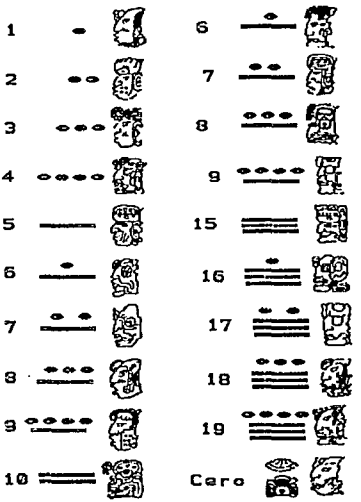
VNIVERSIDAD NACIONAL
AVTÓNOMA DE
MEXICO

Universidad Nacional Autónoma de México

Facultad de Ciencias

1
2 ej.

"Notas para un curso de análisis numérico para estudiantes de ingeniería"



TESIS
que para obtener el título de:
MATEMATICO
Presenta: *Alberto Alonzo Alvarez*

Julio de 1992

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

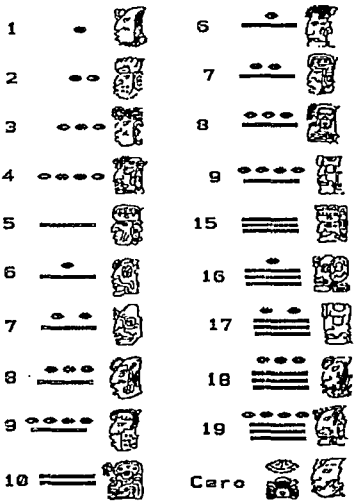
VNIVERSIDAD NACIONAL
AVFNOMA DE
MEXICO

Universidad Nacional Autónoma de México

Facultad de Ciencias

1
2 ej.

"Notas para un curso de análisis numérico para estudiantes de ingeniería"



TESIS
que para obtener el título de:
MATEMATICO
Presenta: *Alberto Alonzo Alvarez*

Julio de 1992

PRESENTACIÓN

¿Por qué unas notas para un curso de análisis numérico para estudiantes de ingeniería? Se preguntarán. ¿No hay acaso en el mercado nacional textos que se puedan adaptar casi a cualquier curso, en virtud del amplio rango que abarcan? ¡Es cierto! En el mercado nacional existen textos de análisis numérico; desde los clásicos, como los de Conte / de Boor, Ralston, Isaacson y Henrici, hasta los breves, introductorios y tal vez menos conocidos, como los de Yakowitz / Szidarovszky, Scraton y Atkinson; con énfasis en la teoría, en las aplicaciones o en el análisis del software; o con combinaciones de aplicaciones y software, etc. Esto es, aparentemente bastaría buscar en el amplio mercado de textos de análisis numérico, para encontrar el adecuado a un curso específico.

En la práctica, todos los profesores lo sabemos, no existe "*el*" texto ideal que se ajuste a las condiciones específicas de un grupo de alumnos, profesor y programa en un tiempo e institución dados. Es por ello que todos los profesores elaboramos notas para cada clase, y son excepcionalmente raros -creo- los maestros que a nivel de licenciatura imparten sus cursos siguiendo total y literalmente un libro de texto. Generalmente, adecuamos los temas a nuestras particulares tendencias y tratamos de ajustarnos a los programas y condiciones del grupo, lo que nos obliga a usar enfoques o desarrollos de diversos autores.

En mi caso, aprovechando la oportunidad que he tenido de impartir cursos de análisis numérico, en la carrera de Ingeniería Industrial, durante varios semestres, y en virtud de que esta área, el análisis numérico, no constituyó parte del curriculum en mi formación como matemático, en la facultad de Ciencias de la UNAM, surgió en mí el interés por el estudio más sistemático del análisis numérico y, posteriormente, la necesidad de escribir notas que se adecuaran a lo que creo deben aprender los alumnos de mis cursos. Esa fue la motivación básica que originó el presente trabajo.

Por otro lado, debo confesar mi debilidad por la edición de materiales para la enseñanza de la matemática, que se inicia al principio de la década de los setentas y culmina en los ochentas, con la edición de los libros de texto para escuelas secundarias del estado de México, casi gratuitos, editados por el Gobierno del Estado de México.

Como docente de la matemática durante más de veinte años en distintos niveles, he aprendido que todos los textos son incompletos y, en consecuencia, perfectibles; dependiendo del sujeto que los analiza y del contexto en el que se hace dicho análisis.

Es por ello que espero que este texto se vaya adecuando y, por lo tanto, mejore con las observaciones que se le hagan en el transcurso de los años.

Finalmente, antes de hacer una breve descripción del contenido del libro, quiero dejar constancia de agradecimiento al Dr Pablo Barrera, "culpable" de mi adicción a la docencia, a la difusión y a la matemática aplicada.

En estas notas se desarrollan los temas del Programa de algoritmos cumputacionales que se imparte en la División de Ingeniería y Ciencias del ITESM, Campus Toluca, a excepción del de ecuaciones diferenciales ordinarias. En vitud de que, en general, cuando el alumno cursa esa materia, no lo ha hecho con la de ecuaciones diferenciales, o lo hará de manera paralela.

Por otro lado, hemos considerado de mayor utilidad que se profundicen otros temas, a los que se dedique más tiempo, mismos que pueden dar una mayor comprensión de la filosofía y técnicas del análisis numérico, como en el caso del capítulo IV (sistemas de ecuaciones).

El texto se ha dividido en seis capítulos, en el I se desarrolla el concepto básico de sistema numérico de punto flotante (típico de los dispositivos digitales) y su repercusión en los errores. Se presenta, como antecedente y marco de referencia, una caracterización de problema y de problema numérico. Se concluye el capítulo con un bosquejo histórico, que pretende enfatizar raíces y antecedentes históricos del cómputo numérico.

La evaluación de funciones del capítulo II pretende introducir al alumno en los problemas de la elección, del "mejor" método, a la que no está acostumbrado en virtud de que en general existe la idea de que para resolver un problema matemático sólo hay un método y, en el caso de que haya más de uno, la selección no es trascendente con respecto a los resultados obtenidos.

El capítulo III (solución de ecuaciones) nos parece que complementa el panorama sobre filosofía y técnicas de los métodos numéricos, proporcionando los fundamentos para el desarrollo de los capítulos IV y V, que consideramos el corazón o parte medular tal vez más densa, del curso.

En el capítulo V (interpolación), presentamos diversos métodos, tomando como punto de partida las diferentes bases de un espacio vectorial. Este enfoque es poco común. Pero consideramos que es conveniente y acce-

sible a la comprensión del alumno, con la ayuda del profesor. La conveniencia de este enfoque estriba en que, a través del mismo, es posible dar a los alumnos elementos de juicio para enfatizar la diferencia de métodos, pero no de resultados, ya que todos producen polinomios del mismo grado, pero sobre diferentes bases. Es común que el estudiante piense que los polinomios de interpolación son diferentes por el hecho de tener representaciones diferentes.

Terminamos con la integración numérica, en ella se presenta un desarrollo que consideramos clásico, y seguimos las líneas de desarrollo del libro de Shampine en la sección final de cuadratura adaptativa.

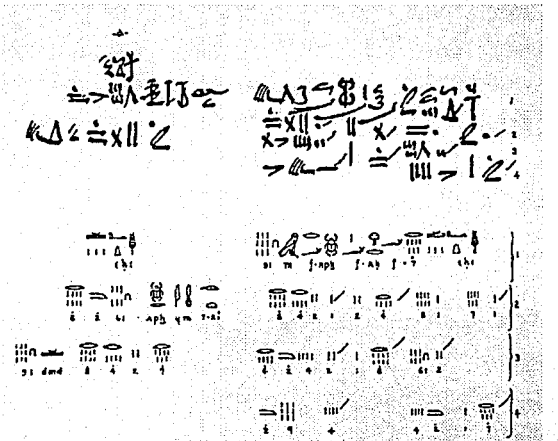
Se incluyen algunos programas de BASIC, cuyo objetivo es presentar un programa que resuelve un ejemplo concreto. Creemos que esto le permite al estudiante comparar el método con el programa propuesto, ya que si bien desde el punto de vista de la programación científica el BASIC no es el mejor lenguaje, por su estructura permite al estudiante comprender la lógica del programa.

Además, existe en el mercado una gran variedad de paquetes de análisis numérico elaborados por profesionales que, obviamente, son los adecuados para el uso práctico.

INDICE

CAPÍTULO I

INTRODUCCIÓN



Problema 24 del papiro del Rhind: "Si el montón (aha) y un séptimo del montón sumados juntos hacen 19. ¿Cuál es el montón?

Suponga que es 7	1 7
	1/7 .. 1
	Total 8
	1 8
	2 16
	1/2 .. 4
	1/4 .. 2
	1/8 .. 1
Total 21/4 1/8	
	1 ..... 1/4 1/8
	2 ..... 4 1/2 1/4
	4 ..... 9 1/2
La cantidad es:	16 1/2 1/8
	1/7 .. 2 1/4 1/8 }19

"La solución de problemas es la característica esencial del pensamiento voluntario."

WILLIAM JAMES

"Divida cada problema en tantas partes como pueda para resolverlo más fácilmente."

RENÉ DESCARTES

PROBLEMAS

Obtener comida generalmente no es problema en la vida moderna. Si tenemos hambre y estamos en casa, tomamos algo de la cocina o vamos a algún restaurante. Sin embargo, si no hay nada de comer en la cocina o no tenemos dinero, la situación cambia radicalmente; en tal caso, la obtención de la comida se convierte en problema. En general, un deseo puede o no conducir a un problema. Si el deseo trae a nuestra mente -de manera inmediata y sin dificultad implícita en la obtención del satisfactor- alguna acción obvia para lograr el objeto deseado, entonces no hay problema.

El problema es un gran problema si es muy difícil; en cambio, sólo es un pequeño problema si nada más tiene una dificultad pequeña. El concepto de dificultad es inherente a la noción de problema: si no existe dificultad no hay problema.

Problema típico es el de encontrar el camino que nos conduzca a un lugar predeterminado en alguna región. Seguramente, ése fue un problema serio y difícil para el hombre primitivo. Eso puede o no ser la razón de que la solución de cualquier problema nos parezca algo así como encontrar un camino que elimine la dificultad, un camino alrededor del obstáculo.

Gran parte de nuestro pensamiento consciente tiene que ver con los problemas. Generalmente nuestros pensamientos están dirigidos hacia algún fin, buscamos formas y medios para alcanzarlo, tratamos de pensar en algo que nos pueda conducir a lograr nuestro objetivo.

Existen problemas cotidianos; la consecución de la comida; personales, ¿con quién ir al cine?; sociales, ¿cómo mejorar la distribución de la riqueza en México?; científicos, ¿cómo obtener una vacuna contra el sida?; de entretenimiento, ¿cómo mejorar mi habilidad para jugar ajedrez?, etc.

En este contexto, podemos caracterizar a un *PROBLEMA* como una función en la que el dominio de la misma está conformado por datos, y el codominio, por las posibles soluciones.[1]

Ejemplo: Consideremos el problema de diagnosticar una enfermedad en un paciente. Los datos son los síntomas y las posibles soluciones son las enfermedades que puede padecer el enfermo.

Si los síntomas son debilidad general, dolor de cabeza, diarrea, abdomen sensible, tos, enrojecimiento de la cara y fiebre, es probable que el médico proponga como solución (diagnostique) tifoidea.

[1]Vandergraft, James S. Introduction to Numerical Computations. Academic. Press. 1978.

En ocasiones, las propuestas de solución (diagnósticos) no son las adecuadas y algunos han sufrido las consecuencias de ello.

La solución

En el marco de referencia establecido -esto es, si aceptamos como modelo matemático de problema el concepto de función- resolver un problema significa:

a) Encontrar la regla de correspondencia entre datos y soluciones posibles.

b) Dada la regla de correspondencia, evaluar la función en el subconjunto dado de datos, para encontrar la imagen de éstos, que es la solución del problema.

Por ejemplo, resolver el problema $ax + b = c$ es encontrar $f(a, b, c) = x$, la imagen de los datos, donde $ax + b = c$

El problema es, entonces, la función que tiene como dominio y codominio a los reales, y que asocia a la terna (a, b, c) el número x, bajo la condición $ax + b = c$

Un método para resolver este problema es el siguiente:

$$(a, b, c) \rightarrow (a, c\text{-}b) \rightarrow \left(\frac{c-b}{a}\right)$$

a la terna (a, b, c) le asociamos la pareja $(a, c\text{-}b)$ y a ésta el real $\frac{c-b}{a}$

otro método es:

$$(a, b, c) \rightarrow \left(a, \frac{b}{a}, c\right) \rightarrow \left(\frac{b}{a}, \frac{c}{a}\right) \rightarrow \frac{c}{a} - \frac{b}{a}$$

a la terna (a, b, c) le asociamos la terna $(a, \frac{b}{a}, c)$, a esta la pareja $\left(\frac{b}{a}, \frac{c}{a}\right)$ y, finalmente, el real $\frac{c}{a} - \frac{b}{a}$

Una observación importante que se desprende del ejemplo anterior es que: *el método, procedimiento o algoritmo para resolver un problema.. ¡no es único!*

En el ejemplo usamos dos métodos, ambos descomponen la función en funciones más sencillas, lo que parece indicar que el método de solución de un problema consiste en descomponer el problema (la función) en problemas (funciones) más sencillos o elementales, donde lo elemental o sencillo es un concepto relativo con respecto al tiempo y a las personas. Así, el problema de encontrar la raíz de un polinomio cuadrático es "sencillo" para un estudiante de ingeniería en la actualidad, pero no lo fue para la mayor parte de los habitantes de Mesopotamia hace 4000 años.

Problemas matemáticos

En el caso de que los elementos del dominio y codominio sean objetos matemáticos, por ejemplo, figuras geometricas, funciones, conjuntos, expresiones algebraicas, etc., diremos que tenemos un problema matemático.

Ejemplos:

1. Encontrar la solución de $2x^2\text{-}6x\text{-}20 = 0$

El conjunto de datos está formado por $\{2, -6, -20\}$ que es un subconjunto del dominio constituído por los números reales. La solución o soluciones, en este caso $\{5, -2\}$, es un subconjunto de los números reales también.

2. Dados $a, b, c,$ números reales, encontrar $x,$ tal que $ax^2 + bx + c = 0$

En este caso, el conjunto de datos pertenece al domino de los reales; las posibles soluciones al conjunto de los complejos; la condición está dada por la ecuación $ax^2 + bx + c = 0$

3. Determinar la suma de 3 y 5. Datos = {3, 5} $\subset \Re$ = dominio, codominio = $\Re$. Si c es la solución, entonces $3 + 5 = c$

4. Dados dos números reales α, β calcular $\alpha+\beta, \alpha-\beta, \alpha\cdot\beta, \alpha/\beta$

En todos estos casos

$$D = \Re x \Re, \qquad C = \Re$$

$$D = \text{dominio}; \qquad C = \text{codominio}$$

$$S(\alpha,\beta) = \alpha+\beta; \; P(\alpha, \beta) = \alpha \cdot \beta$$

$$R(\alpha,\beta) = \alpha-\beta; \; D(\alpha, \beta) = \alpha/\beta$$

Estos cuatro problemas, que pueden parecer triviales para quien ha logrado cierta preparación, no resultan así para quienes inician apenas su formación escolar y -mucho menos- para la mayoría de los seres humanos de hace tres mil años. Si ustedes recuerdan, encontrarán que primero aprendieron a sumar y multiplicar los números {0, 1, 2,..., 9} y, a partir de esto, aprendieron a reducir cualquiera de las operaciones entre dos números reales a combinaciones de sumas y productos de los números {0, 1, 2,..., 9}

Problemas numéricos

En el párrafo anterior, indicamos que problema es una función y que problema matemático es una función cuyo dominio y codominio están constituídos por objetos matemáticos. Si el dominio y el codominio son subconjuntos númericos de espacios vectoriales de dimensión finita, entonces tenemos un problema numérico que resolveremos con un método numérico.

Los problemas numéricos son problemas matemáticos que tienen una característica adicional: la finitud. Es decir, el dominio y contradominio del problema deben poderse describir en forma finita.

De manera más precisa, podemos ahora decir que: *Problema numérico es un problema matemático donde la función asociada está definida en un subconjunto de $\Re^n$ y toma valores en un subconjunto de $\Re^m$. Es decir, en todo problema numérico tenemos:*

$$F: A \to B \quad \text{tal que:} \quad A \subset \Re^n, B \subset \Re^m$$

Resolver un problema numérico es encontrar la imagen de los datos bajo la función; esto es, dado $a \in A$, calcular $b \in B$, tal que $F(a) = b$

Puede suceder que un problema numérico necesite ser reformulado para que sea posible resolverlo.

Consideremos el siguiente ejemplo con todo detalle, ya que esto nos permitirá apreciar claramente la afirmación anterior:

Dado $x > 0$, calcule y_0 tal que $y_0 = \sqrt{x}$

Un método usual de solución consiste en construir una sucesión $\{z_n\}$ que converja a $\sqrt{x}$. Este procedimiento requiere de un límite y_0. Esto no es conveniente en la práctica. Lo usual es tomar una buena aproximación y a $\sqrt{x}$. Lo que necesitamos para poder reformular el problema es precisar qué quiere decir que y sea una buena aproximación de $\sqrt{x}$, y esto usualmente quiere decir que y tenga sus "primeras cifras decimales correctas"; es decir, que y y $\sqrt{x}$ tengan las mismas primeras cifras. Como veremos más adelante, esto se puede escribir pidiendo que y satisfaga

$$\frac{|y - \sqrt{x}|}{|\sqrt{x}|} \leq 10^{-k}$$

donde k es el número de cifras "correctas" que se desean. Así la nueva formulación es la siguiente:

Dado $x > 0$ calcule $y \in R$ tal que

$$|y - \sqrt{x}| < 10^{-k} |\sqrt{x}|$$

Un método numérico usual, derivado del anterior, consiste en escoger el primer término de la sucesión que tenga la propiedad buscada; es decir, tomar como y a la primera z_k, tal que:

$$|z_k - \sqrt{x}| \leq 10^{-k} |\sqrt{x}|$$

pero aún no podemos usar este criterio, ya que requiere de $\sqrt{x}$, que es lo que queremos calcular. Como se verá, dicho criterio se puede sustituir por el siguiente:

$$|z_k - z_{k-1}| < 10^{-k} |z_k|$$

ya que, a partir de cierto momento, $z_k \approx \sqrt{x}$. Está nueva formulación es más adecuada en la práctica; sin embargo, para que el método de solución sea "bueno", es importante que la sucesión $\{z_k\}$ converja rápidamente a $\sqrt{x}$, es decir, que sólo sea necesario calcular un número muy pequeño de aproximaciones para obtener la y que deseamos.

Usualmente en el problema numérico se calcula b tal que $F(\mathbf{a}) = \mathbf{b}$ donde $F: R^n \rightarrow R^m$. A $\mathbf{a}$ se le llama vector de datos o de entrada y a $\mathbf{b} = F(\mathbf{a})$, vector de salida o de resultados.

Algoritmo de solución o método numérico de un problema numérico es la descripción completa de operaciones bien definidas para resolverlo.

SISTEMAS DE NUMERACIÓN

ANTECEDENTES

Los conceptos primigenios de la matemática son, sin lugar a dudas, número, magnitud y forma. El primero de éstos dio lugar a la *logística* y *aritmética* griegas; aquélla es el antecedente inmediato del cálculo numérico.

Según Boyer[1], el concepto de número en la especie humana es tan antiguo como el uso del fuego; esto es, data de unos 400 000 años. Por otro lado, es posible afirmar que el desarrollo de tal concepto se efectuó de manera gradual; de forma similar, probablemente, al desarrollo del lenguaje y la escritura. Para la mayoría de los antropologos y etnólogos, la escritura tuvo un desarrollo posterior al del lenguaje; por ello, se supone que la representación de números es posterior al concepto de número.

El indicador más antiguo de un registro numérico -de que se tenga noticia- fue encontrado en Checoslovaquia, donde se descubrió un hueso de lobo con 55 insiciones agrupadas de cinco en cinco, con una antigüedad aproximada de 30 000 años. Es posible, como lo observó Aristóteles, que los agrupamientos de cinco en cinco o de diez en diez (sistema decimal), sean el resultado del "accidente anatómico de que la mayor parte de nosotros nacemos con cinco dedos en cada mano y pie". Aunque, como lo señala Boyer, tal vez hubiese sido mejor, desde el punto de vista matemático, que el hombre hubiese tenido cuatro o seis dedos en cada mano, ya que eso hubiese dado lugar a sistemas de numeración más funcionales.

De algunos pueblos sólo se tiene referencia de la forma oral que emplearon para expresar números (matlatzincas); de otros, además de la oral, se conoce la forma escrita, la cual imitaba en algunos casos objetos de la vida real (sistemas egipcio y azteca).

Desde la antigüedad, la humanidad se ha valido de diferentes símbolos para representar números; sin embargo, no es suficiente tener símbolos para representar números, ya que sería necesaria una gran cantidad de ellos si para cada número usaramos un símbolo: es necesario crear reglas que permitan combinarlos.

De esta manera, con una cantidad finita de símbolos es posible representar todos aquellos números que sean necesarios.

Así, con un conjunto finito de símbolos y reglas o principios, se han creado diversos sistemas de numeración, que permiten representar números.

[1] Boyer, C. B. A History of Mathematics. John Wiley. New York. 1968.

Sistemas de numeración posicional

Las operaciones (suma, resta, multiplicación, división) que hacemos con los números están íntimamente vinculadas con el sistema numérico usado. Seguramente, para el lector será fácil multiplicar 372.38 por 24.858, usando las propiedades del sistema de numeración decimal; le sugerimos que intente hacer lo mismo con el sistema egipcio o el romano.

La notación posicional de base b está definida por la regla:

(1)

$$(a_n \ldots a_1 a_0 . a_{-1} a_{-2} \ldots)_b =$$

$$= a_n b^n + \ldots + a_1 b + a_0 + \frac{a_{-1}}{b} + \frac{a_{-2}}{b^2} + \ldots$$

Por ejemplo:

$$(520.3)_6 = 5 \cdot 6^2 + 2 \cdot 6 + 0 + \frac{3}{6} = 192\frac{1}{2}$$

Nuestro sistema decimal es, por supuesto, un caso especial, donde b es igual a diez; en este caso, el subíndice b en (1) se omite.

Las generalizaciones más sencillas del sistema decimal se obtienen cuando tomamos a b como un entero mayor que uno y pedimos que las áes sean enteros en el rango $0 \le a_k \le b$. Esto nos da los sistemas binario ($b=2$), ternario ($b=3$), cuaternario ($b=4$), quinario ($b=5$), etc. En general, podemos tomar a b como cualquier número diferente de cero y podemos escoger las áes de un conjunto cualquiera de números.[1]

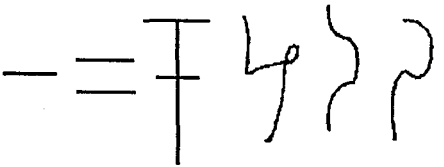
Las áes en (1) se llaman dígitos de la representación. El dígito a_k para una k grande -se dice- es más significativo que el dígito a_k para una k pequeña; de acuerdo con eso, se dice que el primer dígito de la izquierda es el dígito más significativo y el último dígito de la derecha, el dígito menos significativo. En el sistema binario estándar, a menudo los dígitos binarios son denominados bits; en el sistema hexadecimal (base 16), los dígitos se denotan generalmente de la siguiente manera 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F.

Sistema decimal

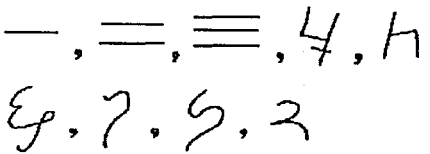
El sistema decimal de numeración emplea símbolos que han sufrido transformaciones desde que aparecieron por primera vez. Así, en las paredes de una cueva situada en una colina llamada Mana Ghat, en la India,

[1] Knuth, D. E. The Art of Computer Programing. Vol. 2. Addison-Wesley Reading Ma. 1981.

se encontraron las siguientes representaciones que datan del siglo III a. C.; representan a los números 1, 2, 4, 6, 7, 9.



En Nasik, también en la India, se descubrieron los siguientes símbolos que representan a los nueve dígitos:



Los árabes aprendieron ese sistema y lo difundieron en Europa; donde se usaron, a partir del siglo XV.

Por su origen y difusión, al sistema decimal de numeración también se le conoce con el nombre de indo-arábigo.

Mediante combinaciones de los símbolos para los diez dígitos, y usando los principios aditivo y posicional, se puede representar cualquier número; así, $125 = 100+20+5. = 1(10)^2+2(10)+5(10)^0$

Sistema binario

El sistema de notación binaria tiene su propia e interesante historia. Se sabe que muchas tribus primitivas de la actualidad usan un sistema binario de conteo (efectuando agrupamientos de dos en lugar de cinco o diez); pero no cuentan en un verdadero sistema binario, ya que no tratan a las potencias de dos de manera específica.

Las unidades de medida inglesas, del siglo XIII, para líquidos son las siguientes:

2 gills = 1 chopin	2 demibushels = 1 bushel or firkin
2 chopins = 1 pint	2 firkins = 1 kilderkin
2 pints = 1 quart	2 kilderkins = 1 barrel
2 quarts = 1 pottle	2 barrels = 1 hogshead
2 pottles = gallon	2 hogsheads = 1 pipe
2 gallons = 1 peck	2 pipes = 1 tun
2 peck = 1 demibushel	

Tal vez, como lo indica Knuth[1], los verdaderos inventores del sistema binario fueron los vinateros ingleses.

La primera noticia que tenemos sobre la notación binaria como tal, data de 1605, ésta aparece en un trabajo de Thomas Harriot (1560-1621)

El primer trabajo publicado sobre el sistema se debe al obispo español Juan Caramuel Lobkowitz, quien desarrolla en el mismo trabajo, *Mathesis biceps I* los sistemas numéricos con bases 2, 3, 4, 5, 6, 7, 8, 9, 10, 12 y 60

Notación científica

1. ¿Sabes cuál es la masa de la Tierra en gramos? ¿Y la de un electrón? Las respuestas a estas preguntas son, respectivamente:

5 980 000 000 000 000 000 000 000 gramos

0.000 000 000 000 000 000 000 000 00091091 gramos

¿Cómo lees estos números? ¿Cómo los sumas, multiplicas o divides? Parece evidente que, a pesar de las bondades que indudablemente tienen el sistema de notación decimal y -en general- los sistemas posicionales para representar un número, al menos en este caso, se presentan serias limitaciones.

[1]Knuth, D. E. The Art of Computer Programing. Vol. 2. Addison-Wesley Reading Ma. 1981.

Por tal razón, se ha desarrollado un sistema de representación de números que se denomina *notación científica*, que consiste fundamentalmente en representar al número como producto. Uno de cuyos factores es una potencia de 10; así, en los ejemplos de la masa en gramos de la Tierra y de un electrón, una representación de dichos números sería la siguiente:

598×10^{25}, para el primero y

91091×10^{-32}, para el segundo.

Con esa representación, las operaciones resultan sumamente sencillas, ya que se aprovechan las propiedades del campo de los números reales; el producto de los dos números será, entonces:

$$(598 \times 10^{25})\,(91091 \times 10^{-32}) =$$

$$(598 \times 91091)\,(10^{25} \times 10^{-32}) =$$

$$(54472418)\,(10^{25-32}) = 54472418 \times 10^{-7}$$

$$= 5.4472418$$

Como podrás observar, los algoritmos, y en general la aritmética de los números reales, están intimamente vinculados con el sistema de representación de los números.

Por lo tanto, a partir de un sistema posicional de representación numérica, un número, $Z \in \Re$, se representa de la siguiente manera:

$$Z = \pm(a_k a_{k-1} \ldots a_0 . a_{-1} a_{-2} \ldots)_q$$

$$= \pm\left(a_k q^k + a_{k-1} q^{k-1} + \ldots + a_0 + \frac{a_{-1}}{q} + \frac{a_{-2}}{q^2} + \ldots\right)$$

Una representación en notación científica de este número será:

$$Z = \pm (. a_k a_{k-1} \ldots a_0 a_{-1} a_{-2} \ldots) q^e$$

$$= \pm \left(\frac{a_k}{q} + \frac{a_{k-1}}{q^2} + \ldots + \frac{a_0}{q^{k+1}} + \frac{a_{-1}}{q^{k+2}} + \ldots\right) q^e$$

con $b_1 \neq 0$
$q = 10$

En dispositivos finitos, como una computadora, esta representación debe tener una longitud finita *t*, por lo que el número se convierte en:

$$Z = \pm(. b_{-1} b_{-2} \ldots b_t) q^e$$

$$Z = \pm q^e \left(\frac{b_{-1}}{q^{-1}} + \frac{b_{-2}}{q^{-2}} + \ldots + \frac{b_{-t}}{q^{-t}}\right)$$

Ejemplo:

$$-36.12 = (3 \cdot 10^{-1} + 6 \cdot 10^{-2} + 1 \cdot 10^{-3} + 2 \cdot 10^{-4}) \cdot 10^2$$

$$= (0.3612)(-10^2)$$

NÚMEROS EN PUNTO FLOTANTE

El término *aritmética de punto flotante* es un concepto de la ciencia de las computadoras, pero también es comprensible para cualquiera que ha usado una regla de cálculo o trabajado con logaritmos; el concepto pudo haber sido tambien familiar para los astrónomos medievales que multiplicaron grandes números mediante el dispositivo denominado *posthaphaeresis*.[1]

La primera operación a que un número es sometido por la computadora es el redondeo, por medio de la cual la computadora guarda la parte más significativa de un número, de acuerdo con sus posibilidades. Esto se debe a que la computadora, por limitaciones físicas, sólo puede almacenar un número finito de dígitos.

[1] "Sherlock Holmes in Babylon". The American Mathematical Monthly. Vol. 87 pp. 335

Abordemos tal descripción, planteándonos ahora la situación de que sólo disponemos de la capacidad de almancenar un número finito de dígitos de un número real: ¿qué es lo que debemos almacenar entonces?

La pregunta anterior nos conduce a la siguiente *¿Cuál es la parte más significativa de un número a partir de su expansión decimal?* Para entender bien esta pregunta es necesario tener presente que en la práctica los números reales representan magnitudes y, por consiguiente, su parte más significativa es aquélla que nos representa mejor su magnitud.

Afortunadamente, la respuesta es simple, ya que generalmente los seres humanos hacemos la misma aproximación que las computadoras, para guardar de una manera efectiva la información más diversa sobre el mundo, la sociedad en que vivimos y nuestra vida personal. ¿Cómo hacemos la operación de redondeo?.. ¡Tomamos los dígitos que están más a la izquierda en la representación decimal!

Ejemplos:

$\pi \equiv 3.1415$ $\quad$ $\$898.95 \equiv \$898.00$

$\sqrt{2} \equiv 1.414$ $\quad$ $1/11 \equiv .0909$

$1/3 \equiv 0.333$ $\quad$ $.000123 \equiv .0001$

Si suponemos que sólo contamos con dos lugares para almacenar los dígitos de un número; entonces, para almacenar los números del ejemplo anterior, podríamos "*redondearlos*" de la siguiente forma.

$\pi \rightarrow 3.1$ $\quad$ $\$898.95 \rightarrow \$90.$

$\sqrt{2} \rightarrow 1.4$ $\quad$ $1/11 \rightarrow .09$

$1/3 \rightarrow 0.3$ $\quad$ $.000123 \rightarrow .00$

Como podemos observar, el "redondeo" aquí deja mucho que desear; pues, en algunos casos, el valor almacenado no se parece en nada al número original. El ejemplo es extremoso pero tiene por objeto mostrar lo inadecuado, en algunos casos, de la notación decimal común y el absurdo a que nos puede conducir su mal uso.

Si nos volvemos esclavos del punto decimal, esto nos hace guardar información irrelevante.

Por otro lado, la notación científica puede ser muy útil para hacer compacta la escritura decimal de un número real.

Ejemplo:

$$38400000000000 = 384 \times 10^{11}$$

$$.0000000000384 = .384 \times 10^{-10}$$

Sin embargo, si la colocación del punto decimal sigue siendo todavía un problema, éste se resuelve cuando logramos que todos los números reales.. *itengan el punto decimal en la misma posición!*

Ejemplos:

$\pi = .31415... \times 10^{1}$ $\quad 898.95 = .89895 \times 10^{3}$

$\sqrt{2} = .1414... \times 10^{2}$ $\quad \frac{1}{11} = .90909 \times 10^{-1}$

$\frac{1}{3} = .333 ... \times 10^{0}$ $\quad .000123 = .123 \times 10^{-3}$

Por lo tanto, todo número real $x > 0$ se puede escribir en la forma:

$$x = \left(\frac{d_1}{10} + \frac{d_2}{10^2} + \frac{d_t}{10^t} + \ldots \right) 10^e \qquad d_1 \neq 0$$

En general, en base arbitraria β:

$$x = \pm (.b_1 b_2 \ldots b_t) \beta^e$$

$$= \pm (b_1 \beta^{-1} + b_2 \beta^{-2} + \ldots + b_t \beta^{t}) \beta^e$$

$$= \pm \left(\sum_{i=1}^{t} b_i \beta^{-i} \right) \beta^e$$

Donde:

β = *base*

e = *exponente*

$. b_1 b_2 \ldots b_t$ = *mantisa*

t = *precisión*

b_t = *dígito*

$b_1 \neq 0$

Denominamos a esta representación de un número real *representación normalizada en base* β.

La ventaja de la formulación anterior es que nos indica la manera adecuada de redondear un número, puesto que se "intuye" que la operación de redondear indica que debemos truncar la serie, pero debemos respetar el exponente e.

Ejemplo. Una forma de redondear a 2 cifras los números del ejemplo anterior:

$\pi \rightarrow .31 \times 10^{1}$ $\quad \$898.95 \rightarrow \$ .89 \times 10^{3}$

$\sqrt{2} \rightarrow .14 \times 10^{1}$ $\quad . \frac{1}{11} . \rightarrow .90 \times 10^{-1}$

$\frac{1}{3} \rightarrow .33 \times 10^{0}$ $\quad .000123 \rightarrow .12 \times 10^{-3}$

Otra forma muy usada de redondear consiste en tomar en cuenta el primer dígito a partir del cual se va a truncar; si éste es mayor o igual que 5, entonces, incrementamos el último dígito que se va a conservar por una unidad.

Ejemplo. Otra forma de redondear:

$$\$898.95 \rightarrow \$.90 \times 10^{3}$$

$$\frac{1}{11} = .909090... \times 10^{-1} \rightarrow .91 \times 10^{-1}$$

Por lo tanto, *redondear un número real es aproximarlo mediante una expansión decimal finita.* Esto se obtiene tomando la parte más significativa de x de su representación decimal normalizada.

Por lo tanto, si denotamos por $r(x)$ al número redondeado de x tenemos que:

$$r(x) = sign(x) \cdot m \cdot \beta^{e}$$

Donde:

$$sign(x) = \begin{cases} -1 & \text{si } x < 0 \\ 0 & \text{si } x = 0 \\ 1 & \text{si } x > 0 \end{cases}$$

y $$m = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots \frac{d_t}{\beta^t}$$

A la forma de redondear, en la cual sólo se trunca la serie, se le llama *redondeo por corte*; y a la que toma en cuenta la magnitud de los dígitos que se van a eliminar se le llama *redondeo simétrico*.

Por lo tanto:

$$r_c(x) = sign(x)\left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots + \frac{d_t}{\beta^t}\right)\beta^{e}$$

$$r_s(x) = \begin{cases} r_c(x) & \text{si } d_t < \beta/2 \\ r_c(x) + \dfrac{1}{\beta^t} & \text{si } d_t \geq \beta/2 \end{cases}$$

Como es natural, esperamos que esta forma de representar números reales sea la más adecuada para emplearse en la computadora digital, ya que la operación de redondeo es simple en su ejecución en esa representación. Sin embargo, la simpleza no es una razón suficiente, lo que queremos es la utilidad.

En primer lugar, la base 10 es poco común en las computadoras, son más comunes las bases 2, 4, 8 y 16. Esto no presenta mayor problema, puesto que los resultados con base 10 se pueden extender muy fácilmente a cualquier otra base.

Errores

Usualmente los datos de un problema provienen de medidas experimentales y éstas están influídas por errores sistemáticos. Además, se presentan errores de redondeo cuando se toma un número finito de dígitos de los números.

Sí se hacen las operaciones en una calculadora que sólo puede manejar un número de t dígitos; entonces, por ejemplo, el producto de dos números usualmente constará de $2t$ dígitos y, por consiguiente, habrá que "redondear" el resultado. Como veremos, el efecto de redondeo puede afectar mucho los resultados o puede ser insignificante; eso dependerá del problema y del método numérico usado.

Errores de truncamiento, son los que se presentan cuando un proceso infinito es truncado. Esto sucede, por ejemplo, cuando una serie infinita se reduce a un número finito de términos o cuando una derivada se aproxima por un cociente de diferencias. En este último caso también se usa el nombre de error de discretización.

Error relativo y error absoluto

Supongamos que cada semana todo individuo que percibe ingresos debe pagar al gobierno mil pesos de impuesto. Es evidente que la proporción que esto representa para el salario individual no es igual para todos: ¡Protesta nacional!

Tratemos de pensar en un invento genial: Imaginemos una balanza de uso universal, es decir, que sirva para pesar objetos tan pequeños como una célula o tan grandes como un camión cargado. ¿Qué características deberá tener?

a) La graduación de la balanza deberá ser extremadamente fina cerca de los pesos pequeños y, por cuestiones prácticas, para pesos grandes podrán estar más separadas las marcas.

b) Lo anterior significa que cerca del cero las marcas deberán estar muy próximas las unas de las otras y, conforme nos alejamos del cero, deberán espaciarse.

Lo anterior motiva la necesidad de juzgar "qué tan cerca estamos del peso correcto".

Supongamos que $r(x)$ aproxima a una cierta cantidad x, ¿qué tan buena es la aproximación?

Redondeemos a 2 dígitos los números:

$$x = 33600.48; \quad y = .003360048$$

Entonces:

$$x = .3360048 \times 10^5; r(x) = .33 \times 10^5$$
$$y = .3360048 \times 10^{-2}; r(y) = .33 \times 10^{-2}$$

Pero:

$$x - r(x) = 600.48 = .60048 \times 10^3$$
$$y - r(y) = .000060048 = .60048 \times 10^{-4}$$

Eso parece indicar que el redondeo "cuesta" mucho a los números grandes y poco a los pequeños. ¿Todavía seguimos pensando que el redondeo es útil?

Consideremos la lista siguiente:

x_i	→	$r(x_i)$
1.000...		$.100 \times 10^1$
1.007		$.100 \times 10^1$
1.009		$.100 \times 10^1$
1.010		$.101 \times 10^1$
2.792		$.279 \times 10^1$
10.25		$.102 \times 10^2$
10.27		$.102 \times 10^2$
10.31		$.103 \times 10^2$
100.1		$.100 \times 10^3$
101.5		101×10^3
102		102×10^3
1000		100×10^4
1009		101×10^4
1010		101×10^4

Observamos que para los números entre 1 y 10 la diferencia entre dos redondeados consecutivos es .01 para números entre 10 y 100 la diferencia es .1 y para números entre 100 y 1000, 1. Geométricamente la situación es que cerca de cero los números están muy cercanos entre sí, y lejos de cero están separados.

Los números representados son:

1, 1.01, 1.02, 1.03, ..., 1.99, 2.00, ..., 9.99
10, 10.1, 10.2, ..., 10.9, 11, ..., 99.9
100, 101, 102, ..., 109, 110, ..., 999
1000;

Para $n = 1$, la situación es crítica, pues la lista sería:

1, 2, 3, ..., 9
10, 20, 30, ..., 90
100, 200, 300, ... 900
1000

Estamos al parecer en un conflicto, pues sabemos que el redondeo se usa en la práctica y vemos que tiene limitaciones. Pero lo más extraño es que no hemos oído queja. ¿Por qué? Seguramente porque debe tener algo bueno, pero, ¿qué es? Es natural sospechar que estamos juzgando mal al redondeo, ya que cuando redondeamos estamos interesados en la parte más significativa y ésta depende del número. Así que para medir su efectividad es conveniente que usemos una medida relativa y no una absoluta.

Si x y y son los números del ejemplo anterior, tenemos

$$\left|\frac{x - r(x)}{x}\right| = \left|\frac{y - r(y)}{y}\right| \approx 2 \times 10^{-2}$$

Vemos que el "error relativo" es una medida más adecuada para juzgar el redondeo.

Si $r(x)$ es un redondeo de x entonces:

El error absoluto de $r(x)$ *es*

$$E_A = |r(x) - x|$$

El error relativo de $r(x)$ *es*

$$E_R = \left|\frac{r(x) - x}{x}\right| \qquad si\ x \neq 0$$

ARITMÉTICA DE COMPUTADORA

A la gente le sorprende descubrir que las computadoras no siempre "calculan la respuesta correcta". Una de las razones por las que las computadoras fallan es que los números, sobre los cuales efectuan sus cálculos, son aproximados o redondeados. Los números reales generalmente son representados en una computadora usando un número fijo de bits; 32, 36 y 60 son los números más comunes de bits.

Una computadora que usa 32 bits para la aritmética real, puede representar solo 2^{32} números reales diferentes. ¿Cuáles, del infinito número de reales, son representables?

La respuesta es: *un conjunto particular de números de punto flotante*. Este conjunto Q_M es un subconjunto finito de los números racionales Q, que -a su vez- son subconjunto de los reales.

El conjunto de números representables depende del formato implementado en la computadora. Un formato de punto flotante, junto con las reglas para efectuar operaciones aritméticas con números en dicho formato, constituye un *sistema de punto flotante*. La mayoría de las grandes computadoras tiene integrado, por lo menos, un sistema de punto flotante en el hardware. La mayoría de las microcomputadoras no tiene hardware de punto flotante, pero hace cálculos de punto flotante a través del software.

El sistema descrito a continuación está basado en el estándar de la aritmética binaria de punto flotante, fue adoptado en 1985 por la ANSI y formulado por el IEEE.

Los números representables que estudiaremos son los de la forma:

$$(-1)^s (.d_1 d_2 \ldots d_t)\, \beta^e$$

En donde:

s	determina el signo del número
d_i	son dígitos tales que $0 \le d_i \le \beta$; $d_1 \neq 0$
e	es el exponente y $E_{min} \le e \le E_{max}$
β	es la base del sistema
t	dígitos de la mantisa o precisión
$[E_{min}, E_{max}]$	es el rango de los exponentes

Es claro que son cuatro parámetros los que determinan este conjunto de números representables en un sistema de punto flotante. Dichos parámetros son:

$$\beta, t, E_{min}, E_{max}$$

En lo sucesivo denotaremos el conjunto de números representables en una computadora de la siguiente manera:

$$Q_M = Q(\beta, t, E_{min}, E_{max})$$

Es fácil ver que éste es un subconjunto finito de números racionales que, en general, podemos denotar como:

$$Q_M = Q(\beta, t, E_{min}, E_{max}) = \{x \mid x = \pm m\beta^e\} \cup \{0\}$$

Para aclarar, consideremos los siguientes ejemplos:

1. Sea $Q_M = Q(2, 3, -1, 2)$, tenemos entonces un conjunto finito de números binarios con tres cifras; por lo que, en general, son números de la forma:

$$\pm .d_1\, d_2\, d_3 \times 2^e$$

En donde:

$$0 \leq d_1 \leq 1$$

$$-1 \leq e \leq 2$$

Los números que constituyen este conjunto son:

$$\left\{\frac{1}{4}, \frac{5}{16}, \frac{3}{8}, \frac{7}{16}, \frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8}, 1, \frac{5}{4}, \frac{3}{2}, \frac{7}{4} \ldots\right\} \cup \{0\}$$

Ya que:

$$(.100)_2 \times 2^{-1} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$(.101)_2 \times 2^{-1} = \left(\frac{1}{2} + \frac{1}{8}\right) \times \frac{1}{2} = \frac{5}{16}$$

$$(.110)_2 \times 2^{-1} = \left(\frac{1}{2} + \frac{1}{4}\right) \times \frac{1}{2} = \frac{3}{8}$$

$$(.111)_2 \times 2^{-1} = \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8}\right) \times \frac{1}{2} = \frac{7}{16}$$

$$(.100)_2 \times 2^{0} = \left(\frac{1}{2}\right) \times 1 = \frac{1}{2}$$

$$(.101)_2 \times 2^{0} = \left(\frac{1}{2} + \frac{1}{8}\right) \times 1 = \frac{5}{8}$$

$$(.110)_2 \times 2^{0} = \left(\frac{1}{2} + \frac{1}{4}\right) \times 1 = \frac{3}{4}$$

$$(.111)_2 \times 2^0 = \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8}\right) \times 1 = \frac{7}{8}$$

$$(.100)_2 \times 2^1 = \frac{1}{2} \times 2 = 1$$

$$(.101)_2 \times 2^1 = \left(\frac{1}{2} + \frac{1}{8}\right) \times 2 = \frac{5}{4}$$

$$(.110)_2 \times 2^1 = \left(\frac{1}{2} + \frac{1}{4}\right) \times 2 = \frac{3}{2}$$

$$(.111)_2 \times 2^1 = \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8}\right) \times 2 = \frac{7}{4}$$

etc.

