} y^{(4)}(c)h^4$	$O(h^3)$
AM4	$y_{n+1} = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})$	$-\frac{19}{720} y^{(5)}(c)h^5$	$O(h^4)$

Métodos predictor – corrector

Los métodos predictor – corrector están constituidos por un método implícito de buena exactitud (que se llama *ecuación correctora*) y un método explícito (*ecuación predictora*) que aporta una aproximación inicial adecuada de y_{n+1} para el proceso iterativo del método implícito.

Los algoritmos de Adams – Bashforth y de Adams – Moulton suelen emplearse para formar pares predictor – corrector. Para ello se seleccionan métodos del mismo orden, de modo que la aproximación inicial sea suficientemente buena para que la ecuación correctora no haya que aplicarla más de una o dos veces.

Como ilustración, y por ser uno de los más empleados, se verá a continuación el método predictor – corrector de Adams de orden cuatro (ABM4):

Ecuación predictora: AB4

$$y_{n+1}^{(0)} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$$

Ecuación correctora: AM4

$$y_{n+1}^{(k+1)} = y_n + \frac{h}{24}[9f(x_{n+1}, y_{n+1}^{(k)}) + 19f_n - 5f_{n-1} + f_{n-2}] \quad k = 0, 1, 2, \dots$$

Si ya han sido calculados $y_0, y_1, y_2, \dots, y_n$ con la exactitud requerida, el cálculo de y_{n+1} requerirá evaluar la función f una vez para hallar $y_{n+1}^{(0)}$ y tantas veces como se utilice la ecuación correctora implícita. Si se requiere aplicar el proceso corrector tres o más veces, entonces cada nuevo punto de la solución necesitará cuatro o más evaluaciones de f y se perderá la principal ventaja del método AM4 sobre RK4, que requiere cuatro evaluaciones de f por cada paso. El método predictor de Adams de orden cuatro se aplica de la siguiente forma:

Se selecciona un paso h suficientemente pequeño para que el proceso corrector converja rápidamente, de manera que solo se requiera aplicar una vez la ecuación correctora. Si esto se consigue, entonces $y_{n+1}^{(1)}$ se puede tomar como la solución de la ecuación correctora, de modo que su error será:

$$y(x_{n+1}) - y_{n+1}^{(1)} = -\frac{19}{720}h^5 y^{(5)}(\tau) \quad (11)$$

Como se ve, se ha supuesto que $y_0, y_1, y_2, \dots, y_n$ poseen un error muy pequeño, de manera que el error total de $y_{n+1}^{(1)}$ puede estimarse como su error local. Por otra parte, el error total de $y_{n+1}^{(0)}$ puede estimarse como el error local de AB4, es decir:

$$y(x_{n+1}) - y_{n+1}^{(0)} = \frac{251}{720}h^5 y^{(5)}(c) \quad (12)$$

Tomando aproximadamente $y^{(5)}(\tau) = y^{(5)}(c) = y^v$ y restando la ecuación (11) de la (12), se tiene:

$$y_{n+1}^{(1)} - y_{n+1}^{(0)} \approx \left(\frac{251}{720} + \frac{19}{720}\right)h^5 y^v = \frac{3}{8}h^5 y^v$$

Este resultado permite estimar el producto $h^5 y^v$:

$$h^5 y^v \approx \frac{8}{3}[y_{n+1}^{(1)} - y_{n+1}^{(0)}]$$

Sustituyendo ahora en la ecuación (11):

$$y(x_{n+1}) - y_{n+1}^{(1)} \approx -\frac{19}{720} \cdot \frac{8}{3} [y_{n+1}^{(1)} - y_{n+1}^{(0)}] = -\frac{19}{270} [y_{n+1}^{(1)} - y_{n+1}^{(0)}]$$

Como $\frac{270}{19} = 14,2105$, se acostumbra a tomar:

$$y(x_{n+1}) - y_{n+1}^{(1)} \approx -\frac{1}{14} [y_{n+1}^{(1)} - y_{n+1}^{(0)}] \quad (13)$$

La fórmula (13) significa que el paso h debe seleccionarse de modo que se cumplan dos requisitos:

1. El proceso corrector converge en un paso.
2. El error introducido en un paso, que puede estimarse como $-\frac{1}{14} [y_{n+1}^{(1)} - y_{n+1}^{(0)}]$ es aceptable de acuerdo con el error total que se puede acumular en la región solución.

Si alguna de estas condiciones no se cumple, debe tomarse un valor menor de h .

Ejemplo 2

Resuelva el problema de Cauchy

$$\frac{dy}{dx} = \frac{1}{2}y \quad y(0) = 0,5$$

en el intervalo $0 \leq x \leq 2$ mediante el método predictor corrector de Adams de orden cuatro (ABM4), con paso $h = 0,2$. Obtenga aproximadamente el error introducido en cada paso y verifique mediante el cálculo exacto del error. La solución exacta es

$$y = \frac{1}{2} e^{\frac{x}{2}}$$

Solución:

En la tabla 2 se muestran los resultados. Observe como el error obtenido en cada paso puede estimarse del orden de $-0,2 \cdot 10^{-6}$, lo cual hace prever que en siete pasos del método predictor – corrector el error total estará entre $-0,000001$ y $-0,000002$. Nótese que en la realidad, el error total acumulado no sobrepasa $-0,000001$.

n	x_n	$y_n^{(0)}$	$y_n^{(1)}$	$-\frac{1}{14} (y_n^{(1)} - y_n^{(0)})$	$y(x_n)$	Error total real
0	0,0	0,500000		0,500000	0,000001	
1	0,2	0,552585		0,552585	0,000000	
2	0,4	0,610701		0,610701	0,000000	
3	0,6	0,674929		0,674929	0,000000	
4	0,8	0,745910	0,745912	-0,00000014	0,745912	0,000000
5	1,0	0,824358	0,824361	-0,00000021	0,824361	0,000000
6	1,2	0,911057	0,911060	-0,00000021	0,911059	-0,000001
7	1,4	1,006874	1,006877	-0,00000021	1,006876	-0,000001
8	1,6	1,112768	1,112771	-0,00000021	1,112770	-0,000001
9	1,8	1,229799	1,229802	-0,00000021	1,229801	-0,000001
10	2,0	1,359138	1,359142	-0,00000029	1,359141	-0,000001

Tabla 2

Algoritmo en seudo código para el método predictor – corrector de Adams

El algoritmo que sigue obtiene la solución aproximada del problema de Cauchy

$$\frac{dy}{dx} = f(x, y)$$

$$y(x_0) = y_0$$

en un intervalo $x_0 \leq x \leq x_F$ con paso $h > 0$. Se suponen conocidos: la función $f(x, y)$, las condiciones iniciales (x_0, y_0) el valor final x_F y el paso h .

```

for  $n = 0$  to 2 {Mediante RK4 se hallan los primeros tres puntos de la solución}
     $(x_n, y_n)$  es un punto de la solución aproximada
     $f_n := f(x_n, y_n)$ 
     $K_1 := h f_n$ 
     $K_2 := h f(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_1)$ 
     $K_3 := h f(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_2)$ 
     $K_4 := h f(x_n + h, y_n + K_3)$ 
     $y_{n+1} := y_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4)$ 
     $x_{n+1} := x_n + h$ 
end
 $n := 3$ 
 $f_n := f(x_n, y_n)$ 
do while  $x_n < x_F$ 
     $(x_n, y_n)$  es un punto de la solución aproximada
     $y_{n+1}^{(0)} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$ 
     $y_{n+1} = y_n + \frac{h}{24}[9f(x_{n+1}, y_{n+1}^{(0)}) + 19f_n - 5f_{n-1} + f_{n-2}]$ 
     $x_{n+1} := x_n + h$ 
     $n := n + 1$ 
     $f_n := f(x_n, y_n)$ 
end
Terminar

```

El método predictor – corrector de Adams de orden cuatro se compara muy bien con RK4 en cuanto a exactitud y estabilidad. En el ejemplo que sigue, se compara la exactitud de ambos procedimientos en la solución de una ecuación diferencial inestable.

Ejemplo 3

Resuelva el problema de Cauchy

$$\frac{dy}{dx} = \frac{1}{2}y \quad y(0) = 0,5$$

en el intervalo $0 \leq x \leq 5$, con paso $h = 0,2$, mediante los métodos RK4 y predictor – corrector ABM4 . Compare la exactitud de ambos resultados.

Solución:

En la tabla 3 se muestran, para $x = 0, 1, 2, 3, 4$ y 5 los resultados obtenidos con cada método y sus respectivos errores calculados por comparación con la solución exacta.

x	Solución de RK4	Solución de ABM4	Solución Exacta	Error de RK4	Error de ABM4
0	0,500000	0,500000	0,500000	0,000000	0,000000
1	0,824360	0,824361	0,824361	0,000001	0,000000
2	1,359140	1,359142	1,359141	0,000001	-0,000001
3	2,240842	2,240847	2,240845	0,000003	-0,000002
4	3,694522	3,694535	3,694528	0,000006	-0,000007
5	6,091235	6,091263	6,091247	0,000012	-0,000015

Tabla 3

Como se aprecia, los resultados poseen una exactitud similar. No obstante, el método predictor – corrector requirió la mitad de las evaluaciones de f (2 por cada paso) que RK4 (4 por cada paso)

Estabilidad de los métodos de Adams

Al aplicar a la ecuación diferencial modelo estable:

$$\frac{dy}{dx} = -Ay \quad A > 0$$

$$\text{el método de AB4: } y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$$

se obtiene, puesto que $f_i = -Ay_i$ para $i = n, n-1, n-2$ y $n-3$,

$$y_{n+1} = y_n + \frac{h}{24}(-55Ay_n + 59Ay_{n-1} - 37Ay_{n-2} + 9Ay_{n-3})$$

$$\text{es decir: } y_{n+1} = y_n + \frac{hA}{24}(-55y_n + 59y_{n-1} - 37y_{n-2} + 9y_{n-3}) \quad (14)$$

Se trata de una ecuación en diferencias lineal y homogénea, cuyas soluciones son del tipo:

$$y_i = \beta^i$$

Los valores de β se obtienen sustituyendo en (14):

$$\beta^{n+1} = \beta^n + \frac{hA}{24}(-55\beta^n + 59\beta^{n-1} - 37\beta^{n-2} + 9\beta^{n-3})$$

Dividiendo por β^{n-3} y trasponiendo:

$$\beta^4 - \beta^3 + \frac{hA}{24}(55\beta^3 - 59\beta^2 + 37\beta - 9) = 0 \quad (15)$$

La ecuación en diferencias (14) es estable si y solo si todas las raíces (reales e imaginarias) de la ecuación característica (15) tienen valor absoluto menor que 1. Dado que la ecuación (15) es de cuarto grado, es muy engoroso obtener una expresión analítica de sus raíces en términos de hA , pero puede ser resuelta numéricamente para diferentes valores de hA . La tabla (4) muestra las cuatro raíces $\beta_1, \beta_2, \beta_3$ y β_4 de (15) para valores de hA con incremento 0,05. Nótese como, a partir de $hA > 0,3$ aparece una raíz β_2 con módulo mayor que 1.

hA	β_1	β_2	β_3, β_4
0,05	0,9512	-0,3713	$0,1527 \pm 0,1725 i$
0,10	0,9048	-0,5234	$0,1947 \pm 0,2032 i$
0,15	0,8697	-0,6530	$0,2242 \pm 0,2232 i$
0,20	0,8189	-0,7729	$0,2479 \pm 0,2389 i$
0,25	0,7792	-0,8879	$0,2679 \pm 0,2525 i$
0,30	0,7418	-1,0000	$0,2853 \pm 0,2650 i$
0,35	0,7069	-1,1104	$0,3007 \pm 0,2771 i$
0,40	0,6745	-1,2198	$0,3143 \pm 0,2890 i$

Tabla 4

De este análisis se concluye que el método de Adams – Bashforth de orden cuatro es inestable para la ecuación estable modelo para valores de hA por encima de 0,3. Como se ve, valores de h relativamente pequeños pueden dar lugar a la inestabilidad del método AB4. Por suerte, el método de Adams – Moulton presenta una mayor estabilidad. En efecto, la ecuación AM4:

$$y_{n+1} = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})$$

aplicada a la ecuación diferencial modelo:

$$\frac{dy}{dx} = -Ay \quad A > 0$$

$$\text{da lugar al problema: } y_{n+1} = y_n + \frac{hA}{24}(-9y_{n+1} - 19y_n + 5y_{n-1} - y_{n-2}) \quad (16)$$

que es una ecuación en diferencias lineal y homogénea de tercer orden. Sus soluciones, del tipo $y_i = \beta^i$ se obtienen, igual que antes, sustituyendo en la última ecuación:

$$\beta^{n+1} = \beta^n + \frac{hA}{24}(-9\beta^{n+1} - 19\beta^n + 5\beta^{n-1} - \beta^{n-2})$$

Dividiendo por β^{n-2} y trasponiendo:

$$\beta^3 - \beta^2 + \frac{hA}{24}(9\beta^3 + 19\beta^2 - 5\beta + 1) = 0 \quad (17)$$

La ecuación (16) es estable si las raíces de la ecuación característica (17) son menores que 1 en valor absoluto. La tabla 5 muestra las tres raíces de la ecuación (17) para valores de hA tomados con incremento 0,5. A partir de $hA = 3$ comienza la zona de inestabilidad del método de AM4.

hA	β_1	β_2	β_3
0,0	1,0000	0,0000	0,0000
0,5	0,6058	0,1285	-0,2255
1,0	0,3333	0,2240	-0,6530
1,5	$0,2309 + 0,1242 i$	$0,2309 - 0,1242 i$	-0,5819
2,0	$0,2039 + 0,1505 i$	$0,2039 - 0,1505 i$	-0,7412
2,5	$0,1874 + 0,1611 i$	$0,1874 - 0,1611 i$	-0,8801
3,0	$0,1765 + 0,1664 i$	$0,1765 - 0,1664 i$	-1,0000
3,5	$0,1688 + 0,1693 i$	$0,1688 - 0,1693 i$	-1,1033
4,0	$0,1631 + 0,1711 i$	$0,1631 - 0,1711 i$	-1,1929

Tabla 5

Al aplicar el método predictor – corrector de Adams de orden 4, puesto que la ecuación predictora se utiliza solamente para hallar una aproximación inicial para comenzar el proceso iterativo de Adams – Moulton, la inestabilidad viene determinada por este último algoritmo y no por la ecuación de Adams – Bashforth.

Ejemplo 4

Resuelva el problema de Cauchy:

$$\frac{dy}{dx} = -y \quad y(0) = 1$$

con paso $h = 0,5$ en el intervalo $0 \leq x \leq 7$ utilizando a) El método de Adams – Bashforth de orden cuatro; b) El método predictor – corrector de Adams de orden cuatro. Compare con la solución exacta $y = e^{-x}$.

Solución:

La tabla 6 muestra todos los resultados obtenidos. En ambos casos, los tres primeros puntos de la solución numérica fueron calculados con RK4. El método predictor – corrector (ABM4) se aplicó utilizando, en cada paso, una sola vez la ecuación correctora de Adams – Moulton. Como se ve, el método de Adams – Bashforth presenta inestabilidad ($hA = 0,5 > 0,3$) lo cual se manifiesta en errores que crecen con gran rapidez hasta el punto de que ya el tercer punto de la solución hallado por AB4 no tienen ninguna cifra significativa exacta (el error alterna su signo debido a que la raíz

de la ecuación característica con módulo mayor que 1 es negativa). El método de Adams – Moulton, por el contrario, se encuentra en su zona de estabilidad ($hA = 0,5 < 3$) y por ello, se obtiene la solución con tres cifras decimales exactas, a pesar de que el paso tomado es mucho mayor que lo recomendable para este ejemplo.

x	$y(x)$	AB4	Error en AB4	ABM4	Error en ABM4
0,0	1,000000	1,000000	0,000000	1,000000	0,000000
0,5	0,606531	0,606771	-0,000240	0,606771	-0,000240
1,0	0,367879	0,368171	-0,000292	0,368171	-0,000292
1,5	0,223130	0,223395	-0,000265	0,223395	-0,000265
2,0	0,135335	0,139746	-0,004411	0,134476	-0,000859
2,5	0,082085	0,084182	-0,002097	0,080918	-0,001167
3,0	0,049787	0,056326	-0,006539	0,048806	-0,000981
3,5	0,030197	0,029425	0,000772	0,029382	-0,000815
4,0	0,018316	0,026255	-0,007939	0,017672	-0,000944
4,5	0,011109	0,004706	0,006403	0,010642	-0,000437
5,0	0,006738	0,019464	-0,012726	0,006407	-0,000331
5,5	0,004087	-0,011755	0,015842	0,003855	-0,000232
6,0	0,002479	0,026937	-0,024458	0,002320	-0,000159
6,5	0,001503	-0,032523	0,034026	0,001397	-0,000106
7,0	0,000912	0,050579	-0,049667	0,000841	-0,000071

Tabla 6

Ejercicios

1. Dada la ecuación diferencial:

$$\frac{dy}{dx} = x^2 - y^2 \quad y(1) = 0,8$$

Obtenga su solución en el intervalo $1 \leq x \leq 3$ mediante el método de Adams – Bashforth con 4 cifras decimales exactas.

2. Dada la ecuación diferencial:

$$\frac{dy}{dx} = x^2 + y^2 \quad y(1) = 0,8$$

Obtenga su solución en el intervalo $1 \leq x \leq 1,5$ mediante el método de Adams – Bashforth con error absoluto menor que 0,0001. Compare con el resultado del ejercicio anterior y explique.

3. Resuelva la ecuación diferencial

$$(e^x + y)dx - (y^2 + 1)dy = 0 \quad y(0) = 1$$

utilizando el método predictor – corrector de Adams de orden 4. Obtenga la solución con cuatro cifras decimales exactas en el intervalo $0 \leq x \leq 2$.

4. Se sabe que las variables ρ y θ están ligadas por la ecuación:

$$\rho^2 \cos \theta d\theta + (\rho + \theta) e^\theta d\rho = 0$$

y que para $\theta = \pi/6$ la variable ρ toma el valor 2. Determine, con cuatro cifras decimales exactas el valor que toma ρ para $\theta = \pi/4$. Utilice el método predictor – corrector de Adams de orden 4.

