Capítulo 1

Cálculo aproximado, algoritmos, errores

Resumen

En este capítulo se realiza un tour corto en los métodos numéricos. Se inicia con la presentación de una metodología para el análisis de problemas y las soluciones aproximadas, la elaboración de algoritmos y algunas nociones de la complejidad de los mismos. A continuación se presentan ejemplos de algoritmos simples así como de algunos problemas elementales que se presentan en el ámbito del álgebra lineal y del análisis matemático, y, métodos simples de resolución numérica. Se hace un corto análisis de los tipos de errores. El uso de instrumentos de cálculo como son las calculadoras de bosillo y los computadores motivan el estudio de la representación en punto flotante, los errores de redondeo y la aritmética en punto flotante, temática que a su vez requiere del análisis de los sistemas de numeración. Luego se realiza un estudio del condicionamiento de funciones de una y varias variables que está relacionado con la amplificación de los errores de redondeo. Particular atención se pone en las operaciones aritméticas, lo que permite establecer una jerarquía en las mismas e identificar que operaciones son las peligrosas y bajo que condiciones y cuales no son peligrosas, lo que constituye una ayuda extremadamente grande cuando se elaboran los algoritmos de cálculo. Mediante algunos ejemplos se analiza el problema de la propagación de los errores así como el de la estabilidad numérica.

1.1. Introducción

Uno de los objetivos importantes del Análisis Numérico es la elaboración de métodos, procedimientos de cálculo y construcción de algoritmos que con la utilización de instrumentos de cálculo como las calculadoras de bolsillo o de instrumentos de cálculo mucho más complejos como los computadores, que requieren de la elaboración de programas computacionales, permitan calcular soluciones exactas o aproximadas de una diversidad de problemas matemáticos de modo que con cualesquiera de estos instrumentos, se deba tener un control sobre los errores cometidos en los cálculos y que los resultados finales sean de calidad.

Por otro lado, los procedimientos de cálculo, los algoritmos numéricos deben ser, en lo posible, los más simples, concisos, de aplicabilidad a una amplia variedad de situaciones. El costo numérico de cada procedimiento o algoritmo y su programa computacional que se construya debe ser, en lo posible, el más pequeño.

La calidad de la solución de un problema dado depende de muchos factores, entre ellos, de los datos de entrada que se requieren para la ejecución del algoritmo, procedimiento o programa computacional construido, así como de los instrumentos de cálculo utilizados, del lenguaje de programación y de la versión del mismo. Es claro que la calidad de la solución depende fuertemente del método numérico empleado y este a su vez depende de dos componentes importantes: el condicionamiento y la estabilidad; y, para problemas cuyas soluciones se aproximan mediante sucesiones, dependen a más de todos los componentes anteriores, de la convergencia.

En este capítulo se tratan algunos elementos de los algoritmos y características de los programas computacionales, los tipos de errores comunes en análisis numérico. Se revisa brevemente los sistemas de numeración entre los que se destacan el binario y el decimal, la representación en punto fijo y punto flotante, los errores de redondeo y la aritmética en punto flotante. Se introducen las nociones elementales de condicionamiento, estabilidad numérica y convergencia que son muy importantes en la construcción de algoritmos, procedimientos de cálculo y de la elaboración de programas computacionales, y que constituyen las bases que deben tenerse siempre presentes para el desarrollo de software en el cálculo científico.

1.2. Cálculo numérico. Algoritmos

Suponemos que un problema (P) ha sido planteado y que requiere de su resolución. Tres situaciones se presentan: la primera en la que la solución del problema (P) podemos encontrarlo directamente y no se requiere del cálculo numérico. La segunda en la que la solución del problema (P) podemos encontrarlo directamente y se requiere de la implementación de un procedimiento de cálculo para aproximar la solución encontrada. La tercera en la que no es posible encontrar directamente la solución y se requiere de un método numérico para aproximar la solución. Son estas dos últimas situaciones que nos interesan. Más aún, en la resolución numérica de un problema matemático (P) se establece la siguiente metodología.

1. Estudio de la existencia de solución del problema (P).
2. Construcción de un método numérico que aproxime la solución del problema (P).
3. Elaboración del respectivo algoritmo o procedimiento de cálculo.
4. Elaboración de un programa o código numérico para el cálculo de la solución aproximada de (P).
5. Realización de pruebas para validar el algoritmo o procedimiento de cálculo y el programa computacional.

Desde el punto de vista práctico, esto es, problemas que surgen en las ciencias y en la industria, la metodología presentada se extiende con la calibración de la solución y luego viene la implementación de la solución. En este curso daremos énfasis fundamentalmente a los puntos 1), 2), 3) y 5) de la metodología precedente.

El punto 4) no lo abordaremos y dejamos al estudiante que elija el lenguaje de programación que le interese para la elaboración de sus propios programas computacionales con los que debe realizar pruebas para verificar resultados mediante la implementación del algoritmo así como verificar la correcta elaboración del programa computacional; o en su defecto, seleccione el paquete de programas de tipo comercial (Matlab, Matemática, etc.) en el que provea la solución del problema planteado con el algoritmo propuesto. Es un error gravísimo el modificar el problema planteado (P) a uno $(\widehat{P})$ cuya solución está implementado en el paquete de programas computacionales. Por otro lado, es también importante el uso de ciertas herramientas informáticas que ayuden a presentar de mejor manera los resultados y permita comprender mejor las soluciones, por ejemplo graficadores para presentar gráficas de curvas 2d, 3d, superficies, flujos, generación de mallas estructuradas y no estructuradas, etc.

El estudio de la existencia de una solución o soluciones del problema (P) es muy importante. Pués en él se deben conocer con precisión las hipótesis con las cuales nuestro problema tiene solución, y bajo que condiciones el problema (P) puede no tener solución. En muchos casos, en el estudio de existencia de soluciones se construye el método que conduce a encontrar la solución de (P). Si el problema no tiene solución, carece de sentido el intentar elaborar un método numérico de solución.

Debido a que los cálculos que se realizan son con números que tienen un número finito de cifras decimales, estos afectan los resultados, por lo que el control de los errores en los cálculos es fundamental, es decir, debemos conocer la precisión con la que obtenemos la solución numérica del problema (P). Este es uno

de los problemas centrales del análisis numérico y que están ligados con las nociones de consistencia y la estabilidad numérica. En el caso en que la solución de (P) se calcula como límite de una sucesión de soluciones de problemas (Pn) más sencillos a resolver, otro de los problemas centrales del análisis numérico es probar o demostrar que las soluciones de esos problemas más sencillos converge a la solución del problema (P), es decir se debe probar la convergencia del método numérico propuesto. La consistencia, estabilidad y convergencia se discutirán más adelante.

Tanto en el estudio de existencia de soluciones como en la elaboración del método numérico se identifican los datos que se requieren para resolver el problema. Una parte de estos datos los conocemos como datos de entrada.

Una vez establecido el método numérico, se pasa enseguida a la elaboración o construcción del algoritmo. En la definición siguiente se establece la noción de algoritmo en su versión la más simple

Definición 1 *Se llama algoritmo a una sucesión finita de operaciones elementales, que organizada como pasos o procedimientos, se describen en forma lógica como calcular la solución de un problema (P) de modo eficaz con datos de entrada dados.*

Un algoritmo contiene los siguientes elementos:

1. **Datos de entrada**: que consisten en valores o datos de partida, los cuales son asignados antes de arrancar la ejecución del algoritmo. Estos datos permiten inicializar el algoritmo para su ejecución.

   Es necesario verificar la lectura correcta de todos los datos de entrada.

   Los datos de entrada dependen obviamente del problema propuesto. Estos pueden ser datos que pertenecen a distintos conjuntos numéricos (enteros, reales, complejos), pueden ser funciones reales como las trigonométricas (seno, coseno, tangente y sus inversas), las funciones exponencial, logarítmica, las funciones hiperbólicas, polinomios, etc, pueden ser datos vectoriales como son los elementos de $\mathbb{R}^n$, pueden ser matrices, etc.

2. **Algoritmo o procedimiento**: constituye la secuencia de todos los pasos o procedimientos de cálculo que se deben ejecutar. Estos deben ser claros, precisos, lógicos. No se deben tener ambigüedades en la descripción de esos pasos o procedimientos. Debe considerarse todas las situaciones posibles que se presenten.

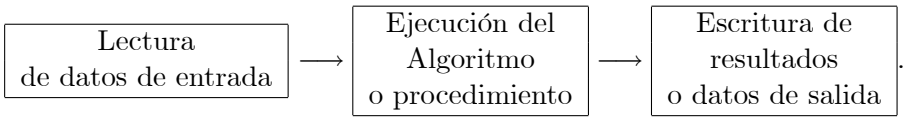
   La ejecución del algoritmo o procedimiento concluye siempre con un número finito de pasos.

3. **Datos de salida**: son una o más cantidades que tiene una relación estrecha con los datos de entrada. Estos resultados están definidos de manera única por los pasos del procedimiento o algoritmo.

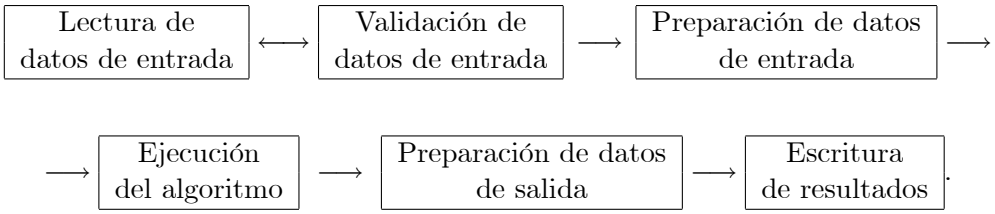
La escritura de un algoritmo contiene los datos de entrada, los datos de salida, y a continuación el procedimiento o la descripción del método a utilizar que constituye el algoritmo propiamente dicho que generalmente se lo expresa en pseodocódigo de modo que facilite la escritura de un programa computacional en cualquier lenguaje de programación.

Más adelante se proponen muchos algoritmos que permiten aclarar todas estas ideas.

En el siguiente esquema se muestra la secuencia de estos tres bloques:



En la práctica estos tres bloques no son suficientes para escribir un programa computacional. Un análisis más detallado de estos tres bloques proponemos en el diagrama siguiente:



En lo posible se busca construir algoritmos que tengan las características siguientes:

1. **Aplicabilidad general**: el algoritmo debe funcionar para una clase de problemas lo más amplia posible, donde las soluciones de un problema específico de la clase resulten solamente por cambios en los datos de entrada.

2. **Simplicidad**: Un algoritmo o procedimiento tiene que ser, en lo posible, simple de programar.

3. **Confiabilidad y seguridad**: el algoritmo no debe ser numéricamente costoso. En lo posible, se debe reducir el número de operaciones elementales a ejecutar. Esto evita que se amplifiquen los errores de redondeo, dando resultados más precisos, y por otro lado, reducen los tiempos de máquina.

   Se debe reducir, en lo posible, el número de variables a utilizar. Igualmente, se debe reducir en lo posible el número de subrutinas o bucles a utilizar, así como la repetición de ciertos cálculos.

   Se deben efectuar tests o pruebas con datos de entrada los más variados a fin de asegurarse que el algoritmo está correctamente elaborado y que los resultados son correctos o muy aceptables. Se buscará, en lo posible, ejemplos que se conozcan las soluciones exactas para compararse con las soluciones numéricas.

   En el estudio de un método numérico y consecuentemente de un algoritmo es importante, siempre que sea posible, determinar el número de operaciones elementales que se realizan para obtener la solución numérica del problema, el número de comparaciones, son menos importantes las reasignaciones . Entenderemos como operaciones elementales a las operaciones aritméticas como la suma, resta, multiplicación, división, raíz n-ésima. Las comparaciones están vinculadas con las relaciones de orden menor que $<$, mayor que $>$, menor o igual que $\leq$, mayor o igual que $\geq$. Prestaremos mayor atención a la determinación del número de operaciones elementales que se requieren para calcular la solución numérica mediante un método o procedimiento está relacionado con la complejidad del algoritmo que análizamos a continuación.

   Complejidad de algoritmos.

   Si para un problema (P) se conocen varios métodos y por lo tanto se pueden proporcionar varios algoritmos, es importante analizar la denominada complejidad del algoritmo. Esta tiene que ver con dos componentes importantes: uno del punto de vista volumen de memoria necesario del instrumento o equipo utilizado para el cálculo, y otro del punto de vista tiempo de máquina que a su vez está relacionado con el número de operaciones elementales (siempre que haya sido posible obtener) que se requieren para calcular la solución. Si se disponen de dos métodos, ¿cómo juzgar que método es mejor? ¿bajo que circunstancias un método es mejor que otro?. Para poder dar respuesta a estas interrograntes debemos estudiar la complejidad de cada algoritmo, esto es, determinar cuánto de memoria se requiere en la ejecución de cada método, el tiempo de máquina requerido para el cálculo de la solución con cada método.

   Cuando un problema (P) puede ser resuelto mediante dos métodos generados por sucesiones de problemas más simples que los notamos $(P_n^{(1)})$ y $(P_n^{(2)})$, el estudio de la convergencia de cada método es importante, esto nos proporcionará un dato que está relacionado con el orden de convergencia, ¿cuál método es mejor?. Para responder a esta interrogante, debemos considerar otro elemento que es la exactitud de la solución numérica que a su vez está relacionada con el orden de convergencia. Desde este punto de vista, el método generado que tenga un orden de convergencia más alto será mejor que el otro, lo que da respuesta a la interrogante.

1.3. Ejemplos de algoritmos y problemas

La metodología arriba propuesta la aplicaremos a algunos ejemplos que proponemos a continuación. Más aún, esta sección está dividida en dos partes: la primera en la que presentamos ejemplos simples de operaciones elementales con vectores y matrices, y luego dos aplicaciones del producto escalar en $\mathbb{R}^2$, y la segunda que está destinada a problemas del análisis matemático como cálculo de valores de funciones polinomiales, funciones con discontinuidad evitable, derivación e integración numérica.

1.3.1. Operaciones elementales con vectores y matrices. Aplicaciones

1. Suma de vectores y producto de escalares por vectores de $\mathbb{R}^n$.

   Sean $\alpha \in \mathbb{R}$, $\overrightarrow{x} = (x_1, \ldots, x_n)$, $\overrightarrow{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$. La suma de $\overrightarrow{x}$ con $\overrightarrow{y}$ se nota $\overrightarrow{x} + \overrightarrow{y}$ y se define como

   $$\overrightarrow{x} + \overrightarrow{y} = (x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n).$$

   El producto del escalar α con el vector $\overrightarrow{x}$ se nota $\alpha\overrightarrow{x}$ definido como

   $$\alpha\overrightarrow{x} = \alpha(x_1, \ldots, x_n) = (\alpha x_1, \ldots, \alpha x_n).$$

   Ponemos $\overrightarrow{z} = \overrightarrow{x} + \overrightarrow{y}$ y $\overrightarrow{w} = \alpha\overrightarrow{x}$. A continuación presentamos un algoritmo en el que se calcula $\overrightarrow{z} = \overrightarrow{x} + \overrightarrow{y}$ y $\overrightarrow{w} = \alpha\overrightarrow{x}$.

   Algoritmo

   Datos de entrada: $n \in \mathbb{Z}^+$, $\overrightarrow{x} = (x_1, \ldots, x_n)$, $\overrightarrow{y} = (y_1, \ldots, y_n)$.

   Datos de salida: $\overrightarrow{z}$, $\overrightarrow{w}$.

   1. $i = 1, \ldots, n$

      $z_i = x_i + y_i$

      $w_i = \alpha x_i$

      Fin bucle i.

   2. Imprimir $\overrightarrow{z}$, $\overrightarrow{w}$.
   3. Fin.

   Note que los datos son la talla n de los vectores $\overrightarrow{x}$ e $\overrightarrow{y}$ así como sus coordenadas. Observe que las operaciones elementales que intervienen en el cálculo de $\overrightarrow{z}$ son adiciones y en el cálculo de $\overrightarrow{w}$ son productos. Se realizan $2n$ operaciones elementales, y, el proceso de cálculo concluye en exactamente n pasos. No contabilizamos la presentación de resultados y el fin.

   La notación $i = 1, \ldots, n$ significa que para $i = 1$ se realizan los cálculos de $z_1 = x_1 + y_1$ y de $w_1 = \alpha x_1$, a continuación $k = 2$ y se realizan los cálculos $z_2 = x_2 + y_2$ y de $w_2 = \alpha x_2$. Se continua con este proceso hasta $k = n$ con lo que se hacen los cálculos $z_n = x_n + y_n$ y de $w_n = \alpha x_n$.

2. Producto escalar en $\mathbb{R}^n$.

   Sean $\overrightarrow{x} = (x_1, ..., x_n)$, $\overrightarrow{y} = (y_1, ..., y_n)$ dos vectores de $\mathbb{R}^n$. El producto escalar de $\overrightarrow{x}$ con $\overrightarrow{y}$ se nota con $\overrightarrow{x} \cdot \overrightarrow{y}$ o también $\overrightarrow{x}^T \cdot \overrightarrow{y}$ (cuando los vectores $\overrightarrow{x}$ e $\overrightarrow{y}$ se escriben como vectores columna) y se define como sigue:

   $$\overrightarrow{x}^T \cdot \overrightarrow{y} = \overrightarrow{x} \cdot \overrightarrow{y} = \sum_{i=1}^{n} x_i y_i.$$

   En el apéndice se resumen algunos resultados de los espacios vectoriales con producto interior.

   Para el cálculo de este producto escalar se requiere de la siguiente información: $n \in \mathbb{Z}^+$ y de los componentes o coordenadas de los vectores $\overrightarrow{x}$, $\overrightarrow{y}$, con lo que el producto escalar que se le denota con p puede calcularse usando el algoritmo que se propone a continuación.

   Algoritmo

   Datos de entrada: $n \in \mathbb{Z}^+$, $\vec{x} = (x_1, \cdots, x_n)$, $\vec{y} = (y_1, \cdots, y_n)$.

Datos de salida: p

1. $p = 0$.
2. $k = 1, \ldots, n$

   $p = p + x_k * y_k$.

   Fin bucle k.
3. Imprimir resultdo p.
4. Fin.

Para $n = 4$, $\overrightarrow{x} = (1, 0, -1, 2)$, $\overrightarrow{y} = (5, 2, -2, -3)$, la aplicación del algoritmo da como resultado $p = 1$.

Observe que las operaciones elementales que intervienen en el cálculo de p son adiciones y productos, se realizan $2n$ operaciones elementales, y, el proceso de cálculo de p concluye en exactamente n pasos.

La notación $k = 1, \ldots, n$ significa que para $k = 1$ se realiza el cálculo de $p + x_1 * y_1$ que se asigna a p, a continuación $k = 2$ y se realiza el cálculo $p + x_2 * y_2$ cuyo resultado se asigna nuevamente a p. Se continua con este proceso hasta $k = n$ con lo que se hace el cálculo $p + x_n * y_n$ que se asigna a p. La escritura $p = p + x_k * y_k$ no es una ecuación, en realidad se trata de una asignación del resultado $p + x_k * y_k$ a la variable p. Este tipo de notación será utilizada únicamente en la escritura de los algoritmos.

3. Norma euclídea en $\mathbb{R}^n$.

   Sea $\overrightarrow{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$. La norma euclídea en $\mathbb{R}^n$ se nota con $\| \cdot \|_2$ y se define como:

$$\| \overrightarrow{x} \|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}.$$

   En el apéndice se resumen algunos resultados de los espacios normados.

   Para el cálculo de $\| \overrightarrow{x} \|_2$ se requiere de la siguiente información: $n \in \mathbb{Z}^+$, las coordenadas x_i, $i = 1, \ldots, n$, del vector $\overrightarrow{x}$. El siguiente algoritmo permite calcular $\| \overrightarrow{x} \|_2$ que se le nota con N_x.

   Algoritmo

   Datos de entrada: $n \in \mathbb{Z}^+$, $\overrightarrow{x} = (x_1, \cdots, x_n)$.

   Datos de salida: N_x

   1. $N_x = 0$.
   2. $i = 1, \ldots, n$

      $N_x = N_x + x_i * x_i$

      Fin bucle i.
   3. $N_x = \sqrt{N_x}$.
   4. Imprimir resultado N_x.
   5. Fin.

   La escritura $N_x = \sqrt{N_x}$, en realidad significa que el cálculo de $\sqrt{N_x}$ se asigna a N_x. Esta notación se utilizará únicamente en la escritura de algoritmos.

   Sean $n = 4$, $\overrightarrow{x} = \left(3, 2, -3, -\sqrt{3}\right)$. La aplicación del algoritmo precedente da como resultado $N_x = 5$.

   Las operaciones elementales que intervienen en el cálculo de N_x son adiciones, productos y una raíz cuadrada, en un total de $2n+1$ operaciones elementales. El proceso de cálculo de N_x concluye luego de $n + 1$ pasos.

4. Suma de matrices reales de $m \times n$.

   Se nota con $M_{m\times n}[\mathbb{R}]$ el espacio vectorial de matrices de $m \times n$ con valores en $\mathbb{R}$. En algunos libros este espacio vectorial se nota como $\mathbb{R}^{mn}$. A una matriz $A \in M_{m\times n}[\mathbb{R}]$ se le nota $A = (a_{ij})_{m\times n}$ y si $m = n$, es decir A es una matriz cuadrada, se escribirá $A = (a_{ij})$.

   Sea $A = (a_{ij})_{m\times n}$, $B = (b_{ij})_{m\times n}$. La suma de las matrices A y B está definida como

$$A + B = (a_{ij})_{m\times n} + (b_{ij})_{m\times n} = (a_{ij} + b_{ij})_{m\times n} = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ & \vdots & \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$$

   A esta matriz suma lo denotamos con $C = (c_{ij})_{m\times n}$, esto es, $C = A + B$. El algoritmo para sumar las matrices A y B se muestra a continuación.

   Algoritmo

   Datos de entrada: $m, n \in \mathbb{Z}^+$, $A = (a_{ij})_{m\times n}$, $B = (b_{ij})_{m\times n}$.

   Datos de salida: $C = (c_{ij})_{m\times n}$.

   1. $i = 1, \ldots, m$
      $j = 1, \ldots, n$
      $c_{ij} = a_{ij} + b_{ij}$
      Fin bucle j.
      Fin bucle i.
   2. Imprimr $C = (c_{ij})_{m\times n}$.
   3. Fin.

   El cálculo de la matriz C requiere de $m \times n$ adiciones. Note que el índice i es utilizado para indicar las filas, el índice j es utilizado para indicar las columnas. El algoritmo muestra que la matriz C se construye fila a fila, esto es, primera fila, a continuación segunda fila, así sucesivamente. Se deja como ejercicio elaborar un algoritmo de cálculo de C por columnas.

5. Producto de matrices.

   Sean $A = (a_{ij})_{m\times n}$, $B = (b_{jk})_{n\times p}$ matrices reales. El producto de la matriz A con B se nota AB y es la matriz $C = (c_{ik})_{m\times p}$ definida como sigue:

$$C_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk} = a_{i1}b_{1k} + \cdots + a_{in}b_{nk}, \quad i = 1, \ldots, m, \quad k = 1, \ldots, p.$$

   Como se puede apreciar, el elemento c_{ik} es el resultado de las sumas de los productos de los elementos de la fila i de la matriz A con los correspondientes de la columna k de la matriz B. Un algoritmo para calcular $C = AB$ se muestra a continuación.

   Algoritmo

   1. $i = 1, \ldots, m$
      $k = 1, \ldots, p$
      $s = 0.$
      $j = 1, \ldots, n$
      $s = s + a_{ij} \times b_{jk}$
      Fin bucle j.
      $c_{ik} = s$
      Fin bucle k.
      Fin bucle i.

2. Imprimir $C = (c_{ik})_{m\times p}$.

3. Fin.

Este algoritmo concluye en un número finito de pasos, exactamente en $m \times p\,(n+2)$ pasos. Las operaciones elementales que se realizan son sumas y productos. Adicionalmente se hacen $2m \times p$ asignaciones. Note que la escritura $s = s + a_{ij}b_{jk}$ no es una ecuación, se trata de una asignación pués el producto $a_{ij}b_{jk}$ se suma a s y este resultado se almacena en s.

6. Intercambio de dos filas de una matriz.

Sea $A = (a_{ij})_{m\times n} \in M_{m\times n}[\mathbb{R}]$. La matriz $B = (b_{ij})_{m\times n}$ obtenida al intercambiar la fila i con la fila j con $i < j$ se define como $B = E_{i\to j}A$, donde $E_{i\to j} = (e_{pq})_{m\times m}$ se obtiene de la matriz identidad $I = (I_{li})_{m\times m}$ al intercambiar la fila i con la fila j, por lo tanto $E_{i\to j} = (e_{pq})_{m\times m}$ está definida como sigue:

$$e_{ik} = \begin{cases} 0, & \text{si } k \neq j \\ 1, & \text{si } k = j, \end{cases} \qquad e_{jk} = \begin{cases} 0, & \text{si } k \neq j \\ 1, & \text{si } k = i, \end{cases} \qquad k = 1, \ldots, m,$$

y para $p = 1, \ldots, m$ con $p \neq i, j$, $\quad e_{pk} = \begin{cases} 0, & \text{si } p \neq k, \\ 1, & \text{si } p = k \end{cases} \qquad k = 1, \ldots, m.$

Un algoritmo que realiza el producto $E_{i\to j}A$ se muestra a continuación.

Algoritmo

Datos de entrada; $m, n \in \mathbb{Z}^+$, $i, j \in \mathbb{Z}^+$, $A = (a_{ij})_{m\times n}$.

Datos de salida: $B = (b_{pr})_{m\times n}$.

1. Si $m = 1$ continuar en 5).

2. $p = 1, \ldots, m$

si $p \neq i$ y $p \neq j$

$r = 1, \ldots, n$

$b_{pr} = a_{pr}$

Fin bucle r.

Fin bucle p.

3. $r = 1, \ldots, n$

$c = a_{ir}$

$b_{ir} = a_{jr}$

$b_{jr} = c$

Fin bucle r.

4. Imprimir $B = (b_{pr})_{m\times n}$. Continuar en 6).

5. Imprimir mensaje: $m \geq 2$.

6. Fin.

La ejecución de este algoritmo implica la realización de asignaciones y de comparaciones, así el número de comparaciones es $2m+1$ y el número de asignaciones $n\,(m+1)$. Obviamente el algoritmo concluye en un número finito de pasos.

Note que se requiere de la siguiente información: talla de la matriz A, esto es, los enteros positivos m, n, los $m \times n$ coeficientes a_{ij} de A, la fila i, la fila j. Esta última información implica que $m \geq 2$. Si $m = 1$ no se realiza intercambio de filas.

7. Producto de una matriz por un vector.

Sean $A = (a_{ij})_{m\times n} \in M_{m\times n}[\mathbb{R}]$, $\quad \overrightarrow{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$. El producto $A\overrightarrow{x}$ se define como sigue:

$$A\overrightarrow{x} = \begin{bmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}x_j \end{bmatrix}.$$

Para elaborar un algoritmo de cálculo del producto de la matriz A por el vector $\overrightarrow{x}$, esto es, $A\overrightarrow{x}$ se requiere de la siguiente información: talla de la matriz A, o sea m, $n \in \mathbb{Z}^+$, de sus componentes a_{ij}, $\ i = 1, \ldots, m,\ \ j = 1, \ldots, n$, y de los componentes o coordenadas $x_i,\ \ i = 1, \ldots, n$ del vector $\overrightarrow{x}$. Con esta información el producto $A\overrightarrow{x}$ puede calcularse con el algoritmo que se propone a continuación.

Algoritmo

Datos de entrada: $m, n \in \mathbb{Z}^+,\ A = (a_{ij})_{m\times n}, \quad \overrightarrow{x} = (x_1, \cdots, x_n)$.

Datos de salida: $\vec{z} = A\vec{x}$.

1. $c = 0$.

2. $i = 1, \ldots, m$

   $j = 1, \ldots, n$

   $c = c + a_{ij} * x_j$

   Fin bucle j.

   $z_i = c$

   $c = 0$.

   Fin bucle i.

3. Imprimir resultado $\overrightarrow{z} = (z_1, \cdots, z_m)$.

4. Fin.

Note que el algoritmo concluye luego de $m \times n$ pasos en los que intervienen productos y adiciones.

8. Vectores colineales. Angulo entre vectores y base ortogonal de $\mathbb{R}^2$.

Sean $\overrightarrow{u}_1 = (a_1, b_1)\,,\ \overrightarrow{u}_2 = (a_2, b_2)$ dos elementos no nulos de $\mathbb{R}^2$. Se considera el siguiente problema: determinar si los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ no son colineales, en tal caso calcular el ángulo que forman dichos vectores y construir una base ortogonal. Elaborar un algoritmo numérico.