2. Si $Q_M = Q(10, 1, -1, 1)$; los números de este sistema son:

$+ .1 \times 10^{-1}$,
$+ .2 \times 10^{-1}$,...
, $+ .9 \times 10^{-1}$,
$+ .1 \times 10^{0}$,
$+ .2 \times 10^{0}$,...
, $+ .9 \times 10^{0}$,
$+ .1 \times 10^{1}$,
$+ .2 \times 10^{1}$,...
, $+ .9 \times 10^{1}$

Además de los negativos y el cero. Este conjunto tiene solamente 55 números.

Por lo tanto, un número representado en computadora, esto es, un número perteneciente a un conjunto Q_M, tiene tres componentes:

1. Un signo aritmético ..... $sign(x)$

2. Un exponente ..... e

3. Una fracción o mantisa ..... $m = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots \frac{d_t}{\beta^t}$

Con las siguientes propiedades:

i) $\frac{1}{\beta} \leq m < 1$

ii) $\beta^{e-1} \leq |x| < \beta^e$ si $x \neq 0$

iii) $x = sign(x) m \beta^e$

En consecuencia, se puede representar como una terna:

$$(sign(x), m, e)$$

Para hacer cálculos numéricos en la computadora es necesario, en virtud de las características de Q_M, definir una función de $\Re$ en Q_M que denominaremos *redondeo.*

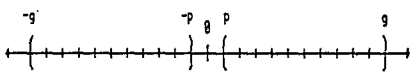
Antes de formular esa función, es necesario hacer notar que existen números reales con valor absoluto muy pequeño que no pertenecen a Q_M, ya que basta que el exponente del número en base β sea menor que E_{min}. Lo mismo sucede para números con valor absoluto muy grande.

Por lo tanto, y en virtud de la finitud de Q_M, existen dos números p y g en Q_M, tales que:

$$|p| \leq |x| \quad \forall x \in Q_M$$

$$|g| \geq |x| \quad \forall x \in Q_M$$

Por lo que la representación en una recta del conjunto de números de computadora Q_M será similar a la siguiente figura:



En todo conjunto Q_M, el número más pequeño y el número más grande se expresan en función de la base, en la siguiente forma:

$$p = \frac{1}{\beta} \times \beta^{E_{min}} = \beta^{E_{min}-1}$$

$$g = \left(1 - \frac{1}{\beta^t}\right)\beta^{E_{max}} = \beta^{E_{max}} - \beta^{E_{max}-t}$$

Ejemplos: Dados $Q(2, 3, -1, 2)$ y $Q(2, 1, -1, 1)$, los números más pequeños en valor absoluto son:

$$2^{-1-1} = \frac{1}{4} \quad \text{y} \quad 2^{-1-1} = \frac{1}{2}$$

Y los números más grandes en valor absoluto son:

$$2^2 - 2^{2-3} = 3\frac{1}{2} \quad \text{y} \quad \left(1 - \frac{1}{2}\right)2 = 1$$

Para definir la función de redondeo, es necesario agregar dos elementos a Q_M que sean las imágenes de todos los números reales más pequeños que p y más grandes que g; denotamos a estos elementos como *overflow* y *underflow*. Definimos la función de redondeo de la siguiente manera:

$$r : \Re \rightarrow Q_M \cup \{o\} \cup \{underflow\} \cup \{overflow\}$$

Tal que:

$$r(x) = \begin{cases} overflow & \text{si} \quad |x| > g \\ r(x) & \text{si} \quad p \leq |x| \leq g \\ \\ o & \text{si} \quad x = 0 \\ underflow & \text{si} \quad 0 < |x| < p \end{cases}$$

Decimos que $x \in \Re$ está en el *l* rango de Q_M si:

$$x = 0 \qquad 6 \qquad p \leq |x| \leq g$$

En este contexto, es evidente que el error relativo que se comete al redondear un número depende del sistema Q_M de una máquina particular; independientemente de ello, podemos expresar este error, que se acostumbra denominar PROPIEDAD FUNDAMENTAL DE Q_M, de la siguiente manera:

Si x está en el rango de Q_M, entonces:

$$\left| \frac{x - r(x)}{x} \right| < \beta^{1-t}$$

para el redondeo por corte.

Y

$$\left| \frac{x - r(x)}{x} \right| < \frac{1}{2} \beta^{1-t}$$

para el redondeo simétrico.

OPERACIONES EN Q_M

Como es natural, queremos usar la computadora para hacer las operaciones aritméticas usuales; sin embargo, como el conjunto Q_M es finito, no es posible realizar las operaciones en forma usual, debido a que Q_M no es cerrado bajo ninguna de las operaciones. Para ejemplificar, consideremos a $Q_M = Q(10, 1, -1. 1)$; dos elementos de este conjunto son: $.9 \times 10^1$ y $.4 \times 10^1$; la suma de ambos es $.13 \times 10^2$, que no pertenece a Q_M; ya que $r(.13 \times 10^2) = overflow$.

Debido a lo anterior, las operaciones aritméticas se hacen sólo de manera aproximada. En la práctica, la mayoría de la gente no lo nota, sobre todo en el campo de las aplicaciones comerciales o contables, donde no se necesita mucha precisión, no así en el campo de la

aplicación científica o numérica. Es por ello que en esta última área es necesario, en ocasiones, usar la denominada *doble precisión*.

El diseño de las operaciones aritméticas en computadora hace que el error relativo de los resultados sea el mínimo posible, dentro de las limitaciones de Q_M.

En general, si denotamos por e la diferencia entre el número real y el número de computadora que se aproxima a aquél, tenemos lo siguiente:

operaciones en $\Re$		operaciones en Q_M
$a+b$	———	$(a+b)(1+e)$
$a-b$	———	$(a-b)(1+e)$
$a \cdot b$	———	$(a \cdot b)(1+e)$
$a \div b$	———	$(a \div b)(1+e)$

PROPIEDADES DE LAS OPERACIONES

Para denotar cualquiera de las operaciones aritméticas en Q_M usaremos el símbolo *, y * para las operaciones en $\Re$; donde:

$$x * y = r(x*y)$$

Expresamos la propiedad fundamental de las operaciones en Q_M en la siguiente forma:

$$\left| \frac{(x*y) - (x*y)}{x*y} \right| < \beta^{-t}$$

Donde:

$$x * y = (x * y)(1+\delta)$$

$$|\delta| < \in_M$$

La suma y la multiplicación en Q_M son conmutativas, esto es:

$$x \oplus y = y \oplus x$$

$$x \otimes y = y \otimes x$$

La múltiplicación es casi asociativa:

$$(x \otimes y) \otimes z \approx x \otimes (y \otimes z)$$

Pero, desafortunadamente, la adición no es asociativa, esto es:

$$(x \oplus y) \oplus z \neq x \oplus (y \oplus z)$$

Consideremos un conjunto Q_M con $\beta = 10$ y $t = 4$ Entonces, si:

$$a = 0.4651 \times 10\text{-}3$$

$$b = 0.2524 \times 10\text{-}3$$

$$c = 0.1333 \times 100$$

Entonces:

$$(a \oplus b) \oplus c = 0.1340 \times 100$$

$$a \oplus (b \oplus c) = 0.1341 \times 100$$

Este ejemplo "parece" no ser muy grave, pero nótese que sólo hemos sumado tres sumandos y en los cálculos numéricos se efectúan miles, y a veces millones, de operaciones.

El problema de que la suma no sea asociativa nos preocupa porque nos plantea la pregunda obvia: ¿COMO SUMAMOS?

Recordemos que si $x, y \in Q_M$ entonces:

$$x \oplus y = (x+y)(1+\delta) = (x+y)+(x+y)\delta \text{ donde } \delta \leq u$$

Esto sugiere, como método práctico para efectuar una suma, ordenar los sumandos de tal manera que:

$$x_1 < x_2 < \ldots < x_n$$

Y efectuar la suma a partir del sumando más pequeño, de tal manera que si la suma de la computadora es:

$$S = x_1 + x_2 + \ldots + x_n$$

Tendremos que, de manera análoga a las propiedades fundamentales del error, enunciadas anteriormente:

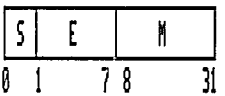
$$\left| \frac{S - S_c}{S} \right| < u$$

Mencionamos en secciones anteriores que podemos representar los números de Q_M como una terna formada por: un signo, $sign(x)$; una mantisa, $d_1\, d_2 \ldots d_t$ y un

exponente, *e*. Por lo que surge de manera natural la siguiente pregunta: ¿Cómo son divididos los bits de una máquina real de 32 bits entre el signo, el exponente y la fracción para propósitos de punto flotante?

La respuesta varía de máquina a máquina, y es instructivo examinar cómo son los números de punto flotante en dos máquinas reales de arquitectura diferente.

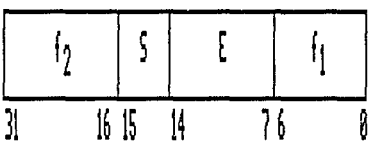
Las computadoras IBM, tales como la 4300 y la 3080, usan números hexadecimales en punto flotante (base $\beta = 16$). Para números de precisión sencilla, el número de dígitos en la fracción es $t = 6$; de los 32 bits, en una palabra sencilla, los primeros 7 bits estan dedicados al exponente. El siguiente bit es el bit del signo, los 24 bits restantes contienen seis grupos de cuatro bits cada uno, los seis dígitos hexadecimales representados en binario. En estas máquinas, no necesitan ser normalizados los números, de tal manera que el primer dígito hexadecimal de la fracción puede o no ser diferente de cero. Se supone que el punto decimal precede al dígito más a la izquierda. Representando el signo por *S*, el exponente por *E* y la mantisa por *M*, el esquema de un número de punto flotante de precisión sencilla en estas máquinas IBM está dado por:



donde los dígitos de la parte inferior del esquema representan la posición del bit en una palabra de 32 bits.

Otras máquinas dividen el espacio disponible de diferentes maneras. Las computadoras VAX-II de Digital Equipment Corporation (DEC) también usan números de punto flotante en precisión sencilla de 32 bits, pero su representación interna es completamente diferente. Las VAX requieren que todos los números sean normalizados, pero el primer bit -el que siempre está activado- no es explicitamente representado, de tal manera que una fracción de *t-bits* realmente representa $t + 1$ bits significativos. La base es $\beta = 2$ y la fracción tiene $t = 23$ bits, mas el bit del signo. El exponente ocupa 8 bits representado en notación *exceso-128*. Un valor especial del exponente es reservado como un caso especial; si los ocho bits del exponente son todos cero, no se considera como -128 (el exponente). La interpretación del número depende del valor del bit del signo. Si el bit del

signo es cero, entonces el número es considerado como el cero de punto flotante, sin importar el valor de los bits restantes de la fracción. Si el bit del signo es uno, entonces el número es considerado como "número invalido", el cual, cuando es identificado, dispara un operando de error. Esta característica fue la precursora de los *NaNs,* definidos mediante las normas actuales de punto flotante. El arreglo de los bits en una palabra de 32-bits es peculiar en la VAX. Denotando al bit del signo por *S*, el exponente por *E* y la parte fraccionaria por $.If_1f_2$, donde f_1 contiene siete dígitos significativos y f_2 contiene los últimos 16 bits. Entonces en la memoria, el número de punto flotante es almacenado como:



ANTECEDENTES HISTÓRICOS

Hasta principios del siglo XIX "la matemática parecía más tarea de físicos o de ingenieros que de matemáticos". Esta aseveración de Babini y la revisión histórica desde la época de las culturas babilónica y egipcia hasta el presente, nos inducen a afirmar que la *matemática*, hasta antes del siglo XIX, *era fundamentalmente numérica y no analítica*. En la clasificación de las ciencias, elaborada por los enciclopedistas franceses, la matemática encabezaba la lista de las ciencias naturales y se decía que"... (su) objeto (de estudio) es la cantidad...".

Así, a lo largo de la historia, y en diversas culturas, encontramos algoritmos para resolver problemas numéricos, que se pueden considerar como antecedentes historicos de los métodos numéricos actuales. La diferencia esencial entre unos y otros es el uso de los dispositivos de cálculo.

Con el objeto de tener una visión panorámica de lo afirmado en el párrafo anterior, presentaremos en esta sección algunos problemas y las soluciones propuestas en diferentes épocas y culturas.

EGIPTO

El conocimiento que tenemos de la matemática egipcia se desprende, fundamentalmente, de dos documentos escritos alrededor de 1850 a.C. y1650 a.C., respectivamente: *El papiro de Moscú* y el *papiro de Rhind.*

La mayor parte de los problemas son de origen práctico, se refieren a alimentación de animales, almacenamiento de granos, etc.

El problema 24 del papiro de Rhind es el prototipo de una serie de problemas que clasificamos en la actualidad como ecuaciones lineales en una incógnita (*aha* ≡ *cantidad* para los egipcios). La traducción sería la siguiente: "Aha, el total, más su séptima parte hacen 19". La ecuación resultante en la notación actual sería:

$$x+\frac{x}{7}=19$$

Los egipcios usaban el método de falsa posición para resolver ecuaciones mediante la asignación de un valor a la incógnita, y comprobaban si se satisfacian las condiciones dadas; en caso de que no, sustituían el valor mediante una proporción simple.

Ejemplo: Para resolver $x+\frac{x}{4}=30$ podemos asignar a x el valor de 4 y tendremos que $x+\frac{x}{4}=5$ en lugar de 30; ya que 5 debe ser multiplicado por 6 para obtener 30; por lo tanto, la respuesta debe ser $24 = 6 \times 4$

MESOPOTAMIA

Las raíces cuadradas parecen haber sido calculadas mediante la siguiente fórmula:

$$\sqrt{A}=\sqrt{a^2}+h=a+\frac{h}{2a}=\frac{1}{2}\left(a+\frac{A}{2}\right)$$

Existen textos cuneiformes con problemas de interés compuesto, tal como el de la cuestión de qué tanto tiempo es necesario para que determinada cantidad de dinero sea duplicada, al 20% de interés. Usaban técnicas de interpolación en la solución de este tipo de problemas.

INDIA

Los textos hindúes más antiguos (en los que existen referencias matemáticas) datan de los primeros siglos de nuestra era.

Entre otras cosas, se encuentran algunas aproximaciones curiosas en términos de fracciones unitarias, tales como:

$$\sqrt{2} = 1+\frac{1}{3}+\frac{1}{3\times4}-\frac{1}{3\times4\times3\times4} = 1.4142156$$

$$\pi = 4\left(1-\frac{1}{8}+\frac{1}{8\times29}-\frac{1}{8\times29\times6}+\frac{1}{8\times29\times6\times8}\right)^2$$

El hecho curioso es que estos datos, encontrados en los *Sulvasutras,* no aparecen en los trabajos hindúes posteriores; lo que nos habla de la falta de continuidad de la tradición en la matemática hindú, a diferencia de sus homólogas egipcia y babilónica.

CHINA

Dinastía Han (206 a.C. - 220 d.C.)

"*La matemática de los nueve capítulos*" consiste principalmente en un conjunto de problemas con reglas generales para su solución; son de característica computacional aritmética y conducen a ecuaciones algebraicas con coeficientes numéricos.

Una serie de tales problemas conduce a sistemas de ecuaciones lineales como el siguiente:

$$3x+2y+z = 39$$

$$2x+3y+2 = 34$$

$$x+2y+3z = 26$$

Que es escrito en forma de matriz de sus coeficientes.

La solución se obtiene mediante lo que ahora denominamos transformación de matrices.

SIGLOS XVII - XIX

Poco después del descubrimiento del cálculo diferencial e integral, se desarrollaron nuevas técnicas que facilitaron el cálculo numérico.

La primera de ellas fue el trabajo de Brook Taylor de la expansión en series de potencias de muchas funciones.

Otra herramienta importante en el cómputo numérico fue el desarrollo del cálculo de diferencias finitas. Muchas de las técnicas o métodos numéricos actuales lo utilizan.

SIGLO XX

Es indiscutible que la gran herramienta del siglo actual es la computadora y, por supuesto, los métodos numéricos no sólo han hecho uso de la misma, sino que son fuertemente determinados por ella. De hecho, en la actualidad no es posible hablar de métodos numéricos sin que éstos estén diseñados para plantearse a través de un lenguaje de alto nivel. -*FORTRAN, Pascal, C*- en las computadoras.

CONCLUSIÓN

Una etapa importante en la aplicación de la matemática es la capacidad de complementar la solución de un problema, lo que en muchos casos significa obtener números para poder continuar con algún proceso de producción o de investigación.

Resolver un "problema numérico" nos conduce, necesariamente, a la obtención de números. Parece sencillo, sin embargo es difícil lograr que esos números se apeguen a la realidad. Es decir, dado que en general sólo podemos obtener soluciones aproximadas, es preciso desarrollar toda una serie de métodos o algoritmos básicos que, combinados adecuadamente, nos lleven a la solución con un costo mínimo y máximo grado de precisión. Para ello, será necesario cubrir los siguientes requisitos:

i) Tener conocimientos teóricos.

ii) Desarrollar cierta capacidad para interpretar adecuadamente un problema dado, o bien traducirlo a un problema soluble numéricamente.

iii) Conocer los procedimientos básicos mencionados y ser capaces de aplicarlos a cada problema.

iv) Manejar un lenguaje que nos permita describir, de manera clara y sin ambigüedad, la forma de resolver un problema particular.

Para una buena cantidad de problemas numéricos no será suficiente con los puntos anteriores, pues si la cantidad de operaciones que haya que hacer rebasa los recursos humanos, será necesario echar mano de una computadora de alta velocidad.

Se puede verificar que la matemática nació asociada con las necesidades de la civilización humana. La aritmética y la geometría elementales fueron los pilares sobre los que se desarrollaron inicialmente la construcción, la navegación, el control de las transacciones comerciales y la administración.

Las actividades humanas presentaron problemas nuevos a la matemática, lo cual estimuló su desarrollo. A su vez, el progreso matemático generó métodos

más efectivos, expandió su área de aplicación y así impulsó un mayor progreso científico y tecnológico.

A través de las computadoras, muchas ciencias están logrando su "matematización". Las computadoras han aumentado el potencial intelectual de la humanidad; eso las hace ocupar un lugar privilegiado entre otras máquinas y permite considerar su invención como uno de los más grandes logros de la humanidad.

La influencia que la matemática puede tener en los diferentes campos de la actividad humana depende:

a) Del nivel de desarrollo del aparato matemático.

b) Del grado de madurez del conocimiento del objeto de estudio.

c) De la capacidad del grupo de individuos interesados en construir un modelo matemático en el que queden representadas las características más importantes del objeto.

Los modelos matemáticos están siempre basados en una simplificación o idealización del objeto, y constituyen una aproximación a él.

Como consecuencia del reemplazo de objetos por modelos, la investigación de los objetos se traduce en la formulación de problemas matemáticos, lo que permite emplear un aparato matemático universal que es independiente de la naturaleza específica del objeto.

La complejidad en la construcción y análisis de un modelo matemático depende generalmente de la complejidad del objeto. Antes de la aparición de las computadoras, los métodos analíticos fueron aplicados a la solución de problemas matemáticos. En esos tiempos, los científicos lucharon por evitar los cálculos laboriosos y trataron de obtener resultados a modo de fórmulas. Aquellos modelos matemáticos para los cuales no era posible obtener una solución en forma explícita no fueron considerados o se les simplificó por medio de hipótesis adicionales. La simplificación del modelo claramente reduce el grado hasta el cual éste corresponde al objeto bajo estudio, haciendo que los resultados sean menos interesantes y, algunas veces erróneos.

La situación ha cambiado bruscamente con la aparición de las computadoras. En los últimos 30 años, la rapidez para efectuar operaciones se ha incrementado por un factor de 100 millones debido a éstas. De igual manera, la aplicación de métodos numéricos basados en las computadoras ha expandido la clase de problemas matemáticos que pueden ser analizados.

En la actualidad, para construir un modelo matemático, el científico no necesita luchar mucho para obtener modelos simplificados. Su atención ahora deberá estar enfocada principalmente en la forma de tomar en cuenta las características más importantes del objeto bajo consideración, para luego expresarlas adecuadamente en el modelo matemático.

Una vez que el modelo está construido, se presenta el problema de desarrollar un método, muchas veces numérico, y usualmente para su planteamiento en una computadora.

De esta forma, las computadoras nos obligan a considerar a la matemática desde otro ángulo, como una herramienta de investigación con nuevas potencialidades, si se usa en forma conveniente. Esto -a su vez- ha producido una reorientación de varias ramas de la matemática y el desarrollo de algunas nuevas.

Así pues, la matemática computacional, es decir, el análisis numérico, se ha convertido en uno de los factores más importantes que han permitido a algunas ciencias alcanzar un alto grado de desarrollo.

EJERCICIOS Y PROBLEMAS*

1. Defina lo que significa "redondear un número real".

   a) Escriba en forma normalizada los siguientes números: 1/7, 1/3, $\sqrt{2}$, e, π.

   b) ¿Cuántas cifras de x deben tomarse en cuenta para tener un error menor que el .01%? Aplique este criterio a los números del inciso (a).

2. a) Transforme los siguientes números de base 10 a base 2: 5847, 82.43, .0028

   b) Transforme los siguientes números de base 2 a base 10: 1001001, 110.111, .00001101

   c) Indique a cuántas cifras significativas en base 10 equivalen a n cifras en base β.

3. Hay cambios de base que se pueden hacer con mucha facilidad. Por ejemplo, se puede usar la siguiente tabla de equivalencias:

base 4	*base 2*
0	00
1	01
2	10
3	11

Para transformar números de base 4 a base 2 y viceversa, simplemente por un proceso de sustitución, como se hace aquí $31202_4 = 1101100010_2$ o bien $1110010100 0_2 = 130220_4$. Nótese que cada dígito en base 4 fue sustituído por una pareja binaria equivalente, y viceversa.

   a) Justifique por qué es válido el proceso.

   b) Generalice el inciso anterior para cambiar números de base β a base β^2 y viceversa.

4. a) Indique cuántos elementos de Q_M hay entre dos potencias consecutivas de la base e interprete geométricamente sobre la recta..

5. ¿Cuáles de los siguientes números no estan en el conjunto $Q(2, 4, -2, 2)$? ¿Por qué? $-.1011 \times 2^1$, $-.11011 \times 2^{-2}$, 1.101×2^2, $.2121 \times 2^2$, $.01101 \times 2^{-2}$, $-.7615 \times 2^2$

6. Describa el conjunto $Q(\beta, t, E_{min}, E_{max})$ para su computadora. Si su computadora tiene doble precisión, describa el conjunto Q_M de doble precisión.

7. Localice en una recta los números positivos del conjunto $Q(2, 3, -2, 1)$

8. ¿Cuál es el intervalo máximo entre dos números consecutivos del conjunto $Q(10, 8, -50, 49)$. ¿Cuál es el mínimo?

*Los ejercicios se tomaron y/o adaptaron de las obras citadas al final de esta sección.

9. Dado $Q(10, 4, -2, 3)$ determine los siguientes valores.

a) $\left(\frac{2}{3}\right) r_s$ b) $\left(-\frac{2}{3}\right) r_s$ c) $(9.99999)r_s$

d) $(9.99999)r_c$ e) $(\pi)r_c$

10. Determine los errores absoluto y relativo por redondeo simétrico y por corte para los números del ejercicio anterior.

11. Determine la unidad de error de redondeo de su computadora.

12. Dados $\beta = 10$ y $t = 4$; encuentre la suma de x y y para los siguientes valores:

a) $x = .1234 \times 10^1$, $y = .1234 \times 10^2$

b) $x = .1234 \times 10^1$, $y = .4321 \times 10^{-3}$

c) $x = .1234 \times 10^2$, $y = 1234 \times 10^{-3}$

d) $x = .1234 \times 10^2$,. $y = -.1235 \times 10^2$

e) $x = .1004 \times 10^2$, $y = -.9987 \times 10^1$

13. "La suma en Q_M no es asociativa". Calcule $\sum_{n=1}^{10} 1/n$ en Q_M *cont* $= 1$ y $\beta = 10$ de dos maneras distintas, como se indica a continuación:

a) Calcule $S_n = S_{n-1} \oplus 1/n$, con $S_1 = 1$, para obtener el resultado en S_{10}

b) Calcule $T_n = T_{n-1} \oplus 1/(11\text{-}n)$, con $T_1 = 1/10$, para obtener el resultado en T_{10}

Escriba una tabla donde aparezcan 1/n, r(1/n), S_n y T_n, para $n = 1, 2, ..., 10$. ¿Qué puede concluir de sus resultados?

14. a) Explique por qué en Q_M, con $t = z$ y $\beta = 10$

i) $\sum_{n=1}^{\infty} = 100$

ii) $\sum_{n=1}^{\infty} k$ es finita, para toda $k \in Q_M$

b) Dado $k \in Q_M$ calcule $\sum_{n=1}^{\infty} k$

15. Dado $a \in Q_M$, $a > 0$

a) Explique por qué la ecuación $x \oplus a = x$ no tiene solución única en Q_M

b) Calcule la solución mínima positiva.

16. ¡Cuidado con la "cancelación desastrosa"!

Use aritmética con $t = 2$ y $\beta = 10$ para calcular las raíces de la ecuación $ax^2 + bx + c = 0$, para los siguientes valores:

$a = .45\ (10^{-4})$

$b = .20\ (10^{1})$

$c = -.22\ (10^{2})$

Usando los siguientes métodos:

a) $r_1 = (-b - \sqrt{b^2 - 4ac}) / (2a)$

$r_2 = (-b + \sqrt{b^2 - 4ac}) / (2a)$

b) r_1 como en el inciso anterior

$r_2 = c / (r_1 a)$ (Justifique esa relación.)

Verifique en ambos casos sus resultados comparándolos con la solución "exacta". ¿Qué puede concluir?

17. El objetivo de este problema es mostrar un camino para determinar el número de dígitos, t, y la base, β, de un sistema de punto flotante cualquiera, $Q(\beta, n, E_{min}, E_{max})$.

a) Demuestre que, en $Q(\beta, n, E_{min}, E_{max}) \sum_{i=1}^{\infty} 1 = \beta^n$

b) Haga ver que $S_n(\beta^n) = \beta^n + \beta$, en donde S_n es la función "sucesor".

c) Demuestre que $\beta^n + \beta = \inf_{k>0}(\beta^n \oplus 2^k \neq b^n)$

d) ¿Cómo se puede calcular β a partir de (a) y (c)? Para calcular n:

e) Verifique que:

i) $\beta^k \oplus 1 = \beta^k + 1 \quad si\ 0 \leq k < n$

ii) $\beta^k \oplus 1 = \beta^k \qquad si\ n \leq k$

De lo anterior se desprende que:

$$n = inf\,(k > 0 / \beta^k \oplus 1 = \beta^k)$$

f) Diga cómo calcular n partiendo de que ya conoce β del inciso (d).

18. Encuentre el (los) punto (s) de inflexión de la función dada por la siguiente tabla:

i	x_i	$f(x_i)$	i	x_i	$f(x_i)$
1	1	373	11	22.0	565
2	5	415	12	22.9	575
3	10	438	13	23	590
4	15	459	14	23.1	620
5	20	491	15	23.2	860

6	21	503	16	23.3	915
7	22	523	17	23.4	944
8	22.5	543	18	23.5	958
9	22.6	550	19	24	986
10	22.7	557	20	26	1067

19. Demuestre que el producto de dos números de t dígitos tiene no más de $2t$ dígitos.

20. En $Q(10, 3, -100, 100)$ y con redondeo por corte, construya ejemplos para demostrar que, en general:

i) $(x \otimes y) \otimes z \neq x \otimes (y \otimes z)$

ii) $(x \oplus y) \oplus z \neq x \oplus (y \oplus z)$

iii) $x \otimes (y \oplus z) \neq (x \otimes y) \oplus (x \otimes z)$

iv) $(x \oplus y) \oplus z$ puede tener un "largo" error relativo

21. Suponiendo que $z = .180 \times 10^2$ es una solución aproximada de $ax = b$; para $a = .111 \times 10^0$ y $b = .200 \times 10^1$. Usando la aritmética del problema anterior calcule el residual. Calcule el residual en doble precisión y en aritmética real. Compare los tres resultados.

22 Usando la aritmética del problema 20, encuentre números u, v, w para los cuales ocurra overflow al calcular $u \otimes (v \otimes w)$ pero no al calcular $(u \otimes u) \otimes w$.

23. Calcule una aproximación para e^x mediante la sucesión de sumas parciales:

$$S_0, S_1, \ldots$$

Donde:

$$S_k = \sum_{j=0}^{k} \frac{x^j}{j!}$$

Nota que:

$$S_{k+1} = S_k + P_{k+1}$$

Donde:

$$P_{k+1} = x P_k / (k+1) \quad \text{y}$$

$$S_0 = 1 \quad ; \quad P_0 = 1$$

Para $x = -10$ encuentre el valor más pequeño de n tal que $S_n = S_{n+1} = S_{n+2} = \ldots$

Calcule los errores absoluto y relativo.

Referencias bibliográficas:

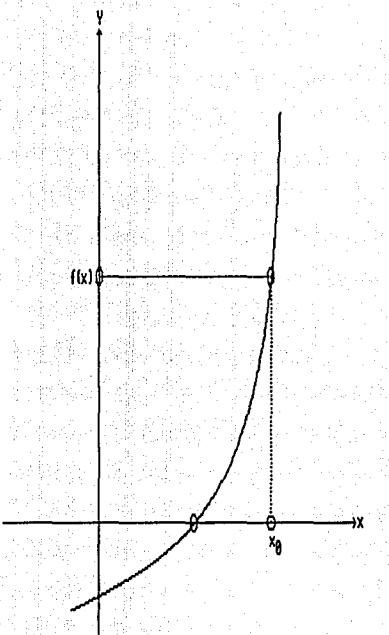
1. *Acton, Forman S.* "Numerical Methods that Work". New York. Harper & Row, 1970.

2. *Shampine, Lawrence F.* "Numerical Computing: an introduction". Philadelphia. W. B. Saunders Company, 1973.

3. *Vandergraft, James S.* "Introduction to Numerical Computations". New York. Academic Press, 1978.

CAPÍTULO II

EVALUACIÓN DE FUNCIONES



"El concepto de función es sugerido por todos los procesos de la naturaleza donde observamos fenómenos que varian de acuerdo con la distancia o el tiempo."

J. T. MERZ

(A History of European Thought in the Nineteenth Century)

"La flor del pensamiento matemático moderno: el concepto de función."

THOMAS J. McCORMACK

(On the Nature of Scientific law and Scientific Explanation)

INTRODUCCIÓN

Aunque el cómputo en ingeniería tiene que ver principalmente con ecuaciones diferenciales, raíces de polinomios y solución de ecuaciones trascedentales, una gran cantidad del trabajo de cómputo se efectúa en la evaluación de funciones.

La computadora está diseñada para sumar y multiplicar números, no para encontrar el coseno de un ángulo o para evaluar la integral de una función, de tal manera que el operador humano tiene que depender de tablas; en virtud de que la memoria de los dispositivos digitales es limitada, es necesario diseñar algoritmos que calculen o evalúen las funciones necesarias directamente de los parámetros de las mismas.

Con tanta frecuencia son usadas las funciones más comunes (senos, cosenos, logaritmos, exponenciales), que la mayoría de los dispositivos digitales, computadoras o calculadoras, tienen subrutinas en lenguaje de máquina que permiten el cálculo directo de las mismas. La facilidad con la que accedemos a esas funciones no debe hacernos olvidar que cada una de ellas requiere de cientos de operaciones y, en consecuencia, está sujeta de manera notable a la propagación de errores de redondeo; por lo tanto, es conveniente usar dichas fun-

ciones de manera adecuada. Por ejemplo, si deseamos evaluar la siguiente expresión:

$$y = a\cos^3 x + b\cos^2 x + c\cos x + d \quad (1)$$

Es conveniente calcularla de la siguiente forma:

$$z = \cos x$$

$$y = ((az + b)z + c)z + d$$

Ya que el cálculo con (1) efectuará más operaciones y en consecuencia tomará más tiempo y tendrá más errores de redondeo.

En la evaluación numérica de una fórmula, no sólo debemos tomar en cuenta la eficiencia y reducir al máximo el número de operaciones elementales, sino que, además, debemos de tomar en cuenta la estabilidad numérica de la misma.

Para poder discutir esta situación, a continuación presentamos el concepto de *condición de una fórmula.*

ESTABILIDAD Y CONDICIÓN

Frecuentemente sucede en un problema numérico que cambios pequeños en los datos producen cambios pequeños en la solución, lo que quiere decir, en cierto sentido, que *el problema es estable.* A esta estabilidad del problema nos referimos cuando decimos que un problema numérico está *bien condicionado.* Como veremos, que un problema esté bien condicionado dependerá de los datos, es decir, para cierto subconjunto del dominio de definición del problema será bien condicionado. Cuando cambios pequeños en los datos produzcan variaciones muy grandes en la solución del problema diremos que éste está *mal puesto* o *mal condicionado.* Saber dónde un problema numérico está mal condicionado es muy importante para la elección del método de solución y para juzgar la calidad de los resultados.

Consideremos el problema de calcular las raíces reales de una ecuación de segundo grado:

$$ax^2 + bx + c = 0$$

Como es del dominio público, las raíces están dadas por:

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad \text{si} \qquad b^2 - 4ac > 0$$

Para analizar la condición de este problema, necesitamos averiguar cómo varía x cuándo cambiamos un poco a, b y c.

Antes de hacer un análisis general, analicemos un ejemplo numérico en el sistema $Q(10, 8, -50, 50)$[1]

Sean: $a = 1.0000000$

$b = -4.0000000$

$c = 3.9999999$

Las raíces reales de la ecuación son:

$$x_1 = 1.999683772$$

$$x_2 = 2.000316228$$

Pero aplicando la fórmula obtenemos:

$$x_1 = x_2 = 2.000000000$$

Con sólo cuatro dígitos correctos; de hecho, las soluciones obtenidas son soluciones exactas de la ecuación que tiene como coeficientes a:

$$a = .999999992$$

$$b = -3.999999968$$

$$c = 3.999999968$$

[1]Forsythe, George. Pitfalls in Computation, or Why a Math Book isn't enough. American Mathematical Monthly.

Este es un ejemplo de inestabilidad del problema, ya que pequeños cambios en los datos, producen cambios mayores en los resultados.

Una forma sencilla de analizar esto es observar la magnitud de las derivadas parciales de x con respecto a a, b, c, y, en aquellos lugares donde éstas sean muy grandes, estarán los casos mal condicionados.

Si hacemos:

$$P(z) = az^2 + bx + c$$

Entonces si x es raíz tenemos:

$$ax^2 + bx + c = 0$$

Derivando parcialmente con respecto a a obtenemos:

$$x^2 + 2ax\frac{\delta x}{\delta a} + b\frac{\delta x}{\delta a} = 0$$

Despejando:

$$\frac{\delta x}{\delta a} = \frac{-x^2}{2ax + b}$$

Como:

$$p'(z) = 2az + b$$

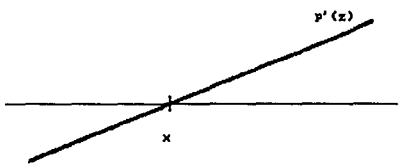
Tenemos:

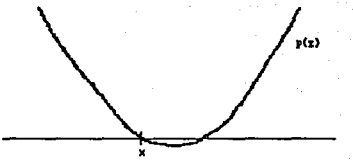
$$\frac{\delta x}{\delta b} = \frac{-x}{p'(x)}$$

$$\frac{\delta x}{\delta c} = \frac{-1}{p'(x)}$$

Ahora observese que no se nota nada extraño en las expresiones, salvo que cuando $p'(x)$ sea muy cercano a cero las variaciones pequeñas en a, b y c producirán cambios grandes en la raíz y, por consiguiente, harán más difícil su cálculo independientemente del método de solución.

Leamos bien el resultado anterior, dice que aquellas raíces x de $p(z)$, en las que la derivada en x, $p'(x)$ es muy pequeña, son muy sensibles a perturbaciones en los coeficientes.





Para localizar los valores correspondientes de a, b y c que dan lugar a esta situación es necesario evaluar $p'(z)$ en la raíz x:

$$x = \frac{b + \sqrt{b^2 - 4ac}}{2a}$$

Lo que es fácil en este caso ya que:

$$p'(x) = 2ax + b$$

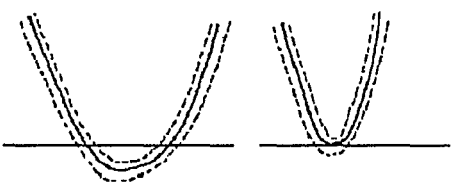
$$p'(x) = 2a\frac{-b + \sqrt{b^2 - 4ac}}{2a} + b$$

$$= -b + \sqrt{b^2 - 4ac} + b$$

$$= \sqrt{b^2 - 4ac}$$

Es decir que aquellos valores a, b, c que hacen que $b^2 - 4ac$ sea muy pequeña son los que hacen al problema mal condicionado. Afortunadamente podemos hacer entender este fenómeno a través de una explicación geométrica.

Notamos que $b^2 - 4ac = 0$ significa que tenemos una sola raíz real, pero doble: $b^2 - 4ac$ pequeño y positivo significa que la ecuación tiene dos raíces reales pero muy cercanas; $b^2 - 4ac$ positivo y muy grande significa que se tienen dos raíces reales pero muy separadas. Veamos geométricamente estos casos y el efecto que se produce por cambios en los coeficientes sobre las raíces.



Observemos que el caso $b^2 - 4ac \approx 0$ manifiesta su mala condición por el hecho de que hay perturbaciones para las cuales la ecuación no tiene raíces (éstas desaparecen).

De este ejemplo podemos inducir que la condición de un problema desempeña un papel importante en el momento de su solución numérica.

Condición de una función

Es necesario contar con una medida de la sensibilidad de una función a variaciones. Una de estas medidas es la derivada de la función a evaluar en el caso de una sola variable, ya que:

$$F(x) - F(x_0) \approx F'(x_0)\,(x - x_0)$$

Si x_0 está muy cerca de x.

Es decir $|F'(x_0)|$ nos mide la sensibilidad de F a variaciones pequeñas en una vecindad de x_0

Sin embargo, en la práctica esta medida no es suficiente, ya que debido al redondeo, es decir, a que no podemos trabajar con toda la expansión decimal de un número, sino sólo con su parte más significativa, es necesario analizar cómo se afectan los cambios relativos por lo que cambiamos la relación anterior a:

$$\left|\frac{F(x) - F(x_0)}{F(x_0)}\right| \approx \left|\frac{F'(x_0)}{F(x_0)}\right| \cdot |x_0| \left|\frac{x - x_0}{x_0}\right|$$

Para x muy cerca de x_0. En este caso, el número:

$$\left| \frac{F'(x_0) \times x_0}{F(x_0)} \right|$$

Nos mide como afecta el cambio relativo en F. A este número se le llama condición de F en x_0

Definición:

$$C(F, x_0) = \left| \frac{x_0 F'(x_0)}{F(x_0)} \right|$$

Los siguientes ejemplos nos permitirán apreciar la diferencia entre $|F'(x_0)|$ y $C(F, x_0)$

Ejemplos:

a) $F_1(x) = x \cdot a \qquad F_1'(x) = a$

$$C(x \cdot a, x_0) = \left| \frac{x_0 \times a}{x_0 \times a} \right| = 1$$

Independientemente de $a \neq 0$, la multiplicación nunca amplifica los errores relativos.

b) $F_2(x) = \sqrt{x} \;\; : \;\; F_2'(x) = \dfrac{1}{2\sqrt{x}}$

$$C(\sqrt{x}, x_0) = \left| \frac{x_0 \cdot \frac{1}{2\sqrt{x_0}}}{\sqrt{x_0}} \right| = \frac{1}{2}$$

Es decir, la extracción de raíz cuadrada reduce los errores relativos a la mitad.

c) $F_3(x) = \dfrac{1}{x} : \quad F_3'(x) = -\dfrac{1}{x^2} \; ; C(1/x, x_0) = 1$

d) $F_4(x) = x + a: \quad F_4'(x) = 1: \quad a \neq 0$

$$C(x + a, x_0) = \left| \frac{x_0}{x_0 + a} \right|$$

Si x_0 esta muy cercana de a, F es muy sensible a errores.

Es interesante observar que, paradójicamente, la operación de suma tiene una condición muy grande en la velocidad de $-a$.

Observemos que si:

$$x_0 = -a(1 + 10^{-k})$$

Entonces:

$$x_0 + a = -a \cdot 10^{-k}$$

$$\frac{1}{x_0 + a} = \frac{-10^k}{a}$$

$$\left| \frac{x_0}{x_0 + a} \right| = \frac{a\,(1 + 10^{-k})}{a} \cdot 10^k$$

Por lo tanto:

$$C(x + a, -a(1 + 10^k)) \approx 10^k$$

En consecuencia:

$$\frac{|\, x_0 + a - (x_0 + a)\,|}{|\, x_0 + a\,|} \approx 10^k \cdot epsmch$$

Nota: epsmch es la epsilón de la máquina.

Si $epsmch \approx 10^{-t}$

Entonces:

$$\frac{|\, x_0 + a - (x_0 + a)\,|}{|\, x_0 + a\,|} \approx 10^{-(t-k)}$$

Lo que quiere decir que se pierden k cifras en el resultado.

Este fenómeno es conocido como cancelación desastrosa, es importante evitarlo en la evaluación de funciones y en general en el cálculo numérico.

Inestabilidad en la evaluación de funciones

Para explicar la inestabilidad de algunos métodos numéricos es importante observar que, debido al redondeo, cuando queremos evaluar una función elemental $F(x)$ lo que obtenemos es sólo un valor aproximado $F_c(x)$ que es usualmente bueno en terminos relativos, ya que satisface:

$$\left| \frac{F_c(x) - F(x)}{F(x)} \right| \approx C(F, x) \cdot epsmch$$

Donde *epsmch* es un número muy pequeño, asociado con el redondeo, y $C(F, x)$ es la condición de F en x. Este valor $F_c(x)$ es el que usaremos en cálculos intermedios cuando $F(x)$ sea requerido.

Veamos ahora cómo afecta esto un caso particular. Supongamos que tenemos el problema de evaluar una función $F(x)$ en x_0 y que para ello contamos con dos expresiones:

$$F(x) = f(x) - g(x)$$

$$F(x) = h(x) \cdot k(x)$$

Queremos ver cuál es la más adecuada. Supongamos también que f, g, h y k están bien condicionadas en x_0, es decir, que se pueden evaluar en x_0 con un buen grado de aproximación.

Consideremos primero el caso:

$$F(x) = f(x) - g(x)$$

Al calcular $f(x)$ y $g(x)$ lo que obtenemos es $f_c(x)$ y $g_c(x)$, debido al redondeo, y estos valores nos producen el valor:

$$F_c(x) \approx f_c(x) - g_c(x)$$

Este valor no será muy preciso en los casos en que $f(x)$ sea muy parecido a $g(x)$, pues como sabemos la condición de la operación de diferencia es muy grande en este caso, es decir, que la mencionada operación de diferencia de los números amplifica mucho los errores relativos, cuando estos son muy parecidos. Lo que nos permite concluir que este procedimiento es potencialmente inestable.

Consideremos ahora el caso:

$$F(x) = h(x) \cdot k(x)$$

De manera semejante, al calcular $h(x)$ y $k(x)$, obtendremos $h_c(x)$ y $k_c(x)$; éstos nos producen el valor:

$$F_c(x) \approx h_c(x) \cdot k_c(x)$$

En este caso el valor obtenido será de la misma precisión que los factores, debido a que la multiplicación no amplifica nunca los errores relativos. Lo anterior nos permite concluir que este caso es estable, pues errores pequeños no son amplificados.

Ejemplos:

a) $F(x) = x - \sqrt{x^2 + 1} \; : \; x > 0$

Si $x \gg 1$ se generaran problemas, ya que $f_c(x)$ y $g_c(x)$ serán similares: habrá cancelación desastrosa.

Es decir para las x mucho mayores que 1 ($x \gg 1$), el valor obtenido tendrá mucho menor precisión que x. Eso nos motiva a buscar otra expresión para $F(x)$

$$F(x) = x - \sqrt{x^2 + 1} = x - \sqrt{x^2 + 1} \cdot \frac{x + \sqrt{x^2 + 1}}{x + \sqrt{x^2 + 1}}$$

$$F(x) = \begin{cases} \dfrac{x^2 - (x^2 + 1)}{x + \sqrt{x^2 + 1}} = \dfrac{-1}{x + \sqrt{x^2 + 1}} & si\ x \gg 1 \\ x - \sqrt{x^2 + 1} & si\ x < 1 \end{cases}$$

Observemos que esta expresión es estable siempre, ya que la suma en el denominador no presenta problema porque estamos considerando positiva a x y la división tampoco amplifica los errores:

b) $F(x) = 1 - \cos x$

Cuando x cercano a cero:

$$\cos x \approx 1$$

Por lo que anticipamos que habrá problemas numéricos, es decir, la evaluación directa amplificará muchos errores pequeños en una vecindad de $x = 0$. Busquemos otra expresión:

$$F(x) = 1 - \cos x = (1 - \cos x)\frac{1 + \cos x}{1 + \cos x}$$

$$= \frac{1 - \cos^2 x}{1 + \cos x} = \frac{sen^2 x}{1 + \cos x}$$

Es fácil ver que esta expresión es estable para x cercana a cero:

c) $x = \dfrac{-b + \sqrt{b^2 - 4ac}}{2a}$

Si b^2 es mucho mayor que $4ac$ ($b^2 \gg 4ac$) y además $b > 0$, entonces:

$$\sqrt{b^2 - 4ac} \approx b$$

Por lo que la evaluación de la función será inestable debido a la cancelación.

Por lo tanto, es necesario representar la raíz x de manera diferente.