5. Un cubito de azúcar de 10 g se coloca en un gran recipiente de agua. Se quiere conocer su velocidad al cabo de dos segundos, sabiendo que la fuerza de empuje es 0,4 veces la fuerza peso, que la fuerza resistiva es proporcional a la velocidad (coeficiente 0,8) y que la masa del cubito disminuye a razón de dos gramos por segundo debido a la disolución del azúcar en el agua. Obtenga la respuesta en cm/s mediante el método predictor – corrector de Adams de orden cuatro con dos cifras decimales exactas. Nota: La segunda ley de Newton para la Mecánica, en su forma general, establece que

$$F = \frac{d(mv)}{dt}$$

6. Una curva pasa por el punto (1, 2). Se sabe, además, que esta curva posee la propiedad de que la recta tangente a ella en un punto cualquiera P corta al eje x en un punto A y al eje y en un punto B tales que, P es el punto medio del segmento AB . Se quiere conocer en qué punto esta curva corta a la recta $x = 3$. Obtenga la solución con cuatro cifras decimales exactas mediante el método predictor – corrector de Adams de orden cuatro.
7. Elabore un algoritmo en seudo código para el método predictor – corrector tomando como ecuación predictora la de Adams – Bashforth de orden 2 y como ecuación correctora la de Adams – Moulton de orden 2.

7.5 Ecuaciones diferenciales con condiciones iniciales

En esta sección se trata acerca de la solución de las ecuaciones diferenciales ordinarias de orden superior y de otros problemas equivalentes a ellas. Para resolver estas ecuaciones se requiere conocer tantos valores de la variable dependiente como orden posee la ecuación. Si todas las condiciones vienen dadas para el mismo valor de la variable independiente, este valor se toma como *inicial* y el problema se llama *de condiciones iniciales*. Este será el caso que será tratado ahora.

La estrategia que se sigue para resolver numéricamente este tipo de problema es transformarlo en un cierto tipo de sistema de ecuaciones diferenciales de primer orden con condiciones iniciales, el cual, después, se resuelve aplicando cualquiera de los métodos numéricos desarrollados en las secciones anteriores, a los cuales se les modifica ligeramente. Por esta razón se verá a continuación que características deben poseer estos sistemas y cómo una ecuación diferencial de orden superior se puede transformar en tal sistema.

Problema de Cauchy de orden m

Considérese el sistema de m ecuaciones diferenciales del siguiente tipo:

$$\begin{aligned}\frac{du_1}{dx} &= f_1(x, u_1, u_2, \dots, u_m) \\ \frac{du_2}{dx} &= f_2(x, u_1, u_2, \dots, u_m) \\ &\vdots \\ \frac{du_m}{dx} &= f_m(x, u_1, u_2, \dots, u_m)\end{aligned}$$

$$u_1(x_0) = u_{10}$$

$$u_2(x_0) = u_{20}$$

con las condiciones iniciales:

⋮

$$u_m(x_0) = u_{m0}$$

donde u_1, u_2, \dots, u_m son funciones desconocidas de la variable independiente x mientras que $x_0, u_{10}, u_{20}, \dots, u_{m0}$ son números reales conocidos. Observe que no se trata de un sistema cualquiera de ecuaciones diferenciales; debe tener las siguientes características:

- El sistema contiene tantas ecuaciones diferenciales como variables dependientes.
- Existe solamente una variable independiente.
- Todas las ecuaciones son de primer orden.
- En la ecuación número i existe una sola derivada que corresponde a la variable u_i .
- La derivada que contiene cada ecuación aparece despejada en uno de los miembros de la ecuación.
- Para un valor x_0 de la variable independiente se conoce el valor que toman todas las variables dependientes.

A un sistema de ecuaciones diferenciales con todas estas características se le llamará un *problema de Cauchy de orden m*.

Ejemplo 1

De los siguientes sistemas de ecuaciones, diga cuáles son problemas de Cauchy y cuáles no. Cuando no lo sea, analice si se puede transformar en un problema de Cauchy.

a) $\begin{aligned} \frac{dx}{dt} &= x - ty + 2 & x(3) &= 0 \\ \frac{dy}{dt} &= xy - e^t \cos xy - y & y(3) &= -1,8 \end{aligned}$

b) $\begin{aligned} \frac{dv_1}{dx} &= v_1 v_3 + x & v_1(0) &= 2 \\ \frac{dv_2}{dx} &= v_1 - 2 & v_2(0) &= 5 \\ \frac{dv_3}{dx} &= x^2 - v_3 & v_3(0,4) &= 1 \end{aligned}$

c) $\begin{aligned} \frac{dy}{dx} + 4 \frac{dz}{dx} &= 3xy - z & y(1) &= 3 \\ \frac{dz}{dx} - \frac{dy}{dx} &= \tan x & z(1) &= 4 \end{aligned}$

Solución:

- a) Se trata de un problema de Cauchy de orden 2. Las variables dependientes son x y y mientras la variable independiente es t . El valor inicial de la variable independiente es $t_0 = 3$.
- b) No es un problema de Cauchy porque no satisface la condición de que los valores conocidos de las tres variables dependientes ocurran para el mismo valor inicial de la variable independiente. Si se pudiera conocer el valor de $v_3(0)$ se tendría un problema de Cauchy de orden 3.
- c) No es un problema de Cauchy, aunque es un sistema de ecuaciones diferenciales de primer orden con condiciones iniciales. Las ecuaciones no tienen la forma requerida ya que en ellas aparece más de una derivada. Sin embargo, el sistema se puede convertir en un problema de Cauchy de la siguiente forma:

Sumando miembro a miembro ambas ecuaciones, se obtiene:

$$5 \frac{dz}{dx} = 3xy - z + \tan x$$

Es decir:

$$\frac{dz}{dx} = \frac{3xy - z + \tan x}{5} \quad (1)$$

Restando a la primera ecuación, 4 veces la segunda, resulta:

$$5 \frac{dy}{dx} = 3xy - z - 4 \tan x$$

o, lo que es igual:

$$\frac{dy}{dx} = \frac{3xy - z - 4 \tan x}{5} \quad (2)$$

El sistema formado por (1) y (2) con las condiciones iniciales dadas:

$$\begin{aligned}\frac{dy}{dx} &= \frac{3xy - z - 4\tan x}{5} \\ \frac{dz}{dx} &= \frac{3xy - z + \tan x}{5} \\ y(1) &= 3 \\ z(1) &= 4\end{aligned}$$

es un problema de Cauchy de orden 2, equivalente al sistema original.

Transformación de una ecuación de orden m en un problema de Cauchy

Casi todas las ecuaciones diferenciales de orden m con condiciones iniciales pueden ser transformadas en problemas de Cauchy de orden m . Todo lo que se requiere es que la ecuación diferencial pueda ser expresada en la forma que sigue:

$$\begin{aligned}\frac{d^m y}{dx^m} &= G\left(x, y, \frac{dy}{dx}, \frac{d^2y}{dx^2}, \dots, \frac{d^{m-1}y}{dx^{m-1}}\right) \\ y(x_0) &= c_1 \\ y'(x_0) &= c_2 \\ &\vdots \\ y^{(m-1)}(x_0) &= c_m\end{aligned}\tag{3}$$

lo cual se puede resumir en dos características:

- La ecuación posee condiciones iniciales.
- La derivada de mayor orden (m) puede ser despejada.

Para transformar una ecuación del tipo (3) en un problema de Cauchy de orden m se introducen m nuevas variables u_1, u_2, \dots, u_m definidas como sigue:

$$\begin{aligned}u_1 &= y \\ u_2 &= \frac{dy}{dx} \\ u_3 &= \frac{d^2y}{dx^2} \\ &\vdots \\ u_m &= \frac{d^{m-1}y}{dx^{m-1}}\end{aligned}\tag{4}$$

Derivando respecto a x en ambos miembros de la primera ecuación y teniendo en cuenta la segunda, se tiene:

$$\frac{du_1}{dx} = u_2\tag{5}$$

Derivando respecto a x en ambos miembros de la segunda ecuación de (4) y teniendo en cuenta la tercera, se tiene:

$$\frac{du_2}{dx} = u_3 \quad (6)$$

Este procedimiento se repite hasta derivar en ambos miembros la penúltima de las ecuaciones (4) tomando en cuenta la última, para obtener:

$$\frac{du_{m-1}}{dx} = u_m \quad (7)$$

Ahora, se deriva en ambos miembros la última de las ecuaciones (4) y resulta:

$$\frac{du_m}{dx} = \frac{d^m y}{dx^m}$$

y, utilizando la ecuación diferencial (3), se puede escribir:

$$\frac{du_m}{dx} = G\left(x, y, \frac{dy}{dx}, \frac{d^2y}{dx^2}, \dots, \frac{d^{m-1}y}{dx^{m-1}}\right)$$

La función G puede expresarse en términos de las nuevas variables u_1, u_2, \dots, u_m definidas en (4) y se tiene:

$$\frac{du_m}{dx} = G(x, u_1, u_2, u_3, \dots, u_m) \quad (8)$$

Las ecuaciones (5), (6), ..., (7), (8) forman el sistema:

$$\begin{aligned} \frac{du_1}{dx} &= u_2 \\ \frac{du_2}{dx} &= u_3 \\ &\vdots \\ \frac{du_{m-1}}{dx} &= u_m \\ \frac{du_m}{dx} &= G(x, u_1, u_2, \dots, u_m) \end{aligned} \quad (9)$$

Por su parte, las condiciones iniciales del problema original, dan lugar, teniendo presentes de nuevo las fórmulas (4), en condiciones iniciales para las variables u_1, u_2, \dots, u_m :

$$\begin{aligned} u_1(x_0) &= c_1 \\ u_2(x_0) &= c_2 \\ &\vdots \\ u_m(x_0) &= c_m \end{aligned} \quad (10)$$

Es evidente que el sistema (9) con las condiciones iniciales (10) forman un problema de Cauchy de orden m .

Ejemplo 2

Transforme la siguiente ecuación diferencial con condiciones iniciales en un problema de Cauchy.

$$\frac{d^4y}{dx^4} + e^x \frac{d^2y}{dx^2} \frac{dy}{dx} - x \left(\frac{d^3y}{dx^3} \right)^2 = \cos y - 4xy$$

$$y(2) = 1,5$$

$$y'(2) = 3,1$$

$$y''(2) = -0,8$$

$$y'''(2) = 3,7$$

Solución:

A pesar de que se trata de una ecuación diferencial no lineal de gran complejidad, cuya solución analítica resulta imposible, es fácil expresarla en la forma requerida, pues la cuarta derivada puede ser despejada sin dificultad:

$$\frac{d^4y}{dx^4} = \cos y - 4xy - e^x \frac{d^2y}{dx^2} \frac{dy}{dx} + x \left(\frac{d^3y}{dx^3} \right)^2$$

Las nuevas variables u_1, u_2, u_3, u_4 se definen como:

$$u_1 = y$$

$$u_2 = y'$$

$$u_3 = y''$$

$$u_4 = y'''$$

Derivando en ambos miembros de cada una de estas ecuaciones y tomando en cuenta la ecuación siguiente y la propia ecuación diferencial, se obtiene el nuevo sistema:

$$\begin{aligned} \frac{du_1}{dx} &= u_2 \\ \frac{du_2}{dx} &= u_3 \\ \frac{du_3}{dx} &= u_4 \\ \frac{du_4}{dx} &= \cos u_1 - 4xu_1 - e^x u_2 u_3 + x(u_4)^2 \end{aligned} \tag{11}$$

con las condiciones iniciales:

$$\begin{aligned} u_1(2) &= 1,5 \\ u_2(2) &= 3,1 \\ u_3(2) &= -0,8 \\ u_4(2) &= 3,7 \end{aligned} \tag{12}$$

El sistema (11) con las condiciones iniciales (12) constituyen un problema de Cauchy de cuarto orden. ■

No solo las ecuaciones diferenciales de orden superior con condiciones iniciales pueden ser transformadas en problemas de Cauchy, en muchos casos también puede hacerse con sistemas de ecuaciones diferenciales de orden superior con condiciones iniciales, como se indica en el siguiente ejemplo.

Ejemplo 3

Transforme en un problema de Cauchy el siguiente sistema de ecuaciones diferenciales con condiciones iniciales:

$$\begin{aligned} y'''(t) + z'(t)y''(t) &= z(t)y(t) + t^3 y'(t) \\ z''(t)y''(t) &= y'(t)z'(t) + y(t)\ln t \end{aligned}$$

$$y(1,5) = 3$$

$$y'(1,5) = 2$$

$$y''(1,5) = -1$$

$$z(1,5) = 0$$

$$z'(1,5) = 1$$

Solución:

Despejando $y'''(t)$ en la primera ecuación y $z''(t)$ en la segunda, el sistema queda:

$$y'''(t) = z(t)y(t) + t^3 y'(t) - z'(t)y''(t)$$

$$z''(t) = \frac{y'(t)z'(t) + y(t)\ln t}{y''(t)}$$

Sean u_1, u_2, u_3, u_4, u_5 , definidas como:

$$u_1 = y$$

$$u_2 = y'$$

$$u_3 = y''$$

$$u_4 = z$$

$$u_5 = z'$$

Ahora se deriva en ambos miembros de cada ecuación y, teniendo en cuenta o bien la ecuación siguiente o una de las ecuaciones del sistema, resulta:

$$\begin{aligned}
\frac{du_1}{dt} &= u_2 \\
\frac{du_2}{dt} &= u_3 \\
\frac{du_3}{dt} &= u_1 u_4 + t^3 u_2 - u_3 u_5 \\
\frac{du_4}{dt} &= u_5 \\
\frac{du_5}{dt} &= \frac{u_2 u_5 + u_1 \ln t}{u_3}
\end{aligned}$$

Con las condiciones iniciales:

$$\begin{aligned}
u_1(1,5) &= 3 \\
u_2(1,5) &= 2 \\
u_3(1,5) &= -1 \\
u_4(1,5) &= 0 \\
u_5(1,5) &= 1
\end{aligned}$$

que es un problema de Cauchy de orden cinco.

Solución numérica de un problema de Cauchy

Cualquiera de los métodos estudiados para resolver ecuaciones diferenciales de primer orden puede ser adaptado para resolver problemas de Cauchy de orden m , solamente hay que realizar el siguiente cambio. Si el método que se va a emplear consta de varias etapas en cada paso (por ejemplo, RK2 consta de tres etapas en cada paso: calcular K_1 , calcular K_2 y hallar y_{n+1}) entonces en cada paso de la solución del problema de Cauchy, se aplica la etapa 1 a las m ecuaciones, después, la etapa 2 a todas las ecuaciones, etc. hasta aplicar la última etapa a todas las ecuaciones y con esto queda terminado un paso.

Debido a que poseen una estructura más simple, aquí solamente serán empleados los métodos de Runge – Kutta de orden 2 y 4 para la solución de problemas de Cauchy de orden m , que son los más frecuentemente usados.

El método RK2 para un problema de Cauchy de orden m

Sea el problema de Cauchy:

$$\begin{aligned}
\frac{du_1}{dx} &= f_1(x, u_1, u_2, \dots, u_m) & u_1(x_0) &= u_{10} \\
\frac{du_2}{dx} &= f_2(x, u_1, u_2, \dots, u_m) & u_2(x_0) &= u_{20} \\
&\vdots &&\vdots \\
\frac{du_m}{dx} &= f_m(x, u_1, u_2, \dots, u_m) & u_m(x_0) &= u_{m0}
\end{aligned}$$

Suponga que las variables u_1, u_2, \dots, u_m , han sido calculadas hasta el paso número n , es decir que se conocen los valores $u_{1n}, u_{2n}, \dots, u_{mn}$ y se trata de determinar $u_{1,n+1}, u_{2,n+1}, \dots, u_{m,n+1}$ mediante un paso del método RK2. Este paso consta de tres etapas:

Etapa 1: Calcular las K_1 :

$$\begin{aligned} K_{11} &= hf_1(x_n, u_{1n}, u_{2n}, \dots, u_{mn}) \\ K_{12} &= hf_2(x_n, u_{1n}, u_{2n}, \dots, u_{mn}) \\ &\vdots \\ K_{1m} &= hf_m(x_n, u_{1n}, u_{2n}, \dots, u_{mn}) \end{aligned}$$

Etapa 2: Calcular las K_2 :

$$\begin{aligned} K_{21} &= hf_1(x_n + h, u_{1n} + K_{11}, u_{2n} + K_{12}, \dots, u_{mn} + K_{1m}) \\ K_{22} &= hf_2(x_n + h, u_{1n} + K_{11}, u_{2n} + K_{12}, \dots, u_{mn} + K_{1m}) \\ &\vdots \\ K_{2m} &= hf_m(x_n + h, u_{1n} + K_{11}, u_{2n} + K_{12}, \dots, u_{mn} + K_{1m}) \end{aligned}$$

Etapa 3: Hallar la solución en x_{n+1} :

$$\begin{aligned} u_{1,n+1} &= u_{1n} + \frac{1}{2}(K_{11} + K_{21}) \\ u_{2,n+1} &= u_{2n} + \frac{1}{2}(K_{12} + K_{22}) \\ &\vdots \\ u_{m,n+1} &= u_{mn} + \frac{1}{2}(K_{1m} + K_{2m}) \end{aligned}$$

Ejemplo 4

Dado el problema de Cauchy:

$$\begin{aligned} \frac{du_1}{dx} &= x + u_2 - u_1 u_3 & u_1(1,4) &= 3 \\ \frac{du_2}{dx} &= x u_3 & u_2(1,4) &= 1,7 \\ \frac{du_3}{dx} &= u_1 + u_2 - u_3 & u_3(1,4) &= 2,5 \end{aligned}$$

Calcule dos pasos de la solución mediante RK2 tomando $h = 0,2$.