Analicemos la existencia de soluciones.

Consideramos en el plano el sistema de coordenadas rectangulares y sean $\overrightarrow{u}_1 = (a_1, b_1)\,,\ \overrightarrow{u}_2 = (a_2, b_2)$ dos vectores no nulos. Se sabe que $\overrightarrow{u}_1$ y $\overrightarrow{u}_2$ son colineales si y solo si sus coordenadas satisfacen la relación $a_1b_2 - a_2b_1 = 0$. Por lo tanto, $\overrightarrow{u}_1$ y $\overrightarrow{u}_2$ no son colineales si y solo si $d = a_1b_2 - a_2b_1 \neq 0$.

El producto escalar de los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ se nota con $\overrightarrow{u}_1 \cdot \overrightarrow{u}_2$ y está definido como $\overrightarrow{u}_1 \cdot \overrightarrow{u}_2 = a_1a_2 + b_1b_2$. La longitud o norma de un vector $\overrightarrow{u} = (a, b) \in \mathbb{R}^2$ se nota $\|\overrightarrow{u}\|$ y se define como

$$\|\overrightarrow{u}\| = (\overrightarrow{u} \cdot \overrightarrow{u})^{\frac{1}{2}} = \sqrt{a^2 + b^2}.$$

Además, la medida del ángulo que forman los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ es el número real $\theta \in [0, \pi]$ definido como

$$\cos(\theta) = \frac{\overrightarrow{u}_1 \cdot \overrightarrow{u}_2}{\|\overrightarrow{u}_1\|\,\|\overrightarrow{u}_2\|}$$

y de esta relación

$$\theta = \operatorname{arc\,cos}\left(\frac{\overrightarrow{u}_1 \cdot \overrightarrow{u}_2}{\|\overrightarrow{u}_1\|\,\|\overrightarrow{u}_2\|}\right).$$

Recordemos que dos vectores $\overrightarrow{u},\ \overrightarrow{v}$ de $\mathbb{R}^2$ son ortogonales o perpendiculares si y solo si $\overrightarrow{u} \cdot \overrightarrow{v} = 0$. En tal caso escribimos $\overrightarrow{u} \perp \overrightarrow{v}$.

En la figura de la izquierda se muestran los vectores no nulos y no colineales $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ y el ángulo θ que forman dichos vectores. En la figura de la derecha se muestran los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$, la

proyección ortogonal de $\overrightarrow{u}_2$ sobre $\overrightarrow{u}_1$ y el vector $\overrightarrow{c}$ ortogonal a $\overrightarrow{u}_1$, esto es $c \perp \overrightarrow{u}_1$.

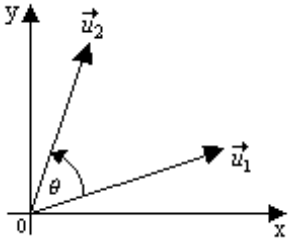


Figura 1

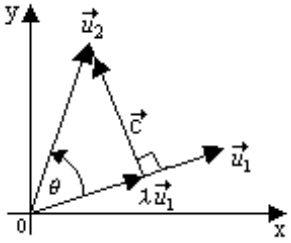


Figura 2

Ponemos $\overrightarrow{v}_1 = \overrightarrow{u}_1$. Para construir una base ortogonal $\{\overrightarrow{v}_1, \overrightarrow{v}_2\}$ consideramos las dos condiciones siguientes:

$$\text{hallar } \lambda \in \mathbb{R} \text{ y } \overrightarrow{c} \in \mathbb{R}^2 \text{ tales que } \begin{cases} \lambda\overrightarrow{u}_1 + \overrightarrow{c} = \overrightarrow{u}_2, \\ \overrightarrow{u}_1 \perp \overrightarrow{c}. \end{cases}$$

Calculemos λ. Multiplicando escalarmente por $\overrightarrow{u}_1$ la primera igualdad, se tiene

$$(\lambda\overrightarrow{u}_1 + \overrightarrow{c}) \cdot \overrightarrow{u}_1 = \overrightarrow{u}_2 \cdot \overrightarrow{u}_1$$

y como el producto escalar es distributivo respecto de la adición de vectores, resulta

$$\lambda\overrightarrow{u}_1 \cdot \overrightarrow{u}_1 + \overrightarrow{c} \cdot \overrightarrow{u}_1 = \overrightarrow{u}_2 \cdot \overrightarrow{u}_1.$$

Tomando en consideración que $\overrightarrow{u}_1 \perp \overrightarrow{c}$ que a su vez es equivalente a $\overrightarrow{u}_1 \cdot \overrightarrow{c} = 0$, se sigue que

$$\lambda\overrightarrow{u}_1 \cdot \overrightarrow{u}_1 = \overrightarrow{u}_2 \cdot \overrightarrow{u}_1.$$

Puesto que $\|\overrightarrow{u}_1\|^2 = \overrightarrow{u}_1 \cdot \overrightarrow{u}_1$ y como el producto escalar es conmutativo, esto es, $\overrightarrow{u}_2 \cdot \overrightarrow{u}_1 = \overrightarrow{u}_1 \cdot \overrightarrow{u}_2$ resulta $\lambda = \dfrac{\overrightarrow{u}_1 \cdot \overrightarrow{u}_2}{\|\overrightarrow{u}_1\|^2}$. El número real λ se llama coeficiente de Fourier.

Una vez calculado λ pasamos a determinar el vector $\overrightarrow{c}$. De la igualdad $\lambda\overrightarrow{u}_1 + \overrightarrow{c} = \overrightarrow{u}_2$ se obtiene $\overrightarrow{c}$:

$$\overrightarrow{c} = \overrightarrow{u}_2 - \lambda\overrightarrow{u}_1 = \overrightarrow{u}_2 - \frac{\overrightarrow{u}_1 \cdot \overrightarrow{u}_2}{\|\overrightarrow{u}_1\|^2}\overrightarrow{u}_1.$$

Definimos $\overrightarrow{v}_2 = \overrightarrow{c}$. Así $\overrightarrow{v}_1 \perp \overrightarrow{v}_2$. En la figura siguiente se muestran los vectores $\overrightarrow{v}_1$, $\overrightarrow{v}_2$ tales que $\overrightarrow{v}_1 \perp \overrightarrow{v}_2$.

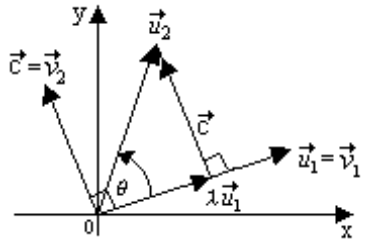


Figura 3

Con todos estos elementos estamos en condiciones de elaborar un algoritmo numérico que permita identificar si dos vectores no nulos son o no colineales. En caso de no ser colineales, calcular el ángulo que forman y obtener una base ortogonal $\{\overrightarrow{v}_1, \overrightarrow{v}_2\}$.

Algoritmo

Datos de entrada: $\overrightarrow{u}_1 = (a_1, b_1)$, $\overrightarrow{u}_2 = (a_2, b_2)$

Datos de salida: Mensaje "vectores colineales", θ, $\overrightarrow{v}_1$, $\overrightarrow{v}_2$.

1. Verificar $a_1 \neq 0$ o $b_1 \neq 0$, y, $a_2 \neq 0$ o $b_2 \neq 0$. Caso contrario $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ son nulos. Continuar en 10)

2. Calcular $d = a_1b_2 - a_2b_1$.

3. Si $d = 0$, continuar en 9).

4. Calcular $p = a_1a_2 + b_1b_2$,

$$n_1 = \left(a_1^2 + b_1^2\right)^{\frac{1}{2}},$$

$$n_2 = \left(a_2^2 + b_2^2\right)^{\frac{1}{2}},$$

$$\theta = \arccos\left(\frac{p}{n_1n_2}\right).$$

5. Poner $\overrightarrow{v}_1 = (a_1, b_1)$.

6. Calcular $\lambda = \dfrac{p}{n_1^2}$,

$$x = a_2 - \lambda a_1,$$

$$y = b_2 - \lambda b_1.$$

7. Poner $\overrightarrow{v}_2 = (x, y)$.

8. Imprimir: ángulo θ, vectores ortogonales $\overrightarrow{v}_1,\ \overrightarrow{v}_2$. Continuar en 11).

9. Imprimir: $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ vectores colineales. Continuar en 11).

10. Imprimir: $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ vectores nulos.

11. Fin.

El número total de operaciones elementales que se realizan en la ejecución de este algoritmo son 22 operaciones, comparaciones 5, asignaciones 4, una evaluación de la función arco coseno. Note que el punto 4) del algoritmo se ejecuta cuando $d \neq 0$.

Verifiquemos el algoritmo con los siguientes datos $\overrightarrow{u}_1 = (3, 1)$, $\overrightarrow{u}_2 = \left(-2, \sqrt{5}\right)$.

Claramente los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ son no nulos. Pasemos a calcular d. Tenemos $d = 3\times\sqrt{5}-(-2)\times 1 = 2 + 3\sqrt{5}$ y $d \neq 0$ con lo que se continua con el cálculo de $p,\ n_1,\ n_2$ y θ. Tenemos

$$\begin{aligned} p &= 3 \times (-2) + 1 \times \sqrt{5} = -6 + \sqrt{5}, \\ n_1 &= \left(3^2 + 1^2\right)^{\frac{1}{2}} = \sqrt{10}, \quad n_2 = \left((-2)^2 + \left(\sqrt{5}\right)^2\right)^{\frac{1}{2}} = 3, \\ \theta &= \arccos\left(\frac{-6 + \sqrt{5}}{3\sqrt{10}}\right) \simeq \arccos\left(-\frac{3{,}763932023}{9{,}48683298}\right) \simeq 1{,}978773429. \end{aligned}$$

Ponemos $\overrightarrow{v}_1 = (3, 1)$.

Calculemos el coeficiente de Fourier λ, y, x e y :

$$\begin{aligned} \lambda &= \frac{p}{n_1^2} = \frac{-6 + \sqrt{5}}{10} \simeq -0{,}3763932023, \\ x &= a_2 - \lambda a_1 \simeq -2 - (-0{,}3763932023) \times 3 = -0{,}870820393, \\ y &= b_2 - \lambda b_1 \simeq \sqrt{5} - (-0{,}3763932023) \times 1 = 2{,}61246118. \end{aligned}$$

El vector $\overrightarrow{v}_2$ está definido como $\overrightarrow{v}_2 = (-0{,}870820393, 2{,}61246118)$.

El símbolo $\simeq$ se utiliza para indicar un valor aproximado.

En la figura siguiente se muestran los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ y los vectores ortogonales $\overrightarrow{v}_1,\ \overrightarrow{v}_2$.

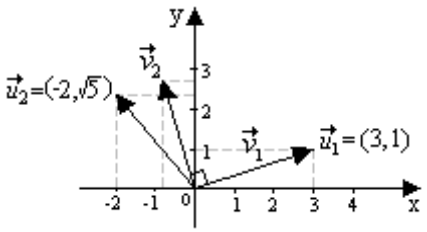


Figura 4

9. En este ejemplo se trata el método de eliminación gaussiana para sistemas de ecuaciones lineales de 3×3. Comencemos observando que los sistemas de tres ecuaciones con tres incógitas más simples de resolver son los sistemas de ecuaciones denominados diagonales, los denominados triangulares superiores y triangulares inferiores que en ese orden se presentan a continuación:

$$\begin{cases} a_1x & & = d_1 \\ & b_2y & = d_2 \\ & & c_3z = d_3 \end{cases}, \quad \begin{cases} a_1x + b_1y + c_1z = d_1 \\ b_2y + c_2z = d_2 \\ c_3z = d_3 \end{cases}, \quad \begin{cases} a_1x & = d_1 \\ a_2x + b_2y & = d_2 \\ a_3x + b_3y + c_3z = d_3, \end{cases}$$

donde $a_i,\ b_i,\ c_i \in \mathbb{R}$ para $i = 1,\ 2,\ 3$, no todos nulos, $d_i \in \mathbb{R}$ para $i = 1,\ 2,\ 3$, donde $x, y, z \in \mathbb{R}$ son las incógnitas del sistema que queremos resolver. Los números reales $a_1,\ b_2,\ c_3$ que figuran en la diagonal de cada uno de los sistemas precedentes se los denomina elementos o coeficientes de la diagonal del sistema de ecuaciones lineales.

En el caso de un sistema de ecuaciones lineales diagonal, la solución es única si y solo si los coeficientes de la diagonal del sistema son no nulos, es decir $a_1 \neq 0,\ b_2 \neq 0,\ c_3 \neq 0$ en cuyo caso la solución es $x = \dfrac{d_1}{a_1},\ y = \dfrac{d_2}{b_2},\ z = \dfrac{d_3}{c_3}$ que lo expresamos como $\left(\dfrac{d_1}{a_1},\ \dfrac{d_2}{b_2},\ \dfrac{d_3}{c_3}\right)$.

Los sistemas de ecuaciones lineales triangulares superiores e inferiores tienen una única solución si y solo si los elementos de la diagonal del sistema son no nulos, esto es, $a_1 \neq 0,\ b_2 \neq 0,\ c_3 \neq 0$.

Ejemplos

1. El sistema de ecuaciones lineales diagonal definido como $(x, y, z) \in \mathbb{R}^3$ tal que $\begin{cases} 2x & = 11 \\ -3y & = 0 \\ 5z = -5, \end{cases}$ tiene como solución $x = \dfrac{11}{2},\ y = -\dfrac{0}{5} = 0,\ z = -\dfrac{5}{5} = -1$, que lo escribimos $(\dfrac{11}{2}, 0, -1)$.

2. Considerar el sistema de ecuaciones lineales definido como sigue: $(x, y, z) \in \mathbb{R}^3$ tal que $\begin{cases} x + 2y + 3z = 11 \\ -y - 2z = 0 \\ 5z = -5. \end{cases}$ Este es un sistema de ecuaciones lineales triangular superior. Para hallar la solución de este sistema, comenzamos por la última ecuación, de la que obtenemos la incógnita $z : z = -\frac{5}{5} = -1$. De la segunda ecuación, se obtiene la incógnita $y : y = -2z = -2 \times (-1) = 2$, y de la primera ecuación, obtenemos $x : x = 11 - 2y - 3z = 11 - 2 \times 2 - 3 \times (-1) = 10$. La solución es $x = 10,\ y = 2,\ z = -1$ que escribimos $(-10,\ 2,\ -1)$.

3. Considerar el sistema de ecuaciones lineales definido por $(x, y, z) \in \mathbb{R}^3$ tal que $\begin{cases} 2x & = 4 \\ 3x + 4y & = 18 \\ -3x + 4y + z = 11. \end{cases}$

Este es un sistema de ecuaciones lineales triangular inferior, cuya solución encontramos resolviendo de la primera a la tercera ecuación. De la primera ecuación obtenemos $x = \frac{4}{2} = 2$. De la segunda ecuación: $y = \frac{1}{4}(18 - 3x) = \frac{1}{4}(18 - 3 \times 2) = 3$, y de la tercera ecuación: $z = 11 + 3x - 4y = 11 + 3 \times 2 - 4 \times 3 = 5$. Así, $x = 2,\ y = 3,\ z = 5$ es la solución que la escribimos $(2,\ 3,\ 5)$.

Pasemos a describir el método de eliminación gaussiana (será tratado con mayor profundidad en el capítulo 6). Para el efecto explicamos mediante tres ejemplos. La idea fundamental en el método

de eliminación gaussiana es transformar el sistema de ecuaciones lineales dado en un sistema de ecuaciones lineales triangular superior, que como hemos visto, esta clase de sistemas son los más simples de resolver.

Ejemplos

1. Resolver el sistema de ecuaciones lineales siguiente: $(x, y, z) \in \mathbb{R}^3$ tal que $\begin{cases} x + 2y + 3z = 7 \\ 2x - y - 2z = 0 \\ 3x - 2y + 5z = 25. \end{cases}$

El procedimiento de la eliminación gaussiana lo dividimos en tres etapas. Las dos primeras que conducen a transformar el sistema de ecuaciones en uno triangular superior; y, la tercera etapa que consiste en resolver el sistema de ecuaciones triangular superior.

a) Primera etapa. Mantenemos fija la primera ecuación. Se trata de eliminar la incógnita x de la segunda y tercera ecuaciones.

Eliminemos x de la segunda ecuación. Para el efecto multiplicamos por $k = -2$ (k se obtiene como el coeficiente de x de la segunda ecuación, dividido para el coeficiente de x de la primera ecuación cambiado de signo) a la primera ecuación y le sumamos el resultado a la segunda ecuación. Obtenemos $\begin{cases} x + 2y + 3z = 7 \\ -5y - 8z = -14 \\ 3x - 2y + 5z = 25. \end{cases}$ Eliminemos x de la tercera ecuación. Para ello multiplicamos por $k = -3$ (k se obtiene como el coeficiente de x de la tercera ecuación, dividido para el coeficiente de x de la primera ecuación cambiado de signo) a la primera ecuación y le sumamos el resultado a la tercera ecuación, resulta $\begin{cases} x + 2y + 3z = 7 \\ -5y - 8z = -14 \\ -8y - 4z = 4. \end{cases}$

b) Segunda etapa. Mantenemos fija la primera y segunda ecuaciones y eliminamos y en la tercera ecuación. Multipliquemos por $k_1 = -\dfrac{-8}{-5} = -\frac{8}{5}$ a la segunda ecuación y el resultado le sumamos a la tercera. k_1 se obtiene como $\begin{cases} x + 2y + 3z = 7 \\ -5y - 8z = -14 \\ \frac{44}{5}z = \frac{132}{5}. \end{cases}$

Note que k_1 se obtiene como el cociente cambiado de signo del coeficiente de y de la tercera ecuación dividido para el coeficiente de y de la segunda ecuación, siempre que este no sea nulo.

c) Tercera etapa: Resolvemos el sistema de ecuaciones triangular superior. Comenzamos con la terera ecuación, obtenemos z: $z = \frac{132}{44} = 3$. De la segunda ecuacion obtenemos y: $y = \frac{14-8z}{5} = \frac{14-8\times 3}{5} = -2$, y de la primera ecuación obtenemos $x = 7 - 2y - 3z = 7 - 2 \times (-2) - 3 \times 3 = 2$. La solución es $x = 2,\ y = -2, z = 3$, que escibimos $(2,\ -2,\ 3)$.

2. Considerar el sistema de ecuaciones lineales siguiente: $(x,\ y,\ z) \in \mathbb{R}^3$ tal que $\begin{cases} 2x + y = -2 \\ 3x + 4y + z = -2 \\ -3x + 4y + z = 4. \end{cases}$ Apliquemos el método de eliminación gaussiana. Mantengamos fija a la primera ecuación, y procedamos a la eliminación de la incógnita x en la segunda y tercera ecuaciones. Multipliquemos a la primera ecuación por $k = -\dfrac{3}{2}$ (coeficiente de x de la segunda ecuación, dividido para el coeficiente de x de la primera ecuación cambiado de signo), el resultado sumamos a la segunda. Obtenemos $\begin{cases} 2x + y = -2 \\ \frac{5}{2}y + z = 1 \\ -3x + 4y + z = 4. \end{cases}$

Sea $k_1 = -\dfrac{-3}{2} = \dfrac{3}{2}$ (k_1 se obtiene dividiendo el coeficiente de x de la tercera ecuación para el coeficiente de x de la primera ecuación, cambiado de signo). Multiplicando a la primera ecuación

por k_1, el resultado sumamos a la tercera ecuación. Tenemos $\begin{cases} 2x+y = -2 \\ \frac{5}{2}y+z=1 \\ \frac{11}{2}y+z=1. \end{cases}$ Para obtener un sistema de ecuaciones triangular superior, mantenemos fijas la primera y segunda ecuaciones del sistema precedente, eliminemos la incógnita y de la tercera ecuación.

Sea $k_2 = -\frac{\frac{11}{2}}{\frac{5}{2}} = -\frac{11}{5}$ (k_2 se obtiene dividiendo el coeficiente de y de la tercera ecuación para el coeficiente de y de la segunda ecuación, cambiado de signo). Multiplicamos a la segunda ecuación por k_2, el resultado sumamos a la tercera, resulta $\begin{cases} 2x+y = -2 \\ \frac{5}{2}y+z=1 \\ -\frac{6}{5}z=-\frac{6}{5}, \end{cases}$ con lo que hemos obtenido un sistema de ecuaciones triangular superior. Determinemos su solución. De la tercera ecuación, obtenemos $z=1$. De la segunda ecuación, se obtiene y : $y=\frac{2}{5}(1-z)=\frac{2}{5}(1-1)=0$. De la primera ecuación se deduce x: $x=\frac{1}{2}(-2-y)=\frac{1}{2}(-2-0)=-1$. La solución del sistema de ecuaciones lineales propuesto es $(-1,\ 0,\ 1)$.

3. Hallar la solución si existe, del sistema de ecuaciones lineales que se propone: $(x,y,z)\in\mathbb{R}^3$ tal que $\begin{cases} y+z=-2 \\ -2x+y-z=6 \\ 5x+y+6z=10. \end{cases}$ Para obtener (siempre que sea posible) un sistema triangular superior, la primera acción que debemos realizar es intercambiar las ecuaciones del modo siguiente: $\begin{cases} 5x+y+6z &= 10 \\ -2x+y-z &= 6 \\ y+z &= -2 \end{cases}$ Mantengamos fija la primera ecuación de este último sistema de ecuaciones lineales. Eliminemos x de la segunda ecuación. Para ello multiplicamos la primera ecuación por $k = -\frac{-2}{5} = \frac{2}{5}$ y el resultado sumamos a la segunda. Obtenemos $\begin{cases} 5x+y+6z=10 \\ \frac{7}{5}y+\frac{7}{5}z=10 \\ y+z=-2. \end{cases}$ Manteniendo fijas las dos primeras ecuaciones, eliminemos y de la tercera ecuación. Multipliquemos por $k_1 = -\frac{\frac{1}{7}}{\frac{5}{}} = -\frac{5}{7}$ a la segunda ecuación y sumemos con la tercera: $\begin{cases} 5x+y+6z=10 \\ \frac{7}{5}y+\frac{7}{5}z=10 \\ 0=-\frac{64}{7}, \end{cases}$ que muestra que la tercera igualdad es contradictoria, es decir que el sistema de ecuaciones propuesto no tiene solución.

1.3.2. Cálculo con funciones

Un polinomio P de grado $\leq n$ con coeficientes reales lo denotamos como sigue:

$$P(x) = a_0 + a_1 x + \cdots + a_n x^n = \sum_{k=0}^{n} a_k x^k \qquad x\in\mathbb{R},$$

donde $a_k \in \mathbb{R}$ con $k = 0, 1, \cdots, n$ son los coeficientes y $a_n \neq 0$.

En orden de complejidad, los más simples son los polinomios constantes $P(x) = c \quad x \in \mathbb{R}$, con $c \in \mathbb{R}$ fijo. A continuación, los polinomios de grado 1 tienen la forma $P(x) = a + bx$, con $a, b, x \in \mathbb{R}$, a, b fijos y $b \neq 0$. Los polinomios de grado 2 se escriben como $P(x) = a + bx + cx^2$ con $a, b, c, x \in \mathbb{R}$, a, b, c fijos y $c \neq 0$. Los polinomios de grado tres se escriben como $P(x) = a + bx + cx^2 + dx^3$ con $a, b, c, d, x \in \mathbb{R}$, a, b, c, d fijos y $d \neq 0$.

Los polinomios son las funciones reales más simples de calcularse en un asignado dato $x \in \mathbb{R}$. Es claro que los más simples son los polinomios constantes y de grado 1, y realizar cálculos con esta clase de polinomios no presenta dificultad alguna. Nos interesamos en los polinomios de grado ≥ 2 que presentan alguna dificultad con los cálculos pués a medida que el grado del polinomio es más grande, el número de operaciones elementales se incrementa y los resultados pueden no ser suficientemente exactos.

1. Esquema de Hörner.

   Con frecuencia requerimos realizar evaluaciones de polinomios de modo que el número total de operaciones elementales a realizar sea el más pequeño posible y que el resultado sea el más exacto posible. Por razones que veremos más adelante y que están relacionadas con el condicionamiento, debemos evitar el cálculo directo de las potencias de x, de los factoriales, sumas y restas alternadas.

   Consideremos el polinomio $P(x) = a_0 + a_1 x + \cdots + a_n x^n = \sum_{k=0}^{n} a_k x^k \quad x \in \mathbb{R}$, donde $a_k \in \mathbb{R}$ con $k = 0, 1, \cdots, n$ son los coeficientes y $a_n \neq 0$. Nos interesamos primeramente en el cálculo de $P(x)$ en un asignado $x \in \mathbb{R}$, de modo que se evite el cálculo directo de las potencias de x y el número de operaciones elementales sea el más pequeño posible. Esto se logra si se escribe $P(x)$ en la forma siguiente:

$$\begin{aligned} P(x) &= a_0 + x\left(a_1 + \cdots + a_n x^{n-1}\right) \\ &\vdots \\ &= a_0 + x(a_1 + x(a_2 + x(a_3 + \cdots + x(a_{n-1} + xa_n)\cdots))). \end{aligned}$$

   A esta forma de calcular $P(x)$ se conoce con el nombre de esquema de Hörner. Utilizando esta escritura, podemos elaborar un algoritmo para calcular $P(x)$ en un punto dado $x \in \mathbb{R}$. Note que el proceso de cálculo de $P(x)$ inicia en el término del paréntesis interior $a_{n-1} + xa_n$ y continua sucesivamente al exterior, que hace el proceso de cálculo sea muy práctico en su aplicación. El número de operaciones elementales (sumas y productos) que se requiere para calcular $P(x)$ es a lo más $2n$.

   Note que si $x \in \mathbb{R}$, el cálculo de $x^2 = x \times x$ significa una operación elemental, el cálculo de $x^3 = x^2 \times x$ significa dos operaciones elementales, entonces para el cálculo del polinomio $P(x) = a + bx + cx^2$ se requieren de 5 operaciones (sumas y productos), mientras que si se escribe en la forma $P(x) = a + x(b + cx)$ se requieren únicamente de 4 operaciones (sumas y productos). Para el cálculo del polinomio $P(x) = a + bx + cx^2 + dx^3$ se requieren de 9 operaciones y con el esquema de Hörner se requieren de 6 operaciones y mejora la exactitud del resultado.

   Para elaborar un algoritmo que permita calcular $P(x)$ requerimos de la siguiente información: grado del polinomio $n \in \mathbb{Z}^+$, coefiecientes $a_0, a_1, \ldots, a_n \in \mathbb{R}$ y del dato $x \in \mathbb{R}$. Con estos elementos proponemos el siguiente algoritmo que se conoce con el nombre de esquema de Hörner.

   Algoritmo

   Datos de entrada: $n \in \mathbb{Z}^+$, $a_0, a_1, \ldots, a_n, x \in \mathbb{R}$.

   Datos de salida: x, $P(x)$.

   1. $b = a_n$
   2. $k = 0, 1, \ldots, n-1$

      $j = n - k$

      $z = a_{j-1} + xb$

      $b = z$

Fin bucle k.

3. Imprimir x, $b = P(x)$.
4. Fin.

Note que el cálculo de $P(x)$ concluye en un número finito de pasos. Verifiquemos el algoritmo propuesto. Para el efecto, consideramos el siguiente polinomio que a su vez lo escribimos usando el esquema de Hörner:

$$P(x) = 0{,}5 + 0{,}3x - 0{,}25x^2 - 2{,}56x^3 + 3x^4 = 0{,}5 + x(0{,}3 + x(-0{,}25 + x(-2{,}56 + 3x))) \quad x \in \mathbb{R}.$$

Calculemos $P(x)$ en los puntos $x = 0$, $0{,}5$ y $-0{,}5$. Utilizando la escritura de $P(x)$, tenemos

$$\begin{aligned} P(0.) &= 0{,}5 + 0.(0{,}3 + 0.(-0{,}25 + 0.(-2{,}56 + 3 \times 0.))) = 0{,}5, \\ P(0{,}5) &= 0{,}5 + 0{,}5(0{,}3 + 0{,}5(-0{,}25 + 0{,}5(-2{,}56 + 3 \times 0{,}5))) = 0{,}455, \\ P(-0{,}5) &= 0{,}5 - 0{,}5(0{,}3 - 0{,}5(-0{,}25 - 0{,}5(-2{,}56 - 3 \times 0{,}5))) = 0{,}795. \end{aligned}$$

2. Consideremos la función E de $\mathbb{R}^+$ en $\mathbb{R}$ definida como $E(x) = \displaystyle\sum_{k=0}^{n} \frac{x^k}{(k+1)(k+2)(k+a)} \qquad x \in \mathbb{R}^+,$

donde $a \in \mathbb{R}^+$ fijo.