Usando el mismo método que en los ejemplos anteriores, tenemos que:

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

$$= \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)\left(\frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}}\right)$$

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

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

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

Estabilidad de métodos numéricos

En la práctica cotidiana de la ingeniería, se presenta la necesidad de evaluar una gran cantidad de fórmulas, algunas muy sencillas y otras muy complejas. Cuando utilizamos la computadora para evaluar una fórmula, es natural separar los cálculos en fórmulas parciales y juntarlas para obtener el resultado final. A continuación tenemos una descripción abstracta de lo anterior, que facilitará la comprensión de lo que ocurre cuando un algoritmo o método numérico es estable.

Métodos estables:

Supongamos que tenemos que evaluar una función $f(x)$; un *método de evaluación* de la función es la "*descomposición*" de $F(x)$ en un número finito de funciones elementales, es decir:

$$F(x) = f_n \circ \ldots \circ f_2 \circ f_1(x)$$

Donde f_i es una función elemental y $\circ$ denota composición de funciones. Así, un algoritmo para el método anterior sería el de la forma siguiente:

Algoritmo:

$$y_0 = x$$

Para $i = 1, 2, \ldots n$

$$y_i = f_i(y_{i-1})$$

Resultado: $x = f(x) = y_n$

Para que este método sea estable, es necesario que la condición de las funciones elementales f_i, en y_{i-1}, sea lo más pequeña posible.

Es decir, el algoritmo anterior es estable si:

$$C(f_i, y_{i-1}) \text{ es pequeña.}$$

Con respecto a la precisión del sistema numérico de punto flotante:

Ejemplos:

a) $f(x) = x - \sqrt{x^2 + 1}$

Método A $y_0 = x$

$$y_1 = \sqrt{y_0{}^2 + 1}$$

$$y_2 = y_0 - y_1$$

Este método es inestable para $x \gg 1$ porque, en ese caso:

$$x \approx \sqrt{x^2 + 1}$$

Método B $y_0 = x$

Si $y_0 < 1$

$$y_1 = \sqrt{y_0{}^2 + 1}$$

$$y_2 = y_0 - y_1$$

Si $y_0 > 1$

$$y_1 = \sqrt{y_0{}^2 + 1}$$

$$y_2 = \frac{-1}{y_0 + y_1}$$

Este método es estable, como se vio en la sección anterior.

SUMATORIAS

Existen muchas situaciones en las que es necesario efectuar sumas con una gran cantidad de sumandos; en ocasiones, esas sumas son infinitas, lo que da origen a las series.

En estadística, por ejemplo, la media y la varianza, conceptos medulares de la misma, se definen de la siguiente manera:

$$\mu_n = \frac{1}{n} \sum_{k=1}^{n} x_k$$

$$\sigma_n{}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \mu_n)^2$$

Por otro lado, algunas funciones, como e^x, se pueden definir mediante series:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Por lo tanto, es prioritario que estudiemos el comportamiento de los errores que se originan cuando se efectúa la sumatoria en una computadora, para determinar métodos que minimicen el error final de la suma. Aparentemente la suma es la operación que menos complicaciones tiene para su cómputo pero, como se verá a través de un par de ejemplos, esto no es necesariamente cierto.

Ejemplo 1

$$\text{Sea} \quad S = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{10{,}000} = \sum_{n=1}^{10{,}000} (1/n)$$

Desde el punto de vista analítico, la suma es la misma, independientemente del orden de los sumandos; esto es:

$$\text{Si} \quad S_F = \frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{10{,}000}$$

$$\text{Y} \quad S_B = \frac{1}{10{,}000} + \frac{1}{9999} + \ldots + \frac{1}{2} + \frac{1}{1}$$

Entonces:

$$S_F = S_B$$

Pero, en la práctica esto no es cierto, ya que si calculamos S_F y S_B en una microcomputadora, con el siguiente programa BASIC, obtenemos valores diferentes:

```
5 REM Cálculo de sumatorias (sumatoia)
10 S=0
20 FOR-I=1 TO 10 000      FOR I=10000 TO 1 STEP-1
30 N=1/I
40 S=S+N
50 NEXT I
60 PRINT S
```

Si usamos el programa original, obtenemos como suma:

$$S_F = 9.787613$$

Si sustituimos la línea 20 para hacer el cálculo del número menor, 1/10000, al número mayor 1, obtenemos:

$$S_B = 9.787604$$

¿Cuál resultado es mejor?

Para responder a esta pregunta, en general, definamos:

$$S = a_1 + a_2 + \ldots + a_n = \sum_{j\,01}^{n} a_j$$

Donde cada a_j es un número en punto flotante. Al sumar estos valores en la computadora, obtenemos una sucesión de sumas parciales, cada una de las cuales tendrá un error de redondeo. Esto es:

$$S_2 = (a_1 + a_2)(1 + \varepsilon_2)$$
$$S_3 = (S_2 + a_3)(1 + \varepsilon_3)$$
$$\vdots$$
$$S_n = (S_{n-1} + a_n)(1 + \varepsilon_n)$$

Mediante transformaciones algebraicas, obtenemos:

$$S_2 = (a_1 + a_2) = (a_1 + a_2)\,\varepsilon_2$$

$$S_3 - (a_1 + a_2 + a_3) = (a_1 + a_2)\varepsilon_2 + (a_1 + a_2)(1 + \varepsilon_2)\varepsilon_3 + a_3\varepsilon_3$$

$$\approx (a_1 + a_2)\varepsilon_2 + (a_1 + a_2 + a_3)\varepsilon_3$$

$$S_4 - (a_1 + a_2 + a_3 + a_4) \approx (a_1 + a_2)\varepsilon_2 + (a_1 + a_2 + a_3)\varepsilon_3 + (a_1 + a_2 + a_3 + a_4)\varepsilon_4$$
$$\vdots$$

$$S_n - \sum a\,i = (a_1 + a_2)\varepsilon_2 + \ldots + (a_1 + a_2 + \ldots + a_n)\varepsilon_n$$
$$= x_1(\varepsilon_2 + \varepsilon_3 + \ldots + \varepsilon_n) + x_2(\varepsilon_2 + \varepsilon_3 + \ldots + \varepsilon_n) + x_3(\varepsilon_3 + \varepsilon_4 + \ldots + \varepsilon_n) + \ldots x_n\varepsilon_n$$

De aquí, concluimos que la mejor estrategia para la evaluación o el cálculo de una sumatoria es efectuar la misma en orden ascendente de magnitud de los sumandos; esto es, ordenar los sumandos de tal forma que:

$$|a_1| \le |a_2| \le \ldots \le |a_n|$$

Ejemplo 2

La siguiente expresión es conocida y válida para todos los números reales x

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

Por lo tanto:

$$e^{-x} = 1 - \frac{x}{1!} + \frac{x^2}{2!} - \frac{x^3}{3!} + \ldots$$

Supongamos que usamos una calculadora o dispositivo digital decimal de precisión 4; esto es, el sistema de números en punto flotante es $Q(10, 4, m, n)$ y queremos calcular:

$$e^{-5} \approx 1 + \frac{(-5)}{1!} + \frac{(-5)^2}{2!} + \frac{(-5)^3}{3!} + \ldots + \frac{(-5)^n}{n!}$$

En la siguiente tabla calculamos los 25 primeros términos de la sumatoria y las respectivas sumas parciales.

Grado	Término	Suma
0	1.000	1.000
1	-5.000	-4.000
2	12.50	8.50
3	-20.83	-12.33
4	26.04	13.71
5	-26.04	-12.33
6	21.70	9.370
7	-15.50	-6.130
8	9.688	3.558
9	-5.382	-1.824
10	2.691	.8670
11	-1.223	-.3560
12	.5097	.1537
13	-.1960	-.04230
14	.7001E-1	.02771
15	-2334E-1	.004370
16	.7293E-2	.01166
17	-.2145E-2	.009518
18	.5958E-3	.01011
19	-.1568E-3	.009957
20	.3920E-4	.009996
21	-.9333E-5	.009987
22	.2121E-5	.009989
23	-.4611E-6	.009989
24	.9607E-7	.009989
25	-.1921E-7	.009989

El valor de e^{-5} con cuatro cifras significativas es de .006738; como se podrá observar en la tabla, este valor es completamente diferente a cualquiera de las sumas parciales, ¿cuál es el problema?

Como seguramente se habrá observado, la serie para calcular e^{-x} es una serie que tiene, de manera alternada, números positivos y negativos; este hecho, como se vio

en secciones anteriores, origina problemas. En este ejemplo, se puede evitar el problema si se calcula:

$$e^{-5} = \frac{1}{e^5}$$

Y se calcula e^5 con una serie de términos positivos, evitando así la cancelación desastrosa; pero, en general, lo que se recomienda es la separación de la serie en dos: S_p y S_n, donde S_p es la sumatoria de los términos positivos y S_n es la de los negativos, por lo que la suma estará dada por:

$$S = S_p - S_n$$

POLINOMIOS

Muchos problemas de ingeniería implican el uso de polinomios y, en particular, la evaluación de los mismos, esto es uno de los cálculos numéricos más comunes y que se efectuan desde la educación secundaria. El hecho parecería indicar que la evaluación de polinomios no representa mayor problema. No es así, y para ilustrarlo consideremos la evaluación del polinomio:

(1) $p(x) = 2x^4 - 19x^3 + 56.98x^2 - 56.834x + 5.1324$

El método más sencillo para calcular el valor del polinomio en un valor específico de x, es el cómputo independiente de cada uno de los términos. Esto es, el término $a_k x^k$ es calculado multiplicando k veces x y finalmente multiplicando por a_k.

En el ejemplo, para calcular $2x^4$ se requiere de 4 multiplicaciones; por lo tanto, para evaluar (1) requerimos de $4 + 3 + 2 + 1 = 10$ multiplicaciones y 3 sumas.

El segundo método de evaluación es más eficiente, ya que sólo requiere de 7 multiplicaciones, y consiste en calcular cada potencia de x mediante la multiplicación por x de la potencia anterior, de la siguiente manera:

$$x^3 = x(x^2) \qquad x^4 = x(x^3)$$

Por lo que cada término $a_k x^k$ sólo requiere de dos multiplicaciones para su cálculo para toda $k > 1$. Esto representa un ahorro considerable en el número de operaciones, sobre todo cuando se tienen polinomios de grados altos.

El tercer método -más eficiente- es el de la multiplicación anidada o método de Horner. Es, a final de cuentas, el de la división sintética. Se deriva de la expresión anidada del polinomio.

Dado:

$$p(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx_n \qquad a_n \neq 0$$

Entonces:

$$p(x) = a_0 + x(a_1 + x(a_2 + \ldots + x(a_{n-1} + a_nx) \ldots)$$

En nuestro ejemplo tenemos que:

$$p(x) = 5.1324 + x\,(-56.834 + x(56.98 + x(-19 + 2x)))$$

En este caso, se efectuan sólo 4 multiplicaciones. El método de la multiplicación anidada se prefiere no solamente por el ahorro de operaciones, sino porque el cálculo de potencias, usado en el primero, es más lento e impreciso y puede ocasionar problemas de cancelación cuando la x es negativa como se vio en la sección anterior. El método de Horner es el recomendado para la evaluación de polinomios.

Fuente común de polinomios es la aproximación de funciones más complicadas a través de polinomios de Taylor.

Así, el polinomio de Taylor $P_5(x)$ para aproximar a $log(x)$ alrededor de $a = 1$ esta dado por la expresión:

$$P_5(x) = (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 - \frac{1}{4}(x-1)^4 + \frac{1}{5}(x-1)^5$$

Si hacemos $z = (x-1)$ tenemos que:

$$P_5(x) = z\,(1 + z\,(-\frac{1}{2} + z\,(\frac{1}{3} + z\,(-\frac{1}{4} + \frac{1}{5}\,z))))$$

El algoritmo para evaluar $p(x)$ en algún número z está definido mediante la siguiente sucesión:

$$\begin{aligned} & b_n = a_n \\ & b_{n-1} = a_{n-1} + zb_n \\ (2) \quad & b_{n-2} = a_{n-2} + zb_{n-1} \\ & \vdots \\ & b_0 = a_0 + zb_1 \end{aligned}$$

Por lo tanto, $p(z) = b_0$

Las b_i son los cálculos sucesivos de cada uno de los paréntesis, que corresponden a la última línea de cálcu-

los en la división sintética, como se muestra a continuación:

Sea $p(x) = 2x^4 - 19x^3 + 56.98x^2 - 56.834x + 5.1324$

Que, en su forma anidada, es:

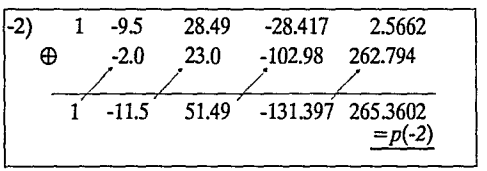
$$p(x) = (((2x - 19)x + 56.98)x - 56.834)x + 5.1324$$

Y dividiendo entre 2 tenemos:

$$p(x) = (((1 \cdot x - 9.5)x - 28.49)x - 28.417)x + 2.5662$$

Si queremos evaluar $p(x)$ en $x = -2$ usando la división sintética, tenemos:

$$p(x) = (((1x - 9.5)x - 28.49)x - 28.417)x + 2.5662$$



Nota 1: *La primera línea corresponde a los coeficientes del polinomio.*

Nota 2: *La tercera línea corresponde a las* b_i

Nota 3: *El último valor de la tercera línea es p(-2)*

Nota 4: *Las diagonales indican multiplicación por -2*

RELACIONES DE RECURRENCIA

Muchos problemas y sus soluciones son formulados frecuentemente en términos de algún proceso infinito, (se vio en la sección anterior) mediante el uso de series para calcular algunas funciones como la exponencial.

Si bien es cierto que las series son de los procesos infinitos más comunes para la evaluación de funciones trascendentales, existen otros, los iterativos o recurrentes. Éstos en general son más eficientes que las series y pueden usarse en situaciones en las que éstas son inadecuadas.

Ejemplos de procesos iterativos:

1. Para el calculo de *e*

$$x_0 = 1$$

$$x_k = x_{k-1} + (1/k!) \qquad k = 1, 2, \ldots$$

2. Para el cálculo de π

$$y_1 = 2$$

$$y_2 = \sqrt{2}$$

$$y_{k+1} = y_k \sqrt{(2y_k)/(y_{k-1} + y_k)}$$

3. Para el cálculo de $\sqrt{2}$

a) $x_0 = 2$

$$x_k = \frac{1}{2}(x_{k-1} + (2/x_{k-1})) \qquad k = 1, 2, \ldots$$

La recursividad genera dos situaciones críticas:

a) Dado que se genera una sucesión $\{x_n\}$ de valores, y los recursos de cómputo son finitos, ¿cuándo y cómo detenemos el proceso?

b) En virtud de que x_n se calcula a partir de los términos anteriores de la sucesión y sabemos que todo valor calculado en una computadora incluye un error de redondeo, surge la pregunta natural: ¿Los errores de los primeros términos son magnificados por el proceso? En otras palabras, ¿es la sucesión convergente al valor deseado?

En esta sección exploraremos y responderemos esta última cuestión. Para ello tomaremos como ejemplo el problema de determinar el valor de la siguiente integral:

$$I_k = \int_0^1 x^k \exp(x-1)\, dx$$

Integrando por partes, tenemos:

$$I_k = 1 - k I_{k-1}$$

Y como:

$$I_0 = \int_0^1 e^{x-1}\, dx$$

$$= e^{-1}(e^x 1_0^1)$$

$$= e^{-1}(e - 1)$$

$$= 1 - e^{-1}$$

Entonces podemos usar las recursión:

$$I_0 = 1 - e^{-1}$$

$$I_k = 1 - k I_{k-1}$$

Para calcular:

$$\int_0^1 x^{20} e^{x-1}\, dx$$

Calculando la sucesión de valores $I_1, I_2, \ldots I_{20}$ Estos se presentan en la siguiente tabla:

k	I_k
1	.3678795
2	.2642411
3	.2072767
4	.1708932
5	.145534
6	.1267958
7	.1124296
8	.1005631
9	9.493256E-02
10	5.067444E-02
11	.4425812
12	-4.310974
13	57.04267
14	-797.5973
15	11964.96
16	-1911438.4
17	3254453
18	-5.858015E+07
19	1.113023E+09
20	-2.226046E+10

Los valores se calcularon en una micro AT, con el siguiente programa en BASIC

```
10 REM calculo integral
20 Y=1-(1/EXP(1))
30 FOR I=1 TO 20
40 Y=1-I*Y
50 PRINT Y
60 NEXT I
70 END
```

Esos resultados contradicen el hecho de que:

$$I_k = \int_0^1 x^k e^{x-1}\, dx > 0$$

Y $\lim_{k \to \infty} I_k = 0$

Muchos procesos iterativos padecen este problema. En este caso, el problema se genera porque los errores de I_{k-1} son multiplicados por k y transmitidos a I_k; por lo que concluimos que: en el cómputo numérico, las fórmulas y algoritmos, aun los sencillos y aparentemente inofensivos, deben ser empleados con precaución. Pero, ¿cómo resolvemos o eliminamos la inestabilidad de nuestro método?

Una relación recursiva que es inestable en un sentido, en orden ascendente, por ejemplo, en general, no lo es en sentido inverso.

Si despejamos I_{k-1} de $I_k = 1 - k I_{k-1}$ tenemos que:

$$I_{k-1} = \frac{1 - I_k}{k}$$

Si usamos esa fórmula, necesitamos también un valor inicial, que en principio parece difícil de encontrar, ya que los valores I_k son precisamente los que tratamos de determinar.

Sin embargo, sabemos que:

$$I_k \to 0 \quad \text{cuando } k \to \infty$$

Y hagamos:

$$I_{20} = 0$$

Usando:

$$I_{20} = 0$$

$$I_{k-1} = \frac{1 - I_k}{k}$$

Para $k = 20, 19, 18, \ldots$ los resultados que obtenemos son:

k	I_k
20	.05
19	.05
18	5.277778E-02
17	5.571896E-02
16	5.901757E-02
15	6.273217E-02
14	6.694771E-02
13	7.177326E-02
12	7.735223E-02
11	8.387707E-02
10	9.161229E-02
9	.100932
8	.1123835
7	.1268024
6	.145533
5	.1708934
4	.2072766
3	.2642411
2	.3678795
1	.6321206

Obtenidos con el programa:

```
10 REM calculo integral
20 Y=0
30 FOR I=20 TO 1 STEP -1
40 Y=(1-Y)/I
50 PRINT Y
60 NEXT I
70 END
```

El último valor coincide con el valor inicial correcto de la iteración ascendente, y todos los demás valores son correctos, a pesar del hecho de que fueron calculados con un valor inicial $I_{20} = 0$ erróneo.

En conclusión, si se tiene un proceso iterativo o recurrente que es inestable en un sentido se corrige si lo calculamos en sentido inverso.

EJERCICIOS Y PROBLEMAS*

1. Use la serie de Taylor para reescribir las siguientes expresiones, de tal manera que puedan ser evaluadas evitando la cancelación desastrosa (o sustractiva).

   (a) $tan\, x - cot\, x, x \approx \pi/4$

   (b) $1 - e^x, x \approx 0$

   (c) $sin^2 x - cos^2 (x + \pi/2), x \approx 0$

2. Reescriba las siguientes expresiones de tal manera que puedan evaluarse sin cancelación sustractiva.

   (a) $\sqrt{1 + x^2} - \sqrt{1 - x^2}$, x cercana a 0

   (b) $1/\left[\sqrt{1 + x^2} - \sqrt{1 - x^2}\right]$, x cercana a 0

   (c) $(1 + x)^2 - (1 - x)^2$, cercana a 0

   (d) $[-b - \sqrt{b^2 - 4ac}\,] / 2a, b < 0,$ a y c son muy pequeñas.

3. Para $f(x) = sen\,(\pi x)$ use la fórmula:

$$f'(1) \approx [f(2 + h) - f(1)]/h$$

   Con diversos valores de h para aproximar $f'(1) = -\pi$. ¿Qué tan pequeña debe ser h antes de que la cancelación sustractiva se convierta en un problema?

*Los ejercicios se tomaron y/o adaptaron de las obras citadas al final de esta sección.

4. Dado $Q_M = Q(10, 3, m, M)$ calcule $\sum_{k=1}^{\infty} 0.3$ asociando los sumandos en el orden natural. ¿Cuántos términos contribuyen a la suma obtenida?

5. Considere la ecuación $a \otimes x = b$, con $a = .111(10^0)$, $b = .200(10^1)$:

   a) Calcule la solución en Q_M.

6. Considere el polinomio $P(x) = x^3 - 3x^2 + 3x - 1$. Evalúelo en Q_M, en la forma:

$$(((x \otimes x) \oplus 3) \otimes x) - ((3 \otimes (x \otimes x)) \oplus 1)$$

   Para llenar la siguiente tabla:

x	.900	.950	1.00	1.05	1.10
$P(x)$					

   Usando exclusivamente los valores de la tabla, describa la gráfica de la función alrededor de $x = 1$. Comente.

7. a) Considere el número 5.15, $Q(10, 5, m, M)$. ¿Cuál es su equivalente en $Q(3, 10, m, M)$?

   b) Si 2.101 es un elemento de $Q(8, 3, m, M)$. ¿Cuál es su equivalente en $Q(4, 10, m, M)$?

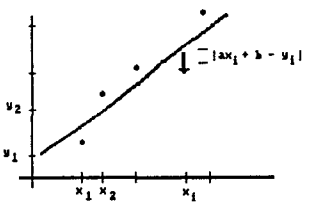
8. Grafique la parte positiva de $Q(10, 3, m, M)$

9. Considere el polinomio de grado n

$$a_0 + a_1x + ... + a_nx^n$$

Describa, por lo menos, dos métodos para evaluarlo en una x dada.

10. Considere dados los m puntos (x_i, y_i), $i = 1, 2, ..., m$. Un problema común consiste en encontrar los coeficientes de la recta $y = ax + b$, que se ajusta mejor a los puntos dados. Una forma usual para resolver este problema es vía el método de "mínimos cuadrados", que consiste en calcular aquellos valores de a y b que minimicen la función $S(a, b) = \sum (ax_i + b - y_i)^2$, según la siguiente figura:



b) Encuentre expresiones para a y b resolviendo el sistema.

11. Calcule $\sum_{k=1}^{\infty} (.129)$ en $Q(10, 3, m, M)$, asociando los sumandos en el orden natural. Justifique su respuesta.

12. Demuestre que el problema que consiste en evaluar la función $f(x) = 1 - \cos x$, en el intervalo $[-\pi, \pi]$, es estable en el sentido de que su condición está acotada por 2 en dicho intervalo, es decir $C_0(f, x_0) \leq 2$, $-\pi \leq x_0 \leq \pi$.

A continuación, se pretende analizar dos métodos distintos para resolver el problema en x "pequeñas":

a) Sea $f_c(x) = 1 - \cos_c(x(1+\delta x))$, y suponga que la evaluación realista de $\cos x$ es tal que $\cos_c(x(1+\delta x)) = (\cos x)(1+\delta_c)$, en donde $|\delta_c| \approx |\delta_x|$ «1. ¿De qué orden es el error relativo al aproximar $f(x)$ por $f^1_c(x)$, para $x \approx \delta_x$?

b) Sea $f^2_c(x) = (x(1+\delta_x))^2/2$, es válido tomar esta aproximación a $f(x)$, pues para x "pequeñas" $\cos x = 1 - x^2/2$. ¿Cómo es en este caso el error relativo?

c) Con base en los resultados anteriores, y tomando en cuenta que el problema es estable, ¿qué puede comentar acerca de los métodos (a) y (b)?

13. Investigue cómo se evalúa $sen(x)$ en su computadora.

14. Demuestre que, para toda $k \in Q(10, n)$, $\sum_{i=1}^{\infty} k$ es la solución mínima de la ecuación $z \oplus k = z$, siguiendo los pasos que se indican: Sea $S = \sum_{i=1}^{\infty} k$, $G = \{z \in F_n(R) - z \oplus k = z\}$, y $z_0 = min(G)$.

a) Demuestre que $S \in G$

b) Si $z \in G$ y $z' \in F_n(R)$ es tal que $z' > z$, entonces $z' \in G$

c) Si $z \in G$ entonces $s_n(z) \in G$, en donde s_n es la función "sucesor"

d) Si z_1 y z_2 son elementos distintos de G entonces $/z_2 - z_1/ > k$

15. Encuentre un método numéricamente estable para evaluar $f(x)$, en los puntos indicados, y diga en cada caso cuál es el error relativo de la evaluación:

a) $f(x) = e^x - 1, x \approx 0$

b) $f(x) = (1 - \cos x) / \operatorname{sen} x, x \approx 0$

c) $f(x) = (x+1)^{1/3} - x^{1/3}, \; x \gg 1$

16. Considere dados los puntos $(x_1 y_1), (x_2 y_2), (x_3 y_3)$, con la propiedad $x_1 < x_2 < x_3$.

a) Construya el polinomio de segundo grado:

$$p(x) = Ax^2 + Bx + C$$

Que satisface las condiciones:

$$p(x_i) = y_i, \qquad i = 1, 2, 3$$

b) ¿Bajo qué condiciones P tiene un mínimo? Suponga que se dan esas condiciones y calcule el mínimo.

c) Exprese A y el mínimo en el caso particular:

$$x_1 = -h, x_2 = 0, x_3 = h$$

Referencias bibliográficas:

1. *Acton, Forman S.* "Numerical Methods that Work". New York. Harper & Row, 1970.

2. *Shampine, Lawrence F.* "Numerical Computing: an introduction". Philadelphia. Saunders Company, 1973.

3. *Vandergraft, James S.* "Introduction to Numerical Computations". New York. Academic Press, 1978

14. Demuestre que, para toda $k \in Q(10, n)$, $\sum_{i=1}^{\infty} k$ es la solución mínima de la ecuación $z \oplus k = z$, siguiendo los pasos que se indican: Sea $S = \sum_{i=1}^{g} k$, $G = \{z \in F_n(R) \text{ - } z \oplus k = z\}$, y $z_0 = min(G)$.

a) Demuestre que $S \in G$

b) Si $z \in G$ y $z' \in F_n(R)$ es tal que $z' > z$, entonces $z' \in G$

c) Si $z \in G$ entonces $s_n(z) \in G$, en donde s_n es la función "sucesor"

d) Si z_1 y z_2 son elementos distintos de G entonces $/z_2 - z_1/ > k$

15. Encuentre un método numéricamente estable para evaluar $f(x)$, en los puntos indicados, y diga en cada caso cuál es el error relativo de la evaluación:

a) $f(x) = e^x - 1, x \approx 0$

b) $f(x) = (1 - \cos x) / \text{sen } x, x \approx 0$

c) $f(x) = (x+1)^{1/3} - x^{1/3}, \; x \gg 1$

16. Considere dados los puntos $(x_1 y_1), (x_2 y_2), (x_3 y_3)$, con la propiedad $x_1 < x_2 < x_3$.

a) Construya el polinomio de segundo grado:

$$p(x) = Ax^2 + Bx + C$$

Que satisface las condiciones:

$$p(x_i) = y_i, \qquad i = 1, 2, 3$$

b) ¿Bajo qué condiciones P tiene un mínimo? Suponga que se dan esas condiciones y calcule el mínimo.

c) Exprese A y el mínimo en el caso particular:

$$x_1 = -h, x_2 = 0, x_3 = h$$

Referencias bibliográficas:

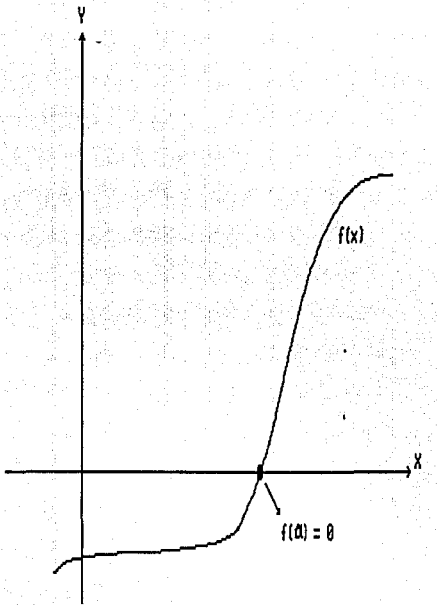
1. *Acton, Forman S.* "Numerical Methods that Work". New York. Harper & Row, 1970.

2. *Shampine, Lawrence F.* "Numerical Computing: an introduction". Philadelphia. Saunders Company, 1973.

3. *Vandergraft, James S.* "Introduction to Numerical Computations". New York. Academic Press, 1978

CAPíTULO III

RAÍCES O CEROS DE FUNCIONES ESCALARES



"La solución de las ecuaciones cúbicas depende de un juicio correcto, ayudado por Dios."

BIJA GANITA

(A Philosophical and Mathematical Dictionary, Londres 1815)

"El objetivo de la matemática concreta es descubrir las ecuaciones que expresan las leyes matemáticas de los fenómenos bajo consideración."

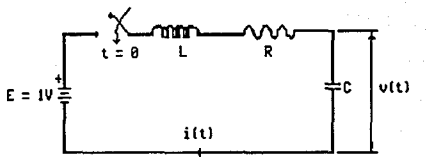
COMTE

(Positive Philosophy)

INTRODUCCION

En el circuito diagramado a continuación, el voltaje a través del condensador *C*, expresado en función del tiempo *t*, satisface la siguiente ecuación:

$$V(t) = 1 - 2e^{-t} + e^{-2t} \qquad (1)$$



L = 1, R = 3, C = 1/2

Si queremos determinar el valor de *t*, que hace que el voltaje se eleve a la mitad de su valor final, tenemos que resolver la siguiente ecuación:

$$V(t) - 0.5 = 0$$

Desafortunadamente, ésta como la mayoría de las ecuaciones, no se puede resolver con métodos analíticos. Ni aún las ecuaciones algebraicas (de grado mayor o igual

a cinco), como seguramente lo sabe el lector, se pueden resolver por métodos analíticos.

Otras ecuaciones que surgen de problemas como el anterior son las siguientes:

$$x \cosh \frac{50}{x} = x + 10$$

$$3.24x^8 - 2.42x^7 + 10.37x^6 + 10.01x^2 + 47.98 = 0$$

$$2^x - 10x + 1 = 0$$

$$\cosh(\sqrt{x^2+1}) - e^x + \log \left| \operatorname{sen} x \right| = 0$$

$$\operatorname{sen} x = x$$

En este capítulo consideraremos algunos métodos para resolverlas; esto es, encontrar las raíces de ecuaciones de la forma:

$$f(x) = 0 \qquad (2)$$

O, dicho de otra manera, determinar los ceros de la función:

$$f(x) \qquad (3)$$

Trataremos de encontrar el valor o los valores, de la variable x que satisfacen la ecuación 2. Hablaremos indistintamente de las raíces de la ecuación 2 ó de los ceros de la función 3

MÉTODOS

Seguramente el lector recuerda, de sus cursos de álgebra, el método analítico para resolver ecuaciones polinomiales de segundo grado; por ejemplo, dada la ecuación:

$$6x^2 - 7x + 2 = 0$$

Usando la "fórmula" obtenemos las dos raíces de la ecuación, a saber:

$$r_1 = \frac{1}{2}$$

$$r_2 = \frac{2}{3}$$

De hecho, existen "fórmulas" (creadas desde el siglo XVI) para obtener las soluciones de ecuaciones polinomiales de tercero y cuarto grados[1]; pero, como lo

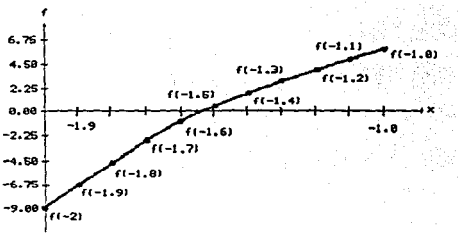
[1] Eves, Howard. An Introduction to the History of Mathematics. Holt, Rinehart

demostró el matemático noruego Niels Henrik Abel, en 1824, las raíces de la ecuación polinomial general de grado mayor o igual al quinto no pueden ser expresadas por medio de radicales, en términos de los coeficientes de la ecuación, como se hace para las ecuaciones de segundo, tercero y cuarto grados.

Por lo tanto, para resolver ecuaciones en general, es necesario recurrir a otros métodos; uno de ellos es la graficación de la función $f(x)$ para determinar el lugar donde esta función corta al eje x. Este método indica cuántas raíces reales hay y da una idea aproximada de sus valores. Mediante esta técnica se calculan algunos valores de la función y se interpolan los valores para los cuales la función cambia de signo.

Para ejemplificar, a continuación se tabulan algunos valores de una función $f(x)$ y se localizan en la gráfica siguiente algunos puntos $(x, f(x))$ de la función en el subintervalo $I = [-2, -1]$, con el objeto de localizar la raíz de dicha función a través de su gráfica.

x	f(x)	x	f(x)	x	f(x)
-10	-1281	-3	-42	4	21
-9	-954	-2	-15	5	54
-8	-687	-1	-2	6	111
-7	-474	0	9	7	198
-6	-309	1	6	8	321
-5	-186	2	3	9	486
-4	-99	3	6	10	699



Trazado de los puntos de una función para la solución grafica de una ecuación

Antes del advenimiento de las computadoras, el método gráfico fue el comúnmente usado para resolver ecuaciones.

En la actualidad, con computadoras y calculadoras a la mano, los métodos iterativos son la norma. Estos generan una sucesión, casi siempre convergente, a partir de una propuesta inicial de localización de la raíz.

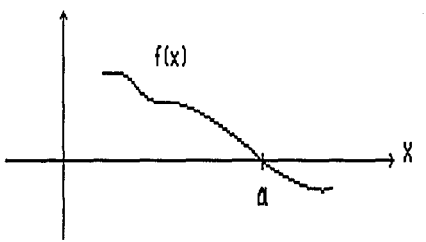
En algunos métodos se inicia con la propuesta inicial de un intervalo I_0 que contiene a la raíz α, se genera una sucesión de intervalos $\{I_n\}$ que contienen esa raíz. Denotamos estos métodos con el nombre génerico de *intervalos de seguridad* I_n

En los *métodos de interpolación polinomial* se genera una sucesión de puntos $\{x_n\}$, que cuando converge a la raíz, lo hace rápidamente.

Para aprovechar las ventajas de ambos métodos se generan *métodos híbridos* que, como su nombre lo indica, aplican técnicas de los métodos anteriores.

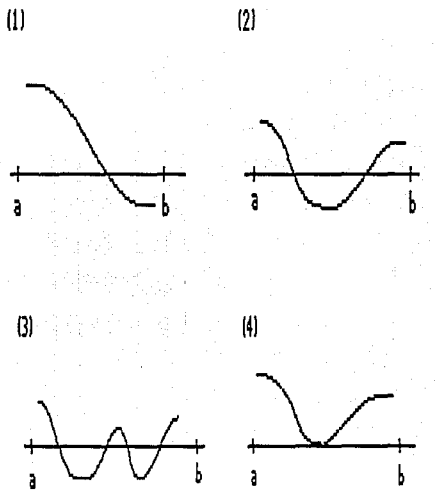
Comentarios generales previos

Dada una función real de variable real *(f:R → R)*, podemos reducir el problema de calcular sus raíces al cálculo de las mismas en un intervalo *[a, b]*.

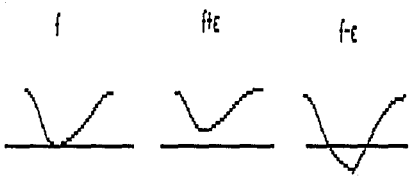


Si *f* es continua y cambia de signo en *[a, b]*, entonces tiene -cuando menos- un cero. Se pueden presentar las cuatro situaciones siguientes:

Algunas situaciones posibles:



En la figura (4) se presenta una situación difícil, en tanto que no es posible darse cuenta de la existencia del cero antes de resolver el problema. Además, el problema es inestable, pues pequeñas perturbaciones a la función cambian cualitativamente el problema: desaparecen los ceros o aparecen dos, en lugar de uno, debido a los errores de redondeo propios de las computadoras.



A) Métodos seguros

i) Se parte de un intervalo inicial I_0 en donde $f(x)$ tiene únicamente un cero α.

ii) Mediante algún procedimiento, se selecciona un subintervalo $I_{k+1} \subset I_k$, con la propiedad $\alpha \in I_{k+1}$, para $k = 0, 1, 2, \ldots$

iii) Generalmente se trata de conseguir que $\bigcap_{k=0}^{\infty} I_k = \{\alpha\}$

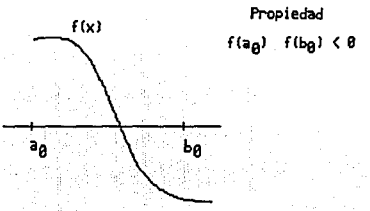
iv) Los métodos usuales son:

1) Bisección.
2) Regla falsa.
3) Regla falsa modificada.

Estos métodos nunca fallan pero pueden ser demasiado lentos.

1) Bisección

Sea $f(x)$ una función continua que tiene un único cero en (a_0, b_0):



Hay que tomar el punto medio $x_0 = \dfrac{a_0 + b_0}{2}$ y averiguar el signo de f en x_0:

si $f(x_0)f(a_0) < 0$ entonces $a_1 = a_0$, $b_1 = x_0$

si $f(x_0)f(a_0) > 0$ entonces $a_1 = x_0$, $b_1 = b_0$

si $f(x_0)f(a_0) = 0$ entonces $\alpha = x_0$

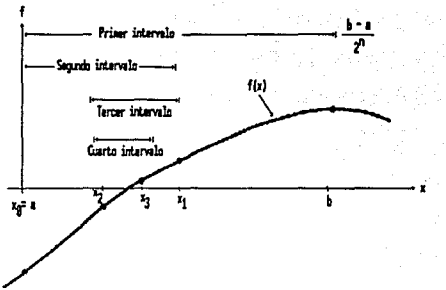
Verificamos la convergencia; es decir, cuando la longitud del intervalo es "suficientemente" pequeña, se detiene el procedimiento; si no, se repite eligiéndose un nuevo intervalo.

Si $l(I_n)$ es la longitud del intervalo I_n entonces:

$$l(I_n) = \frac{l(I_0)}{2^n}$$

$$\Rightarrow l(I_n) \underset{n \to \infty}{\to} 0$$

$$\Rightarrow a_n \underset{n \to \infty}{\to} b_n \underset{n \to \infty}{\to} x_n \underset{n \to \infty}{\to} \alpha$$



Ejemplo 1: La función $f(x) = x - 0.2 \operatorname{sen} x - 0.5$ tiene un cero en $I_0 = [0.5, 1.0]$, ya que $f(0.5)\, f(1.0) < 0$ y $f'(x) \neq 0$ en I_0. Por lo tanto, en el programa siguiente, en la línea 40 se indican los valores inciales del intervalo: $a = .5, b = 1$

Se considera que el intervalo I_n es suficientemente pequeño cuando $l(I_n) < Q = 1E\text{-}06$. (línea 30)

La raíz que se obtiene es:

$$\alpha = .6154685$$

Programa

```
10 REM METODO DE BISECCION
20 DEF FNF(X)=X-.2*SIN(X)-.5
30 LET Q=.000001
40 LET A=.5: LET B=1
50 LET F=FNF(A)
60 LET X0=(A+B)/2 : LET G = FNF(X0)
70 IF F*G<0 THEN LET B=X0: GOTO 90
80 LET A=X0:LET F=G
90 IF ABS(B-A)<Q THEN GOTO 120
100 LPRINT "LIMITES ";A;" Y ";B
110 GOTO 60
120 LPRINT:LPRINT"UNA RAIZ ES ";(A+B)/2
```

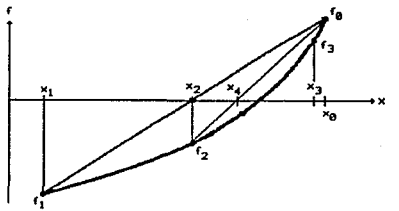
La tabla siguiente muestra los valores de los limites del intervalo A, B y los valores de la función es dichos puntos, $f(A)$ y $f(B)$

A	*B*	*f(A)*	*f(B)*
.5	.75	.0958851	.1136723
.5	.625	.0958851	7.980526E-03
.5626	.625	4.416055E-02	7.980526E-03
.59375	.625	1.814464E-02	7.980526E-03
.609375	.625	5.096019E-03	7.980526E-03
.609375	.6171875	5.096019E-03	1.438737E-03
.6132813	.6171875	1.829505E-03	1.438737E-03
.6152344	.6171875	1.956225E-04	1.438737E-03
.6152344	.616211	1.956225E-04	6.214977E-04
.6152344	.6157227	1.956225E-04	2.129078E-04
.6152344	.6154785	1.956225E-04	8.642674E-06
.6153563	.6154785	9.346008E-05	8.642674E-06
.6154175	.6154785	4.240871E-05	8.642674E-06
.615448	.6154785	1.686812E-05	8.642674E-06
.6154633	.6154785	4.112721E-06	8.642674E-06
.6154633	.6154709	4.112721E-06	2.264977E-06
.6154671	.6154709	9.23872E-07	2.264977E-06
.6154671	.615469	9.23872E-07	6.556511E-07
.615468	.615469	1.192093E-07	6.556511E-07
.615468	.6154685	1.192093E-07	2.980232E-07
.615468	.6154683	1.192093E-07	5.960465E-08
.6154681	.6154683	2.980232E-08	5.960465E-08

2) Regla falsa

Inicialmente, tenemos un cero de la función en $I_0 = [a_0, b_0]$ se generarán intervalos de seguridad $I_k = [a_k, b_k]$, con la propiedad de que $f(a_k) f(b_k) < 0$.

Los intervalos se definen recursivamente, como se ilustra en la figura de la siguiente página.



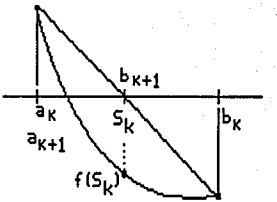
Sea S_k el punto en donde la secante que pasa por $(a_k, f(a_k))$ y $(b_k, f(b_k))$ intersecta al eje de las x.

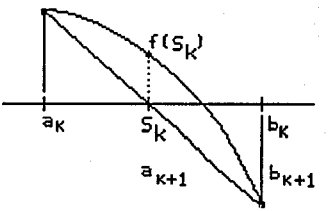
Si $f(a_k)\, f(S_k) < 0$ entonces $a_{k+1} = a_k$, $b_{k+1} = S_k$

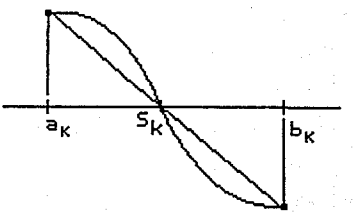
Si $f(a_k)\, f(S_k) > 0$ entonces $a_{k+1} = S_k$, $b_{k+1} = b$.

Si $f(a_k)\, f(S_k) = 0$ entonces S_k es la solución .

Los tres casos se ilustran a continuación:



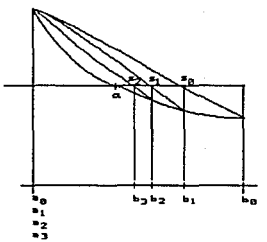




Este método tiene el inconveniente de que el criterio de convergencia no puede ser satisfecho a través de la longitud del intervalo I_k, pues no podemos asegurar que:

$$\lim_{k \to \infty} l(I_{k+1}) = 0$$

La siguiente figura ilustra la situación:



Se observa que $b_k \to \alpha$ pero $a_k = a_0 \ \forall_k$; luego,

$$\lim_{k \to \infty} l(I_k) = |\alpha_0 - a_0| \neq 0 \qquad (4)$$

Este efecto se debe a que $f''(x) > 0$ en $[a_0, b_0]$. El problema de cómo detener el proceso se resuelve evaluando la función en los extremos de I_k.

Para facilitar la descripción del método, no construiremos I_k con a_k del lado izquierdo y b_k del lado derecho. Lo que haremos será llamar b_{k+1} a S_k siempre, y calcular a_{k+1} de acuerdo con los cambios de signo de la función:

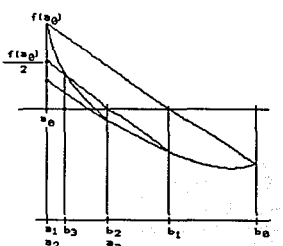
$$b_{k+1} = S_k$$

$$si \ f(a_k)\, f(S_k) < 0 \text{ entonces } a_{k+1} = a_k$$

$$si \ f(a_k)\, f(S_k) > 0 \text{ entonces } a_{k+1} = b_k$$

3) Regla falsa modificada

La deficiencia mencionada en (4) se resuelve con lo que llamaremos "método de la regla falsa modificada", que tiene por objeto evitar que un extremo se quede fijo.



Se trabaja normalmente regla falsa cuando $f(b_k)$ y $f(b_{k+1})$ son de signo contrario; cuando $f(b_{k+1})$ no cambia de signo, se construye la secante que pasa por los puntos

$$(b_k, f(b_k)) \text{ y } (a_k, \frac{1}{2}f(a_k))$$

Ejemplos 2 y 3. Calcular la raíz de la función del ejemplo 1, usando los métodos de regla falsa y regla falsa modificada.

La única diferencia entre la regla falsa y la regla falsa modificada, es que en esta se hace $a_{k+1} = \frac{1}{2}f(a_k)$; esta situación se refleja en la línea 220.