Solución:

$$x_0 = 1,4; \quad u_{10} = 3; \quad u_{20} = 1,7; \quad u_{30} = 2,5$$

Paso 1

Etapa 1: Calcular las K_1 :

$$\begin{aligned} K_{11} &= hf_1(x_0, u_{10}, u_{20}, u_{30}) = hf_1(1,4; 3; 1,7; 2,5) = 0,2(1,4 + 1,7 - (3)(2,5)) = -0,88 \\ K_{12} &= hf_2(x_0, u_{10}, u_{20}, u_{30}) = hf_2(1,4; 3; 1,7; 2,5) = 0,2(1,4)(2,5) = 0,7 \\ K_{13} &= hf_m(x_0, u_{10}, u_{20}, u_{30}) = hf_3(1,4; 3; 1,7; 2,5) = 0,2(3 + 1,7 - 2,5) = 0,44 \end{aligned}$$

Etapa 2: Calcular las K_2 :

$$K_{21} = hf_1(x_0 + h, u_{10} + K_{11}, u_{20} + K_{12}, u_{30} + K_{13}) = hf_1(1,6; 2,12; 2,4; 2,94) = \\ = 0,2(1,6 + 2,4 - (2,12)(2,94)) = -0,4466$$

$$K_{22} = hf_2(x_0 + h, u_{10} + K_{11}, u_{20} + K_{12}, u_{30} + K_{13}) = hf_2(1,6; 2,12; 2,4; 2,94) = \\ = 0,2(1,6)(2,94) = 0,9408$$

$$K_{23} = hf_3(x_0 + h, u_{10} + K_{11}, u_{20} + K_{12}, u_{30} + K_{13}) = hf_3(1,6; 2,12; 2,4; 2,94) = \\ = 0,2(2,12 + 2,4 - 2,94) = 0,316$$

Etapa 3: Hallar la solución en x_1 :

$$u_{11} = u_{10} + \frac{1}{2}(K_{11} + K_{21}) = 3 + \frac{1}{2}(-0,88 - 0,4466) = 2,3367 \\ u_{21} = u_{20} + \frac{1}{2}(K_{12} + K_{22}) = 1,7 + \frac{1}{2}(0,7 + 0,9408) = 2,5204 \\ u_{31} = u_{30} + \frac{1}{2}(K_{13} + K_{23}) = 2,5 + \frac{1}{2}(0,44 + 0,316) = 2,8780$$

$$x_1 = 1,6; \quad u_{11} = 2,3367; \quad u_{21} = 2,5204; \quad u_{31} = 2,8780$$

Paso 2

Etapa 1: Calcular las K_1 :

$$K_{11} = hf_1(x_1, u_{11}, u_{21}, u_{31}) = hf_1(1,6; 2,3367; 2,5204; 2,878) = \\ = 0,2(1,6 + 2,5204 - (2,3367)(2,878)) = -0,5209$$

$$K_{12} = hf_2(x_1, u_{11}, u_{21}, u_{31}) = hf_2(1,6; 2,3367; 2,5204; 2,878) = \\ = 0,2(1,6)(2,878) = 0,9210$$

$$K_{13} = hf_m(x_1, u_{11}, u_{21}, u_{31}) = hf_3(1,6; 2,3367; 2,5204; 2,878) = \\ = 0,2(2,3367 + 2,5204 - 2,878) = 0,3958$$

Etapa 2: Calcular las K_2 :

$$K_{21} = hf_1(x_1 + h, u_{11} + K_{11}, u_{21} + K_{12}, u_{31} + K_{13}) = hf_1(1,8; 1,8158; 3,4414; 3,2738) = \\ = 0,2(1,8 + 3,4414 - (1,8158)(3,2738)) = -0,1406$$

$$K_{22} = hf_2(x_1 + h, u_{11} + K_{11}, u_{21} + K_{12}, u_{31} + K_{13}) = hf_2(1,8; 1,8158; 3,4414; 3,2738) = \\ = 0,2(1,8)(3,2738) = 1,1786$$

$$\begin{aligned}
K_{23} &= hf_3(x_1 + h, u_{11} + K_{11}, u_{21} + K_{12}, u_{31} + K_{13}) = hf_3(1,8; 1,8158; 3,4414; 3,2738) = \\
&= 0,2(1,8158 + 3,4414 - 3,2738) = 0,3967
\end{aligned}$$

Etapa 3: Hallar la solución en x_2 :

$$\begin{aligned}
u_{12} &= u_{11} + \frac{1}{2}(K_{11} + K_{21}) = 2,3367 + \frac{1}{2}(-0,5209 - 0,1406) = 2,0060 \\
u_{22} &= u_{21} + \frac{1}{2}(K_{12} + K_{22}) = 2,5204 + \frac{1}{2}(0,9210 + 1,1786) = 3,5702 \\
u_{32} &= u_{31} + \frac{1}{2}(K_{13} + K_{23}) = 2,878 + \frac{1}{2}(0,3958 + 0,3967) = 3,2742
\end{aligned}$$

$$x_2 = 1,8; \quad u_{12} = 2,0060; \quad u_{22} = 3,5702; \quad u_{32} = 3,2742$$

La solución manual de este ejercicio solo tiene un propósito didáctico pues, como se aprecia, los cálculos son voluminosos pero el algoritmo es sencillo, condiciones ideales para utilizar un programa de computadora.

Algoritmo en seudo código

El algoritmo que sigue obtiene la solución aproximada del problema de Cauchy

$$\begin{array}{ll}
\frac{du_1}{dx} = f_1(x, u_1, u_2, \dots, u_m) & u_1(x_0) = u_{10} \\
\frac{du_2}{dx} = f_2(x, u_1, u_2, \dots, u_m) & u_2(x_0) = u_{20} \\
\vdots & \vdots \\
\frac{du_m}{dx} = f_m(x, u_1, u_2, \dots, u_m) & u_m(x_0) = u_{m0}
\end{array}$$

en un intervalo $x_0 \leq x \leq x_F$ con paso $h > 0$ mediante el método RK2. Se suponen conocidos: las funciones $f_1(x, u_1, u_2, \dots, u_m), f_2(x, u_1, u_2, \dots, u_m), \dots, f_m(x, u_1, u_2, \dots, u_m)$, las condiciones iniciales $(x_0, u_{10}, u_{20}, \dots, u_{m0})$ el valor final x_F y el paso h .

```

n := 0
do while  $x_n < x_F$ 
    ( $x_n, u_{1n}, u_{2n}, \dots, u_{mn}$ ) es un punto de la solución aproximada
    for  $i = 1$  to  $m$ 
         $K_{1i} := hf_i(x_n, u_{1n}, u_{2n}, \dots, u_{mn})$ 
    end
    for  $i = 1$  to  $m$ 
         $K_{2i} := hf_i(x_n + h, u_{1n} + K_{11}, u_{2n} + K_{12}, \dots, u_{mn} + K_{1m})$ 
    end
    for  $i = 1$  to  $m$ 
         $u_{i,n+1} := u_{i,n} + \frac{1}{2}(K_{1i} + K_{2i})$ 
    end
     $x_{n+1} := x_n + h$ 
     $n := n + 1$ 
end
Terminar

```

Estimación del error por doble cálculo

La expresión obtenida para hallar el error por doble cálculo en el caso de una ecuación diferencial de primer orden

$$e_h \approx \frac{y_h - y_{2h}}{2^p - 1}$$

puede ser extendida al problema de Cauchy de orden m . Los detalles de la deducción serán omitidos. Sea \mathbf{u}_h el vector de soluciones obtenidos para un cierto valor x utilizando paso h con un método de orden p (en el caso de RK2, $p = 2$) y sea \mathbf{u}_{2h} el vector de soluciones correspondiente al mismo valor x pero calculado con paso $2h$, entonces, si \mathbf{e}_h representa al vector de errores correspondiente a \mathbf{u}_h , puede probarse que

$$\|\mathbf{e}_h\| \approx \frac{\|\mathbf{u}_h - \mathbf{u}_{2h}\|}{2^p - 1} \quad (13)$$

donde, como se recordará de la sección 3.1, la notación $\|\mathbf{x}\|$ significa la norma del vector \mathbf{x} , es decir, el mayor valor absoluto de las componentes del vector \mathbf{x} .

Observe que la fórmula (13) establece una manera de estimar la mayor componente del vector de errores, esto significa que se obtiene realmente una cota superior para los errores de las variables u_1, u_2, \dots, u_m .

Para el método RK2, como $p = 3$, la fórmula (13) resulta:

$$\|\mathbf{e}_h\| \approx \frac{\|\mathbf{u}_h - \mathbf{u}_{2h}\|}{3} \quad (14)$$

Ejemplo 5

Resuelva el problema de Cauchy

$$\begin{aligned} \frac{du_1}{dx} &= x + u_2 - u_1 u_3 & u_1(1,4) &= 3 \\ \frac{du_2}{dx} &= x u_3 & u_2(1,4) &= 1,7 \\ \frac{du_3}{dx} &= u_1 + u_2 - u_3 & u_3(1,4) &= 2,5 \end{aligned}$$

En el intervalo $1,4 \leq x \leq 3$ con pasos $h = 0,1$ y $h = 0,2$. Muestre las soluciones obtenidas y una cota del error para $x = 1,8; 2,0; 2,2; 2,4; 2,6; 2,8$ y $3,0$.

Solución:

En la tabla 1 se muestran los resultados pedidos. Aparecen marcados con * los valores de las variables que sirvieron para calcular la cota del error debido a que con ellos se obtiene la máxima diferencia entre las soluciones halladas.

x	$h = 0,1$			$h = 0,2$			Cota del error
	u_1	u_2	u_3	u_1	u_2	u_3	
1,4	3,0000	1,7000	2,5000	3,0000	1,7000	2,5000	0,0000
1,6	2,3114*	2,5141	2,8959	2,3367*	2,5204	2,8780	0,0084
1,8	1,9608*	3,5671	3,2985	2,0059*	3,5702	3,2742	0,0150
2,0	1,8524*	4,9132	3,8073	1,9014*	4,9065	3,7825	0,0163
2,2	1,8859*	6,6498	4,4981	1,9256*	6,6282	4,4726	0,0132
2,4	1,9819	8,9240*	5,4392	2,0055	8,8815*	5,4084	0,0142
2,6	2,0889	11,9441*	6,7064	2,0958	11,8706*	6,6628	0,0245
2,8	2,1830	15,9991*	8,4002	2,1751	15,8775*	8,3341	0,0405
3,0	2,2597	21,4922*	10,6629	2,2360	21,2944*	10,5611	0,0659

Tabla 1

El método RK4 para un problema de Cauchy de orden m

Se trata de nuevo del problema de Cauchy:

$$\begin{aligned} \frac{du_1}{dx} &= f_1(x, u_1, u_2, \dots, u_m) & u_1(x_0) &= u_{10} \\ \frac{du_2}{dx} &= f_2(x, u_1, u_2, \dots, u_m) & u_2(x_0) &= u_{20} \\ &\vdots & &\vdots \\ \frac{du_m}{dx} &= f_m(x, u_1, u_2, \dots, u_m) & u_m(x_0) &= u_{m0} \end{aligned}$$

Se supone que las variables u_1, u_2, \dots, u_m , han sido calculadas hasta el paso número n , es decir que se conocen los valores $u_{1n}, u_{2n}, \dots, u_{mn}$ y se trata de determinar $u_{1,n+1}, u_{2,n+1}, \dots, u_{m,n+1}$ mediante un paso del método RK4. Este paso consta de cinco etapas:

Etapa 1: Calcular las K_1 :

$$\begin{aligned} K_{11} &= hf_1(x_n, u_{1n}, u_{2n}, \dots, u_{mn}) \\ K_{12} &= hf_2(x_n, u_{1n}, u_{2n}, \dots, u_{mn}) \\ &\vdots \\ K_{1m} &= hf_m(x_n, u_{1n}, u_{2n}, \dots, u_{mn}) \end{aligned}$$

Etapa 2: Calcular las K_2 :

$$\begin{aligned} K_{21} &= hf_1(x_n + \frac{1}{2}h, u_{1n} + \frac{1}{2}K_{11}, u_{2n} + \frac{1}{2}K_{12}, \dots, u_{mn} + \frac{1}{2}K_{1m}) \\ K_{22} &= hf_2(x_n + \frac{1}{2}h, u_{1n} + \frac{1}{2}K_{11}, u_{2n} + \frac{1}{2}K_{12}, \dots, u_{mn} + \frac{1}{2}K_{1m}) \\ &\vdots \\ K_{2m} &= hf_m(x_n + \frac{1}{2}h, u_{1n} + \frac{1}{2}K_{11}, u_{2n} + \frac{1}{2}K_{12}, \dots, u_{mn} + \frac{1}{2}K_{1m}) \end{aligned}$$

Etapa 3: Calcular las K_3 :

$$\begin{aligned} K_{31} &= hf_1(x_n + \frac{1}{2}h, u_{1n} + \frac{1}{2}K_{21}, u_{2n} + \frac{1}{2}K_{22}, \dots, u_{mn} + \frac{1}{2}K_{2m}) \\ K_{32} &= hf_2(x_n + \frac{1}{2}h, u_{1n} + \frac{1}{2}K_{21}, u_{2n} + \frac{1}{2}K_{22}, \dots, u_{mn} + \frac{1}{2}K_{2m}) \\ &\vdots \\ K_{3m} &= hf_m(x_n + \frac{1}{2}h, u_{1n} + \frac{1}{2}K_{21}, u_{2n} + \frac{1}{2}K_{22}, \dots, u_{mn} + \frac{1}{2}K_{2m}) \end{aligned}$$

Etapa 4: Calcular las K_4 :

$$\begin{aligned} K_{41} &= hf_1(x_n + h, u_{1n} + K_{31}, u_{2n} + K_{32}, \dots, u_{mn} + K_{3m}) \\ K_{42} &= hf_2(x_n + h, u_{1n} + K_{31}, u_{2n} + K_{32}, \dots, u_{mn} + K_{3m}) \\ &\vdots \\ K_{4m} &= hf_m(x_n + h, u_{1n} + K_{31}, u_{2n} + K_{32}, \dots, u_{mn} + K_{3m}) \end{aligned}$$

Etapa 5: Hallar la solución en x_{n+1} :

$$\begin{aligned} u_{1,n+1} &= u_{1n} + \frac{1}{6}(K_{11} + 2K_{21} + 2K_{31} + K_{41}) \\ u_{2,n+1} &= u_{2n} + \frac{1}{6}(K_{12} + 2K_{22} + 2K_{32} + K_{42}) \\ &\vdots \\ u_{m,n+1} &= u_{mn} + \frac{1}{6}(K_{1m} + 2K_{2m} + 2K_{3m} + K_{4m}) \end{aligned}$$

Ejemplo 6

Dado el problema de Cauchy:

$$\begin{aligned} \frac{du_1}{dx} &= x + u_2 - u_1 u_3 & u_1(1,4) &= 3 \\ \frac{du_2}{dx} &= x u_3 & u_2(1,4) &= 1,7 \\ \frac{du_3}{dx} &= u_1 + u_2 - u_3 & u_3(1,4) &= 2,5 \end{aligned}$$

Calcule un paso de la solución mediante RK4 tomando $h = 0,2$.

Solución:

$$x_0 = 1,4; \quad u_{10} = 3; \quad u_{20} = 1,7; \quad u_{30} = 2,5$$

Etapa 1: Calcular las K_1 :

$$\begin{aligned} K_{11} &= hf_1(x_0, u_{10}, u_{20}, u_{30}) = hf_1(1,4; 3; 1,7; 2,5) = 0,2(1,4 + 1,7 - (3)(2,5)) = -0,88 \\ K_{12} &= hf_2(x_0, u_{10}, u_{20}, u_{30}) = hf_2(1,4; 3; 1,7; 2,5) = 0,2(1,4)(2,5) = 0,7 \\ K_{13} &= hf_m(x_0, u_{10}, u_{20}, u_{30}) = hf_3(1,4; 3; 1,7; 2,5) = 0,2(3 + 1,7 - 2,5) = 0,44 \end{aligned}$$

Etapa 2: Calcular las K_2 :

$$K_{21} = hf_1(x_0 + \frac{1}{2}h, u_{10} + \frac{1}{2}K_{11}, u_{20} + \frac{1}{2}K_{12}, u_{30} + \frac{1}{2}K_{13}) = hf_1(1,5; 2,56; 2,05; 2,72) =$$

$$= 0,2(1,5 + 2,05 - (2,56)(2,72)) = -0,6826$$

$$\begin{aligned} K_{22} &= hf_2(x_0 + \frac{1}{2}h, u_{10} + \frac{1}{2}K_{11}, u_{20} + \frac{1}{2}K_{12}, u_{30} + \frac{1}{2}K_{13}) = hf_2(1,5; 2,56; 2,05; 2,72) = \\ &= 0,2(1,5)(2,72) = 0,816 \end{aligned}$$

$$\begin{aligned} K_{23} &= hf_3(x_0 + \frac{1}{2}h, u_{10} + \frac{1}{2}K_{11}, u_{20} + \frac{1}{2}K_{12}, u_{30} + \frac{1}{2}K_{13}) = hf_3(1,5; 2,56; 2,05; 2,72) = \\ &= 0,2(2,56 + 2,05 - 2,72) = 0,378 \end{aligned}$$

Etapa 3: Calcular las K_3 :

$$\begin{aligned} K_{31} &= hf_1(x_0 + \frac{1}{2}h, u_{10} + \frac{1}{2}K_{21}, u_{20} + \frac{1}{2}K_{22}, u_{30} + \frac{1}{2}K_{23}) = hf_1(1,5; 2,6587; 2,108; 2,689) = \\ &= 0,2(1,5 + 2,108 - (2,6587)(2,689)) = -0,7082 \\ K_{32} &= hf_2(x_0 + \frac{1}{2}h, u_{10} + \frac{1}{2}K_{21}, u_{20} + \frac{1}{2}K_{22}, u_{30} + \frac{1}{2}K_{23}) = hf_2(1,5; 2,6587; 2,108; 2,689) = \\ &= 0,2(1,5)(2,689) = 0,8067 \end{aligned}$$

$$\begin{aligned} K_{33} &= hf_3(x_0 + \frac{1}{2}h, u_{10} + \frac{1}{2}K_{21}, u_{20} + \frac{1}{2}K_{22}, u_{30} + \frac{1}{2}K_{23}) = hf_3(1,5; 2,6587; 2,108; 2,689) = \\ &= 0,2(2,6587 + 2,108 - 2,689) = 0,4155 \end{aligned}$$

Etapa 4: Calcular las K_4 :

$$\begin{aligned} K_{41} &= hf_1(x_0 + h, u_{10} + K_{31}, u_{20} + K_{32}, u_{30} + K_{33}) = hf_1(1,6; 2,2918; 2,5067; 2,9155) = \\ &= 0,2(1,6 + 2,5067 - (2,2918)(2,9155)) = -0,5150 \end{aligned}$$

$$\begin{aligned} K_{42} &= hf_2(x_0 + h, u_{10} + K_{31}, u_{20} + K_{32}, u_{30} + K_{33}) = hf_2(1,6; 2,2918; 2,5067; 2,9155) = \\ &= 0,2(1,6)(2,9155) = 0,9333 \end{aligned}$$

$$\begin{aligned} K_{43} &= hf_3(x_0 + h, u_{10} + K_{31}, u_{20} + K_{32}, u_{30} + K_{33}) = hf_3(1,6; 2,2918; 2,5067; 2,9155) = \\ &= 0,2(2,2918 + 2,5067 - 2,9155) = 0,3766 \end{aligned}$$