Dado $x \in \mathbb{R}$, para calcular $E(x)$, primeramente debemos expresar en forma explícita el sumatorio y luego escribirle en forma del esquema de Hörner como a continuación se muestra:

$$\begin{aligned} E(x) &= \frac{1}{1 \times 2 \times a} + \frac{x}{2 \times 3 \times (1+a)} + \frac{x^2}{3 \times 4 \times (2+a)} + \cdots + \frac{x^{n-1}}{n(n+1)(n-1+a)} + \\ &\quad + \frac{x^n}{(n+1)(n+2)(n+a)} \\ &= \frac{1}{2a} + x\left(\frac{1}{2 \times 3(1+a)} + x\left(\frac{1}{3 \times 4(2+a)} + \cdots + x\left(\frac{1}{n(n+1)(n-1+a)} + \right.\right.\right. \\ &\quad \left.\left.\left. + \frac{x}{(n+1)(n+2)(n+a)}\right)\cdots\right)\right). \end{aligned}$$

Note que en la última igualdad se evitan los cálculos directos de las potencias x^k, $k = 2, \ldots, n$, lo que reduce el número de operaciones elementales, facilita la escritura de un algoritmo para su cálculo.

Ponemos $a_k = \dfrac{1}{(k+1)(k+2)(k+a)}$, $k = 0, 1, 2, ..., n$. En el cálculo de a_k intervienen 3 adiciones, 3 productos y una división.

Algoritmo

Datos de entrada: n, a, x.

Datos de salida: x, $E(x)$.

1. $y = \dfrac{1}{(n+1)(n+2)(n+a)}$.
2. $k = 1, ..., n$

   $j = n - k$

   $y = \dfrac{1}{(k+1)(k+2)(k+a)} + y * x.$

   Fin bucle k.
3. Imprimir resultado $y = E(x)$.
4. Fin.

Observe que el cálculo de $E(x)$ concluye en un número finito de pasos.

3. Para cada $n \in \mathbb{Z}^+$ con $n \geq 3$ impar, se considera la función φ_n definida como sigue:

$$\varphi_n(x) = \sum_{k=0}^{n} \frac{(-1)^k x^{\frac{k}{2}}}{k! 3^k} \quad x \geq 0.$$

Se trata de calcular $\varphi_n(x)$ de modo que el número de operaciones elementales sea el más pequeño posible y elaborar un algoritmo de cálculo que permita calcular $\varphi_n(x_k) \quad k = 0, 1, \ldots, m$ en puntos x_k igualmente espaciados en el intervalo $[0, 100]$.

Sigamos la metodología utilizada para resolver problemas. Primeramente debemos constatar que se tienen soluciones. En efecto, la función φ_n está bien definida para todo $x \geq 0$. Además, de la definición de $\varphi_n(x)$ se tiene

$$\varphi_n(x) = 1 - \frac{x^{\frac{1}{2}}}{1! \times 3} + \frac{x}{2! \times 3^2} - \frac{x^{\frac{3}{2}}}{3! \times 3^3} + \cdots + \frac{(-1)^{n-1}}{(n-1)! \times 3^{n-1}} + \frac{(-1)^n}{n! \times 3^n} x^{\frac{n}{2}}.$$

Más adelante veremos que la resta de números positivos muy próximos entre sí es una operación peligrosa pués los errores de redondeo son amplificados, más aún, la realización de sumas y restas alternadas es muy peligrosa ya que los errores de redondeo provocan grandes errores en los datos de salida. Vemos que en el cálculo de $\varphi_n(x)$ debemos realizar este tipo de operaciones, además, se deben calcular los factoriales $k! \quad k = 1, 2, \ldots, n$, las potencias 3^k, $x^{\frac{k}{2}}$.

Puesto que n es impar, $n-1$ es par y en consecuencia $p = \dfrac{n-1}{2}$ es un entero positivo, asociamos todos los términos positivos y todos los términos negativos. Cada grupo contiene exactamente $p+1$ términos. Así

$$\begin{aligned}
\varphi_n(x) &= 1 + \frac{x}{2! \times 3^2} + \cdots + \frac{x^{\frac{n-1}{2}}}{(n-1)! \times 3^{n-1}} - \left[\frac{x^{\frac{1}{2}}}{1! \times 3} + \frac{x^{\frac{3}{2}}}{3! \times 3^3} + \cdots + \frac{x^{\frac{n}{2}}}{n! \times 3^n} \right] \\
&= \sum_{k=0}^{p} \frac{x^k}{(2k)! \times 3^{2k}} - \sum_{k=0}^{p} \frac{x^{k+\frac{1}{2}}}{(2k+1)! \times 3^{2k+1}} \\
&= \sum_{k=0}^{p} \frac{1}{(2k)!} \left(\frac{x}{9}\right)^k - \frac{\sqrt{x}}{3} \sum_{k=0}^{p} \frac{1}{(2k+1)!} \left(\frac{x}{9}\right)^k \quad x \geq 0.
\end{aligned}$$

Definimos

$$\theta_1(x) = \sum_{k=0}^{p} \frac{1}{(2k)!} \left(\frac{x}{9}\right)^k \quad x \geq 0, \qquad \theta_2(x) = \sum_{k=0}^{p} \frac{1}{(2k+1)!} \left(\frac{x}{9}\right)^k \quad x \geq 0.$$

En forma explícita, $\theta_1(x)$ se escribe como sigue:

$$\begin{aligned}
\theta_1(x) &= 1 + \frac{x}{2! \times 9} + \frac{x^2}{4! \times 9^2} + \cdots + \frac{x^{p-1}}{[2(p-1)]! \times 9^{p-1}} + \frac{x^p}{(2p)! \times 9^p} \\
&= 1 + \frac{1}{2}\frac{x}{9}\left(1 + \frac{1}{3 \times 4}\frac{x}{9}\left(1 + \cdots + \frac{1}{(2p-3)(2p-2)}\frac{x}{9}\left(1 + \frac{1}{(2p-1)(2p)}\frac{x}{9}\right)\cdots\right)\right).
\end{aligned}$$

Procediendo en forma similar con $\theta_2(x)$, obtenemos

$$\begin{aligned}
\theta_2(x) &= \frac{1}{1!} + \frac{1}{3!}\frac{x}{9} + \frac{1}{5!}\frac{x^2}{9^2} + \cdots + \frac{1}{(2p-1)!}\frac{x^{p-1}}{9^{p-1}} + \frac{1}{(2p+1)!}\frac{x^p}{9^p} \\
&= 1 + \frac{1}{3!}\frac{x}{9}\left(1 + \frac{1}{4 \times 5}\frac{x}{9}\left(1 + \cdots + \frac{1}{(2p-2)(2p-1)}\frac{x}{9}\left(1 + \frac{1}{(2p)(2p+1)}\frac{x}{9}\right)\cdots\right)\right).
\end{aligned}$$

Si ponemos $y = \dfrac{x}{9}$, $\theta_1(x)$ y $\theta_2(x)$ se escriben como

$$\begin{aligned}
\theta_1(x) &= 1 + \frac{y}{2}\left(1 + \frac{y}{3 \times 4}\left(1 + \cdots + \frac{y}{(2p-3)(2p-2)}\left(1 + \frac{y}{(2p-1)(2p)}\right)\cdots\right)\right), \\
\theta_2(x) &= 1 + \frac{y}{6}\left(1 + \frac{6}{4 \times 5}\left(1 + \cdots + \frac{6}{(2p-2)(2p-1)}\left(1 + \frac{6}{(2p)(2p+1)}\right)\cdots\right)\right).
\end{aligned}$$

La escritura de θ_1 y θ_2 es una variante del esquema de Hörner que evita el cálculo directo de los factoriales $k!\quad k=1,2,\ldots,n$, de las potencias x^k y 3^k. De esta manera reduce significativamente el número de operaciones elementales y permite elaborar un algritmo numérico en forma muy simple. Además,

$$\varphi_n(x)=\theta_1(x)-\frac{\sqrt{x}}{3}\theta_2(x)\qquad x\geq 0,$$

el cálculo de $\varphi_n(x)$ implica una sola resta y no sumas y restas como originalmente se tenía en el cálculo de $\varphi_n(x)$. Note también que los cocientes $\dfrac{x}{9}$ que tanto en $\theta_1(x)$ como en $\theta_2(x)$ se realizan, se evitan con el cálculo de $y=\dfrac{x}{9}$. El número de operaciones elementales que se realizan para calcular $\theta_1(x)$ y $\theta_2(x)$ son a lo más de $2\times(6p)=6(n-1)$, y el cálculo de $\varphi_n(x)$ requiere de a lo más $6(n-1)+5=6n-1$ operaciones elementales.

Si se debe calcular $\varphi_n(x)$ en la forma original se deben realizar a lo más $-1+\dfrac{n}{2}(1+3n)$ operaciones elementales.

Observe que el cálculo de $\dfrac{(\sqrt{x})^k}{k!\times 3^k}\qquad k=3,\ldots,n$ requiere de $3k-2$ operaciones elementales pués $(k-1)!k=k!$ corresponde cada una operación elemental. Análogamente $a^{k-1}a=a^k$ es una operación elemental.

Para $n=7$ se requieren 77 operaciones elementales, mientras que con la forma simplificada se requieren a lo más 41 operaciones elementales. Para $n=15$, en la forma original se requieren aproximadamente 345 operaciones elementales, mientras que en la forma simplificada requieren aproximadamente 89 operaciones elementales.

Cuando n es grande se presenta otras dificultades de cálculo de $n!$, 3^n, x^n, por ejemplo $15!\simeq 1{,}30767436\times 10^{12}$, $\quad 5^{15}\simeq 3{,}051757812\times 10^{10}$.

Finalmente, como se debe calcular $\varphi_n(x_k)$ en puntos igualmente espaciados x_k de $[0,100]$, se define $h=\dfrac{100}{m}$ y $x_k=kh\qquad k=0,1,\ldots,m.$

Con todos estos resultados se propone el siguiente algoritmo de cálculo de $\varphi_n(x_k)\quad k=0,1,\ldots,m$, y, $n\in\mathbb{Z}^+$ con $n\geq 3$ impar.

Algoritmo

Datos de entrada: $m,n\in\mathbb{Z}^+,\ x_k\quad k=0,1,\ldots,m.$

Datos de salida: x_k, $\varphi_n(x_k)\qquad k=0,1,\ldots,m.$

1. Verificar $n\geq 3$ impar. Caso contrario continuar en 6).
2. $h=\dfrac{100}{m}$.
3. Para $x=0$, poner $\varphi_n(0)=1$.
4. Para $k=1,\ldots,m$

   $b=1.$

   $c=1.$

   $x_k=kh$

   $y=\dfrac{x_k}{9}$

   $j=0,1,\ldots,p-1$

   $i=p-j,$

   $b=1+\dfrac{1}{(2i-1)(2i)}\times y\times b,$

   $c=1+\dfrac{1}{(2i)(2i+1)}\times y\times b,$

   Fin de bucle j.

$$\varphi_n(x_k) = b - \frac{\sqrt{x_k}}{3}c.$$

Fin de bucle k.

5. Imprimir $x_k,\ \varphi_n(x_k) \quad k = 0, 1, \ldots, m.$
6. Mensaje: Error de lectura de n.
7. Fin.

Para $n = 3$, la función $\varphi_3(x)$ está definida como

$$\varphi_3(x) = 1 - \frac{\sqrt{x}}{1! \times 3} + \frac{x}{2! \times 3^2} - \frac{\left(\sqrt{x}\right)^3}{3! \times 3^3} \quad x \geq 0.$$

Calculemos $\varphi_3(10)$. Tenemos

$$\begin{aligned} \varphi_3(10) &= 1 - \frac{\sqrt{10}}{1! \times 3} + \frac{10}{2! \times 3^2} - \frac{\left(\sqrt{10}\right)^3}{3! \times 3^3} \\ &= 1 - 1{,}054092553 + 0{,}555555556 - 0{,}1952023247 = 0{,}3062606775. \end{aligned}$$

Para esta cálculo se requieren de 15 operaciones elementales. Note las molestias en la realización de los cálculos en $\varphi_3(x)$. Apliquemos el algoritmo. Ponemos $y = \dfrac{10}{9} \simeq 1{,}111111111$.

$$\begin{aligned} \varphi_3(x) &= 1 + \frac{y}{2} - \frac{\sqrt{10}}{3}\left(1 + \frac{y}{6}\right), \\ \varphi_3(10) &= 1{,}555555556 - \frac{\sqrt{10}}{3} \times 1{,}185185185 = 0{,}306260678. \end{aligned}$$

Se requieren de 9 operaciones lementales.

Para $n = 7$, $\varphi_7(x)$ está definido como

$$\varphi_7(x) = 1 - \frac{\sqrt{x}}{1! \times 3} + \frac{x}{2! \times 3^2} - \frac{\left(\sqrt{x}\right)^3}{3! \times 3^3} + \frac{x^2}{4! \times 3^4} - \frac{\left(\sqrt{x}\right)^5}{5! \times 3^5} + \frac{x^3}{6! \times 3^6} - \frac{\left(\sqrt{x}\right)^7}{7! \times 3^7}$$

que se escribe como

$$\varphi_7(x) = 1 + \frac{y}{2}\left(1 + \frac{y}{12}\left(1 + \frac{y}{30}\right)\right) - \frac{\sqrt{x}}{3}\left(1 + \frac{y}{6}\left(1 + \frac{y}{20}\left(1 + \frac{y}{42}\right)\right)\right)$$

donde $y = \dfrac{x}{9}$. Para $x = 10$, $y = \dfrac{10}{9} \simeq 1{,}111111111$, y aplicando el algoritmo, obtenemos

$$\varphi_7(x) = 1{,}608901082 - 1{,}260426345 = 0{,}348474737.$$

En el siguiente capítulo se tratan las series de potencias, las mismas que se aproximan con sumas finitas, las que a su vez se escriben siguiendo un procedimiento similar al discutido en el ejemplo que acabamos de presentar.

Note que $\varphi(x) = \exp(-\dfrac{\sqrt{x}}{3}) = \sum\limits_{k=0}^{\infty} \dfrac{(-1)^k x^{\frac{k}{2}}}{k! 3^k} \quad x \geq 0$, y la función φ_n es la suma parcial del desarrollo en serie de potencias de φ. Para $x = 10$, los valores que hemos calculado $\varphi_n(10)$ son aproximaciones de $\varphi(10)$:

$$\varphi(10) = \exp(-\frac{\sqrt{10}}{3}) \simeq 0{,}3485085369.$$

4. Este es un ejemplo de una función que posee una discontinuidad evitable.

Se define la función real φ como sigue: $\varphi(x) = \dfrac{\left(1 + x^4\right)^{\frac{1}{3}} - \left(1 - x^4\right)^{\frac{1}{3}}}{x^4} \quad 0 < |x| \leq 1$. Se desea calcular $\varphi(x)$ para $x \in]0, 1[$.

Suponemos que con una calculadora de bolsillo (calculadora hipotética) se tiene $10^{-100} \simeq 0$ pero $10^{-99} \not\simeq 0$. Entonces, para $0 < |x| \leq 10^{-25}$ se tiene $0 < x^4 \leq 10^{-100} \simeq 0$ y no podemos calcular

$\varphi(x)$. Para resolver este inconveniente, aplicamos el binomio de Newton con exponente racional que se define a continuación:

$$(1+a)^r = 1 + ra + \frac{r(r-1)}{2!}a^2 + \frac{r(r-1)(r-2)}{3!}a^3 + \cdots,$$

donde $r \in \mathbb{Q}$ con $r \neq 0,\ |a| < 1$.

Apliquemos el binomio de Newton a $\left(1+x^4\right)^{\frac{1}{3}}$ y $\left(1-x^4\right)^{\frac{1}{3}}$ para $0 < |x| < 1$, tenemos

$$\begin{aligned}
\left(1+x^4\right)^{\frac{1}{3}} &= 1 + \frac{1}{3}x^4 + \frac{\frac{1}{3}\left(\frac{1}{3}-1\right)}{2!}x^8 + \frac{\frac{1}{3}\left(\frac{1}{3}-1\right)\left(\frac{1}{3}-2\right)}{3!}x^{12} \\
&\quad + \frac{\frac{1}{3}\left(\frac{1}{3}-1\right)\left(\frac{1}{3}-2\right)\left(\frac{1}{3}-3\right)}{4!}x^{16} + \frac{\frac{1}{3}\left(\frac{1}{3}-1\right)\left(\frac{1}{3}-2\right)\left(\frac{1}{3}-3\right)\left(\frac{1}{3}-4\right)}{5!}x^{20} \\
&\quad + \cdots \\
&= 1 + \frac{1}{3}x^4 - \frac{1}{9}x^8 + \frac{5}{81}x^{12} - \frac{10}{243}x^{16} + \frac{22}{729}x^{20} + \cdots, \\
\left(1-x^4\right)^{\frac{1}{3}} &= 1 - \frac{1}{3}x^4 - \frac{1}{9}x^8 - \frac{5}{81}x^{12} - \frac{10}{243}x^{16} - \frac{22}{729}x^{20} + \cdots.
\end{aligned}$$

Entonces

$$\left(1+x^4\right)^{\frac{1}{3}} - \left(1+x^4\right)^{\frac{1}{3}} = \frac{2}{3}x^4 + \frac{10}{81}x^8 + \frac{44}{729}x^{16} + \cdots,$$

de donde

$$\varphi(x) = \frac{\left(1+x^4\right)^{\frac{1}{3}} - \left(1-x^4\right)^{\frac{1}{3}}}{x^4} = \frac{2}{3} + \frac{10}{81}x^8 + \frac{44}{729}x^{16} + \cdots \quad 0 < |x| < 1.$$

En esta nueva formulación de la función φ, vemos que se ha eliminado el inconveniente de cálculo que arriba señalamos. En realidad se tiene una discontinuidad evitable en $x = 0$. Tenemos

$$\lim_{x\to 0}\varphi(x) = \frac{2}{3}, \quad \text{luego } \varphi(x) \simeq \frac{2}{3} \text{ si } 0 < |x| \leq 10^{-25}$$

Más aún, si $0 < |x| \leq 10^{-\frac{25}{2}}$, $0 < |x|^8 \leq 10^{-100} \simeq 0$, y $\varphi(x)$ se aproxima como $\varphi(x) \simeq \frac{2}{3}$.

Para x tal que $10^{-\frac{25}{2}} < |x| \leq 10^{-\frac{25}{4}}$, se tiene $10^{-200} < |x|^{16} \leq 10^{-100} \simeq 0$, luego $\varphi(x)$ se aproxima como $\varphi(x) \simeq \frac{2}{3} + \frac{10}{81}x^8$.

Si $10^{-\frac{25}{4}} < |x| \leq 10^{-\frac{100}{24}} \simeq 6{,}812920691 \times 10^{-5}$, entonces $\varphi(x)$ se aproxima como

$$\varphi(x) \simeq \frac{2}{3} + \frac{10}{81}x^8 + \frac{44}{729}x^{16} = \frac{2}{3} + x^8\left(\frac{10}{81} + \frac{44}{729}x^8\right).$$

Para x tal que $10^{-\frac{100}{24}} < |x| \leq 10^{-\frac{100}{32}}$ se tiene $10^{-\frac{400}{3}} < |x|^{32} \leq 10^{-100} \simeq 0$, en cuyo caso

$$\varphi(x) \simeq \frac{2}{3} + \frac{10}{81}x^8 + \frac{44}{729}x^{16} + \frac{718}{19683}x^{24} = \frac{2}{3} + x^8\left(\frac{10}{81} + x^8\left(\frac{44}{729} + \frac{718}{19683}x^8\right)\right).$$

Así sucesivamente.

Para x tal que $10^{-1} < |x| \leq 1$, calculamos $\varphi(x)$ con la expresión que se definió originalmente. Así,

$$\begin{aligned}
\varphi(0{,}1) &= \frac{\left(1+(0{,}1)^2\right)^{\frac{1}{3}} - \left(1-(0{,}1)^4\right)^{\frac{1}{3}}}{(0{,}1)^4} \simeq \frac{1{,}000033332 - 0{,}9999666656}{(0{,}1)^4} \simeq 0{,}6666664, \\
\varphi(0{,}2) &= \frac{\left(1+(0{,}2)^2\right)^{\frac{1}{3}} - \left(1-(0{,}2)^4\right)^{\frac{1}{3}}}{(0{,}1)^4} \simeq \frac{1{,}000533049 - 0{,}999466382}{(0{,}2)^4} \simeq 0{,}666666875.
\end{aligned}$$

Note que si se utiliza el desarrollo de $\varphi(x) = \frac{2}{3} + \frac{10}{81}x^8 + \cdots$, se obtiene $\varphi(0{,}2) \simeq 0{,}6666669827$ que es mucho más exacto que el precedente.

5. En este ejemplo se trata un método de derivación numérica.

Sea f una función real derivable en el intervalo $]a,b[\,,\ x_0 \in]a,b[\,.$ La derivada de f en x_0 se nota $f'(x_0)$ o también $\dfrac{df}{dx}(x_0)$ y se define como

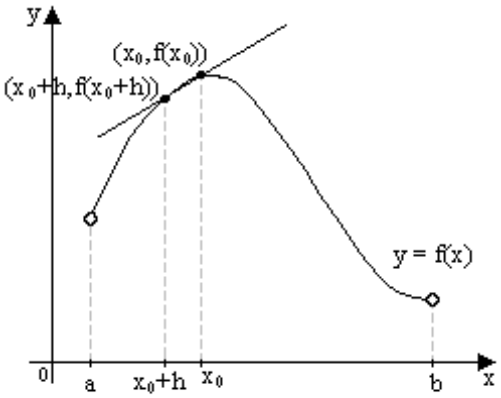
$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h).f(x_0)}{h},$$

siempre que el límite exista. El cociente $\dfrac{f(x_0+h).f(x_0)}{h} \quad h \neq 0$, se llama cociente incremental.

Admitiremos que la función f es derivable en algún intervalo abierto $]a,b[$ de $\mathbb{R}$ y nos proponemos calcular numéricamente $f'(x_0)$. Sea $h \in \mathbb{R}$ con $h \neq 0$ suficientemente pequeño. De la definición de $f'(x_0)$ surge inmediatamente la idea de aproximar $f'(x_0)$ mediante el cociente incremental, esto es,

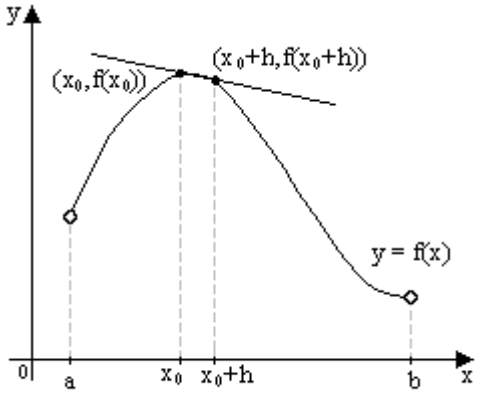
$$f'(x_0) \simeq \frac{f(x_0+h).f(x_0)}{h}.$$

En la figura siguiente se muestra la gráfica de una función f definida en $]a,b[$ y la recta secante que une los puntos $(x_0, f(x_0))$ y $(x_0+h, f(x_0+h))$ en los casos $h<0$ y $h>0$.



h<0

Figura 5



h>0

Figura 6

Ponemos $y_0 = f(x_0)\,,\ y_1 = f(x_0+h)$ y y_0' una aproximación de $f'(x_0)$ definida como

$$y_0' = \frac{y_1 - y_0}{h},$$

y_0' la denominaremos derivada numérica de $f'(x_0)$.

Supongamos que f posee derivada segunda en $]a,b[$. El polinomio de Taylor con resto de f está definido como

$$f(x_0+h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(\xi)\,,$$

con $h \neq 0$ y ξ entre x_0 y x_0+h, entonces

$$f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h} - \frac{h}{2!}f''(\xi)\,.$$

La aproximación de $f'(x_0)$ se escribe como

$$y_0' = \frac{y_1 - y_0}{h} + 0(h)\,.$$

Con frecuencia $y_0 = f(x_0)\,,\ y_1 = f(x_0+h)$ no se calculan exactamente, consideraremos $y_0,\ y_1$ aproximaciones de $f(x_0)$ y $f(x_0+h)$ respectivamente.

Ejemplo

Consideremos la función real definida como $f(x) = x^4 \quad x \in \mathbb{R}$. Claramente f es derivable. Calculemos numéricamente $f'(1{,}5)$. Ponemos $x_0 = 1{,}5$. En la tabla siguiente se muestran valores de $h,\ x_0 + h,\ y_0 = f(x_0),\ y_1 = f(x_0 + h),\ f'(x_0) \simeq y_0' = \dfrac{y_1 - y_0}{h}$.

h	$x_0 + h$	y_0	y_1	$f'(x_0) \simeq y_0' = \dfrac{y_1 - y_0}{h}$.
$-0{,}1$	1,4	5,0625	3,8416	12,209
$-0{,}005$	1,495	5,0625	4,995336751	13,4326498
$-0{,}00005$	1,499995	5,0625	5,0624325	13,5
0,05	1,55	5,0625	5,77200625	14,190125
0,0005	1,5005	5,0625	5,069253376	13,506752
0,000005	1,500005	5,0625	5,0625675	13,5

El valor exacto de $f'(1{,}5)$ es 13,5.

En base a la definición de derivada numérica así como al proceso de cálculo seguido en el ejemplo, se propone como ejercicio elaborar el algoritmo correspondiente.

En un capítulo más adelante se verán otros métodos numéricos de cálculo de $f'(x_0)$ y de derivadas de orden superior. Igualmente, se tratará el cálculo aproximado de derivadas parciales.

6. En este ejemplo se considera un método de integración aproximada.

Sean $a, b \in \mathbb{R}$ con $a < b$, u una función real continua definida en $[a, b]$. Se considera el siguiente problema:

$$\text{hallar } I(u) = \int_a^b u(x)\, dx.$$

Como es conocido, la integral definida de una función continua está bien definida. Para el cálculo de $I(u)$ se consideran dos casos: el primero en el que podemos encontrar una función primitiva F de u, esto es, una función F tal que $F'(x) = u(x) \quad \forall x \in [a, b]$ y en consecuencia $I(u) = F(b) - F(a)$. En el segundo caso, no podemos encontrar una función primitiva de u, con lo que el cálculo de $I(u)$ debemos realizarlo en forma aproximada. Para el efecto, elegimos el método conocido como la regla del rectángulo que describimos a continuación.

Sean $m \in \mathbb{Z}^+$ y $\tau(m) = \{x_0 = a,\ x_1, \ldots, x_m = b\}$ una partición de $[a, b]$, esto es, $x_{i-1} < x_i \quad i = 1, \ldots, m$. Ponemos $h_i = x_i - x_{i-1} \quad i = 0, 1, \ldots, m$, y $\widehat{h}$=max$\{h_i \mid i = 1, \ldots, m\}$. Si se elige $h = \dfrac{b-a}{m}$ y $x_i = ih \quad i = 0, 1, \ldots, m$, $\tau(m)$ se dice partición uniforme. Se tiene $h_i = h$ y $\widehat{h} = h$. En general estas particiones son las más comunes.

Se define la función real v_n sobre $[a, b]$ como sigue

$$\begin{cases} v_m(x) = u(t_i) & x \in [x_{i-1}, x_i[\,, \quad i = 1, \ldots, m, \\ v_m(b) = u(b)\,, \end{cases}$$

donde $t_i = x_{i-1} + \dfrac{1}{2} h_i$ es el punto medio del intervalo $[x_{i-1}, x_i]$. La función v_m se le llama interpolante de u.

En la figura siguiente se muestra la gráfica de una función u definida en $[a, b]$, la partición $\tau(m)$

de $[a,b]$ con $m=5$ y la función interpolante v_m de u.