Programa

```
100 DEF FNF(X) = X-.2*SIN(X)-.5
110 LET A = 0.5: LET B = 1 LET K = 20
150 LET K = 0:LET FA = FNF(A): LET FB = FNF(B)
160 LET K = K+1: LET M = (FB-FA)/(B-A): LET S = B-(FB/M): LET FS = FNF(S)
170 IF FS = 0 THEN 230
180 IF FS*FA > 0 THEN 200
190 GOTO 220
200 LET A = B: LET FA = FB: LET B = S: LET FB = FS
210 REM ENDIF
220 LET B = S: LET FB = FS      :LET FA = FA/2
221 PRINT A,B,S
222 IF ABS(B-A) > (.001*ABS(B) + .00001) AND K < MAX THEN 160
230 PRINT "LA RAIZ ES ";S
```

Regla falsa (resultados)

A	B	S
1	.6121225	.6121225
1	.6153677	.6153677
1	.6154652	.6154652
1	.6154681	.6154681
1	.6154681	.6154681

LA RAIZ ES .6154682

EL VALOR DE LA FUNCION EN S ES 0

EL NUMERO DE ITERACIONES FUE 6

Regla falsa modificada (resultados)

A	B	S
1	.6121225	.6121225
.6121225	.6185591	.6185591
.6185591	.6143821	.6143821
.6143821	.6161058	6161058
6161058	.6151749	6151749
.6151749	.615621	.615621

LA RAIZ ES .615621

EL VALOR DE LA FUNCION EN S ES 1.277924E-04

EL NUMERO DE ITERACIONES FUE 6

B) Métodos que usan interpolación

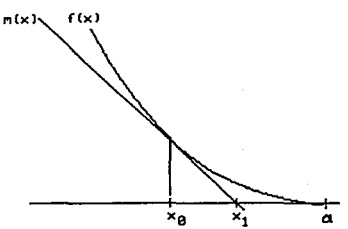
Calcular α tal que $f(\alpha) = 0$

A partir de una colección pequeña de aproximaciones a la solución, se encuentra una mejor aproximación, como el cero del polinomio que interpola a dichos puntos.

La idea general que está detrás de esto consiste en aproximar a $f(x)$, en la vecindad de x_i, por una función $m_i(x)$, con la propiedad de que $m_i(x) = 0$ sea un problema fácil; $m(x)$ podría ser:

i) *lineal*
ii) *cuadrático*

1) Método de la tangente o de Newton



$$m_0(x) = f(x_0) + (x - x_0) f'(x_0)$$

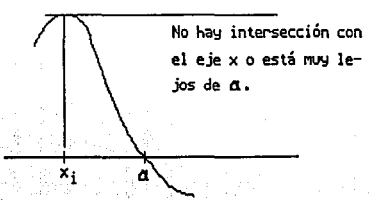
$$m_0(x_1) = 0 \quad \Rightarrow \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

en general $m_i(x) = f(x_i) + (x - x_i) f'(x_i)$ y x_{i+1} queda definido por:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

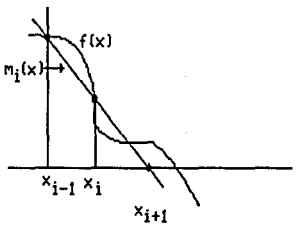
Este es el método de Newton y tiene la propiedad de que, cuando converge, lo hace con mucha rapidez.

Tiene la desventaja de ser costoso porque implica el cálculo de la función y su derivada a cada paso; además no siempre converge, como lo ilustra la figura siguiente:



2) Método de la secante

Consiste en aproximar a $f(x)$ por su secante en dos puntos, los últimos de la iteración.



$$m_i(x) = f(x_i) + \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} (x - x_i)$$

$$m_i(x_{i+1}) = 0$$

$$x_{i+1} = x_i - \frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})} f(x_i)$$

Ejemplos 3 y 4: Usando los métodos de la tangente y de la secante, encuentre la raíz de $x - 0.2 \operatorname{sen} x - 0.5 = 0$

Programa

```
10 REM metodo de la secante
20 DEF FNF(X) = X-.2*SIN(X)-.5
30 LET Q = 1E-09
40 LET X = 0.5: LET Y = 1
50 LET F = FNF(X): LET G = FNF(Y)
60 LET I = 1
70 LET I = I + 1
80 LET Z = Y-(X-Y) *G/(F-G)
90 LET X = Y: LET Y = Z: LET F = G: LET G = FNF(Y)
100 PRINT I,Y
110 IF ABS(Y-X) > Q THEN GOTO 70
120 PRINT: PRINT "una raiz es ";Y
```

Resultados

I	X
2	.5095649
3	.5980207
4	.6155891
5	.615468
6	.6154681
7	.6154682
8	.6154682

una raíz es .6154682

Programa

```
10 REM metodo de newton
20 INPUT "x0";X
30 LET I = 0: LET Q = 1E-09
40 LET I = I+1
50 LET F = X-.2*SIN(X)-.5
60 LET G = 1-.2*COS(X)
70 LET Y = F/G: LET X = X-Y
80 PRINT I,X
90 IF ABS(Y) > Q THEN GOTO 40
100 PRINT: PRINT "una raiz es ";X
```

I	X
1	2.602959E-02
2	.6249485
3	.6154745
4	.6154682
5	.6154682

una raiz es .6154682

Iteraciones de punto fijo y convergencia

Los dos métodos anteriores, Newton y Secante, cuando convergen, convergen rápidamente, en este contexto surge un cuestionamiento:

¿Cómo nos aseguramos que con una selección inicial x_0, el método escogido sea convergente?

Para responder a la cuestión, reconsideremos la fórmula iterativa del método de Newton.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

En general, podemos representar esta expresión de la siguiente manera:

$$x_{n+1} = g(x_n) \qquad (5)$$

Donde $g(xn) = x - \dfrac{f(x_{n-1})}{f'(x_{n-1})}$

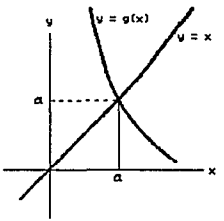
Si la sucesión $x_0, x_1, \ldots$ generada mediante (5) converge a algún punto α y $g(x)$ es continua, entonces:

$$\alpha = \lim_{n \to \infty} x_{n+1} = \lim_{n \to \infty} g(x_n) = g(\lim_{n \to \infty} x_n) = g(\alpha)$$

Por lo tanto:

$$\alpha = g(\alpha)$$

Y decimos que α es un punto fijo de la función g, esto es, α es invariante bajo g y por lo tanto la gráfica de g pasa por el punto (α, α).



Si α es un punto fijo de la función iterativa $g(x)$ para el método de Newton, entonces α es una solución de la ecuación.

$$f(x) = 0$$

Ya que:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = g(x_n)$$

Veremos ahora que $g(\alpha) = \alpha$ implica que α es una raíz de f.

$$\alpha = \alpha - \frac{f(\alpha)}{f'(\alpha)} = g(\alpha) = \alpha$$

$$\therefore \frac{f(\alpha)}{f'(\alpha)} = 0 \Leftrightarrow f(\alpha) = 0$$

$\therefore$ *α es una raíz de:*

$$f(x) = 0 \qquad (6)$$

Es posible escoger diversas funciones $g(x)$, con la propiedad de que los puntos fijos de $g(x)$ son raíces de la ecuación $f(x) = 0$. Considere, por ejemplo, el problema de calcular la raíz cuadrada de un número a; las siguientes funciones podrían ser utilizadas:

i) $f_1(x) = x^2 - a$

ii) $f_2(x) = 1 - \frac{a}{x^2}$ *(7)*

iii) $f_3(x) = x - \frac{a}{x}$

Aplicando el método de Newton a cada función, obtenemos las iteraciones siguientes:

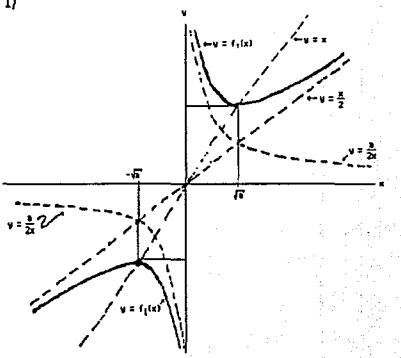
$$i)\ x_{n+1} = \frac{x_n}{2} + \frac{a}{2x_n} = g_1(x_n)$$

$$ii)\ x_{n+1} = \frac{3}{2}\, x_n - \frac{x_n^{\,3}}{2a} = g_2(x_n) \qquad (8)$$

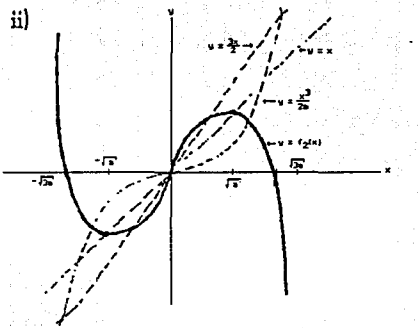
$$iii)\ x_{n+1} = \frac{2x_n\, a}{x^2{}_n + a} = g_3(x_n)$$

Cuyas gráficas son las siguientes:

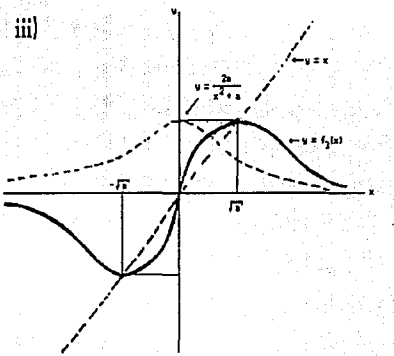
i)



ii)



iii)



Para cada selección, podemos calcular las sucesiones $x^1_1, x^1_2, \ldots$ mediante las fórmulas en (8) y esperar que éstas converjan, si lo hacen, éstas convergen a las raíces de la ecuación $x - \sqrt{a} = 0$

En conclusión, tenemos que la iteración de punto fijo es un método para generar una sucesión que nos permite resolver la ecuación $f(x) = 0$ a partir de la ecuación $g(x) = x$; de tal manera que cualquier solución de $g(x) = x$ es solución de $f(x) = 0$

C) Híbridos

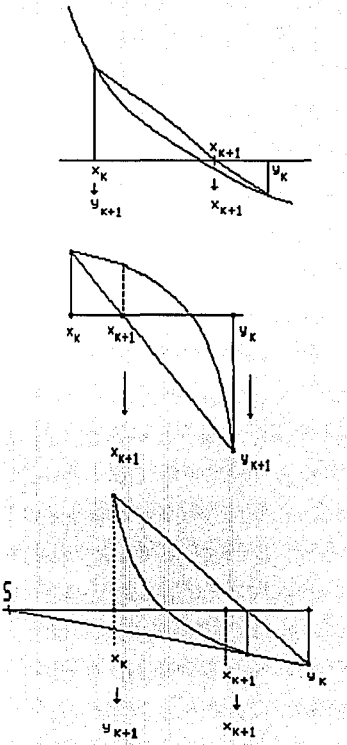
Como su nombre lo indica, consisten en combinar un método seguro con uno veloz.

Aquí veremos solamente aquél que usa los métodos de bisección y secante:

1) Se toma inicialmente un intervalo de seguridad $[x_0, x_1]$ (es decir $f(x_0)\, f(x_1) < 0$).

2) Llamemos x_k , y_k a los extremos del intervalo de seguridad en cada paso, (es decir $y_0 = x_1$)

3) x_{k+1} se calcula a partir de x_k y y_k por el método de la secante, siempre que x_{k+1} esté dentro del intervalo de seguridad; si no, por bisección aplicada a x_k , y_k.



La figura anterior ilustra cómo puede ocurrir que el punto obtenido por medio de la secante salga del intervalo de seguridad; entonces, se calcula $x_3 = \frac{x_2 + y_2}{2}$

4) Calcular y_{k+1} de tal modo que la raíz esté entre x_{k+1} y y_{k+1}

$$si \ f(x_k) f(x_{k+1}) < 0 \text{ entonces } y_{k+1} = x_k$$

$$si \ f(x_k) f(x_{k+1}) > 0 \text{ entonces } y_{k+1} = y_k$$

CONVERGENCIA

Suponga que tenemos el proceso iterativo:

$$x_{n+1} = g(x_n)$$

Que genera la sucesión $x_0, x_1, x_2, \ldots$ y deseamos saber si la sucesión $\{x_n\}$ así generada, converge o no a un punto fijo.

Supongamos que x_n es un valor cercano a la raíz α, por lo tanto:

$$x_n = \alpha + \varepsilon_n$$

Donde ε_n es pequeño. Si ε_n disminuye cuando n crece, entonces decimos que el proceso converge a α. Si ε_n aumenta, entonces la sucesión diverge. En el primer caso tenemos:

$$x_{n+1} = g(x_n)$$

$$= g(\alpha + \varepsilon_n)$$

$$= g(\alpha) + \varepsilon_n g'(\alpha) + \frac{1}{2}\varepsilon^2_n g''(\alpha) + \ldots$$

(S. Taylor)

Como $\alpha = g(\alpha)$ y si ε_n es pequeño entonces ε^2_n es insignificante, tenemos que:

$$x_{n+1} \approx \alpha + \varepsilon_n g'(\alpha)$$

$$x_{n+1} - \alpha \approx \varepsilon_n g'(\alpha)$$

$$\varepsilon_{n+1} \approx \varepsilon_n g'(\alpha)$$

Por lo tanto, el error ε_{n+1} es un múltiplo del error ε_n.

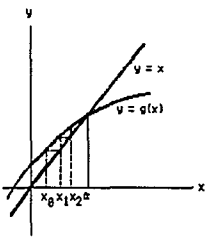
En consecuencia, ε_{n+1} disminuye cuando n crece si y sólo si $0 \leq | g'(\alpha) | < 1$

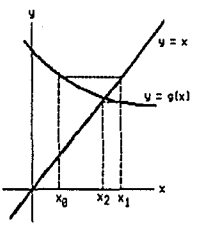
Si ε_{n+1} aumenta entonces $| g'(\alpha) | > 1$ en cuyo caso la sucesión $\{x_n\}$ es divergente.

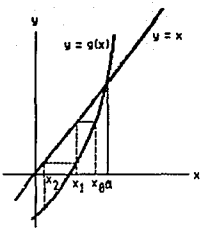
Por lo tanto, $x_{n+1} = g(x_n)$ genera un proceso iterativo convergente si y sólo si:

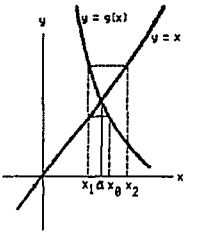
$$0 \leq | g'(\alpha) | < 1$$

y x_0 esta "cerca" de α









A primera vista el resultado anterior parece no ser útil de inmediato, ya que es necesario evaluar la función $g'(x)$ en la raíz α, que desconocemos. Sin embargo, a menudo se tiene una idea aproximada del valor de la raíz y por lo general esto será suficiente.

Rapidez de convergencia

Dados dos métodos convergentes, ¿cómo seleccionamos el mejor? ¿Eligiendo al más rápido? ¿Cómo medimos rapidez de convergencia?

Anteriormente hemos dicho que un método iterativo es convergente si $0 \leq | g'(\alpha) | < 1$ y si además x_0 está lo suficientemente cerca de α.

Consideremos la ecuación:

$$2x^2 - 4x + 1 = 0 \qquad (9)$$

Y derivemos dos funciones iterativas para esta ecuación, a saber:

i) $2x^2 - 4x + 1 = 0$

$2x^2 = 4x - 1$

$12x^2 = 24x - 6$

$-12x^2 = -24x + 6$

$-5x^2 = 7x^2 - 24x + 6$

$$-5x = 7x - 24 + \frac{6}{x}$$

$$x = \frac{-1}{5}\left(7x - 24 + \frac{6}{x}\right)$$

$$g_1(x_{n+1}) = -\frac{1}{5}\left(7x_n - 24 + \frac{6}{x_n}\right)$$

Con un procedimiento análogo, obtenemos:

$$ii)\ g_2(x_{n+1}) = \frac{2x^2{}_n - 1}{4(x_n - 1)}$$

Una raíz de la ecuación (9) es:

$$\alpha = 1.707107$$

con ii) la raíz se obtiene para $n = 6$ y $x_0 = 2$ pero para i) con $n = 280$ se obtiene $\alpha = 1.713879$, esto es, se tienen sólo dos dígitos de precisión; y después de $n \approx 3000$ se obtienen valores alternados de 1.707098 y 1.707116, las oscilaciones de este tipo ocurren con cierta frecuencia en aquellos procesos que convergen con lentitud.

Por lo tanto i) es una función iterativa que genera una sucesión convergente (¡pero apenas!)

Esto se debe a que está muy cercana a 1

$$g'_1(1.707107) = -0.9882252 < 1$$

En cambio, para:

$$g'_2(x_{n+1}) = \frac{2x^2{}_n - 4x_n + 1}{4(x - 1)^2}$$

Tenemos que:

$$g'_2(1.707107) = 0$$

Por lo que es de esperarse, como sucede, que la convergencia de ii) sea muy rápida.

Convergencia cuadrática

Supongamos que $g'(\alpha) = 0$, entonces, sabemos que:

$$x_{n+1} = g(\alpha) + \frac{1}{2}\varepsilon^2{}_n g''(\alpha) + \ldots$$

Por lo tanto:

$$x_{n+1} = \alpha + \frac{1}{2}\varepsilon^2{}_n g''(\alpha) + \ldots$$

$$x_{n+1} - \alpha = \frac{1}{2}\varepsilon^2{}_n g''(\alpha) + \ldots$$

$$\varepsilon_{n+1} = \frac{1}{2}\varepsilon^2{}_n g''(\alpha) + \ldots$$

$$\varepsilon_{n+1} \approx \frac{1}{2}\varepsilon^2{}_n g''(\alpha)$$

Por lo que el error ε_{n+1} es un múltiplo del cuadrado del error ε_n; de manera que si ε_n es pequeño, ε_{n+1} es más pequeño.

En general, si:

$$\varepsilon_{n+1} \approx c\varepsilon^p{}_n \qquad p \geq 1$$

Entonces la convergencia dependerá del valor de p.

Donde c es una constante, y decimos que el proceso es de convergencia p.

Resumiendo, tenemos que:

a) Un proceso iterativo es lineal o de primer orden si

$$g'(\alpha) \neq 0 \text{ y } |g'(\alpha)| < 1 \qquad p = 1$$

b) Un proceso es cuadrático o de segundo orden si

$$g'(\alpha) = 0 \text{ y } g''(\alpha) \neq 0 \qquad p = 2$$

c) Un proceso es de tercer orden si

$$g'(\alpha) = 0 \quad g''(\alpha) = 0 \text{ y } g'''(\alpha) \neq 0 \qquad p = 3$$

Ya que en los desarrollos de Taylor, tenemos que los valores de ε, serán ε, ε^2, ε^3 respectivamente.

Observaciones:

1) El método de Newton converge cuadráticamente.

2) La secante converge con orden:

$$p = \frac{1+\sqrt{5}}{2} = 1.6180334\ldots$$

3) La regla falsa modificada es, usualmente, superlineal:

$$\varepsilon_{n+1} = c\varepsilon^{p}_{n} \qquad \text{donde } p > 1$$

4) El método de bisección no tiene convergencia siquiera lineal.

EJERCICIOS Y PROBLEMAS*

1. Encuentre un intervalo de longitud uno que contenga una raíz de $p(x) = x^3 - 6x^2 + 9x - 5$

2. Proporcione una evidencia gráfica de que las siguientes funciones tienen infinidad de raíces reales; localize la raíz positiva más pequeña de cada una de ellas dentro de un intervalo de longitud 1.

   a) $x - tanx - 1$

   b) $sen\ 2x - senx$

   c) $sen\ 3x - senx$

   d) $x\ senx - 1$

   e) $ln\ x - tanx$

3. Demuestre que la función:

$$f(x) = senx - 1$$

   Tiene una sola raíz en el intervalo $(1, \pi/2)$

4. Demuestre que la función:

$$f(x) = x\ tanx - 1$$

   Tiene una sola raíz en el intervalo (0, 1)

5. Localize las raíces de las siguientes ecuaciones en intervalos de longitud uno.

   a) $x^2 - \sqrt{x} = 2$

   b) $x - e^x = -2$

   c) $ln(x) + \sqrt{x} = 2$

   d) $2x = e^{-x}$

   e) $e^x - 5x = 0$

6. Encuentre la raíz más pequeña de:

$$x^2 - 10\,000x + 1 = 0$$

   mediante la fórmula cuadrática y algunos de los métodos desarrollados en el texto.

7. Usando el método de Newton, encuentre las primeras tres raíces positivas de:

   a) $sen\ x - \frac{b}{x} cos\ x = 0$

   b) $tan\ x - \frac{b}{x} = 0$

   Explique las diferencias en comportamiento de ambas ecuaciones.

* Los ejercicios se tomaron y/o adaptaron de las obras citadas al final de esta sección.

8. Aplique el método de bisección para determinar las raíces de las siguientes ecuaciones con tres cifras significativas.

a) $x^4 - 2x^3 - x - 3 = 0$ $(2, 3)$

b) $x^5 - x - 0.46 = 0$ $(1, 2)$

c) $x^5 + x + 1 = 0$ $(-1, 0)$

d) $x^5 - 3x^2 - 100 = 0$ $(2, 3)$

9. Usando tres métodos diversos, encuentre las raíces reales de:

a) $x^6 - 12.1x^5 + 59.5x^4 - 151.85x^3 + 1212.6625x^2 - 156.6x + 48.5625 = 0$

b) $x^4 - 3.0x^3 + 3.37x^2 - 1.680x + 0.3136 = 0$

c) $x^3 + 1.5x^2 - 5.75x + 4.37 = 0$

d) $x^4 - 1.73x^2 + 0.46x + 1.275 = 0$

10. Las funciones siguientes tienen raíces en los intervalos indicados a la derecha. Calcule las raíces de estas funciones con una precisión de tres cifras mediante el método de bisección.

a) $x \tan x - 1$ $(0, 1)$

b) $\text{sen } 2x - \text{sen} x$ $(0.7, 1.7)$

c) $\text{sen } 3x - \text{sen} x$ $(0.5, 1.5)$

d) $x \text{ sen} x - 1$ $(1, 2)$

e) $\ln x - \tan x$ $(3.8, 4.8)$

11. Aproxime el valor de π resolviendo las siguientes ecuaciones, mediante el método de bisección y con una precisión de cinco cifras. ¿Coinciden las raíces con el valor de $\pi = 3.14159265$?

a) $\tan \frac{x}{4} - 1 = 0$

b) $\text{sen } \frac{x}{4} - \cos \frac{x}{4} = 0$

12. Encuentre las raíces de las funciones del ejercicio 10. Usando el método de Newton. Use los extremos del intervalo como aproximaciones iniciales.

13. Encuentre el valor de $\sqrt{2}$ con precisión de cinco cifras, usando el método de la secante.

14. Determine las raíces de las funciones del ejercicio 9 usando los métodos de regla falsa y el de Newton.

15. ¿Qué sucede si se aplica el método de Newton a la función $f(x) = \text{arc tan} x$ con $x_0 = 2$?

16. ¿Cuál es el límite de la sucesión

$$x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}$$

17. ¿Para que valores iniciales converge el método de Newton si la función es $f(x) = x^2/(1 + x^2)$?

18. Analice qué sucede cuando el método de Newton es aplicado a la función $f(x) = 2x^3 - 9x^2 + 12x + 15$ con los siguientes valores iniciales:

a) $x = 3$

b) $x > 3$

c) $x < 3$

19. El recíproco de un número R puede calcularse sin recurrir a la división mediante la siguiente expresión:

$$x_{n+1} = x_n (2 - x_n R)$$

Empezando con $x_0 = 0.2$, calcule el recíproco de 4 con seis cifras de precisión. Tabule el error en cada paso y determine el tipo de convergencia.

20. Use la fórmula:

$$x_{n+1} = \frac{1}{2}[x_n + (\frac{R}{x_n})]$$

Para encontrar $\sqrt{R}$. Efectúe tres iteraciones para $R=2$, con $x_0=1$. Use el método de bisección en el intervalo $[1, 2]$ para encontrar $\sqrt{R}$. ¿Cuántas iteraciones son necesarias en cada método para obtener una precisión de 10^{-6}?

21. Todo polinomio de grado n tiene n raíces en el plano complejo ¿Se concluye de esto que toda función de la forma:

$$f(x) = \sum_{n=0}^{\infty} a_n x^n$$

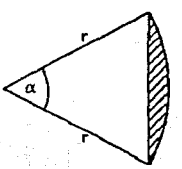
tiene un número infinito de raíces?

Problemas

1. El área sombreada de la figura depende de los valores del radio r y el ángulo α (en radianes) de acuerdo con la siguiente fórmula:

$$A = \frac{r^2}{2}[\alpha - sen\,\alpha]$$

Determine el valor de α si $r = 1.25$ y $A = 2$.



2. La ecuación de Kepler para la determinación de las órbitas es:

$$M = \phi - e \operatorname{sen} \phi$$

donde M se denomina anomalía media.

ϕ se denomina anomalía de la excentricidad.

e es la excentricidad.

Si $M = a(t - t_0)$. Determine el valor del ángulo ϕ si $t_0 = 0, e = 0.4, a = 0.5, t = 1.0$

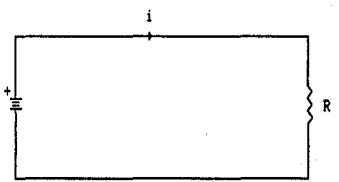
3. En el circuito del diagrama, la resistencia R cambia su valor, en respuesta al calor, de acuerdo con la siguiente fórmula:

$$R(i) = 100 + 10\sqrt{i}$$

Además la corriente i, satisface la siguiente ecuación.

$$V = R(i)\, i$$

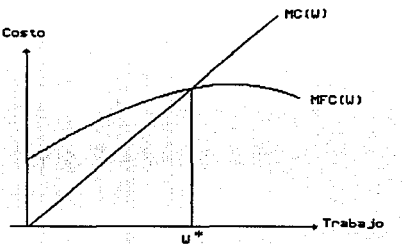
Determine el valor de la corriente i.



4. Una compañía debe decidir si compra una computadora grande o varias microcomputadoras para satisfacer sus necesidades de cómputo. El costo para satisfacer la demanda de trabajo W con micros crece de manera lineal, de la siguiente forma $MC(W) = W$ donde $MC(W)$ es el costo en función de la carga de trabajo W; en cambio el costo usando una computadora grande esta dado por la función.

$$MFC(W) = 100 + \log(W + 1)$$

Encuentre el valor W^* que hace que el costo de usar micros y el costo de usar una computadora grande sean iguales.



5. Supóngase que se desea comprar un automóvil y se está limitado a dos opciones. El costo anual neto de poseer cualquiera de los dos vehículos está compues-

to por el costo de compra, costo de mantenimiento y de las ganacias:

	Modelo de lujo	Modelo económico
Costo de compra, $	-15,000	-5000
Costo de mantenimiento, $/año/año	-400	-200
Ganacias anuales y beneficios, $	7500	3000

Si la tasa de interés es del 12.5% *(i = 0.125)*, calcule el punto de equilibrio *(n)* para los automóviles.

6. Si se compra una pieza de equipo en $20 000 en abonos, pagando $5 000 durante 5 años. ¿Qué tasa de interés se está pagando? La fórmula que relaciona el costo actual *(P)*, los pagos anuales *(A)*, el número de años *(n)* y la tasa de interés es:

$$A = P\frac{i(1+i)^n}{(1+i)^n}$$

7. Un centro de diversiones cuesta $10 millones de dólares y produce una ganacia de $2 millones. Si la deuda se debe pagar en 10 años ¿A qué tasa de interés debe hacerse el préstamo? El costo anual *(P)*, el pago anual *(A)* y la tasa de interés *(i)* se relacionan entre sí mediante la siguiente fórmula:

$$\frac{P}{A} = \frac{(1+i)^n - 1}{i(1+i)^n}$$

donde n es el número de pagos anuales. Para este problema.

$$\frac{P}{A} = \frac{10000000}{2000000} = 5$$

Por lo tanto, la ecuación se transforma en:

$$5 = \frac{(1+i)^{10} - 1}{i(1-i)^{10}}$$

La tasa de interés que satisface esta ecuación se puede determinar encontrando la raíz de:

$$f(i) = \frac{(1+i)^{10} - 1}{i(1+i)^{10}} - 5$$

a) Dibújese *f(i)* contra i y para obtener una estimación gráfica de la raíz.

b) Calcúlese i usando el método de bisección (contar las iteraciones).

c) Calcúlese i usando el método de la regla falsa (contra las iteraciones).

En los incisos (b) y (c) úsense los valores iniciales de $i = 0.1$ y 0.2. Obténgase un nivel del error del 2% en ambos casos.

8. La temperatura (en grados Kelvin) de un sistema, varía durante el día de acuerdo con:

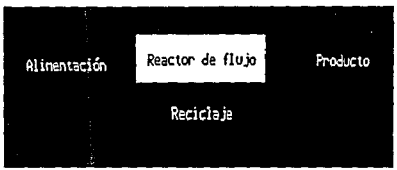
$$T = 400 + 200 \cos \frac{2\pi t}{1440}$$

En donde t se expresa en minutos. La presión sobre el sistema esta dada por $p = e^{-t/1440}$. Desarróllese un programa que calcule el volumen molar del oxígeno en intervalos de un minuto a lo largo del día. Grafíquense los resultados. Si se tiene capacidad gráfica en la computadora grafíquense los datos. Si no es así, grafíquense los resultados a intervalos de 60 minutos.

9. En ingeniería química, los reactores de flujo (es decir, aquéllos en que un fluido va de un extremo al otro con una mezcla mínima a lo largo del eje longitudinal) se usan a menudo para convertir reactivos en productos. Se ha determinado que la eficiencia de la conversión se puede mejorar a veces reciclando una parte del flujo del producto de manera que regrese a la entrada para un paso adicional a través del reactor. La tasa de reciclaje se define como:

$$\mathrm{R} = \frac{\textit{volumen de fluido regresado a la entrada}}{\textit{volumen de fluido que deja el sistema}}$$

Supóngase que se está procesando una sustancia química A para generar un producto B. Para el caso en que B de acuerdo con una reacción autocatalítica.



(esto es, en la que uno de los productos actúa como catalizador o de estimulante en la reacción, o

$$A + B \rightarrow B + B$$

se puede demostrar que una tasa óptima de reciclaje debe satisfacer:

$$In \frac{1 + R(1 - X_{Af})}{R(1 - X_{Af})} = \frac{R + 1}{R[1 + R(1 - X_{Af})]}$$

en donde X_{Af} es la fracción del reactante A que se convierte en el producto B. La tasa óptima de reciclaje corresponde a un reactor de tamaño mínimo, necesario para alcanzar el nivel de conversión deseado.

Usese el método de bisección para determinar las tasas de reciclaje necesarias que minimicen al tamaño del reactor en conversiones fraccionales de:

a) $X_{Af} = 0.99$

b) $X_{Af} = 0.995$

c) $X_{Af} = 0.999$

10. La concentración de la bacteria contaminante C en un lago decrece de acuerdo con la relación:

$$C = 80e^{-2t} + 20e^{-0.1t}$$

Determínese el tiempo requerido para que la bacteria se reduzca a 10, usando a) un método gráfico y b) el método de Newton.

11. Muchos campos de la ingeniería requieren estimaciones exactas de la población. Por ejemplo, para la transportación, los ingenieros consideran necesario determinar por separado la tendencia del crecimiento demográfico de una ciudad y de los suburbios adyacentes. La población del área urbana declina en función del tiempo de acuerdo con:

$$P_u(u) = P_{u,máx}\, e^{-k_u t} + P_{u,mín}$$

Mientras que la población suburbana crece, de acuerdo con:

$$P_s(t) = \frac{P_{s,máx}}{1 + \frac{P_{s,máx}}{P_0} - 1\, e^{-k_s t}}$$

En donde $P_{u,máx}$, k_u, $P_{u,mín}$, $P_{s,máx}$, P_0 y k_s son parámetros derivados de forma empírica.

Determínese el tiempo y los valores correspondientes de $P_u(t)$ y de $P_s(t)$ cuando las poblaciones son iguales. Los valores de los parámetros son $P_{u,máx} = 60\,000$; $K_u = 0.04$ año^{-1}; $P_{u,mín} = 12\,000$; $P_{s,máx} = 5\,000$ y $k_s = 0.06$ año^{-1}. Para obtener las soluciones, úsese a) un método gráfico y b) el método de la regla falsa.

12. El movimiento de una estructura se define mediante la siguiente ecuación para una oscilación amortiguada:

$$y = 10e^{-kt} \cos wt$$

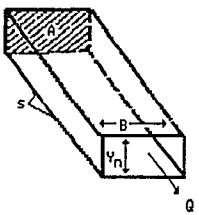
donde $k = 0.5$ y $w = 2$

a) Usese el método gráfico, para obtener una estimación inicial del tiempo necesario para que el desplazamiento baje hasta 4

b) Usese el método de Newton para determinar la raíz hasta un error de 0.01%

c) Usese el método de la secante para determinar la raíz hasta un $\in_s = 0.01\%$

13. La figura muestra un canal abierto de dimensiones constantes con un área transversal A. Bajo condiciones de flujo uniforme, se cumple la siguiente relación basada en la ecuación de Manning:



$$Q = \frac{y_n B}{n} \left(\frac{y_n B}{B + 2y_n}\right)^{2/3} S^{1/2}$$

En donde Q es el flujo, y_n es la profundidad normal, B es el ancho del canal, n es un coeficiente de rugosidad usado para medir los efectos de la fricción del material en el canal y S es la pendiente del canal. La ecuación se usa en ingeniería de fluidos y recursos de agua para determinar la profundidad normal. Si este valor es menor que la profundidad crítica:

$$y_c = \left(\frac{Q^2}{B^2 g}\right)^{1/3}$$

En donde g es la aceleración de la gravedad (980 cm/s^2), entonces el flujo es subcrítico.

Usese un método gráfico y el método de bisección para determinar y_n si $Q = 14.15\ m^3/s$; $B = 4.572\ m$; $n = 0.017$ y $S = 0.0015$. Señálese si el flujo es sub o supercrítico.

14. Una corriente oscilatoria en un circuito eléctrico se describe mediante.

$$I = 10e^{-t} sen\ (2\pi t)$$

en donde t está dado en segundos. Determínense todos los valores de t tales que $I = 2$

Problemas teóricos

1. La supercomputadora Cray-1 no divide directamente; en lugar de ello, para calcular el recíproco de un número $R > 0$ $(1/R)$, calcula una "aproximación recíproca" y usa ésta como el primer valor para la iteración de Newton. Dado $f(x) = (1/x) - R$, demuestre que la fórmula para el método de Newton es:

$$X_{k+1} = x_k\,(2 - Rx_k)$$

a) Usando esta iteración calcule $1/R$ para $R = 1, 2, ..., 10$ Tabule el número de iteraciones necesarias para generar un resultado exacto a seis dígitos. Usa $x_0 = 0.01$

b) Usando la misma función $f(x)$ repita los cálculos con el algoritmo de bisección y compare el número de iteraciones con (a). Use como primer intervalo a $[.01, 2]$

2. Un método clásico para la solución de ecuaciones cúbicas es el usado por Cardano. La ecuación cúbica:

$$x^3 + ax^2 + bx + c = 0$$

Es transformada a la forma reducida:

$$y^3 + py + q = 0$$

Mediante la sustitución $x = y - a/3$. Los coeficientes en esta forma son:

$$P = b - \frac{a^2}{2}$$

$$q = c - \frac{ab}{3} + 2\left(\frac{a}{3}\right)^3$$

Una raíz real de la forma reducida puede determinarse de la siguiente manera:

$$S = [(\frac{p}{3})^3 + (\frac{q}{2})^2]^{1/2}$$

$$y_1 = [-\frac{q}{2} + S]^{1/3} + [-\frac{q}{2} - S]^{1/3}$$

Y, por lo tanto, la raíz real de la ecuación original es:

$$x_1 = y_1 - \frac{a}{3}$$

a) Aplica el método de Cardano para encontrar la raíz real de:

$$x^3 + 3x^2 + c^2x + 3c^2 = 0$$

Para diversos valores de c, investigue la pérdida de exactitud por redondeo para valores grandes de c, digamos para c de tal manera que $1/c \approx \varepsilon$ de la computadora.

b) Aplique el método de Newton para la misma ecuación y para los mismos valores de c. Investigue los efectos del error de redondeo y de la selección del valor inicial.

3. Use el método de Newton para calcular la raíz única de:

$$x + e^{-bc} \cos x = 0 \quad B > 0$$

Con diversos valores de b, por ejemplo, $b = 1, 5, 10, 25, 50$. Escoja a $x_0 = 0$ y explique cualquier comportamiento anómalo. Teóricamente, el método de Newton convergerá para cualesquiera valores de x_0 y B.

Referencias bibliográficas

1. *Atkinson, Kendall E.* "An Introduction to Numerical Analysis". New York. John Wiley & Sons, 1989.

2. *Acton, Forman S.* "Numerical Methods that Work". New York. Harper & Row, 1970.

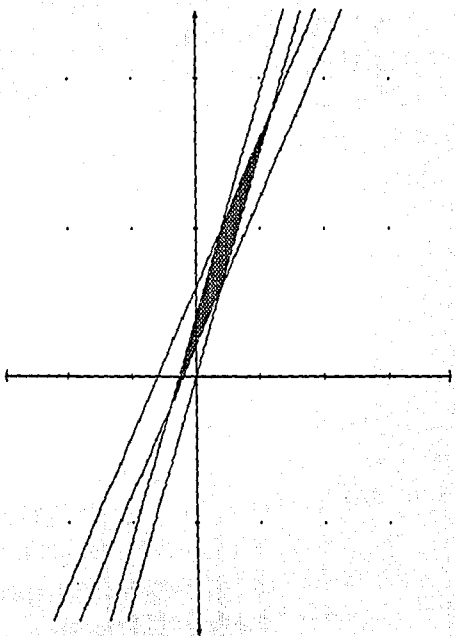
3. *Chapra, Steven C; Canale, Raymond P.* "Métodos Numéricos para Ingenieros". México. McGraw-Hill, 1987.

4. *Cheney, Ward; Kincaid, David.* "Numerical Mathematics and Computing". Monterey USA. Brooks/Cole Publishing, 1980.

5. *Yakowitz, Sidney; Szidarovszky, Ferenc.* "An Introduction to Numerical Computations". New York. MacMillan, 1990.

CAPÍTULO IV

SISTEMAS DE ECUACIONES LINEALES



"Casi todo lo que la matemática de nuestro siglo ha producido en forma de ideas científicas originales, está relacionado con el nombre de Gauss."

L. KRONECKER

(Zahlentheorie)

"Gauss fue el gigante desde cuya altura se abarcan de un vistazo las estrellas y los abismos."

W. BOLYAI

(Kurzer Grundriss eines Versuchs)

INTRODUCCION

La ley de Ohm establece una relación funcional entre el voltaje V, la resistencia R y la corriente i, a través de un circuito, mediante la siguiente fórmula:

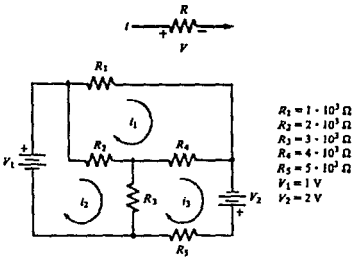
$$V = iR \qquad (1)$$

Por otro lado, la ley de Kirchhoff establece que la suma algebraica de las corrientes sobre un nodo debe ser cero, lo que establecemos de la siguiente manera:

$$\sum i_k = 0 \qquad (2)$$

A partir de esas dos leyes podemos determinar las corrientes y voltajes en cualquier punto de un circuito eléctrico. De hecho, estas leyes conducen a sistemas de ecuaciones lineales.

En el diagrama siguiente se presenta un circuito eléctrico con cinco resistencias y dos fuentes de poder.



La aplicación de las leyes de Ohm y Kirchhoff conduce al siguiente sistema de ecuaciones lineales.

$$
(3)\qquad
\begin{aligned}
(R_1 + R_2 + R_4)i_1 \quad & -R_2 i_2 \quad && -R_4 i_3 = 0 \\
-R_2 i_1 + (R_2 + R_3)\, i_2 \quad & && -R_3 i_3 = V_1 \\
-R_4 i_1 \quad & -R_3 i_2 + (R_3 + R_4 + R_5) i_3 = V_2
\end{aligned}
$$

Este es un sistema de 3 ecuaciones lineales de tres incógnitas, se puede resolver con alguno de los métodos que se enseñan en la escuela secundaria.

El sistema de ecuaciones (3) puede ser escrito de manera compacta en forma matricial de la siguiente manera:

$$\mathbf{A\,x = b} \qquad (4)$$

Este problema, si bien ha quedado resuelto desde el punto de vista analítico desde hace mucho tiempo, presenta dificultades que han ocupado la atención de muchos analistas numéricos durante los últimos 40 años. Actualmente se siguen publicando artículos al respecto.

La dificultad principal consiste en que para resolver un sistema se tienen que invertir muchos recursos, tanto en tiempo de computación como en el espacio requerido para almacenar la información.

Consideremos nuevamente el sistema:

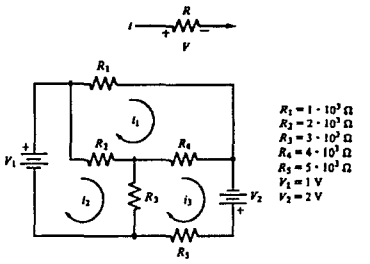
$$\mathbf{A\,x = b}$$

en donde **A** es una matriz cuadrada, de orden *n*.

Un primer resultado es que el problema tiene solución si y sólo si:

$$\det(\mathbf{A}) \neq 0$$

En el diagrama siguiente se presenta un circuito eléctrico con cinco resistencias y dos fuentes de poder.



La aplicación de las leyes de Ohm y Kirchhoff conduce al siguiente sistema de ecuaciones lineales.

$$
(3)\qquad
\begin{aligned}
(R_1 + R_2 + R_4)i_1 \quad -R_2 i_2 \quad\quad\quad -R_4 i_3 &= 0 \\
-R_2 i_1 + (R_2 + R_3)\, i_2 \quad\quad -R_3 i_3 &= V_1 \\
-R_4 i_1 \quad -R_3 i_2 + (R_3 + R_4 + R_5) i_3 &= V_2
\end{aligned}
$$

Este es un sistema de 3 ecuaciones lineales de tres incógnitas, se puede resolver con alguno de los métodos que se enseñan en la escuela secundaria.

El sistema de ecuaciones (3) puede ser escrito de manera compacta en forma matricial de la siguiente manera:

$$\mathbf{A\,x = b} \qquad (4)$$

Este problema, si bien ha quedado resuelto desde el punto de vista analítico desde hace mucho tiempo, presenta dificultades que han ocupado la atención de muchos analistas numéricos durante los últimos 40 años. Actualmente se siguen publicando artículos al respecto.

La dificultad principal consiste en que para resolver un sistema se tienen que invertir muchos recursos, tanto en tiempo de computación como en el espacio requerido para almacenar la información.

Consideremos nuevamente el sistema:

$$\mathbf{A\,x = b}$$

en donde A es una matriz cuadrada, de orden *n*.

Un primer resultado es que el problema tiene solución si y sólo si:

$$\det(A) \neq 0$$

Pero calcular un determinante por su desarrollo en menores, por ejemplo, es un problema que requiere de mucho tiempo. De hecho, no es posible calcular un determinante de orden 20, ya que tomaría miles de años (haciendo cada una de las operaciones en fracciones de segundo). En efecto, hacerlo implica calcular 20 determinantes de orden 19. Si denotamos con # op (n) el número de operaciones requeridas para calcular un determinante de orden n, para calcular uno de orden 20 se tendría que hacer el siguiente número de operaciones:

$$\begin{aligned} \#\ \text{op}\ (20) &= 20\ \#\ \text{op}\ (19) \\ &= 20 \times 19 \times \#\ \text{op}\ (18) \\ &\vdots \\ &= 20 \times 19 \times \ldots \times 3\ \#\ \text{op}\ (2) \\ &= 20! \\ &\approx 2.4329 \times 10^{18} \end{aligned}$$

Si se hace una multiplicación cada centésimo de segundo, se necesitarían aproximadamente.. 1771 millones de años para calcular tan sólo un determinante de orden 20! Por lo tanto, el método de Cramer no es práctico.

Considere los cuatro sistemas siguientes:

a)
$$\begin{aligned} 3x + 5y - 8z &= 0 \\ 2x - y - z &= 10 \\ x + y - 5z &= 1 \end{aligned}$$

b)
$$\begin{aligned} 5x &= 9 \\ 3x + y + z &= 10 \\ 10x + z &= 3 \end{aligned}$$

c)
$$\begin{aligned} x + y + z &= 0 \\ x - y - z &= 0 \\ x - y + z &= 1 \end{aligned}$$

d)
$$\begin{aligned} 3x + y + z &= 5 \\ x + y + z &= 0 \\ z &= 3 \end{aligned}$$

El ejemplo (b) parece un problema fácil, o menos difícil que los otros tres.

La propiedad que hace *fácil* resolver el problema (b) es que el sistema de 3 ecuaciones se reduce a tres ecuaciones en una sola incógnita.

En general, diremos que un sistema de n ecuaciones simultáneas resulta fácil cuando equivale a n ecuaciones en una incógnita.

Definiciones:

1) Los *elementos diagonales* de una matriz $A = (\alpha_{ij})$ son aquéllos para los cuales sus subíndices son iguales α_{ii}, $i = 1, 2, \ldots, n$.

2) *Sistema triangular* es aquél que tiene asociada una matriz triangular.

3) Una *matriz es triangular inferior* si $\alpha_{ij} = 0$ para $j > i$; *triangular superior* si $\alpha_{ij} = 0$ para $i > j$.