Etapa 5: Hallar la solución en x_1 :

$$\begin{aligned} u_{11} &= u_{10} + \frac{1}{6}(K_{11} + 2K_{21} + 2K_{31} + K_{41}) = 2,3039 \\ u_{21} &= u_{20} + \frac{1}{6}(K_{12} + 2K_{22} + 2K_{32} + K_{42}) = 2,5131 \\ u_{31} &= u_{30} + \frac{1}{6}(K_{13} + 2K_{23} + 2K_{33} + K_{43}) = 2,9006 \end{aligned}$$

$$x_1 = 1,8; \quad u_{11} = 2,3039; \quad u_{21} = 2,5131; \quad u_{31} = 2,9006$$

Algoritmo en seudo código

El algoritmo que sigue obtiene la solución aproximada del problema de Cauchy

$$\begin{aligned}\frac{du_1}{dx} &= f_1(x, u_1, u_2, \dots, u_m) & u_1(x_0) &= u_{10} \\ \frac{du_2}{dx} &= f_2(x, u_1, u_2, \dots, u_m) & u_2(x_0) &= u_{20} \\ &\vdots && \vdots \\ \frac{du_m}{dx} &= f_m(x, u_1, u_2, \dots, u_m) & u_m(x_0) &= u_{m0}\end{aligned}$$

en un intervalo $x_0 \leq x \leq x_F$ con paso $h > 0$ mediante el método RK4. Se suponen conocidos: las funciones $f_1(x, u_1, u_2, \dots, u_m), f_2(x, u_1, u_2, \dots, u_m), \dots, f_m(x, u_1, u_2, \dots, u_m)$, las condiciones iniciales $(x_0, u_{10}, u_{20}, \dots, u_{m0})$ el valor final x_F y el paso h .

```

n := 0
do while  $x_n < x_F$ 
    ( $x_n, u_{1n}, u_{2n}, \dots, u_{mn}$ ) es un punto de la solución aproximada
    for  $i = 1$  to  $m$ 
         $K_{1i} := hf_i(x_n, u_{1n}, u_{2n}, \dots, u_{mn})$ 
    end
    for  $i = 1$  to  $m$ 
         $K_{2i} := hf_i(x_n + \frac{1}{2}h, u_{1n} + \frac{1}{2}K_{11}, u_{2n} + \frac{1}{2}K_{12}, \dots, u_{mn} + \frac{1}{2}K_{1m})$ 
    end
    for  $i = 1$  to  $m$ 
         $K_{3i} := hf_i(x_n + \frac{1}{2}h, u_{1n} + \frac{1}{2}K_{21}, u_{2n} + \frac{1}{2}K_{22}, \dots, u_{mn} + \frac{1}{2}K_{2m})$ 
    end
    for  $i = 1$  to  $m$ 
         $K_{4i} := hf_i(x_n + h, u_{1n} + K_{31}, u_{2n} + K_{32}, \dots, u_{mn} + K_{3m})$ 
    end
    for  $i = 1$  to  $m$ 
         $u_{i,n+1} := u_{i,n} + \frac{1}{6}(K_{1i} + 2K_{2i} + 2K_{3i} + K_{4i})$ 
    end
     $x_{n+1} := x_n + h$ 
     $n := n + 1$ 
end
Terminar

```

Estimación del error por doble cálculo

Como RK4 es un método de orden 4 ($p = 4$), la fórmula (13) resulta en:

$$\|\mathbf{e}_h\| \approx \frac{\|\mathbf{u}_h - \mathbf{u}_{2h}\|}{15} \quad (15)$$

cuyos términos poseen el mismo significado que en la fórmula (14).

Ejemplo 7

Resuelva el problema de Cauchy

$$\begin{aligned}\frac{du_1}{dx} &= x + u_2 - u_1 u_3 & u_1(1,4) &= 3 \\ \frac{du_2}{dx} &= x u_3 & u_2(1,4) &= 1,7 \\ \frac{du_3}{dx} &= u_1 + u_2 - u_3 & u_3(1,4) &= 2,5\end{aligned}$$

En el intervalo $1,4 \leq x \leq 3$ con pasos $h = 0,1$ y $h = 0,2$. Muestre las soluciones obtenidas y una cota del error para $x = 1,8; 2,0; 2,2; 2,4; 2,6; 2,8$ y $3,0$. Compare con los resultados obtenidos utilizando RK2 (Ejemplo 5).

Solución:

La tabla 2 muestra los resultados pedidos. Aparecen marcados con * los valores de las variables que sirvieron para calcular la cota del error debido a que con ellos se obtiene la máxima diferencia entre las soluciones halladas.

x	$h = 0,1$			$h = 0,2$			Cota del error
	u_1	u_2	u_3	u_1	u_2	u_3	
1,4	3,00000	1,70000	2,50000	3,00000	1,70000	2,50000	0,00000
1,6	2,30398	2,51286*	2,90072	2,30387	2,51306*	2,90061	0,00002
1,8	1,94825	3,56751*	3,30503	1,94848	3,56788*	3,30471	0,00003
2,0	1,83980*	4,91718	3,81401	1,84042*	4,91762	3,81360	0,00005
2,2	1,87693*	6,65901	4,50568	1,87771*	6,65943	4,50532	0,00006
2,4	1,97788*	8,94079	5,44932	1,97839*	8,94112	5,44914	0,00004
2,6	2,08906*	11,97229	6,72194	2,08872*	11,97251	6,72190	0,00003
2,8	2,18545*	16,04538	8,42471	2,18360*	16,04549	8,42480	0,00013
3,0	2,26299*	21,56742	10,70128	2,25883*	21,56736	10,70144	0,00028

Tabla 2

Si se comparan los errores de la tabla 2 con los obtenidos en la tabla 1, se verá que, ahora se han obtenido resultados mucho más exactos. Mientras en RK2 (vea la tabla 1) la mayor parte de los resultados solo tenían una cifra decimal exacta, en RK4 se obtuvo tres y hasta cuatro cifras decimales exactas. Por supuesto, debe tenerse en cuenta que el método RK4 requiere dos veces más cálculos que RK2.

Ejemplo 8

Según se estudió en la sección 7.1 (ejemplo 3) cuando un péndulo ideal de longitud L se mueve, lo hace de acuerdo con la ecuación diferencial:

$$-g \sin \theta = L \frac{d^2 \theta}{dt^2}$$

donde g es la aceleración producida por la fuerza de gravedad y θ es el ángulo de deflexión (en radianes) respecto a la posición de equilibrio (vea la figura 1). Un péndulo de $L = 0,5\text{ m}$ se desplaza 45° de la vertical y se suelta. Halle su posición al cabo de dos segundos con un error menor que $0,0001$ radian. Utilice el método RK4. Suponga $g = 9,8\text{ m/s}^2$. Compare gráficamente con la solución que se obtendría tomando la aproximación $\sin \theta \approx \theta$ que se hace frecuentemente para convertir la ecuación en lineal.

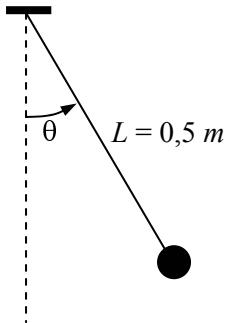


Figura 1

Solución:

Además de la ecuación diferencial, se requieren las condiciones iniciales del problema que, de acuerdo con el enunciado, son: posición inicial: $\theta = 45^\circ = \pi/4$ Rad. = $0,785398$ y velocidad angular inicial: cero. Por tanto, el problema a resolver es:

$$\begin{aligned} -9,8 \sin \theta &= 0,5 \frac{d^2\theta}{dt^2} \\ \theta(0) &= 0,785398 \\ \theta'(0) &= 0 \end{aligned}$$

y se desea determinar $\theta(2)$ con error menor que $0,0001$.

Primero, el problema se debe transformar en un problema de Cauchy de orden 2. Para ello, se introducen las nuevas variables u_1 y u_2 definidas como:

$$\begin{aligned} u_1 &= \theta \\ u_2 &= \theta' \end{aligned}$$

$$\begin{aligned} \text{de donde: } \frac{du_1}{dt} &= u_2 & u_1(0) &= 0,785398 \\ \frac{du_2}{dt} &= -19,6 \sin u_1 & u_2(0) &= 0 \end{aligned}$$

En la tabla 3 se muestran los valores de u_1 y u_2 para $t = 2$, obtenidos con diferentes pasos, así como los errores estimados mediante doble cálculo.

h	$u_1(2)$	$u_2(2)$	Cota del error
0,2	-0,461014	-2,612360	
0,1	-0,484048	-2,637867	0,001700
0,05	-0,485493	-2,636478	0,000097

Tabla 3

Como se ve, al cabo de dos segundos el péndulo se encuentra formando un ángulo de 0,48549 radianes con la vertical y del lado opuesto (negativo) al lado en que se soltó. Como la variable u_2 coincide con θ' (velocidad angular), en el instante $t = 2$ el péndulo se está moviendo hacia la izquierda (contrario al sentido positivo de referencia de la figura 1) con una velocidad de 2,63648 radianes por segundo. En la figura 2 se encuentran, en un mismo sistema de ejes, la gráfica de $u_1(t)$ en el intervalo $0 \leq t \leq 2$ (con puntos más pequeños), construida a partir de los resultados numéricos obtenidos y la gráfica de la solución $u_1(t)$ (con puntos mayores) si se hubiera hecho la aproximación usual de $\sin \theta \approx \theta$ en el modelo del péndulo. Obsérvese que la diferencia entre ambos resultados es muy significativa.

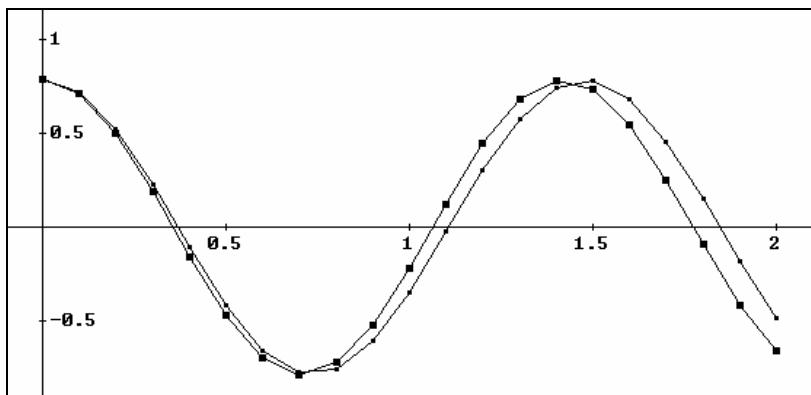


Figura 3

Ejercicios

1. Dado el siguiente problema de Cauchy:

$$\begin{aligned} \frac{dv}{dx} &= x - w & v(0) &= 0 \\ \frac{dw}{dx} &= v & w(0) &= 1 \end{aligned}$$

Halle $v(0,6)$ y $w(0,6)$ con paso $h = 0,1$ mediante RK4 y estime el error cometido.

2. Dado el sistema:

$$\begin{aligned}\frac{dy}{dt} &= x^2 + yt - 2 & x(1) &= 3 \\ \frac{dx}{dt} &= \sin x - xy & y(1) &= 2,5\end{aligned}$$

Halle $x(2)$ y $y(2)$ con cuatro cifras decimales exactas.

3. Dada la ecuación diferencial

$$\frac{d^2y}{dx^2} + y \frac{dy}{dx} - x^2 y = xe^x \quad y(0) = 1; \quad y'(0) = 1,45$$

Halle la solución en el intervalo $[0, 1]$ de modo que $y(1)$ tenga tres cifras decimales exactas.

4. Resuelva la siguiente ecuación diferencial en el intervalo $1 \leq x \leq 3$. Muestre la solución en 1,5; 2,0; 2,5 y 3,0. La solución debe poseer cuatro cifras decimales exactas.

$$\frac{d^3y}{dx^3} + y \frac{dy}{dx} = x \quad y(1) = 2; \quad y'(1) = 3; \quad y''(1) = 1$$

5. Halle la solución de la ecuación diferencial:

$$\frac{d^3y}{dx^3} + 3xy \frac{dy}{dx} - \frac{d^2y}{dx^2} = \sin x \quad y(0) = 1; \quad y'(0) = 0; \quad y''(0) = 2$$

en el intervalo $[0, 2]$ con cuatro cifras decimales exactas.

6. Dado el sistema:

$$\begin{aligned}\frac{d^2y}{dx^2} + z \frac{dy}{dx} - \cos y &= e^{-z} & y(1,2) &= 1 \\ \frac{dz}{dx} - x \frac{dy}{dx} &= xyz & y'(1,2) &= 2 \\ && z(1,2) &= 0,5\end{aligned}$$

Halle la solución con tres cifras decimales exactas en el intervalo $[1,2; 2,1]$.

7. En el ejemplo 5 de la sección 7.1 se llegó al siguiente modelo matemático para el movimiento de un planeta alrededor del sol.

$$\begin{cases} -\frac{km_s x}{(x^2 + y^2)^{3/2}} = \frac{d^2x}{dt^2} & x(0) = r_0 \\ -\frac{km_s y}{(x^2 + y^2)^{3/2}} = \frac{d^2y}{dt^2} & y(0) = 0 \\ & v_x(0) = v_{0x} \\ & v_y(0) = v_{0y} \end{cases}$$

donde m_s es la masa del sol, k es la constante de gravitación universal y v_x y v_y son las componentes de la velocidad del planeta en el instante inicial. Transforme este sistema en un problema de Cauchy.

8. En el ejemplo 4 de la sección 7.1, se mostró el modelo de Lotka – Volterra para dos poblaciones $x(t)$ de presas y $y(t)$ de depredadores:

$$\begin{cases} \frac{dx}{dt} = ax - bxy \\ \frac{dy}{dt} = -cy + dxy \end{cases}$$

Suponga que en una isla hay un área protegida donde habitan lobos y alces. Los alces se alimentan del bosque y los lobos se alimentan de los alces. A partir de datos históricos se han determinado las constantes del modelo para esta situación específica: $a = 0,3$; $b = 0,01111$; $c = 0,2106$; $d = 0,00002632$. Se sabe, por un censo realizado este año, que existen en la isla 577 alces y 50 lobos. Haga una predicción de cuantos alces y lobos habrá dentro de 15 años y dentro de 28 años.

9. Un bloque de masa 1 Kg se encuentra sobre un piso horizontal. Entre el bloque y la pared hay un resorte ($k = 10$ newton/m). En el instante $t = 0$ el bloque es desplazado de manera que el muelle se comprime 20 cm y entonces se suelta el bloque. Suponga que la fuerza resistiva del medio es proporcional al cuadrado de la velocidad (coeficiente 0,3). Halle cuánto se ha desplazado el bloque al cabo de un segundo y cuál es su velocidad en ese instante. Obtenga su respuesta con tres cifras decimales exactas.
10. El algoritmo en seudo código que sigue es una variación al mostrado en esta sección para el método RK4 para problemas de Cauchy de orden m . Explique por qué este algoritmo no funciona.

```

n := 0
do while  $x_n < x_F$ 
     $(x_n, u_{1n}, u_{2n}, \dots, u_{mn})$  es un punto de la solución aproximada
    for  $i = 1$  to  $m$ 
         $K_{1i} := hf_i(x_n, u_{1n}, u_{2n}, \dots, u_{mn})$ 
         $K_{2i} := hf_i(x_n + \frac{1}{2}h, u_{1n} + \frac{1}{2}K_{11}, u_{2n} + \frac{1}{2}K_{12}, \dots, u_{mn} + \frac{1}{2}K_{1m})$ 
         $K_{3i} := hf_i(x_n + \frac{1}{2}h, u_{1n} + \frac{1}{2}K_{21}, u_{2n} + \frac{1}{2}K_{22}, \dots, u_{mn} + \frac{1}{2}K_{2m})$ 
         $K_{4i} := hf_i(x_n + h, u_{1n} + K_{31}, u_{2n} + K_{32}, \dots, u_{mn} + K_{3m})$ 
         $u_{i,n+1} := u_{i,n} + \frac{1}{6}(K_{1i} + 2K_{2i} + 2K_{3i} + K_{4i})$ 
    end
     $x_{n+1} := x_n + h$ 
     $n := n + 1$ 
end
Terminar

```

7.6 Ecuaciones diferenciales con condiciones de frontera

Cuando las condiciones que determinan la solución de una ecuación diferencial de orden mayor que 1, vienen especificadas para valores diferentes de la variable independiente, se dice que se trata de un problema con condiciones de frontera. Son, por lo general, problemas más difíciles que los de condiciones iniciales estudiados en la sección 7.5 y aquí solo se hará una breve introducción al tema, analizando un método de solución sumamente elemental pero que, en algunos casos, resulta aplicable. El lector interesado en ampliar acerca del asunto puede consultar la bibliografía que se recomienda al final de este capítulo.

La primera dificultad surge en cuanto a la existencia y unicidad de la solución. No todo problema con condiciones de frontera tiene solución y, cuando la solución existe, puede no ser única. En lo que sigue siempre se supondrá que los problemas planteados poseen una y solo una solución que satisface las condiciones de frontera dadas.

Los ejemplos que siguen son problemas de ecuaciones diferenciales ordinarias con condiciones de frontera, los cuales serán resueltos posteriormente.

Ejemplo 1

Plantee el modelo matemático del siguiente problema. Un péndulo de longitud 0,5 m se encuentra en la posición de equilibrio y se le imprime una cierta velocidad angular, de tal modo que en 0,35 segundos alcanza una deflexión de 45°. Se desea conocer la posición en cada instante en el intervalo $0 \leq t \leq 0,35$ y, además, cuál fue la velocidad angular inicial.

Solución:

Como se sabe la ecuación diferencial que rige el movimiento de un péndulo simple (vea el ejemplo 8 de la sección 7.5) es:

$$-g \operatorname{sen} \theta = L \frac{d^2\theta}{dt^2}$$

Como $L = 0,5$ m y tomando $g = 9,8$ m/s², la ecuación queda como:

$$\frac{d^2\theta}{dt^2} = -19,6 \operatorname{sen} \theta$$

Las condiciones conocidas del problema son:

$$\begin{aligned}\theta(0) &= 0 \\ \theta(0,35) &= \frac{\pi}{4}\end{aligned}$$

Se quiere conocer $\theta'(0)$ y $\theta(t)$ para $0 \leq t \leq 2$. Como se ve, se trata de un problema de condiciones de frontera.

Ejemplo 2

Resuelva la ecuación diferencial

$$\frac{d^3y}{dx^3} + y \frac{dy}{dx} - y^2 = 0$$

$$y(0) = 1$$

con las condiciones:

$$y(2) = 7$$

$$y'(0) = 2$$

El método de los disparos

Cuando en un problema de orden m con condiciones de frontera, se tienen $m - 1$ condiciones conocidas para el mismo valor de la variable independiente y una que corresponde a otro valor, el problema puede resolverse por un procedimiento de prueba y error. El método debe su nombre al hecho de que la forma de proceder recuerda la manera en que los artilleros realizan la corrección del tiro de mortero (figura 1).

