CAPÍTULO 1 Matemática Numérica, 2da Edición
Manuel Álvarez, Alfredo Guerra, Rogelio Lau

CONCEPTOS INICIALES

Objetivos

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

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

1.1 Introducción

¿Qué es la Matemática Numérica?

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

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

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

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

Una breve historia

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

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

es decir:

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

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

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

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

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

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

1.2 Fuentes de error en la solución de un problema

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

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

La modelación y los errores de modelación

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

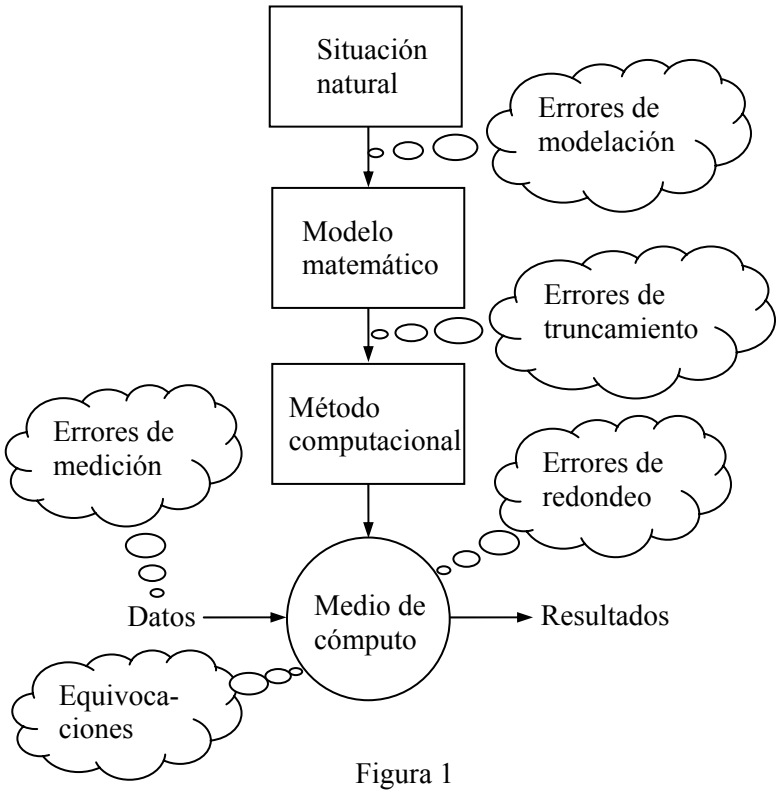

Figura 1

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

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

Métodos computacionales y errores de truncamiento

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

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

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

Errores en el proceso de cálculo

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

- De medición u observación

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

- Equivocaciones

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

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

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

- Errores de redondeo

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

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

Para ilustrar los conceptos anteriores, considérese el siguiente ejemplo

Ejemplo 1

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

Solución:

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

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

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

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

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

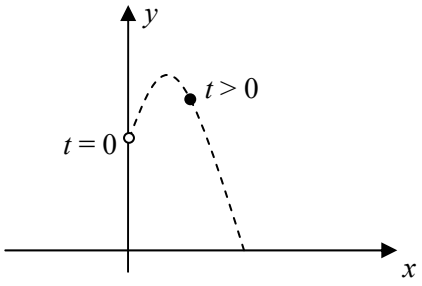

Figura 2

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

$$F_x = ma_x \qquad (1)$$
$$F_y = ma_y \qquad (2)$$

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

$$ma_x = 0 \qquad (3)$$
$$ma_y = -mg \qquad (4)$$

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

$$a_x = 0 \qquad (5)$$
$$a_y = -g \qquad (6)$$

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

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

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

donde $t \geq 0$.

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

$$x(0) = 0 \qquad (9)$$
$$y(0) = h \qquad (10)$$
$$v_x(0) = v_{0x} \qquad (11)$$
$$v_y(0) = v_{0y} \qquad (12)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Es decir:

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

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

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

Sustituyendo en la integral (13):

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

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

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

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

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

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

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

Ejercicios

1. En un relato breve debido al escritor argentino Jorge Luis Borges, se cuenta de un imaginario país en que los cartógrafos eran tan meticulosos y fieles a los detalles que para hacer el mapa de la nación habían necesitado todo un estado. Analice cómo se relaciona esta historia con el arte de modelar y los errores de modelación.

2. En el ejemplo 1 se hicieron varias hipótesis respecto al problema del lanzamiento de una partícula. Si se quisiera acercar un poco más el modelo a la realidad, analice que hipótesis podría ser modificada razonablemente.

3. Aplique el modelo matemático elaborado en el ejemplo 1, a la caída de una gota de lluvia desde una nube que se encuentra a 5 000 metros de altura. Según este modelo, calcule la velocidad (en kilómetros por hora) de la gota de lluvia al llegar a la tierra. Suponga que la velocidad inicial de la gota es cero y todas las demás hipótesis del modelo. Analice si el resultado obtenido es razonable e indique qué hipótesis deberían ser modificadas para que el modelo se pueda utilizar en este caso.