4) Un *sistema de ecuaciones lineales es fácil* si y sólo si se puede transformar en sistema triangular. Por lo tanto, dado un sistema de ecuaciones $\mathbf{Ax = b}$, es conveniente transformar dicho sistema en un *sistema fácil.*

En el ejemplo (b) tenemos la siguiente transformación, permutando columnas o renglones.

$$\begin{array}{cc} & (1\ \ 2\ \ 3) \\ \begin{array}{c} 5x = 9 \\ 3x + y + z = 10 \\ 10x + z = 3 \end{array} & \begin{pmatrix} 5 & 0 & 0 \\ 3 & 1 & 1 \\ 10 & 0 & 1 \end{pmatrix} \begin{array}{l} (1) \\ (2) \\ (3) \end{array} \end{array}$$

$$\begin{array}{c} (1\ \ 2\ \ 3) \\ \begin{pmatrix} 5 & 0 & 0 \\ 10 & 0 & 1 \\ 3 & 1 & 1 \end{pmatrix} \begin{array}{l} (1) \\ (2) \\ (3) \end{array} \end{array}$$

$$\begin{array}{cc} & (1\ \ 2\ \ 3) \\ \begin{array}{c} 5x = 9 \\ 10x + z = 10 \\ 3x + z + y = 3 \end{array} & \begin{pmatrix} 5 & 0 & 0 \\ 10 & 1 & 0 \\ 3 & 1 & 1 \end{pmatrix} \begin{array}{l} (1) \\ (2) \\ (3) \end{array} \end{array}$$

Nota: Los números que aparecen arriba y a la derecha de cada matriz indican las permutaciones.

Así, el nuevo sistema quedaría:

$$\begin{pmatrix} 5 & 0 & 0 \\ 10 & 1 & 0 \\ 3 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ z \\ y \end{pmatrix} = \begin{pmatrix} 9 \\ 3 \\ 10 \end{pmatrix}$$

Cuando el original es:

$$\begin{pmatrix} 5 & 0 & 0 \\ 3 & 1 & 1 \\ 10 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 9 \\ 10 \\ 3 \end{pmatrix}$$

Observamos que al permutar las columnas, la solución queda permutada de igual manera. Al permutar renglones, el lado derecho debe permutarse de igual manera para no alterar la solución.

Las permutaciones a una matriz se pueden representar como transformaciones tales que, al ser aplicadas por la izquierda, alteran el orden de los renglones y al aplicarse por la derecha, el de las columnas.

Si **A'** es una matriz obtenida mediante permutaciones de renglones y columnas de **A**, entonces:

$$\mathbf{A' = P A Q\ ,}$$

en donde **P** y **Q** son matrices de permutación.

Definición:

Matriz de permutación es aquélla que se obtiene mediante permutaciones de la matriz identidad.

Ejemplos:

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Consideremos el siguiente ejemplo:

$$\alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{16}x_6 = b_1$$
$$\alpha_{26}x_6 = b_2$$
$$\alpha_{34}x_4 + \alpha_{35}x_5 + \alpha_{36}x_6 = b_3$$
$$\alpha_{45}x_5 + \alpha_{46}x_6 = b_4$$
$$\alpha_{52}x_2 + \alpha_{53}x_3 + \alpha_{54}x_4 + \alpha_{55}x_5 + \alpha_{56}x_6 = b_5$$
$$\alpha_{63}x_3 + \alpha_{64}x_4 + \alpha_{65}x_5 + \alpha_{66}x_6 = b_6$$

Que, escrito en forma matricial, queda:

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} & \alpha_{15} & \alpha_{16} \\ 0 & 0 & 0 & 0 & 0 & \alpha_{26} \\ 0 & 0 & 0 & \alpha_{34} & \alpha_{35} & \alpha_{36} \\ 0 & 0 & 0 & 0 & \alpha_{45} & \alpha_{46} \\ 0 & \alpha_{52} & \alpha_{53} & \alpha_{54} & \alpha_{55} & \alpha_{56} \\ 0 & 0 & \alpha_{63} & \alpha_{64} & \alpha_{65} & \alpha_{66} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{pmatrix}$$

El orden para resolver este sistema sería:

(2 4 3 6 5 1)

Una posibilidad para resolverlo sería permutando los renglones de **A**, de acuerdo con las transposiciones $\mathbf{P}_{ij}$, que intercambian los renglones *i* y *j*, en alguna de las formas siguientes:

1) $\mathbf{P}_{23}\,\mathbf{P}_{34}\,\mathbf{P}_{45}\,\mathbf{P}_{26}\,\mathbf{A} = \mathbf{U}$

2) $\mathbf{P}_{65}\,\mathbf{P}_{64}\,\mathbf{P}_{63}\,\mathbf{P}_{52}\,\mathbf{A} = \mathbf{U}$

En ambos casos, tenemos una permutación **P**, dada por:

$$\mathbf{P} = \mathbf{P}_{23}\,\mathbf{P}_{34}\,\mathbf{P}_{45}\,\mathbf{P}_{26} = \mathbf{P}_{65}\,\mathbf{P}_{64}\,\mathbf{P}_{63}\,\mathbf{P}_{52}$$

Una forma compacta de guardar esta información es a través de un vector que nos indique el orden en que quedan los renglones después de la permutación:

$$\begin{pmatrix}1\\2\\3\\4\\5\\6\end{pmatrix} \xrightarrow{P_{52}} \begin{pmatrix}1\\5\\3\\4\\2\\6\end{pmatrix} \xrightarrow{P_{63}} \begin{pmatrix}1\\5\\6\\4\\2\\3\end{pmatrix} \tilde{A}\ \xrightarrow{P_{64}} \begin{pmatrix}1\\5\\6\\3\\2\\4\end{pmatrix} \xrightarrow{P_{65}} \begin{pmatrix}1\\5\\6\\3\\4\\2\end{pmatrix}$$

El último vector indica que el sistema es triangular superior si se deja la primera ecuación en su lugar, la quinta se pasa al segundo lugar, la sexta al tercero, la tercera al cuarto, la cuarta al quinto y al último queda la segunda ecuación.

El proceso anterior se puede estructurar en una de dos maneras:.

1) **P A** se calcula explícitamente.

2) **P A** no se calcula explícitamente.

La razón de (1) es por comodidad, la razón de (2) es por las dificultades que implica hacer las permutaciones en cuanto a uso de memoria y tiempo de búsqueda de los elementos.

MÉTODOS DE SOLUCIÓN

Para resolver $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ existen dos clases de métodos de solución:

a) Métodos directos

En un número finito de pasos se obtiene la solución.

b) Métodos iterativos

Construyen una sucesión $\{\mathbf{x}_n\}$ que converge a la solución **x**.

Métodos directos

Factorización LU y eliminación gaussiana

El proceso de transformar una matriz **A** a la forma escalonada es, en realidad, una factorización de la matriz en dos triangulares de la siguiente manera:

$$\mathbf{A = L\ U}$$

De tal modo que resolviendo primero:

$$\mathbf{L\,y = b}$$

Y después:

$$\mathbf{U\,x = y}$$

Tendremos que **x** es solución de:

$$\mathbf{A\,x = b}$$

Consideremos por el momento el problema:

$$\mathbf{L\,y = b}$$

En donde $\mathbf{L} = (\lambda_{ij})$, $\mathbf{y} = (\eta_1, \eta_2, ..., \eta_n)^t$ y $\mathbf{b} = (\beta_1, \beta_2, ..., \beta_n)^t$. Dado que **L** es triangular inferior, tenemos el sistema:

$$\begin{aligned}
&\lambda_{11}\eta_1 + 0 + \ldots\ldots\ldots + 0 = \beta_1\\
&\lambda_{21}\eta_1 + \lambda_{22}\eta_2 + 0 + \ldots + 0 = \beta_2\\
&\vdots\qquad\qquad\qquad\qquad\qquad\vdots\\
&\lambda_{n1}\eta_1 + \lambda_{n2}\eta_2 + \ldots + \lambda_{nn}\eta_n = \beta_n
\end{aligned}$$

Suponiendo que $\lambda_{kk} \neq 0\ \forall_k$, calculamos la solución:

$$\begin{aligned}
&\eta_1 = \beta_1/\lambda_{11}\\
&\eta_2 = (\beta_2 - \lambda_{21}\eta_1)/\lambda_{22}\\
&\vdots\\
&\eta_n = (\beta_n - (\lambda_{n1}\eta_1 + \ldots + \lambda_{n,n-1}\eta_{n-1}))/\lambda_{nn}
\end{aligned}$$

En general, la solución se expresa por:

$$\eta_1 = \beta_1/\lambda_{11}$$

$$\eta_k = (\beta_k - \sum_{j=1}^{k-1} \lambda_{kj}\eta_j)/\lambda_{kk},\ k = 2, 3, ..., n$$

(sustitución hacia adelante.)

Veamos ahora el sistema:

$$\mathbf{U}\,\mathbf{x} = \mathbf{y}$$

En donde $\mathbf{U} = (\mu_{ij})$, $\quad \mathbf{x} = (\xi_1, \xi_2, \ldots \xi_n)^t$

Escrito en términos de *n* ecuaciones tenemos:

$$\begin{aligned}
\mu_{11}\xi_1 + \mu_{12}\xi_2 + \ldots + \mu_{1n}\xi_n &= \eta_1 \\
\mu_{22}\xi_2 + \ldots + \mu_{2n}\xi_n &= \eta_2 \\
\vdots \qquad\qquad & \vdots \\
\mu_{nn}\xi_n &= \eta_n
\end{aligned}$$

Si $\mu_{kk} \neq 0 \ \forall_k$ tenemos:

$$\xi_n = \eta_n / \mu_{nn_n}$$

$$\xi_k = (\eta_k - \sum_{j=k+1} \mu_{kj}\,\xi_j)/\mu_{kk}, \quad k = n\text{-}1, n\text{-}2, \ldots, 1.$$

(Sustitución hacia atrás.)

Descripción esquemática del proceso de factorización LU en el caso en que no aparezcan ceros en la diagonal durante el proceso:

$$Sea\ A_1 = \left(\begin{array}{cccc|c}
\alpha_{11}^{(1)} & \alpha_{12}^{(1)} & \ldots & \alpha_{1n}^{(1)} & \beta_1^{(1)} \\
\alpha_{21}^{(1)} & \alpha_{22}^{(1)} & \ldots & \alpha_{2n}^{(1)} & \beta_2^{(1)} \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\alpha_{n1}^{(1)} & \alpha_{n2}^{(1)} & \ldots & \alpha_{nn}^{(1)} & \beta_n^{(1)}
\end{array}\right)$$

Trabajamos con la matriz aumentada porque las transformaciones aplicadas a la matriz deben aplicarse al vector del lado derecho.

Definimos: $\alpha_{ij}^{(1)} = \alpha_{ij}$ y $\beta_i^{(1)} = \beta_i$

Al hacer la eliminación de la primera columna, obtenemos:

$$A_2 = \left(\begin{array}{cccc|c}
\alpha_{11}^{(1)} & \alpha_{12}^{(1)} & \ldots & \alpha_{1n}^{(1)} & \beta_1^{(1)} \\
0 & \alpha_{22}^{(2)} & \ldots & \alpha_{2n}^{(2)} & \beta_2^{(2)} \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
0 & \alpha_{n2}^{(2)} & \ldots & \alpha_{nn}^{(2)} & \beta_n^{(2)}
\end{array}\right)$$

En donde:

$$\alpha_{ij}^{(2)} = \alpha_{ij}^{(1)} - \lambda_{i1}\,\alpha_{1j}^{(1)}$$

para $j = 2, 3, \ldots, n,\ i = 2, 3, \ldots, n$

En particular:

$0 = \alpha_{21}^{(2)} = \alpha_{21}^{(1)} - \lambda_{21}\,\alpha_{11}^{(1)}$, de donde:

$$\lambda_{i1} = \frac{\alpha_{i1}^{(1)}}{\alpha_{11}^{(1)}}$$

Observamos que para eliminar la columna, el elemento correspondiente de la diagonal debe ser distinto de cero, en este caso es necesario que $\alpha_{11} \neq 0$

En general, encontraremos:

$$A_k = \left(\begin{array}{ccccccc|c} \alpha_{11}^{(1)} & \alpha_{12}^{(1)} & \ldots & \alpha_{1k}^{(1)} & \alpha_{1,k+1}^{(1)} & \ldots & \alpha_{1n}^{(1)} & \beta_1^{(1)} \\ 0 & \alpha_{22}^{(2)} & \ldots & \alpha_{2k}^{(2)} & \alpha_{2,k+1}^{(2)} & \ldots & \alpha_{2n}^{(2)} & \beta_2^{(2)} \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & \ldots & \alpha_{kk}^{(k)} & \alpha_{k,k+1}^{(k)} & \ldots & \alpha_{kn}^{(k)} & \beta_k^{(k)} \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & \ldots & \alpha_{nk}^{(k)} & \alpha_{n,k+1}^{(k)} & \ldots & \alpha_{nn}^{(k)} & \beta_n^{(k)} \end{array}\right)$$

En donde:

$$\alpha_{ij}^{(k)} = \alpha_{ij}^{(k-1)} - \lambda_{i,k-1}\,\alpha_{k-1,j}^{(k-1)}$$

y $$\beta_i^{(k)} = \beta_i^{(k-1)} - \lambda_{i,k-1}\,\beta_{k-1}^{(k-1)}$$

para $j = k, k+1, \ldots, n$ y para $i = k, k+1, \ldots, n$

Dado que, en particular:

$$0 = \alpha_{i,k-1}^{(k)} = \alpha_{i,k-1}^{(k-1)} - \lambda_{i,k-1}\,\alpha_{k-1,k-1}^{(k-1)}$$

si $\alpha_{k-1,k-1}^{(k-1)} \neq 0$ tenemos:

$$\lambda_{i,k-1} = \frac{\alpha_{i,k-1}^{(k-1)}}{\alpha_{k-1,k-1}^{(k-1)}}$$

O bien:

$$\lambda_{ik} = \frac{\alpha_{ik}^{(k)}}{\alpha_{kk}^{(k)}}, \text{ para } k = 1,2,...,n-1, i = k+1,...,n$$

Observemos también que:

$$\lambda_{kk} = 1, \forall_k$$

De este modo, obtenemos:

$$A_n = \begin{pmatrix} \alpha_{11}^{(1)} & \alpha_{12}^{(1)} & \dots & \alpha_{1n}^{(1)} \\ 0 & \alpha_{22}^{(2)} & \dots & \alpha_{2n}^{(2)} \\ . & 0 & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 0 & \dots & \alpha_{nn}^{(n)} \end{pmatrix} \quad b_n = \begin{pmatrix} \beta_1^{(1)} \\ \beta_2^{(2)} \\ . \\ . \\ . \\ \beta_n^{(n)} \end{pmatrix}$$

tales que $A_n x = b_n$

Si definimos:

$$U = A_n$$

$$c = b_n$$

Obtenemos el sistema equivalente:

$$U x = c$$

En donde $c = b_n = (\beta_1^{(1)}, \beta_2^{(2)}, \dots \beta_n^{(n)})^t$

$$= (\gamma_1, \gamma_2, \dots, \gamma_n)$$

y $\beta_1^{(1)} = \beta_1$

$$\beta_2^{(2)} = \beta_2^{(1)} - \lambda_{21}\beta_1^{(1)} = \beta_2 - \lambda_{21}\beta_1^{(1)}$$

$$\beta_3^{(3)} = \beta_3^{(2)} - \lambda_{32}\beta_2^{(2)} = \beta_3^{(1)} - \lambda_{31}\beta_1^{(1)} - \lambda_{32}\beta_2^{(2)}$$

$$\beta_4^{(4)} = \beta_4^{(3)} - \lambda_{43}\beta_3^{(3)} = \beta_4^{(2)} - \lambda_{42}\beta_2^{(2)} - \lambda_{43}\beta_3^{(3)}$$

$$= \beta_4 - \lambda_{41}\beta_1^{(1)} - \lambda_{42}\beta_2^{(2)} - \lambda_{43}\beta_3^{(3)}$$

$$\vdots$$

$$\beta_n^{(n)} = \beta_n - \sum_{j=1}^{n-1} \lambda_{nj}\beta_j^{(j)}$$

Sustituyendo, podemos escribir las ecuaciones anteriores como:

$$\gamma_1 = \beta_1$$

$$\lambda_{21}\gamma_1 + \gamma_2 = \beta_2$$

$$\lambda_{31}\gamma_1 + \lambda_{32}\gamma_2 + \gamma_3 = \beta_3$$

$$\vdots$$

$$\lambda_{n1}\gamma_1 + \lambda_{n2}\gamma_2 + \ldots + l_{n,n-1}\gamma_{n-1} + \gamma_n = \beta_n$$

O bien:

$$\mathbf{L\,c = b},$$

$$\mathbf{c = L^{-1}b}$$

En donde:

$$\mathbf{L} = \begin{pmatrix} 1 & & & \\ \lambda_{21} & 1 & & \\ \lambda_{31} & \lambda_{32} & 1 & \\ \vdots & \vdots & & \ddots \\ \lambda_{n1} & \lambda_{n2} & \lambda_{n3} & l_{n,n-1}\ 1 \end{pmatrix}$$

Entonces, podemos escribir el sistema original como:

$$\mathbf{U\,x = L^{-1}\,b}$$

O bien:

$$\mathbf{L\,U\,x = b}$$

Es decir, L y U son matrices triangulares:

$$L = \begin{pmatrix} 1 & & & \\ * & 1 & & \\ * & \ldots & * & 1 \end{pmatrix} \quad U = \begin{pmatrix} * & \ldots & * \\ & * & \ldots & * \\ & & & * \end{pmatrix}$$

Tales que:

$$\mathbf{A = L\,U}$$

Además $\det(A) = \det(L)\det(U) = 1 \times \prod_{k=1}^{n} \mu_{kk} = \prod_{k=1}^{n} \mu_{kk}$

Nota: Esto es válido si no aparecen ceros en la diagonal durante el proceso.

Eliminacion gausiana: algunos comentarios

Se presenta un problema que afecta al algoritmo, cuando:

$$\alpha_{kk}^{(k)} = 0$$

para alguna k.

La solución usual es permutar el renglón k-ésimo con alguno inferior, de tal manera que el nuevo pivote no sea cero; si esto no es posible, es decir:

$$\text{si} \quad \alpha_{ik}^{(k)} = 0 \quad \text{para } i = k, \ldots, n$$

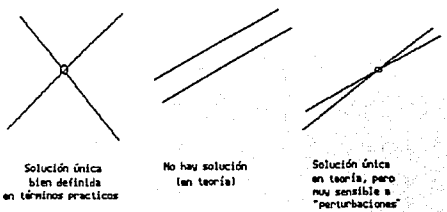
Entonces el problema no tiene solución única, pues la matriz es singular.

Otro problema frecuente es que si se hace eliminación gausiana a una matriz singular, usando aritmética de punto flotante, puede ocurrir que elementos que en teoría son cero aparecerán como cantidades pequeñas y esto cambia cualitativamente el problema. Las matrices siguientes son singulares, sin embargo, con aritmética de punto flotante con 7 cifras decimales al hacer la factorización LU, se contradice el hecho de que son singulares.

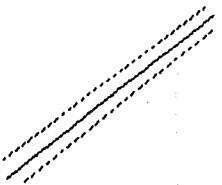
$$\begin{pmatrix} 2 & 6 & 2 & 8 \\ 1 & 4 & 2 & 2 \\ 1 & 1 & 2 & 4 \\ 2 & 5 & 4 & 6 \end{pmatrix} \qquad \begin{pmatrix} 6 & 4 & 5 & 2 \\ 4 & 2 & 1 & 1 \\ 2 & 2 & 4 & 1 \\ 8 & 2 & 6 & 2 \end{pmatrix}$$

Eso plantea la necesidad de revisar nuestros resultados y, de hecho, un buen algoritmo debe someterse a un criterio que juzgue la solución obtenida.
Afortunadamente este problema, y otros parecidos, no afecta de igual manera a todos los sistemas de ecuaciones, es decir, solo algunos sistemas son muy sensibles a perturbaciones en los datos.

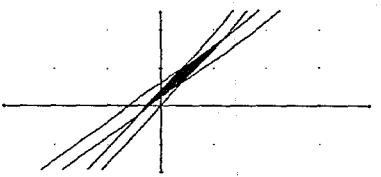
Resolver un sistema 2×2 es equivalente a encontrar la intersección de dos rectas y podemos reconocer tres situaciones.



En el último caso hablamos de *sensibilidad a perturbaciones*; ilustramos esto representando una posible perturbación de la recta en la gráfica siguiente:

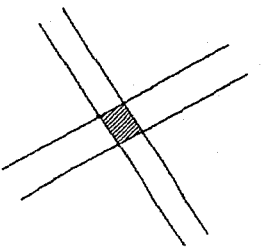


Así, representamos el efecto de tales perturbaciones mediante la siguiente ilustración:



En donde vemos que el *punto de intersección* de las rectas perturbadas puede ser *cualquiera de los puntos del área sombreada*, y la dificultad (sensibilidad) radica en el hecho de que esa área sombreada es muy alargada; es decir, hay problemas para los cuales existe una gran probabilidad de error grande en la solución, producido por un error pequeño en los datos.

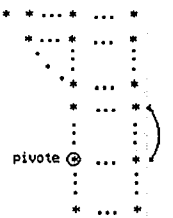
Lo anterior no ocurre en el primer caso; ya que el área de intersección de las "rectas" será similar a la siguiente gráfica.



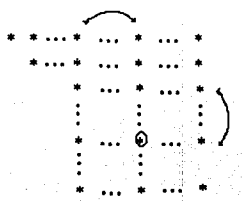
Un problema más grave todavía, en tanto que depende de nosotros y no del problema mismo, es que un problema estable puede *enfermarse* si es atacado por un método inadecuado. Por ello consideramos una versión modificada de la eliminación gausiana (EG).

Modificaciones que hacen a la EG estable.

A) En cada paso, tómese como pivote al elemento más grande en valor absoluto de los restantes de la columna donde se van a introducir los ceros (pivoteo parcial).



B) En cada paso, tómese como pivote, al elemento más grande en valor absoluto de toda la submatriz restante (pivoteo total).



El pivoteo total es mejor que el pivoteo parcial. Aquél es muy bueno pero poco usual, porque resulta mucho más caro que el parcial. La razón es que debe localizarse un elemento de entre n^2 y el parcial sólo busca entre n elementos.

Eliminacion gausiana con pivoteo parcial (egpp)

La idea que está detras de esta técnica es una respuesta a la dificultad, tanto teórica como práctica, que aparece en la eliminación gausiana (EG).

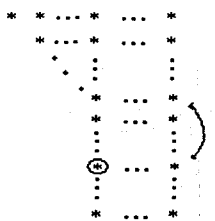
Dificultad de carácter teórico es aquélla que consiste en que, para llevar a cabo la EG, es necesario calcular las cantidades:

$$\lambda_{ij} = \frac{\alpha_{ij}^{(j)}}{\alpha_{jj}^{(j)}}$$

De manera que este procedimiento sólo puede tener buen éxito cuando $\alpha_{jj}^{(j)} \neq 0$ para $j = 1, 2, \ldots, n-1$. $\alpha_{jj}^{(j)} = 0$ no implica que el problema no tiene solución, ya que puede existir una permutación $P_{j,\, mj}$, que intercambia los renglones j y $m_j \geq j$, tal que:

$$\alpha_{mj,\, j} \neq 0$$

Por otro lado, desde el punto de vista práctico, se ha visto la inconveniencia de hacer divisiones con *denominador pequeño* cuando se trabaja en aritmética de punto flotante; en este sentido es conveniente, para cada paso k de la EG, intercambiar el renglón k-ésimo con aquel en el que se encuentre el elemento de mayor magnitud, sobre la columna k-ésima, es decir, si $|\alpha_{m,k}| = \max_{k \le i \le n} |\alpha_{ik}|$ se intercambian los renglones k y m. Si representamos por "*" un elemento diferente de cero de A_k y por " * " el correspondiente máximo, describimos la parte medular de EGPP con el siguiente esquema:



Una causa para que EGPP fracase es que, para alguna k, $1 \le k \le n-1$, se tenga que $\alpha_{i,k}^{(k)} = 0$, para toda $i = k, k+1, ..., n$, lo que significa, al menos en teoría, que la matriz en cuestión es singular.

A continuación, veremos cómo afectan las permutaciones al problema original:

$$\mathbf{A x = b,}$$

Y en particular a la factorización de:

$$\mathbf{A = L U}$$

Supongamos que $\mathbf{A} = (\alpha_{ij})$ es una matriz para la cual EGPP tiene éxito y denotemos por:

$$\mathbf{P}_{k, mk} \; ; \; k \le m_k \le n \; , \; 1 \le k \le n-1$$

a la permutación elemental que lleva al máximo elemento de la columna k a la posición de pivote.

Definimos entonces:

$$\mathbf{A}_1 = \mathbf{A}$$

$$\overline{\mathbf{A}}_1 = \mathbf{P}_{1m1} \mathbf{A}_1$$

Y para construir:

$$A_2 = \begin{pmatrix} \alpha_{11}^{(1)} & & \dots & \alpha_{1n}^{(1)} \\ 0 & \alpha_{22}^{(2)} & \dots & \alpha_{2n}^{(2)} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 0 & \alpha_{n2}^{(2)} & \dots & \alpha_{nn}^{(2)} \end{pmatrix}$$

A diferencia de EG, para la cual:

$$\alpha_{ij}^{(2)} = \alpha_{ij}^{(1)} - \lambda_{i1}\alpha_{1j}^{(1)} \ , \ \lambda_{i1} = \frac{\alpha_{i1}^{(1)}}{\alpha_{11}^{(1)}} \ , \ i \geq 2$$

Tendremos que:

$$\alpha_{1j}^{(2)} = \alpha_{m1,j}$$

$$\overline{\alpha}_{1j}^{(1)} = \alpha_{m1,j}^{(1)}$$

$$\overline{\alpha}_{m1,j}^{(1)} = \alpha_{1j}^{(1)} \qquad j = 1, \dots, n$$

$$\overline{\alpha}_{ij}^{(1)} = \alpha_{ij}^{(1)} \qquad i \neq 1; \ m_1 = 1, \dots n; \ j = 1, \dots n$$

$$\alpha_{ij}^{(2)} = \overline{\alpha}_{ij}^{(1)} - \lambda_{ij}\overline{\alpha}_{ij}^{(1)}$$

Es decir, los multiplicadores λ_{i1}, se calculan en correspondencia, no a A_1, sino a $P_{1m1}A_1$.

En general, para pasar del paso k al $k+1$

$$\overline{A}_k = P_k A_k$$

$$\overline{\alpha}_{kj}^{(k)} = \alpha_{mkj}^{(k)}$$

$$\overline{\alpha}_{mkj}^{(k)} = \alpha_{kj}^{(k)}$$

$$\overline{\alpha}_{ij}^{(k)} = \alpha_{ij}^{(k)}$$

$$i \neq k, mk$$

$$i = k+1, \dots, n$$

$$j = k, k+1, \dots, n$$

Para construir $A_{k+1} = (\alpha_{ij}^{(k+1)})$

$$\alpha_{ij}^{(k+1)} = \overline{\alpha}_{ij}^{(k)} - \lambda_{ik}\overline{\alpha}_{kj}^{(k)}$$

$$j = k+1, \dots, n$$

$$i = k+1, \dots, n$$

$$\lambda_{ik} = \frac{\alpha_{ik}^{(k)}}{\alpha_{kk}^{(k)}}$$

$$i = k+1, \dots, n$$

Así que, al final, obtenemos:

$$A = \begin{pmatrix} \alpha_{11}^{(n)} & \alpha_{12}^{(n)} & \dots & \alpha_{1n}^{(n)} \\ 0 & \alpha_{22}^{(n)} & \dots & \alpha_{2n}^{(n)} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & \cdot & \cdot \cdot & \alpha_{nn}^{(n)} \end{pmatrix} = U$$

Tendremos que proceder de manera similar con el vector **b**, por lo tanto:

$$\mathbf{b}_1 = \mathbf{b}$$

$$\overline{\mathbf{b}}_1 = \mathbf{P}_1 \mathbf{b}$$

$\mathbf{b}_2$ se obtiene de la siguiente manera:

$$\beta_i^{(2)} = \overline{\beta}_i^{(1)} - \lambda_{i1} \overline{\beta}_1^{(1)}$$

$$i = 2, \dots, n$$

En general:

$$\overline{\mathbf{b}}_k = \mathbf{P}_k \mathbf{b}_k$$

Y $\mathbf{b}_{k+1}$ se obtiene:

$$\beta_i^{(k+1)} = \overline{\beta}_i^{(k)} - \lambda_{ik} \overline{\beta}_k^{(k)}$$

$$i = k+1, \dots, n$$

Obtenemos finalmente un sistema triangular de ecuaciones:

$$\mathbf{A}_n \mathbf{x} = \mathbf{b}_n$$

Que se puede resolver fácilmente.

Descripción formal del proceso de egpp

Si denotamos por $\mathbf{M}_k$ a la matriz que nos permite pasar de $\mathbf{A}_k$ a $\mathbf{A}_{k+1}$, el proceso se puede escribir en la forma siguiente:

$$\begin{array}{ll} \mathbf{A}_1 = \mathbf{A} & \mathbf{b}_1 = \mathbf{b} \\ \mathbf{A}_2 = \mathbf{M}_1 \mathbf{P}_1 \mathbf{A}_1 & \mathbf{b}_2 = \mathbf{M}_1 \mathbf{P}_1 \mathbf{b}_1 \\ \mathbf{A}_3 = \mathbf{M}_2 \mathbf{P}_2 \mathbf{A}_2 & \mathbf{b}_3 = \mathbf{M}_2 \mathbf{P}_2 \mathbf{b}_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \mathbf{A}_{k+1} = \mathbf{M}_k \mathbf{P}_k \mathbf{A}_k & \mathbf{b}_{k+1} = \mathbf{M}_k \mathbf{P}_k \mathbf{b}_k \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \mathbf{A}_n = \mathbf{M}_{n-1} \mathbf{P}_{n-1} \mathbf{A}_{n-1} & \mathbf{b}_n = \mathbf{M}_{n-1} \mathbf{P}_{n-1} \mathbf{b}_{n-1} \end{array}$$

Sustituyendo, podemos escribir todo lo anterior en forma compacta, mediante la siguiente expresión:

$$U = A_n = M_{n-1} P_{n-1} M_{n-2} P_{n-2} \cdots M_2 P_2 M_1 P_1 A$$

$$c = b_n = M_{n-1} P_{n-1} \cdots M_2 P_2 M_1 P_1 b$$

El proceso anterior es muy fácil de realizar en una computadora digital.

Almacenamiento de P_i, M_i

Es fácil demostrar que las M_i tienen la forma:

$$M_1 = \begin{pmatrix} 1 & & & \\ -\lambda_{21} & 1 & & \\ -\lambda_{31} & 0 & 1 & \\ \vdots & \vdots & & \ddots \\ -\lambda_{n1} & 0 & \cdots & 1 \end{pmatrix}$$

$$M_2 = \begin{pmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & -\lambda_{32} & 1 & & & \\ 0 & -\lambda_{42} & 0 & 1 & & \\ \vdots & \vdots & & & \ddots & \\ 0 & -\lambda_{n2} & \cdots & & & 1 \end{pmatrix}$$

$$M_k = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & -\lambda_{k+1,k} & 1 & & \\ & & & -\lambda_{k+2,k} & 0 & 1 & \\ & & & \vdots & & & \ddots \\ & & & -\lambda_{n,k} & & & 1 \end{pmatrix}$$

$$M_{n-1} = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ \cdot & & \cdot & & \\ \cdot & & & \cdot & \\ \cdot & & & & \cdot \\ 0 & \cdots & -\lambda_{n,n-1} & & 1 \end{pmatrix}$$

Debido a que casi todos los elementos de las matrices M_i son cero y los de la diagonal son unos, para guardarlos sólo es necesario guardar las λ_{ij}, pero éstas se pueden guardar en los lugares donde se introdujeron los ceros; así que, en un caso típico de orden 6, tenemos que:

$$L = \begin{pmatrix} * & \cdots & * & * \\ -\lambda_{21} & * & & * \\ -\lambda_{31} & -\lambda_{32} & * & \cdot \\ -\lambda_{41} & -\lambda_{42} & * & \cdot \\ -\lambda_{51} & \cdot & & \cdot \\ -\lambda_{61} & \cdot & \cdot \cdot & * \end{pmatrix}$$

Las matrices de permutación se pueden guardar fácilmente en un vector:

$$P = \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ \cdot \\ \cdot \\ \cdot \\ k_{n-1} \\ -^{+}1 \end{pmatrix}$$

Donde k_1 es el renglón que se intercambio en el primer paso, k_2 el correspondiente al segundo y así sucesivamente. El signo ($\pm$) depende del número de permutaciones efectuadas.

Ejemplo: Considere el siguiente sistema de ecuaciones:

i) $2x + y + 3z = 1$

ii) $4x + 4y + 7z = 1$

iii) $2x + 5y + 9z = 3$

Para *eliminar* la incógnita x de las dos últimas ecuaciones, restamos 2 veces la ecuación i) a la ecuación ii) y la ecuación i) de la ecuación iii). Obtenemos el siguiente sistema:

$$2x + y + 3z = 1$$
$$2y + z = -1$$
$$4y + 6z = 2$$

Repitiendo el proceso:

$$2x + y + 3z = 1$$
$$2y + z = -1$$
$$4z = 4$$

Sustituyendo las variables *hacia atrás* tenemos que la última ecuación, da el valor de la variable z. De la penúltima ecuación $z = \frac{4}{4} = 1$ con $z = 1$ obtenemos:

$$y = \frac{-1 - 1}{2} = -1$$

Finalmente, con los valores z y y, de la primera ecuación:

$$x = \frac{1 - 3 - (-1)}{2} = -\frac{1}{2}$$

La representación matricial del sistema (1) es:

$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 4 & 7 \\ 2 & 5 & 9 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}$$

$$\mathbf{A\,x = b}$$

Que hemos transformado mediante eliminación gaussiana, cuya representación matricial es:

$$\begin{pmatrix} 2 & 1 & 3 \\ 0 & 2 & 1 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 4 \end{pmatrix}$$

$$\mathbf{U\,x = c}$$

Factorización LU y eliminación gaussiana

Los métodos más populares para resolver sistemas de ecuaciones son variantes del algoritmo de eliminación gaussiana. Por eliminación, entendemos el procedimiento desarrollado en el ejemplo, mediante el cual los múltiplos de una ecuación son sumados a otras ecuaciones, de tal manera que los coeficientes de una de las variables de dichas ecuaciones se hagan ceros. Esto se repite hasta que el sistema final de ecuaciones tiene como matriz de coeficientes una matriz triangular; cuando el sistema tiene esta forma, entonces es muy sencillo de resolver.

Algunos autores prefieren presentar la solución de sistemas de ecuaciones lineales empleando la idea de eliminación; otros, a través de la idea de factorización triangular. Resulta importante observar que los dos enfoques son equivalentes desde el punto de vista matemático: la factorización triangular es un concepto importante en el trabajo más avanzado con matrices. Por otro lado, los métodos de Cholesky, Doolitle y otros aprovechan las características de la matriz: diagonal, de banda, esparcida, etcétera, para hacer más eficientes los cálculos y el almacenamiento en la computadora, pero esencialmente son métodos de eliminación o factorización.

Ejemplo: Resolver el siguiente sistema de ecuaciones mediante eliminación gaussiana con pivoteo parcial.

$$2x_0 + 4x_1 + x_2 + 2x_3 = 5$$
$$4x_0 + 14x_1 - x_2 + 6x_3 = 11$$
$$x_0 - x_1 + 5x_2 - x_3 = 9$$
$$-4x_0 + 2x_1 - 6x_2 + x_3 = -2$$

Programa de eliminación gaussiana con pivoteo parcial

```
10 REM EGPP
20 REM ***** ENTRADA DE DATOS *****
30 INPUT "NUMERO DE ECUACIONES ";N
40 LET N = N-1:PRINT
50 DIM A(20,21),X(20)
60 FOR R = 0 TO N
70 FOR S = 0 TO N
80 PRINT "A(";R;",";S;") = ";
90 INPUT A(R,S)
100 NEXT S
110 PRINT :PRINT "B(";R;") = ";
120 INPUT A(R,N + 1)
```

```
130 PRINT:PRINT:NEXT R
200 REM ***** ELIMINACION *****
210 FOR Z=0 TO N-1
215 FOR R=Z+1 TO N
220 LET W=0
230 FOR R=Z TO N
240IF ABS (A(R,Z))>W THEN LET U=R: LET W=ABS(A(R,Z))
250 NEXT R
270 FOR S=Z TO N+1
280 LET P=A(U,S)
290 LET A(U,S)=A(Z,S)
300 LET A(Z,S)=P
310 NEXT S
410 LET P=A(R,Z)/A(Z,Z)
420 FOR S=Z+1 TO N+1
430 LET A(R,S)=A(R,S)-P*A(Z,S)
440 NEXT S
450 NEXT R
500 NEXT Z
600 REM ***** SUSTITUCION DE VALORES *****
610 FOR R=N TO 0 STEP-1
620 LET P=A(R,N+1)
630 IF R=N THEN GOTO 670
640 FOR S=R+1 TO N
650 LET P=P-A(R,S)*X(S)
660 NEXT S
670 LET X(R)=P/A(R,R)
680 NEXT R
700 REM ***** SE IMPRIMEN RESULTADOS *****
710 FOR R=0 TO N
720 PRINT "X(";R;") =";X(R)
730 NEXT R
```

La solución es: x(0) = -3; x(1) =1; x(2) =3; x(3) =2

Sensibilidad de un sistema de ecuaciones

Medida de la condición

Resulta obvio que la solución de:

$$A x = b$$

es sensible o no, dependiendo de la matriz A.

En esta sección, desarrollamos la forma de cuantificar la sensibilidad de un sistema de ecuaciones lineales.

Para llevar acabo esta tarea, necesitamos medir la "magnitud" de los vectores y las matrices; hacemos esto a través de funciones conocidas como normas.

Estas funciones son la generalización del valor absoluto que usamos para números reales, por esa razón utilizamos una notación parecida, la norma de **x** se denotará como $\| x\|$; veremos varias formas de definir la magnitud o norma de un vector, que dependerá de la aplicación que hagamos de la misma.

Debido a lo anterior, la comunidad matemática ha llegado a un acuerdo sobre las propiedades que debe tener una función que quiera asignar magnitudes a vectores; llamaremos normas a las funciones que satisfagan esos requisitos.

Norma: es una función real definida en R^n que satisface las propiedades siguientes:

1) $\| x\| \geq 0 \ \forall x \in R^n$ y $\| x\| = 0 \Leftrightarrow x \equiv 0$.

2) $\| \alpha\ x\| = |\alpha|\ \| x\|$, $\forall\ x \in R^n$ y $\alpha \in R$.

3) $\| x + y\| \leq \| x\| + \| y\|$, $\forall\ x, y \in R^n$.

Así, para definir una norma, basta escoger una función que satisfaga las propiedades anteriores.

Los siguientes son ejemplos conocidos y de frecuente aplicación:

1. Norma euclidiana $\| x\|_2 = (\sum_{i=1}^{n} x_i^{\,2})^{1/2}$

Esta norma es interesante por el hecho de que es inducida por un producto escalar, definido según:

$$< x, y > \ \equiv \ x^t y = \sum_{i=1}^{n} x_i\, y_i,$$

De modo que:

$$x^t x = \sum_{i=1}^{n} x_i^{\,2}$$

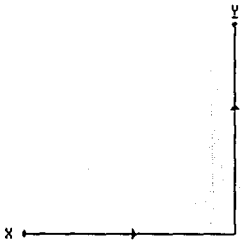
Y entonces: $\| x\|_2^{\,2} = x^t x$

Por brevedad, no demostraremos que $\| * \|_2$ es una norma, esperamos que eso resulte familiar.

2. Norma "del taxista" $\| x\|_1 = \sum_{i=1}^{n} x_i$

Su "alias" proviene del hecho de que si tomamos $\| x - y\|_1$ como una medida de la distancia entre **x** e **y**, esa medida corresponde a la distancia real, medida no

en línea recta, sino dando rodeos, como cuando uno se desplaza en automóvil de un punto a otro de la ciudad, como se indica en la siguiente figura:



3) $\| x \|_{\infty} = \max_{1 \le i \le n} x_i$

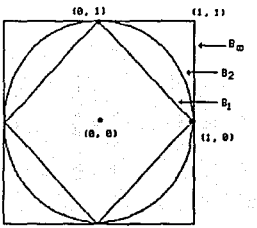
La elección de una norma u otra será un asunto de conveniencia, de acuerdo con la aplicación que se quiera hacer. Estamos acostumbrados a pensar que la norma "buena" es la norma euclidiana, ya que está basada en el teorema de Pitágoras; sin embargo, observamos que la norma del taxista es más realista cuando se trata de medir distancias recorridas. Una conveniencia de la elección de una norma está en la geometría, ya que los conjuntos de puntos que están a la misma distancia de un punto dado tendrán forma distinta. Para ejemplificar lo anterior, definamos los siguientes conjuntos:

$$B_2 = \{x \in R^2 \mid \| x \|_2 = 1\}$$

$$B_1 = \{x \in R^2 \mid \| x \|_1 = 1\}$$

$$B_{\infty} = \{x \in R^2 \mid \| x \|_{\infty} = 1\},$$

Llamaremos "bolas unitarias" a estos conjuntos, de acuerdo con cada una de las normas que corresponden a la figura siguiente:



En general, se pueden definir infinidad de normas de manera similar a las anteriores.

Definición:

Una norma p, para $p = 1, 2, 3, \ldots$ es de la forma:

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

Resultado interesante es el siguiente:

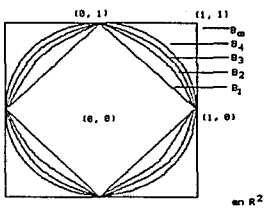
$$\| x \|_{\infty} = \lim_{p \to \infty} \| x \|_p \ , \ \forall \ x \in R^n.$$

La demostración no es difícil, sin embargo se requiere de resultados intermedios, que no veremos aquí.

Podemos interpretar geométricamente el significado de las normas p dibujando las bolas unitarias, para cada p.
Sea:

$$B_p = \{ x \in R^n \mid \| x \|_p = 1 \} \ ,$$

gráficamente, obtenemos:



Normas de matrices

Existe una variedad amplia de normas para matrices, con la propiedad de que pueden obtenerse a partir de las normas de R^n.

Algunas de esas normas son las siguientes:

1) $\| A \|_1 = \max \{ \| a_1 \|_1, \| a_2 \|_1, ..., \| a_n \|_1 \}$ donde a_1, a_2, ..., a_n son las columnas de A.

2) $\| A \|_{\infty} = \max \{ \| r_1 \|_1, \| r_2 \|_1, ..., \| r_n \|_1 \}$ donde r_1, r_2, ..., r_n son los renglones de A.

3) $\| A \|_2$ = El valor característico más grande de la matriz $A^t A$."

Nótese que la norma euclidiana de una matriz es más costosa de evaluar que las otras.

Ejemplo: Si $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ entonces:

$\| A \|_1 = 6$

$\| A \|_{\infty} = 7$

Propiedades fundamentales:

i) $\| Ax \| \leq \| A \| \cdot \| x \|$

ii) $\dfrac{\| x \|}{\| A \|} = \| A^{-1} x \|$

iii) $\| AB \| \leq \| A \| \cdot \| B \|$

iv) $\| A\| \cdot \| A^{-1}\| \geq 1$

Ahora estamos en posibilidades de analizar el efecto de las perturbaciones en la solución de los sistemas de ecuaciones.

Dado el sistema:

$$A x = b$$

Y suponiendo que:

$$A \bar{x} = b + e$$

Surge la pregunta: ¿Qué tanto difieren x y $\bar{x}$? Para responderla, hacemos el siguiente análisis.

Dado que:

$$A \bar{x} = b + e$$

Y $\quad b = Ax$

Entonces:

$$A \bar{x} = Ax + e$$

$$\Rightarrow \quad A (\bar{x} - x) = e$$

$$\Rightarrow \quad \bar{x} - x = A^{-1} e$$

$$\| x - x\| = \| A^{-1} e\| = \| A^{-1}\| \cdot \| e\|$$

Por lo tanto: $\| x - x\| \leq \| A^{-1}\| \cdot \| e\|$

Sin embargo, esta expresión no es muy útil ya que necesitamos el error relativo:

$$\frac{\| x - x \|}{\| x \|} \leq \frac{\| A^{-1} \| \cdot \| e \|}{\| x \|}$$

Reescribiendo esto en términos del error relativo en **b**, obtenemos:

$$\frac{\| x - x \|}{\| x \|} \leq \frac{\| A^{-1} \| \cdot \| b \|}{\| x \|} \cdot \frac{\| e \|}{\| b \|}$$

Esta expresión no es fácil de calcular, ya que depende de A^{-1} y de la solución x, por lo que sería bueno contar con alguna otra expresión que no dependa de la solución exacta, que no conocemos en los casos prácticos.

Obsérvese que:

$$\| b\| = \| Ax\| = \| A\| \cdot \| x\|$$

Por lo que, si sustituimos en la expresión anterior, tenemos que:

$$\frac{\| x - \bar{x} \|}{\| x \|} \leq \frac{\| A^{-1} \| \cdot \| A \| \cdot \| x \|}{\| x \|} \cdot \frac{\| e \|}{\| b \|}$$

$$\frac{\| x - \bar{x} \|}{\| x \|} \leq \| A^{-1} \| \cdot \| A \| \cdot \frac{\| e \|}{\| b \|}$$

Así, obtenemos que el error relativo en **x** depende del error relativo en **b** y del número $\| A \| \cdot \| A^{-1} \|$. Este número es muy importante y lo conocemos como el número de condición de **A**, que expresamos de la siguiente forma:

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

Este nos mide la tendencia de **A** a amplificar los errores en los datos y $k(A) \geq 1$

Usualmente un problema es más sensitivo a error en los datos, o durante el proceso de solución, conforme mayor sea el número $k(A)$.