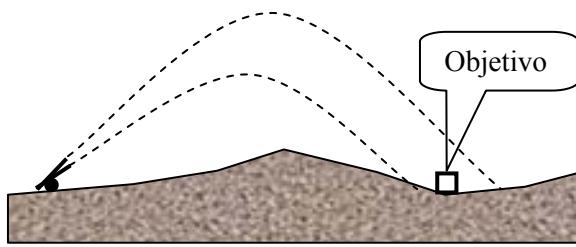


Figura 1

Considérese la ecuación de orden m :

$$\frac{d^m y}{dx^m} = G\left(x, y, \frac{dy}{dx}, \frac{d^2y}{dx^2}, \dots, \frac{d^{m-1}y}{dx^{m-1}}\right)$$

con $m - 1$ condiciones conocidas en x_0
1 condición conocida en x_F

La idea general del método es:

1. Suponer un valor arbitrario para la condición que no se conoce en x_0
2. Resolver el problema resultante de condiciones iniciales hasta x_F
3. Determinar si en x_F se satisface la condición conocida en x_F . Si es así, el problema está resuelto.
4. Si la condición conocida en x_F no se satisface, modificar adecuadamente el valor supuesto y regresar a 2.

Naturalmente que, cuando se tiene una idea del significado físico o geométrico de las variables que intervienen en la ecuación diferencial, será más simple tomar una aproximación inicial adecuada y realizar la corrección en cada paso. En cualquier caso, suponiendo que el resultado en x_F es una

función continua del valor tomado en x_0 , puede comenzarse tratando de hallar dos valores de la condición supuesta que produzcan errores de diferente signo en la condición desconocida en x_F .

Ejemplo 3

Resuelva, por el método de los disparos, el problema de condiciones de frontera del ejemplo 1. Utilice el algoritmo RK4 y obtenga las aproximaciones con tres cifras decimales exactas.

Solución:

Se trata del problema:

$$\frac{d^2\theta}{dt^2} = -19,6 \operatorname{sen} \theta \quad \begin{aligned} \theta(0) &= 0 \\ \theta(0,35) &= \frac{\pi}{4} \end{aligned}$$

El cual puede reformularse del siguiente modo:

$$\frac{d^2\theta}{dt^2} = -19,6 \operatorname{sen} \theta \quad \begin{aligned} \theta(0) &= 0 \\ \theta'(0) &= \omega \end{aligned} \quad (1)$$

Hallar ω de manera que $\theta(0,35) = \pi/4$

El problema (1) se transforma fácilmente en un problema de Cauchy de orden dos. Para ello, se define:

$$u_1 = \theta \quad y \quad u_2 = \theta'$$

y resulta:

$$\begin{aligned} \frac{du_1}{dt} &= u_2 & u_1(0) &= 0 \\ \frac{du_2}{dt} &= -19,6 \operatorname{sen} u_1 & u_2(0) &= \omega \end{aligned}$$

y se trata de encontrar el valor de ω para el cual es $u_1(0,35) = \pi/4 = 0,78540$

Disparo 1: Sea $\omega = 2$ (radian/s).

Aplicando RK4 con $h = 0,05$ (con este paso la cota del error es suficientemente pequeña), se obtienen los resultados que muestra la tabla 1.

t	$u_1(t)$	$u_2(t)$
0,00	0,00000	2,00000
0,05	0,09918	1,95124
0,10	0,19354	1,80779
0,15	0,27851	1,57778
0,20	0,35008	1,27371
0,25	0,40491	0,91116
0,30	0,44051	0,50776
0,35	0,45530	0,08169

Tabla 1

Como se aprecia, con esta velocidad angular, al cabo de los 0,35 segundos el péndulo está a punto de detenerse (u_2 está aproximándose a cero) y el ángulo de deflexión es de solo 0,45530 radianes en lugar de los 0,7854 que se desea. Estas circunstancias aconsejan aumentar la velocidad angular inicial.

Disparo 2: Sea $\omega = 3$ (radian/s).

La tabla 2 muestra los resultados obtenidos. Nótese que ahora el péndulo avanza con más rapidez, pero para $t = 0,35$ el ángulo de deflexión θ solo es de 0,68977 radianes en lugar de 0,7854 que se precisan. Es aconsejable aumentar aun más la velocidad angular inicial.

t	$u_1(t)$	$u_2(t)$
0,00	0,00000	3,00000
0,05	0,14878	2,92693
0,10	0,29033	2,71282
0,15	0,41794	2,37190
0,20	0,52575	1,92506
0,25	0,60908	1,39670
0,30	0,66448	0,81214
0,35	0,68997	0,19619

Tabla 2

Disparo 3: Sea $\omega = 4$ (radian/s).

En la tabla 3 se encuentran los resultados en este caso. Ahora la velocidad inicial es más alta de lo requerido pues, como se aprecia, para $t = 0,35$ el péndulo ha alcanzado un ángulo de 0,93238, mayor que lo que se necesita. Es evidente que el verdadero valor de ω se halla entre 3 y 4 radianes por segundo.

t	$u_1(t)$	$u_2(t)$
0,00	0,00000	4,00000
0,05	0,19837	3,90272
0,10	0,38715	3,61920
0,15	0,55755	3,17223
0,20	0,70217	2,59358
0,25	0,81529	1,91770
0,30	0,89286	1,17702
0,35	0,93238	0,39986

Tabla 3

A partir de aquí, podría utilizarse el sencillo algoritmo de bisección, haciendo el próximo disparo con $\omega = 3,5$. Sin embargo, si se tiene en cuenta lo ineficiente de dicho algoritmo, es preferible seguir otra estrategia. Llamando $H(\omega)$ a la discrepancia entre la deflexión obtenida al cabo de 0,35 segundos para la velocidad angular ω y el valor deseado, $\pi/4$, esto es:

$$H(\omega) = u_1(0,35) - 0,78540$$

los resultados obtenidos en los tres primeros disparos permiten escribir:

$$\begin{aligned} H(2) &= 0,45530 - 0,78540 = -0,33010 \\ H(3) &= 0,68997 - 0,78540 = -0,09543 \\ H(4) &= 0,93238 - 0,78540 = 0,14698 \end{aligned}$$

Como lo que se requiere es hallar una solución de la ecuación

$$H(\omega) = 0$$

puede emplearse algun método más eficiente que bisección, por ejemplo el de las secantes. La ecuación del proceso iterativo del método de las secantes (Sección 2.6, fórmula (2)) es:

$$x_n = x_{n-1} - \frac{x_{n-1} - x_{n-2}}{f(x_{n-1}) - f(x_{n-2})} f(x_{n-1})$$

La cual, adaptada a la notación de este problema, se transforma en:

$$\omega_n = \omega_{n-1} - \frac{\omega_{n-1} - \omega_{n-2}}{H(\omega_{n-1}) - H(\omega_{n-2})} H(\omega_{n-1}) \quad (2)$$

Según la ecuación (2) se tomará:

$$\begin{aligned} \omega_4 &= \omega_3 - \frac{\omega_3 - \omega_2}{H(\omega_3) - H(\omega_2)} H(\omega_3) \\ &= 4 - \frac{4 - 3}{H(4) - H(3)} H(4) = 4 - \frac{1}{0,14698 - (-0,09543)} (0,14698) \end{aligned}$$

$$\omega_4 = 3,39367$$

Disparo 4: Sea $\omega = 3,39367$ (radian/s).

Se obtiene, para $t = 0,35$, los valores $u_1(0,35) = 0,78416$ y $u_2(0,35) = 0,26403$. Con ello se tiene:

$$H(3,39367) = 0,78416 - 0,78540 = -0,00124$$

Aplicando de nuevo la fórmula (2) de las secantes se calcula la nueva velocidad angular inicial:

$$\begin{aligned} \omega_5 &= \omega_4 - \frac{\omega_4 - \omega_3}{H(\omega_4) - H(\omega_3)} H(\omega_4) \\ &= 3,39367 - \frac{3,39367 - 4}{-0,00124 - 0,14698} (-0,00124) \\ &= 3,39874 \end{aligned}$$

Disparo 5: Sea $\omega = 3,39874$ (radian/s).

Se obtiene, para $t = 0,35$, los valores $u_1(0,35) = 0,78539$ y $u_2(0,35) = 0,26501$. Con ello se tiene:

$$H(3,39874) = 0,78539 - 0,78540 = -0,00001$$

Aproximación suficiente, ya que solamente se necesitan 3 cifras decimales exactas en los resultados. Se concluye entonces que:

La velocidad angular inicial fue 3,39874 radian/s. La tabla 4 muestra los resultados finales en el intervalo pedido.

t	$u_1(t)$	$u_2(t)$
0,00	0,00000	3,39847
0,05	0,16855	3,31601
0,10	0,32893	3,07404
0,15	0,47358	2,69017
0,20	0,59599	2,18926
0,25	0,69102	1,59956
0,30	0,75494	0,94930
0,35	0,78539	0,26501

Tabla 4

Ejemplo 4

Resuelva la ecuación diferencial del ejemplo 2 mediante el método de los disparos utilizando el algoritmo RK4 en el intervalo $0 \leq x \leq 2$ con tres cifras decimales exactas.

$$\frac{d^3y}{dx^3} + y \frac{dy}{dx} - y^2 = 0$$

$$y(0) = 1$$

con las condiciones: $y(2) = 7$
 $y'(0) = 2$

Solución:

El problema puede plantearse del siguiente modo:

Dada la ecuación diferencial:

$$\frac{d^3y}{dx^3} + y \frac{dy}{dx} - y^2 = 0$$

con las condiciones iniciales:

$$y(0) = 1$$

$$y'(0) = 2$$

$$y''(0) = k$$

Halle k de manera que $y(2) = 7$.

La ecuación se puede transformar en un problema de Cauchy de orden 3, haciendo:

$$u_1 = y$$

$$u_2 = y'$$

$$u_3 = y''$$

$$\frac{du_1}{dx} = u_2 \quad u_1(0) = 1$$

$$\frac{du_2}{dx} = u_3 \quad u_2(0) = 2$$

$$\frac{du_3}{dx} = (u_1)^2 - u_1 u_2 \quad u_3(0) = k$$

y resulta:

La función $H(k)$ se define como la discrepancia entre el valor $u_1(2)$ obtenido tomando $u_3(0) = k$ y el valor requerido, $u_1(2) = 7$, esto es:

$$H(k) = u_1(2) - 7$$

y el propósito que se persigue es determinar el valor de k tal que $H(k) = 0$.

Disparo 1: $k = 2$

Tomando paso $h = 0,1$ se obtiene: $u_1(2) = 8,50538$; $u_2(2) = 7,20606$; $u_3(2) = 8,16104$

Tomando paso $h = 0,05$ se: $u_1(2) = 8,50537$; $u_2(2) = 7,20610$; $u_3(2) = 8,16111$

De donde queda claro que puede utilizarse como paso $h = 0,05$ con entera confianza en todo lo que sigue. En cuanto a la función H :

$$H(2) = 8,50537 - 7 = 1,50537$$

Disparo 2: $k = 3$

Se obtienen como resultados en $t = 2$:

$$u_1(2) = 9,76862$$

de modo que:

$$H(3) = 9,76862 - 7 = 2,76862$$

En lugar de continuar el proceso de tanteo, el próximo valor de k se buscará mediante la fórmula de las secantes (2):

$$k_3 = k_2 - \frac{k_2 - k_1}{H(k_2) - H(k_1)} H(k_2)$$

$$= 3 - \frac{3 - 2}{H(3) - H(2)} H(3)$$

$$= 3 - \frac{3-2}{2,76862-1,50537}(2,76862)$$

$$k_3 = 0,80834$$

Disparo 3: $k = 0,80834$

Se obtiene:

$$u_1(2) = 6,89796$$

de modo que:

$$H(0,80834) = 6,89796 - 7 = -0,10204$$

Aplicando de nuevo la fórmula de las secantes para determinar el próximo valor de k :

$$\begin{aligned} k_4 &= k_3 - \frac{k_3 - k_2}{H(k_3) - H(k_2)} H(k_3) \\ &= 0,80834 - \frac{0,80834 - 3}{H(0,80834) - H(3)} H(0,80834) \\ &= 0,80834 - \frac{0,80834 - 3}{-0,10204 - 2,76862} (-0,10204) \end{aligned}$$

$$k_4 = 0,88624$$

Disparo 4: $k = 0,88624$

Resulta:

$$u_1(2) = 7,00684$$

De donde:

$$H(0,88624) = 7,00684 - 7 = 0,00684$$

Como el valor de $H(k_4)$ es todavía mayor de lo permisible, se aplica la fórmula (2) de las secantes para determinar el próximo valor de k :

$$\begin{aligned} k_5 &= k_4 - \frac{k_4 - k_3}{H(k_4) - H(k_3)} H(k_4) \\ &= 0,88624 - \frac{0,88624 - 0,80834}{H(0,88624) - H(0,80834)} H(0,88624) \\ &= 0,88624 - \frac{0,88624 - 0,80834}{0,00684 - (-0,10204)} (0,00684) \end{aligned}$$

$$k_5 = 0,88135$$

Disparo 4: $k = 0,88135$

Se obtuvo:

$$u_1(2) = 7,00002$$

De donde:

$$H(0,88135) = 7,00002 - 7 = 0,00002$$

Como la discrepancia es menor que 0,0005 (tres cifras decimales exactas), se toma como resultado final. De manera que:

$$y''(0) = k = 0,88135$$

En la tabla 5 se muestran los resultados obtenidos para valores de $x = 0; 0,4; 0,8; 1,2; 1,6; 2,0$.

x	$u_1(x)$	$u_2(x)$	$u_3(x)$
0,0	1,00000	2,0000	0,88135
0,4	1,85945	2,27023	0,48087
0,8	2,80160	2,43956	0,48087
1,2	3,83411	2,78820	1,46372
1,6	5,11757	3,77662	3,67029
2,0	7,00002	5,84088	6,73486

Tabla 5

Ejercicios

1. Dada la ecuación diferencial

$$\frac{d^2y}{dx^2} + y \operatorname{sen} x = 1 \quad y(1) = 0 \\ y'(2) = 2$$

halle la solución en el intervalo $[0, 2]$ con tres cifras decimales exactas.

2. Resuelva la ecuación diferencial

$$\frac{d^2y}{dx^2} + x \frac{dy}{dx} + y^2 = e^x \quad y(0) = 1 \\ y(2) = 2$$

en el intervalo $[1, 2]$ con tres cifras decimales exactas.

3. Dada la ecuación diferencial

$$\frac{d^3y}{dx^3} + x \frac{dy}{dx} = y \cos x \quad y(2) = 7 \\ y'(0) = 3 \\ y''(0) = 1$$

halle $y'(2)$ con tres cifras decimales exactas.

4. Mediante un instrumento de precisión se determinó que un péndulo de dos metros de longitud demoró exactamente medio segundo en pasar desde la posición vertical hasta un ángulo de deflexión de 30° . Determine, con tres cifras decimales exactas, la posición del péndulo en cada décima de segundo en ese intervalo de tiempo. La ecuación diferencial correspondiente aparece en el ejemplo 1 de esta sección.

5. En el ejemplo 4 de la sección 7.1, se mostró el modelo de Lotka – Volterra para dos poblaciones $x(t)$ de presas y $y(t)$ de depredadores:

$$\begin{cases} \frac{dx}{dt} = ax - bxy \\ \frac{dy}{dt} = -cy + dxy \end{cases}$$

Suponga que en una isla hay un área protegida donde habitan lobos y alces. Los alces se alimentan del bosque y los lobos se alimentan de los alces. A partir de datos históricos se han determinado las constantes del modelo para esta situación específica: $a = 0,3$; $b = 0,01111$; $c = 0,2106$; $d = 0,00002632$. Se sabe, por un censo realizado hace 15 años, que en aquella época había en la isla 500 alces y una cantidad indeterminada de lobos. Después de disminuir drásticamente, la población de alces comenzó a crecer y hoy ya hay 600 de ellos. Determine, a partir de esta información, cuantos lobos había hace 15 años y cuantos debe haber ahora.

Otras lecturas recomendadas

El tema de ecuaciones diferenciales es uno de los más amplios dentro de la Matemática Numérica y por ello, existen obras completas dedicadas solamente al mismo. En este libro, el análisis se ha limitado a las ecuaciones diferenciales ordinarias. El tema de las ecuaciones diferenciales parciales resulta tan abundante en métodos y resultados que se ha preferido no tratarlo ni siquiera en forma elemental. El lector interesado en este importante asunto, puede consultar las obras clásicas: “Finite Difference methods for partial differential equations” de Forsythe y Wasow y o “Métodos en diferencias para las ecuaciones elípticas” de Samarski y Andréiev. En la extensa obra “Computing Methods” de Berezin y Zhidkov se dedican unas 400 páginas al tratamiento del tema de las ecuaciones diferenciales (ordinarias y parciales) y el lector puede allí encontrar algunas demostraciones y muchos métodos que en este texto no ha sido posible contemplar.

Principales ideas del capítulo

- Una ecuación diferencial es aquella en la que aparecen derivadas. Cuando hay una sola variable independiente respecto a la cual se plantean todas las derivadas, entonces las derivadas se llaman ordinarias y también la ecuación diferencial.
- Cuando se requiere más de una condición particular para determinar la solución que se desea, pueden darse dos situaciones muy diferentes: Que se especifiquen todas las condiciones particulares para un mismo valor de la variable independiente (*condiciones iniciales*) o que se incluyan condiciones particulares que deberá satisfacer la solución para dos o más valores de la variable independiente (*condiciones de frontera*).
- Las principales limitaciones de los métodos analíticos para resolver ecuaciones diferenciales ordinarias son que cada método analítico se ocupa de un tipo especial de ecuación diferencial ordinaria y es inaplicable en otros casos y que, a pesar de la diversidad de métodos analíticos, la mayoría de las ecuaciones diferenciales ordinarias no puede resolverse por esta vía.
- El campo de direcciones brinda una idea cualitativa muy abarcadora acerca de cómo se comportarán las soluciones de una ecuación diferencial de primer orden. Si $y = y(x)$ es una solución particular de la ecuación diferencial, entonces cuando su gráfica pase por un punto del campo de direcciones, lo hará en la dirección del segmento que corresponde a ese punto.
- En el campo de las ecuaciones diferenciales, se considerará que el problema de Cauchy:

$$\frac{dy}{dx} = f(x, y) \quad y(x_0) = y_0$$

es inestable si pequeños cambios en y_0 producen grandes cambios en la solución de la ecuación para valores de x alejados de x_0 .