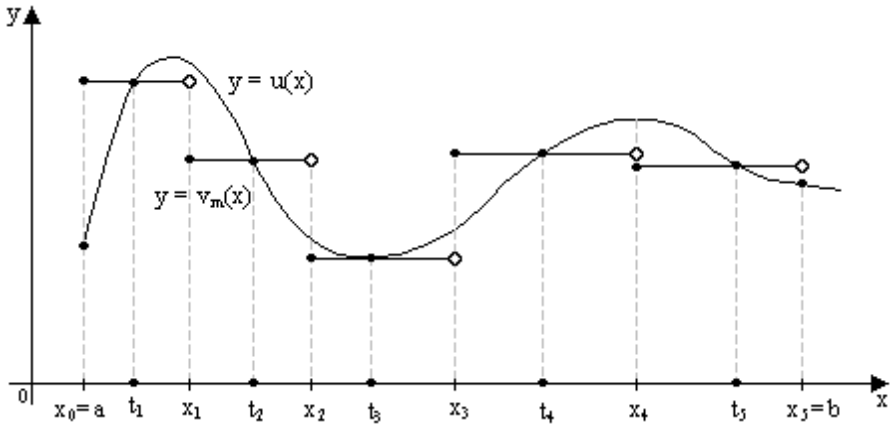


Figura 7

Entonces

$$\begin{aligned} I(v_m) &= \int_a^b v_m(x)\,dx = \sum_{i=1}^{m}\int_{x_{i-1}}^{x_i} v_m(x)\,dx = \sum_{i=1}^{m}\int_{x_{i-1}}^{x_i} u\left(x_{i-1}+\frac{1}{2}h_i\right)dx \\ &= \sum_{i=1}^{m} h_i u\left(x_{i-1}+\frac{1}{2}h_i\right). \end{aligned}$$

La aproximación $I(v_m)$ de $I(u)$ se llama regla del rectángulo. Note que el problema de cálculo de la integral $I(u)$ se le ha tansformado en uno más sencillo que es calcular $I(v_m)=\sum_{i=1}^{m} h_i u\left(x_{i-1}+\frac{1}{2}h_i\right)$

Puesto que la función u se ha discretizado según la partición $\tau(m)$ de $[a,b]$, se tiene un conjunto de puntos $u(a)$, $u(t_i)\quad i=1,\ldots,m$, $u(b)$; en el cálculo de $I(u)$ se comete un error de discretización. En análisis numérico interesa mucho estimar el error de discretización en el cálculo numérico de integrales definidas y particularmente del método de la regla del rectángulo, esto es, estimar $|I(u)-I(v_m)|$ y probar que $I(v_m) \underset{m\to\infty}{\to} I(a)$, en un cápitulo posterior se tratarán todos estos problemas.

Algoritmo

Datos de entrada: $a,b\in\mathbb{R}$, $m\in\mathbb{Z}^+$, función u.

Datos de salida: $I(v_m)$, mensaje.

1. Verificar $a<b$. Caso contrario, continuar en 8).
2. Calcular $h=\dfrac{b-a}{m}$.
3. $j=0,1,\ldots,m$

   $x_j=a+jh$

   Fin de bucle j.
4. $S=0$.
5. $j=1,\ldots,m$

   $t_j=x_{j-1}+0{,}5h$

   $S=S+u(t_j)$

   Fin de bucle j.
6. $I(v_m)=hS$.
7. Imprimir $I(v_m)$. Continuar en 9).

8. Mensaje: $a < b$.

9. Fin.

Como aplicación de la regla del rectángulo consideramos la función u definida como $u(x) = x^3 \quad x \in [1,2]$ y $m = 10$. Se considera la partición uniforme. Se define $h = \dfrac{1}{10} = 0{,}1$, la partición $\tau(10)$ del intervalo $[1,2]$ está definida como $\tau(10) = \{1, 1{,}1, 1{,}2, \ldots, 1{,}9, 1\}$. Entonces

$$\begin{aligned} I(v_{10}) &= \sum_{i=1}^{10} h_i u(t_i) = \sum_{i=1}^{10} h_i u\left(x_{i-1} + \frac{1}{2}h_i\right) = \sum_{i=1}^{10} hu(x_{i-1} + 0{,}05) \\ &= 0{,}1\left[u(1{,}05) + u(1{,}15) + \cdots + u(1{,}95)\right] = 3{,}74625. \end{aligned}$$

El valor exacto es

$$I(u) = \int_1^2 u(x)\,dx = \int_1^2 x^3 dx = \frac{1}{4}x^4\Big|_1^2 = \frac{15}{4} = 3{,}75.$$

Note que $|I(u) - I(v_{10})| = 0{,}00375$.

7. Ecuaciones diferenciales

Sean $T > 0$, f una función real definida en $[0,T] \times \mathbb{R}$. Suponemos que f es continua, más aún, se supone que f satisface la condición de Lipschitz que se indica a continuación

$$\exists M > 0 \text{ tal que } |f(t,y_1) - f(t,y_2)| \leq M|y_1 - y_2| \quad \forall y_1, y_2 \in \mathbb{R} \text{ y } t \in [0,T].$$

Se considera el problema de Cauchy de valor inicial siguiente:

$$\text{hallar } u \in C^1([0,T]) \text{ solución de } \begin{cases} u'(t) = f(t,u(t)) & t \in]0,T[, \\ u(0) = u_0. \end{cases}$$

Por la hipótesis impuesta sobre f, se sabe que dicho problema tiene solución única. En la generalidad de los casos, la función u no puede determinarse explícitamente, esta viene representada como una integral de una función que no puede integrarse con funciones elementales lo que dificulta el cálculo numérico de $u(t) \quad t \in [0,T]$. Frente a estos dos hechos, la idea es aproximar la solución de la ecuación diferencial en forma numérica.

Sean $m \in \mathbb{Z}^+$, $\tau(m) = \{t_0 = 0, t_1, \ldots, t_m = T\}$ una partición de $[0,T]$ donde $t_{j-1} < t_j \quad j = 1, \ldots, m$. Ponemos $h_j = t_{j-1} - t_j \quad j = 1, \ldots, m$ y $\widehat{h} = \max\limits_{j=1,\ldots,m} h_j$. Si se elige una partición uniforme, se tiene $t_j = jh \quad j = 0, 1, \ldots, m$ con $h = \dfrac{T}{m}$.

De la definición de $u'(t) = \lim\limits_{h \to 0} \dfrac{u(t+h) - u(t)}{h}$ se sigue que para h suficientemente pequeño y no nulo,

$$u'(t) \simeq \frac{u(t+h) - u(t)}{h} \quad t \in]0,T[,$$

y como $u'(t) = f(t,u(t))$ entonces

$$\frac{u(t+h) - u(t)}{h} \simeq f(t,u(t)) \quad t \in]0,T[,$$

luego

$$u(t+h) \simeq u(t) + hf(t,u(t)),$$

y en $t = t_j$, se obtiene

$$u(t_{j+1}) \simeq u(t_j) + hf(t_j, u(t_j)) \quad j = 0, 1, \ldots, m-1.$$

Denotamos con u_j un aproximación de $u(t_j) \quad j = 0, 1, \ldots, m$. Consideramos una partición uniforme de $[0,T]$. Se define

$$\begin{cases} u_0 \text{ dado}, \\ u_{j+1} = u_j + hf(t_j, u_j) & j = 0, 1, \ldots, m. \end{cases}$$

que se conoce como esquema numérico de Euler explícito, lo que a su vez da lugar al siguiente algoritmo.

Algoritmo

Datos de entrada: m, función $f(t,u(t))$, u_0, T.

Datos de salida: $t_j,\ u_j,\ j=0,1,\ldots,m$.

1. Poner $h=\dfrac{T}{m}$.
2. $\tau(m)=\{t_j=jh \mid j=0,1,\ldots,m\}$.
3. Para $j=0,1,\ldots,m-1$

   $u_{j+1}=u_j+hf(t_j,u_j)$

   Fin de bucle j.
4. Imprimir resultados: $t_j,\ u_j \quad j=0,1,\ldots,m$.
5. Fin.

Apliquemos el método de Euler explícito al siguiente ejemplo: $\begin{cases} u'(t)=u(t)+t & t\in\,]0,0{,}5[\,, \\ u(0)=0. \end{cases}$

Tenemos $f(t,u(t))=u(t)+t$. En este caso la solución $u(t)$ se determina mediante el conocido método de separación de variables, se obtiene $u(t)=-(t+1)+e^t \quad t\in[0,0{,}5]$.

Sean $m=5,\ h=\dfrac{0{,}5}{5}=0{,}1$ y $\tau(5)=\{0,0{,}1,\ldots,0{,}5\}$ una partición de $[0,0{,}5]$, $u_0=0$. Los resultados del método de Euler explícito se muestran a continuación.

$$\begin{aligned} u_1 &= u_0+h(u_0+t_0)=0+0{,}1(0+0)=0, \\ u_2 &= u_1+h(u_1+t_1)=0+0{,}1(0+0{,}1)=0{,}01, \\ u_3 &= u_2+h(u_2+t_2)=0{,}01+0{,}1(0{,}01+0{,}2)=0{,}031, \\ u_4 &= u_3+h(u_3+t_3)=0{,}03+0{,}1(0{,}031+0{,}3)=0{,}0641, \\ u_5 &= u_4+h(u_4+t_4)=0{,}0641+0{,}1(0{,}0641+0{,}4)=0{,}11051. \end{aligned}$$

En la figura siguiente se muestra la gráfica de la solución $u(t)$ con línea continua y la aproximada se muestra con $*$ sobre la partición de $[0,0{,}5]$.

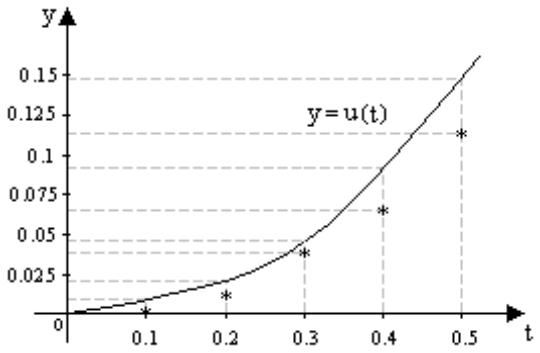


Figura 8

En la tabla siguiente se muestran los valores exactos de u, los calculados u_j sobre la partición $\tau(5)$ así como $|u(t_j)-u_j|$.

j	t_j	$u(t_j)$	u_j	$\lvert u(t_j)-u_j\rvert$
0	0	0	0	0
1	0,1	0,005170918	0	0,005170918
2	0,2	0,021402758	0,01	0,011402758
3	0,3	0,049858808	0,031	0,018858808
4	0,4	0,091824698	0,0641	0,027724698
5	0,5	0,148721271	0,11051	0,038211271

.

Note como se incrementa el error. Esta clase de problemas se abordarán en el último capítulo, donde se hará un análisis de la convergencia de los métodos y en particular de este método de Euler explícito.

1.4. Sistemas de Numeración.

Entre los sistemas de numeración más usados tenemos el sistema decimal o de base 10 cuyas cifras decimales son los enteros comprendidos entre 0 y 9. El número 10 es llamado base de dicho sistema.

Sea $M \in \mathbb{Z}^+$, para indicar su representación decimal escribimos $M = m_n m_{n-1} \dots m_o$, donde $m_i \in \{0, 1, 2, \dots, 9\} \quad \forall i = 0, 1, \dots, n$. A la representación decimal de M le asociamos el polinomio $P(x) = \sum_{k=0}^{n} m_k x^k \quad x \in \mathbb{R}$, donde los coeficientes $m_k \quad k = 0, \cdots, n$ son las cifras decimales del número entero positivo M. Entonces $M = P(10) = \sum_{k=0}^{n} m_k \times 10^k$. Así por ejemplo si $M = 2165$, el polinomio asociado a M está definido como $P(x) = 5 + 6x + x^2 + 2x^3 \quad x \in \mathbb{R}$. En $x = 10$ se tiene

$$P(10) = 5 \times 10^0 + 6 \times 10^1 + 1 \times 10^2 + 2 \times 10^3 = 2165 = M.$$

Otro de los sistemas de numeración más utilizados es el binario o de base 2, cuyas cifras binarias son los dígitos 0 y 1. En este sistema, un número entero positivo A lo representaremos como $A = (a_n a_{n-1} \dots a_0)_2$, donde $a_i \in \{0, 1\} \quad i = 0, \dots, n$. ¿Cuál es la representación de este número A en el sistema decimal? A continuación abordamos el problema de la conversión entre estos dos sistemas de numeración.

Es claro que el número entero 0 se representa como 0 en el sistema decimal y como $(0)_2$ en el sistema binario.

1.4.1. Conversión de binario a decimal y viceversa.

Conversión de binario a decimal.

Sea $A = (a_n a_{n-1} \dots a_0)_2$ un número binario. Para convertir el número A al sistema decimal le asociamos el polinomio $P(x) = \sum_{k=0}^{n} a_k x^k$, donde $a_k \in \{0, 1\} \quad k = 0, \dots, n$ son las cifras binarias del número A, y evaluamos $P(2)$ usando el esquema de Hörner. Así $A = P(2)$ en el sistema de numeración decimal.

Ejemplo

Sea $M = (1101101)_2$. Determinemos M en base 10. Para el efecto, asociamos a M el polinomio $P(x) = 1 + x^2 + x^3 + x^5 + x^6 \quad x \in \mathbb{R}$. Utilizando el esquema de Hörner, se tiene que $P(2) = 109$, por lo tanto $(1101101)_2 = 109$.

Conversión de decimal a binario.

Sea $M \in \mathbb{Z}^+$ en base 10. Supongamos que M tiene la siguiente representación en binario $M = (a_n a_{n-1} \dots a_2 a_1 a_0)_2$ cuyo polinomio asociado es $P(x) = \sum_{k=0}^{n} a_n x^k$ y evaluado en $x = 2$ se expresa como sigue:

$$P(2) = \sum_{k=0}^{n} a_n 2^k = a_0 + a_1 \times 2 + a_2 \times 2^2 + \cdots + a_n \times 2^n.$$

Sea $u \in \mathbb{Z}$. Recordemos que un número entero u se dice par si y solo si existe $j \in \mathbb{Z}$ tal que $u = 2j$, y u se dice impar si y solo si existe $j \in \mathbb{Z}$ tal que $u = 2j + 1$.

Para determinar las cifras binarias $a_0, a_1, \dots, a_n$, procedemos como sigue: $a_1 \times 2 + a_2 + 2^2 + \dots a_n \times 2^n$ es par, entonces

$$M \text{ impar} \Leftrightarrow a_0 = 1, \quad \text{y}, \quad M \text{ par} \Leftrightarrow a_0 = 0.$$

Determinada la cifra a_0, pasamos a determinar la cifra a_1. Definimos $M_1 = \frac{M-a_0}{2} = a_1 + a_2 \times 2 + \ldots + a_n \times 2^{n-1}$. Luego,

$$M_1 \text{ impar} \Leftrightarrow a_1 = 1, \quad \text{y}, \quad M_1 \text{ par} \Leftrightarrow a_1 = 0.$$

De manera análoga a la precedente, definimos $M_2 = \frac{M_1-a_1}{2} = a_2 + a_3 \times 2 + \ldots + a_n \times 2^{n-2}$, entonces

$$M_2 \text{ impar} \Leftrightarrow a_2 = 1, \quad \text{y}, \quad M_2 \text{ par} \Leftrightarrow a_2 = 0.$$

Continuando con este proceso n veces obtenemos las cifras binarias $a_k,\ k = 0, 1, \ldots, n$.

Para determinar n, observamos que si para todo k, $a_n = 1$, entonces $\sum_{k=0}^{n} 2^k = 2^{n+1} - 1 < 2^{n+1}$, y si para $k = 0, 1, \ldots, n-1, \quad a_k = 0$ y $a_n = 1$, entonces $P(2) = 2^n$. Por lo tanto $n \in \mathbb{N}$ debe verificar la desigualdad

$$2^n \leq M < 2^{n+1}.$$

Dado un número entero positivo en el sistema numérico decimal, el procedimiento descrito precedentemente permite obtener las cifras binarias de dicho número. El algoritmo de conversión de decimal a binario es el siguiente:

Algoritmo

Dato de entrada: M.

Dato de salida: $(a_n a_{n-1} \ldots a_0)_2$.

1. Determinar n tal que $2^n \leq M < 2^{n+1}$.
2. $M_0 = M$.
3. $i = 0, 1, \ldots, n-1$

   $$a_i = \begin{cases} 0, & \text{si } M_i \text{ es par,} \\ 1, & \text{si } M_i \text{ es impar.} \end{cases}$$

   $$M_{i+1} = \frac{M_i - a_i}{2}$$

   Fin de bucle i.
4. $a_n = M_n$.
5. Imprimir número binario $(a_n a_{n-1} \ldots a_0)_2$.
6. Fin.

Ejemplos

1. Sea $M = 2412$. Apliquemos el algoritmo precedente. Determinemos el número de cifras requeridas en la representación binaria. Tenemos la desigualdad $2^{11} = 2048 < M < 2^{12} = 4096$, se sigue que $n = 11$. En la siguiente tabla se ilustran los resultados de la aplicación del algoritmo de conversión de decimal a binario del número 2412.

i	0	1	2	3	4	5	6	7	8	9	10	11
M_i	2412	1206	603	301	150	75	37	18	9	4	2	1
a_i	0	0	1	1	0	1	1	0	1	0	0	1

   Por lo tanto $2412 = (100101101100)_2$.
2. Sea $M = 729$. Se tiene que $n = 9$ y la aplicación del algoritmo nos da $729 = (1011011001)_2$.

Caso fraccionario.

Consideramos ahora el caso de la conversión de decimal a binario para números racionales.

Definición 2

i. La serie $\sum_{k=1}^{\infty} a_k 2^{-k}$ *se llama fracción binaria, donde* $a_k \in \{0,1\} \quad \forall k \in \mathbb{Z}^+$.

ii. La fracción binaria se dice finita si $\exists n_0 \in \mathbb{Z}^+$ *tal que* $\forall n \geq n_0,\ a_n = 0$. *De otro modo, la fracción binaria se dice infinita.*

Observación**:** la serie $\sum_{k=1}^{\infty} a_k 2^{-k}$ es convergente, pués $a_k 2^{-k} \leq 2^{-k}, \quad \forall k \in \mathbb{Z}^+$, y

$$\sum_{k=1}^{\infty} a_k 2^{-k} \leq \sum_{k=1}^{\infty} a_k 2^{-k} \leq \sum_{k=1}^{\infty} 2^{-k} = 1.$$

Sea $S = \sum_{k=1}^{\infty} a_k 2^{-k}$ una fracción binaria. En el sistema binario escribimos $S = (0.a_1 a_2 a_3 \ldots)_2$.

Ejemplo

El número $(0{,}00111\ldots)_2$ es una fracción binaria infinita y representa a 0,25, mientras que el número $(0{,}01)_2$ es una fracción binaria finita y también representa a 0,25. Este último es consecuencia del redondeo del primero, que se tratará más adelante.

Dada una fracción binaria finita $(0.a_1 a_2 a_3 \ldots a_n)_2$, asociamos a la misma el polinomio $P(x) = \sum_{k=1}^{n} a_k x^k$ $x \in \mathbb{R}$. Para determinar el valor del número binario en el sistema decimal, calculamos el valor de $P(0{,}5)$ usando el esquema de Hörner. Por tanto $(0.a_1 a_2 \ldots a_n)_2 = P(0{,}5)$ en el sistema decimal.

Ejemplos

1. Si $(0{,}01101)_2$, entonces $P(x) = x^2(1 + x(1 + x^2))$, de donde

$$P(0{,}5) = (0{,}5)^2(1 + 0{,}5(1 + (0{,}5)^2)) = 0{,}40625.$$

2. Sea $b = (0{,}101101)_2$. El polinomio asociado al número b es $P(x) = x + x^3 + x^4 + x^6 \quad x \in \mathbb{R}$. Entonces $b = P(0{,}5) = 0{,}703125$.

Veamos el problema recíproco, es decir la conversión de una fracción decimal a binario.

Primeramente, se debe tener presente que, en general, un número real no admite una representación binaria finita. Por ejemplo, un número real que admite una representación decimal infinita seguramente su representación binaria no es finita. Además, si un número real tiene representación decimal finita no siempre admite representación binaria finita, así los números 0,1 y 0,01 no admiten representación binaria finita pero si periódica.

Sea $b \in \mathbb{R}$ tal que $0 < b < 1$. Fijado $n \in \mathbb{Z}^+$, determinemos las n primeras cifras binarias de b, esto es, determinamos la fracción binaria finita $(0.a_1 a_2 \ldots a_n)_2$ que lo notamos $b_1 = (0.a_1 a_2 \ldots a_n)_2$. Tenemos

$$\begin{aligned} b_1 &= a_1 \times 2^{-1} + a_2 \times 2^{-2} + a_3 \times 2^{-3} + \ldots + a_n \times 2^{-n} \iff \\ 2b_1 &= a_1 + a_2 \times 2^{-1} + a_3 \times 2^{-2} + \ldots + a_n \times 2^{-n+1}, \end{aligned}$$

Entonces $a_2 \times 2^{-1} + a_3 \times 2^{-2} + \ldots + a_n \times 2^{-n+1} < 1$, luego

$$a_1 = 0 \Leftrightarrow 2b < 1, \quad \text{y,} \quad a_1 = 1 \Leftrightarrow 2b \geq 1.$$

Determinada la cifra binaria a_1 pasamos a determinar la cifra binaria a_2. Se define

$$b_2 = 2(2b_1 - a_1) = a_2 + a_3 \times 2^{-2} + \ldots + a_n \times 2^{-n+2}.$$

Razonando como en la parte previa, tenemos

$$a_2 = 0 \Leftrightarrow b_2 < 1, \quad \text{y}, \quad a_2 = 1 \Leftrightarrow b_2 \geq 1.$$

En la k-ésima etapa, tenemos

$$b_k = a_k + a_{k+1}2^{-1} + \ldots + a_n 2^{-n+k},$$

de donde

$$a_k = 0 \Leftrightarrow b_k < 1, \quad \text{y}, \quad a_k = 1 \Leftrightarrow b_k \geq 1.$$

Tenemos el siguiente algoritmo de conversión de decimal a binario.

Algoritmo

Datos de entrada: b, n.

Datos de salida: $(0.a_1a_2 \ldots a_n)_2$.

1. $b_1 = b$.
2. $k = 1, \ldots, n-1$

   $$a_k = \begin{cases} 1, & \text{si } 2b_k \geq 1, \\ 0, & \text{si } 2b_k < 1. \end{cases}$$

   $$b_{k+1} = 2b_k - a_k$$

3. $a_n = \begin{cases} 1, & \text{si } 2b_n \geq 1, \\ 0, & \text{si } 2b_n < 1. \end{cases}$
4. Imprimir la fracción binaria $(0.a_1a_2 \ldots a_n)_2$.
5. Fin.

Ejemplos

1. Sean $b = \dfrac{1}{3}$. Determinemos las primeros cinco cifras binarias de b. Tenemos $n = 5$. En la tabla siguiente se ilustran los resultados de la aplicación del algoritmo precedente.Tenemos,

i	1	2	3	4	5
b_i	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$
$2b_i$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{2}{3}$
a_i	0	1	0	1	0.

   Las cinco primeras cifras binarias de b son: $(0.\ 0\ 1\ 0\ 1\ 0)_2$. La fracción binaria finita $(0.\ 0\ 1\ 0\ 1\ 0)_2$ es una aproximación de b.

2. En la tabla siguiente se muestran los resultados de la aplicación del algoritmo de conversión de decimal a binario para $b = 0{,}1$ con 10 cifras binarias:

i	1	2	3	4	5	6	7	8	9	10
b_i	0,1	0,2	0,4	0,8	0,6	0,2	0,4	0,8	0,6	0,2
$2b_i$	0,2	0,4	0,8	1,6	1,2	0,4	0,8	1,6	1,2	0,4
a_i	0	0	0	1	1	0	0	1	1	0.

   Obtenemos $\tilde{b} = (0{,}0001100110)_2$ aproximación de b con 10 cifras binarias.

Observación. Supongamos que $b = \sum_{k=1}^{\infty} a_k 2^{-k}$, y $\widetilde{b} = (0.a_1 \dots a_n)_2$. Sea $0 < Tol < 1$ suficientemente pequeño. Determinemos n tal que $\left|b - \widetilde{b}\right| < Tol$. Sea $x_n = \sum_{k=1}^{n} a_k 2^{-k}$. Entonces

$$b - x_n = \sum_{k=n+1}^{\infty} a_k 2^{-k} \leq 2^{-n} < Tol.$$

Basta elegir n como el más pequeño número entero positivo tal que $2^{-n} < Tol$. Para tal n resulta que la fracción binaria finita $\widetilde{b} = (0.a_1 \dots a_n)_2$ en decimal es $x_n = \sum_{k=1}^{n} a_k 2^{-k}$. A esta fracción binaria finita la denominaremos aproximación de b con una precisión Tol.

Ejemplo

Sean $b = (0{,}001111\dots)_2$, $\widetilde{b} = (0{,}01)_2$, $Tol = 10^{-5}$. Entonces $\left|b - \widetilde{b}\right| = \dfrac{1}{2^n} < Tol = 10^{-5}$. Se tiene $n = 19$ y consecuentemente

$$\widetilde{b} = \sum_{k=1}^{19} a_k 2^{-k} = \frac{1}{4} - \frac{1}{2^{19}} = 0{,}25.$$

Note que $b = (0{,}001111\dots)_2 = \sum_{k=3}^{\infty} 2^{-k} = \dfrac{1}{4} = 0{,}25$, y $\widetilde{b} = (0{,}00111\dots 1)_2 \simeq (0{,}01)_2 = 0{,}25$.

El algoritmo de conversión de decimal a binario para el caso entero puede ser utilizado para determinar las n primeras cifras binarias de un número real $b \in]0, 1[$. En efecto, sea $b \in \mathbb{R}^+$ con $0 < b < 1$. Supongamos que $b = \sum_{k=1}^{\infty} a_k 2^{-k}$ una fracción binaria. Para $n \in \mathbb{Z}^+$ buscamos una aproximación binaria finita de la forma $\widetilde{b} = (0.a_1 a_2 \dots a_n)_2$. Entonces

$$\widetilde{b} = P\left(\frac{1}{2}\right) = \sum_{k=1}^{n} a_k 2^{-k} = a_1 \times 2^{-1} + \dots + a_n \times 2^{-n},$$

de donde

$$2^n \widetilde{b} = a_n + a_{n-1} \times 2 + \dots a_1 \times 2^{n-1}.$$

Puesto que $2^n b$, en general, no es un entero y como

$$\begin{aligned} 2^n b &= a_1 \times 2^{n-1} + \dots + a_{n-1} \times 2 + a_n + a_{n+1} \times 2^{-1} + a_{n+2} \times 2^{-2} + \dots \\ &= 2^n \times \tilde{b} + a_{n+1} \times 2^{-1} + a_{n+2} \times 2^{-2} + \dots, \end{aligned}$$

entonces $M = [2^n b] = 2^n \widetilde{b}$, donde $[\cdot]$ denota la función mayor entero menor o igual que x y para números reales positivos coincide con la parte entera de dicho número. Resulta que $M = (a_1 a_2 \dots a_n)_2$ cuyas cifras binarias pueden ser determinadas por aplicación del algoritmo de conversión de decimal a binario para $M \in \mathbb{Z}^+$ ya descrito anteriormente.

1.4.2. Conversión de decimal a cualquier base y viceversa

Sea $N \in \mathbb{Z}^+$ con $N > 1$. En el sistema de numeración de base N, los dígitos de dicho sistema son $0, 1, \dots, N-1$ si $N \leq 10$, y si $N > 10$ los dígitos de dicho sistema son $0, 1, \dots, 9$ y para las $N - 10$ cifras sucesivas se utilizan otros símbolos, por ejemplo las letras $A, B, C, \dots$ en ese orden.

Sistema	Base	Dígitos
Binario	2	$0, 1$
Octal	8	$0, 1, \dots, 7$
Decimal	10	$0, 1, \dots, 9$
Hexadecimal	16	$0, 1, \dots, 9, A, B, \dots, F.$

Sea $M \in \mathbb{Z}^+$, al número M lo representamos en base N de la manera siguiente: $M = (m_n m_{n-1} \dots m_0)_N$, donde $m_k \in \begin{cases} \{0,1,\dots,N-1\}, & \text{si } N \leq 10, \\ \{0,1,\dots 9,A,B,\dots\}, & \text{si } N > 10. \end{cases}$ A este número entero positivo M representado en el sistema de numeración de base N le asociamos el polinomio real $P(x) = \sum_{k=0}^{n} m_k\, x^k \quad x \in \mathbb{R}$. Entonces M en el sistema de numeración decimal se determina mediante la evaluación del polinomio $P(x)$ en $x = N$, esto es, $M = P(N)$ en base 10.

Ejemplos

1. Sea $M = (7347)_8$. El polinomio asociado a M se escribe como:

$$P(x) = 7 + 4x + 3x^2 + 7x^3 = 7 + x(4 + x(3 + 7x)) \quad x \in \mathbb{R}.$$

   Luego $P(8) = 7 + 8(4 + 8(3 + 7 \times 8)) = 3815$ en base 10. Así $(7347)_8 = 3815$.