4. La barras de acero para la construcción se fabrican en diámetros estándar de ¼", ⅜", ½", ⅝", ¾", etc. En un modelo destinado a calcular el diámetro de las barras de acero que se colocarán en un elemento de hormigón armado, analice la magnitud de los errores que se deben tolerar.

5. Las resistencias eléctricas y los capacitores que se utilizan en el trabajo con circuitos electrónicos corrientes, se fabrican con errores de hasta 10%. En la modelación de circuitos electrónicos con vistas al diseño, analice la magnitud de los errores que se deben permitir.

6. En el diseño mecánico es usual trabajar con láminas de acero y tornillos. Estos elementos se fabrican industrialmente en espesores y diámetros discretos como 3mm, 4mm, 5mm, 6mm, etc. Analice con qué errores se puede realizar la modelación destinada al diseño de elementos mecánicos que utilizan elementos de este tipo.

7. En la pantalla de un monitor de una computadora personal los textos y los gráficos que aparecen se producen mediante puntos de colores llamados pixels. En la actualidad, la cantidad de pixels es del orden de 1000 en sentido horizontal y un poco menos en el sentido vertical. Si se está modelando un proceso con el objetivo final de mostrar geométricamente en una pantalla el resultado, analice la magnitud de los errores numéricos que se pueden permitir.

8. Los ingenieros hidráulicos realizan diseños con sistemas de tuberías. Si se tiene en cuenta que las tuberías se venden con diámetros interiores muy específicos como: ⅜", ½", ¾", 1", 1¼", 1½", 2", etc., analice que precisión se necesita en un modelo destinado a determinar el diámetro de un sistema complejo de tuberías.

1.3 Medidas del error

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

Definición 1

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

$$error(x) = x^* - x$$

■

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

Definición 2

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

$$E(x) = |error(x)|$$ ■

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

Definición 3

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

$$e(x) = \frac{E(x)}{|x^*|}$$ ■

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

Ejemplo 1

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

Solución:

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

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

b) Aquí se tiene $x^* = 11{,}5$ volt y $x = 11{,}6$ volt. Por tanto:

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

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

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

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

Definición 4

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

$$E_m(x) \geq E(x)$$ ■

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

Algo similar puede hacerse con los errores relativos:

Definición 5

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

$$e_m(x) \geq e(x)$$ ■

En la tabla 1 se resumen las cinco definiciones anteriores.

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

Tabla 1

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

El mínimo error absoluto máximo

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

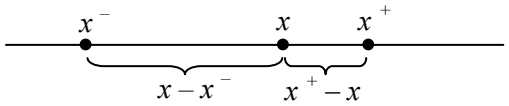

Figura 1

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

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

Si se toma $$x = \frac{x^- + x^+}{2}$$

entonces se minimiza el error absoluto máximo, el cual será

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

Ejemplo 2

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

Solución:

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

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

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

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

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

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

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

Relación entre $E_m(x)$ y $e_m(x)$

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

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

se tiene que:

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

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

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

Similarmente, la igualdad (1) se puede escribir:

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

de donde resulta:

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

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

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

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

Ejemplo 3

Arquímedes obtuvo para el número π la desigualdad:

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

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

Solución:

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

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

El error absoluto máximo viene dado por:

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

Para hallar el error relativo máximo se usará la relación:

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

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

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

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

Hallando aproximaciones por defecto y por exceso

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

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

Se tiene que: $$-E_m(x) \leq x^* - x \leq E_m(x)$$

y, sumando *x* en los tres miembros: $$x - E_m(x) \leq x^* \leq x + E_m(x) \qquad (4)$$

Es decir, $$x^- = x - E_m(x) \qquad \text{y} \qquad x^+ = x + E_m(x)$$

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

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

Ejemplo **4**

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

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

Solución:

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

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

Por tanto, según (4): $R - E_m(R) \le R^* \le R + E_m(R)$
Es decir: $44{,}65\text{k}\Omega \le R^* \le 49{,}35\text{k}\Omega$

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

Ejercicios

1. Calcule los errores absolutos y relativos que se cometen al aproximar las siguientes constantes matemáticas y físicas por los valores indicados a su derecha.
   a) $e = 2{,}7182818...$ (base de los logaritmos neperianos) $e_A = 2{,}7$
   b) $c = 2{,}99793 \cdot 10^8$ m/s (velocidad de la luz en el vació) $c_A = 3 \cdot 10^8$ m/s
   c) $g = 9{,}8$ m/s^2 (aceleración de la gravedad) $g_A = 10$ m/s^2
   d) $C = 0{,}577216$ (constante de Euler) $C_A = 0{,}58$
   e) $m_p = 1{,}67239 \cdot 10^{-24}$ g (masa del protón) $m_{pA} = 1{,}7 \cdot 10^{-24}$ g
   f) $m_e = 9{,}1983 \cdot 10^{-28}$ g (masa del electrón) $m_{eA} = 10^{-27}$ g

2. Suponga que usted no conoce el valor de $\sqrt{2}$. Como $1{,}4^2 = 1{,}96 < 2$ y $1{,}5^2 = 2{,}25 > 2$, se puede asegurar que $1{,}4 < \sqrt{2} < 1{,}5$. A partir de esta conclusión obtenga una aproximación para $\sqrt{2}$ que tenga el mínimo error absoluto máximo. Halle también el error relativo máximo.