- La estabilidad de la ecuación diferencial

$$\frac{dy}{dx} = f(x, y)$$

dependen del signo de $f_y(x, y)$ en la región considerada del plano xy . El problema es estable cuando esta derivada es negativa en toda la región considerada.

- La ecuación diferencial $\frac{dy}{dx} = -Ay$ con $A > 0$ se llama *ecuación estable modelo*.
- El método de Euler para resolver una ecuación diferencial de primer orden consiste en las ecuaciones: $x_{n+1} = x_n + h$; $y_{n+1} = y_n + hf(x_n, y_n)$ para $n = 0, 1, 2, \dots$
- El método de Euler posee un error local de orden h^2 y un error total de orden h , por lo cual se dice que es de primer orden.
- El error total de un algoritmo de orden p puede estimarse mediante la fórmula $e_h = \frac{y_h - y_{2h}}{2^p - 1}$.
- El método de Euler es estable con la ecuación modelo, si se toma un paso h tal que $hA < 2$.
- Las fórmulas de Runge - Kutta de orden dos (RK2) son:

$$\text{RK2 : } \begin{cases} K_1 = hf(x_n, y_n) \\ K_2 = hf(x_n + h, y_n + K_1) \\ y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2) \end{cases} \quad n = 0, 1, 2, 3, \dots$$

- El método RK2 es estable con la ecuación modelo, si se toma un paso h tal que $hA < 2$
- Las fórmulas de Runge - Kutta de orden cuatro (RK4) son:

$$\text{RK4 : } \begin{cases} K_1 = hf(x_n, y_n) \\ K_2 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_1) \\ K_3 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_2) \\ K_4 = hf(x_n + h, y_n + K_3) \\ y_{n+1} = y_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4) \end{cases} \quad n = 0, 1, 2, 3, \dots$$

- Si se usa los últimos $p + 1$ valores de la solución: $y_n, y_{n-1}, \dots, y_{n-p}$, para hallar y_{n+1} . se dice que se trata de un método de paso $p + 1$, que se simboliza: $y_{n+1} = G(y_n, y_{n-1}, \dots, y_{n-p})$
- Como en los métodos de paso múltiple se utiliza una mayor cantidad de información acerca de la solución ya calculada, se logra una mayor eficiencia computacional, en el sentido de una menor cantidad de operaciones para obtener una exactitud similar.
- Un método de paso $p + 1$ no puede funcionar mientras no se conozcan $p + 1$ valores de la solución: $y_0, y_1, y_2, \dots, y_p$. Por esto se dice que los métodos de paso múltiple no son capaces de “arrancar”.
- El método de Adams – Bashforth de orden 4 consiste en la fórmula

$$y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$$

y es un método de paso cuádruple.

- El método de Adams – Moulton de orden 4 consiste en la fórmula

$$y_{n+1} = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})$$

y es un método de paso triple e implícito.

- Los métodos de Adams – Bashforth y de Adams – Moulton se usan frecuentemente como una pareja predictor – corrector donde

$$\text{Ecuación predictora: AB4} \quad y_{n+1}^{(0)} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$$

$$\text{Ecuación correctora: AM4} \quad y_{n+1}^{(k+1)} = y_n + \frac{h}{24}[9f(x_{n+1}, y_{n+1}^{(k)}) + 19f_n - 5f_{n-1} + f_{n-2}]$$

- El método de Adams – Bashforth de orden cuatro es inestable para la ecuación estable modelo, para valores de hA por encima de 0,3 mientras que, a partir de $hA = 3$ comienza la zona de inestabilidad del método de AM4.
- A un sistema de ecuaciones diferenciales se llama un *problema de Cauchy de orden m* cuando presenta las siguientes características: El sistema contiene tantas ecuaciones diferenciales como variables dependientes; existe solamente una variable independiente; todas las ecuaciones son de primer orden; en la ecuación número i existe una sola derivada que corresponde a la variable u_i ; la derivada que contiene cada ecuación aparece despejada en uno de los miembros de la ecuación; para un valor x_0 de la variable independiente se conoce el valor que toman todas las variables dependientes.
- Para transformar una ecuación de orden m con condiciones iniciales, en un problema de Cauchy de orden m se introducen m nuevas variables u_1, u_2, \dots, u_m definidas como sigue:

$$u_1 = y; \quad u_2 = \frac{dy}{dx}; \quad u_3 = \frac{d^2y}{dx^2}; \dots; \quad u_m = \frac{d^{m-1}y}{dx^{m-1}}$$

- Cualquiera de los métodos estudiados para resolver ecuaciones diferenciales de primer orden puede ser adaptado para resolver problemas de Cauchy de orden m , solamente hay que realizar el siguiente cambio. Si el método que se va a emplear consta de varias etapas en cada paso, entonces en cada paso de la solución del problema de Cauchy, se aplica la etapa 1 a las m ecuaciones, después, la etapa 2 a todas las ecuaciones, etc. hasta aplicar la última etapa a todas las ecuaciones y con esto queda terminado un paso.
- Si \mathbf{u}_h es el vector de soluciones de un problema de Cauchy obtenido para un cierto valor x utilizando paso h con un método de orden p y \mathbf{u}_{2h} el vector de soluciones correspondiente al mismo valor x pero calculado con paso $2h$, entonces, si \mathbf{e}_h representa al vector de errores correspondiente a \mathbf{u}_h , se cumple que

$$\|\mathbf{e}_h\| \approx \frac{\|\mathbf{u}_h - \mathbf{u}_{2h}\|}{2^p - 1}$$

- La idea general del método de los disparos para resolver una ecuación diferencial con condiciones de frontera consiste en: 1) Suponer un valor arbitrario para la condición que no se conoce en x_0 ; 2) Resolver el problema resultante de condiciones iniciales hasta x_F ; 3) Determinar si en x_F se satisface la condición conocida en x_F . Si es así, el problema está resuelto; 4) Si la condición conocida en x_F no se satisface, modificar adecuadamente el valor supuesto y regresar a 2).

Auto examen

1. Muestre dos ejemplos de ecuaciones diferenciales: uno de condiciones iniciales y el otro con condiciones de frontera.
2. Cite tres diferencias importantes entre los métodos analíticos y los numéricos que se utilizan para resolver ecuaciones diferenciales.
3. ¿A qué se llama campo de direcciones de una ecuación diferencial de primero orden y para qué sirve?
4. Analice el campo de direcciones de la ecuación diferencial $\frac{dy}{dx} = x^2 - y^2$ y diga algunas de las propiedades que poseen sus soluciones.
5. ¿Cuándo se dice que una ecuación diferencial es estable?
6. ¿Cuál es la interpretación geométrica de los métodos de Euler y de Runge – Kutta orden 2?
7. Dada la ecuación diferencial $\frac{dy}{dx} = \frac{x+y}{y+1}$; $y(1) = 2,5$; obtenga su solución en el intervalo $[1, 3]$ mediante los métodos RK2 y RK4 con 3 cifras decimales exactas y compare la cantidad de evaluaciones de la función que fueron necesarias en cada caso.
8. Resuelva mediante el método predictor – corrector de Adams de orden 4 el ejercicio anterior.
9. ¿Por qué se dice que los métodos de Adams – Bashforth son poco estables? ¿A qué se debe esta poca estabilidad?
10. Un bloque de hielo con una masa de 10 Kg, se deja caer desde una gran altura. Desprecie la resistencia del aire y determine su posición durante los primeros cinco segundos de la caída si, debido a la temperatura, el bloque va perdiendo masa a razón de 5 g/s. Obtenga la solución con error menor que un milímetro.
11. Resuelva la ecuación diferencial:

$$\frac{d^2y}{dx^2} + xy = e^x \quad y(1) = 2 \\ y(2) = 4,5$$

en el intervalo $[1, 2]$ con tres cifras decimales exactas.

RESPUESTAS A LOS EJERCICIOS

Sección 1.2

2. Podría considerarse la fricción entre el aire y la piedra.
3. 1227 km/h
4. Hasta del orden de $\frac{1}{16}$ pulgada.
5. Hasta 5%.
6. Hasta 0,5 mm
7. Ancho de la pantalla/2000
8. Hasta del orden de $\frac{1}{16}$ pulgada.

Sección 1.3

1. a) $E(e_A) = 0,01828$; $e(e_A) = 0,67 \%$
b) $E(c_A) = 2,07 \cdot 10^5 \text{ m/s}$; $e(c_A) = 0,069 \%$
c) $E(g_A) = 0,2 \text{ m/s}^2$; $e(g_A) = 2,04 \%$
d) $E(C_A) = 0,002784$; $e(C_A) = 0,048 \%$
e) $E(m_{pA}) = 8,017 \cdot 10^{-29} \text{ g}$; $e(m_{pA}) = 1,65 \%$
f) $E(m_{eA}) = 2,761 \cdot 10^{-26} \text{ g}$; $e(m_{eA}) = 8,7 \%$
2. $\sqrt{2} \approx 1,45$. Error relativo máximo: 3,6 %
3. 1,415625. Error absoluto menor que 0,0008
4. Error absoluto máximo: $3\mu\text{F}$. Error relativo máximo: 9,4 %
5. Entre 18,018 km y 18,584 km
6. En ambas direcciones: 0,137 mm
7. Error absoluto máximo: 0,225 v. Intervalo de seguridad: [224,77; 225,23]
8. $x = 0,1817$
9. $x = 0,6555$
10. Error absoluto: 0,00439 Error relativo: 0,14 %
11. Error absoluto: 0,0189 Error relativo: 0,602 %
12. La resistencia de 50 k Ω es un error del fabricante.

Sección 1.4

1. a) Los primeros cuatro ceros son no significativos.
b) Todos son significativos.
c) Todos son significativos.
2. a) Probablemente, todos los ceros que aparecen son no significativos.
b) Todos son significativos.
c) Lo mas probable es que todos los ceros sean no significativos.
d) Lo mas probable es que todos los ceros (o, al menos dos) sean no significativos.
e) Todos son significativos.
3. a) Son exactas 1 y 7. Dos cifras exactas. Una cifra decimal exacta.
b) Son exactas 4 y 4. Dos cifras exactas. Cinco cifras decimales exactas.
c) Son exactas 2 y 3. Dos cifras exactas. Ninguna cifra decimal exacta.

- d) Son exactas 6 y 7. Dos cifras exactas. Tres cifras decimales exactas.
e) Son exactas 3, 1 y 4. Tres cifras exactas. Dos cifras decimales exactas.
4. a) Cuatro cifras exactas. Tres cifras decimales exactas. No hay cifras dudosas.
b) Tres cifras exactas. Cinco cifras decimales exactas. No hay cifras dudosas.
c) Dos cifras exactas. Seis cifras decimales exactas. No hay cifras dudosas.
d) Tres cifras exactas. Una cifra decimal exacta. 3 y 1 son cifras dudosas.
e) Cinco cifras exactas. Ninguna cifra decimal exacta. 3, 5 y 2 son cifras dudosas.
f) Dos cifras exactas. Ninguna cifra decimal exacta. No hay cifras dudosas.
g) Tres cifras exactas. Cuatro cifras decimales exactas. 6 y 3 son cifras dudosas.
h) Dos cifras exactas. Seis cifras decimales exactas. 5 es una cifra dudosa.
5. a) $E_m(x) = 0,005$; $e_m(x) = 7,17 \cdot 10^{-6}$; 8 y 7 son dudosas.
b) $E_m(x) = 0,005$; $e_m(x) = 0,0001$; 6 es dudosa.
c) $E_m(x) = 0,000005$; $e_m(x) = 0,000913$; 6 y 8 son dudosas.
d) $E_m(x) = 0,000005$; $e_m(x) = 0,00837$; 7 y 3 son dudosas.
e) $E_m(x) = 0,0005$; $e_m(x) = 0,0001$; los dos últimos 9 son dudosas.
f) $E_m(x) = 50$; $e_m(x) = 0,000128$; 8 y 5 son dudosas.
g) $E_m(x) = 0,0005$; $e_m(x) = 0,00506$; 7 y 8 son dudosas.
h) $E_m(x) = 0,0000005$; $e_m(x) = 0,0141$; 4 y 3 son dudosas.
6. a) $x_A = 58,547$; $E(x_A) = 0,00046$; $e(x_A) = 7,86 \cdot 10^{-6}$.
b) $x_A = 0,0454$; $E(x_A) = 0,000035$; $e(x_A) = 0,000771$.
c) $x_A = 6,550$; $E(x_A) = 0,000127$; $e(x_A) = 0,0000194$.
d) $x_A = 67800$; $E(x_A) = 45,675$; $e(x_A) = 0,000674$.
e) $x_A = 0,0066$; $E(x_A) = 0,0000213$; $e(x_A) = 0,00322$.

Sección 1.6

1. $S = 43,292$; $E_m(S) = 0,673$; $e_m(S) = 1,6\%$; una cifra exacta.
2. $L = 3,5066$; tres cifras exactas.
3. $x = 2,0485$; $E_m(S) = 0,00048$; cuatro cifras exactas.
4. $S = 77088$; $e_m(S) = 0,5\%$;
5. 0,2234; tres cifras exactas.
6. $V = 932\,188 \text{ cm}^3$; $E_m(V) = 54\,000 \text{ cm}^3$; $e_m(V) = 5,8\%$. Para reducir $E_m(V)$ a la mitad, una posibilidad es hacer: $E_m(h) = 1 \text{ cm}$ y $E_m(r) = 0,5 \text{ cm}$.
7. Una posibilidad: π con cinco cifras exactas, r con tres cifras exactas.
8. El error absoluto máximo por truncamiento es 0,08214. El error absoluto máximo debido al redondeo es 0,00114. El error absoluto máximo total es 0,084.
9. a) $E_m(y) = 0,00003$; b) $E_m(y) = 0,000018$; c) $E_m(y) = 0,000011$.

Sección 1.7

2. El determinante del sistema es pequeño en comparación con sus coeficientes.
4. $f(x) \approx -\frac{1}{3} + \frac{1}{30}x^2$
5. $f(x) \approx -\frac{5}{24} + \frac{11}{48}x$

Sección 2.2

1. a) 1 ó 3 raíces positivas en $(0; 2,42)$. Una raíz negativa en $(-3; 0)$
 b) 0, 2 ó 4 raíces positivas en $(0; 61)$. No hay raíces negativas.
 c) 0, 2 ó 4 raíces positivas en $(0; 14)$. No hay raíces negativas.
 d) No hay raíces reales.
 e) No hay raíces positivas. 1 ó 3 raíces negativas en $(-10; 0)$.

2. a) Raíces en: $(0,8; 1,3)$ y $(-2,2; -1,7)$.
 b) Raíces en: $(0,8; 1,3)$, $(-1,3; 1,8)$, $(3,5; 4)$ y $(4,5; 5)$.
 c) Raíces en: $(0,3; 0,8)$ y $(0,8; 1,3)$.
 d) No hay raíces reales.
 e) Raíz en $(-4,5; -4)$.

3. Ceros en $(0,7; 1)$, $(2,1; 2,4)$, $(2,6; 2,9)$, $(4; 4,3)$ y $(4,7; 5)$.
 Puntos de extremo en $(1,3; 1,6)$, $(2,3; 2,6)$, $(3,5; 3,8)$ y $(4,5; 4,8)$.
 Puntos de inflexión en: $(1,7; 2)$, $(2,8; 3,1)$ y $(4; 4,3)$.

4. Hay una raíz real en $(2; 2,5)$. En ese intervalo la derivada es negativa.
5. Raíces en $(-5,3; -4,8)$ y $(0,7; 1,2)$
6. Raíces en $(-8; -7,5)$, $(-5; -4,5)$, $(-1,3; -0,8)$, $(-0,25; 0,25)$, $(0,8; 1,3)$, $(4,5; 5)$ y $(7,5; 8)$.
7. Una raíz en $(0,1; 0,5)$. En este intervalo la primera derivada es positiva y la segunda negativa.
8. Raíces en $(-2,4; -2,2)$ y $(2,6; 2,8)$.
9. Una raíz en $(0,75; 1,25)$.
10. Raíces en $x = 0$, $(0,6; 0,8)$, $(2,1; 2,3)$, $(4,1; 4,3)$ y $(5; 5,2)$.
11. La mayor raíz está en $(0,8; 1)$.
12. Abscisas de los puntos en $(0,5; 0,6)$, $(1,9; 2)$, $(-0,6; -0,5)$ y $(-2, -1,9)$.
13. Raíces en $(0,6; 0,7)$ y $(2; 2,1)$.

Sección 2.3

1. a) -2 y 1
 b) $1,267949$; $1,381966$; $3,618034$ y $4,732051$.
 c) $0,705547$ y 1 .
 d) No hay raíces reales.
 e) $-4,147899$.

2. Ceros: $0,961505$; $2,209266$; $2,724166$; $4,150984$ y $4,954080$.
 Puntos de extremo: $1,355567$; $2,456088$; $3,543912$ y $4,644433$.
 Puntos de inflexión: $1,775255$; 3 y $4,224745$.

3. a) $2,219107$.
 b) $-4,982864$ y $0,850182$.
 c) $-7,837964$; $-4,754761$; $-1,102506$; $1,102506$; $4,754761$ y $7,837964$.
 d) $0,314836$.
 e) $-2,274803$ y $2,733240$.
 f) $0,905085$.
 g) $0,694899$; $2,167455$; $4,261276$ y $5,128225$
 h) $2,043790$.

4. $0,897540$

5. Ceros: $-0,162375$; $0,975030$ y $6,343047$.
 Puntos de extremo: $0,383465$ y $4,387028$.
 Punto de inflexión: $2,440804$.
6. Puntos de intersección: $(-1,995373; 0,135963)$ y $(0,639263; 1,895084)$
 7. Vértices: $(0,89754; 0,89754)$ y $(0,93613; 0,87635)$
 8. Toma el valor 2 en $1,559610$. La pendiente es 2 en $1,363670$.
 9. $y = 2,306964 + 3,153388(x - 1,519855)$.
 10. Ancho del pasillo: $1,5988$ m.
 11. Radios: $5,054559$ cm; $6,054559$ cm y $7,054559$ cm.