Un criterio práctico para saber que tan buena es la solución de un sistema de ecuaciones lineales es el siguiente:

Si $\bar{x}$ es la solución obtenida con eliminación gaussiana y pivoteo parcial, y estamos trabajando con una computadora que tiene *t* cifras decimales en la mantisa, entonces:

$$\frac{\| x - \bar{x} \|}{\| x \|} \approx k(A) \cdot 10^{-t}$$

Afortunadamente la mayoría de las rutinas que utilizan EGPP para calcular la solución de **A x = b** pueden calcular aproximadamente $k(A)$ en forma económica.

Es fundamental utilizar este criterio; de otra forma no podemos tener idea de la bondad de la solución de un sistema de ecuaciones lineales. A través de la siguiente expresión, podemos estimar el número de cifras correctas de la solución.

$$\frac{\| x - \bar{x} \|}{\| x \|} \approx k(A)\, 10^{-t} \approx 10^{-r}$$

r es el número de cifras correctas, al menos, de la solución.

Ejemplos: Dadas las matrices **N**, **C** y **H**, sus números de condición son: $k(N) = 3.9454$, $k(C) = 3.1865 \cdot 10^{17}$, $k(H) = 524.0568$ respectivamente, lo que indica que la matriz **C**, por la magnitud de $k(C)$ conduce a un sistema de ecuaciones **Cx = b** inestable.

La matriz mejor condicionada es la matriz **N**, ya que $k(N) < k(H) < k(C)$

$$N = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 5 & 1 \\ 2 & 1 & 6 \end{pmatrix}$$

$$H = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

Métodos iterativos

Para sistemas de orden moderado ($n \leq 200$) la eliminación gaussiana es eficiente y es el método recomendado en general. Para sistemas lineales de órdenes altos, que aparecen, por ejemplo, en la solución de ecuaciones diferenciales, los métodos iterativos son más eficientes.

Método de Jacobi

En el capítulo anterior se utilizaron en forma eficiente los métodos iterativos para resolver ecuaciones con una incógnita. Por lo tanto, resulta natural preguntarse si pueden utilizarse para resolver sistemas de ecuaciones.

Considere el siguiente sistema de ecuaciones:

$$5x + y + z = 10$$

$$x + 6y - 2z = 7$$

$$x - 3y + 7z = 16$$

Las ideas del capítulo anterior sugieren el proceso iterativo siguiente:

$$x_{n+1} = \frac{1}{5}(10 - y_n - z_n)$$

$$y_{n+1} = \frac{1}{6}(7 - x_n + 2z_n)$$

$$z_{n+1} = \frac{1}{7}(16 - x_n + 3y_n)$$

En general, consideremos un sistema:

$$\mathbf{A}\,\mathbf{x} = \mathbf{b}$$

Donde la matriz A es no singular de grado n y en la que se han intercambiado los renglones de la matriz, de tal manera que:

$$a_{ii} \neq 0 \quad \forall\, i$$

Entonces, el sistema $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ puede reescribirse de la siguiente forma:

$$x_1 = -\frac{1}{a_{11}}(a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n - b_1)$$

$$x_2 = -\frac{1}{a_{22}}(a_{21}x_1 + a_{23}x_3 + \ldots + a_{2n}x_n - b_2)$$

.
.
.

$$x_n = -\frac{1}{a_{nn}}(a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{n,n1}x_{n-1} - b_n)$$

Suponiendo que tenemos una aproximación inicial a la solución:

$$\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots x_n^{(0)})^t$$

Sustituyendo $\mathbf{x}^{(0)}$ en el lado derecho del sistema anterior y evaluando, los elementos del vector resultante proporcionan la siguiente aproximación.

Las expresiones recursivas generales se dan a continuación:

$$x_1^{(n+1)} = -\frac{1}{a_{11}}(a_{12}x_2^{(n)} + a_{13}x_3^{(n)} + \ldots + a_{1n}x_n^{(n)} - b_1)$$

$$x_2^{(n+1)} = -\frac{1}{a_{22}}(a_{21}x_1^{(n)} + a_{23}x_3^{(n)} + \ldots + a_{2n}x_n^{(n)} - b_2)$$

.
.
.

$$x_n^{(n+1)} = -\frac{1}{a_{nn}}(a_{n1}x_1^{(n)} + a_{n2}x_2^{(n)} + \ldots + a_{n,n1}x_{n-1}^{(n)} - b_n)$$

El método dado por las fórmulas recursivas anteriores define el *método de Jacobi.*

Se puede demostrar que, bajo ciertas condiciones,

$$\{\mathbf{x}^{(n)}\} \to \mathbf{x} \quad \text{cuando } n \to \infty$$

Una de tales condiciones es que:

$$|a_{ii}| > \sum |a_{ij}|\,;\ i,j = 1,2,\ldots n;\ i \neq j$$

Si se satisface esta condición, se dice que A es *diagonalmente dominante.*

El criterio para detener el proceso iterativo es:

1) El número de iteraciones ha excedido a un número predeterminado *m*.

2) La diferencia entre valores sucesivos de x es menor que alguna tolerancia ε.

Método de Gauss-Seidel

Si consideramos el método iterativo de Jacobi, observamos que para calcular el nuevo valor de $x_2^{(n+1)}$, usamos el valor previo de x_1, esto es $x_1^{(n)}$; a pesar de que ya conocemos un valor *más actualizado*, esto es $x_1^{(n+1)}$ que obtuvimos en el paso anterior. De manera análoga, para obtener $x_3^{(n+1)}$, usamos los valores $x_1^{(n)}, x_2^{(n)}$, a pesar de que nuevos, y presumiblemente más exactos, valores de $x_1^{(n+1)}, x_2^{(n+1)}$ ya estan a nuestra disposición.

Una modificación al método de Jacobi, que generalmente es más rápida en su convergencia, y que usa los valores $x_i^{(n+1)}$ previamente calculados da lugar al método de Gauss-Seidel, es el definido por las siguientes fórmulas recursivas.

$$x_1^{(n+1)} = -\frac{1}{a_{11}}(a_{12}x_2^{(n)} + a_{13}x_3^{(n)} + \ldots + a_{1n}x_n^{(n)} - b_1)$$

$$x_2^{(n+1)} = -\frac{1}{a_{21}}(a_{21}x_1^{(n+1)} + a_{23}x_3^{(n)} + \ldots + a_{2n}x_n^{(n)} - b_2)$$

$$\vdots$$

$$x_n^{(n+1)} = -\frac{1}{a_{nn}}(a_{n1}x_1^{(n+1)} + a_{n2}x_2^{(n+1)} + \ldots + a_{n,n-1}x_{n-1}^{(n+1)} - b_n)$$

El método de Gauss-Seidel tiene la ventaja de que sólo un valor de cada incógnita necesita almacenarse en cualquier momento.

Puede demostrarse que el método de Gauss-Seidel será convergente para el sistema $Ax = b$ si la matriz A es diagonalmente dominante.

Ejemplo: Aplique los métodos de Jacobi y Gauss-Seidel en la solución del siguiente sistema:

$$64x - 3y - z = 14$$

$$x + y + 40z = 20$$

$$2x - 90y + z = -5$$

La matriz de coeficientes no es diagonalmente dominante, ya que en la segunda ecuación tenemos:

$$|1| < |1| + |40|$$

Pero, intercambiando la segunda y tercera ecuaciones, tenemos un sistema con una matriz de coeficientes diagonalmente dominante.

$$64x - 3y - z = 14$$

$$2x - 90y + z = -5$$

$$x + y + 40z = 20$$

El proceso iterativo de Jacobi se ha codificado en el siguiente programa:

```
10 REM PROCESO DE JACOBI
20 LET R=0:LET X=0:LET Y=0:LET Z=0
30 LET X1=(10-Y-Z)/5
40 LET Y1=(7-X+2*Z)/6
50 LET Z1=(16-X+3*Y)/7
60 LET R=R+1:LET X=X1:LET Y:Y1:LET Z=Z1
65 CONDICION DE PARO
70 PRINT "X(";R;") =";X
80 PRINT "Y(";R;") =";Y
90 PRINT "Z(";R;") =";Z
100 PRINT:GOTO 30
```

La solución es: X(10) = .22949 y(10) = 6.609454 E-02 z(10) = .4892138 Codificación del método de Gauss-Seidel

```
10 REM PROCESO DE GAUSS-SEIDEL
20 LET R=0:LETX=0:LET Y=0:LET Z=0
30 LET X=(10-Y-Z)/5
40 LET Y=(7-X+2*Z)/6
50 LET Z=(16-X+3*Y)/7
60 LET R=R+1
70 PRINT "X(";R;") =";X
80 PRINT "Y(";R;") =";Y
90 PRINT "Z(";R;") =";Z
100 PRINT:GOTO 30
```

La solución es: x(4) = 2294949 y(4) = 6.609454 E-02 z(4) = .4892138

EJERCICIOS Y PROBLEMAS*

1. Resuelva los siguientes sistemas de ecuaciones usando eliminación gaussiana. Trabaje con cuatro cifras significativas.

(a)
$$\begin{aligned} 3x_1 + 4x_2 + 3x_3 &= 16 \\ x_1 + 5x_2 - x_3 &= -12 \\ 6x_1 + 3x_2 + 7x_3 &= 102 \end{aligned}$$

(b)
$$\begin{aligned} 3x + 2y - 5z &= 4 \\ 2x - 3y + z &= 8 \\ x + 4y - z &= -3 \end{aligned}$$

(c)
$$\begin{pmatrix} 1 & -1 & 2 & 1 \\ 3 & 2 & 1 & 4 \\ 5 & 8 & 6 & 3 \\ 4 & 2 & 5 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \end{pmatrix}$$

2. Aplique el método de eliminación gaussiana a los siguientes sistemas. Comente los problemas y resuélvaUn rlos por otros métodos.

(a)
$$\begin{aligned} 3x_1 + 2x_2 &= 4 \\ -x_1 - \frac{2}{3}x_2 &= 1 \end{aligned}$$

(b)
$$\begin{aligned} 6x_1 - 3x_2 &= 6 \\ -2x_1 + x_2 &= -2 \end{aligned}$$

(c)
$$\begin{aligned} 0x_1 + 2x_2 &= 4 \\ x_1 - x_2 &= 5 \end{aligned}$$

(d)
$$\begin{aligned} x_1 + x_2 + 2x_3 &= 4 \\ x_1 + x_2 + 0x_3 &= 2 \\ 0x_1 + x_2 + x_3 &= 0 \end{aligned}$$

3. La matriz A puede ser factorizada como A = LU, donde L es una matriz triangular inferior y U es una matriz triangular superior. Encuentre la descomposición LU para la matriz.

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 \\ -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix}$$

4. Dada la matriz:

$$A = \begin{pmatrix} 2 & -1 & 2 \\ 2 & -3 & 3 \\ 6 & -1 & 8 \end{pmatrix}$$

a) Determine la factorización de la matriz **A = LDU** donde **L** es triangular inferior, **D** es diagonal y **U** es triangular superior.

*Los ejercicios se tomaron y/o adaptaron de las obras citadas al final de esta sección.

b) Use la descomposición de **A** para resolver el sistema $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ donde $\mathbf{b} = (-2, -5, 0)^t$

5. Repita el problema anterior para:

$$A = \begin{pmatrix} 2 & 1 & -2 \\ -4 & 3 & -3 \\ 2 & 2 & 4 \end{pmatrix} \quad b = \begin{pmatrix} 1 \\ 4 \\ 4 \end{pmatrix}$$

6. Considere la matriz y vectores siguientes:

$$A = \begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix} \quad b = \begin{pmatrix} 0.217 \\ 0.254 \end{pmatrix}$$

$$x = \begin{pmatrix} 0.999 \\ 1.001 \end{pmatrix} \quad x = \begin{pmatrix} 0.341 \\ -0.087 \end{pmatrix}$$

Calcule los vectores residuales:

$$\mathbf{r}^* = \mathbf{A}\,\mathbf{x}^* - \mathbf{b}$$

$$\bar{\mathbf{r}} = \mathbf{A}\,\bar{\mathbf{x}} - \mathbf{b}$$

Y decida qué vector $\bar{\mathbf{x}}$, $\mathbf{x}^*$ es el vector solución. Calcule los vectores error.

$$\bar{\mathbf{e}} = \mathbf{x} - \bar{\mathbf{x}}$$

$$\mathbf{e}^* = \mathbf{x} - \mathbf{x}^*$$

Donde $\mathbf{x} = (1, -1)^t$ es la solución exacta.

7. Aplique el método de eliminación gaussiana con pivoteo parcial a los sistemas de los problemas 1 y 2.

8. Resuelva el siguiente sistema con EGPP.

$$\begin{pmatrix} 0.4096 & 0.1234 & 0.3678 & 0.2943 \\ 0.2246 & 0.3872 & 0.4015 & 0.1129 \\ 0.3645 & 0.1920 & 0.3781 & 0.0643 \\ 0.1784 & 0.4002 & 0.2786 & 0.3927 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0.4043 \\ 0.1550 \\ 0.4240 \\ 0.2557 \end{pmatrix}$$

9. Si el elemento $a_{31} = 0.3645$ de la matriz del ejemplo anterior se cambia por $a_{31} = 0.3345$. Resuelva el sistema y comente el efecto que este *pequeño error* tiene en el resultado.

10. ¿Cuál es el contenido final de la siguiente matriz, después de aplicar el proceso de eliminación gaussiana?

$$\begin{pmatrix} 1 & 3 & 2 & 1 \\ 4 & 2 & 1 & 2 \\ 2 & 1 & 2 & 3 \\ 1 & 2 & 4 & 1 \end{pmatrix}$$

11. Considera el sistema:

$$10^{-4}x_1 + x_2 = b_1$$

$$x_1 + x_2 = b_2$$

Donde $b_1 \neq 0$ y $b_2 \neq 0$, su solución exacta es:

$$x_1 = \frac{-b_1 + b_2}{1 - 10^{-4}} \qquad x_2 = \frac{b_1\ 10^{-4} b_2}{1 - 10^{-4}}$$

a) Resuelva el sistema mediante eliminación gaussiana, con $b_1 = 1$, $b_2 = 2$. Compare sus resultados con la solución exacta $x_1 = 1.00010$ $x_2 = 0.999899$

b) Repita el proceso usando eliminación gaussiana con pivoteo parcial.

c) Encuentre valores para b_1 y b_2, de tal manera que la eliminación gaussiana no produzca pobres resultados.

12. Usando eliminación gaussiana con pivoteo parcial, reduzca la siguiente matriz, mostrando las matrices intermedias.

$$\begin{pmatrix} 1 & 0 & 3 & 0 \\ 0 & 1 & 3 & -1 \\ 3 & -3 & 0 & 6 \\ 0 & 2 & 4 & -6 \end{pmatrix}$$

13. Demuestre que el siguiente sistema de ecuaciones posee una solución cuando $\alpha = 0$, no tiene solución si $\alpha = -1$ y una infinidad cuando $\alpha = 1$

$$x_1 + 4x_2 + \alpha x_3 = 6$$
$$2x_1 - x_2 + 2\alpha x_3 = 3$$
$$\alpha x_1 + 3x_2 + x_3 = 5$$

14. Dada la matriz A:

a) Determine la matriz triangular inferior M y la matriz triangular superior U tal que $\mathbf{MA} = \mathbf{U}$.

b) Determine $\mathbf{M}^{-1} = \mathbf{L}$ tal que $\mathbf{A} = \mathbf{LU}$

$$A = \begin{pmatrix} 25 & 0 & 0 & 0 & 1 \\ 0 & 27 & 4 & 3 & 2 \\ 0 & 54 & 58 & 0 & 0 \\ 0 & 108 & 116 & 0 & 0 \\ 100 & 0 & 0 & 0 & 24 \end{pmatrix}$$

15. Considere a la matriz

$$A = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix}$$

a) Muestre que **A** no puede ser factorizada en el producto de una matriz triangular inferior y una matriz triangular superior.

b) Intercambie los renglones de **A**, de tal manera que **A** pueda factorizarse.

16. Considere la matriz:

$$A = \begin{pmatrix} a & 0 & 0 & z \\ 0 & b & 0 & 0 \\ 0 & x & c & 0 \\ w & 0 & y & d \end{pmatrix}$$

a) Determine la matriz triangular inferior **M** y en la triangular superior **U** tal que **MA** = **U**

b) Determine la matriz triangular inferior **L** y la matriz triangular superior **U** tal que **A** = **LU**

17. Considere la matriz:

$$A = \begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix}$$

Factorice **A** de las siguientes maneras

a) **A** = **LU** donde **L** es triangular inferior y **U** es triangular superior

b) **A** = **LDU** **L** triangular inferior, **D** diagonal, **U** triangular superior

18. Evalúe el determinante de **A** del ejercicio anterior.

19. Considere la matriz de Hilbert de orden 3

$$A = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$$

Repita los problemas 17 y 18 usando esta matriz.

20. Resuelva el siguiente sistema de ecuaciones, reteniendo solamente cuatro cifras significativas en cada paso y compare la solución, cuando en lugar de cuatro se retienen ocho cifras significativas.

$$0.1026x + 0.2122y = 0.7381$$

$$0.2081x + 0.4247y = 0.9327$$

21. ¿Para qué valores de α, la eliminación Gaussiana produce respuestas erroneas del siguiente sistema? (explique qué sucede en la computadora).

$$x_1 + x_2 = 2$$

$$\alpha x_1 + x_2 = 2 + \alpha$$

22. Use el método de Jacobi para resolver:

$$10x_1 - 3x_2 + 6x_3 = 24.5$$

$$1x_1 + 8x_2 - 2x_3 = -9$$

$$-2x_1 + 4x_2 - 9x_3 = -50$$

23. Resuelva el problema 22 usando el método de Gauss-Seidel con un criterio de paro $\varepsilon = 10\%$

24. Use el método de Gauss-Seidel para resolver ($\varepsilon = 5\%$):

$$x_1 + 7x_2 - 3x_3 = -51$$

$$4x_1 - 4x_2 + 9x_3 = 61$$

$$12x_1 - x_2 + 3x_3 = 8$$

25. Use el método de Gauss-Seidel para resolver ($\varepsilon = 5\%$):

$$-6x_1 + 12x_3 = 60$$

$$4x_1 - x_2 - x_3 = -2$$

$$6x_1 + 8x_2 = 44$$

26. Resuelva el siguiente conjunto de ecuaciones:

$$4x_1 - 2x_2 - x_3 = 39$$

$$x_1 - 6x_2 + 2x_3 = -28$$

$$x_1 - 3x_2 + 12x_3 = -86$$

Usando a) eliminación gaussiana, b) el método de Jacobi y c) el método de Gauss-Seidel ($\varepsilon = 5\%$).

27. Resuelva el siguiente sistema de ecuaciones:

$$x_1 - 3x_2 + 12x_3 = 10$$

$$5x_1 - 12x_2 + 2x_3 = -33$$

$$x_1 - 14x_2 = -103$$

Usando a) eliminación gaussiana, b) el método de Jacobi y c) el método de Gauss-Seidel ($\varepsilon = 5\%$).

28. Un ingeniero supervisa la producción de tres tipos de automóviles. Se requiere de tres clases de materiales -metal, plástico y caucho- para la producción. La cantidad necesaria para producir cada automóvil es de:

Automóvil	Metal, kg/auto	Plástico, kg/auto	Caucho, kg/auto
1	1500	25	100
2	1700	33	120
3	1900	42	160

Si se dispone de un total de 106 toneladas de metal, 2.17 toneladas de plástico y 8.2 toneladas de caucho diariamente, ¿cuántos automóviles se pueden producir por día?

29. Un ingeniero requiere 4 800m^3 de arena, 5 810m^3 de grava fina y 5690m^3 de grava gruesa para la construcción de un proyecto. Existen tres bancos donde se pueden obtener estos materiales. La composición en cada banco es de:

Banco %	Arena %	Grava fina %	Grava gruesa %
banco 1	52	30	18
banco 2	20	50	30
banco 3	25	20	55

¿Cuántos metros cúbicos se debe tomar de cada banco para cumplir con las necesidades del ingeniero?

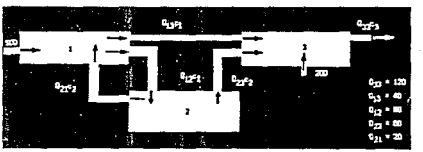
30. La figura muestra tres reactores ligados por tubos. Como se puede ver, la velocidad de tranferencia de sustancias químicas a través de los tubos es igual a la velocidad de flujo (Q, con unidades de metros cúbicos por segundo) multiplicada por la concentración del reactor del cual surge el flujo (c, con unidades de milígramos por metro cúbico). Si el sistema es estacionario, la transferencia en cada reactor balancea la transferencia de salida. Por ejemplo, en el reactor 1, (entrada) = (salida), o:

$$500 + Q_{21}c_2 = Q_{12}c_1 + Q_{13}c_1$$

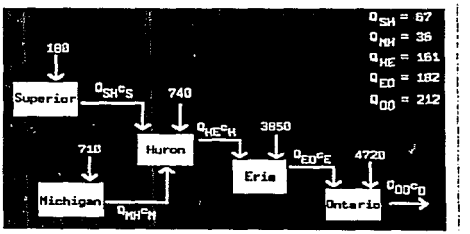
O, usando las velocidades de flujo especificadas como en la figura:

$$500 + 20c_2 = 80c_1 + 40c_1$$

En donde 500 es una entrada directa (milígramos por segundo). Desarróllense ecuaciones de balance de masas comparables para cada uno de los otros reactores y resuélvanse las tres ecuaciones algebraicas lineales simultáneas para la concentración en los reactores.



31. Empleando el mismo planteamiento básico del problema 30, determínese la concentración del cloruro en cada uno de los Grandes Lagos con la información de la figura.



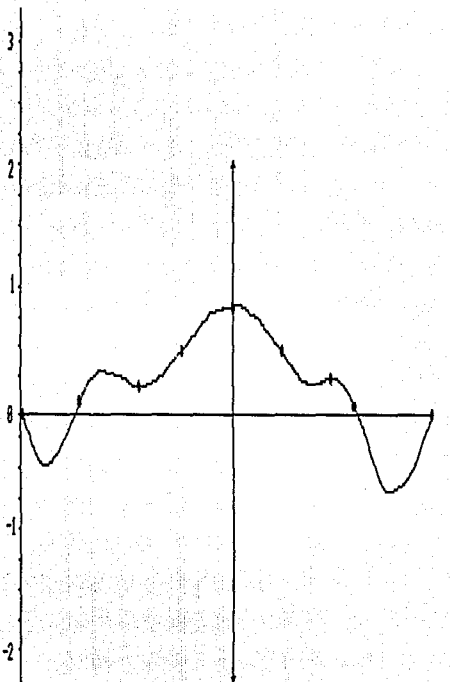
Referencias bibliográficas

1. *Chapra, Steven C; Canale, Raymond P.* "Métodos Numéricos para Ingenieros". México. McGraw-Hill, 1987.

2. *Cheney, Ward; Kincaid, David.* "Numerical Mathematics and Computing". Monterey USA. Brooks/Cole Publishing, 1980.

CAPÍTULO V

INTERPOLACIÓN



INTRODUCCIÓN

Es común encontrar en libros, revistas, periódicos, etc., tablas como las siguientes:

Comprensión	Eficiencia volumétrica
2	87.3
2.2	86.0
2.4	84.9
2.6	83.5
2.8	82
3	80.8
3.2	79.5
3.4	78.3
3.6	77.2
3.8	76.0

Temperatura	Presión
0	2.555
4	2.852
5	2.931
8	3.179
12	3.534
16	3.923
20	4.342
24	4.801
28	5.294
32	5.830
36	6.411

Este tipo de tablas lleva implícita una relación funcional:

$$y = f(x)$$

De tal forma que $y_i = f(x_i)$

Algunas preguntas que surgen naturalmente son:

a) ¿Qué es $f(x)$?

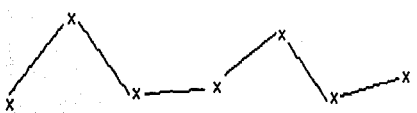
b) ¿Cuánto vale $\int_a^b f(x)\ dx$, para algún (a,b)?

c) ¿Cuánto vale $f(x)$ en puntos que no aparecen en la tabla?

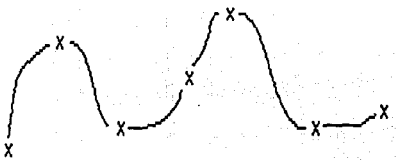
d) ¿Qué tan rápido crecen las y_i, en términos de las x_i?

e) ¿Cómo es la gráfica de $f(x)$?

Con respecto a esta última cuestión, la forma más natural de graficar f es a través de rectas por pedazos:



Otra forma, menos sencilla pero quizás más apegada a la realidad, consiste en trazar una curva que sea lo más "suave" posible y que pase por los puntos (x_i, y_i).



En general, un primer acercamiento consiste en determinar cuál es la familia más pequeña a la que pertenece f. ¿Qué sentido tiene lo anterior? Podemos decir que la familia más grande es el conjunto de funciones reales de variable real. Pero es obvio que esto no nos dice nada, por lo que es deseable reducir el tamaño de la familia. De hecho si llamamos F a tal familia, siempre esperaremos poder determinar a F a través de un número pequeño de parámetros.

Ejemplos:

1) $F = \{f: R \to R \mid f(x) = mx + b,\ m,b \in R\}$ luego F depende de dos parámetros.

2) $F = \{f: R \rightarrow R \mid f(x) = mx + b, b \in R, m > 0\}$ Esta familia es más pequeña que la del ejemplo (1)

Dado que, sin ningún antecedente, la elección de f resulta arbitraria, trataremos de buscar a f como un elemento de cierta familia F de funciones, que depende de ciertos parámetros. Podemos distinguir al menos 4 casos diferentes:

1) $F \subset \{f \mid f(x_i) = y_i\}$
Esta propiedad es indispensable cuando la información dada en la tabla no está sujeta a error, o bien éste es despreciable.

2) $F \subset \{f \mid f(x_i) \approx y_i\}$
Se presupone que los datos llevan errores considerables, entonces se busca un "ajuste", tal que $f(x_i) \approx y_i$

3) Una combinación de (1) y (2)
$F \subset \{f \mid f(x_i) = y_i$ para algunas $i\}$
Cuando, por alguna razón, se duda de algunos datos y de otros no.

4) Dada una tabla, hay que construir una tabla nueva a partir de la anterior, de tal manera que se quiten ciertos "defectos".

El problema (1) es conocido como "*interpolación*"
El problema (2) es conocido como "*ajuste*"
El problema (3) es conocido como "*interpolación-ajuste*"
El problema (4) es conocido como "*alisamiento*"

INTERPOLACION

En este capítulo desarrollaremos solamente el problema 1, esto es, el de interpolación, que consiste en determinar una función f, que pase por los puntos $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$, esto es:

$$f(x_i) = y_i \qquad \forall_i = 1, 2, \ldots n$$

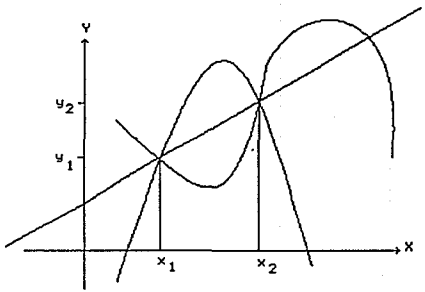
Supongamos que está dada la familia F, a la que pertenece nuestra función f de interpolación, en términos de n parámetros:

$$f \in F = \{g \mid y_i = F(x, a_1, a_2, \ldots, a_n) \qquad g(x_i) = y_i\}$$

En donde cada elemento de F es de la forma:

$$g = \phi(x, a_1, a_2, \ldots, a_n)$$

Ejemplo: Dar un polinomio, de grado menor o igual que 3, que pase por los puntos (x_1, y_1) (x_2, y_2)



Como ilustra la figura anterior, la solución no es única. Analíticamente definimos a la familia F, de polinomios cúbicos g, que satisfacen:

$$g(x_i) = y_i, \qquad i = 1, 2$$

Mediante la propiedad:

$$a_0 + a_1 x_1 + a_2 x_1^2 + a_3 x_1^3 = y_1$$

$$a_0 + a_1 x_2 + a_2 x_2^2 + a_3 x_2^2 = y_2$$

Lo que nos lleva (por eliminación de dos parámetros) a expresar cada miembro de la familia como:

$$g(x) = \phi\ (x, a_2, a_3)$$

Lo que indica que necesitamos imponer condiciones adicionales para determinar g.

A continuación, damos algunas familias de funciones que pueden usarse para interpolar un conjunto de datos.

i) $p(x) = a_0 + a_1 x + \ldots + a_n x^n$ *polinomiales*

ii) $T(x) = a_0 + a_1 \cos x + a_2 \cos 2x + \ldots + a_r \cos rx + b_1 \operatorname{sen} x + \ldots + b_r \operatorname{sen} rx$ *polinomiales trigonométricas*

iii) $R(x) = \dfrac{a_0 + a_1 x + \ldots + a_n x^n}{b_0 + b_1 x + \ldots + b_k x^k}$ *racionales*

iv) $E(x) = a_1 + exp(a_2(x - a_3))$ *exponenciales*

v) $G(x) = a_0 exp(-a_1(x - a_2)^2)$ *gausianas*

La elección de la familia depende del origen de los datos y ésta tiene que hacerse por la persona que quiere hacer la interpolación.

Para que el problema de interpolación sea soluble se requiere que exista al menos una colección $a_1, a_2, \ldots, a_r$ que satisfaga la propiedad:

$$g(x_i, a_1, a_2, \ldots, a_r) = y_i \quad , \quad i = 1, 2, \ldots, m$$

Si g depende de r parámetros $a_1, a_2, \ldots, a_r$ y tenemos m parejas $\{(x_i, y_i)\}$ de datos; entonces, para que el sistema anterior tenga cuando menos una solución, es necesario que:

num. de parámetros $\geq$ *número de puntos*

Es relevante el hecho de que el problema de interpolación siempre genera un problema de solución de ecuaciones:

$$g(x_i) = y_i \quad , \quad i = 1, 2, \ldots, m$$

Que admite, obviamente, una clasificación, dependiendo de la forma en que aparecen los parámetros en $g(x)$: *lineal o no lineal:*

A) *Caso lineal: si* $g(x) = a_1 g_1(x) + a_2 g_2(x) + \ldots + a_n g_n(x)$ donde cada $g_i(x)$ no depende de parámetros desconocidos, entonces:

$$g(x_i) = y_i \qquad i = 1, 2, \ldots, m$$

Nos queda en la forma:

$$a_1 g_1(x_i) + a_2 g_2(x_i) + \ldots + a_n g_n(x_i) = y_i$$

$$\begin{pmatrix} g_1(x_1) & g_2(x_1) & \cdots & g_n(x_1) \\ g_1(x_2) & g_2(x_2) & \cdots & g_n(x_2) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ g_1(x_m) & g_2(x_m) & \cdots & g_n(x_m) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{pmatrix}$$

Es decir:

$$\mathbf{A}\,\mathbf{a} = \mathbf{y}$$

Donde: *grado de* $\mathbf{A} = m \times n$

$$\mathbf{A} = (a_{ij}) \qquad a_{ij} = g_j(x_i)$$
$$\mathbf{a} = (a_1, a_2, \ldots, a_n)$$
$$\mathbf{y} = (y_1, y_2, \ldots, y_m)$$

Ejemplo:

$$g(x) = a_0 + a_1x + \ldots + a_nx^n$$

$$\Rightarrow g_j(x) = x^{j-1},\ j = 1, 2, \ldots, n+1$$

$$\Rightarrow A = \begin{pmatrix} g_1(x_1) & g_2(x_1) & \ldots & g_{n+1}(x_1) \\ g_1(x_2) & g_2(x_2) & \ldots & g_{n+1}(x_2) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ g_1(x_m) & g_2(x_m) & \ldots & g_{n+1}(x_m) \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x_1 & \ldots & x_1^{\,n} \\ 1 & x_2 & \ldots & x_2^{\,n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & x_m & \ldots & x_m^{\,n} \end{pmatrix}$$

B) *Caso no lineal.* Si la función g no depende linealmente de los parámetros, entonces el sistema:

$$g(x_i) = y_i \qquad i = 1, 2, \ldots, m$$

Es un sistema de ecuaciones no lineales, que en general es más complicado de resolver:

Ejemplo:

$$g(x) = a_1 + a_2\,exp(a_3x)$$

$$g(x_1) = y_1 = a_1 + a_2\,exp(a_3x_1)$$

$$g(x_2) = y_2 = a_1 + a_2\,exp(a_3x_2)$$

$$g(x_3) = y_3 = a_1 + a_2\,exp(a_3x_3)$$

Interpolación polinomial

Como hemos visto, la función de interpolación *f* puede pertenecer a diversas familias de funciones: polinomiales, trigonométricas, racionales, exponenciales, etc.

La familia más común, por razones obvias, es la de polinomios. En este caso, hablamos de interpolación polinomial y el problema se reduce a encontrar un polinomio p de grado r, tal que:

$$p(x_i) = y_i \qquad \forall_i = 1, \ldots n\text{-}1$$

Por lo que debemos determinar los $n+1$ parámetros $a_0, a_1, \ldots, a_n$ de tal manera que el polinomio $p(x) = a_0 + a_1x + \ldots + a_nx^n$ interpole a los puntos (x_i, y_i), $i = 1, 2, \ldots, m$

Si $n+1 \geq m$ y si $x_i \neq x_j$ $(i \neq j)$ entonces existe, cuando menos, un polinomio con la propiedad:

$$p(x_i) = y_i, \qquad i = 1, 2, ..., m$$

El sistema de ecuaciones a que da lugar este problema es:

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & x_m & x_m^2 & \dots & x_m^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ . \\ . \\ . \\ a_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_m \end{pmatrix}$$

Observamos que si una abscisa se repite, la matriz tendría dos renglones repetidos. Supongamos que $x_l = x_k$, entonces el sistema sería inconsistente, a menos de que $y_l = y_k$

Si $x_i \neq x_j$ $\forall\, i \neq j$ los renglones de la matriz son linealmente independientes, y dado que el número de renglones m, es menor o igual que el número de columnas $n+1$, la matriz es de rango máximo m, y por lo tanto, siempre tiene solución.

En el caso $n+1 = m$ tenemos un sistema cuadrado con matriz no singular, luego este problema tiene solución única, es decir, *si el número de parámetros (grado del polinomio + 1) es igual al número de puntos, entonces existe un único polinomio que interpola a dichos puntos, siempre que las abscisas de los puntos no se repitan.*

$$\begin{pmatrix} 1 & x_1 & \dots & x_1^{m-1} \\ 1 & x_2 & \dots & x_2^{m-1} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 1 & x_m & \dots & x_m^{m-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ . \\ . \\ . \\ a_{m-1} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_m \end{pmatrix}$$

El determinante del sistema es el determinante de Vandermode, y esta dado por:

$$det\,(V) = \prod_{1 \leq i \leq j \leq m} (x_i - x_j) \neq 0 \text{ si } (x_i \neq x_j\ ,\ i \neq j)$$

Por lo tanto, existe un único polinomio **de grado menor o igual que m-1** *que interpola a los m puntos* (x_i, y_i) *siempre que* $x_i \neq x_j$ *para* $i \neq j$.

Vale la pena recalcar el hecho de que el grado del polinomio de interpolación no es necesariamente *m-1*, aunque esto resulta obvio del sistema de ecuaciones, que bien podría arrojar como solución $a_{m-1} = 0$. Un caso extremo sería aquel en que $y_1 = y_2 = ... = y_m = 0$, para

el cual la solución sería $a_0 = a_1 = ... = a_{m-1} = 0$, es decir, tendríamos un polinomio de interpolación de grado 0:

$$p(x) \equiv 0$$

Un caso frecuente es cuando $m=3$, y tenemos el siguiente sistema:

$$(1) \qquad \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$$(2) \qquad p(x) = a_0 + a_1 x + a_2 x^2$$

¿Es necesario resolver el sistema (1)?

La respuesta a esta pregunta es: sí, siempre y cuando queramos obtener el polinomio expresado en la forma (2)

Bases para polinomios

Recordemos que cada polinomio de grado ≤ 2 es un elemento de un espacio vectorial de dimensión 3, el cual no tiene una base única.

Sea $S_2 = \{p(x) \mid p(x) =$ polinomio de grado $\leq 2\}$, sabemos que si $B = \{p_1(x), p_2(x), p_3(x) \mid p_i \in S_2$ y p_i son linealmente independientes$\}$ entonces B *es una base de* S_2. Luego, podemos expresar todo polinomio de grado ≤ 2 en la forma:

$$(3) \qquad p(x) = c_1 p_1(x) + c_2 p_2(x) + c_3 p_3(x)$$

Observemos que, en particular, la forma (2) queda incluída en este contexto, pues $\{1, x, x^2\}$ es una base de S_2. De los distintos tipos de base que nos sean accesibles, se desprenderán diferentes métodos numéricos para resolver el problema de interpolación polinomial.

1) $\{1, x, x^2, ..., x^{m-1}\}$

Esta base puede causar problemas numéricos cuando $m \geq 4$

2) $\{1, (x-a), (x-a)^2, ..., (x-a)^{m-1}\}$

Resuelve parcialmente la dificultad planteada en (1) si se toma, por ejemplo $a = \sum_{i=1}^{m} x_i/m$, pero tiene la deficiencia anterior también.

3) $\{1, \left(\frac{x-a}{b}\right), \left(\frac{x-a}{b}\right)^2, ..., \left(\frac{x-a}{b}\right)^{m-1}\}$

4)$\{1, (x - x_1), (x - x_1)(x - x_2), ..., (x - x_1)(x - x_2)...(x - x_m)\}$

5) $\{L_1(x), L_2(x), ..., L_m(x)\}$

donde:

$$L_k(x) = \frac{(x - x_1) ... (x - x_{k-1}) (x - x_{k+1}) ... (x - x_m)}{(x_k - x_1) ... (x_k - x_{k-1}) (x_k - x_{k+1}) ... (x_k - x_m)}$$

$k = 1, 2, ..., m$

Observación: La elección de una base nos indica la *forma* en que estamos representando el polinomio de interpolación, que es "único".

Por ejemplo, al polinomio $p(x) = 3 + 2x + x^2$ en el que empleamos la base (1) para su representación, lo podemos representar de manera diferente, usando otra base; dicho polinomio tiene la representación $p(x) = -32 + 12(x - 5) + (x - 5)^2$ empleando la base (2)

Métodos de interpolación

Al abordar el problema de interpolación polinomial usando las bases **1, 2 ó 3**, obtenemos un sistema de ecuaciones:

$$\mathbf{A\,c = y}$$

Que podría ser resuelto mediante eliminación gaussiana *con pivoteo parcial.*
Los casos interesantes, los generan las bases **5 y 4** que dan lugar a las fórmulas de Lagrange y de Newton para interpolación polinomial.

Fórmula de Lagrange

La representación del polinomio de interpolación en la "*formula de Lagrange*", tiene su origen en el hecho conocido de que los ceros de un polinomio determinan de manera única, excepto por un factor constante, a dicho polinomio, cuya obtención es fácil.

Por ejemplo, si queremos determinar el polinomio de *grado* $\leq n-1$ tal que $p(x_i) = 0$. $i = 1, 2, ..., n-1$, es decir, que tiene como raíces a las x_i, hacemos a:

$$p(x) = A(x - x_1)(x - x_2) ... (x - x_{n-1})$$

Para determinar la constante A necesitamos una condición adicional, ésta puede ser el valor de $p(x)$ en otro punto; supongamos que queremos que:

$$p(x_n) = 1$$

Entonces:

$$p(x_n) = 1 = A(x_n - x_1)(x_n - x_2) \dots (x_n - x_{n-1})$$

$$\Rightarrow A = \frac{1}{(x_n - x_1)(x_n - x_2) \dots (x_n - x_{n-1})}$$

Por lo tanto:

$$p(x) = \frac{(x - x_1)(x - x_2) \dots (x - x_{n-1})}{(x_n - x_1)(x_n - x_2) \dots (x_n - x_{n-1})}$$

Este es el origen de los polinomios $L_k(x)$, donde:

$$L_k(x) = \frac{(x - x_1)(x - x_2) \dots (x - x_{k-1})(x - x_{k+1})}{(x_k - x_1)(x_k - x_2) \dots (x_k - x_{k-1})(x_k - x_{k+1})} \frac{\dots (x - x_m)}{\dots (x_k - x_m)}$$

Que tienen las siguientes propiedades:

i) $L_k(x_i) = 0$ si $i \neq k$

ii) $L_k(x_k) = 1$

iii) $L_k(x)$ tiene *m-1* ceros reales distintos, y dado que se expresa como una constante por el producto de las diferencias $(x - x_i)$, para $i = 1, 2, \dots, m$ $i \neq k$, resulta ser $L_k(x)$ un polinomio de grado *m-1* para toda $k = 1, 2, \dots, m$.

iv) El conjunto $\{L_k(x) \mid k = 1, 2, \dots, m\}$ es linealmente independiente; por lo tanto, es base del espacio vectorial de los polinomios de grado menor o igual que *m-1*..

De las observaciones anteriores, concluimos que hay un único polinomio que interpola a un conjunto de datos (x_i, y_i) y está dado por:

$$P(x) = \sum_{k=1}^{m} y_k L_k(x)$$

La expresión desarrollada más común es:

$$p(x) = y_1 L_1(x) + y_2 L_2(x) + \dots + y_m L_m(x)$$

En donde:

$$L_k(x) = \prod_{1=k}^{m+1} \frac{x - x_i}{x_k - x_i} \quad k = 1, 2, \dots m$$

Ejemplo: Determina el polinomio de grado 2 que interpola a los puntos (0, -1), (1, -1), (2, 7).

$$L_1(x) = \frac{(x-1)(x-2)}{(0-1)(0-2)} = \frac{(x-1)(x-2)}{2}$$

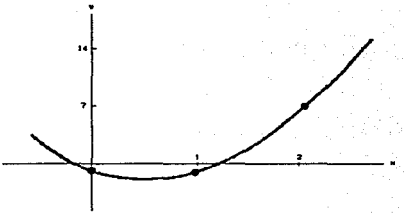
$$L_2(x) = \frac{(x-0)(x-2)}{(1-0)(1-2)} = \frac{x(x-2)}{-1}$$

$$L_3(x) = \frac{(x-0)(x-1)}{(2-0)(2-1)} = \frac{x(x-1)}{2}$$

Por lo tanto, el polinomio es:

$$p(x) = (-1)\frac{(x-1)(x-2)}{2} + (-1)\frac{x(x-2)}{-1} + (7)\frac{x(x-1)}{2}$$

Cuya gráfica es la siguiente:



Obviamente, este procedimiento se puede automatizar para calcular los términos $L_k(x)$

A continuación, se presenta un programa para tal fin.

Programa

```
100 DIM FX(10,10),X(10)
110 INPUT N
120 FOR I = 1 TO N
130 INPUT X(I),FX(I,1)
140 NEXT I
150 FOR J = 1 TO N - 1
160 K = J + 1
170 FOR I = 1 TO N - J
180 FX(I,K) = (FX(I + 1, J) - FX(I,J)) / (X(I + J) - X(I))
```

```
190 NEXT I
200 NEXT J
210 FOR J = 1 TO N
220 PRINT FX(1,J)
230 NEXT J
240 INPUT XI
250 FA = 1
260 Y = 0
270 FOR J = 1 TO N
280 Y = Y + FX(1,J) * FA
290 PRINT Y
300 FA = FA * (XI - X(J))
310 IF IF J > = N THEN 350
320 EA = FA * FX(1,J + 1)
330 PRINT EA
340 NEXT J
350 END
```

Fórmula de Newton

La idea detrás de esta formulación es la construcción de manera gradual, del polinomio de interpolación.

Supóngase que el polinomio $p_k(x)$ interpola a los datos $p\{(x_i, y_i)\}$ i = 1, 2, ..., k y queremos un polinomio $p_{k+1}(x)$ que interpole los mismos datos más uno (x_{k+1}, y_{k+1}); queremos aprovechar el polinomio $p_k(x)$ para obtener de manera fácil $p_{k+1}(x)$ a partir de $p_k(x)$ mediante una "*simple corrección*", es decir, queremos que:

$$p_{k+1}(x) = p_k(x) + c_{k+1}(x)$$

Donde $c_{k+1}(x)$ es un polinomio que vamos a determinar.