3. Si la aproximación que usted halló en el ejercicio anterior se eleva al cuadrado, se puede saber si ella es menor o mayor que $\sqrt{2}$. De ese modo usted puede determinar un intervalo de menor amplitud que el ofrecido en ese ejercicio, donde se encuentre $\sqrt{2}$. A partir de este nuevo intervalo se puede encontrar una nueva aproximación y su error absoluto máximo. Siguiendo esta idea, calcule una aproximación para $\sqrt{2}$ que posea un error absoluto menor que 0,001.

4. En el capacitor de arranque de un motor eléctrico aparece su capacidad como: 32 ± 3 μF. Determine el error absoluto máximo y el error relativo máximo del valor nominal de 32 μF.

5. Se quiere calcular la distancia entre dos puntos de un territorio a partir de un mapa de escala 1 Km = 1 cm. Midiendo con una cinta métrica se obtuvo una distancia de 18,3 cm. Si los editores del mapa garantizan un error relativo menor que 1% en la confección del mapa y el error en la medición pudiera haber sido hasta de 1mm, calcule entre qué valores debe hallarse la distancia real que se busca.

6. En un programa que produce una gráfica sobre la pantalla de una computadora, se calculan las coordenadas (x, y) de un punto del display. Estas coordenadas son números reales, pero para dibujarlas en la pantalla primero hay que redondearlas al valor entero más cercano obteniendo (x_p, y_p) donde x_p y y_p son pixels. Si se está utilizando una resolución de 1024 por 768 pixels y la pantalla mide 28 cm de ancho y 21 cm de altura, halle el error absoluto máximo que se produce en dirección vertical y en dirección horizontal cuando el punto (x, y) se representa en el display.

7. Se mide el voltaje en un circuito eléctrico con un voltímetro, cuyo fabricante garantiza un error relativo máximo de 0,1 %. Si el voltaje medido es de 225 v, determine el error absoluto máximo de la medición realizada y halle un intervalo donde se encuentra el verdadero valor con toda seguridad.

8. A veces se utiliza la aproximación $\text{sen } x \approx x$ para valores pequeños de x. Grafique (mejor si usa algún programa para realizar la gráfica) las funciones $y^* = \text{sen } x$ y $y = x$ en un mismo sistema coordenado y compárelas. Determine el máximo valor x_{max} que puede tomar x de modo que el error absoluto que se comete en la aproximación sea menor que 0,001.

9. Una mejor aproximación que en el ejercicio anterior se alcanza si se toma $\text{sen } x \approx x - \frac{x^3}{6}$.
   Repita en este caso el enunciado del ejercicio anterior.

10. En la antigua babilonia se conocía una forma para hallar aproximadamente la raíz cuadrada de un número que fuera cercano a un cuadrado perfecto. En la notación actual, se escribiría así:
$$\sqrt{a^2 + x} \approx a + \frac{x}{2a}$$
   Halle el error absoluto y el error relativo al calcular $\sqrt{10}$ por esta vía.

11. En el papiro de Rihn aparece una fórmula para calcular aproximadamente el área de un círculo como la de un cuadrado cuyo lado fuera 8/9 del diámetro del círculo. Demuestre que esto equivale a tomar para π la aproximación 256/81. Calcule el error absoluto y el error relativo de esta aproximación.

12. De un sobre con resistencias eléctricas corrientes (10% de error relativo máximo) correspondientes a un valor nominal de 56 KΩ se midieron algunos ejemplares y se encontró una resistencia de 50 KΩ y otra de 60 KΩ. Determine si alguna de ellas constituye una equivocación del fabricante.

1.4 Cifras significativas y cifras exactas

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

Cifras significativas de un número

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

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

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

Definición 1

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

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

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

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

Ejemplo 1

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

Solución:

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

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

■

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

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

Definición 2

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

Ejemplo 2

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

Solución:

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

$$0{,}000\underline{2030}$$

■

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

Ejemplo 3

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

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

Solución:

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

La notación científica

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

Ejemplo 4

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

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

Solución:

Número de Avogadro: $N = 602\ 497\ 000\ 000\ 000\ 000\ 000\ 000\ \text{mol}^{-1}$
Masa del electrón: $m_e = 0{,}000\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 910\ 83\ \text{g}$
Velocidad de la luz en el vacío $c_0 = 29\ 979\ 300\ 000\ \text{cm/s}$

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

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

Cifra exacta

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

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

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

Ejemplo 5

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

a) $x = 3{,}1416$. Se sabe que $x^* = \pi = 3{,}141592653\ldots$
b) $x = 3{,}99999$. Se sabe que $x^* = 4$
c) $x = 4{,}20457$. Se sabe que $x^* = 4{,}20451$
d) $x = 0{,}00046384$. Se sabe que $E(x) = 0{,}000002$
e) $x = 23{,}01241$. Se sabe que $E_m(x) = 0{,}04$

Solución:

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

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

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

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

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

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

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

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

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

Como los demás dígitos de *x* poseen mayor valor posicional, ellos también serán exactos. En resumen, en este caso las cifras exactas de *x* son las que aparecen subrayadas a continuación: 3,99999. El último 9 no es una cifra exacta.