Sección 2.4

1. a) 1 ó 3 raíces positivas en $(0; 2,42)$. Una raíz negativa en $(-3; 0)$
 b) 0, 2 ó 4 raíces positivas en $(0; 61)$. No hay raíces negativas.
 c) 0, 2 ó 4 raíces positivas en $(0; 14)$. No hay raíces negativas.
 d) No hay raíces reales.
 e) No hay raíces positivas. 1 ó 3 raíces negativas en $(-10; 0)$.
2. a) $2,219107$.
 b) $-4,982864$ y $0,850182$.
 c) $-7,837964$; $-4,754761$; $-1,102506$; $1,102506$; $4,754761$ y $7,837964$.
 d) $0,314836$.
 e) $-2,274803$ y $2,733240$.
 f) $0,905085$.
 g) $0,694899$; $2,167455$; $4,261276$ y $5,128225$
 h) $2,043790$.
3. Punto de intersección: $(1,28533; 2,12347)$.
 4. Pendiente $-0,221500$
 5. $1,49022$.
 6. $a = 1,256431$.
 7. $a = 25,32649$.

Sección 2.5

1. $x_0 = 0$; $x_0 = 0,8$; $x_0 = 2,3$ y $x_0 = 4$.
2. a) -2 y 1
 b) $1,267949$; $1,381966$; $3,618034$ y $4,732051$.
 c) $0,705547$ y 1 .
 d) No hay raíces reales.
 e) $-4,147899$.
3. a) $2,219107$.
 b) $-4,982864$ y $0,850182$.
 c) $-7,837964$; $-4,754761$; $-1,102506$; $1,102506$; $4,754761$ y $7,837964$.
 d) $0,314836$.
 e) $-2,274803$ y $2,733240$.
 f) $0,905085$.

- g) 0,694899; 2,167455; 4,261276 y 5,128225
 h) 2,043790.

4. $r = 5,363858$ cm.
5. Área: 1,368454 unidades cuadradas.
6. $y = 0,21723(x - 2\pi)$.
7. Debe darse el corte a 7,378 cm del centro.
8. Radio: 1,814544.

Sección 2.6

1. $x_0 = 4,5$ y $x_1 = 4,4$.
2. a) -2 y 1
 b) 1,267949; 1,381966; 3,618034 y 4,732051.
 c) 0,705547 y 1.
 d) No hay raíces reales.
 e) -4,147899.
3. a) 2,219107.
 b) -4,982864 y 0,850182.
 c) -7,837964; -4,754761; -1,102506; 1,102506; 4,754761 y 7,837964.
 d) 0,314836.
 e) -2,274803 y 2,733240.
 f) 0,905085.
 g) 0,694899; 2,167455; 4,261276 y 5,128225
 h) 2,043790.
4. Radio: 0,494539.
5. Volumen (litros) Altura de la marca (mm) Volumen (litros) Altura de la marca (mm)

1000	312,0	6000	1157,7
2000	508,1	7000	1319,7
3000	680,3	8000	1491,9
4000	842,3	9000	1688,0
5000	1000,0	10000	2000,0
6. Volumen (litros) Altura de la marca (mm) Volumen (litros) Altura de la marca (mm)

1000	523,4	6000	1515,8
2000	767,5	7000	1702,0
3000	971,1	8000	1905,5
4000	1157,2	9000	2149,6
5000	1336,5	10000	2673,0
7. $a = 8,17436$ unidades.
 8. El poste se partió a 1,1282 m del suelo.
 9. Debe colocarse a 4,5526 m ó a 0,9423 m del cuadro.
 10. El otro extremo se encuentra en el punto (2,20036 cm; 4,84158 cm).
 13. $a = 0,70138$ y $b = -0,53876$ (o viceversa).

Sección 2.7

1. $x = 1,99676; y = 0,11362$ y $x = -1,67439; y = 1,09381$
2. $x = 0,72855; y = 1,08981$
3. Puntos de intersección: $(0,43472; -0,80959)$ y $(1,33920; -0,44272)$.
4. $a = 0,70138$ y $b = -0,53876$ (o viceversa).
5. $0,32132 + 0,41060 i$
6. $0,32132 + 0,41060 i$
7. $3,85592 + 0,82665 i; 3,85592 - 0,82665 i; 0,64408 + 0,31556 i$ y $0,64408 - 0,31556 i$.
8. $-0,28562 + 0,84042 i; -0,28562 - 0,84042 i; 0,28562 + 1,56744 i$ y $0,28562 - 1,56744 i$.

Sección 3.2

1. a) $x = 1,595571; y = 1,300911; z = 1,131726$.
b) $x = -43; u = -10; v = 111; z = -53$.
c) $x = -0,610831; y = 1,338060; z = 2,805128$.
d) $x = 2,110213; y = 2,25; z = -1,428899$.
e) $x = -0,123009; y = 0,514663; z = 0,121067$.
2. $a = -12,579365; b = 79,710317; c = -159,439286; d = 102,997500$.
3. $x = 2,857143; y = 1,785714; z = 0,5$.
4. $x = 1,53125; y = 0,375; z = -2,15625$.
5. $k = 0,221093; A = 1,418117; B = 0,772147; C = 1,727054$.
6. $x + y = 0$.
7. Una posibilidad: $\begin{cases} x = t^2 - t + 1 \\ y = 1,5t^2 - 0,5t + 2 \end{cases} \quad -1 \leq t \leq 1$
10. En litros por minuto: $x_1 = 150; x_2 = 50; x_3 = 300; x_4 = 150; x_5 = 350; Q_3 = 500$.

Sección 3.3

1. a) $x_1 = 0,332315; x_2 = 0,891265; x_3 = 0,012220; x_4 = 0,336993; x_5 = 0,732601$.
b) $x_1 = 0,814852; x_2 = 0,962870; x_3 = 0,641746; x_4 = 0,383091; x_5 = 0,970146$.
c) $x_1 = -1,679549; x_2 = 0,961353; x_3 = 0,103060; x_4 = 0,710145; x_5 = -0,140097$.
2. a) 6,275 s (sin contar operaciones de sumar); b) 0,016 s (contando todas las operaciones).
4. a) -626; b) -204; c) 105.
6. a) Orden 4: 3,55 min; orden 6: 12 min; orden 8: 28,44 min; orden 10: 55,56 min.
b) Orden 4: 7,92 min; orden 6: 6 horas; orden 8: 89,6 horas; orden 10: 420 días.

7. a)
$$\begin{bmatrix} 1 & 1,6 & -1,7 \\ 0 & -0,4 & 0,3 \\ -1 & -1,4 & 1,8 \end{bmatrix}$$
 b)
$$\begin{bmatrix} 0,00000 & 0,35000 & 0,05000 & -0,15000 \\ 0,22222 & -2,00000 & 0,17778 & 0,02222 \\ -0,22222 & -0,05000 & 0,07222 & 0,22778 \\ 0,11111 & 0,00000 & -0,11111 & 0,11111 \end{bmatrix}$$

c)
$$\begin{bmatrix} -0,16667 & 0,16667 & 0,00000 & 0,50000 \\ -0,83333 & -0,83333 & 1,00000 & 0,50000 \\ 0,33333 & 0,83333 & -0,50000 & -0,50000 \\ 0,16667 & -0,33333 & 0,50000 & 0,00000 \end{bmatrix}$$

9.
$$\begin{aligned} x_1 &= -0,42y_1 + 0,64y_2 - 0,18y_3 + 0,8y_4 \\ x_2 &= 0,7y_1 - 0,4y_2 + 0,3y_3 - y_4 \\ x_3 &= -0,34y_1 + 0,28y_2 + 0,14y_3 - 0,6y_4 \\ x_4 &= -0,8y_1 + 0,6y_2 - 0,2y_3 + y_4 \end{aligned}$$

Sección 3.4

1. a) Número de condición: 22,34; mal condicionado.
b) Número de condición: 73633; muy mal condicionado.
2. a) Por ejemplo, al cambiar a_{44} de 27 a 25, la solución en x_4 cambia en más del 100 %.
b) Por ejemplo, al cambiar a_{44} de 30 a 31, la solución en x_4 cambia de 1442 a 10,014.

Sección 3.5

1. a) Diagonal no predominante. $\alpha = 1,1$.
b) Diagonal predominante; $\alpha = 0,9$; $\beta = 0,4$;
 Detener el algoritmo de Jacobi cuando $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq 5,56 \cdot 10^{-6}$;
 Detener el algoritmo de Seidel cuando $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq 7,46 \cdot 10^{-5}$.
c) Diagonal predominante; $\alpha = 0,9$; $\beta = 0,9$;
 En ambos métodos, detener el algoritmo cuando $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq 5,56 \cdot 10^{-6}$;
 Detener el algoritmo de Seidel cuando $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq 5,56 \cdot 10^{-6}$.

2. a) $5x - y + 2z = 25$
 $4x - 15y + z = -7$ $\alpha = 0,6$; $\beta = 0,6$.
 $x + 3y - 9z = 3$

b) $8x - y + 2z = 14$
 $3x - 7y - z = 5$ $\alpha = 0,5714$; $\beta = 0,371$.
 $-x + 3y + 9z = 4$

c) $5x + 2y + z = 4$
 $-x + 4y - 2z = 5$ $\alpha = 0,75$; $\beta = 0,67$.
 $3x + 2y + 9z = 25$

6. Para $n = 10$: El método de Gauss requiere 333 operaciones. El método de Seidel necesita: 600 ($\beta = 0,2$); 1400 ($\beta = 0,5$); 4200 ($\beta = 0,8$). En los tres casos el método de Gauss es más rápido.
- Para $n = 100$: El método de Gauss requiere 333000 operaciones. El método de Seidel necesita: 60000 ($\beta = 0,2$); 140000 ($\beta = 0,5$); 420000 ($\beta = 0,8$). En los dos primeros casos el método de Seidel es más rápido.
- Para $n = 1000$: El método de Gauss requiere $333 \cdot 10^6$ operaciones. El método de Seidel necesita: $6 \cdot 10^6$ ($\beta = 0,2$); $14 \cdot 10^6$ ($\beta = 0,5$); $42 \cdot 10^6$ ($\beta = 0,8$). En los tres casos el método de Seidel es más rápido.

Sección 3.6

1. x es valor propio de C para $\lambda = 11$.
 y es valor propio de A para $\lambda = 3$ y de B para $\lambda = 2$.
 z es valor propio de A para $\lambda = 1$ y de B para $\lambda = -2$.
2. El valor propio de mayor valor absoluto es 0,705777.
3. Los valores propios son: $-0,5; -0,456941; 0,251164$ y $0,705777$.
4. a) Valores propios: 9,198523; 4,463077; 2,338400.
b) Valores propios: 5,522920; $-3,670893$; 0,147972.
c) Valores propios: $-5,027598$; 3,635654; 3,391944.
d) Valores propios: 2,248298; $-0,124149 + 1,882242 i$; $-0,124149 - 1,882242 i$.

Sección 4.3

1. a) $y = 11$
b) $y = 5$ (tomando nodos 0; 1 y 3).
2. a) 1,533333
b) $-0,06667x^3 + 1,66667x^2 - 2,56667x + 3$.
3. $\tan 1 \approx 1,57721$; error absoluto menor que 0,046.
4. Tomando nodos 0; 1; 4; 9: $p(x) = 0,016667x^3 - 0,25x^2 + 1,23333x$. $p(2) = 1,6$
Tomando nodos 1; 4; 9: $p(x) = -0,016667x^2 + 0,416667x + 0,6$. $p(2) = 1,366667$.
5. Tomando nodos 0; 0,5; 1: $p(x) = 0,575657x^2 - 0,032577x + 1$; Error absoluto menor que 0,01 para $0 \leq x \leq 1$.
6. $p(x) = 0,175353x^3 - 0,984884x^2 + 2,440212x - 1,630681$ (tomando nodos 1; 1,166667; 1,333333; 1,5); error absoluto menor que 0,0002 para $1 \leq x \leq 1,5$.
7. $x = 0,749995$.
8. $\frac{1}{6}n(2n^2 + 3n + 1)$
10. Tomando nodos $10^\circ; 20^\circ; 30^\circ$: $g = 9,7881$.
11. Tomando nodos 20; 25; 30: viscosidad = 0,00865 poises.
12. Tomando nodos 1; 1,5; 2,5; 4: corriente permisible = 17,1 ampere.

Sección 4.4

1. $f(0,38) = 1,30134; f(0,5) = 1,41421$.
2. $f(1) = 0,7642; f(2) = 0,2242; f(4) = -0,3953$.
3. $\cos 10^\circ \approx 0,9826$; $\cos 20^\circ \approx 0,9379$; $\cos 40^\circ \approx 0,7659$; $\cos 50^\circ \approx 0,6427$; $\cos 70^\circ \approx 0,3460$; $\cos 80^\circ \approx 0,1794$.
4. Grado 3.

5. $f(2,7) = 2,9870$.
6. $\Phi(0,323) = 0,2533$.
7. Longitud: 22,10374.
8. 1,00162 cal/g·grado.
10. Tomando nodos 0; 30; 60: 21,552 dina/cm; error estimado: - 0,0022.
11. Tomando nodos 125; 250; 500: 0,02971; error estimado: 0,00027.

Sección 4.5

1. Para $k = 1$:

$$f(x) = \begin{cases} 4,2x - 10,4 & 3 \leq x \leq 3,5 \\ x + 0,8 & 3,5 \leq x \leq 4 \\ -1,4x + 10,4 & 4 \leq x \leq 4,5 \\ -0,4x + 5,9 & 4,5 \leq x \leq 5 \\ 2,8x - 10,1 & 5 \leq x \leq 5,5 \\ 5,2x - 23,3 & 5,5 \leq x \leq 6 \end{cases}$$

Para $k = 2$:

$$f(x) = \begin{cases} -3,2x^2 + 25x - 44 & 3 \leq x \leq 4 \\ x^2 - 9,9x + 28,4 & 4 \leq x \leq 5 \\ 2,4x^2 - 22,4x + 55,9 & 5 \leq x \leq 6 \end{cases}$$

Para $k = 3$:

$$f(x) = \begin{cases} 0,53333x^3 - 8,8x^2 + 44,46667x - 66,4 & 3 \leq x \leq 4,5 \\ -0,53333x^3 + 11,2x^2 - 70,66667x + 143,9 & 4,5 \leq x \leq 6 \end{cases}$$

$$2. \quad s(x) = \begin{cases} -2,71692x^3 + 24,45231x^2 - 68,47769x + 60,91923 & 3 \leq x \leq 3,5 \\ 0,78462x^3 - 12,31385x^2 + 60,20385x - 89,20923 & 3,5 \leq x \leq 4 \\ 2,77846x^3 - 36,24000x^2 + 155,9085x + 216,8154 & 4 \leq x \leq 4,5 \\ 1,70154x^3 - 21,70154x^2 + 90,48538x - 118,6808 & 4,5 \leq x \leq 5 \\ -0,78462x^3 + 15,59077x^2 - 95,97615x + 192,0885 & 5 \leq x \leq 5,5 \\ -1,76308x^3 + 31,73538x^2 - 184,77154x + 354,8800 & 5,5 \leq x \leq 6 \end{cases}$$

$$3. \quad s(x) = \begin{cases} -0,07964x^3 + 0,00000x^2 - 0,22036x + 5,00000 & 0 \leq x \leq 1 \\ -0,00180x^3 - 0,23350x^2 + 0,01314x + 4,92216 & 1 \leq x \leq 2 \\ 0,18686x^3 - 1,36546x^2 + 2,27706x + 3,41286 & 2 \leq x \leq 3 \\ 0,05438x^3 - 0,17320x^2 - 1,29974x + 6,98969 & 3 \leq x \leq 4 \\ -0,00438x^3 + 0,53196x^2 - 4,12036x + 10,75052 & 4 \leq x \leq 5 \\ -0,23686x^3 + 4,01907x^2 - 21,55593x + 39,80979 & 5 \leq x \leq 6 \\ -0,04820x^3 + 0,62320x^2 - 1,18067x - 0,94072 & 6 \leq x \leq 7 \\ 0,12964x^3 - 3,11134x^2 + 24,96108x - 61,93814 & 7 \leq x \leq 8 \end{cases}$$

$$4. \quad s(x) = \begin{cases} 7,18359x^3 - 215,5076x^2 + 2151,788x - 7143,412 & 10 \leq x \leq 10,2 \\ -1,98876x^3 + 65,1663x^2 - 711,0850x + 2590,357 & 10,2 \leq x \leq 11 \\ -0,17713x^3 + 5,38249x^2 - 52,88848x + 177,7556 & 11 \leq x \leq 12,2 \\ 0,78453x^3 - 29,81421x^2 + 376,4958x - 1573,282 & 12,2 \leq x \leq 13,5 \\ -3,69910x^3 + 151,7727x^2 - 2074,635x + 9454,163 & 13,5 \leq x \leq 14 \\ 12,25912x^3 - 518,4728x^2 + 7308,803x - 34335,21 & 14 \leq x \leq 14,5 \\ -49,33042x^3 + 2160,673x^2 - 31538,33x + 153421,3 & 14,5 \leq x \leq 14,6 \end{cases}$$

$$5. \quad s(x) = \begin{cases} 0,01225x^3 + 0,00000x^2 - 0,50338x + 3,17000 & 0 \leq x \leq 3 \\ -0,02348x^3 + 0,32160x^2 - 1,46838x + 4,13481 & 3 \leq x \leq 5,5 \\ 0,000549x^3 - 0,15647x^2 + 1,21615x - 0,98896 & 5,5 \leq x \leq 9,5 \end{cases}$$

Unidades: cm.

6.

x (m)	50	150	250	350	450	550	650	750	850	950
h (m)	120,61	107,03	111,13	114,19	115,72	115,44	111,77	105,99	100,39	99,45

7.

t (min)	5	15	25	35	45	55	65	75
T ($^{\circ}$ C)	34,01	49,23	70,73	88,84	97,43	105,79	107,51	90,41

8.

Nodos del spline:

i	0	1	2	3	4	5	6	7	8	9
x_i (km)	0,00	3,57	5,00	10,00	11,25	15,00	20,00	25,00	28,89	30,00
y_i (km)	7,50	10,00	11,43	11,43	10,00	7,86	7,50	7,92	5,00	4,00

$$s(x) = \begin{cases} 0,01061x^3 + 0,00000x^2 + 0,56501x + 7,50000 & 0 \leq x \leq 3,57 \\ -0,06522x^3 + 0,81216x^2 - 2,33439x + 10,95029 & 3,57 \leq x \leq 5 \\ -0,00802x^3 - 0,04581x^2 + 2,09072x + 3,12421 & 5 \leq x \leq 10 \\ 0,15230x^3 + -4,85558x^2 + 50,39605x - 159,2778 & 10 \leq x \leq 11,25 \\ -0,02657x^3 + 1,18142x^2 - 17,75893x + 98,09346 & 11,25 \leq x \leq 15 \\ 0,00522x^3 - 0,24899x^2 + 3,81587x - 10,96690 & 15 \leq x \leq 20 \\ -0,01418x^3 + 0,91486x^2 - 19,46100x + 144,2122 & 20 \leq x \leq 25 \\ 0,01028x^3 - 0,91990x^2 + 26,38416x - 237,4356 & 25 \leq x \leq 28,89 \\ 0,00858x^3 - 0,77252x^2 + 22,26405x - 200,4110 & 28,89 \leq x \leq 30 \end{cases}$$

9.

$$\begin{bmatrix} \frac{1}{3} & \frac{1}{6} & 0 & 0 \\ \frac{1}{6} & 1 & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & 1 & \frac{1}{6} \\ 0 & 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \end{bmatrix} = \begin{bmatrix} -2 \\ \frac{3}{2} \\ -\frac{3}{2} \\ 0 \end{bmatrix}$$

10.