2. Sea $M = (3AF)_{16}$. El polinomio asociado es $P(x) = F + Ax + 3x^2$ cuyo valor en $x = 16$ es

$$P(16) = F + A \times 16 + 3 \times 16^2 = 943.$$

   Tenemos $(3AF)_{16} = 943$.

Para elaborar un algoritmo de conversión de base 10 a base $\mathbb{N}$, debemos primeramente revisar el algoritmo de la división de Euclides y las clases residuales.

El algoritmo de la división de Euclides se establece en los siguientes términos: dados $a, b \in \mathbb{Z}^+$, existen $c, r \in \mathbb{N}$ tales que $0 \leq r < b$, y $a = bc + r$. El número natural c se llama cociente y r se llama residuo. Por ejemplo, si $a = 23$, $b = 5$, entonces $23 = 4 \times 5 + 3$, donde $c = 4$ y $r = 3$.

Por otro lado, la congruencia módulo m se define como a continuación se indica: dados $m \in \mathbb{Z}^+$, $a, b \in \mathbb{Z}$, se dice que a y b son congruentes módulo m que se escribe $a \equiv b \bmod(m)$ si y solo si existe $c \in \mathbb{Z}$ tal que $a - b = cm$. Es inmediato verificar que la relación de congruencia módulo m es una relación de equivalencia y que dicha relación define una partición de $\mathbb{Z}$ en clases residuales notadas como $[0]$, $[1], \cdots, [m-1]$ tales que

$$\begin{aligned} [j] &= \{cm + j \,|\, c \in \mathbb{Z}\}, \\ [i] \cap [j] &= \emptyset \text{ para } i \neq j,\ i, j = 0, \dots, m-1, \\ \mathbb{Z} &= \bigcup_{j=0}^{m-1} [j]\,. \end{aligned}$$

El procedimiento descrito para obtener las cifras binarias de un número en base 10 equivale a aplicar el algoritmo de la división de Euclides: $a = 2c + r$, donde $c \in \mathbb{N}$ y $r \in \{0, 1\}$.

Este procedimiento puede extenderse de modo similar a otras bases. En efecto, sea $M \in \mathbb{Z}^+$ expresado en el sistema decimal, $N \in \mathbb{Z}^+$ con $N > 1$ la nueva base. Por el algoritmo de la división de Euclides, se tiene que $M = cN + r$, donde $r \in \mathbb{N}$ tal que $0 \leq r < N$. Las clases residuales módulo N son: $[0], [1], \dots, [N-1]$.

El algoritmo de conversión de base 10 a base N se describe a continuación.

Algoritmo

Datos de entrada: M, N

Dato de salida: $(a_n a_{n-1} \dots a_0)_N$.

1. Determinar n tal que $N^n \leq M < N^{n+1}$.

2. $M_0 = M$.

3. $i = 0, 1, \dots, n-1$

$a_i = r \quad \text{si } M_i \in [r],$

$M_{i+1} = \dfrac{M_i - a_i}{N}$

Fin de bucle i.

4. $a_n = M_n$.

5. Imprimir $M = (a_n a_{n-1} \dots a_0)_N$.

6. Fin.

Ejemplo

1. Sean $M = 35$, $N = 3$. Se verifica inmediatamente que $3^3 \leq 35 < 3^4$, luego $n = 3$ y consecuentemente M tiene cuatro cifras en base 3. Ponemos $M = (a_3 a_2 a_1 a_0)_3$. Debemos determinar las cifras $a_0, a_1, a_2, a_3 \in \{0, 1, 2, \}$. En la siguiente tabla se muestran los resultados de la aplicación del algoritmo de conversión de decimal a base 3.

$$\begin{aligned} M_0 &= M = 35 \in [2], \ a_0 = 2, \qquad M_1 = \frac{M_0 - a_0}{3} = 11 \in [2], \quad a_1 = 2, \\ M_2 &= \frac{M_1 - a_1}{3} = 3 \in [0], \quad a_2 = 0, \qquad M_3 = \frac{M_2 - a_2}{3} = 1 = a_3. \end{aligned}$$

Luego $35 = (1022)_3$.

Relación del número de cifras entre dos sistemas de numeración

La relación del número de cifras para la representación de un número en dos bases distintas lo podemos determinar de la siguiente manera. Sean $a, b \in \mathbb{R}^+$ con $a \neq 1$, $b \neq 1$ dos bases distintas y $M \in \mathbb{Z}^+$ expresado en base 10. Supongamos que M se representa con respecto de estas dos bases como $M = (a_n a_{n-1}, \dots a_0)_a$, y $M = (b_m b_{m-1} \dots b_0)_b$. Entonces $m, n \in \mathbb{N}$ satisfacen las siguientes desigualdades:

$$\begin{aligned} a^n &\leq M < a^{n+1}, \\ b^m &\leq M < b^{m+1}, \end{aligned}$$

y en consecuencia

$$n \leq log_a(M) < n + 1,$$

$$m \leq log_b\,(M) < m + 1.$$

De la relación $\log_b(M) = \log_b(a) \times \log_a(M))$, se sigue que $m \leq \log_b(a) \times \log_a(M) < m + 1$, de donde

$$\frac{m}{\log_a\,(M)} \leq log_b(a) < \frac{m+1}{\log_a\,(M)}.$$

Tomando en consideración que $n < \log_a(M) < n + 1$, se obtiene la siguiente desigualdad

$$\frac{m}{n+1} < \frac{m}{\log_a\,(M)} \leq log_b(a) < \frac{m+1}{\log_a\,(M)} \leq \frac{m+1}{n}.$$

Para M suficientemente grande de modo que $\dfrac{1}{n}$ sea despreciable, se deduce que

$$log_b(a) \simeq \frac{m}{n} \Longleftrightarrow n log_b(a) \simeq m,$$

es decir que M en el sistema de numeración de base a requiere de aproximadamente $n log_b(a)$ cifras en el sistema de numeración de base b.

Para $b = 2$ y $a = 10$ se tiene que $\dfrac{m}{n} \simeq 3$, de donde $m \simeq 3n$. La representación binaria requiere aproximadamente cerca de 3 veces del número de cifras necesarias en la representación decimal. Para $b = 2$ y $a = 16$, se tiene $\dfrac{m}{n} = 4$.

1.5. Representación en punto flotante.

La representación en punto flotante normalizada de un número real no nulo x en el sistema decimal (comúnmente conocida como notación científica) se expresa como $x = \pm a \times 10^p$, donde $a \in [\frac{1}{10}, 1[$, $p \in \mathbb{Z}$. El número a se denomina mantisa de x y el exponente p se denomina característica de x. Este tipo de escritura de los números reales se utiliza en algunos instrumentos de cálculo como por ejemplo las calculadoras de bolsillo, los computadores portátiles. Más precisamente, un número de máquina $d \neq 0$ en una calculadora o en un computador es un número real que tiene su representación en punto flotante normalizada de la forma:

$$d = sign(d) \times a \times 10^p,$$

donde $sign(d)$ denota el signo de d, $a = 0.d_1 \cdots d_m \quad \text{con} \quad d_i \in \{0, 1, \cdots, 9\}, \quad i = 1, \cdots, m, \quad d_1 \neq 0$; y el exponente p, por ejemplo para ciertos tipos de calculadoras de bosillo, pertenece al conjunto $\{-100, -99, \cdots, 98, 99\}$.

Si $b = 0$, entonces $d_i = 0, \quad i = 1, \cdots, m$. La condición $d_1 \neq 0$ asegura que $a \geq 10^{-1}$. Además, para una aplicación numérica, el número de cifras decimales m es, en general, fijo.

Ejemplos

1. El número $x = \dfrac{1685}{3} = 561{,}6666\ldots$ se escribe en punto flotante normalizado $0{,}5616666\ldots \times 10^3$ y como número de máquina en punto flotante (por ejemplo para una calculadora de bolsillo) $0{,}5616666667 \times 10^3$.

2. El número $x = -0{,}0000000001$ como número en punto flotante normalizado (y como número de máquina) se escribe como $-0{,}1 \times 10^{-10}$.

Fijado el número de cifras decimales m, debe notarse que la mantisa más pequeña es $a = 0{,}1$ y no $a = 0{,}0 \cdots 01$ y la mantisa más grande es $a = 0{,}9 \cdots 9$.

En el sistema binario, la representación de números reales es análoga. Un número real $x \neq 0$ tiene una representación binaria normalizada que se escribe como sigue:

$$x = sign(x) \times a \times 2^p,$$

donde $a \in [\frac{1}{2}, 1[$ y $p \in \mathbf{Z}$. El número a se expresa como una serie binaria $\sum_{i=1}^{\infty} a_i 2^{-i}$ y $|p|$ se escribe como un entero en binario, esto es:

$$|p| = (p_n p_{n-1} \ldots p_0)_2 = \sum_{i=0}^{n} p_i \, 2^i.$$

Ejemplos

1. $x = \dfrac{1685}{3} = \left(\dfrac{1685}{3} \times 2^{-10}\right) 2^{10} = \dfrac{1685}{3072} \times 2^{10}.$

2. $-0{,}001 = -\dfrac{2^9}{1000} 2^{-9} = -0{,}512 \times 2^{-9}.$

Una cifra binaria se denomina bit. Si el número de bits asignados para almacenar el número es fijo, un número de máquina tiene una forma de punto flotante binaria normalizada $b = sign(b) \times a \times 2^p$, que puede almacenar exactamente usando los siguientes grupos de bits: $\begin{cases} \text{un bit para el signo de } b, \\ r \text{ bits para el exponente } |p|, \\ m \text{ bits para la mantisa } a. \end{cases}$

Es decir que un número de máquina b en punto flotante binario normalizado requiere para su representación de $m+r+1$ bits, a condición de que dicho número de bits sea fijo. Además,

$$a = (0.\, b_1 \dots b_m)_2, \qquad b_i \in \{0,1\}, \quad i = 1, \dots, m, \quad b_1 \neq 0,$$

$$|p| = (p_{r-1} \dots p_0)_2, \qquad p_j \in \{0,1\}, \quad j = 0, 1, \dots, r-1.$$

La mantisa más pequeña es $a = (0{,}10\dots0)_2 = \frac{1}{2}$ y la más grande es $a = (0{,}11\dots1)_2 = 1 - 2^{-m}$. Para el exponente se tiene que

$$0 \leq |p| \leq 2^r - 1.$$

En consecuencia, el número más grande a representarse en punto flotante binario normalizado es

$$(1 - 2^{-m}) \times 2^{2^r - 1}$$

y el número positivo más pequeño es 2^{-2^r}. Este último se obtiene del modo siguiente:

$$b = (0.b_1 \dots b_m)_2 \times 2^{-q} \geq (0{,}10\dots0) = \frac{1}{2} \times 2^{-q} = 2^{-q-1},$$

donde $q \geq 0$. Como $0 \leq q \leq 2^r - 1$, entonces $-q \geq -2^r + 1$, y en consecuencia $b \geq 2^{-q-1} = 2^{-2^r}$.

Así, un número de máquina en punto flotante binario normalizado satisface la desigualdad:

$$2^{-2^r} \leq |b| \leq \left(1 - 2^{-m}\right) 2^{2^r - 1}.$$

Además, el conjunto de números de máquina es finito.

Por ejemplo,si se tiene $m = 24$, $r = 7$. Entonces $0 \leq p \leq 127$ y $2^{2^r-1} = 2^{127} \simeq 10^{38}$, de modo que

$$10^{-38} \simeq 2^{-2} \leq |b| \leq 2^{2^7 - 1} \simeq 10^{38}.$$

En la actualidad se tiene los siguientes tipos de representación de números reales: simple precisión, doble precisión y doble precisión extendida. El estudiante debe conocer cuantos bits se requieren para la mantisa y cuantos para el exponente. Por ejemplo para ciertos tipos de máquinas se tiene 1 bit para el signo, para doble precisión se tienen 23 bits para la mantisa y 8 bits para el exponente, para doble precisión extendida se tienen 47 bits para la mantisa y 16 bits para el exponente. Si se dispone de 64 bits para representar un número en doble precisión extendida, estos se dispone de la manera siguiente: 47 bits para la mantisa, 1 bit para el signo y 16 bits para el exponente.

Observación.

1. En una calculadora de bolsillo, el número de bits para representar la mantisa así como su equivalente en base 10 es, generalmente fijo. En un computador se tiene alguna flexibilidad para almacenar mantisas de diferentes tallas. Además, la aritmética que utilizan puede ser binaria (base 2), octal (base 8), hexadecimal (base 16).

2. En las calculadoras se almacena el exponente p con un corrimiento de 100, de tal manera que el exponente corrido $E + 100$ es un entero no negativo entre 0 y 199. De este modo se evita utilizar un bit para el signo del exponente. Así por ejemplo:

   $$x = \frac{1685}{3} = 0{,}5616666667 \times 10^3, \quad p + 100 = 103.$$

   En binario, el exponente corrido de r bits tiene una representación de la forma

   $$0 \leq p + p_0 = (b_{r-1} \dots b_0)_2 = \sum_{i=0}^{r-1} b_i 2^i \leq 2^r - 1.$$

   El corrimiento p_0 se le toma como 2^{r-1}, en cuyo caso

   $$-2^{r-1} \leq p \leq 2^r - 1 - p_0 = 2^r - 1 - 2^{r-1} = 2^{r-1} - 1.$$

Los enteros m (para la mantisa), p_0 y r son fijos.

Por ejemplo, el grupo de bits: $-1\,|\;\;101100110\ldots0\;|\;\;0000110$, donde $m = 24,\ r = 7,$ indica que el número es negativo (primer bit -1), la mantisa $a = (0{,}101100110\ldots0)_2 = 0{,}69922$, $p + p_0 = (0000110)_2 = 6$. Puesto que $p_0 = 2^{r-1} = 2^{7-1} = 64$, entonces $p = 6 - p_0 = -58$, $x = -0{,}69922 \times 2^{-58}$. Note que para $r = 7,\ -64 \leq p \leq 63$.

Para el grupo de bits $0\,|\;\;101100110\ldots0\;|\;\;1000010$ se tiene: mantisa $a = (0{,}101100110\ldots0)_2 = 0{,}69922$, exponente con corrimiento: $p + p_0 = (1000010)_2 = 66,\ \ p_0 = 64$ entonces $p = 66 - p_0 = 2$. Luego $x = 0{,}69922 \times 2^2 = 2{,}57688$.

1.6. Tipos de Errores

En la sección 3 hemos visto algunos ejemplos de cálculo de soluciones numéricas de problemas relativamente sencillos que surgen en los ámbitos del álgebra lineal, el análisis matemático, las ecuaciones diferenciales en los que se evidencian los resultados aproximados obtenidos en instrumentos de cálculo como son las calculadoras de bolsillo y los computadores. Con cualquiera de estos instrumentos, los resultados mostrados están sujetos a errores.

En el análisis numérico, uno de los problemas fundamentales es el estudio o análisis del error cometido en cada uno de los métodos de aproximación que se proponen. Interesa establecer la exactitud y precisión en el cálculo de la solución de un problema (P), la minimización de los errores cometidos en el cálculo de la solución aproximada de (P). Se distinguen varios tipos de errores que limitan la exactitud. Estos pueden clasificarse en tres grupos: errores en los datos de entrada o errores inherentes; errores de redondeo y errores de aproximación.

1. **Errores en los datos de entrada o errores inherentes.** Se deben a esquematizaciones hechas para la reducción de términos matemáticos de cierto modelo. Pueden deberse también a errores debidos en las medidas experimentales de una magnitud física o a observaciones de cualquier otra índole (de tipo económico, social, etc.). Pueden tener también su origen como resultados de un cálculo realizado previamente. Nótese que estos errores aparecen antes de iniciar el cálculo de un cierto problema (P). En el estudio que nosotros haremos no nos ocuparemos de este tipo de errores.

2. **Errores de redondeo.** Estos errores son debidos a la necesidad de trabajar con números de máquina. Dependen casi exclusivamente del instrumento de cálculo a disposición. La evaluación rigurosa es, a menudo, muy complicada. Para el cálculo de la solución de ciertos problemas que consideraremos más adelante, los errores de redondeo tienen una influencia enorme que puede arruinar los resultados. Por tanto es de mucha importancia el poder controlarlos.

   A continuación presentamos tres números reales y sus aproximaciones con 8, 16, 24, 32 cifras luego del punto decimal, las mismas que han sido obtenidas con el programa de Matemática. El símbolo $\simeq$ se utilizará en lo sucesivo para indicar aproximación. Tenemos,

$$\begin{aligned}
\frac{805}{111} &\simeq 7{,}25225225, \qquad \frac{805}{111} \simeq 7{,}252252252522522525,\\
\frac{805}{111} &\simeq 7{,}25225225225225225225225,\\
\frac{805}{111} &\simeq 7{,}25225225225225225225225225225225,
\end{aligned}$$

$$\begin{aligned}
\pi &\simeq 3{,}14159265, \qquad \pi \simeq 3{,}1415926535897932,\\
\pi &\simeq 3{,}14159265358979323846,\\
\pi &\simeq 3{,}14159265358979323846264338327950,
\end{aligned}$$

$$\begin{aligned}\sqrt{2} &\simeq 1{,}41421356, \qquad \sqrt{2} \simeq 1{,}4142135623730950,\\ \sqrt{2} &\simeq 1{,}41421356237309504880168872420970,\\ \sqrt{2} &\simeq 1{,}41421356237309504880168872420970.\end{aligned}$$

Note que el número $\dfrac{805}{111}$ es un número racional, su representación decimal es infinita y periódica. Basta conocer el primer período de su representación decimal y con esta puede escribirse el número con el número de cifras que se desee.

Al representar los números reales como aproximaciones con un número determinado de cifras decimales, se comete un error de redondeo que trataremos más adelante.

3. **Errores de aproximación.** Este tipo de errores se dividen en dos grupos: los errores de truncamiento y los errores de discretización.

a) Errores de truncamiento. Consideremos los siguientes ejemplos.

1. Sean (a_n) una sucesión numérica real y $\sum_{n=0}^{\infty} a_n$ una serie que suponemos convergente. Denotamos con S su suma, esto es, $S = \sum_{n=0}^{\infty} a_n$. Con frecuencia, el cálculo exacto de S es muy difícil de obtener, por lo que se recurre al cálculo aproximado. La idea es aproximar la suma S a través de un número finito m de términos, digamos $S_m = \sum_{n=0}^{m} a_n$. Este procedimiento produce un error denominado de truncamiento. La determinación del número de términos necesarios para la aproximación de la solución S es importante, pués evita la ejecución de cálculos que no mejoran la precisión de la solución, y, disminuyen los costos numéricos.

2. Sean (f_n) una sucesión de funciones definidas en un intervalo $[a,b]$ de $\mathbb{R}$ y $\sum_{n=0}^{\infty} f_n$ una serie de funciones que suponemos converge uniformemente en el intervalo $[a,b]$. Se define la función f como sigue: $f(x) = \sum_{n=0}^{\infty} f_n(x) \qquad x \in [a,b]$. Nos interesamos en trazar la gráfica de la función f. En la generalidad de situaciones resulta complicado, y en ocasiones imposible, el cálculo de cada $f(x)$. Una forma de resolver este problema es aproximar cada $f(x)$ con una suma finita de términos de la serie $\sum_{n=0}^{\infty} f_n(x)$, la misma que se elige apropiadamente en función de la precisión que deseamos obtener. Este proceso de aproximación provoca un error denominado de truncamiento. Por otro lado, dado el volumen de cálculo a realizar es conveniente elaborar un algoritmo numérico para calcular cada $f(x)$. Cuando la serie converge rápidamente este es el camino a seguir. Lastimosamente, en ocasiones, el solo hecho de limitar a un número finito de términos no basta sobre todo en el caso de series que convergen lentamente, esto conduce a proponer otro tipo de problema que consiste en la búsqueda de un método para acelerar la convergencia de la serie, una vez logrado esto, se pasa a calcular los valores aproximados de $f(x)$. En un capítulo posterior se estudian esta clase de problemas.

b) Errores de discretización. Consideremos los siguientes ejemplos.

1) Sea v una función real continua en el intervalo cerrado $[a,b]$. Queremos calcular $v(x)$ con $x \in [a,b]$, lastimosamente la función v no es conocida en todo el intervalo $[a,b]$ sino en un conjunto finito de $m+1$ puntos de una partición $\tau(m) = \{x_0 = a,\ x_1, \cdots,\ x_m = b\}$ de $[a,b]$, donde $x_{i-1} < x_i$ $i = 1, \cdots, m$, digamos $S = \{(x_i, v(x_i)) \in \mathbb{R}^2 \mid i = 0, 1, \cdots, m\}$. Este problema (P) se presenta con mucha frecuencia y se le conoce como problema de interpolación. La idea es aproximar $v(x)$ mediante $v_h(x)$ de una función v_h definida en $[a,b]$ que sea mucho más simple de calcular de modo que $v_h(x_i) = v(x_i) \quad i = 0, 1, \cdots, m$, esto conduce a construir la función v_h como $v_h = \sum_{i=0}^{i=m} v(x_i)\varphi_i$, donde $\{\varphi_0, \cdots, \varphi_m\}$ es un conjunto de funciones que se construyen apropiadamente. La función v_h así definida se conoce con el nombre de función interpolante de v. Este proceso produce un error denominado de discretización. En un capítulo posterior se estudia este tipo de problemas.

En la figura siguiente se muestra la gráfica de una función continua v definida en el intervalo $[0,a]$ con $a > 0$, y la de una función interpolante v_h (segmentos de recta) de v. Se muestran también los

puntos de la partición de $\tau(m)$ de $[0,a]$.

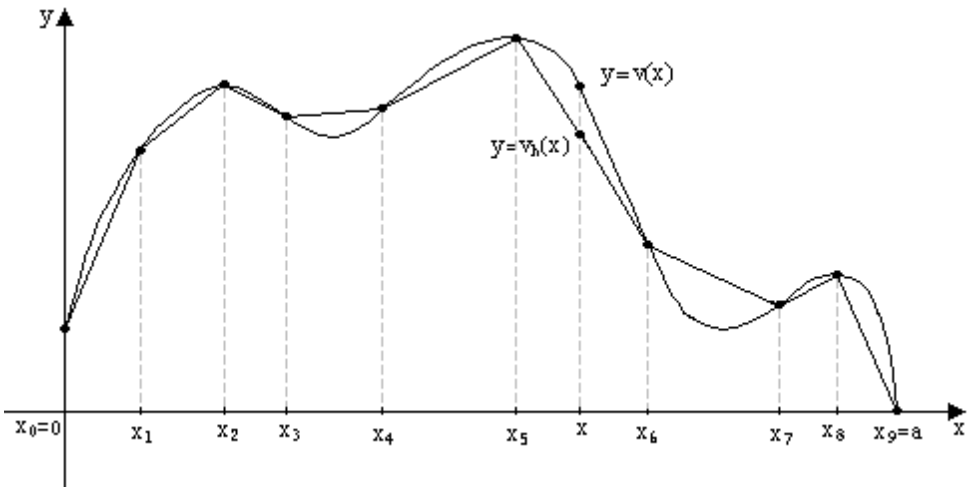


Figura 9

2) Sea $L > 0$. Se considera el siguiente problema:

$$\text{hallar una función } u \in C^2([0,L]) \text{ solución de } \begin{cases} -u''(x) + u(x) = f(x) & \text{si } x \in]0,L[, \\ u(0) = 0, \quad u(L) = 0, \end{cases}$$

donde $f \in C(\ [0,L])$, con $C^k([0,L])$ el espacio de funciones que poseen derivada continua en el intervalo $[0,L]$, $k = 0, 1, \cdots$, y se pone $C([0,L]) = C^0([0,L])$. Este problema es parte de la familia de los denominados problemas unidimensionales de valores en la frontera. Con la hipótesis sobre f, se demuestra que este problema tiene solución única $u \in C^2([0,L])$. Para ciertos tipos de funciones f se puede encontrar soluciones explícitas que no se representan como integrales, para otros tipos de funciones f las soluciones se expresan como integrales de las que no pueden calcularse sus primitivas o que resultan difíciles de calcularse. Por otro lado, se tiene interés en calcular numéricamente la solución $u(x) \quad x \in [0,L]$. Todos estos argumentos nos conducen a resolver el problema de valores en la frontera en forma numérica, es decir, proceder a discretizar dicho problema. Para el efecto, sea $m \in \mathbb{Z}^+$, $\tau(m) = \{x_0 = 0, x_1, \cdots, x_m = L\}$ una partición de $[0,L]$, donde $x_{j-1} < x_j \quad j = 1, \cdots, m$. Ponemos $h_j = x_j - x_{j-1} \quad j = 1, \cdots, m, \quad h = \text{máx}\{h_j \mid i = 1, \cdots, m\}$. En el caso de una partición uniforme, se define $h = \dfrac{L}{m}$ y $x_j = jh \quad j = 0, 1, \cdots, m$. Consideremos una partición uniforme. Se denota con u_j a una aproximación de $u(x_j) \quad j = 0, 1, \cdots, m$. En el capítulo 2 mostraremos que $u''(x)$ se aproxima mediante el cociente denominado diferencias finitas centrales de segundo orden que se indica a continuación:

$$u''(x_j) \simeq \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} \qquad j = 1, \cdots, m-1.$$

Con esta aproximación, el problema propuesto de valores en la frontera se aproxima como

$$\begin{cases} -\dfrac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + u(x_j) \simeq f(x_j) \qquad j = 1, \cdots, m-1, \\ u(x_0) = 0, \qquad u(x_m) = 0, \end{cases}$$

por lo que el problema discreto es el siguiente:

$$\text{hallar } \overrightarrow{u} \;=\; (u_1, \cdots, u_{m-1}) \in \mathbb{R}^{m-1} \quad \text{solución del sistema de ecuaciones lineales}$$
$$\begin{cases} -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + u_j = f(x_j) \qquad j = 1, \cdots, m-1, \\ u_0 = 0, \qquad u_m = 0. \end{cases}$$

Este proceso de discretización del problema de valores en la frontera, produce un error denominado error de discretización.

La estimación de los errores de discretización es fundamental en el Análisis Numérico.

Exacto, inexacto, precisión, imprecisión.

En el lenguaje corriente, los términos exactitud y precisión se usan indistintamente como sinónimos. En el contexto del análisis numérico es importante establecer la diferencia que existe entre estos dos téminos.

El témino exactitud se refiere a que tan cercano está un valor calculado o medido con el verdadero valor; mientras que el término inexacto se define como una desviación del verdadero valor. También se considerará como exacto aquel resultado o método riguroso, conforme a la lógica.

El término precisión se refiere a que tan cercano está un valor individual calculado o medido con cualquier otro; mientras que el término imprecisión se refiere a una magnitud que se aleja una de otra. Se comprenderá como precisión aquello que no deja incertidumbre, determinado rigurosamente.

Por ejemplo, con 9 cifras luego del punto decimal, $\sqrt{3}$ se calcula como 1,732050808. En este caso hablamos de una exactitud de 10^{-9}. Cuando $\sqrt{3}$ se aproxima como 1,7320, 1,7320508, 1,732050808 hablamos de una precisión de 4, 7, 9 cifras luego del punto decimal.

Si $I = \int_0^2 \sqrt{x}dx = \frac{2}{3}(\sqrt{2})^3 \simeq 1{,}88562$, hablamos en este caso del cálculo de I con una exactitud de 10^{-5}, y aplicando la regla del rectángulo siguiente: $I_5 = 0{,}4(\sqrt{0{,}2} + \sqrt{0{,}6} + \sqrt{1{,}0} + \sqrt{1{,}4} + \sqrt{1{,}8})$, resulta $I_5 \simeq 1{,}898667$, hablamos en este caso de un cálculo de I_5 con una precisión de 10^{-6}, y un cálculo aproximado de I con una precisión de 10^{-2}.

Con error de cálculo entenderemos tanto la inexactitud como la imprecisión. Tenemos

$$\text{Verdadero valor} = \text{Valor aproximado} + \text{error},$$

donde el error puede deberse a los errores de redondeo, de aproximación (truncamiento, discretización) o ambos. De manera general, al verdadero valor lo conoceremos como solución exacta y al valor aproximado lo denominaremos solución numérica o también solución aproximada.

De estas observaciones tenemos que los métodos numéricos deben ser suficientemente exactos y precisos.