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

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

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

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

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

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

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

d) En este caso se ilustra el hecho de que no es necesario conocer el valor exacto *x** para determinar los dígitos exactos de *x*, basta conocer el error absoluto. Como el error contiene un 2 en la sexta cifra decimal, está claro que la sexta cifra de *x* no puede ser exacta. Considérese el 6 que se halla en el quinto lugar decimal, su valor posicional es

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

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

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

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

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

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

Contando las cifras exactas

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

Definición 3

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

Ejemplo 6

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

a) $x = 3{,}1416$
b) $x = 3{,}99999$
c) $x = 4{,}20457$
d) $x = 0{,}00046384$
e) $x = 23{,}01241$

Solución:

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

El redondeo

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

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

Ejemplo 7

Considérese el número exacto

$$x^* = e = 2{,}718\ 281\ 828\ 459\ 045\ldots$$

y el valor aproximado

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

El error absoluto es

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

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

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

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

$$x = \underline{2{,}718286}325411$$

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

$$x_1 = 2{,}71829$$

El error absoluto de x_1 es

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

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

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

Cifras decimales exactas y error absoluto

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

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

Cifras exactas y error relativo

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Ejemplo 13

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

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

Solución:

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

$$e_m(x) = \frac{E_m(x)}{|x|} = \frac{50}{673500} = 0{,}000074$$

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

$$e_m(x) = \frac{E_m(x)}{|x|} = \frac{0{,}005}{67{,}35} = 0{,}000074$$

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

$$e_m(x) = \frac{E_m(x)}{|x|} = \frac{0{,}5 \cdot 10^{-7}}{0{,}0006735} = 0{,}000074$$

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

x	$E_m(x)$	$e_m(x)$
673 500	50	0,000074
67,35	0,005	0,000074
0,0006735	$0{,}5\cdot10^{-7}$	0,000074

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

Ejercicios

1. Diga qué dígitos de los números que siguen no son significativos: a) 0,00048003; b) 12004; c) 4,54600.

2. A continuación aparecen, dentro de determinados contextos, números enteros con varios ceros finales. Determine, si fuera posible, a partir del contexto, cuales de dichos ceros son significativos y cuales no.
   a) De un folleto de Datos sobre Cuba: "... Su longitud es de 1250 kilómetros y sus costas suman alrededor de 7000 kilómetros..."
   b) De la leyenda de un mapa: Escala 1:30 000 000.
   c) De un plegable sobre la ciudad de Baracoa: "El municipio tiene una población de 80000 habitantes, de ella 35000 en la ciudad cabecera".
   d) De una revista turística: "En los alrededores de la ciudad de Sucre (Bolivia) se encuentra la mayor cantidad (5000) de huellas de dinosaurios del mundo"
   e) El año 2000 se considera el último del siglo XX.

3. En los siguientes casos se dan números x aproximados a números exactos x^*. Determine cuales cifras de x son exactas y diga la cantidad de cifras exactas y de cifras decimales exactas de x.
   a) $x^* = 2{,}718281...$ $x = 2{,}71$
   b) $x^* = 0{,}000436$ $x = 0{,}00044$
   c) $x^* = 236\ 578$ $x = 236\ 000$
   d) $x^* = 0{,}066666...$ $x = 0{,}067$
   e) $x^* = \pi$ $x = 3{,}14$

4. En los siguientes casos se dan números aproximados y alguna información sobre su error. Determine en cada caso la cantidad de cifras exactas y de cifras decimales exactas del número aproximado. Diga qué cifras son dudosas.
   a) $x = 5{,}839\ 362$ $E(x) = 0{,}0002$
   b) $x = 0{,}0045387$ $e(x) = 0{,}03\ \%$
   c) $x = 7{,}8876\cdot10^{-5}$ $E(x) = 4\cdot10^{-7}$
   d) $x = 54{,}831$ $e_m(x) = 0{,}001$
   e) $x = 45\ 846\ 352$ $E_m(x) = 300$
   f) $x = 33\ 786$ $e(x) = 0{,}002$
   g) $x = 0{,}055763$ $E_m(x) = 0{,}00005$
   h) $x = 0{,}0000785$ $e_m(x) = 0{,}1\ \%$

5. A continuación aparecen números aproximados con su cantidad de cifras exactas o de cifras decimales exactas. Determine en cada caso el error absoluto máximo y el error relativo máximo del número y diga qué cifras son dudosas.

a) $x = 697{,}6587$ con 2 cifras decimales exactas
b) $x = 50{,}006$ con 4 cifras exactas
c) $x = 0{,}0054768$ con 3 cifras exactas
d) $x = 0{,}0005973$ con 5 cifras decimales exactas
e) $x = 4{,}99999$ con 3 cifras decimales exactas
f) $x = 390\ 785$ con 4 cifras exactas
g) $x = 0{,}09878$ con 2 cifras exactas
h) $x = 0{,}00003543$ con 6 cifras decimales exactas