Como:

$$p(x_i) = y_i \qquad i = 1, 2, ..., n$$

Entonces:

$$p_{k+1}(x_i) = p_k(x_i) + c_{k+1}(x_i)$$
$$= y_i + c_{k+1}(x_i)$$
$$i = 1, 2, ..., k$$

Entonces, para que:

$$p_{k+1}(x_i) = y_i \qquad i = 1, 2, ..., k$$

Es necesario que: $c_{k+1}(x_i) = 0 \qquad i = 1, 2, ..., k$

Entonces $c_{k+1}(x)$ es de la forma:

$$c_{k+1}(x) = A(x - x_1)(x - x_2) \ldots (x - x_k)$$

Donde A es una constante por determinar. Ahora bien, como queremos que:

$$p_{k+1}(x_{k+1}) = y_{k+1}$$

Esto implica:

$$y_{k+1} = p_k(x_{k+1}) + A(x_{k+1} - x_1) \ldots (x_{k+1} - x_k)$$

Despejando:

$$A = \frac{y_{k+1} - p_k(x_{k+1})}{(x_{k+1} - x_1) \ldots (x_{k+1} - x_k)}$$

Y entonces obtenemos la solución a nuestro problema:

$$p_{k+1}(x) = p_k(x) + \frac{(y_{k+1} - p_k(x_{k+1}))}{(x_{k+1} - x_1) \ldots (x_{k+1} - x_k)} (x - x_1)(x - x_2) \ldots (x - x_k)$$

Esta es la idea detrás de la base $\{1, (x - x_0), (x - x_0)(x - x_1), \ldots, (x - x_0)(x - x_1) \ldots (x - x_{n-1})\}$; lo interesante en este caso es que se puedan obtener fórmulas bonitas y fáciles de calcular mediante un esquema de recurrencia; ya que:

$$\text{Si } p(x) = a_0 \cdot 1 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots + a_n(x - x_0) \ldots (x - x_{n-1})$$

$$\Rightarrow p(x_0) = a_0 \equiv y_0$$

$$p(x_1) = y_0 + a_1(x_1 - x_0) \equiv y_1 \Rightarrow a_1 = \frac{y_1 - y_0}{x_1 - x_0}$$

$$p(x_2) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) \equiv y_2$$

Entonces:

$$a_2 = \frac{(y_2 - y_0) - \dfrac{y_1 - y_0}{x_1 - x_0}(x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)}$$

$$= \frac{y_2 - y_0}{(x_2 - x_0)(x_2 - x_1)} - \frac{y_1 - y_0}{(x_2 - x_1)(x_1 - x_0)}$$

$$= \frac{\dfrac{y_2 - y_0}{x_2 - x_0} - \dfrac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_1}$$

Por lo tanto, si usamos la base de Newton:

$$\{1, (x-x_0), (x-x_0)(x-x_1), ..., (x-x_0)(x-x_1)...(x-x_{n-1})\}$$

El polinomio de interpolación será:

$$p(x) = a_0 \cdot 1 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + ... + a_n(x-x_0)(x-x_1)...(x-x_{n-1})$$

En donde, si hacemos $y_i = f_i$, tenemos:

$$a_0 = f_0$$

$$a_1 = \frac{f_1 - f_0}{x_1 - x_0}$$

$$a_2 = \frac{\frac{f_2 - f_0}{x_2 - x_0} - \frac{f_1 - f_0}{x_1 - x_0}}{x_2 - x_1}$$

Diferencias divididas

Consideremos ahora los puntos ordenados, como en la siguiente tabla:

x	y
x_1	f_1
x_0	f_0
x_2	f_2

$$p(x) = A_0 \cdot 1 + A_1(x-x_1) + A_2(x-x_1)(x-x_0)$$

$$\Rightarrow A_0 = f_1$$

$$A_1 = \frac{f_0 - f_1}{x_0 - x_1}$$

$$A_2 = \frac{\frac{f_2 - f_1}{x_2 - x_1} - \frac{f_0 - f_1}{x_0 - x_1}}{x_2 - x_0}$$

Dado que el polinomio de grado 2 que interpola a tres puntos es único, resulta:

$$a_2 = A_2$$

Pues, en ambos casos, es el coeficiente del término de segundo grado.

Inspeccionando esas expresiones, observamos que:

$$a_2 = \frac{(x_1 - x_0)(f_2 - f_0) - (x_2 - x_0)(f_1 - f_0)}{(x_2 - x_1)(x_2 - x_0)(x_1 - x_0)}$$

$$a_2 = \frac{(x_2 - x_0 - x_1 + x_0) f_0 + (x_0 - x_2) f_1 + (x_1 - x_0) f_2}{(x_2 - x_1)(x_2 - x_0)(x_1 - x_0)}$$

$$= \frac{(x_2 - x_1) f_0 - (x_2 - x_0) f_1 + (x_1 - x_0) f_2}{(x_2 - x_1)(x_2 - x_0)(x_1 - x_0)}$$

Simplificando, obtenemos:

$$a_2 = \frac{f_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{f_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{f_2}{(x_2 - x_0)(x_2 - x_1)}$$

La simetría de la expresión para a_2 proviene del hecho de que el polinomio de interpolación es único, y no depende del orden en que se escriban los puntos. La expresión anterior es un caso particular de una *"diferencia dividida"*.

$$f_1[x_0\ x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = a_1$$

Observamos que $f_1 : R^2 \to R$ es una función simétrica con respecto a la diagonal $y = x$, en donde toma el valor de la derivada de f en x: $f_1[x, x] = f'(x)$. Llamamos a esta cantidad *"primera diferencia dividida"*. Construímos la *"segunda diferencia dividida"* a partir de la primera.

$$f_2[x_0\ x_1\ x_2] = \frac{f_1[x_0\ x_1] - f_1[x_1\ x_2]}{x_0 - x_2} = a_2$$

Por la simetría vista anteriormente, tenemos que:

$$f_2[x_0\ x_1\ x_2] = f_2[x_0\ x_2\ x_1] = \ldots$$

En general:

$$f_n[x_0\ x_1\ \ldots\ x_n] = \frac{f_{n-1}[x_0 \ldots x_{n-1}] - f_{n-1}[x_1 \ldots x_n]}{x_0 - x_n}$$

Observamos que las diferencias divididas se pueden escribir en forma simétrica, de donde vemos que ninguna diferencia cambia al hacer permutaciones en los elementos:

$$f[x_0 x_1] = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}$$

$$f[x_0 x_1 x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

En general:

$$f[x_0 x_1 \dots x_n] = \frac{f(x_0)}{\prod_{\substack{j=0 \\ j=0}} (x_0 - x_j)} + \dots + \frac{f(x_n)}{\prod_{\substack{j=0 \\ j=n}} (x_n - x_j)} = \sum_{i=0}^{n} \frac{f(x_i)}{\prod_{\substack{j=0 \\ j=i}} (x_i - x_j)}$$

Volviendo al problema del polinomio de interpolación, tenemos que, para la base de Newton:

$$1, (x - x_0), (x - x_0)(x - x_1), \dots, (x - x_0)(x - x_1), \dots, (x - x_{n-1})$$

El polinomio de interpolación queda dado por:

$$p(x) = a_0 \cdot 1 + a_1(x - x_0) + \dots + a_n(x - x_0)(x - x_1) \dots (x - x_{n-1})$$

Imponiendo las condiciones $p(x_i) = f_i$, $i = 0, \dots, n$ obtenemos:

$$a_0 = f(x_0) = f[x_0]$$

$$a_1 = \frac{f(x_0) - f(x_1)}{x_0 - x_1} = f[x_0 x_1]$$

$$a_2 = \frac{f[x_0\, x_1] - f[x_1\, x_2]}{x_0 - x_2} = f[x_0 x_1 x_2]$$

.

.

.

$$a_n = \frac{f[x_0\, x_1 \ldots x_{n-1}] - f[x_1\, x_2 \ldots x_n]}{x_0 - x_n} = f[x_0 x_1 \ldots x_n]$$

La simetría observada en las diferencias divididas tiene una gran ventaja en la práctica. Supongamos que tenemos una tabla con una gran cantidad de datos ordenados $x_0 < x_1 < x_2 < \ldots < x_n$ (n grande) y nuestro problema consiste en dar una estimación de $p(x^*)$, para alguna x^* entre x_k y x_{k+1}. Al calcular las diferencias divididas en el orden original puede ocurrir que se haga trabajo innecesario para resolver nuestro problema. Partiendo de la idea intuitiva de que el valor de $p(x^*)$ depende más fuertemente de los valores cercanos a x^*

Interpolación en tablas equiespaciadas

A continuación, consideremos el caso particular (bastante común) en que los puntos se encuentran igualmente espaciados, es decir:

$$x_k = x_0 + kh, \qquad k = 0, 1, \ldots, n$$

$$f_k = f(x_k) = f(x_0 + kh)$$

Entenderemos por $f_{-k} \equiv f(x_0 - kh)$. Las diferencias divididas toman ahora la forma:

$$f[x_k] = f(x_k) = f_k$$

$$f[x_0 x_1] = \frac{f[x_0] - f[x_0 + h]}{-h} = \frac{f(x_0 + h) - f(x_0)}{h}$$

$$f[x_0 x_1 x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

$$= \frac{f(x_0)}{2h^2} + \frac{f(x_0 + h)}{-h^2} + \frac{f(x_0 + 2h)}{2h^2}$$

$$= \frac{f(x_0) - 2f(x_0 + h) + f(x_0 + 2h)}{2h^2}$$

$$f[x_0 x_1 x_2] = \frac{f(x_0 + 2h) - 2f(x_0 + h) + f(x_0)}{2h^2}$$

$$= \frac{[f(x_0 + 2h) - f(x_0 + h)] - [f(x_0 + h) - f(x_0)]}{2h^2}$$

A partir de la última expresión convenimos en usar la siguiente notación, en términos de Δ

$$\Delta_h f(x) = f(x+h) - f(x)$$

Dado que h es fija, podemos simplificar la notación anterior suprimiendo el subíndice. Por lo tanto:

$$f[x_0 x_1 x_2] = \frac{\Delta f(x_0 + h) - \Delta f(x_0)}{2h^2}$$

Algunas propiedades del operador Δ

$$\Delta(f \pm g) \equiv \Delta f \pm \Delta g$$

$$\Delta(\Delta f) \equiv \Delta^2 f(x)$$

$$f[x_0 x_1 x_2] = \frac{\Delta^2 f(x_0)}{2h^2}$$

Esta última es una expresión muy práctica, como veremos adelante.

Si definimos los operadores E, I de la siguiente manera:

$$E(f(x)) = f(x + h)$$

$$I(f(x)) = f(x)$$

Tendremos:

$$\Delta = E - I$$

De manera que:

$$f[x_0 x_1 x_2] = \frac{\Delta^2 f(x_0)}{2h^2} = \frac{(E - I)^2 f(x_0)}{2h^2}$$

$$= \frac{E^2 f(x_0) - 2Ef(x_0) + If(x_0)}{2h^2} \qquad (3)$$

Podemos generalizar las diferencias, como se ve a continuación:

$$f[x_0 x_1 x_2 x_3] = \frac{f[x_0\ x_1\ x_2] - f[x_1\ x_2\ x_3]}{x_0 - x_3}$$

$$= \frac{\frac{\Delta^2 f(x_0)}{2h^2} - \frac{\Delta^2 f(x_1)}{2h^2}}{-3h}$$

$$= \frac{\Delta^2 (f(x_1) - f(x_0))}{(2 \times 3) h^3} = \frac{\Delta^3 f(x_0)}{3! h^3}$$

Por un argumento inductivo, llegamos a que:

$$f[x_0 x_1 \ldots x_k] = \frac{\Delta^k f(x_0)}{k! h^k} \equiv \frac{\Delta^k f_0}{k! h^k}$$

Recordemos ahora la expresión para el polinomio de interpolación, para reescribirla en términos de Δ:

$$p(x) = f(x_0) + f[x_0 x_1](x - x_0) + f[x_0 x_1 x_2](x - x_0)(x - x_1)$$
$$+ \ldots + f[x_0 x_1 \ldots x_n](x - x_0) \ldots (x - x_{n-1})$$

$$= f_0 + \frac{\Delta f_0}{h}(x - x_0) + \frac{\Delta^2 f_0}{2! h^2}(x - x_0)(x - x_1) + \ldots$$

$$+ \frac{\Delta^n f_0}{n! h^n}(x - x_0)(x - x_1) \ldots (x - x_{n-1})$$

Introducimos la notación siguiente:

$$s = (x - x_0)/h$$

De manera que:

$$x - x_1 = x - (x_0 + h) = \left(\frac{x - x_0}{h} h + h\right)$$

$$= \left(\frac{x - x_0}{h} + 1\right) h = (s - 1) h$$

Análogamente:

$$(x - x_2) = (s - 2)h$$

En general:

$$(x - x_i) = (s - i)h, \quad i = 0, 1, \ldots, n$$

Con la notación anterior, el polinomio queda:

$$p(x) = f_0 + \frac{\Delta f_0}{1!} s + \frac{\Delta^2 f_0}{2!} s(s-1) + \ldots +$$
$$+ \frac{\Delta^n f_0}{n!} s(s-1) \ldots (s-(n-1))$$

La cual se parece mucho a la fórmula de Taylor.

Ahora veamos cómo, a partir de la base nueva y el operador Δ, obtenemos una propiedad interesante:

$$\Delta 1 = 0$$

$$\Delta s = (s+1) - s = 1$$

$$\Delta (s(s\text{-}1)) = (s+1)s - s(s\text{-}1)$$

$$= 2s$$

$$\Delta (s\,(s\text{-}1)\,(s\text{-}2)) = (s+1)s\,(s\text{-}1) - s\,(s\text{-}1)\,(s\text{-}2)$$

$$= 3s\,(s\text{-}1)$$

En general:

$$\Delta (s\,(s\text{-}1) \ldots (s\text{-}k)) = (k+1)s\,(s\text{-}1) \ldots (s\text{-}(k\text{-}1))$$

Introduciendo la notación:

$$x^{[n]} = x(x-1)(x-2) \ldots (x-(n-1))$$

La última propiedad queda así:

$$\Delta x^{[n]} = n x^{[n-1]}$$

Y la expresión para el polinomio:

$$p(x) = f_0 + \frac{\Delta f_0}{1!} s + \ldots + \frac{\Delta^n f_0}{n!} s^{[n]}$$

La sigiente notación se usa mucho por la facilidad de recordarla:

$$\binom{s}{n} = \frac{s\,(s-1) \ldots (s-(n-1))}{n!} = \frac{s^{[n]}}{n!}$$

Luego:

$$\Delta \binom{s}{n} = \Delta \frac{s^{[n]}}{n!} = \frac{1}{n!} \Delta s^{[n]}$$

$$= \frac{m}{n!} s^{[n-1]}$$

$$= \frac{s^{[n-1]}}{(n-1)!}$$

$$= \binom{s}{n-1}$$

Así, el polinomio se expresa como sigue, en términos de los coeficientes binomiales:

$$p(x) = f_0 + \Delta f_0 \binom{s}{1} + \Delta^2 f_0 \binom{s}{2} + \ldots + \Delta^n f_0 \binom{s}{n} = \sum_{i=0}^{n} \Delta^i f_0 \binom{s}{i}$$

Una aplicación práctica de los conceptos anteriores es la construcción de una "tabla de diferencias", como la que se da a continuación:

x	f	Δf	$\Delta^2 f$ ...	$\Delta^n f$	
x_0	f_0				
		Δf_0			
x_1	f_1		$\Delta^2 f_0$		
		Δf_1	.		
x_2	f_2	.	.	$\Delta^n f_0$	→ *Coeficientes del polinomio de interpolación, centrado en* x_0
.	.	.	.		
.	.	.	$\Delta^2 f_{n-2}$		
.	.	Δf_{n-1}			
x_n	f_n				

Otra propiedad interesante, desde el punto de vista práctico, es:

$$\Delta^{n+1} P_n(x) = 0$$

En donde P_n es un polinomio de grado n. Es importante porque si se tienen datos para interpolar un polinomio de grado n, con $n \ll m$, la tabla de diferencias sólo tendrá valores significativos en las primeras n columnas, el resto será de ceros, y eso simplifica el trabajo.

Se tiene una aplicación importante cuando se desea calcular $p(x^*)$, para alguna x^* entre x_k y x_{k+1}, en cuyo caso se calcula la tabla de diferencias hasta obtener una columna constante y de ahí se determina el grado y el polinomio de interpolación.

Calcular $(2.3)^3$ *usando la tabla siguiente:*

	x	$f(x) = x^3$	Δ	Δ^2	Δ^3	
x_{-5}	-3	-27				
x_{-4}	-2	-8				
x_{-3}	-1	-1				
x_{-2}	0	0	1			
x_{-1}	1	1	7	3	1	
x_0	2	8	19	6	1	← coeficientes del polinomio
x^*						
x_1	3	27	37	9	1	
x_2	4	64	61	12		
x_3	5	125				

↑ polinomio de grado 3.

$$p(x) = 8 + 19\,(x-2) + 6\,(x-2)\,(x-3) + (x-1)\,(x-2)\,(x-3)$$

Algoritmo

$P_0 = 0$ $\qquad \pi_1 = 1$

$P_1 = P_0 + \delta\pi_1 = 8$ $\qquad \pi_2 = \pi_1\,(x-2) = 1\,(2.3-2) = .3$

$P_2 = P_1 + 19\pi_2 =$ $\qquad \pi_3 = .3(2.3-3)$

$= 8 + 19\,(.3) = 13.7$ $\qquad = .3\,(-.7) = -.21$

$P_3 = 13.7 + 6\,(-.21) = 12.44$ $\qquad \pi_4 = (-.21)\,(2.3-1) = -.273$

$P_4 = 12.44 + 1\,(.273) = 12.167$

Por lo tanto $(2.3)^3 = 12.167$

Interpolación por tramos: Splines

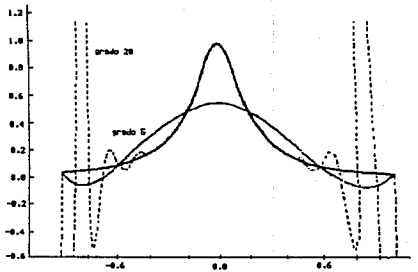
Limites de la interpolación polinomial

En general, desde el punto de vista práctico, no es conveniente usar polinomios de grado muy grande; además de que algunas veces la aproximación por medio de polinomios de interpolación no converge a la función.

Ejemplo: si la función $f(x) = \dfrac{1}{1 + 25x^2}$ $x \in [-1, 1]$

Se aproxima por polinomios de interpolación $p_n(x)$ tomando n puntos $\{(x_i, f_i(x_i)\}$ con las x_i uniformemente distribuidas en $[-1, 1]$, se ha demostrado que la sucesión de polinomios de interpolación $\{p_n(x)\}$ no converge a $f(x)$ más que en una parte del intervalo.

Es decir, esta función es adecuadamente aproximada por polinomios de interpolación en la parte central, pero la aproximación de dichos polinomios a la función es pésima en los extremos, como puede observarse en la siguiente gráfica.



Interpolación spline

Las limitaciones prácticas de la interpolación, sobre todo cuando se tienen muchos puntos $\{(x_i, y_i)\}$ i = 1, 2, ..., m; obligan a buscar alternativas. Una de las más efectivas es la conocida como aproximación spline; esta consiste en aproximar una función o una tabla por una familia de polinomios, es decir, se subdivide el intervalo de interés en intervalos pequeños y en cada uno de sus subintervalos un polinomio de la familia se encarga de aproximar a la función. Esto es, dada $f(x)$ en $[a, b]$, se construye una partición:

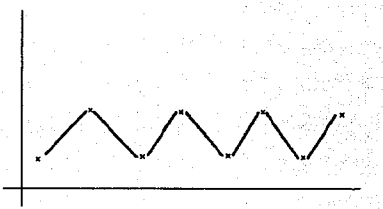
$$a = x_0 < x_1 < \ldots < x_n = b$$

Y una familia de polinomios $\{p_i(x)\}$ i = 1, 2, ..., n. La función aproximante *spline* $A(x)$ se define como:

$$A(x) = p_i(x) \qquad x \in [x_{i-1}, x_i]$$

Spline lineal

Interpolación lineal por pedazos



Dados $\{(x_i, y_i)\}$ i = 1, 2, ..., n; una forma sencilla de construir una función interpolante que aproxime esos datos es la que se obtiene al unir los puntos de la gráfica con rectas; así, si $l_i(x)$ es la función lineal que tiene la propiedad:

$$l_i(x_{i-1}) = y_{i-1}$$

$$l_i(x_i) = y_i$$

Entonces el "*spline lineal*" está dado por:

$$s(x) = l_i(x) \qquad x \in [x_{i-1}, x_i]$$

Una desventaja del *spline lineal* es que la función $s(x)$ "*tiene picos*"; uno de los descubrimientos interesantes es que es posible construir splines "*sin picos*" que interpolen una tabla.

Spline cúbico

Un spline cúbico es una función $s(x)$ definida por:

i) Una partición:

$$x_0 < x_1 < \ldots < x_n$$

ii) Una familia $\{p_i(x)\}$ de polinomios cúbicos tales que:

$$s(x) = p_i(x) \qquad x \in I_i = [x_{i-1}, x_i]$$

Quizás entre todas las funciones polinomiales por tramos, las más populares debido a sus propiedades son los polinomios cúbicos por tramos y en especial los splines cúbicos. En esta sección, vamos a estudiar diferentes tipos de funciones cúbicas por tramos, que nos permitirán resolver otra vez, el problema de construir una curva que pase por un conjunto de puntos. Supongamos que tenemos la siguiente tabla de valores:

x	x_1	x_2	x_3	x_4	...	x_n
g	g_1	g_2	g_3	g_4	...	g_n
g'	s_1	s_1	s_3	s_4	...	s_y

Esto es, tenemos los valores de una función y sus derivadas en los puntos $\{x_i\}$, i = 1, ..., n. Vamos a ver a continuación que es fácil construir un polinomio cúbico por tramos que coincida con g en sus valores y en sus derivadas en los mismos puntos.

Antes de resolver el problema anterior, abordaremos un caso especial, a partir del cual la expresión de la solución del problema general sera inmediata.

Para construir un polinomio cúbico $p(y)$ tal que:

$$p(0) = p_0, \qquad p(1) = p_1$$
$$p'(0) = p'_0, \qquad p'(1) = p'_1$$

Expresamos el polinomio $p(y)$ en la forma:

$$p(y) = p_0 c_1(y) + p_1 c_2(y) + p'_0 c_3(y) + p'_1 c_4(y)$$

Donde:

$$c_i(y); \qquad i = 1, ..., 4$$

Son polinomios cúbicos. Observemos que si:

$$p_1 = p'_0 = p'_1 = 0 \quad y \quad p_0 = 1,$$

Entonces:

$p(y) = c_1(y)$, es decir $c(y)$ debe satisfacer:

$$c_1(0) = 1; \qquad c'_1(0) = 0;$$

$$c_1(1) = 0; \qquad c'_1(1) = 1;$$

De estas condiciones es muy fácil obtener $c_1(y)$ puesto que, como $y=1$ es una raíz doble, tenemos que debe ser de la forma:

$$c_1(y) = (a + by)(y - 1)^2$$

Donde a, b son coeficientes que quedan determinados por las condiciones $c_1(0) = 1$ y $c'_1(0) = 0$. Por lo tanto:

$$a = 1; \quad b = 2,$$

$$c_1(y) = (1 + 2y)(y - 1)^2$$

Procediendo de manera similar, obtenemos:

$$c_2(y) = y_2(3 - 2y),$$

$$c_3(y) = (y - 1)2y,$$

$$c_4(y) = y^2(y - 1),$$

Entonces, el polinomio $p(y)$ queda como:

$$p(y) = p_0(1 + 2y)(y - 1)^2 + p_1y^2(3 - 2y) + p'_0(y - 1)^2y + p'_1y^2(y - 1);$$

Ahora, es fácil obtener el polinomio $p_i(x)$ tal que:

$$p_i(x_i) = g_i \qquad p_i(x_{i+1}) = g_{i+1}$$

$$p'_i(x_i) = s_i \qquad p'_i(x_{i+1}) = s_{i+1}$$

Consideremos el cambio de variable:

$$y = \frac{y - x_i}{\Delta x_i} \qquad \Delta x_i = x_{i+1} - x_i$$

Que transforma el intervalo $[x_i, x_{i+1}]$ en el intervalo $[0, 1]$; la idea es que podemos obtener la expresión para $c_i(x)$ a partir de $p(y)$, sustituyendo y, pero esto es posible

si los términos que contienen p'_0 y p'_1 los multiplicamos por Δx_i ya que:

$$p'_i(x) = \frac{d}{dx} p(y(x)) = p'(y)y' = p'(y)\frac{1}{\Delta x_i}$$

$$p'_i(x_i) = p'(0)\frac{1}{\Delta x_i} = s_i \qquad \Rightarrow p'_0 = p'(0) = s_i \Delta x_i$$

$$p'_i(x_{i+1}) = p'(1)\frac{1}{\Delta x_i} = s_{i+1} \Rightarrow p_{'1} = p'(1) = s_{i+1}\Delta x_i$$

$$p_i(x_i) = p(0) = g_i \qquad \Rightarrow p_0 = g_i$$

$$p_i(x_{i+1}) = p(1) = g_{i+1} \qquad \Rightarrow p_1 = g_{i+1}$$

Por lo tanto:

$$p_i(x) = g_i c_i(y) + g_{i+1}c_2(y)$$

$$+ s_i \Delta x_i c_3(y) + s_{i+1}\Delta x_i c_4(y) \qquad y = \frac{x - x_i}{\Delta x_i}$$

Sustituyendo y, obtenemos:

$$p_i(x) = g_i \frac{(x - x_{i+1})^2 [2(x - x_i) + \Delta x_i]}{\Delta x_i^2}$$

$$+ g_{i+1} \frac{(x - x_i)^2 [2(x_{i+1} - x) + \Delta x_i]}{\Delta x_i^3}$$

$$+ s_i \frac{(x_{i+1} - x)^2 (x - x_i)}{\Delta x_i^2}$$

$$- s_{i+1} \frac{(x - x_i)^2 (x_{i+1} - x)}{\Delta x_i^2} \qquad (0)$$

$$x_i \leq x \leq x_{i+1} \qquad i = 1, \ldots, n-1$$

Así, hemos obtenido finalmente una familia de polinomios $\{p_i(x)\}$, cada uno definido en el intervalo $[x_i, x_{i+1}]$, por lo que la función definida por:

$$f(x) = p_i(x), \qquad \text{para } x_i \leq x \leq x_{i+1}, \qquad i = 1, \ldots, n-1$$

Es una función continuamente diferenciable, polinomial por tramos. Esta función es conocida con el nombre de *función interpolante cúbica de Hermite*. Se puede demostrar que f aproxima muy bien a la función g y mejora a medida que las longitudes de los intervalos $[x_i, x_{i+1}]$ se hacen más pequeñas.

Ejemplos:

1) Supongamos que $g(x) = 1/(1 + x^2)$ y queremos calcular el interpolante cúbico de Hermite que coincide con g en los puntos -3, -1, 0, 1 y 3. Como:

$$g'(x) = \frac{-2x}{(1+x^2)^2}$$

Entonces, los datos que necesitamos para construir el interpolante f son los siguientes:

x	-3	-1	0	1	3
g	0.1	0.5	1	0.5	0.1
g'	0.06	0.5	0	-0.5	-0.5

Tabla 1. Valores de la función $g(x) = 1/(1 + x^2)$ y de su primera derivada.

Teniendo en cuenta que $s_i = g'(x_i)$, i = 1, ..., 5 y utilizando (0) obtenemos, entonces, que:

$$p_1(x) = s_1 \frac{(x_2 - x)^2(x - x_1)}{(\Delta x_1)^2} - s_2 \frac{(x - x_1)^2(x_2 - x)}{(\Delta x_1)^3}$$

$$+ g(x_1)\frac{(x_2 - x)^2\,[2(x - x_1) + \Delta x_1]}{(\Delta x_1)^3}$$

$$+ g(x_2)\frac{(x - x_1)^2(2(x_2 - x) + \Delta x_1)}{(\Delta x_1)^3} \qquad (1)$$

Sustituyendo cada término en (1) por su valor y simplificando llegamos a que:

$$p_1(x) = \frac{(x + 3)^3}{8} + \frac{(-1 - x)^2}{4}(0.58 + 0.16x)$$

De manera similar se obtienen las siguientes expresiones para polinomios p_2, p_3 y p_4:

$$p_2(x) = -0.5x^3 - x^2 + 1$$

$$p_3(x) = 0.5x^3 - x^2 + 1$$

$$p_4(x) = \frac{(3 - x)^2}{8} + \frac{(x - 1)^2}{4}(0.58 + 0.16x)$$

Por lo tanto, el interpolante cúbico de Hermite para la función $g(x) = 1/(1 + x^2)$ está dado por:

$$f(x) = \begin{cases} \dfrac{(x+3)^3}{8} + \dfrac{(-1-x)^2}{4}(0.58 + 0.16x) & -3 \le x \le -1 \\ -0.5x^3 - x^2 + 1 & -1 \le x \le 0 \\ 0.5x^3 - x^2 + 1 & 0 \le x \le 1 \\ \dfrac{(3-x)^2}{8} + \dfrac{(x-1)^2}{4}(0.58 + 0.16x) & 1 \le x \le 3 \end{cases}$$

El gráfico de la función f aparece en la figura *a*:

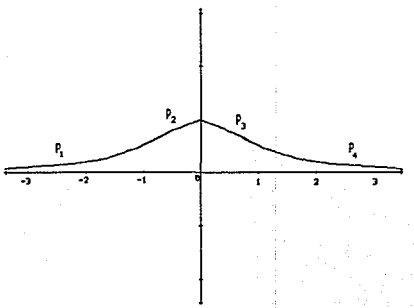


Fig. a) Interpolante cúbico de Hermite para $g(x) = 1/(1+x^2)$ x-puntos de interpolación.

2) Consideremos ahora la función:

$$g(x) = x^0_+ = max(x^0, 0) = \begin{cases} 1 & \text{si} \quad x \ge 0 \\ 0 & \text{si} \quad x < 0 \end{cases}$$

Y construyamos el interpolante cúbico de Hermite que coincide con g en los puntos -2, -1, -0 y 2. Como $g(x)$ no tiene primera derivada en 0, vamos a experimentar con diferentes valores s_3 para ver cómo influye esto en la forma del interpolante. Organicemos nuevamente nuestros datos en una tabla.

x	-2	-1	0	2
g	0	0	1	0
g'	0	0	s_3	0

Tabla 2. Valores de la función $g = x^0_+$ y de su primera derivada.

De la expresión (0) podemos concluir que, independientemente del valor de s_3, el polinomio p_1 que constituye el interpolate f es siempre el mismo y coincide con el polinomio nulo, ya que s_1, s_2 y $g(x_2)$ valen 0.

Por otro lado, de (0) obtenemos que:

$$p_2(x) = (x + 1)^2((s_3 - 2)x + 1),$$

$$p_3(x) = \frac{(2-x)^2}{4}\left((s_3 + 1)x + 1\right) + \frac{x^2}{4}(3-x)$$

Por ejemplo, si tomamos $s_3 = 0$, entonces:

$$f(x) = \begin{cases} 0 & -2 \le x \le 1 \\ (x+1)^2(-2x+1) & -1 \le x \le 0 \\ 1 & 0 \le x \le 2 \end{cases}$$

Mientras que si $s_3 = 1$, entonces:

$$f(x) = \begin{cases} 0 & -2 \le x \le 1 \\ (x+1)^2(-x+1) & -1 \le x \le 0 \\ \frac{(2-x)^2}{4}(2x+1) + \frac{x^2}{4}(3-x) & 0 \le x \le 2 \end{cases}$$

Los gráficos de estas funciones se pueden apreciar en las figuras *b* y *c* respectivamente.

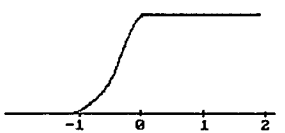


Fig. b) Interpolante cúbico de Hermite para x^0_+ con $s_3 = 0$

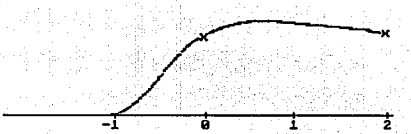


Fig. c) Interpolante cúbico de Hermite para x^0_+ con $s_3 = 1$

Nótese que el carácter local de interpolación de Hermite se aprecia en las mismas, ya que las irregularidades en una vecindad de 0, no se trasladan al resto del intervalo.

b) Interpolación mediante spline cúbico

Si los valores de g' en los puntos de interpolación se desconocen, entonces las inclinaciones $(s_i)_2^{n-1}$ se pueden escoger de modo que el interpolante f tenga hasta segunda derivada continua. En tal caso, se dice que f es un spline cúbico de interpolación. Por lo tanto, el spline cúbico es una función formada por secciones de polinomios cúbicos, que se enlazan con la mayor suavidad

posible (sin que necesariamente sea único el polinomio). Derivando dos veces la expresión (0) se obtiene:

$$p''(x) = -2s_i \frac{2x_{i+1} + x_i - 3x}{(\Delta x_i)^2}$$

$$-2\, s_{i+1} \frac{2x_i + x_{i+1} - 3x}{(\Delta x_i)^2}$$

$$+ 6 \frac{g(x_{i+1}) - g(x_i)}{(\Delta x_i)^3} (x_{i+1} + x_i - 2x)$$

Por lo tanto:

$$f''(x_i^-) = p''_{i-1}(x_i) = \frac{2s_{i-1}}{\Delta x_{i-1}} + \frac{4s_i}{\Delta x_{i-1}} - 6\frac{g(x_i) - g(x_{i-1})}{(\Delta x_{i-1})^2}$$

$$f''(x_i^+) = p''_i(x_i) = \frac{-4s_i}{\Delta x_i} - \frac{2s_{i+1}}{\Delta x_i} + 6\frac{g(x_{i+1}) - g(x_i)}{(\Delta x_i)^2} \qquad (2)$$

Y la condición de continuidad de f'':

$$f''(x_i^-) = f''(x_i^+), \qquad i = 2, ..., n-1$$

Se expresa mediante la ecuación:

$$\frac{1}{\Delta x_{i-1}} s_{i-1} + 2\left(\frac{1}{\Delta x_{i-1}} + \frac{1}{\Delta x_i}\right) s_i + \frac{1}{\Delta x_i} s_{i+1}$$

$$= 3\left[\frac{(g(x_i) - g(x_{i-1})}{(\Delta x_{i-1})^2} + \frac{g(x_{i+1}) - g(x_i)}{(\Delta x_i)^2}\right]$$

O en forma más conveniente como:

$$\Delta x_i s_{i-1} + 2(\Delta x_{i-1} + \Delta x_i)s_i + \Delta x_{i-1} s_{i+1} = b_i \quad i = 2, ..., n-1 \qquad (3)$$

Donde:

$$b_i = 3\left[\Delta x_i \frac{g(x_i) - g(x_{i-1})}{\Delta x_{i-1})} + \Delta x_{i-1} \frac{g(x_{i+1}) - g(x_i)}{\Delta x_i}\right]$$

Suponiendo entonces que s_1 y s_n se seleccionan de alguna forma (3), representa un sistema de n-2 ecuaciones lineales para calcular las n-2 incógnitas s_2, ..., s_{n-1}. La matriz de este sistema es tridiagonal y de diagonal dominante por filas. Por lo tanto (3) tiene exactamente una solución que se puede hallar por el método de eliminación de Gauss sin pivoteo. Los parámetros s_1 y s_n se pueden escoger de diferentes formas. A continua-

ción presentamos algunas de ellas:

i) Si se conoce el valor de g' en x_1 y x_n entonces resulta muy natural tomar $s_1 = g'(x_1)$ y $s_n = g'(x_n)$. En tal caso el spline cúbico de interpolación no sólo interpola a g en los puntos $x_1, ..., x_n$, sino que además interpola a g' en x_1 y x_n. Este spline se conoce como *spline cúbico completo de interpolación.*

Si hacemos $u_{i-1} = \dfrac{\Delta x_i}{\Delta x_i + \Delta x_{i-1}}$, $1 - u_{i-1} = \dfrac{\Delta x_i}{\Delta x_i + \Delta x_{i-1}}$

entonces el sistema se puede escribir en la forma:

$$\begin{pmatrix} 2 & 1-u_1 & & & & \\ u_2 & 2 & 1-u_2 & & & \\ 0 & u_3 & 2 & 1-u_3 & & \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & 1-u_{n-3} & \\ 0 & 0 & 0 & \ldots\ u_{n-2} & 2 & \end{pmatrix} \begin{pmatrix} s_2 \\ s_3 \\ s_4 \\ \cdot \\ \cdot \\ \cdot \\ s_{n-2} \\ s_{n-1} \end{pmatrix} = \begin{pmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \\ \cdot \\ \cdot \\ \cdot \\ \beta_{n-2} \\ \beta_{n-1} \end{pmatrix}$$

Donde:

$$\beta_i = 3\,[u_{i-1}\, g\,[x_{i-1}, x_i] + (1 - u_{i-1})\, g[x_i, x_{i+1}]\,] \quad i = 2, ..., n-1$$

$$\bar{\beta}_2 = \beta_2 - u_1 s_1$$

$$\bar{\beta}_{n-1} = \beta_{n-1} - (1 - u_{n-2})\, s_n$$

ii) Otra posible selección de las condiciones en los extremos surge de exigir que:

$$f''(x_1) = f''(x_n) = 0 \qquad (4)$$

Utilizando (2) la condición (3) se expresa de la siguiente forma, en término de las inclinaciones.

$$2s_1 + s_2 = 3\,\frac{g(x_2) - g(x_1)}{\Delta x_1}$$

$$s_{n-1} + 2s_n = 3\,\frac{gx_n - g(x_{n-1})}{\Delta x_{n-1}}$$

El sistema de ecuaciones queda:

$$\begin{pmatrix} 2 & 1 & & & & \\ u_1 & 2 & 1-u_1 & & & \\ 0 & u_2 & 2 & 1-u_2 & & \\ \cdot & & \cdot & \cdot & & \\ \cdot & & \cdot & \cdot & & \\ \cdot & & u_{n-2} & 2 & 1-u_{n-2} & \\ 0 & \cdot & \cdot & \cdot & u_{n-1} & 2 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ \cdot \\ \cdot \\ \cdot \\ s_{n-1} \\ s_n \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \cdot \\ \cdot \\ \cdot \\ \beta_{n-1} \\ \beta_n \end{pmatrix}$$

Donde:

$$\beta_1 = 3g[x_1, x_2] \quad \beta_n = 3g[x_{n-1}, x_n]$$

El spline que satisface (4) se conoce como *spline natural de interpolación* y, desde cierto punto de vista, no es muy recomendable, ya que la imposición arbitraria de la condición (4) puede provocar que cerca de los extremos x_1 y x_n el error aumente (a menos que realmente $g''(x_1) = g''(x_n) = 0$).

iii) Si uno no conoce nada acerca de las derivadas de g en los puntos extremos, entonces una posiblidad es escoger s_1 y s_n, de manera que p_1 coincida idénticamente con p_2 y p_{n-1} coincida con p_{n-2}. En otras palabras se trata de escoger s_1 y s_n de modo que los puntos x_2 y x_{n-1} no sean puntos de ruptura activos.

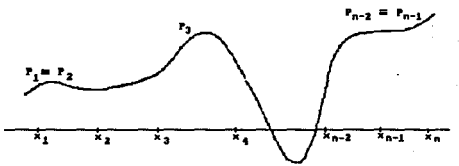


Fig. d) Los puntos x_2 y x_{n-1} no son puntos de ruptura activos

Esta condición es equivalente a exigir f''' sea continua en x_2 y x_{n-1} y significa añadir la siguiente ecuación al inicio del sistema (3)

$$s_1\Delta x_2 + s_2(x_3 - x_1) = \frac{(\Delta x_1 + 2(x_3 - x_1))\Delta x_2(g(x_2) - g(x_1))}{\Delta x_1 \quad x_3 - x_1} + \frac{(\Delta x_1)^2(g(x_3) - g(x_2))}{\Delta x_2 \quad x_3 - x_1}$$

Y la ecuación:

$$s_{n-1}(x_n + x_{n-2}) + s_n\Delta x_{n-2} = \frac{(\Delta x_{n-1})^2(g(x_{n-2}) - g(x_{n-1}))}{\Delta x_{n-2} \quad x_n - x_{n-2}} + \frac{2(x_n - x_{n-2}) + \Delta x_{n-1})\Delta x_{n-2}(g(x_n) - g(x_{n-1}))}{\Delta x_{n-1} \quad x_n - x_{n-2}}$$

Al final del mismo. Esta condición de frontera significa que la primera y la última sección polinomial interpolan

a g, en un punto adicional que no es de ruptura. En otras palabras, en vez de *n-1* secciones polinomiales tenemos entonces *n-3*, de modo que f coincide con p_1 en $[x_1, x_3]$ y $p_1(x_i) = g(x_i)$, i =1, 2, 3, mientras que $p'_1(x_3) = s_3$. Similarmente f coincide con p_{n-3} en $[x_{n-2}, x_n]$ y $p_{n-3}(x_i) = g(x_i)$ i = n-2, n-1, n, mientras que $p'_{n-3}(x_{n-2}) = s_{n-2}$ Si planteamoes la condición de frontera en estos términos, entonces el sistema lineal (3) (y la notación) varía ligeramente. Sin embargo esta claro que la función f resultantes es idéntica a la obtenida por el planteamiento anterior. Por otro lado, esta forma de interpretar las condiciones de frontera, nos ilustra nuevamente el hecho de que en la interpolación polinomial por tramos, los puntos de interpolación y los de ruptura no necesariamente tienen que coincidir.

Ejemplos:

Calculemos ahora el spline cúbico completo que interpola la función $g(x) = \dfrac{1}{1+x^2}$ en los puntos -3, -1, 0, 1 y 3. Utilizando los datos de la tabla 1 construimos el sistema de ecuaciones (3) con las condiciones de frontera i). Este sistema se expresa matricialmente de la siguiente forma:

$$\begin{pmatrix} 1 & & & & \\ 1 & 6 & 2 & & \\ & 1 & 4 & 1 & \\ & & 2 & 6 & 1 \\ & & & & 1 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{pmatrix} = \begin{pmatrix} 0.06 \\ 3.6 \\ 0 \\ -1.8 \\ -0.06 \end{pmatrix}$$

Obteniéndose que:

$s_1 = 0.06$, $s_2 = 0.62$, $s_3 = -0.09$, $s_4 = -0.26$ y $s_5 = -0.06$

Sustituyendo en la expresión (0) podemos calcular los cuatro polinomios que constituyen el spline cúbico f. De este modo, llegamos a que:

$$f(x) = \begin{cases} 0.07x^3 + 0.56x^2 + 1.53x + 1.54 & -3 \le x \le -1 \\ 0.47x^3 - 1.06x^2 - 0.09x + 1 & -1 \le x \le 0 \\ 0.65x^3 - 1.06x^2 - 0.09x + 1 & 0 \le x \le 1 \\ 0.02x^3 - 0.07x^2 - 0.18x + 0.73 & -1 \le x \le 3 \end{cases}$$

El gráfico de esta función se aprecia en la figura *e*:

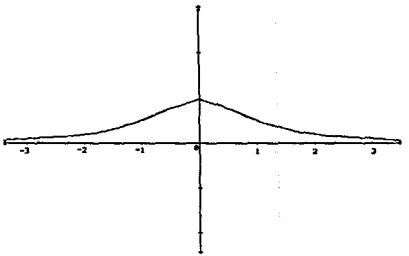


Fig. e) Spline cúbico que interpola a la función $g(x)=1/(1+x^2)$

Para finalizar esta sección, construyamos el spline cúbico natural que interpola a la función x^0_+ en los puntos -2, -1, 0 y 2. Utilizando esta vez los datos de la tabla 2 y las condiciones de frontera ii), construimos el sistema de ecuaciones (3), que nos permite calcular los coeficientes del spline:

$$\begin{pmatrix} 2 & 1 & & \\ 1 & 4 & 1 & \\ & 2 & 6 & 1 \\ & & 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ 6 \\ 0 \end{pmatrix}$$

Redondeando a dos lugares decimales, resulta que:

$$s_1 = -0.3, \quad s_2 = 0.61, \quad s_3 = 0.87, \quad s_4 = 0.43$$

Mediante la expresión (0), calculamos entonces los polinomios que forman al spline *f* y obtenemos finalmente.

$$f(x) = \begin{cases} 0.3x^3 + 1.83x^2 + 3.35x + 1.83 & -2 \le x \le -1 \\ -0.52x^3 - 0.65x^2 + 0.87x + 1 & -1 \le x \le 0 \\ 0.11x^3 - 0.65x^2 + 0.87x + 1 & 0 \le x \le 2 \end{cases}$$

Donde los coeficientes de cada polinomio también se aproximan a dos lugares decimales. La figura *f* muestra el gráfico de *f*:

Fig. f) Spline cúbico que interpola a la función $g(x) = x^0_+$
x-puntos de interpolación

Observemos con detenimiento las figuras *a*, *b*, *e* y *f*. Como se puede apreciar, en general el interpolante cúbico de Hermite aproxima mejor a las funciones de prueba que el spline cúbico. Esto se debe esencialmen-

te a que el spline cúbico invierte parte de sus posibilidades para aproximar datos, en ser una función suave (tiene hasta segunda derivada continua), mientras que el interpolante cúbico de Hermite, al ser una función menos suave, tiene más libertad para modelar un conjunto de datos. Sin embargo, hay que tener en cuenta que la construcción del interpolante cúbico de Hermite, requiere que conozcamos la primera derivada en los puntos de interpolación de la función que vamos a aproximar. Pero, muchas veces, en la práctica esto no se sabe, porque los datos se obtienen de forma experimental y, en realidad, la función g es desconocida. Por eso, en tales problemas el spline cúbico es una herramienta fundamental.