1.7. Errores de redondeo

Hemos visto que los números de máquina en punto flotante satisfacen la desigualdad:

$$2^{-2^r} \leq |b| \leq \left(1 - \frac{1}{2^m}\right) \times 2^{2^r - 1},$$

donde r, m son enteros positivos. Además, todo número de máquina se escribe en la forma

$$b = (0.a_1 ... a_m)_2 \times 2^{(a_r a_{r-1} ... a_0)_2}.$$

Sea A el conjunto de tales números. Se sigue que el conjunto de números de máquina es finito. El problema que se presenta es el siguiente: ¿cómo aproximar un número $x \notin A$ por un número $y \in A$?

Consideremos lo tres casos siguientes:

1. $x > \left(1 - \frac{1}{2^m}\right) \times 2^{2^r - 1}$.
2. $2^{-2^r} \leq x \leq \left(1 - \frac{1}{2^m}\right) \times 2^{2^r - 1}$.
3. $0 < x < 2^{-2^r}$.

Comenzaremos con el caso 2). Sea $M = \left[2^{-2^r}, \left(1 - \frac{1}{2^m}\right) 2^{2^r - 1}\right] \subset \mathbb{R}$. Se tiene que $A \subset M$.

La métrica usual d en $\mathbb{R}$ está definida como $d(x, y) = |x - y| \quad \forall x, y \in \mathbb{R}$.

Sea $x \in M$, como A es un conjunto cerrado, $\exists \widetilde{x} \in A$ tal que $d(x, A) = d(x, \widetilde{x}) = |x - \widetilde{x}|$, donde la distancia del punto x al conjunto A se define como $d(x, A) = \min\limits_{y \in A} |x - y|$. Resulta que

$$|x - \widetilde{x}| = d(x, A) \leq |x - y| \quad \forall y \in A.$$

Por tanto la aproximación de cualquier número $x \in (M \setminus A)$ por un número notado como $rd(x) \in A$ debe satisfacer la siguiente condición:

$$|x - rd(x)| \leq |x - y| \quad \forall y \in A.$$

El número $rd(x)$ aproximación de x se lo obtiene por redondeo y se denomina redondeado de x.

Ejemplo

Supongamos que nuestro conjunto A está constituido por números reales de la forma $0.a_1a_2a_3a_4 \times 10^p$, donde $a_i \in \{0, 1, \cdots, 9\}$, $p \in \mathbb{Z}$ y $a_1 \neq 0$. Note que la mantisa de los elementos del conjunto A únicamente tienen 4 dígitos, que escribiremos $t = 4$. Entonces $rd(0{,}14285 \times 10^0) = 0{,}1429 \times 10^0$, pués

$$| \, 0{,}14285 \times 10^0 - 0{,}1429 \times 10^0 \, | = 0{,}5000 \times 10^{-4} \leq | \, 0{,}14285 \times 10^0 - y \, | \quad \forall y \in A.$$

De manera similar se obtienen los siguientes resultados

$$\begin{aligned} rd(0{,}8423 \times 10^0) &= 0{,}8423 \times 10^0, \\ rd(3{,}14159 \times 10^0) &= 0{,}3142 \times 10^1, \\ rd(0{,}142842 \times 10^2) &= 0{,}1428 \times 10^2. \end{aligned}$$

En general, para encontrar $rd(x)$ con t dígitos, se procede del modo siguiente: el número $|x| \in (M \setminus A)$ es representado en forma normalizada: $|x| = a \times 10^b$, de modo que $\frac{1}{10} \leq |a| \leq 1$. Sea $a = 0.\alpha_1\alpha_2 \ldots \alpha_t\alpha_{t+1} \ldots$ la representación decimal de a, donde $0 \leq \alpha_i \leq 9 \quad \forall_i = 1, 2, \ldots, \alpha_1 \neq 0$. Definimos

$$\tilde{a} = \begin{cases} 0.\alpha_1\alpha_2 \ldots \alpha_t, & \text{si } 0 \leq \alpha_{t+1} \leq 4, \\ 0, \alpha_1\alpha_2 \ldots \alpha_t + 10^{-t}, & \text{si } 5 \leq \alpha_{t+1} \leq 9. \end{cases}$$

Como $1 \leq \alpha_1 \leq 9$ entonces $|a| \geq 0.\alpha_1 \geq \frac{1}{10} = 0{,}1$. Se pone $sign(x) = \begin{cases} -1, & \text{si } x < 0, \\ 1, & \text{si } x > 0, \end{cases}$ y $rd(x) = sign(x) \times \tilde{a} \times 10^b$.

Definición 3 *El error de redondeo de x se define como $x - rd(x)$. El número no negativo $|x - rd(x)|$ se llama error absoluto.*
El error relativo de x se define mediante la relación:

$$\varepsilon_x = \frac{rd(x) - x}{x} \quad x \neq 0.$$

En algunos textos, el error de redondeo se le denomina error inherente.

Se tiene la siguiente mayoración del error relativo ε_x:

$$\begin{aligned} |\varepsilon_x| &= \left|\frac{rd(x) - x}{x}\right| = \frac{\left|(sign(x)\tilde{a}10^b - sign(x)a{,}10^b\right|}{|sign(x)a \times 10^b|} = \frac{|\tilde{a} - a|}{|a|} \\ &\leq \frac{5 \times 10^{-(t+1)}}{|a|} \leq 5 \times 10^{-(t+1)+1} = 5 \times 10^{-t} = eps. \end{aligned}$$

Luego $|\varepsilon_x| \leq eps = 5 \times 10^{-t}$.

Definición 4 *El número* $eps = 5 \times 10^{-t}$ *se llama precisión de máquina .*

Se tiene que si

$$\frac{rd(x) - x}{x} = \varepsilon_x \Rightarrow rd(x) = x(1 + \varepsilon_x), \text{ con } |\varepsilon_x| \leq eps.$$

El número $rd(x) \in A$ tiene la propiedad:

$$|x - rd(x)| \leq |x - y| \quad \forall y \in A.$$

En el sistema binario, $rd(x)$ está definido de modo análogo. Comenzamos con la escritura de x en la forma. $|x| = a \times 2^b$, donde $a = 0.\alpha_1\alpha_2 \ldots \alpha_t\alpha_{t+1} \ldots, \alpha_i \in \{0, 1\}, \quad i = 1, 2, \ldots,$ y $\alpha_1 = 1$. Se tiene $1 > a \geq \frac{1}{2}$. Se define

$$\tilde{a} = \begin{cases} 0.\alpha_1 \ldots \alpha_t, & \text{si } \alpha_{t+1} = 0, \\ 0.\alpha_1 \ldots \alpha_t + 2^{-t}, & \text{si } \alpha_{t+1} = 1. \end{cases}$$

Entonces $rd(x) = sign(x) \times \tilde{a} \times 2^b$, y $\quad |\varepsilon_x| \leq eps = 2^{-t}$. Resulta que $rd(x) = x(1 + \varepsilon_x)$ con $|\varepsilon_x| \leq eps$.

Ahora analizamos los casos 1) y 3).

Puesto que un número finito de números b son útiles para expresar los exponentes en aritmética de punto flotante, hay desgraciadamente números $x \notin A$ tales que $rd(x) \notin A$. Así, si $t = 4$ y $b = 2$, consideramos los siguientes ejemplos:

1. $rd(0{,}31794 \times 10^{110}) = 0{,}3179 \times 10^{110} \notin A.$
2. $rd(0{,}99997 \times 10^{99}) = 0{,}1 \times 10^{100} \notin A.$
3. $rd(0{,}012345 \times 10^{-99}) = 0{,}1235 \times 10^{-100} \notin A.$
4. $rd(0{,}54321 \times 10^{-115}) = 0{,}5432 \times 10^{-115} \notin A.$

En los ejemplos 1) y 2), el exponente positivo es demasiado grande para almacenarlo en la memoria del computador, en estas condiciones se dice que el exponente está excedido de la capacidad de representación de exponentes en el computador, este caso se lo conoce como exponente en overflow. En el ejemplo 2) existe un overflow solo después de redondear. Situación análoga a la descrita precedentemente se presenta con los ejemplos 3) y 4); en tales casos se dice exponentes en underflow. En caso de underflow u overflow, pueden ser controlados, si por ejemplo se escribe:

$$\begin{aligned} rd(0{,}012345 \times 10^{-99}) &= 0{,}0123 \times 10^{-99} \in A, \\ rd(0{,}54321 \times 10^{-110}) &= 0 \in A, \\ rd(0{,}31794 \times 10^{110}) &= 0{,}3179 \times 10^{15} \times 10^{95}. \end{aligned}$$

Note que los números $0{,}3179 \times 10^{15}$, y 10^{95} pertenecen al conjunto A pero $0{,}3179 \times 10^{15} \times 10^{95} \notin A$. Para estos casos no se satisface la relación $rd(x) = x(1 + \varepsilon) \quad |\varepsilon| \leq eps$. En los computadores digitales estos números x que no pertenecen al conjunto M son tratados como irregularidades o errores en los datos. En el caso de underflow, $rd(x)$ puede ser indicado por 0 o se produce una detención en la ejecución del programa. En el caso de overflow, $rd(x)$ es indicado como un error en x y la inmediata detención del programa en ejecución.

Para evitar estos problemas es necesario incorporar en los programas contraseñas especiales o reescalar los datos de modo apropiado, lo que se traduce en elaborar programas especiales. Por lo dicho precedentemente y por abuso de lenguaje, podemos decir que existe una función $rd : \mathbb{R} \to \mathbb{A}$ definida por $rd(x) = x(1 + \varepsilon) \quad |\varepsilon| \leq eps$.

1.8. Aritmética de punto flotante

Operaciones aritméticas

Hemos denotamos con A el conjunto de números de máquina. Sean $x, y \in A$, en general, $x + y$, $x - y$, $x \times y$, $x/y \quad y \neq 0$, no son números de máquina. Definimos las operaciones aritméticas $\oplus$, $\ominus$, $\otimes$, $\oslash$ llamadas operaciones de punto flotante, como sigue:

$$\begin{aligned} x \oplus y &= rd(x+y), \qquad x \ominus y = rd(x-y), \\ x \otimes y &= rd(x \times y), \qquad x \oslash y = rd(x/y) \quad y \neq 0. \end{aligned}$$

De la definición de la función rd, resulta que

$$\begin{aligned} x \oplus y &= (x+y)(1+\varepsilon_1), \qquad x \ominus y = (x-y)(1+\varepsilon_2), \\ x \otimes y &= (x \times y)(1+\varepsilon_3), \qquad x \oslash y = (x/y)(1+\varepsilon_4), \end{aligned}$$

con $|\varepsilon_i| \leq eps \quad i = 1, 2, 3, 4.$

Las operaciones en punto flotante pueden no ser asociativas o distributivas, así : si $a, b, c \in A$, en general,

$$a \oplus (b \oplus c) \neq (a \oplus b) \oplus c, \qquad a \otimes (b \oplus c) \neq a \oplus b \oplus a \oplus c.$$

Comprobemos con un ejemplo. Sean $t = 5$ el número de cifras de la mantisa, $a = 0{,}21345 \times 10^{-2}$, $b = 0{,}33456 \times 10^{2}$, $c = -0{,}33341 \times 10^{2}$. Con estos datos, verifiquemos que $a \oplus (b \oplus c) \neq (a \oplus b) \oplus c$. Tenemos

$$b \oplus c = rd(0{,}33456 \times 10^{2} - 0{,}33341 \times 10^{2}) = rd(0{,}00115 \times 10^{2}) = 0{,}115 \times 10^{0},$$

luego

$$a \oplus (b \oplus c) = rd(0{,}21345 \times 10^{-2} + 0{,}115 \times 10^{0}) = rd(0{,}11714 \times 10^{0}) = 0{,}11714 \times 10^{0},$$

Calculemos $(a \oplus b) \oplus c$. Para ello calculamos primeramente $a \oplus b$. Tenemos

$$a \oplus b = rd(0{,}21345 \times 10^{-2} + 0{,}33456 \times 10^{2}) = rd(0{,}33458 \times 10^{2}) = 0{,}33458 \times 10^{2},$$

a continuación

$$(a \oplus b) \oplus c = rd(0{,}33458 \times 10^{2} - 0{,}33341 \times 10^{2}) = rd(0{,}00117 \times 10^{2}) = 0{,}117 \times 10^{0},$$

El valor exacto es $a + b + c = 0{,}1171345$. Note que $a \oplus (b \oplus c)$ se aproxima mejor a $a + b + c$.

Con la misma información verifiquemos que $a \otimes (b \oplus c) \neq a \otimes b \oplus a \otimes c$. Calculemos el lado izquierdo. Tenemos $b \oplus c = 0{,}115 \times 10^{0}$

$$\begin{aligned} a \otimes (b \oplus c) &= rd(0{,}21345 \times 10^{-2} * 0{,}115 \times 10^{0}) = rd(0{,}02454675 \times 10^{-2}) \\ &= rd(0{,}2454675 \times 10^{-3}) = 0{,}24547 \times 10^{-3}, \end{aligned}$$

Pasemos a calcular $a \otimes b \oplus a \otimes c$. Entonces

$$a \otimes b = rd(0{,}21345 \times 10^{-2} * 0{,}33456 \times 10^{2}) = rd(0{,}071411832) = 0{,}71412 \times 10^{-1},$$

$$a \otimes c = rd(-0{,}21345 \times 10^{-2} * 0{,}33341 \times 10^{2}) = -rd(0{,}0711663645 \times 10^{0}) = -0{,}71167 \times 10^{-1},$$

Con estos resultados calculemos $a \otimes b \oplus a \otimes c$:

$$a \otimes b \oplus a \otimes c = rd(0{,}71412 \times 10^{-1} - 0{,}71167 \times 10^{-1}) = rd(0{,}00245 \times 10^{-1}) = 0{,}245 \times 10^{-3},$$

Claramente $a \otimes (b \oplus c) = 0{,}24547 \times 10^{-3} \neq a \otimes b \oplus a \otimes c = 0{,}245 \times 10^{-3}$. Valor exacto $a(b + c) = 0{,}0002454675 = 0, 2454675 \times 10^{-3}$.

Expresiones aritméticas y funciones.

Sea E una expresión aritmética. En punto flotante la evaluación de E se nota con $fl(E)$. Sea E una función real definida en un subconjunto I de $\mathbb{R}$. El valor $E(x)$ en $x \in I$ en punto flotante se nota con $E_*(x)$ y se define por

$$E_*(x) = fl(E(x)).$$

Ejemplos

1. Sean $a, b, c \in \mathbb{R}$.
   a) Si $E = a + (b + c)$, entonces $fl(E) = a \oplus (b \oplus c)$.
   b) Si $E = (ab)c$, $fl(E) = (a \otimes b) \otimes c$.
   c) Si $E = a(b + c)$, $fl(E) = a \otimes (b \oplus c)$.

2. Sean $a = 0{,}18\times10^2$,$b = 0{,}3596\times10^0$,$c = 0{,}1\times10^1$,$t = 4$ el número de cifras de la mantisa, calculemos $E = \dfrac{a}{b} + c$ en punto flotante. De la definición de punto flotante de E, tenemos

$$\begin{aligned} fl(E) &= rd([0{,}18 \times 10^2 \oslash 0{,}3596 \times 10^0] \oplus 0{,}1 \times 10^1) = rd(0{,}50055611 \times 10^2 \oplus 0{,}1 \times 10^1) \\ &= rd(0{,}5006 \times 10^2 \oplus 0{,}1 \times 10^1) = 0{,}5106 \times 10^2. \end{aligned}$$

3. Sea $E(x) = sen(x)$ $x \in \mathbb{R}$. Entonces $E_*(x) = fl(sen(x))$ que lo notaremos $sen_*(x)$. Para $t = 5$, y $x = \dfrac{\pi}{6}$,se tiene $sen(\dfrac{\pi}{6}) = 0{,}5$, mientras que $rd(\dfrac{\pi}{6}) = 0{,}5236 \times 10^0$ y en consecuencia $\text{sen}_*(0{,}5236 \times 10^0) = 0{,}5$. En el capítulo 3 se propone un algoritmo de cálculo de sen(x) $x \in \mathbb{R}$.

4. Sea $E(x) = e^x$ $x \in \mathbb{R}$. Entonces $E_*(x) = fl(e^x)$ que lo notaremos e^x_*. Si $x = 0{,}5$ y $t = 5$, entonces $e^{0{,}5} = 1{,}648721171\ldots$, $e^{0{,}5}_* = 0{,}16487 \times 10^1$. En el capítulo 3 se propone un algoritmo de cálculo de e^x $x \in \mathbb{R}$.

5. $fl(\sqrt{x}) = \sqrt{x}^*$, $x \geq 0$. Para $t = 5$, $x = 0{,}14567$, $\sqrt{x} = 0{,}381667395\ldots$, $\sqrt{x}^* = 0{,}38167 \times 10^0$. Más adelante se muestra un algoritmo de cálculo de $\sqrt{x}$ $x > 0$.

Observación. Sean $a, b \in \mathbb{R}$. Las operaciones aritméticas en punto flotante se expresan de la manera siguiente.

$$\begin{aligned} a \oplus b &= rd(rd(a) + rd(b)), \qquad a \ominus b = rd(rd(a) - rd(b)), \\ a \otimes b &= rd(rd(a) * rd(b)), \qquad a \oslash b = rd(rd(a)/rd(b)) \quad rd(b) \neq 0. \end{aligned}$$

Por abuso de lenguaje, a las operaciones en punto flotante las notaremos del mismo que las operaciones aritméticas habituales con números reales.

1.9. Condicionamiento de funciones reales.

La calidad de la solución numérica de un problema (P) depende fuertemente del método numérico empleado y este a su vez depende de dos componentes importantes: el condicionamiento y la estabilidad; y, para problemas cuyas soluciones se aproximan mediante sucesiones, dependen a más de los componentes anteriores, de la convergencia.

En esta sección se introduce la noción de condicionamiento que es muy importantes en la construcción de algoritmos, procedimientos de cálculo y de la elaboración de programas computacionales, y que constituyen las bases que deben tenerse siempre presentes para el desarrollo de software en el cálculo científico. Trataremos primeramente el condicionamiento de funciones reales de una sola variable, a continuación trataremos el condicionamiento de funciones reales en varias variables.

1.9.1. Condicionamiento de funciones reales de una sola variable.

Sea $\varphi : [a,b] \to \mathbb{R}$ una función derivable en $]a,b[$. Ponemos $y = \varphi(x) \quad x \in [a,b]$. Investiguemos como el error absoluto $\triangle x$ de x influye en el cálculo de y, donde $\triangle x = \widetilde{x} - x$ y $\widetilde{x} = rd(x)$. Se pone $\widetilde{y} = \varphi(\widetilde{x})$. Nos interesa determinar la influencia de los errores (redondeo, truncamiento) del dato de entrada x, esto es, de Δx en el dato de salida $y = \varphi(x)$, es decir en $\widetilde{y} = \varphi(\widetilde{x})$ y como medir esa influencia.

Supongamos que la función φ es al menos dos veces derivable en $]a,b[$ y que $\mid \varphi'' \mid$ es acotada en $]a,b[$. Por el desarrollo de Taylor, se tiene $\varphi(\widetilde{x}) = \varphi(x) + \varphi'(x)(\widetilde{x} - x) + \frac{1}{2}(\widetilde{x} - x)^2\varphi''(\xi)$ con ξ entre x y $\widetilde{x}$, y $(\widetilde{x} - x)^2 = (\triangle x)^2 < eps$, lo que implica que $\frac{1}{2}(\widetilde{x} - x)^2 \mid \varphi''(\xi) \mid \simeq 0$.

Definición 5 *El error relativo de y se define mediante la relación*

$$\varepsilon_y = \frac{\widetilde{y} - y}{y} \quad y \neq 0,$$

donde $\widetilde{y} = \varphi(\widetilde{x})$.

Usando el desarrollo de Taylor en primera aproximación, tenemos

$$\triangle y = \widetilde{y} - y = \varphi(\widetilde{x}) - \varphi(x) = \varphi'(x)(\tilde{x} - x) = x\ \varphi'(x)\ \frac{\widetilde{x} - x}{x} = x\varphi'(x)\varepsilon_x \quad x \neq 0.$$

Luego,

$$\varepsilon_y = \frac{\widetilde{y} - y}{y} = \frac{\Delta y}{y} = \frac{x\varphi'(x)}{\varphi(x)}\varepsilon_x, \quad \varphi(x) \neq 0.$$

Note que si $\varphi''(x)$ existe, el término $\frac{1}{2}(\Delta x)^2\ \varphi''(x)$ se redondea por 0 debido a que $(\Delta x)^2$ se redondea por 0.

Definición 6 *El número real $c(x) = \frac{x\varphi'(x)}{\varphi(x)}$ con $\varphi(x) \neq 0$ se llama número de condicionamiento de la función φ en el punto x.*

El número de condicionamiento $c(x)$ indica cuán grande es el error relativo de y ante variaciones del dato de entrada x. Cuando $\mid c(x) \mid > 1$ el error relativo ε_y se amplifica y cuando $\mid c(x) \mid \leq 1$ el error relativo ε_y se contrae.

Definición 7 *Diremos que $y = \varphi(x)$ está bien condicionado si $|c(x)| \leq 1$. En el caso contrario, diremos que $y = \varphi(x)$ está mal condicionado.*

Ejemplos

1. Consideremos la función f definida por $f(x) = e^x \quad x \in \mathbb{R}$. Es conocido que la función f es derivable, y $f'(x) = e^x \quad \forall x \in \mathbb{R}$. El número de condicionamiento de esta función está definido como $c(x) = \frac{xf'(x)}{f(x)} = \frac{xe^x}{e^x} = x \quad \forall x \in \mathbb{R}$. Luego

   $$|c(x)| \leq 1 \Leftrightarrow |x| \leq 1 \Leftrightarrow x \in [-1,1]\,.$$

   Por lo tanto $y = e^x$ está bien condicionado si y solo si $x \in [-1,1]$, en el caso contrario la función f está mal condicionada.
   Definimos $g(x) = \left(e^{\frac{x}{n}}\right)^n \quad x \in \mathbb{R}$. Entonces

   $$c(x) = \frac{xg'(x)}{g(x)} = \frac{x\left[n(e^{\frac{x}{n}})^{n-1}e^{\frac{x}{n}} \cdot \frac{1}{n}\right]}{(e^{\frac{x}{n}})^n} = x.$$

La utilización de $g(x) = (e^{\frac{x}{n}})^n = e^x$ con $n \in \mathbb{Z}^+$, $x \in \mathbb{R}$ es más ventajoso del punto de vista de la elaboración de un algoritmo que permita evaluar e^x.

2. Sea $n \in \mathbb{Z}^+$ y f la función dada por $f(x) = x^n \quad x \in \mathbb{R}$. Para $x \neq 0$, tenemos

$$c(x) = \frac{xf'(x)}{f(x)} = \frac{x(nx^{n-1})}{x^n} = n.$$

Luego $y = x^n$ está bien condicionado si y solo si $n = 1$. Para $n > 1$, la función f está mál condicionada, esto significa que la potencia x^n está mal condicionada cuando $n > 1$. Es por esta razón que se evita el cálculo directo de las potencias. Anteriormente vimos algunos ejemplos de cálculos con polinomios en los que se evitan los cálculos directos de las potencias, de esta manera se mejora el condicionamiento con lo que se logra mejorar los resultados.

3. Considerar la función f definida por $f(x) = x^{\frac{1}{n}}$ con $n \in \mathbb{Z}^+$, $x > 0$. El número de condicionamiento está definido como

$$c(x) = \frac{xf'(x)}{f(x)} = \frac{x\frac{1}{n}x^{\frac{1}{n}-1}}{x^n} = \frac{1}{n}.$$

Resulta que $y = x^{1/n}$ está bien condicionado $\forall x \in \mathbb{R}^+, \quad n \in \mathbb{Z}^+$.

4. Sea $f(x) = sen\,(x) \quad x \in \mathbb{R}$. El número de condicionamiento de esta función está definido como

$$c(x) = \frac{x\cos(x)}{sen(x)} \qquad x \neq k\pi, \quad k \in \mathbb{Z}.$$

Para el análisis del número de condicionamiento $c(x)$ lo dividimos en dos partes.

a) Puesto que $c(x) = \dfrac{\cos(x)}{\dfrac{sen(x)}{x}}$. Entonces

$$\lim_{x\to 0} c(x) = \frac{\lim\limits_{x\to 0}\cos(x)}{\lim\limits_{x\to 0}\dfrac{sen(x)}{x}} = \frac{1}{1} = 1,$$

que muestra que en $x = 0$ se tiene una discontinuidad evitable. Además,

$$\lim_{x\to \frac{\pi}{2}} c(x) = \frac{\lim\limits_{x\to \frac{\pi}{2}} x\cos(x)}{\lim\limits_{x\to \frac{\pi}{2}} sen(x)} = 0.$$

b) Por otro lado, si escribamos $c(x) = \dfrac{x}{\tan(x)}$. Tenemos

$$|c(x)| \leq 1 \Leftrightarrow |x| \leq |\tan(x)| \quad x \in \left]-\frac{\pi}{2}, 0\right[\cup \left]0, \frac{\pi}{2}\right[.$$

Definimos $g(x) = -x + \tan(x)$. Resulta que $g'(x) = -1 + \sec^2(x)$. Entonces

$$g'(x) = 0 \Leftrightarrow \cos^2(x) = 1 \Leftrightarrow x = 2k\pi, \quad k \in \mathbb{Z}.$$

Además $g''(x) = 2\sec^2(x)\tan(x)$. Luego, $g''(x) > 0 \quad$ si $x \in \left]0, \frac{\pi}{2}\right[$, y $g''(x) < 0$ si $x \in \left]-\frac{\pi}{2}, 0\right[$. Adicionalmente, g es creciente en $\left]0, \frac{\pi}{2}\right[$, luego $g(x) > g(0) = 0$ con lo cual $\tan(x) > x$.

En conclusión

$$|c(x)| \leq 1 \Leftrightarrow x \in \left]-\frac{\pi}{2}, 0\right[\cup \left]0, \frac{\pi}{2}\right[.$$

Sea $\widetilde{c}(x) = \begin{cases} 1, \text{ si } x = 0, \\ c(x), \text{ si } x \in \left]-\frac{\pi}{2}, 0\right[\cup \left]0, \frac{\pi}{2}\right[. \end{cases}$ Entonces $|\widetilde{c}(x)| \leq 1 \quad \forall x \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$, es decir que $\text{sen}(x)$ está bien condicionado en el intervalo $\left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$. Esta propiedad será utilizada para aproximar $\text{sen}(x)$ mediante la serie de Taylor que se verá en el capítulo posterior.

5. Sea $\varphi : \mathbb{R}^+ \to \mathbb{R}$ función dada por $\varphi(x) = y^x$, donde $y > 0$ es fijo. Entonces

$$c(x) = \frac{x}{e^{x\ln(y)}} e^{x\ \ln(y)} \ln(y) = x\ln(y).$$

Luego

$$|c(x)| \leq 1 \Leftrightarrow |x| \leq \frac{1}{|\ln(y)|} \Leftrightarrow x \in \left[\frac{-1}{|\ln(y)|}, \frac{1}{|\ln(y)|}\right] \quad \text{con } y > 0,\ y \neq 1.$$

6. Sea $\varphi : \mathbb{R}^+ \to \mathbb{R}$ la función definida por $\varphi(x) = x^y$ con $y \in \mathbb{R}^+$ fijo. Se tiene

$$c(x) = \frac{x}{e^{y\ln(x)}} e^{y\ln(x)} \times \frac{y}{x} = y.$$

La función φ está bien condicionada si $|y| \leq 1$. Note que si $y \in \mathbb{N}$, $c(x) = y$ fue obtenido anteriormente.

7. Se desea calcular $f = (\sqrt{2}-1)^6$. Se da una aproximación de $\sqrt{2} \simeq 1{,}414$ y seis algoritmos para su cálculo

$$\begin{aligned} f_1 &= (\sqrt{2}-1)^6, \qquad f_2 = \frac{1}{(\sqrt{2}+1)^6}, \\ f_3 &= (3-2\sqrt{2})^3, \qquad f_4 = \frac{1}{(3+2\sqrt{2})^3}, \\ f_5 &= \frac{1}{99+70\sqrt{2}}, \qquad f_6 = 99-70\sqrt{2}. \end{aligned}$$

¿Qué algoritmo está bien condicionado?