6. Redondee los números que aparecen a continuación de acuerdo con las especificaciones que se indican. Determine en cada caso el error absoluto y el error relativo introducido por el redondeo.
   a) $x = 58{,}54654$ Conservar 5 cifras significativas
   b) $x = 0{,}045365$ Conservar 4 cifras decimales
   c) $x = 6{,}549873$ Conservar hasta las milésimas
   d) $x = 67\ 845{,}675$ Conservar hasta las centenas
   e) $x = 0{,}00657873$ Redondear a partir de las diezmilésimas

1.5 Los números en la computadora

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

Propiedad 1

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

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

Representación de números enteros

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

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

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

Tabla 1

Representación de números reales

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

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

Entonces, para almacenar el número x, se utiliza:

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

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

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

Propiedad 2

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

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

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

Propiedad 3

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

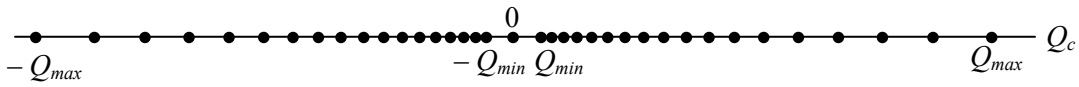

Figura 1

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

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

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

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

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

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

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

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

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

Recomendaciones

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

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

  donde ε es un número pequeño, pero grande en comparación con los errores de redondeo que pueda haber introducido la imprecisión de la máquina.
- Piense siempre que los números que almacena la máquina *no son los mismos* con los que trabaja su algoritmo manual sino aproximaciones de aquellos, aun cuando usted no haya realizado ninguna operación aritmética con ellos.

1.6 Propagación del error

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

Una ley general

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

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

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

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

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

El error en el resultado es la diferencia:

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

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

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

multiplicando ambos miembros de la igualdad por –1:

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

que se puede escribir en términos de errores como:

$$error(R) = f_x(x^*,y^*)error(x) + f_y(x^*,y^*)error(y) \qquad (1)$$

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

$$error(R) = f_x(x,y)error(x) + f_y(x,y)error(y)$$

Si se toma valor absoluto en cada miembro queda:

$$\left|error(R)\right| = \left|f_x(x,y)error(x) + f_y(x,y)error(y)\right|$$

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

$$\left|error(R)\right| \le \left|f_x(x,y)\right| \cdot \left|error(x)\right| + \left|f_y(x,y)\right| \cdot \left|error(y)\right|$$

Es decir:

$$E(R) \le \left|f_x(x,y)\right| \cdot E(x) + \left|f_y(x,y)\right| \cdot E(y)$$

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

$$E(R) \le \left|f_x(x,y)\right| \cdot E_m(x) + \left|f_y(x,y)\right| \cdot E_m(y)$$

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

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

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

Propagación del error en sumas y diferencias

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

$$R = x + y$$

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

$$E_m(R) = E_m(x) + E_m(y) \tag{3}$$

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

$$E_m(x \pm y) = E_m(x) + E_m(y) \tag{4}$$

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

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

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

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

Ejemplo **1**

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

Solución:

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

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

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

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

Según el enunciado: $E_m(c) = 0{,}002$

Por tanto: $S = 121{,}373$

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

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

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

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

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

Redondeando: $e_m(S) = 0{,}01 = 1\%$

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

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

Por tanto: $120{,}173 \leq S \leq 122{,}573$

■

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

Ejemplo 2

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

Solución

Como $P_e = \pi D_e$

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

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

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

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

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

El grosor aproximado x de la pared será:

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

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

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

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

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

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

Propagación del error en el producto

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

$$R = xy$$

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

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

$$E_m(R) = \left|y\right| \cdot E_m(x) + \left|x\right| \cdot E_m(y) \qquad (6)$$

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

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

Es decir: $e_m(R) = e_m(x) + e_m(y)$

En resumen, $e_m(xy) = e_m(x) + e_m(y)$ (7)

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

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

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

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

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

Si k es exacto: $e_m(kx) = e_m(x)$ (9)

Ejemplo 3

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

Solución:

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

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

$$E_m(G) = G \cdot e_m(G) = 288000 \cdot 0{,}21 = 60480 \text{ USD}$$

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

Propagación del error en el cociente

En este caso, será: $R = \dfrac{x}{y}$

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

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

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

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

La fórmula anterior es poco adecuada por su complejidad. Si en ambos miembros se divide por $|R| = \dfrac{|x|}{|y|}$ se obtiene:

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

Es decir:

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

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

Ejemplo 4

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

Solución:

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

Valor	**Error absoluto máximo**		**Error relativo máximo**
x = 51,254	0,0005		
y = 23,978	0,0005		
x – y = 27,276	0,001	$\rightarrow$	$\dfrac{0{,}001}{27{,}276} = 0{,}000037$
x + y = 75,232	0,001	$\rightarrow$	$\dfrac{0{,}001}{75{,}232} = 0{,}000014$
$w = \dfrac{x-y}{x+y} = \dfrac{27{,}276}{75{,}232}$	(0,3625585)(0,000051) =	$\leftarrow$	0,000037 + 0,000014 = 0,000051
= 0.3625585	0,000019		

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

Otro modo de solución:

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

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

Utilizando esta igualdad:

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

De donde:

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

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

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

Propagación del error en la potencia y la exponencial

Como último caso particular, considérese que

$$R = x^y$$

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

En este caso la fórmula (2):

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

se convierte en:

$$E_m(R) = \left|y\cdot x^{y-1}\right|\cdot E_m(x) + \left|x^y \ln x\right|\cdot E_m(y) \qquad (11)$$

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

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

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

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

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

O sea: $$e_m(x^k) = |k| e_m(x) \qquad (12)$$

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

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

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

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

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

Es decir: $$e_m(b^y) = \left| \ln b \right| \cdot E_m(y) \qquad (13)$$

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

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

Ejemplo 5

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

Solución:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

■

Ejercicios

1. Si $x = 23{,}76$; $y = 45{,}74$ y $z = 65{,}272$ todos con error relativo máximo de 0,5 %, halle el valor de $S = x - y + z$, su error relativo máximo y la cantidad de cifras exactas que posee.

2. Se sabe que $L = xy - uz$ y se conocen valores aproximados $x = 0{,}5487$; $y = 6{,}7855$; $z = 0{,}07824$; $u = 2{,}76803$ todos con 3 cifras decimales exactas. Calcule el valor de *L* y determine cuántas cifras decimales exactas posee este resultado.

3. Se tiene la ecuación de segundo grado $4{,}3276x^2 - 9{,}776x + 1{,}8655 = 0$ y se sabe que los coeficientes están calculados con 4 cifras exactas. Determine el valor de la mayor de las raíces, aplicando la fórmula general para la ecuación de segundo grado, y diga cuales de sus cifras son exactas.

4. Se conocen aproximadamente los tres lados de un triángulo: $a = 435{,}87$ (error relativo menor que 0,0002) $b = 355{,}75$ (con todas sus cifras exactas) y $c = 532{,}31$ (con error menor que 0,1%). Calcule el área del triangulo mediante la fórmula $S = \sqrt{p(p-a)(p-b)(p-c)}$ donde *p* es el semiperímetro del triangulo y determine el error relativo máximo del resultado.

5. Calcule el valor de $\dfrac{34{,}785 - 22{,}873}{65{,}872 - 12{,}543}$ si todos los números son aproximados y poseen cuatro cifras exactas. Halle las cifras exactas que posee el resultado.

6. Para calcular el volumen de un cilindro se miden el radio y la altura. Se obtienen las mediciones: $r = 45{,}6 \pm 1$ cm y $h = 142{,}7 \pm 2$. Determine el volumen, su error absoluto

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

7. Suponiendo la tierra como una esfera de unos 6400 km de radio, proponga cuántas cifras exactas tomar de π y del radio para calcular el volumen de la tierra con un error menor que 1%.

8. Para calcular el valor de $e^{-\pi}$ se utilizarán los primeros 8 términos de la serie de Maclaurin para la exponencial: $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$ Si el valor de π se toma con cuatro cifras decimales exactas, calcule $e^{-\pi}$, el error de truncamiento, el error debido al redondeo y el error total del resultado obtenido.

9. Se conoce que $x = 4{,}7684$ con todas sus cifras exactas. Determine el error absoluto máximo que se comete si se calcula y como:

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

1.7 Errores e inestabilidad

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

Problemas estables y problemas inestables

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

Ejemplo **1**

Sea la ecuación algebraica de grado 10:

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

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

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

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

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

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

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

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

Ejemplo 2

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

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

cuya solución puede fácilmente comprobarse que es $x = 2$ y $y = 1$. En este problema, los datos son los 6 coeficientes de las ecuaciones y el resultado los valores de x y y que forman la solución. Nótese cómo el simple cambio del coeficiente 3,01 por 3,02 (un cambio inferior al 1%) transforma el sistema en

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

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

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

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

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

Pérdida de significación

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

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

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

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

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

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

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

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

Ejemplo **3**

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

Solución:

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

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

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

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

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

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

■

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

Ejemplo 4

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

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

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

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

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

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

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

```
   1,000 000 0
 – 9,000 000 0
  40,500 000
– 121,500 00
  273,375 00
– 492,075 00
  738,112 50
– 949,001 79
 1067,627 0
– 1067,627 0
  960,364 31
```

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

Métodos inestables para problemas estables

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

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

Ejemplo **5**

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

Solución 1

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

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

Trabajando con 5 cifras exactas, se obtiene para x_1:

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

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

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

Solución 2

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

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

que posee 5 cifras exactas. ■

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

Ejemplo **6**

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

Solución 1:

Primero el problema se generaliza de la siguiente manera:

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

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

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

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

para $$I_n = \int_0^1 x^n e^{x-1} dx = x^n e^{x-1}\Big|_0^1 - \int_0^1 nx^{n-1} e^{x-1} dx$$

$$I_n = \int_0^1 x^n e^{x-1} dx = 1 - n\int_0^1 x^{n-1} e^{x-1} dx$$

Esto es: $$I_n = 1 - nI_{n-1} \qquad \text{para } n = 2, 3, 4,... \qquad (3)$$

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

$$I_9 = -\ 0{,}06848$$

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

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

Solución 2:

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

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

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

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

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

que tiene todas sus cifras exactas.