EJERCICIOS Y PROBLEMAS*

1. Use los polinomios interpolantes de Lagrange apropiados, de grado uno, dos, tres y cuatro para aproximar:

   a) $f(2.5)$ si $f(2.0) = 0.5103757$, $f(2.2) = 0.5207843$, $f(2.4) = 0.5104147$, $f(2.6) = 0.4813306$, $f(2.8) = 0.4359160$

   b) $f(0)$ si $f(-0.3) = -0.20431$, $f(-0.1) = 0.08993$, $f(0.1) = 0.11007$, $f(0.3) = 0.39569$, $f(0.5) = 0.79845$

   c) $f(1.25)$ si $f(1.0) = 0.24255$, $f(1.1) = 0.48603$, $f(1.2) = 0.86160$, $f(1.3) = 1.59751$, $f(1.4) = 3.76155$

   d) $f(0.5)$ si $f(0.2) = 0.9798652$, $f(0.4) = 0.9177710$, $f(0.6) = 0.8080348$, $f(0.8) = 0.6386093$, $f(1.0) = 0.3843735$

   e) $f(0.2)$ si $f(0.1) = 1.2314028$, $f(0.3) = 1.9121188$, $f(0.4) = 2.3855409$, $f(0.5) = 2.9682818$, $f(0.6) = 3.6801169$

2. Use los valores de abajo para construir un polinomio de Lagrange de grado dos o menor. Encuentre una aproximación para *sen 0.34*.

   $sen\ 0.30 = 0.29552$ $\quad$ $sen\ 0.32 = 0.31457$

   $sen\ 0.3 = 0.34290$

3. Agregue el valor $sen\ 0.33 = 0.32404$ a los datos en el ejercicio 2 y construya un polinomio de Lagrange de grado tres o menor. Aproxime el *sen 0.34*.

4. Sea $f(x) = 3xe^x - 2e^x$. Aproxime $f(1.03)$ usando el polinomio interpolante de grado menor o igual a dos, con $x_0 = 1$, $x_1 = 1.05$ y $x_2 = 1.07$.

5. Use los valores de abajo para construir una aproximación polinómica de Lagrange de tercer grado para $f(1.09)$. La función que se está aproximando es $f(x) = \log_{10} \tan x$.

   $f(1.00) = 0.1924$ $\quad$ $f(1.05) = 0.2414$

   $f(1.10) = 0.2933$ $\quad$ $f(1.15) = 0.3492$

6. Use el polinomio interpolante de Lagrange de grado tres o menor para aproximar $\cos 0.750$ usando los valores de abajo.

   $\cos 0.698 = 0.7661$ $\quad$ $\cos 0.768 = 0.7193$

   $\cos 0.733 = 0.7432$ $\quad$ $\cos 0.803 = 0.6946$

*Los ejercicios se tomaron y/o adaptaron de las obras citadas al final de esta sección.

7. Use los valores de abajo para construir una aproximación polinómica de Lagrange de cuarto grado para $f(1.25)$. La función que se está aproximando es $f(x) = e^{x^2-1}$.

$f(1.0) = 1.00000 \quad f(1.2) = 1.55271 \quad f(1.4) = 2.61170$

$f(1.1) = 1.23368 \quad f(1.3) = 1.99372$

8. Sea $f(x) = (4x - 7)/(x - 2)$ y $x_0 = 1.7$, $x_1 = 1.8$, $x_2 = 1.9$ y $x_3 = 2.1$.

   a) Aproxime $f(1.75)$ usando el polinomio interpolante de grado a lo más dos en los nodos x_0, x_1, y x_2.

   b) Aproxime $f(1.75)$ y $f(2.00)$ usando el polinomio interpolante en x_0, x_1, x_2 y x_3.

9. Sea $f(x) = e^x$, $0 \le x \le 2$. Usando los valores dados en la tabla, efectúe los siguientes cálculos;

   a) Aproxime $f(0.25)$ usando interpolación lineal con $x_0 = 0$ y $x_1 = 0.5$

   b) Aproxime $f(0.75)$ usando interpolación lineal con $x_0 = 0.5$ y $x_1 = 1$

   c) Aproxime $f(0.25)$ y $f(0.75)$ usando el polinomio interpolante de segundo grado con $x_0 = 0, x_1 = 1$ y $x_2 = 2$

   d) ¿Cuáles aproximaciones son mejores? ¿Por qué?

x	0.0	0.5	1.0	2.0
$f(x)$	1.00000	1.64872	2.71828	7.38906

10. La siguiente tabla muestra la población de los Estados Unidos de América desde 1930 hasta 1980.

Año	1930	1940	1950	1960	1970	1980
Población (en miles)	123,203	131,669	150,697	179,323	203,212	226,505

Encuentre el polinomio de Lagrange de grado 5 que ajusta estos datos y use este polinomio para estimar la población en los años 1920, 1965 y 2000. La población en 1920 fue aproximadamente de 105,711,000. ¿Qué tan exactos piensa usted que son sus resultados de 1965 y 2000?

11. Calcúlese el logaritmo de 4 en base 10 (*log 4*) usando interpolación lineal. a) Interpolar entre *log 3 = 0.477 121 3* y *log 5 = 0.698 970 0*. b) Interpolar entre *log 3* y *log 4.5 = 0.653 212 5*. Para cada una de las interpolaciones calcúlese el error relativo porcentual basado en el valor verdadero de *log 4 = 0.602 060 0*.

12. Ajústese un polinomio de interpolación de Newton de segundo orden para aproximar *log 4*, usando los

datos del problema 11. Calcúlese el error relativo porcentual.

13. Ajústese un polinomio de interpolación de Newton de tercer orden para calcular *log 4* usando los datos del problema 11 además del punto adicional, *log 3.5 = 0.544 068 0*. Calcúlese el error relativo porcentual.

14. Dados los datos:

x	0	0.5	1.0	1.5	2.0	2.5
$f(x)$	1	2.119	2.910	3.945	5.720	8.695

Calcúlese $f(1.6)$ usando polinomio de interpolación de Newton de orden 1 hasta el 3. Escójase la secuencia de puntos de las aproximaciones para lograr exactitud.

15. Dados los datos:

x	1	2	3	5	6
$f(x)$	4.75	4	5.25	19.75	36

Calcúlese $f(3.5)$ usando polinomios de interpolación de Newton de orden 1 hasta el 4. Escójanse los puntos base para obtener una buena aproximación. ¿Qué indican los resultados respecto al orden del polinomio que se usa para generar los datos en la tabla?

16. Repítanse los problemas 11 al 13 usando polinomios de Lagrange.

17. Repítase el problema 14 usando interpolación de Lagrange.

18. Repítase el problema 15 usando polinomios de Lagrange de orden 1 hasta el 3.

19. Desarróllese la interpolación usando splines cuadráticos para los datos del problema 15 y calcúlese $f(3.5)$

20. Desarróllese la interpolación mediante splines cúbicos para los datos del problema 15 y calcúlese $f(3.5)$

21. Use interpolación de spline cúbico natural para encontrar una aproximación a:

a) $f(2.5)$ dado que:

x	$f(x)$
2.2	0.5207843
2.4	0.5104147
2.6	0.4813306

b) $f(5.3)$ dado que:

x	$f(x)$
5.0	2.168861
5.2	1.797350
5.4	1.488591

c) $f(0)$ dado que:

x	$f(x)$
-0.3	-0.20431
-0.1	-0.08993
0.1	0.11007
0.3	0.39569

d) $f(1.25)$ dado que:

x	$f(x)$
1.1	0.48603
1.2	0.86160
1.3	1.59751
1.4	3.76155

e) $f(0.5)$ dado que:

x	$f(x)$
0.2	0.9798652
0.4	0.9177710
0.6	0.8080348
0.8	0.6386093
1.0	0.3843735

f) $f(0.2)$ dado que:

x	$f(x)$
0.1	1.2314028
0.3	1.9121188
0.4	2.3855409
0.5	2.9682818
0.6	3.6801169

22. Repita el ejercicio 21 usando interpolación de spline cúbico y el hecho de que:

a) $f'(2.2) = -0.0014878, f'(2.6) = -0.1883635$

b) $f'(5.0) = -1.495067, f'(5.4) = 1.070309$

c) $f'(-0.3) = 0.32213, f'(0.3) = 1.67787$

d) $f'(1.1) = 2.90986, f'(1.4) = 41.13928$

e) $f'(0.2) = 0.20271, f'(1.0) = 1.55741$

f) $f'(0.1) = 2.64281, f'(0.6) = 7.84023$

23. Construya un spline cúbico natural para aproximar $f(x) = \cos \pi x$ usando los valores dados por $f(x)$ en $x = 0, 0.25, 0.5, 0.75, 1.0$. Integre el spline sobre $[0, 1]$, y compare el resultado con $\int_0^1 \cos \pi x \, dx = 0$. Use las derivadas del spline para aproximar $f'(0.5)$ y $f''(0.5)$. Compare estas aproximaciones con los valores reales.

24. Construya un spline cúbico para aproximar $f(x) = e^{-x}$ usando los valores dados por $f(x)$ en $x = 0, 0.25, 0.75, 1.0$. Integre el spline sobre $[0, 1]$ y compare el resultado con $\int_0^1 e^{-x} \, dx = (1/e)(e - 1)$. Use las derivadas del spline para aproximar $f'(0.5)$ y $f''(0.5)$. Compare estas aproximaciones con los valores reales.

25. Repita el ejercicio 23, construyendo ahora el spline cúbico con $f'(0) = f'(1) = 0$.

26. Repita el ejercicio 24, construyendo en su lugar el spline cúbico con $f'(0) = -1, f'(1) = -e^{-1}$.

27. a) Use los valores de abajo para construir un spline cúbico natural para aproximar *sen 0.34*.

x	$sen\, x$	$D_x(sen\, x) = cos\, x$
0.30	0.29552	0.95534
0.32	0.31457	0.94924
0.35	0.34290	0.93937

b) Use el spline construido en (a) para aproximar *cos 0.34*

c) Use el spline construido en (a) para aproximar:

$$\int_{0.30}^{0.35} sen\, x\, dx$$

28. Añada *sen 0.33 = 0.32404* y *cos 0.33 = 0.94604* a los datos del ejercicio 27 y rehaga los cálculos de ese ejercicio.

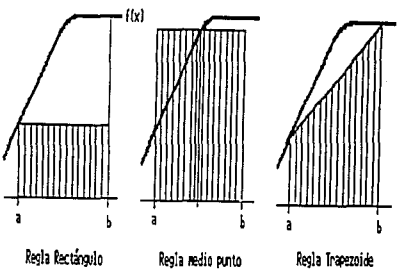
Referencias bibliográficas:

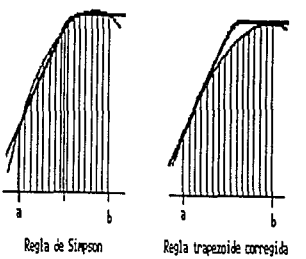
1. *Burden, Richard L.; Faires, J. Douglas.* "Análisis Numérico". México. Grupo Editorial Iberoamérica, 1985.

2. *Chapra, Steven C; Canale, Raymond P.* "Métodos Numéricos para Ingenieros". México. McGraw-Hill, 1987.

CAPÍTULO VI

INTEGRACIÓN





"La verificación experimental de una teoría sobre un fenómeno natural, generalmente descansa en el resultado de una integración."

J. W. MELLOR

(Higher Mathematics for Students of Chemistry and Physics)

"El cálculo es la mayor ayuda que tenemos para apreciar las verdades físicas en el más amplio sentido de la palabra."

W. F. OSGOOD

(Bulletin American Mathematical Society Vol. 13)

INTRODUCCION

La física nos informa que la energía E, requerida para mover un cuerpo de la posición a a la posición b, a lo largo de una línea recta, se expresa así:

$$E = \int_a^b f(x)\ dx$$

Donde $f(x)$ es la fuerza reactiva. Esta fórmula es importante para estimar la carga explosiva necesaria para obtener de un cañón un rango preestablecido; también permite estimar el combustible necesario para colocar un satélite en órbita. En ambos casos, las fuerzas reactivas son la fricción atmosférica y la gravitación terrestres.

Los métodos de integración numérica desarrollados en este capítulo son herramientas para evaluar, entre otros fenómenos, la energía E cuando no existe la antiderivada de f o cuando sólo se conocen algunos valores de f.

Además, la integración numérica es importante en el estudio de modelos probabilísticos; ya que muchos de

estos conducen a integrales, para las que no existen fórmulas simples, como en los siguientes casos:

$$\int_a^b e^{-x^2}\,dx \qquad \int_a^b x^{\alpha}(1-x)^{\beta}\,dx$$

Los métodos de interpolación del capítulo anterior proporcionan el fundamento de la integración numérica, ya que dada la función f que se desea integrar, se aproxima ésta mediante un polinomio de interpolación $p(x)$ y se integra $p(x)$ en lugar de f, obteniéndose un valor aproximado de $\int f$ a partir de $\int p$. La integración de un polinomio se reduce a una secuencia de operaciones aritméticas, que se pueden efectuar mediante una computadora.

FÓRMULAS GENERALES

Una de las formas usuales de integrar numéricamente una función $f(x)$ consiste en aproximar a f, mediante un polinomio $p(x)$por interpolación, e integrar el polinomio. La mayor parte de las fórmulas de interpolación se desprende de este método:

$$\int_a^b f(x)\,dx \approx \int_a^b p(x)\,dx$$

En donde $p(x) \approx f(x)$ *en* $[a, b]$.

En particular, si $p(x)$ está representado en la forma de Lagrange, entonces:

$$p(x) = f_0 l_0(x) + f_1 l_1(x) + \ldots + f_n l_n(x)$$

Por lo tanto:

$$\int_a^b f(x)\,dx \approx f_0\int_a^b l_0(x)\,dx + \ldots + f_n\int_a^b l_n(x)\,dx$$

En general:

$$\int_a^b p(x)\,dx = f_0 w_0 + f_1 w_1 + \ldots + f_n w_n$$

$$\int_a^b p(x) = \sum_{i=0}^{n} w_i f_i$$

Nótese que:

$$\int_a^b l_i(x) = w_i$$

Es independiente de la función f. A la siguiente aproximación se le denomina *Regla de cuadratura:*

$$\int_a^b f(x)\,dx \approx \sum f(x_i)\,w_i$$

Esta regla da lugar a diversos métodos, dependiendo de los puntos de interpolación $(x_i, f(x_i))$ y, en consecuencia, del grado del polinomio interpolante.

Cuadratura de Newton-Cotes

Los puntos x_i se denominan nodos de integración y se eligen, generalmente, en el intervalo $[a, b]$.

Los métodos que emplean interpolación polinomial con *nodos igualmente espaciados* se denominan, en general *Fórmulas de Newton-Cotes.*

En general, una Regla de Cuadratura de orden $n+1$ de Newton-Cotes es de la forma:

$$\int_a^b f(x)\,dx \approx \sum_{k=0}^{n} p(x_k)\,w_k$$

En donde $x_0 = a$ y $x_k = x_0 + hk$, $\quad k = 1, 2, \ldots, n$

Ejemplos:

1) Si $n = 1$, obtenemos la regla del trapecio:

$$\int_a^b f(x)\,dx \approx \frac{(b-a)}{2}[f(a) + f(b)]$$

2) Si $n = 2$, obtenemos la regla de Simpson:

$$\int_a^b f(x)\,dx \approx \frac{(b-a)}{b}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

3) Si $n = 3$; obtenemos la regla de los tres octavos.

$$\int_a^b f(x)\,dx \approx \frac{3(b-a)}{8 \times 4}\left[f(a) + 3f_1\left(a + \frac{b-a}{3}\right) + 3f\left(a + 2\frac{(b-a)}{3}\right) + f(b)\right]$$

Cuadratura de Gauss

Es posible obtener reglas de cuadratura proponiendo, desde el principio, que se quiere la mayor precisión posible con un número de puntos dado:

Ejemplo:

$$\int_{-1}^{1} f(x)\,dx \approx w_1 f(-1) + w_2 f(0) + w_3 f(1)$$

Queremos que esta aproximación sea "exacta" para polinomios de grado ≤ 2, entonces w_1, w_2, w_3 se calculan a partir de esta condición:

si $f(x) = 1 \int_{-1}^{1} f\,dx = 2 = w_1 + w_2 + w_3$

si $f(x) = x \int_{-1}^{1} x\,dx = 0 = w_1(-1) + 0.w_2 + w_3(1)$

si $f(x) = x^2 \int_{-1}^{1} x^2\,dx = \frac{2}{3} = w_1 f + w_2.0 + w_3.(1)$

Resolviendo el sistema para w_1, w_2, w_3 obtenemos:

$$w_1 = \frac{1}{3} = w_3 \qquad w_2 = \frac{4}{3}$$

Así, obtenemos la regla de cuadratura:

$$\int_{-1}^{1} f(x)\,dx \approx \frac{1}{3}f(-1) + \frac{4}{3}f(0) + \frac{1}{3}f(1)$$

Ejemplo:

$$\int_{-1}^{1} f(x)\,dx \approx w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3)$$

Donde queremos determinar $x_1, x_2, x_3, w_1, w_2, w_3$, de tal forma que la aproximación sea "exacta" para polinomios del mayor grado posible, en este caso sería para polinomios de grado menor o igual que 5

Procediendo como en el ejemplo anterior obtenemos el sistema de ecuaciones:

$$2 = w_1 + w_2 + w_3$$

$$0 = w_1x_1 + w_2x_2 + w_3x_3$$

$$\frac{2}{3} = w_1x_1^2 + w_2x_2^2 + w_3x_3^2$$

$$0 = w_1x_1^3 + w_2x_2^3 + w_3x_3^3$$

$$\frac{2}{5} = w_1x_1^4 + w_2x_2^4 + w_3x_3^4$$

$$0 = w_1x_1^5 + w_2x_2^5 + w_3x_3^5$$

Este es un sistema de ecuaciones no lineales en $w_1, w_2, w_3, x_1, x_2, x_3$, que afortunadamente se puede resolver, obteniéndose:

$$w_1 = w_3 = \frac{5}{9} : w_2 = \frac{8}{9}$$

$$x_1 = -\sqrt{3/5} : x_2 = 0 : x_3 = -x_1$$

Por lo que la regla de cuadratura que se obtiene es:

$$\int_{-1}^{1} f(x)\,dx \approx \frac{5}{9}f(-\sqrt{3/5}) + \frac{8}{9}f(0) + \frac{5}{9}f(\sqrt{3/5})$$

Este tipo de regla de cuadratura es conocido como cuadratura gaussiana.

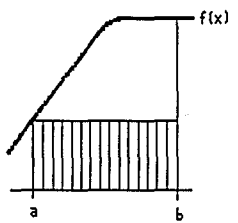
METODOS ELEMENTALES

1. Reglas del rectángulo.

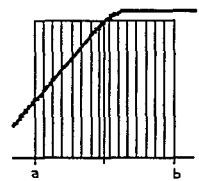
Este método de integración numérica, considera como polinomio de interpolación a una constante, esto es, un polinomio de grado cero y a un sólo punto del intervalo $[a,b]$; generalmente se considera $x = a, x = b$ ó $x = \frac{a+b}{2}$.

Dependiendo del punto considerado, tenemos tres posibles aproximaciones, que son las siguientes:

i) si $x = a \Rightarrow \int_a^b f(x)\,dx \approx f(a)\,(b-a)$



ii) si $x = \frac{a+b}{2} \Rightarrow \int_a^b f(x)\,dx \approx f\left(\frac{a+b}{2}\right)(b-a)$



iii) si $x = b \Rightarrow \int_a^b f(x)\,dx \approx f(b)\,(b-a)$ estos son los tres casos típicos de la regla del rectángulo.

Ejemplo: Aplique las tres reglas anteriores para calcular:

$$I = \int_0^1 \exp(-x^2)\,dx$$

$a = 0$ $\quad b = 1$ $\quad (a+b)/2 = 1/2$

$f(0) = 1$ $\quad f(1) = 0.36788$ $\quad f(1/2) = 0.77880$

Por lo tanto:

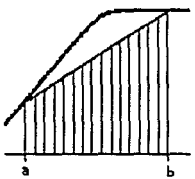
i) $I = f(a)\,(b-a) = 1(1) = 1$

ii) $I = f\left(\frac{a+b}{2}\right)(b-a) = (0.77880)\,1 = 0.77880$

iii) I = f(b) (b-a) = (0.36788) (1) = 0.36788

2. Regla del trapecio

Esta regla considera como polinomio de interpolación la recta que pasa por los puntos $(a, f(a))$ y $(b, f(b))$ y en consecuencia a los puntos a y b del intervalo $[a, b]$.



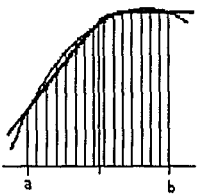
La regla de cuadratura toma la forma:

$$\int_a^b f(x)\,dx \approx \frac{f(a)+f(b)}{2}\ (b-a)$$

Ejemplo: (ver página 182)

3. Regla de Simpson

El polinomio de interpolación es, en este caso, una parábola vertical, esto es, un polinomio de segundo grado; por lo tanto, se requieren tres puntos x_1, x_2, x_3 en el intervalo $[a, b]$.



La fórmula de cuadratura que se obtiene:

$$\int_a^b f(x)\,dx \approx \frac{1}{3}\left(\frac{b-a}{2}\right)\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right)$$

Ejemplo: (ver página 182)

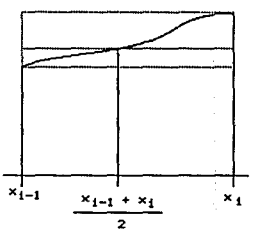
Las reglas anteriores son simples o elementales y todas ellas se desprenden de interpolar a la función mediante un polinomio de grado cero, uno y dos respectivamente. A partir de estas, se desprenden las denominadas reglas compuestas, que son la aplicación por segmentos de cada una de las reglas simples, imitando la interpolación por splines.

REGLAS COMPUESTAS

Consideremos ahora el intervalo $[a, b]$ dividido en n segmentos de la misma longitud h.

Sea uno de estos segmentos el representado en la gráfica siguiente:

Entonces, tenemos las siguientes posibilidades:



a) $f(\frac{x_{i-1} + x_i}{2})h_i$ *Regla del punto medio*

b) $\frac{f(x_{i-1}) + f(x_i)}{2} h_i$ *Regla del trapecio*

c) $\frac{h_i}{6}(f(x_{i-1}) + 4f(\frac{x_{i-1} + x_i}{2}) + f(x_i))$ *Regla de Simpson*

A partir de las anteriores, podemos expresar las siguientes fórmulas de integración:

$$\sum_{i=1}^{n} f\left(\frac{x_{i-1} + x_i}{2}\right) h_i \qquad \textit{Punto medio}$$

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} \frac{f(x_{i-1}) + f(x_i)}{2} h_i$$

$$= [\frac{f(x_0)}{2} h_1 + f(x_1)\frac{h_1 + h_2}{2} + \ldots + f(x_{n-1})\frac{h_{n-1} + h_n}{2} + \frac{f(x_n)}{2} h_n]$$

$$= \frac{1}{2}[f(x_0)\,h_1 + f(x_1)\,(h_1 + h_2) + \ldots + f(x_{n-1})\,(h_{n-1} + h_n)$$

$$+ f(x_n)\,h_n] \qquad \textit{Trapecio}$$

$$\sum_{i=1}^{n} \frac{h_i}{6}\,(f(x_{i-1}) + 4f\left(\frac{x_{i-1} + x_i}{2}\right) + f(x_i)) \qquad \textit{Simpson}$$

Tomando $h_i = h\ \forall_i$, las fórmulas anteriores se simplifican para dar:

$$h \sum_{i=1}^{n} f\left(\frac{x_{i-1} + x_i}{2}\right) \qquad \textit{Punto medio}$$

$$\frac{h}{2}[f(x_0) + 2f(x_1) + \ldots + 2f(x_{n-1}) + f(x_n)] \quad \textit{Trapecio}$$

$$\frac{h}{6}[f(x_0) + 4f\left(\frac{x_0 + x_1}{2}\right) + 2f(x_1) + 4f\left(\frac{x_1 + x_2}{2}\right) + 2f(x_2)$$

$$+ \ldots + 4f\left(\frac{x_{n-1} + x_n}{2}\right) + f(x_n)] \qquad \textit{Simpson}$$

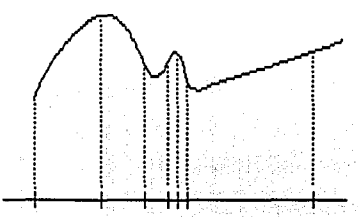
Las reglas compuestas requieren de una partición del intervalo $[a, b]$.

Cuadratura adaptativa

En este método, la división del intervalo $[a, b]$ para determinar los puntos de cuadratura x_i se hace de tal manera que se va refinando la subdivisión de $[a, b]$, hasta obtener la precisión deseada. Recordemos que en las fórmulas compuestas de Newton-Cotes la partición del intervalo I se hace de tal manera que la longitud h de los subintervalos es la misma.

Esta idea resulta bien cuando la función tiene un comportamiento similar en todo el intervalo, pero se vuelve ineficiente cuando la función es más lisa en unos puntos del intervalo que en otros.

Esta situación se representa en la gráfica siguiente, en la que es obvio que el refinamiento del intervalo debe hacerse en la región o regiones donde la función es menos suave, esto es, hay que refinar sólo donde sea necesario:



El procedimiento, se define de la siguiente manera:

Calcular $Q_1 = \int f$ sobre el intervalo $[a, b] = I_0$ subdividir el intervalo $[a, b]$ de tal manera que:

$$I_1 = \left[a, \frac{a+b}{2}\right]$$

$$I_2 = \left[\frac{a+b}{2}, b\right]$$

Calcular Q_{11} y Q_{12} integrales de f en I_1 e I_2. Comparar Q_1 con $Q_{11} + Q_{12}$. Esto es:

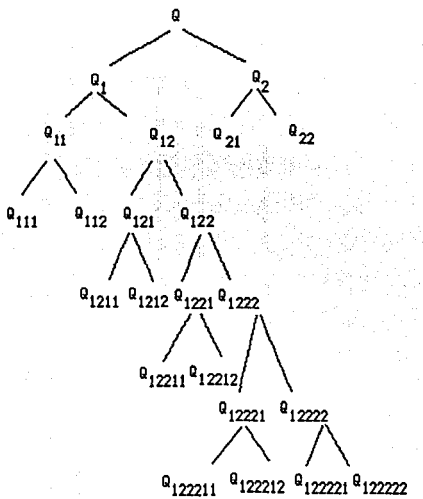
$$| Q_1 - (Q_{11} + Q_{12}) | < \varepsilon$$

Si es así, entonces $Q_{11} + Q_{12}$ es la solución. En caso contrario, repetimos el procedimiento para I_1 haciendo la comparación entre Q_{11} y $Q_{111} + Q_{112}$, esto es:

$$| Q_{11} - (Q_{111} + Q_{112}) | \leq \varepsilon/2$$

En caso de ser cierta la relación anterior, detenemos el procedimiento, en caso de ser falsa, subdividimos el intervalo derecho y repetimos el proceso.

Si las Q's representan las diferentes aproximaciones de la integral en los distintos intervalos I, el esquema gráfico del proceso es el siguiente:



Ejemplo:

Evaluar $\int_1^2 \exp\left(-\frac{1}{2}\ x\right) dx$

Usando la regla del trapecio (compuesta):

El programa es el siguiente:

```
10 REM REGLA DEL TRAPECIO
20 DEF FNF(X) = EXP(-X/2)
30 DATA 1,2
40 READ A,B
50 INPUT "NUMERO DE DIVISIONES ";N
60 LET H = (B-A)/N
70 LET P = (FNF(A) + FNF(B))/2
80 FOR R = 1 TO N-1
90 LET P = P + FNF(A + R*H)
100 NEXT R
110 PRINT "LA INTEGRAL ES ";H*P
```

Resultados representativos de este programa son:

```
NUMERO DE DIVISIONES ? 10
LA INTEGRAL ES .477402

NUMERO DE DIVISIONES ? 100
LA INTEGRAL ES .477303
NUMERO DE DIVISIONES ? 1000
LA INTEGRAL ES .477302
```

El valor correcto de la integral es 0.477 302 437, de manera que el último de los valores anteriores tiene una exactitud de siete cifras decimales: pero el lector descubrirá que se necesita un tiempo bastante prolongado para correr el programa. Desde luego, casi nunca se encuentra una respuesta "correcta". Uno de los problemas más difíciles en la integración numérica es decir qué tan grande debe tomarse n a fin de obtener la exactitud necesaria. Este punto se considerará más adelante en este capítulo.

Ejemplo:

Evaluar $\int_1^2 \exp\left(-\frac{1}{2}\ x\right) dx$

Usando la regla de Simpson compuesta.

```
10 REM REGLA DE SIMPSON
20 DEF FNF(X) = EXP(-X/2)
30 DATA 1,2
40 READ A,B
50 INPUT "NUMERO DE DIVISIONES ";N
60 IF N = 2*INT(N/2) THEN GOTO 100
70 PRINT "EL NUMERO DEBE SER POSITIVO"
80 GOTO 50
100 LET H = (B-A)/N
110 LET P = FNF(A) + FNF(B):LET Z = 4
120 FOR R = 1 TO N-1
130 LET P = P + Z*FNF(A + R*H)
```

```
140 LET Z=6-Z

150 NEXT R

160 PRINT "LA INTEGRAL ES ";H*P/3
```

Obsérvese que se hace una verificación en la línea 60 para asegurar que N se par. Asimismo, la variable Z toma los valores 4, 2, 4, 2, ..., conforme se repite el ciclo FOR-NEXT; esto se logra de manera muy simple por medio de la línea 140. Son resultados representativos del programa:

```
NUMERO DE DIVISIONES ? 4
LA INTEGRAL ES .477303

NUMERO DE DIVISIONES ? 9
NUMERO DE DIVISIONES ? 10
LA INTEGRAL ES .477302

NUMERO DE DIVISIONES ? 20
LA INTEGRAL ES .477302
```

La respuesta correcta es 0.477 302 437. Con cuatro bandas se ha obtenido un resultado mejor que con cien bandas de la regla del trapecio; y con veinte bandas prácticamente se obtiene una exactitud de máquina. ¡No es de extrañar que la regla de Simpson tenga un uso tan extenso!

EJERCICIOS Y PROBLEMAS*

1. Calcule $\int_0^2 x^2\,dx$ con las reglas del rectángulo, trapecio y Simpson.

2. Dada la información:

x	0	0.5	1
f	7	19	13

Deduzca una buena estimación para $\int_0^1 f(x)\,dx$.

3. Dada la información:

x	-2	-1	0
f	60	12	3

Determine una buena estimación para $\int_{-2}^0 f(x)\,dx$.

4. Utilizando $N=2$ subintervalos, determine las estimaciones de las reglas trapezoidal y de Simpson, para $\int_{-1}^1 (1+x^3)^{1/2}\,dx$.

5. Aplique la regla de Simpson a $\int_4^{10} (x+1)^{-1}\,dx$ con:

   (a) $N = 2$ subintervalos

   (b) $N = 4$ subintervalos

6. Utilice la regla de Simpson con $\int_0^1 (1+x^2)^{-1}\,dx$ con $N = 4$ subintervalos.

7. Empleando $N = 4$ subintervalos, determine las estimaciones de las reglas trapezoidal y Simpson para $\int_{-1}^1 (1+x^4)^{-1}\,dx$.

8. (a) Demuestre que la regla de Simpson es exacta para $f(x) = Ax^3 + Bx^2 + Cx + D$ considerando cada uno de los cuatro términos por separado sobre el intervalo $a \le x \le b$, con cualquier número (par) de subintervalos N a su elección. (Si hace los cálculos a mano, la elección $N = 2$ es satisfactoria.)

   (b) Aplique la regla de Simpson con $N = 2$ subintervalos a:

$$\int_0^2 \left(\frac{7}{4}x^3 - 0.5 + 3\right) dx$$

   (c) Evalúe la integral definida de la parte (b) empleando antiderivadas; luego, opine cuál método le pareció más rápido.

*Los ejercicios se tomaron y/o adaptaron de las obras citadas al final de está sección.

9. Demuestre que existe un punto η en el intervalo $(0, 1)$, de modo que $N = 4$ subintervalos, la integral de x^5 de $x = 0$ a $x = 1$ es igual a S_4 (la regla de Simpson con cuatro subdivisiones) más $ET_{S4}(\eta)$.

10. (a) Deduzca el polinomio de interpolación $p(x)$ con $f(x)$ en $x = a, x = c = \frac{1}{2}(a + b)$ y $x = b$. Utilice el enfoque de Lagrange (o cualquier otro) para escribir el polinomio en la forma $\sum f_k L_k(x)$, en donde $f_0 = f(a), f_1 = f(c)$ y $f_2 = f(b)$.

(b) Integre el polinomio $p(x)$ de la parte (a) desde $x = a$ hasta $x = b$ y verifique que el resultado sea equivalente a la regla de Simpson aplicada a un par de subintervalos.

11. (a) Verifique que $\int_0^2 s(s-1)(s-2)\,ds = 0$.

(b) Verifique que $\int_0^2 s(s-1)(s-2)(s-3)\,ds = -4/15$

12. Pruebe el "teorema del valor medio para sumatorias":

Si $g(x)$ es continua y las constantes $\{c_k\}$ no son negativas, entonces:

$$\sum c_k g(x_k) = g(\theta) \sum c_k$$

para alguna constante θ en el intervalo determinado por las $\{x_k\}$.

[Sugerencia: aplique el teorema del valor intermedio a $F(t)$ definida como $c_k g(x_k) - g(t) \sum c_k$.]

13. Existe la denominada regla del punto medio, en la cual la función $f(x)$ se evalúa en el punto medio de cada subintervalo. La fórmula es:

$$M_N = h \sum_{k=0}^{N-1} f\left(\frac{x_k + x_{k+1}}{2}\right)$$

Y la expresión del error está dada por:

$$ET_{MN} = \frac{b-a}{24} h^2 f''(\eta).$$

(a) Demuestre que, si aplicamos tanto la regla del punto medio como la regla trapezoidal para integrar $f(x)$ desde $x = 0$ hasta $x = h$ (es decir, un subintervalo), y si tomamos dos terceras partes del resultado del punto medio M_N y una tercera parte del resultado trapezoidal T_N (este particular promedio ponderado motivado por las respectivas expresiones del error) obtenemos la regla de Simpson con una longitud de paso $\frac{1}{2}h$.

(b) Obtenga la expresión del error. [Sugerencia: reemplace $f(x)$ por un polinomio de Taylor de grado dos alrededor de: $c = \frac{1}{2}(x_k + x_{k+1}) = x_k + \frac{1}{2}h$].

14. Existe la denominada "regla trapezoidal corregida" que utiliza la derivada de f, evaluada en los dos puntos extremos. La fórmula es:

$$CT_N = T_N + (h^2/12)\,[f'(a) - f'(b)]$$

El error de truncamiento es proporcional a $h^4 f^{(4)}(\eta)$, por lo que este método es comparable en precisión de la regla de Simpson para la mayoría de las integrales.

Utilice la regla trapezoidal corregida con longitud de paso $h = 2/3$ (de modo que $N = 3$) a fin de estimar:

$$\int_0^2 \frac{1}{x + 1/3}\,dx$$

15. Empleando $N = 2$ subintervalos, estime $\int_{-1}^{5} (x^3 - 4x - 5)\,dx$ aplicando:

(a) La regla trapezoidal

(b) La regla trapezoidal corregida

(c) La regla Simpson

16. Utiliza medios analíticos para evaluar:

(a) $\int_0^{10} (10 + 2x - 6x^2 + 5x^4)\,dx$

(b) $\int_{-3}^{5} (1 - x - 4x^3 + 3x^5)\,dx$

(c) $\int_0^{\pi} (8 + 5\,sen\,x)\,dx$

17. Utilice una aplicación simple de la regla trapezoidal y evalúense las integrales del problema 16.

18. Evalúe las integrales del problema 16 con la regla trapezoidal de segmentos múltiples, con $n = 2, 4$ y 6.

19. Evalúe las integrales del problema 16 con una aplicación simple de la regla Simpson.

20. Evalúe las integrales del problema 16 con una regla de Simpson de segmentos múltiples, con $n = 4$ y 6.

21. Integre la siguiente función analíticamente y usando la regla trapezoidal, con $n = 1, 2, 3$ y 4:

$$\int_0^{3\pi/20} [sen\,(5x + 1)]\,dx$$

22. Integre la siguiente función analíticamente y usando las reglas de Simpson, con $n = 4$ y 5:

$$\int_{-4}^{6} [(4x + 8)^3]\,dx$$

Comente los resultados.

23. Integre la siguiente función analítica y numéricamente. Use la regla trapezoidal y la regla de Simpson para integrar la función. En ambos casos, use la versión de segmentos múltiples, con $n = 4$:

$$\int_0^4 xe^{2x}\,dx$$

Compare el error relativo porcentual de los resultados numéricos.

24. Integre la siguiente función analítica y numéricamente. Utilice la regla trapezoidal y la regla de Simpson.

$$\int_0^1 15.3^{2.5x}\,dx$$

Calcule el error relativo porcentual de los resultados numéricos.

25. Evalúe la integral:

$$\int_0^{\pi} (4 + 2\,sen\,x)\,dx$$

a) Analíticamente.

b) Mediante la aplicación simple de la regla trapezoidal.

c) Mediante la aplicación múltiple de la regla trapezoidal ($n = 5$).

d) Mediante la aplicación simple de la regla de Simpson.

e) Mediante la aplicación múltiple de las reglas de Simpson ($n = 5$). En los casos b) a e), calcule el error relativo porcentual basado en a).

26. Evalúe la integral de los siguientes datos tabulares mediante la regla trapezoidal:

x	0	0.1	0.2	0.3	0.4	0.5
$f(x)$	1	7	4	3	5	9

27. Efectúe las mismas evaluaciones del problema 26 usando las reglas de Simpson.

28. Evalúe la integral de los siguientes datos tabulares usando la regla trapezoidal:

x	-3	-1	1	3	5	7	9	11
$f(x)$	1	-4	-5	2	4	8	6	-3

29. Efectúe la misma evaluación del problema 28 usando las reglas de Simpson.

30. La función:

$$f(x) = 10 - 38.6x + 74.07x^2 - 40.1x^3$$

Se usa en el cálculo de la siguiente tabla de datos que no están igualmente espaciados:

x	0	0.1	0.3	0.5	0.7	0.95	1.2
$f(x)$	10	6.84	4	4.20	5.51	5.77	1

Evalúe la integral desde $a = 0$ y $b = 1.2$ usando:

a) Medios analíticos

b) La regla trapezoidal

c) Una combinación de las reglas de Simpson y la regla trapezoidal; utilice las reglas de Simpson en donde sea posible, para obtener la mayor exactitud posible.

En b y c) calcula el error relativo porcentual.

31. Evalúe la siguiente integral doble:

$$\int_{-2}^{2} \int_{0}^{4} (x^3 - 3y^2 + xy^3)\, dx\, dy$$

a) Analíticamente

b) Usando la regla trapezoidal con segmentos múltiples ($n = 2$).

c) Usando una aplicación simple de la regla de Simpson de 1/3.

En b) y c) calcúlese el error relativo porcentual.

32. Evalúe la integral triple:

$$\int_{-4}^{4} \int_{0}^{6} \int_{-1}^{3} (x^4 - 2yz)\, dx\, dy\, dz$$

a) Analíticamente.

b) Usando una aplicación simple de la regla de Simpson de 1/3.

En b) calcule el error relativo porcentual.

33. Al efectuar un estudio de la línea de ensamble de una planta de automóviles, en un periodo de 24 horas, se visitan dos puntos sobre la línea y en instantes diferentes durante el día se verifica el número de autos que pasa por ahí en un minuto. Los datos son:

Punto A		Punto B	
Tiempo	Carros/minuto	Tiempo	Carros/minuto
Medianoche	3	Medianoche	3
2 A.M.	3	1 A.M.	3
3 A.M.	5	4 A.M.	5
6 A.M.	4	5 A.M.	2
9 A.M.	5	7 A.M.	1
11 A.M.	6	10 A.M.	4
2 P.M.	2	1 P.M.	3
5 P.M.	1	3 P.M.	4
6 P.M.	1	9 P.M.	6
7 P.M.	3	10 P.M.	1
8 P.M.	4	11 P.M.	3
Medianoche	6	Medianoche	6

Utilice integración numérica y la ecuación *media* $= \dfrac{\int_a^b f(x)\,dx}{b-a}$ para determinar el número total de carros que pasa por día en cada punto.

Referencias bibliográficas:

1. *Allen Smith W.*. "Análisis Numérico". México. Prentice-Hall Hispanoamericana, 1988.

2. *Chapra, Steven C; Canale, Raymond P.* "Métodos Numéricos para Ingenieros". México. McGraw-Hill, 1987.

BIBLIOGRAFÍA

A

Acton, Forman S. **Numerical Methods that Work.** New York. Harper & Row, Publishers. 1970.

Atkinson, Kendall E. **An Introduction to Numerical Analysis.** New York. John Wiley & Sons. 1989.

Atkinson, Kendall. **Elementary Numerical Analysis.** New York. John Wiley & Sons. 1985.

Atkinson/Harley. **Introducción a los Métodos Numéricos con Pascal.** Addison-Wesley Iberoamericana. Sheffield, EE.UU. 1987.

B

Barnett, Stephen. **Matrix Methods for Engineers and Scientists.** London. McGraw-Hill. 1979.

Burden, Richard L.; Faires, J. Douglas. **Análisis Numérico.** México. Grupo Editorial Iberoamerica. 1985.

C

Conte, S. D.; de Boor, Carl. **Análisis Numérico.** México. McGraw-Hill. 1974.

Conte, Samuel D.; de Boor Carl. **Elementary Numerical Analysis. An Algorithmic Approach.** Tokyo. McGraw-Hill Kogakusha, LTD. 1980.

CH

Chapra, Steven C.; Canale, Raymond P. **Métodos Numéricos para Ingenieros.** México. McGraw-Hill. 1987.

Cheney, Ward; Kincadi, David. **Numerical Mathematics and Computing.** Monterey, California. Brooks/Cole Publishing Company. 1980.

D

Dorn, William S.; Greenberg, Herbert J. **Matemáticas y Computación: con programación FORTRAN.** México. Editorial Limusa-Wiley, S. A. 1970.

E

Ellis, L. E.; Goldstein G.; Tinsley, J. D. **Computers and the Teaching of Numerical Mathematics in the Upper Secondary School.** Gran Bretaña. G. Bell & Sons LTD. 1971.

F

Forsythe, George E.; Moler, Cleve B. **Solución Mediante Computadoras de Sistemas Algebráicos Lineales.** Argentina. Editorial Universitaria de Buenos Aires. 1973.

G

Goldstine, Herman H. A **History of Numerical Analysis from the 16th through the 19th Century.** New York. Springer-Verlag. 1977.

H

Hamming, R. W. **Numerical Methods for Scientists and Engineers.** New York. McGraw-Hill Book Company, Inc. 1962.

Henrici, Peter. **Elementos de Análisis Numérico.** México. Editorial Trillas. 1972.

Henrici, Peter. **Essentials of Numerical Analysis with Pocket Calculator Demonstrations.** New York. John Wiley & Sons. 1982.

K

Kahaner, David; Moler, Cleve; Nash, Stephen. **Numerical Methods and Software.** New Jersey. Prentice Hall. 1989.

King, Thomas J. **Introduction to Numerical Computation.** New York. McGraw-Hill Book Company. 1984.

Kuo, Shan S. **Computer Applications of Numerical Methods.** Reading, Massachusetts. Addison-Wesley Publishing Company. 1972.

L

Leinbach, L. Carl. **Calculus with the Computer.** New Jersey. Prentice-Hall, Inc. 1974.

Luthe, Rodolfo; Olivera, Antonio; Schutz, Fernando. **Métodos Numéricos.** México. Limusa. 1988.

M

Maron, M. J. **Numerical Analysis. A Practical Approach.** New York. Macmillan Publishing Company. 1987

McCalla, Thomas Richard. **Introduction to Numerical Methods and FORTRAN Programming.** New Yor. John Wiley & Sons, Inc. 1967.

Metropolis, N.; Howlett, J.; Rota Gian-Carlo. **A History of Computing in the Twentieth Century.** New York. Academic Press. 1980.

P

Pizer, Stephen M. **Numerical Computing and Mathematical Analysis.** Chicago. Sciencie Research Associates, Inc. 1975.

Press, William H.; Flannery, Brian P.; Teukolsky, Saul A.; Vetterling, William T. **Numerical Recipes.** New York. Cambridge University Press. 1986.

R

Ralston, Anthony. **Introducción al Análisis Numérico.** México. Editorial Limusa-Wiley, S. A. 1970.

Rice, John R. **Numerical Methods, Software, and Analysis.** Auckland. McGraw-Hill. 1983.

S

Scheid, Francis. **Análisis Numérico.** México. McGraw-Hill. 1972.

Scraton, R. E. **Métodos Numéricos Básicos.** México. McGraw-Hill. 1987.

Shampine, Lawrence F.; Allen Jr., Richard C. **Numerical Computing: an introduction.** Philadelphia. W. B. Saunders Company. 1973.

Shampine, Lawrence; Pruess, Steven; Allen, Richard. **Fundamentals of Numerical Computing.** México. Publicaciones del Departamento de Matemáticas de la Facultad de Ciencias de la UNAM. 1989.

Smith, W. Allen. **Análisis Numérico.** México. Prentice-Hall Hispanoamericana, S.A. 1988.

Stoer, J; Bulirsch, R. **Introduction to Numerical Analysis.** New York. Springer-Verlag. 1980.

V

Vandergraft, James S. **Introduction to Numerical Computations.** New York. Academic Press. 1978.

Volkov, E. A. **Métodos Numéricos.** URSS. Editorial Mir Moscú. 1990.

Y

Yakowitz, Sidney; Szidarovszky, Ferenc. **An Introduction to Numerical Computations.** New York. Macmillan Publishing Company. 1990.