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & \frac{1}{6} \\ \frac{1}{6} & 1 & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & 1 & \frac{1}{6} \\ \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \end{bmatrix} = \begin{bmatrix} -2 \\ \frac{3}{2} \\ -\frac{3}{2} \\ 2 \end{bmatrix}$$

11. Spline para $x(t)$:

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix} = \begin{bmatrix} -2 \\ -5 \\ 4 \end{bmatrix} \quad M_0 = M_4 = 0$$

Spline para $y(t)$:

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \\ -3 \end{bmatrix} \quad M_0 = M_4 = 0$$

12. Spline periódico para $x(t)$:

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ M_2 \end{bmatrix} = \begin{bmatrix} 9 \\ -9 \\ 0 \end{bmatrix} \quad M_3 = M_0$$

Spline periódico para $y(t)$:

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ M_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \\ -12 \end{bmatrix} \quad M_3 = M_0$$

Sección 4.6

- Polinomio de grado 1: $p(x) = 0,26350x + 15,03919$; $D = 723,95$.
 Polinomio de grado 2: $p(x) = 1,20279x^2 - 9,87351x + 27,58344$; $D = 4,55$.
 Polinomio de grado 2: $p(x) = 0,00754x^3 + 1,10774x^2 - 9,57587x + 27,45678$; $D = 4,42$.
- Polinomio: $p(x) = -0,00103x^2 + 0,11732x + 6,62798$;
 Velocidad óptima: $56,95 \text{ km/h}$.

3. $f_r = 50,1 v^{1,60586}$

4. $P = 1172,6 + 1,3059 h$; P : Precipitación en *mm* de lluvia
 h : Altitud en *m*.

5. $L = 3,66 + 0,01083 v + 0,00868 v^2$ L : Distancia de frenado (*m*).
 v : Velocidad (*km/h*).

6. $R_0 = 0,017506$; $\alpha = 0,00388$.

7. Modelo: $y = 632,2 + 4,483t + 131,86 \cos\left(\frac{\pi}{6}t\right) + 140,12 \sin\left(\frac{\pi}{6}t\right)$
Pronóstico: Legarán 2758 grupos en el primer trimestre del próximo año.

9. b) $Q = 32,83963 s^{0,5374}$

10. $\delta_0 = 1,51007$; $\beta = 0,0001284$.

Sección 5.3

1. a) Mediante trapecios:

<i>n</i>	4	8	16	32	64
Resultado	0,77869	0,75728	0,75183	0,75046	0,75011
Error	-0,02869	-0,00728	-0,00183	-0,00046	-0,00011

b) Mediante Simpson:

<i>n</i>	4	8	16	32	64
Resultado	0,75199138	0,75014884	0,75000982	0,75000062	0,75000004
Error	-0,00199138	-0,00014884	-0,00000982	-0,00000062	-0,00000004

El error disminuye aproximadamente en 16 veces cuando el paso se reduce a la mitad.

2. 2,1556 ampere-hora.
3. 10,035919 unidades cúbicas.
4. 3,1415925
5. $115,2 m^2$. Error estimado: -0,142 m^2 .
6. a) $128/3 = 42,66667$
b) 42,875.
c) 42,66667.
7. 1,324684
8. 40,5825 *m*.
9. 22,103492.
10. $21500 cm^2$.
11. 186 megawatt-hora. Error estimado: -0,133 megawatt-hora.

Sección 5.4

1. 3,703874 unidades cuadradas.
2. El método de Gauss de 4 puntos.
3. Medir distancias verticales:

d_1 : a 0,422 m del centro.

d_2 : a 2,077 m del centro.

d_3 : a 4,500 m del centro.

d_4 : a 6,923 m del centro.

d_5 : a 8,578 m del centro.

$$A = 9[0,2369(d_1 + d_5) + 0,4786(d_2 + d_4) + 0,5689d_3]$$

4. 9,472851 unidades cúbicas.
5. 7,6403
6. 1,40217274; (Gauss de cuatro puntos). Error estimado: -0,000003.
7. Medir los diámetros:

d_1 : a 0,6762 cm del fondo.

d_2 : a 3,0000 cm del fondo.

d_3 : a 5,3238 cm del fondo.

$$V = \frac{3\pi}{4}(0,5556d_1^2 + 0,8889d_2^2 + 0,5556d_3^2)$$

8. 3,05911654
9. 0,69314717 (mediante Gauss de cuatro puntos).

Sección 5.5

1. 3,972620.
2. 2,003497.
3. 4,338322 unidades cuadradas.
4. 6,669944.
5. 0,3413448.
6. $a = 0,6745$.
7. 1,635409 unidades cuadradas.
8. 219,646 m.

Sección 5.6

1. 0,233178
2. 0,232963 (Simpson con $n = 16$).
3. 0,869844.
4. Mediante Gauss de tres puntos se obtiene 0,869842.
5. 23,415661.
6. 23,0911; Error estimado: 0,032.
7. 1,903826.
8. 0,285399.

Sección 6.2

1. 0,6667

2. 0,3409
3. 1,4616
4. Radio: 5,364 cm; altura: 11,063 cm.
5. $a = 0,978$; $b = 1,478$.

Sección 6.3

1. a) 0,7854
b) 2,3470
c) 0,5671
d) 4,8105
2. 1,4337
3. Cada cateto: 3,162278
4. Base: 1,7207; altura: 0,6522
5. $y = 1,0286x$

Sección 6.5

1. a) Mínimo en (3,7500; 3,5000). Se obtiene en tres iteraciones partiendo de A o de B .
b) Mínimo en (3,7500; 1,0500). Se obtiene en tres iteraciones partiendo de A o de B .
c) Mínimo en (2,6153; 2,7884). Se obtiene en 11 iteraciones partiendo de A o de B .
d) Mínimo en (5,0858; 2,0515). Se obtiene en 23 iteraciones partiendo de A o de B .
2. Tomando $\mathbf{x}_0 = (1,3; 0,8)$ se obtiene: Mínimo en (1,3182; 0,8608).
Tomando $\mathbf{x}_0 = (2,2; 1,6)$ se obtiene: Mínimo en (2,2355; 1,5354).
3. $y = 0,8000x + 1,7000$
4. Distancia mínima: 3,9560
5.
$$\frac{x}{6,3001} + \frac{y}{2,7000} + \frac{z}{3,3000} = 1$$

Sección 6.6

1. a) Mínimo en (3,7500; 3,5000). Se obtiene en cinco iteraciones partiendo de A y en siete iteraciones partiendo de B .
b) Mínimo en (3,7498; 1,0500). Se obtiene en 11 iteraciones partiendo de A y en 26 iteraciones partiendo de B .
c) Mínimo en (2,6154; 2,7884). Se obtiene en ocho iteraciones partiendo de A y en cuatro iteraciones partiendo de B .
d) Mínimo en (5,0860; 2,0516). Se obtiene en 32 iteraciones partiendo de A y en cinco iteraciones partiendo de B .
2. Tomando $\mathbf{x}_0 = (1,3; 0,8)$ se obtiene: Mínimo en (1,3180; 0,8607).
Tomando $\mathbf{x}_0 = (2,2; 1,6)$ se obtiene: Mínimo en (2,2355; 1,5354).
3. $g(x) = 1,095 e^{0,0720x}$
4. Diámetro: 0,7604

Sección 6.7

1. a) Mínimo en $(3,7500; 3,5000)$. Se obtiene en tres iteraciones partiendo de A o de B .
b) Mínimo en $(3,7498; 1,0500)$. Se obtiene en tres iteraciones partiendo de A o de B .
c) Mínimo en $(2,6154; 2,7885)$. Se obtiene en siete iteraciones partiendo de A o de B .
d) Mínimo en $(5,0860; 2,0516)$. Se obtiene en siete iteraciones partiendo de A o de B .
2. Tomando $\mathbf{x}_0 = (1,3; 0,8)$ se obtiene: Mínimo en $(1,3182; 0,8608)$.
Tomando $\mathbf{x}_0 = (2,2; 1,6)$ se obtiene: Mínimo en $(2,2355; 1,5354)$.
3. Distancia mínima: 2,6622
4. $P = (3,0868; 2,4551)$
5. En el punto $(0,4357; 0,3048; 1,6360)$

Sección 6.8

1. a) Mínimo en $(3,7500; 3,4998)$ en 59 evaluaciones partiendo de A . Mínimo en $(3,7497; 3,5000)$ en 64 evaluaciones partiendo de B .
b) Mínimo en $(3,7497; 1,0500)$ en 107 evaluaciones partiendo de A . Mínimo en $(3,7508; 1,0501)$ en 146 evaluaciones partiendo de B .
c) Mínimo en $(2,6154; 2,7885)$ en 57 evaluaciones partiendo de A . Mínimo en $(2,6152; 2,7883)$ en 45 evaluaciones partiendo de B .
d) Mínimo en $(5,0861; 2,0516)$ en 57 evaluaciones partiendo de A . Mínimo en $(5,0862; 2,0516)$ en 66 evaluaciones partiendo de B .
2. Tomando $\mathbf{x}_0 = (1,3; 0,8)$ se obtiene: Mínimo en $(1,3181; 0,8609)$.
Tomando $\mathbf{x}_0 = (2,2; 1,6)$ se obtiene: Mínimo en $(2,2356; 1,5356)$.
3. $g(x) = 2,9911 - 2,0599 e^{-0,8346 x}$
4. Distancia más corta: 4,0980

Sección 7.1

1. Ecuación diferencial homogénea. Mediante el cambio de variables $y = ux$ se transforma en una ecuación de variables separables.
2. No pertenece a ninguno de los tipos usuales.
3. Ecuación diferencial lineal de primer orden. Multiplicando por e^x se hace exacta.
4. No pertenece a ninguno de los tipos usuales.
5. Ecuación diferencial de Bernoulli. Mediante el cambio de variables $v = 1/y$ puede ser resuelta.
6. Mediante el cambio de variables $x = X + h$; $y = Y + k$, escogiendo h y k adecuadamente, se transforma en una homogénea que mediante un cambio de variables (ver ejercicio 1) se convierte en variables separables.
7. No pertenece a ninguno de los tipos usuales.
8. Ecuación diferencial lineal con coeficientes constantes. Puede resolverse por el método de variación de parámetros.
9. Ecuación diferencial lineal con coeficientes variables. En este caso los coeficientes se presentan de tal modo que la ecuación se reduce a una de coeficientes constantes mediante el cambio de variables $x = e^z$ (Euler).

10. Ecuación diferencial lineal con coeficientes variables. Puede resolverse mediante series de potencias. Para valores de n enteros y no negativos, las soluciones son los polinomios de Legendre. Se denomina ecuación de Legendre.
11. Ecuación diferencial lineal de segundo orden. No corresponde a ninguno de los tipos usuales.
12. Ecuación diferencial no lineal de segundo orden. Como la variable x no aparece, la ecuación puede ser resulta mediante 1 cambio de variables: $y' = v$; $y'' = \frac{dy}{dx} = \frac{dv}{dy} \frac{dy}{dx} = \frac{dv}{dy} v$, el cual la transforma en variables separables.
13. Condiciones de frontera.
14. Condiciones iniciales.
15. Condiciones iniciales.
16. Condiciones de frontera.
17. Condiciones iniciales.
18. Condiciones iniciales.
19. Condiciones iniciales.
20. Condiciones de frontera.

Sección 7.2

1. $\frac{dy}{dx} = \frac{x+y}{y-x}$
2. $\frac{dy}{dx} = \frac{\cos x}{x} - \frac{3y}{x^2}$
3. $\frac{dy}{dx} = (3x+2)^2 - y$
4. $\frac{dy}{dx} = \frac{1}{2}(x + \sqrt{x^2 - 4})$ ó $\frac{dy}{dx} = \frac{1}{2}(x - \sqrt{x^2 - 4})$

Sección 7.3

1. a) $f_y(x, y)$ es negativa para $2 \leq x \leq 6$. En ese intervalo la ecuación es estable.
c)

x	Euler		RK2		RK4	
	y	Error	y	Error	y	Error
2	1,0000	0,0000	1,0000	0,0000	1,0000	0,0000
3	2,0466	-0,0004	2,0460	0,0003	2,0460	0,0004
4	2,8757	-0,0002	2,8754	0,0002	2,8754	0,0002
5	3,6872	-0,0001	3,6870	0,0001	3,6870	0,0001
6	4,5104	-0,0001	4,5104	0,0001	4,5103	0,0001
	$h = 0,0025$		$h = 0,1$		$h = 0,5$	
	1600 evaluaciones		80 evaluaciones		32 evaluaciones	

2. b) Mediante RK4 con $h = 0,25$:

<i>x</i>	<i>y</i>	Error
0	4,000000	0,00000
1	2,148456	- 0,00005
2	1,786310	- 0,00003
3	1,724191	- 0,00002
4	1,762076	- 0,00002
5	1,843780	- 0,00001

c) Mediante RK4 con $h = 0,25$:

<i>x</i>	<i>y</i>
0	4,001000
1	2,148870
2	1,786548
3	1,724355
4	1,762199
5	1,843881

3. a) Mediante RK4 con $h = 0,1$ b) Mediante RK4 con $h = 0,1$

<i>x</i>	<i>y</i>
0	0,999000
1	1,997282
2	2,992611
3	3,979915
4	4,945402
5	5,851587

<i>x</i>	<i>y</i>
0	1,001000
1	2,002718
2	3,007389
3	4,020085
4	5,054598
5	6,148413

6. $y = 0,891802$

7. b) $P(10) = 89,9$

8. Mediante RK4 con $h = 0,5$

<i>x</i>	<i>y</i>
0	2,0000
1	1,3398
2	0,9348
3	0,6616
4	0,4712
5	0,3366
6	0,2408

9. Mediante RK4 con $h = 0,1$

x	y
-1	-1,0000
-0,8	-0,9896
-0,6	-0,9565
-0,4	-0,8960
-0,2	-0,7993
0	-0,6448

Sección 7.4

1. Con $h = 0,05$

x	y
1	0,800000
1,4	1,049434
1,8	1,454869
2,2	1,917452
2,6	2,374858
3,0	2,814269

2. Con $h = 0,0125$

x	y
1	0,800000
1,1	0,989912
1,2	1,246322
1,3	1,603810
1,4	2,130129
1,5	2,979916

3. Con $h = 0,1$

x	y
0	1,00000
0,4	1,39666
0,8	1,78357
1,2	2,16708
1,6	2,55795
2,0	2,96741

4. Con $h = \pi/48$

θ	ρ
$\pi/6$	2,000000
$\pi/4$	1,844027
$\pi/3$	1,766722
$5\pi/12$	1,733777
$\pi/2$	1,725678

5. 1394,030 cm/s.

6. $y(3) = 0,666664$

Sección 7.5

1. $v(0,6) = -0,389978$; $w(0,6) = 0,860694$; error menor que 0,000001.
2. Mediante RK4, con $h = 0,05$: $x(2) = 1,04244$; $y(2) = 0,43985$.
3. Mediante RK4, con $h = 0,1$:

x	y
0	1,0000
0,2	1,2621
0,4	1,4775
0,6	1,6681
0,8	1,8643
1,0	2,1012

4. Mediante RK4, con $h = 0,05$:

x	y
1,0	2,00000
1,5	3,49698
2,0	4,38663
2,5	4,10880
3,0	3,17926

5. Mediante RK4, con $h = 0,025$:

x	y
0,0	1,00000
0,5	1,29613
1,0	2,28533
1,5	3,13088
2,0	2,08273

6. Mediante RK4, con $h = 0,05$:

x	y	z
1,2	1,0000	0,5000
1,5	1,5613	1,8695
1,8	1,8880	5,3769
2,1	1,9487	16,9281

7.

$$\begin{cases} \frac{du_1}{dt} = u_2 \\ \frac{du_2}{dt} = -\frac{km_s u_1}{(u_1^2 + u_3^2)^{3/2}} \\ \frac{du_3}{dt} = u_4 \\ \frac{du_4}{dt} = -\frac{km_s u_3}{(u_1^2 + u_3^2)^{3/2}} \end{cases} \quad \begin{array}{l} u_1(0) = r_0 \\ u_2(0) = v_{0x} \\ u_3(0) = 0 \\ u_4(0) = v_{0y} \end{array} \quad \text{donde: } \begin{array}{l} u_1 = x \\ u_2 = \frac{dx}{dt} \\ u_3 = y \\ u_4 = \frac{dy}{dt} \end{array}$$

8. Dentro de 15 años: 3562 alces y 3 lobos. Dentro de 28 años: 7540 alces y 101 lobos. Se utilizó RK4 con $h = 0,25$.
9. El bloque se ha desplazado 38,514 cm desde su posición inicial, de modo que se encuentra a 18,514 cm de la posición de equilibrio. Su velocidad es de 0,0109 m/s acercándose a la pared. Se utilizó RK4 con $h = 0,05$.

Sección 7.6

1. Mediante RK4 con $h = 0,05$:

x	y	y'
1,0	0,0000	2,0838
1,2	0,4342	2,2450
1,4	0,8920	2,3172
1,6	1,3547	2,2933
1,8	1,8034	2,1805
2,0	2,2223	2,0000

2. Mediante RK4 con $h = 0,05$:

x	y	y'
0,0	1,0000	-0,3809
0,4	0,8705	-0,2116
0,8	0,8671	0,2264
1,2	1,0639	0,7565
1,6	1,4605	1,2007
2,0	2,0000	1,4670

3. Mediante RK4 con $h = 0,05$: $y(0) = 0,00333$; $y'(0) = 2,2965$

4. Mediante RK4 con $h = 0,05$. $w(0) = 1,2914$ rad/s.

t	θ
0	0,0000
0,1	0,1281
0,2	0,2499
0,3	0,3597
0,4	0,4523
0,5	0,5236

5. Inicialmente había unos 83 u 84 lobos. Actualmente solo quedan 4.