Ejercicios

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

Ecuación:	Raíz positiva:
sen $x = 0{,}999x$	$x = 0{,}0775$
sen $x = 0{,}998x$	$x = 0{,}1096$
sen $x = 0{,}997x$	$x = 0{,}1342$
sen $x = 0{,}996x$	$x = 0{,}1550$
sen $x = 0{,}995x$	$x = 0{,}1733$

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

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

Sistema original:	Sistema modificado:
$\begin{cases} 30x_1 + 15x_2 + 10x_3 = 55 \\ 15x_1 + 10x_2 + 7.5x_3 = 32.5 \\ 10x_1 + 7.5x_2 + 6x_3 = 23.5 \end{cases}$	$\begin{cases} 30x_1 + 15x_2 + 10x_3 = 55 \\ 15x_1 + 10x_2 + 7.5x_3 = 32.5 \\ 10x_1 + 7.5x_2 + 6x_3 = 23.4 \end{cases}$
Solución: $x_1 = 1$, $x_2 = 1$, $x_3 = 1$	Solución: $x_1 = 0{,}9$, $x_1 = 1{,}6$, $x_1 = 0{,}4$

3. La función $f(x) = \dfrac{x\cos x - \operatorname{sen} x}{x^3}$ no está definida para $x = 0$ pero debe tender a $-\frac{1}{3}$ cuando x tiende hacia cero. Evalúela para valores cercanos a $x = 0$, mediante su calculadora o con un asistente matemático, y observe lo que ocurre cuando le asigna a x valores del orden de 0,00001. Explique el comportamiento que observe.

4. Exprese la función $f(x)$ del ejercicio anterior de una forma más conveniente para valores muy pequeños de x. (Sugerencia: exprese las funciones sen x y cos x mediante sus series de Maclaurin y simplifique la expresión teniendo en cuenta que x tomará valores muy pequeños.

5. Muestre que la función $g(x) = \dfrac{\ln(1-x) + xe^{\frac{x}{2}}}{x^3}$ presenta pérdida de significación cuando se evalúa para valores de x cercanos a cero, a pesar de que aparentemente, no se presentan diferencias de números parecidos. Proponga una forma alternativa estable para evaluar la función para valores de x próximos a cero.

6. Para trazar en el display de una computadora personal la recta tangente a la grafica de la función $y = x^7$ en el punto de abscisa $x = 4$, se calcula aproximadamente el valor de la pendiente de la recta como:

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

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

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

1.8 Seudo código para la escritura de algoritmos

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

El operador de asignación

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

La estructura alternativa

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

If ⟨condición⟩ **then**
    ⟨secuencia 1 de acciones⟩
else
    ⟨secuencia 2 de acciones⟩
end

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

Estructuras repetitivas

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

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

do while ⟨condición⟩
    ⟨secuencia de acciones⟩
end

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

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

repeat
    ⟨secuencia de acciones⟩
until ⟨condición⟩

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

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

La estructura **for** tiene la forma:

for ⟨variable⟩ = ⟨valor inicial⟩ **to** ⟨valor final⟩
    ⟨secuencia de acciones⟩
end

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

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

Ejemplo 1

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

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

utilizando la fórmula general.

Solución:

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

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

El algoritmo de cálculo será como sigue:

Se supone conocidos los coeficientes de la ecuación: *a*, *b* y *c*
if $a \neq 0$ **then**
    $D := b^2 - 4ac$
    if $D \geq 0$ **then**
        $x_1 := \frac{-b + \sqrt{D}}{2a}$
        $x_2 := \frac{-b - \sqrt{D}}{2a}$
    else
        No hay raíces reales
    end
else
    La ecuación no es de segundo grado
end

Ejemplo 2

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

Solución:

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

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

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

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

El algoritmo de cálculo adoptará la forma:

Se supone conocidos la función f y los números a, b y n
$h := \frac{b-a}{n-1}$
for $i = 1$ **to** n
    $x_i := a + (i-1)h$
    $y_i := f(x_i)$
end

Ejemplo 3

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

Solución:

Los elementos de la progresión se obtendrán mediante la fórmula:

$$x_k = x_{k-1} + d \quad \text{para } k = 2, 3, 4,...$$
$$\text{con} \quad x_1 = a$$

El algoritmo de cálculo puede expresarse así:

Se supone conocidos a, d y x_{max}
$k := 1$
$x_1 := a$
do while $x_k \le x_{max} - d$
    $k := k + 1$
    $x_k = x_{k-1} + d$
end

Ejemplo 4

Una cierta sucesión convergente está definida mediante la fórmula recurrente:

$$x_0 = a$$
$$x_i = \phi(x_{i-1}) \quad \text{para} \quad i = 1, 2, 3,...$$

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

Solución

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

Se supone conocidos la función ϕ y los números a y ε
$x_{anterior} := a$
repeat
    $x_{actual} := \phi(x_{anterior})$
    $Diferencia := |x_{actual} - x_{anterior}|$
    $x_{anterior} := x_{actual}$
until $Diferencia \leq \varepsilon$
$L := x_{actual}$

Ejercicios

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

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

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

2. Obtener todos los términos x_n de una progresión geométrica de razón $r > 0$ y primer término $a > 0$ que sean menores que un valor $M > a$. Los valores de a, r y M son conocidos.