Para responder a esta pregunta, primeramente vamos a cálcular los números de condicionamiento asociados con los algoritmos propuestos. Para el efecto denotamos con ε_x el error relativo al dato de entrada $x = \sqrt{2}$.

a) Sea f_1 la función dada por $f_1(x) = (x-1)^6$, entonces

$$f_1'(x) = 6(x-1)^5, \qquad f_1(1{,}414) = (1{,}414-1)^6 = 0{,}005,$$

luego

$$|\varepsilon_{f_1}| = \left|\frac{x}{f_1(x)} f_1'(x)\varepsilon_x\right| = \left|\frac{1{,}414}{0{,}005} \times 6 \times (0{,}414)^5\right| |\varepsilon_x| = 20{,}636\,|\varepsilon_x|\,.$$

b) Sea $f_2(x) = \dfrac{1}{(x+1)^6}$. Entonces $f_2'(x) = -\dfrac{6}{(x+1)^7}$. Resulta $f_2(1{,}414) = 0{,}005$ y

$$|\varepsilon_{f_2}| = \left|\frac{x}{f_2(x)} f_2'(x)\varepsilon_x\right| \left|\frac{1{,}414}{0{,}005}\right| \left(-\frac{6}{(1{,}414+1)^7}\right) |\varepsilon_x| = 3{,}552\,|\varepsilon_x|\,.$$

c) Consideramos la función f_3 definida como $f_3(x) = (3-2x)^3$. Entonces $f_3'(x) = -6(3-2x)^2$. Tenemos $f_3(1{,}414) = 0{,}005$, y

$$|\varepsilon_{f_3}| = \left|\frac{1{,}414}{0{,}005}\right| \left(-6(3-2\times 1{,}414)^2\right) |\varepsilon_x| = 50{,}198\,|\varepsilon_x|\,.$$

d) Tal como en los casos precedentes, sea $f_4(x) = \dfrac{1}{(3+2x)^3} \Longrightarrow f_4'(x) = -\dfrac{6}{(3+2x)^4}$. Se tiene $f_4(1{,}414) = 0{,}005$, y

$$|\varepsilon_{f_4}| = \left|\frac{1{,}414}{0{,}005}\right| \left(-\frac{6}{(3+2\times 1{,}414)^4}\right) |\varepsilon_x| = 1{,}471\,|\varepsilon_x|\,.$$

e) Sea $f_5(x) = \dfrac{1}{99+70x} \Longrightarrow f_5'(x) = -\dfrac{70}{(99+70x)^2}$. Entonces

$$|\varepsilon_{f_5}| = \left|\frac{1{,}414}{0{,}005} \times \frac{-70}{(99+70\times 1{,}414)^2}\right| |\varepsilon_x| = 0{,}5\,|\varepsilon_x|\,.$$

f) De la definición del algoritmo f_6, se sigue que $f_6(x) = 99 - 70x \Longrightarrow f_6'(x) = -70$. Resulta$f_6(1{,}414) = 0{,}020$,

$$|\varepsilon_{f_6}| = \left|\frac{1{,}414}{0{,}020} \times (-70)\right| |\varepsilon_x| = 4949{,}0\,|\varepsilon_x|\,.$$

Comparando los números de condicionamiento de cada una de las funciones, observamos que f_5 tiene el más pequeño número de condicionamiento, esto es, f_5 está bien condicionado, mientras que f_6 tiene el más grande número de condicionamiento, es decir que f_6 está mál condicionado y de hecho es el peor algoritmo que se puede utilizar para calcular el valor aproximado de $(\sqrt{2}-1)^6$. En conclusión, el mejor algoritmo para el cálculo de $(\sqrt{2}-1)^6$ con $\sqrt{2} \simeq 1{,}414$ es $f_5 = \dfrac{1}{99+70\sqrt{2}}$.

1.9.2. Condicionamiento de funciones reales en varias variables

Sean $\Omega \subset \mathbb{R}^n$ un abierto y $\overrightarrow{\varphi}$ una función de Ω en $\mathbb{R}^m$ que suponemos diferenciable en todo punto de Ω, la función $\overrightarrow{\varphi}$ se denomina campo vectorial. Ponemos $\vec{y} = \vec{\varphi}(\vec{x}) = \begin{bmatrix} \varphi_1(\overrightarrow{x}) \\ \vdots \\ \varphi_m(\overrightarrow{x}) \end{bmatrix}$ $\vec{x}^T = (x_1 \ldots, x_n) \in \Omega$, donde $\varphi_1, \cdots, \varphi_m$ son campos escalares diferenciables en Ω.

Para $\vec{x}^T = (x_1 \ldots, x_n) \in \Omega$, ponemos $\widetilde{x}_i = rd(x_i) \quad i = 1, \ldots, n$, y definimos $\widetilde{x}^T = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$, $\Delta\overrightarrow{x}^T = \widetilde{x}^T - \overrightarrow{x}^T$. El error relativo de x_i está definido mediante la relación:

$$\varepsilon_{x_i} = \frac{\tilde{x}_i - x_i}{x_i} \qquad x_i \neq 0, \quad i = 1, \ldots, n.$$

Se define $\widetilde{y} = \overrightarrow{\varphi}(\widetilde{x}) = \begin{bmatrix} \varphi_1(\widetilde{x}) \\ \vdots \\ \varphi_m(\tilde{x}) \end{bmatrix}$, y, $\triangle\overrightarrow{y} = \widetilde{y} - \overrightarrow{y} = \begin{bmatrix} \widetilde{y}_1 - y_1 \\ \vdots \\ \widetilde{y}_m - y_m \end{bmatrix} = \begin{bmatrix} \triangle y_1 \\ \vdots \\ \triangle y_n \end{bmatrix}$.

Determinemos el error relativo $\varepsilon_{y_i} = \dfrac{\Delta y_i}{y_i} \quad i = 1, \ldots, m$. Usando el desarrollo de Taylor en primera aproximación, eliminando los desarrollos de orden superior, tenemos

$$\Delta y_i = \widetilde{y}_i - y_i = \varphi_i(\tilde{x}) - \varphi_i(\vec{x}) = \nabla\varphi_i(\vec{x}).\Delta\vec{x} = \sum_{j=1}^{n} (\widetilde{x}_j - x_j)\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j}.$$

Para $\vec{x}^T = (x_1 \ldots, x_n) \in \Omega$ tal que $x_j \neq 0, \quad j = 1, \ldots, n$, se tiene la siguiente relación

$$\widetilde{x}_j - x_j = \frac{\widetilde{x}_j - x_j}{x_j} x_j = x_j\varepsilon_{x_j} \qquad x_j \neq 0, \quad j = 1, \ldots, n,$$

y en consecuencia

$$\Delta y_i = \sum_{j=1}^{n} (\widetilde{x}_j - x_j)\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j} = \sum_{j=1}^{n} x_j \frac{\partial\varphi_i(\overrightarrow{x})}{\partial x_j}\varepsilon_{x_j} \qquad i = 1, \ldots, m.$$

Luego, para $y_i \neq 0 \quad i = 1, \ldots, m$, se tiene

$$\varepsilon_{y_i} = \frac{\Delta y_i}{y_i} = \frac{1}{y_i}\sum_{j=1}^{n} x_j \frac{\partial\varphi_i(\overrightarrow{x})}{\partial x_j}\varepsilon_{x_j} = \sum_{j=1}^{n} \frac{x_j}{y_i}\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j}\varepsilon_{x_j} = \sum_{j=1}^{n} \frac{x_j}{\varphi_i(\overrightarrow{x})}\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j}\varepsilon_{x_j} \qquad i = 1, \ldots, m.$$

Observamos que cada ε_{y_i} depende de los factores de amplificación $\frac{x_j}{\varphi_i(\overrightarrow{x})}\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j}$ de ε_{x_j}, $\quad i=1,\ldots,m$, $j=1,\ldots,n$,

Definición 8 *El conjunto de números reales* $\{C_{ij}(\overrightarrow{x}) \mid i=1,\ldots,m, \quad j=1,\ldots,n\}$, *donde* $C_{ij}(\overrightarrow{x})$ *está definido como*

$$C_{ij}(\overrightarrow{x}) = \frac{x_j}{\varphi_i(\overrightarrow{x})}\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j} \text{ con } \varphi_i(\overrightarrow{x}) \neq 0, \quad i=1,\ldots,m, \quad j=1,\ldots,n,$$

se llaman números de condicionamiento de la función $\overrightarrow{\varphi}$ *en* $\overrightarrow{x} \in \Omega$.
La matriz $C(\overrightarrow{x}) = (C_{ij}(\overrightarrow{x}))_{m\times n}$ *se llama matriz de condicionamiento de* $\overrightarrow{\varphi}$ *en* $\overrightarrow{x} \in \Omega$.

En el caso en que $m=1$, esto es, φ es un campo escalar, la matriz de condicionamiento de φ en $\overrightarrow{x} \in \Omega$ se identifica con el vector fila $C(\overrightarrow{x})$ definido como

$$C(\overrightarrow{x}) = \left(\frac{x_i}{\varphi(\overrightarrow{x})}\frac{\partial\,\varphi}{\partial x_i}(\overrightarrow{x}), \ldots, \frac{x_n}{\varphi(\overrightarrow{x})}\,\frac{\partial\varphi}{\partial x_n}(\overrightarrow{x})\right),$$

al que lo denominaremos vector de condicionamiento de φ en $\overrightarrow{x} \in \Omega$.

Definimos

$$\overrightarrow{\varepsilon}_{\overrightarrow{y}} = \begin{bmatrix} \varepsilon_{y_1} \\ \vdots \\ \varepsilon_{y_n} \end{bmatrix} = \begin{bmatrix} \displaystyle\sum_{j=1}^{n} \frac{x_j}{\varphi_1(\overrightarrow{x})}\,\frac{\partial\varphi_1}{\partial x_j}(\overrightarrow{x})\varepsilon_{x_j} \\ \vdots \\ \displaystyle\sum_{j=1}^{n} \frac{x_j}{\varphi_m(\overrightarrow{x})}\,\frac{\partial\varphi_m(\overrightarrow{x})}{\partial x_j}\varepsilon_{x_j} \end{bmatrix} = C(\overrightarrow{x})\overrightarrow{\varepsilon}_{\overrightarrow{x}},$$

donde $\overrightarrow{\varepsilon}_{\overrightarrow{x}} = \begin{bmatrix} \varepsilon_{x_1} \\ \vdots \\ \varepsilon_{xn} \end{bmatrix}$. Así, $\overrightarrow{\varepsilon}_{\overrightarrow{y}} = C(\overrightarrow{x})\overrightarrow{\varepsilon}_{\overrightarrow{x}}$.

Definición 9 *Se dice que* $\overrightarrow{y} = \overrightarrow{\varphi}(\overrightarrow{x})$ *está bien condicionado en* $\overrightarrow{x} \in \Omega$ *si y solo si* $\mid C_{ij}(\overrightarrow{x}) \mid \leq 1$ $\forall i=1,\ldots,m, \quad j=1,\ldots,n$. *En el caso contrario, se dice que* $\overrightarrow{y} = \overrightarrow{\varphi}(\overrightarrow{x})$ *está mal condicionado en* $\overrightarrow{x} \in \Omega$.

Para determinar el condicionamiento de un campo vectorial diferenciable $\overrightarrow{\varphi}$ en $\overrightarrow{x} \in \Omega$ se deben estudiar todos los números de condicionamiento $C_{ij}(\overrightarrow{x}) \quad i=1,\ldots,m, \quad j=1,\ldots,n$. Hacemos notar que solo en pocos casos es posible determinar $\overrightarrow{x}$ tal que $\mid C_{ij}(\overrightarrow{x}) \mid \leq 1$. En la generalidad de los casos, es muy difícil y casi imposible determinar $\overrightarrow{x}$ tal que $\mid C_{ij}(\overrightarrow{x}) \mid \leq 1$, por lo que se recurre a otros procedimientos para estimar el condicionamiento. Así, en algunos casos el número de condicionamiento C se define por la desigualdad (cociente de Raileygh-Ritz):

$$\frac{\parallel \overrightarrow{\varphi}(\widetilde{x}) - \overrightarrow{\varphi}(\overrightarrow{x}) \parallel}{\parallel \overrightarrow{\varphi}(\overrightarrow{x}) \parallel} \leq C\frac{\parallel \widetilde{x} - \overrightarrow{x} \parallel}{\parallel \overrightarrow{x} \parallel} \qquad \overrightarrow{\varphi}(\overrightarrow{x}) \neq 0, \quad \overrightarrow{x} \neq 0,$$

donde $C > 0$ es el número de condicionamiento.

Ejemplos

1. Sea φ la función de $\mathbb{R}^2$ en $\mathbb{R}$ definida como $\varphi(x,y) = x+y \quad (x,y) \in \mathbb{R}^2$. Supongamos que para $(x,y) \in \mathbb{R}^2$ se tiene $z = \varphi(x,y) \neq 0$. Determinemos el error relativo de z. Este está dado como sigue:

$$\varepsilon_z = \frac{x}{\varphi(x,y)}\,\frac{\partial\varphi}{\partial x}(x,y)\;\varepsilon_x + \frac{y}{\varphi(x,y)}\,\frac{\partial\varphi}{\partial y}(x,y)\;\varepsilon_y = \frac{x}{x+y}\varepsilon_x + \frac{y}{x+y}\varepsilon_y.$$

Los números de condicionamiento de φ en $(x,y) \in \mathbb{R}^2$ están definidos como $C_x = \dfrac{x}{x+y}$, $C_y = \dfrac{y}{x+y}$, y el vector de condicionamiento de φ en $(x,y) \in \mathbb{R}^2$ está definido como $C(x,y) = \left(\dfrac{x}{x+y}, \dfrac{y}{x+y}\right)$.

Analicemos los números de condicionamiento C_x y C_y.

a) Si $\begin{cases} x>0, & y>0, \\ x<0, & y<0, \end{cases}$ o entonces $\mid C_x(x,y) \mid < 1$, $\mid C_y(x,y) \mid < 1$. Luego $z = \varphi(x,y)$ está bien condicionado.

b) Si $x>0$ e $y<0$ tal que $x \neq -y$, entonces al menos uno de los números $|C_x|$ o $|C_y|$ es mayor que 1; en cuyo caso $z = \varphi(x,y)$ está mal condicionado.

En conclusión, la suma de dos números positivos (respectivamente negativos) está bien condicionada, mientras que la suma de dos números uno positivo y otro negativo está mal condicionada, esto equivale a decir que si x, y son números reales positivos, la resta $x-y$ está mal condicionada y consecuentemente la resta de dos números positivos es una operación peligrosa fundamentalmente si $x \neq -y$, $x \simeq -y$.

El resultado que acabamos de obtener se puede extender a sumas de tres o más números reales. Así, sean $x_1, \cdots, x_m \in \mathbb{R}$ y $z = x_1 + \cdots + x_m$. Entonces z está bien condicionado si y solo si $x_i > 0$ $\forall i = 1, \cdots, m$ (respectivamente $x_i < 0 \quad \forall i = 1, \cdots, m$), en el caso contrario tenemos que z está mal condicionado, más áun, las sumas y restas alternadas de números reales positivos está mal condicionada, por lo que este tipo de cálculos son peligrosos ya que amplifican los errores. Para aclarar más estas ideas, sean $x_1, \cdots, x_{2m} \in \mathbb{R}^+$ y $z = x_1 - x_2 + x_3 - x_4 + \cdots + x_{2m-1} - x_{2m}$. Esta suma está mal condicionada, ¿cómo mejorar el resultado? Ecribimos z en la siguiente forma

$$z = x_1 + x_3 + \cdots + x_{2n-1} - x_2 - x_4 - \cdots - x_{2m} = x_1 + x_3 + \cdots + x_{2n-1} - (x_2 + x_4 + \cdots + x_{2m})$$

La sumas $z_1 = x_1 + x_3 + \cdots + x_{2n-1}$, $z_2 = x_2 + x_4 + \cdots + x_{2m}$ están bien condicionadas, luego $z = z_1 - z_2$ con lo que se mejora el resultado. Más adelante se exhiben ejemplos.

2. Sea φ la función de $\mathbb{R}^2$ en $\mathbb{R}$ definida como $\varphi(x,y) = xy \quad (x,y) \in \mathbb{R}^2$. Supongamos que para $(x,y) \in \mathbb{R}^2$ se tiene $z = \varphi(x,y) \neq 0$, esto es $x \neq 0$, $y \neq 0$, entonces,

$$C(x,y) = \left(\frac{x}{\varphi(x,y)}\,\frac{\partial \varphi}{\partial x}(x,y),\ \frac{y}{\varphi(x,y)}\frac{\partial \varphi}{\partial y}(x,y)\right) = \left(\frac{x}{xy}y,\ \frac{y}{xy}x\right) = (1,1).$$

Se tiene que $C_x(x,y) = 1$, $C_y(x,y) = 1$, por lo que el producto de dos números reales no nulos está bien condicionado y por tanto el producto de dos números no es una operación peligrosa.

3. Sean $p, q \in \mathbb{R}$ tales que $p^2 - 4q \geq 0$. Consideramos la ecuación: $x \in \mathbb{R}$ tal que $x^2 + px + q = 0$ cuyas raíces reales son

$$x_1 = \frac{1}{2}\left(-p + \sqrt{p^2-4q}\right), \qquad x_2 = \frac{1}{2}\left(-p - \sqrt{p^2-4q}\right), \quad \text{donde } p^2 - 4q \geq 0.$$

Estas raíces dependen de p y q, lo que nos permite definir las funciones reales φ, θ como

$$\begin{cases} \varphi(p,q) = \dfrac{1}{2}\left(-p + \sqrt{p^2-4q}\right), \\ \theta(p,q) = -\dfrac{1}{2}\left(p + \sqrt{p^2-4q}\right), \end{cases} \quad (p,q) \in \mathbb{R}^2 \text{ tal que } p^2 - 4q \geq 0.$$

La función φ está asociada a la raíz x_1 mientras que la función θ está asociada a la raíz x_2. Estudiemos el condicionamiento de la primera raíz, esto es, el condicionamiento de la función φ. Tenemos

$$\begin{cases} \dfrac{\partial \varphi}{\partial p}(p,q) = \dfrac{-p + \sqrt{p^2-4q}}{2\sqrt{p^2-4q}}, \\ \dfrac{\partial \varphi}{\partial q}(p,q) = -\dfrac{1}{\sqrt{p^2-4q}}, \end{cases} \quad (p,q) \in \mathbb{R}^2 \text{ tal que } p^2 - 4q > 0.$$

Luego

$$\varepsilon_\varphi = \frac{p}{\varphi(p,q)}\frac{\partial\varphi}{\partial p}\varepsilon_p + \frac{q}{\varphi(p,q)}\frac{\partial\varphi}{\partial q}\varepsilon_q = -\frac{p}{\sqrt{p^2-4q}}\varepsilon_p + \frac{p+\sqrt{p^2-4q}}{2\sqrt{p^2-4q}}\varepsilon_q.$$

Los números de condicionamiento de φ están definidos como $\begin{cases} C_p(p,q) = -\dfrac{p}{\sqrt{p^2-4q}}, \\ C_q(p,q) = \dfrac{p+\sqrt{p^2-4q}}{2\sqrt{p^2-4q}}, \end{cases}$ siempre que $(p,q) \in \mathbb{R}^2$ tal que $p^2-4q>0$. Analicemos cada uno de estos números de condicionamiento. Si $q<0$, se tiene

$$\left|\frac{p}{\sqrt{p^2-4q}}\right| < 1, \qquad \left|\frac{p+\sqrt{p^2-4q}}{2\sqrt{p^2-4q}}\right| < 1,$$

con lo cual $x_1 = \varphi(p,q)$ está bien condicionado. Si $q>0$ tal que $p^2+4q>0$, φ está mal condicionado. El número de condicionamiento $\mid C_p(p,q) \mid$ es mucho más grande aún en la situación siguiente: $(p,q) \in \mathbb{R}^2$ tal que $q>0$, $p^2 \simeq 4q$ de modo que $p^2+4q>0$. Esto nos muestra que no es conveniente calcular x_1 con la fórmula arriba propuesta, sino con la que se obtiene del modo siguiente:

$$x_1 = \frac{1}{2}\left(-p+\sqrt{p^2-4q}\right) = \frac{1}{2}\frac{\left(-p+\sqrt{p^2-4q}\right)\left(-p-\sqrt{p^2-4q}\right)}{-p-\sqrt{p^2-4q}} = -\frac{2q}{p+\sqrt{p^2-4q}}.$$

Veamos un ejemplo numérico de esta situación. Consideremos la ecuación $x \in \mathbb{R}$ solución de $x^2+62{,}10x+1=0$. Entonces $x_1 = \frac{1}{2}\left(-62{,}10+\sqrt{(62{,}10)^2-4}\right) = -0{,}1610723\times 10^{-1}$. Efectuemos el cálculo de x_1 con 4 cifras decimales en aritmética de punto flotante, se tiene

$$\begin{aligned} \tilde{x}_1 &= fl(x_1) = 0{,}5\times 10^0 * \left(-0{,}6210\times 10^2 + \sqrt{(0{,}6210\times 10^2) - 0{,}4\times 10^{1^*}}\right) \\ &= 0{,}5\times 10^0 * (-0{,}6210\times 10^2 + 0{,}6206\times 10^2) = -0{,}2\times 10^{-2}. \end{aligned}$$

Utilicemos ahora la nueva escritura de x_1. Obtenemos

$$\widetilde{t_1} = fl(x_1) = -\frac{0{,}2\times 10^1 \ * \ 0{,}1\times 10^1}{0{,}6210\times 10^2 + 0{,}6207\times 10^2} = -\frac{0{,}2\times 10^1}{0{,}1242\times 10^3} = -0{,}1610\times 10^1.$$

Se observa que $\widetilde{t_1}$ es una mejor aproximación de x_1.

Nota. Tomando en consideración el valor absoluto de los números de condicionamiento, de los ejemplos se establece la jerarquía de las operaciones siguientes: la radicación de números reales positivos, la suma de números reales positivos (respectivamente suma de números reales negativos) son consideradas operaciones no peligrosas. A continuación se tiene el producto y cociente de números reales. La potenciación está bien condicionada si el exponente es igual a 1, por este motivo el esquema de Hörner evita el cálculo directo de las potencias. La suma de números reales de signos opuestos es una operación peligrosa ya que al menos un número de condicionamiento es mayor que 1 lo que amplifica los errores. Por esta razón debe evitarse sumas sucesivas con números reales de signos opuestos. De preferencia deben escribirse los algoritmos de modo que se tengan sumas de números positivos y reducir como sea posible las sumas de números con signos opuestos. Igualmente, debe evitarse el cálculo directo de las potencias con exponentes mayores que 1.

1.10. Propagación de los errores.

Ejemplos

1. Sean $a,b,c, \in \mathbb{R}$, y $E = a+b+c$. Se tiene $E = a+(b+c) = (a+b)+c = (a+c)+b$, y por tanto se disponen de tres algoritmos para evaluar E.

Primer algoritmo. Tenemos $E = a + (b + c)$, que puede verse como la composición de funciones siguiente:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} \xrightarrow{\varphi_0} \begin{bmatrix} a \\ \\ b+c \end{bmatrix} \xrightarrow{\varphi_1} a + (b + c),$$

donde $\varphi_0 : \mathbb{R}^3 \longrightarrow \mathbb{R}^2$ es la función definida por $\varphi_0(x, y, z) = \begin{bmatrix} x \\ y + z \end{bmatrix}$, y $\varphi_1 : \mathbb{R}^2 \rightarrow \mathbb{R}$ es la función dada por $\varphi_1(u, v) = u + v$.

Luego

$$E = (\varphi_1 \circ \varphi_0)(a, b, c) = \varphi_1(\varphi_0(a, b, c)) = \varphi_1(a, b + c) = a + (b + c).$$

Segundo algoritmo. En este caso $E = (a + b) + c$, que puede expresarse como el resultado de la siguiente composición de funciones:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} \xrightarrow{\varphi_0} \begin{bmatrix} a+b \\ \\ c \end{bmatrix} \xrightarrow{\varphi_1} (a + b) + c,$$

donde $\varphi_0 : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ es la función definida por $\varphi_0(x, y, z) = \begin{bmatrix} x + y \\ z \end{bmatrix}$, $\varphi_1 : \mathbb{R}^2 \rightarrow \mathbb{R}$ es la función definida por $\varphi_1(u, v) = u + v$. Se tiene $E = (a + b) + c = (\varphi_1 \circ \varphi_0)(a, b, c)$.

De manera análoga se formula el tercer algoritmo.

Consideremos el segundo algoritmo, esto es $E = (a+b)+c$. Pongamos $\widetilde{E} = fl((a+b)+c)$. Según las operaciones elementales en punto flotante, tenemos: $\eta = fl(a + b) = (a + b)(1 + \varepsilon_1)$,

$$\begin{aligned} \widetilde{E} &= fl(\eta + c) = (\eta + c)(1 + \varepsilon_2) = [(a + b)(1 + \varepsilon_1) + c](1 + \varepsilon_2) \\ &= a + b + c + (a + b)\varepsilon_1 + (a + b + c)\varepsilon_2 + (a + b)\varepsilon_1\varepsilon_2 \\ &= E + (a + b)\varepsilon_1 + y\ \varepsilon_2 + (a + b)\varepsilon_1\ \varepsilon_2. \end{aligned}$$

Luego,

$$\varepsilon_E = \frac{\widetilde{E} - E}{E} = \varepsilon_2 + \frac{a + b}{a + b + c}(1 + \varepsilon_2)\varepsilon_1 \quad \text{si } E = (a + b) + c \neq 0.$$

Puesto que $|\varepsilon_1| \leq eps$, $|\varepsilon_2| \leq eps$, se tiene que $\varepsilon_1\varepsilon_2 \simeq 0$, entonces

$$\varepsilon_E = \varepsilon_2 + \frac{a + b}{a + b + c}\varepsilon_1.$$

Para el primer y tercer algoritmos, procediendo en forma similar al segundo, se obtienen respectivamente los resultados siguientes:

$$\widetilde{\varepsilon}_E = \widetilde{\varepsilon}_2 + \frac{b + c}{a + b + c}\widetilde{\varepsilon}_1, \qquad \widehat{\varepsilon}_E = \widehat{\varepsilon}_2 + \frac{a + c}{a + b + c}\widehat{\varepsilon}_1.$$

Si a, b, c son positivos o todos negativos, los 3 algoritmos están bien condicionados. Mientras que si, por ejemplo, $a < 0$, y, b, c son positivos, la evaluación de y dependerá del algoritmo.

Sean $a = -0{,}33341 \times 10^2$, $b = 0{,}21345 \times 10^{-2}$, $c = 0{,}33456 \times 10^2$. Calculando con 5 cifras decimales de precisión, tenemos

$$\begin{aligned} \frac{a + b}{a + b + c} &= \frac{-0{,}33341 \times 10^2 + 0{,}21345 \times 10^{-2}}{-0{,}33341 \times 10^2 + 0{,}21345 \times 10^{-2} + 0{,}33456 \times 10^2} = -284{,}6, \\ \frac{b + c}{a + b + c} &= 285{,}63, \\ \frac{a + c}{a + b + c} &= 0{,}982. \end{aligned}$$

El algoritmo a elegir es $(a+c)+b$. Observe los resultados siguientes:

$$(a+b)+c = 0{,}117, \qquad (b+c)+a = 0{,}117, \qquad ,(a+c)+b = 0{,}11713.$$

Valor exacto $E = a+b+c = 0{,}1171345$.

2. De manera más general, sean $a_1, \ldots, a_n \in \mathbb{R}$ y $E = \sum_{i=1}^{n} a_i$.

El algoritmo (no eficiente) para la evaluación de E es el siguiente:

Algoritmo

Datos de entrada: n, $a_1, \cdots, a_n$.

Datos de salida: E.

1. $E = a_1$.
2. Para $k = 2, \ldots, n$

   $E = E + a_k$.

   Fin de bucle k
3. Fin.