3. Evaluar un polinomio $p(x)$ de grado n escrito en la forma $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$. Los coeficientes del polinomio, el grado n y el valor de x son conocidos.

4. Repita el ejercicio 3 pero ahora con el polinomio escrito en la forma de Horner:

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

   Analice la ventaja de efectuar las operaciones de esta manera.

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

6. Multiplicar una matriz cuadrada $A = \{a_{ij}\}$ de orden n por una matriz columna $B = \{b_i\}$ de orden n. Tanto n como las componentes de las matrices son conocidos.

7. Dadas dos matrices $A = \{a_{ij}\}$ de orden $p \times n$ y $B = \{b_{ij}\}$ de orden $q \times m$, determinar si son conformes para el producto. En caso negativo escribir: "Operación imposible" y en caso positivo hallar la matriz $A \cdot B$. Los números p, n, q, m y los elementos de ambas matrices son conocidos.

Otras lecturas recomendadas

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

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

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

Principales ideas del capítulo

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

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

- Los números reales que se encuentran en el rango permisible, es decir, que son cero o están en $[-Q_{max}, -Q_{min}]$ o en $[Q_{min}, Q_{max}]$ pueden ser tratados por la computadora, aunque, en la mayoría de los casos, sufrirán una aproximación para sustituirlos por elementos de Q_c
- En el trabajo con números reales en una máquina, en lugar de verificar si $x = y$, verifique si $|x - y| < \varepsilon$ donde ε es un número pequeño, pero grande en comparación con los errores de redondeo que pueda haber introducido la imprecisión de la máquina.
- Si $R = f(x, y)$ entonces $E_m(R) = \left|f_x(x,y)\right| \cdot E_m(x) + \left|f_y(x,y)\right| \cdot E_m(y)$
- Las leyes principales que rigen la propagación de errores son: $E_m(x \pm y) = E_m(x) + E_m(y)$

  $e_m(xy) = e_m(x) + e_m(y)$; $e_m\left(\frac{x}{y}\right) = e_m(x) + e_m(y)$; $e_m(x^k) = |k| e_m(x)$; $e_m(e^y) = E_m(y)$
- Se entiende por un problema inestable aquel en el cual pequeños cambios en los datos producen grandes cambios en los resultados. Cuando un error alcanza a los datos de un problema inestable, este error se amplifica repentinamente y, a partir de ahí, este error desproporcionado contamina todos los pasos siguientes del algoritmo.
- En el caso de algoritmos donde se suman varios números de diferentes signos, el problema se suele hacer inestable cuando los datos conducen a una suma próxima a cero, debido que en el error absoluto se suman los errores absolutos de los sumandos y el error relativo se hace muy grande al dividir por una suma muy pequeña.
- Se utilizará el signo := para indicar que la expresión que aparezca a la derecha debe ser asignada a la variable que aparece a la izquierda
- La estructura **if** – **then** – **else** se utiliza para indicar que, si se cumple una condición, se ejecute una secuencia de acciones y en caso contrario, se ejecute otra.
- Las estructuras repetitivas sirven para que se ejecute repetidamente una secuencia de acciones. La secuencia se repite hasta que se satisface una cierta condición. Se emplean tres tipos de estructuras repetitivas: **do – while**; **repeat – until** y **for.**

Auto examen

1. Explique las diferencias esenciales entre los métodos analíticos y los métodos numéricos en cuanto a:
   a) Tipo de operaciones que utilizan.
   b) Exactitud de sus resultados.
   c) Generalidad de sus métodos.
   d) Posibilidad de implementarse en una computadora.

2. ¿Qué son los errores de truncamiento y por qué reciben este nombre?

3. Se sabe que el número aproximado $x = 26{,}8768$ tiene un error relativo máximo de 2%. Determine un intervalo donde se encuentre con toda seguridad el número exacto x^*.

4. Al resolver una ecuación por un procedimiento numérico se llegó a la conclusión de que la raíz buscada estaba comprendida en el intervalo (2,56482; 2,56494). Si se toma como aproximación de la raíz el valor $x = 2{,}5649$ halle el error absoluto máximo de x, su error relativo máximo y la cantidad de cifras decimales exactas que posee.

5. Explique por qué se puede afirmar que en la memoria de una computadora no se pueden almacenar ningún número irracional ni todos los números racionales. ¿Cuáles son, en definitiva, los números que sí se pueden guardar?

6. Se sabe que $a = 55{,}24$ con 3 cifras exactas, $b = 0{,}85674$ con error absoluto menor que 0,00001 y $c = 0{,}045386$ con un error relativo máximo de 0,05 %. Si con estos datos se calcula $S = ab + bc + ca$ halle el error relativo máximo de S y determine cuales de sus dígitos son exactos.

7. ¿Qué significa la afirmación de que calcular las raíces de un polinomio de grado alto suele ser un problema inestable?

8. ¿Por qué en los algoritmos numéricos debe evitarse la operación de restar números similares?

9. Se sabe que la función sen x puede ser aproximada por su polinomio de Taylor de grado $2n + 1$:

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

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