Como en cada paso del bucle del algoritmo, se suma un dato, este procedimiento se formula usando funciones como sigue:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \xrightarrow{\varphi_1} \begin{bmatrix} a_1+a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix} \xrightarrow{\varphi_2} \begin{bmatrix} a_1+a_2+a_3 \\ a_4 \\ \vdots \\ a_n \end{bmatrix} \longrightarrow \cdots \xrightarrow{\varphi_{n-2}}$$

$$\begin{bmatrix} a_1+a_2+\cdots+a_{n-1} \\ a_n \end{bmatrix} \xrightarrow{\varphi_{n-1}} \sum_{i=1}^{n} a_i,$$

donde $\varphi_1, \varphi_2, ..., \varphi_{n-1}$ se denominan funciones elementales definidas como sigue:

$$\begin{aligned} \varphi_1 &: \mathbb{R}^n \to \mathbb{R}^{n-1}, \quad \varphi_1(a_1, \ldots, a_n) = (a_1+a_2, a_3 \ldots, a_n), \\ \varphi_2 &: \mathbb{R}^{n-1} \to \mathbb{R}^{n-2}, \quad \varphi_2(x_1, \ldots, x_{n-1}) = (x_1+x_2+x_3, \ldots, x_{n-1}), \\ &\vdots \\ \varphi_{n-1} &: \mathbb{R}^2 \to \mathbb{R}, \quad \varphi_{n-1}(u,v) = u+v. \end{aligned}$$

Entonces

$$\begin{aligned} E &= \left(\varphi_{n-1} \circ \varphi_{n-2} \circ \cdots \circ \varphi_2 \circ \varphi_1\right)(a_1, \ldots, a_n) = \varphi_{n-1} \circ \varphi_{n-2} \circ \cdots \circ \varphi_2\left(\varphi_1(a_1, \ldots, a_n)\right) \\ &= \varphi_{n-1} \circ \varphi_{n-2} \circ \cdots \circ \varphi_2(a_1+a_2, a_3, \ldots, a_n) = \varphi_{n-1} \circ \varphi_{n-2} \circ \cdots \circ \varphi_3(a_1+a_2+a_3, a_4, \ldots, a_n) \\ &\vdots \\ &= \varphi_{n-1}\left(\varphi_{n-2}(a_1+a_2+\cdots+a_{n-2}, a_{n-1}, a_n)\right) = \varphi_{n-1}(a_1+a_2+\cdots+a_{n-1}, a_n) \\ &= \sum_{i=1}^{n} a_i. \end{aligned}$$

3. Sean $a, b \in \mathbb{R}$. Supongamos que debemos calcular $E = a^2 - b^2 = (a+b)(a-b)$. Sabemos que $a^2 - b^2 = (a+b)(a-b)$, por lo tanto E puede calcularse de dos maneras: $E = a^2 - b^2$ y $E = (a+b)(a-b)$. Podemos describir estos dos procesos de cálculo mediante funciones reales apropiadas que describan cada operación elemental que se realiza.

Primer procedimiento: el cálculo de $E = a^2 - b^2$ podemos realizarlo mediante la siguiente secuencia de funciones:

$$\begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{\varphi_0} \begin{bmatrix} a^2 \\ b^2 \end{bmatrix} \xrightarrow{\varphi_1} a^2 - b^2,$$

donde φ_0 es la función de $\mathbb{R}^2$ en sí mismo definida como $\varphi_0(a,b) = \left(a^2, b^2\right) \quad (a,b) \in \mathbb{R}^2$, φ_1 es la función de $\mathbb{R}^2$ en $\mathbb{R}$ definida por $\varphi_1(u,v) = u - v \quad (u,v) \in \mathbb{R}^2$. Entonces, por la composición de funciones, tenemos

$$E = (\varphi_1 \circ \varphi_0)(a,b) = \varphi_1(a^2, b^2) = a^2 - b^2 \quad \forall (a,b) \in \mathbb{R}^2.$$

Segundo procedimiento: el cálculo de $E = (a+b)(a-b)$ se ejecuta mediante la aplicación de las siguientes funciones

$$\begin{bmatrix} a \\ b \end{bmatrix} \overset{\varphi_0}{\to} \begin{bmatrix} a+b \\ a-b \end{bmatrix} \overset{\varphi_1}{\to} (a+b)(a-b),$$

con φ_0 la función de $\mathbb{R}^2$ en $\mathbb{R}^2$ definida como $\varphi_0(a,b) = (a+b, a-b) \quad (a,b) \in \mathbb{R}^2$, φ_1 función de $\mathbb{R}^2$ en $\mathbb{R}$ definida $\varphi_1(x,y) = x * y \quad \forall (x,y) \in \mathbb{R}^2$. Mediante la composición de funciones se verifica inmediatamente que $E = (\varphi_1 \, o \, \varphi_o)(a,b) \quad \forall (a,b) \in \mathbb{R}^2$.

Observación. Supongamos que un problema consiste en calcular $y_1, \ldots, y_m$ a partir de datos de entrada $x_1, \ldots, x_n$. Ponemos $\overrightarrow{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, $\quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$. Supongamos que existe una función $\overrightarrow{\varphi} : D \to \mathbb{R}^m$ tal que

$$\overrightarrow{y} = \overrightarrow{\varphi}(\overrightarrow{x}) = \begin{bmatrix} \varphi_1(\overrightarrow{x}) \\ \vdots \\ \varphi_m(\overrightarrow{x}) \end{bmatrix} \quad \overrightarrow{x} \in D,$$

donde $D \subset \mathbb{R}^n$, $\varphi_j : D \to \mathbb{R}, \quad j = 1, \ldots, m$.

En cada etapa del cálculo hay un conjunto de números a operarse a partir de datos de entrada x_i, $i = 1, \ldots n$ y cada operación corresponde a la transformación del nuevo conjunto a operarse. Escribamos secuencialmente el conjunto de datos a operarse como un vector.

$$\overrightarrow{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_{n_i}^{(i)} \end{bmatrix} \in R^{n_i},$$

y asociamos la operación elemental con una función.

$$\overrightarrow{\varphi}^{(i)} : D_i \to R^{n_{i+1}}, \qquad D_i \subset R^{n_i},$$

de modo que $\overrightarrow{x}^{(i+1)} = \vec{\varphi}^{(i)}\left(\vec{x}^{(i)}\right)$, donde $\overrightarrow{x}^{(i+1)}$ es el resultado de la transformación del conjunto operado y la función $\varphi^{(i)}$ está definida de modo único salvo permutaciones en las operaciones $\overrightarrow{x}^{(i)}$ y $\overrightarrow{x}^{(i+1)}$.

Dado un algoritmo para el cálculo de $\overrightarrow{y} = \overrightarrow{\varphi}(\overrightarrow{x})$, la secuencia de operaciones elementales de la descomposición de $\overrightarrow{\varphi}$ en una secuencia de funciones elementales:

$$\begin{aligned} \overrightarrow{\varphi}^{(i)} &: D_i \to D_{i+1} \quad i = 0, 1, \ldots r, \qquad D_i \subset \mathbb{R}^{n_i}; \\ \overrightarrow{\varphi} &= \vec{\varphi}^{(r)} \circ \overrightarrow{\varphi}^{(r-1)} \circ \cdots \circ \vec{\varphi}^{(0)}, \\ D_0 &= D, \quad D_{r+1} \subset \mathbb{R}^{n_{r+1}} = R^m, \end{aligned}$$

que caracterizan al algoritmo. Así

$$\begin{aligned} \overrightarrow{y} &= \overrightarrow{\varphi}(\overrightarrow{x}) = \left(\overrightarrow{\varphi}^{(r)} \circ \overrightarrow{\varphi}^{(r-1)} \circ \cdots \circ \overrightarrow{\varphi}^{(0)}\right)(\overrightarrow{x}) = \overrightarrow{\varphi}^{(r)} \circ \overrightarrow{\varphi}^{(r-1)} \circ \cdots \circ \overrightarrow{\varphi}^{(1)}\left(\overrightarrow{\varphi}^{(0)}(\overrightarrow{x})\right) \\ &= \overrightarrow{\varphi}^{(r)} \circ \overrightarrow{\varphi}^{(r-1)}\left(\overrightarrow{x}^{(r-1)}\right) = \overrightarrow{\varphi}^{(r)}(x^r). \end{aligned}$$

Para mejor comprensión observe los ejemplos 1), 2) y 3) de esta sección.

1.11. Estabilidad numérica. Convergencia.

Sean $D \subset \mathbb{R}^n$ y $\overrightarrow{\varphi} : D \to \mathbb{R}^m$ una función. Ponemos $\overrightarrow{y} = \overrightarrow{\varphi}(\overrightarrow{x}) \quad \overrightarrow{x} \in D$. Dado un algoritmo de cómputo de $\vec{y} = \varphi(\vec{x})$, digamos

$$\overrightarrow{y} = \overrightarrow{\varphi}_{r-1} \circ \overrightarrow{\varphi}_{r-2} \circ \cdots \circ \overrightarrow{\varphi}_2 \circ \overrightarrow{\varphi}_1(\overrightarrow{x}),$$

en aritmética de punto flotante, errores en los datos de entrada y errores de redondeo en los resultados intermedios perturbarán los mismos y en consecuencia afectarán en el resultado final.

Sea $\overrightarrow{\varepsilon} \in \mathbb{R}^n$ y $\overrightarrow{x}_\varepsilon = \vec{x} + \vec{\varepsilon}$ el dato de entrada perturbado. Sea $\overrightarrow{y}_\varepsilon = \left(\overrightarrow{\varphi}_{r-1} \circ \varphi_{r-2} \circ \cdots \circ \varphi_2 \circ \overrightarrow{\varphi}_1\right)(\overrightarrow{x}_\varepsilon)$. Interesa comparar los resultados obtenidos $\overrightarrow{y} = \overrightarrow{\varphi}(\overrightarrow{x})$, y $\overrightarrow{y}_\varepsilon = \left(\overrightarrow{\varphi}_{r-1} \circ \varphi_{r-2} \circ \cdots \circ \varphi_2 \circ \overrightarrow{\varphi}_1\right)(\overrightarrow{x}_\varepsilon)$, es decir, como los errores de redondeo, de truncamiento afectan en el resultado final mediante la ejecución de la secuencia indicada $\overrightarrow{\varphi}_{r-1} \circ \varphi_{r-2} \circ \cdots \circ \varphi_2 \circ \overrightarrow{\varphi}_1$.

Definición 10 *De manera general, diremos que un algoritmo es estable numéricamente con respecto de otro si pequeñas variaciones en los datos de entrada producen pequeñas variaciones en los datos de salida. Un algoritmo será inestable si pequeñas variaciones en los datos de entrada producen grandes variaciones en los datos de salida.*

En Análisis Numérico, el estudio de la estabilidad numérica tiene mucha importancia, pués para construir un algoritmo, entre uno de los requerimientos a verificar es el de la estabilidad numérica. Si este requisito no es verificado no puede aceptarse al algoritmo como buen algoritmo y puede ser desechado.

Definición 11 *Sea V un espacio normado provisto de la norma $\| \cdot \|$. Supongamos que la solución S de un problema (P) propuesto en V se aproxima mediante un algoritmo que genera a S_n aproximación de S, $n = 1, 2, \ldots,$ en el sentido siguiente:*

$$\forall \varepsilon > 0,\ \exists n_0 \in \mathbb{Z}^+ \ \ tal\ que\ \ \forall n \geq n_0 \Longrightarrow \| S_n - S \| < \varepsilon,$$

en tal caso diremos que el algoritmo es convergente.

Dado un problema (P) y propuesto un algoritmo de solución, este debe ser bien condicionado y numéricamente estable. Si además el algoritmo genera una sucesión (S_n), debe verificarse que la sucesión (S_n) converge a S. Por lo tanto, la elaboración de un algoritmo implica el estudio del condicionamiento, estabilidad numérica y convergencia.

Notemos que un resultado importante del análisis numérico es el siguiente: si un algoritmo está bien condicionado y es numéricamente estable entonces el algoritmo es convergente. Usaremos la notación:

$$\text{condicionamiento} + \text{estabilidad} \Longrightarrow \text{convergencia}.$$

Cuando hay dos o más formas o métodos de construcción de sucesiones $(S_n^{(1)})$, $(S_n^{(2)})$ que convergen a S, es importante estudiar no solo la convergencia sino el orden de convergencia de cada método, con lo que se puede precisar las bondades y las limitaciones de cada uno de ellos.

Ejemplos

1. Sea $\{a_i \mid i = 1, \ldots, 10000\}$ un conjunto de números reales tales que $a_1 = 1, \ \ a_2 = \dfrac{1}{2}, \ldots, a_{10} = \dfrac{1}{10}$ y para $\ i = 11, \ldots, 10000, \quad \mid a_i \mid \simeq 6 \times 10^{-3}$. Sea $S = \sum_{i=1}^{10000} a_i^2$. Calcular S con una precisión de máquina $eps = 5 \times 10^{-4}$ y que se adapte a la estabilidad numérica.

   Consideremos dos algoritmos para la evaluación de S.

Primer algoritmo. Ponemos $S = 1$, y para $\begin{cases} i = 2, \ldots, 10000 \\ S = S + a_i^2. \end{cases}$ Como para $i = 11, \ldots, 10000$, $|a_i| \simeq 6 \times 10^{-3}$, entonces $a_i^2 \simeq 0{,}36 \times 10^{-4}$ siendo $eps = 5 \times 10^{-4}$ resulta que $a_i^2 \simeq 0$, con lo cual

$$\sum_{i=1}^{10000} a_i^2 = 1 + \frac{1}{2^2} + \ldots + \frac{1}{10^2} = 1{,}5498 = S_1.$$

Segundo algoritmo. Sea $|b_i| = 100\,|a_i| \simeq 6 \times 10^{-1}$, entonces $b_i^2 \simeq 0{,}36 \times 10^o > eps,\ i = 11, \ldots, 10000$. Ponemos

$$S_2 = 1 + \frac{1}{2^2} + \ldots + \frac{1}{10^2} + 10^{-4} \sum_{i=11}^{10000} (100a_i)^2.$$

Ahora bien,

$$\sum_{i=11}^{10000} (100a_i)^2 \simeq 10000 \times 0{,}36 \times 10^o = 0{,}36 \times 10^4.$$

Luego

$$S_2 = 1 + \frac{1}{2^2} + \ldots + \frac{1}{10^2} + 10^{-4} \sum_{i=11}^{10000} (100a_i)^2 \simeq 1{,}9098.$$

El primer algoritmo es numéricamente inestable, el segundo algoritmo es numéricamente estable. La razón de la inestabilidad numérica del primer algoritmo se encuentra en el cálculo de a_i^2 que está mal condicionado, pués

$$\varepsilon_{a_i^2} = \frac{a_i}{a_i^2} \times 2a_i \times \varepsilon_{a_i} = 2\varepsilon_{a_i},$$

donde ε_{a_i} es el error relativo de a_i. Así, el número de condicionamiento de cada a_i^2 es mayor que 1 lo cual amplifica el error relativo y lo vuelve inestable para a_i muy pequeño.

Determinemos el error relativo en porcentaje para cada algoritmo. Se tiene

$$S = \sum_{i=1}^{10000} a_i^2 \simeq 1{,}9094.$$

Para el primer algoritmo

$$|\varepsilon_1| = \left| \frac{S_1 - S}{S} \right| \times 100 = \frac{|1{,}5498 - 1{,}9094|}{1{,}9094} \times 100 = 18{,}8\,\%.$$

Para el segundo algoritmo

$$|\varepsilon_2| = \left| \frac{S_2 - S}{S} \right| \times 100 = \frac{|1{,}9098 - 1{,}9094|}{1{,}9094} = 0{,}00019\,\%.$$

Se observa claramente que el segundo algoritmo es mejor que el primero.

2. Se define la función real φ como $\varphi(x) = \dfrac{\operatorname{senh}(x)}{x} \quad x \neq 0$. Se desea calcular valores de $\varphi(x)$.

Primeramente, para $x \in \mathbb{R}$ tal que $|x| \geq \dfrac{1}{2}$ no se presenta ninguna dificultad en el cálculo de $\varphi(x)$. Obviamente,

$$\lim_{x \to -\infty} \frac{\operatorname{senh}(x)}{x} = \lim_{x \to \infty} \frac{\operatorname{senh}(x)}{x} = \infty.$$

Nos interesamos en calcular $\varphi(x)$ para $x \in \left]-\dfrac{1}{2}, 0\right[\cup \left]0, \dfrac{1}{2}\right[$. Con el uso de una calculadora de bolsillo, se tienen los siguientes resultados: para $x = 10^{-100}$, $\operatorname{senh}\left(10^{-100}\right) = 0$, luego

$$\varphi\left(10^{-100}\right) = \frac{\operatorname{senh}\left(10^{-100}\right)}{10^{-100}} = 0,$$

lo que es falso.

Para $x \in \left]0, \frac{1}{2}\right[$ suficientemente pequeño, podemos suponer $x < eps$, ¿cómo calcular $\varphi(x)$ si $\operatorname{senh}(x) \simeq 0$ y $x \simeq 0$? Sabemos que $\lim_{x\to\infty} \frac{\operatorname{senh}(x)}{x} = 1$. Para responder a la pregunta, recurrimos a la definición de la función seno hiperbólico y por el polinomio de Taylor con resto (véase el apéndice) se tiene para $x \in \mathbb{R}$

$$\operatorname{senh}(x) = \frac{1}{2}\left(e^x - e^{-x}\right) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{x^{2n+1}}{(2n+1)!} + E_{2n+1}(x),$$

con $E_m(x)$ el error de aproximación del polinomio de Taylor definido como

$$E_m(x) = \frac{1}{m!}\int_0^x (x-t)^m g^{(m+1)}(t)\,dt = \frac{g^{(m_1)}(c)}{(m+1)!}x^{m+1} \quad 0 \le c < x \le \frac{1}{2},$$

y $g(x) = \operatorname{senh}(x)$, $m \in \mathbb{Z}^+$. Como $g^{(m)}(x) = \begin{cases} \operatorname{senh}(x), & \text{si } m \text{ es par,} \\ \cosh(x), & \text{si } m \text{ es impar,} \end{cases}$ entonces $\left|g^{(m+1)}(x)\right| \le \frac{1}{2}\left(e^{\frac{1}{2}} + e^{-\frac{1}{2}}\right)$ si $x \in \left[0, \frac{1}{2}\right]$ y en consecuencia

$$|E_m(x)| \le \left|\frac{g^{(m+1)}(c)}{(m+1)!}x^{m+1}\right| \le \frac{1}{2}\left(e^{\frac{1}{2}} + e^{-\frac{1}{2}}\right)\frac{x^{m+1}}{(m+1)!}.$$

Por lo tanto, para $x \in \left]0, \frac{1}{2}\right[$ se tiene

$$\varphi(x) = \frac{\operatorname{senh}(x)}{x} = 1 + \frac{x^2}{3!} + \frac{x^4}{5!} + \cdots + \frac{x^{2n}}{(2n+1)!} + R_{2n}(x) \quad x \in \left]0, \frac{1}{2}\right[,$$

y

$$|R_{2n}(x)| = \left|\frac{E_{2n}(x)}{x}\right| \le \frac{1}{2}\left(e^{\frac{1}{2}} + e^{-\frac{1}{2}}\right)\frac{x^{2n}}{(2n+1)!} \quad x \in \left]0, \frac{1}{2}\right[.$$

Se supone que en una calculadora de bolsillo $10^{-100} \simeq 0$ pero 10^{-99} no se redondea por cero, se tiene

$$\varphi\left(10^{-100}\right) = 1 + R_2\left(10^{-100}\right) \simeq 1 \quad \text{pués} \quad R_2\left(10^{-100}\right) \simeq 0,$$

que es un resultado mucho más apegado a la realidad, pués $\frac{\operatorname{senh}(x)}{x} \longrightarrow 1$ cuando $x \longrightarrow 0$. De la representación de la función φ como polinomio de Taylor con resto arriba indicada, para $x = 10^{-40}$, 10^{-20}, 10^{-10}, se obtienen los siguientes resultados:

$$\begin{aligned}
\varphi\left(10^{-40}\right) &\simeq 1 + \frac{1}{6}\times 10^{-80}, \quad R_2(10^{-40}) \simeq 0,\\
\varphi\left(10^{-20}\right) &\simeq 1 + \frac{1}{6}\times 10^{-40} + \frac{1}{120}\times 10^{-80}, \quad R_4(10^{-40}) \simeq 0,\\
\varphi\left(10^{-10}\right) &\simeq 1 + \frac{1}{6}\times 10^{-20} + \frac{1}{120}\times 10^{-40} + \frac{1}{5040}\times 10^{-60} + \frac{1}{362880}\times 10^{-80}, \quad R_8(10^{-40}) \simeq 0.
\end{aligned}$$

Para $x = 0{,}005$, veamos los siguientes resultados. De la definición de φ, tenemos

$$\varphi(0{,}005) = \frac{\operatorname{senh}(0{,}005)}{0{,}005} \simeq 1{,}000004165.$$

Si utilizamos la definición de seno hiperbólico, se tiene

$$\varphi(x) = \frac{\operatorname{senh}(x)}{x} = \frac{1}{2x}\left(e^x - e^{-x}\right) \quad x \neq 0,$$

y resulta que $\varphi\,(0{,}005) \simeq 1{,}00000418$. Si

$$\varphi\,(x) = 1 + \frac{x^2}{3!} + \frac{x^4}{5!} + R_5\,(x) = 1 + \frac{x^2}{6}\left(1 + \frac{x^2}{20}\right) + R_4\,(x)\,,$$

se tiene $\varphi\,(0{,}005) \simeq 1{,}000004167$, y si

$$\varphi\,(x) = 1 + \frac{x^2}{3!} + \frac{x^4}{5!} + \frac{x^6}{7!} + R_6\,(x) = 1 + \frac{x^2}{6}\left(1 + \frac{x^2}{20}\left(1 + \frac{x^2}{42}\right)\right) + R_6\,(x)\,,$$

se tiene $\varphi\,(0{,}005) \simeq 1{,}000004167$. Con una presición del orden de 10^{-10}, de estos resultados, la mejor aproximación es $\varphi\,(0{,}005) = 1{,}000004167$. Pués

$$|R_6\,(x)| \leq \frac{\left(5 \times 10^{-3}\right)^6}{7!} \simeq 0{,}31002 \times 10^{-17} < 10^{-10}.$$

Resultados similares se obtienen en el caso $x \in \left]-\frac{1}{2}, 0\right[$.

Para $x \in \left]-\frac{1}{2}, 0\right[\cup \left]0, \frac{1}{2}\right[$, se dice que el cálculo de $\varphi\,(x)$ mediante el polinomio de Taylor con error se adapta a la estabilidad numérica y por lo tanto $\varphi\,(x)$ es numéricamente estable frente a las formas de cálculo de $\varphi\,(x)$ como $\varphi\,(x) = \dfrac{\operatorname{senh}(x)}{x}$, o de $\varphi\,(x) = \dfrac{1}{2x}\left(e^x - e^{-x}\right) \quad x \neq 0$.

3. Sean $x_0, x_1, \ldots, x_n$ números de máquina positivos, cuyo número de máquina es ε. Entonces el error de redondeo relativo al calcular $\sum_{k=0}^{n} x_k$ de la manera usual es $(1+\varepsilon)^n - 1 \simeq n\,\varepsilon$.

Sea $S_n = \displaystyle\sum_{i=0}^{n} x_i$ y $\widetilde{S}_n$ el resultado de la suma en el computador. Se tiene $\begin{cases} S_o = x_o \\ S_{k+1} = S_k + x_{k+1}, \end{cases}$ y $\begin{cases} \widetilde{S}_o = x_o \\ \widetilde{S}_{k+1} = fl\left(\widetilde{S}_k + x_{k+1}\right). \end{cases}$ Definimos

$$\varepsilon_{s_k} = \frac{\widetilde{S}_k - S_k}{S_k} \quad \text{y} \quad \varepsilon_k = \frac{\widetilde{S}_{k+1} - (\widetilde{S}_k + x_{k+1})}{\widetilde{S}_k + x_{k+1}}.$$

Entonces

$$\begin{aligned} \varepsilon_{S_{k+1}} &= \frac{\widetilde{S}_{k+1} - S_{k+1}}{S_{k+1}} = \frac{(\widetilde{S}_k + x_{k+1})(1+\varepsilon_k) - (S_k + x_{k+1})}{S_{k+1}} \\ &= \frac{(S_k(1+\varepsilon_{S_k}) + (x_{k+1}))(1+\varepsilon_k) - (S_k + x_{k+1})}{S_{k+1}} \\ &= \varepsilon_k + \varepsilon_{S_k}\left(\frac{S_k}{S_{k+1}}\right)(1+\varepsilon_k). \end{aligned}$$

Puesto que $S_k < S_{k+1}$ y $|\varepsilon_k| \leq \varepsilon$, resulta

$$|\varepsilon_{S_{k+1}}| \leq \varepsilon + |\varepsilon_{S_k}|\,(1+\varepsilon) = \varepsilon + \sigma\,|\varepsilon_{S_k}|\,,$$

donde $\sigma = 1 + \varepsilon$. Se tiene

$$\begin{aligned} |\varepsilon_{So}| &= 0, \\ |\varepsilon_{S_1}| &\leq \varepsilon, \\ |\varepsilon_{s_2}| &\leq \varepsilon + \sigma\varepsilon = \varepsilon(1+\sigma), \\ \left|\varepsilon_{S_3}\right| &\leq \varepsilon + \sigma(\varepsilon + \sigma\varepsilon) = \varepsilon(1 + \sigma + \sigma^2), \\ &\vdots \\ |\varepsilon_{Sn}| &\leq \varepsilon + \sigma\varepsilon + \sigma^2\varepsilon + \ldots + \sigma^{n-1}\varepsilon = \varepsilon(1 + \sigma + \sigma^2 + \ldots + \sigma^{n-1}) \\ &= \varepsilon\frac{\sigma^n - 1}{\sigma - 1} = \varepsilon\frac{(1+\varepsilon)^n - 1}{\varepsilon} = (1+\varepsilon)^n - 1. \end{aligned}$$

Por el binomio de Newton.

$$(1+\varepsilon)^n - 1 = 1 + n\varepsilon + \frac{n(n+1)}{2!}\varepsilon^2 + \ldots + \varepsilon^n - 1 \simeq n\varepsilon.$$

4. Sea $S = \sum_{k=1}^{\infty} a_k$ una serie real convergente. Intentamos aproximar S siguiendo 2 etapas. Primero calculamos la suma parcial $S_n = \sum_{k=1}^{n} a_n$ para n grande y a continuación redondeamos S_n reteniendo una cierta cantidad de dígitos después del punto decimal. Digamos que se han retenido m dígitos. ¿Se puede asegurar que el último dígito es el correcto?.

   Sea $\widetilde{S}_n = rd(S_n)$. Deseamos que $\left|\widetilde{S}_n - S\right| \leq \frac{1}{2} \times 10^{-m}$. Si S_n se ha redondeado correctamente para obtener $\widetilde{S}_n$, entonces $\left|\widetilde{S} - S_n\right| \leq \dfrac{1}{2} \times 10^{-m}$. Pero

   $$\left|\widetilde{S} - S\right| \leq \left|\widetilde{S} - S_n\right| + |S_n - S| \leq \frac{1}{2} \times 10^{-m} + |S_n - S|\,.$$

   Esta desigualdad no se puede mejorar a menos que $S_n = S$, o también $a_k = a$, $k = 1, \ldots, n$, por tanto no se puede lograr $\left|\widetilde{S} - S\right| \leq \dfrac{1}{2} \times 10^{-m}$. Si imponemos que $\left|\widetilde{S} - S\right| < \dfrac{6}{10} \times 10^{-m}$, entonces

   $$\frac{1}{2} \times 10^{-m} + |S_n - S| < \frac{6}{10} \times 10^{-m}.$$

   de donde $|S_n - S| < 10^{-m-1}$. Luego $\left|\sum_{k=n+1}^{\infty} a_k\right| < 10^{-m-1}$.

5. Sea $E(n) = \int_0^1 x^n e^{x-1} dx,\ n = 0, 1, 2, \ldots$.Para cada n, se desea calcular valores aproximados de $E(n)$. Con este propósito se deben elaborar algoritmos para aproximar $E(n)$. Con este ejemplo se obtendrán un algoritmo mal condicionado y otro bien condicionado, numéricamente estable y convergente.

   Primeramente analicemos el problema.

   Sea $f_n(x) = x^n e^{x-1} \quad x \in [0,1], \quad n = 0, 1, \cdots$. Se tiene que $f_n(x) \geq 0 \quad \forall x \in [0,1], \quad \forall n = 0, 1, 2, \ldots,$ y como la integral de una función no negativa es no negativa, se sigue que $E(n) = \int_0^1 f_n(x)dx \geq 0$. Por otra parte, si $0 < x < 1, \quad x^n \underset{n\to\infty}{\longrightarrow} 0$, luego $f_n(x) \underset{n\to\infty}{\longrightarrow} 0 \quad \forall x \in [0,1[$, entonces

   $$\lim_{n\to\infty} E(n) = \lim_{n\to\infty} \int_0^1 x^n e^{x-1} = 0.$$

   En conclusión $E(n)$ es no negativo, $E(n+1) < E(n) \quad$ con $n = 0, 1, \ldots$ que muestra que $E(n)$ es decreciente y acotada por 0.

   Primer algoritmo. Para $n = 0$, se tiene

   $$E(0) = \int_0^1 e^{x-1} dx = e^{x-1}\,|_0^1 = 1 - e^{-1} = 0{,}6321205588\ldots.$$

   Para $n > 0$, aplicamos el método de integración por partes, se obtiene

   $$E(n) = x^n e^{x-1}\,|_0^1 - n\int_0^1 x^{n-1} e^{x-1} dx = 1 - n\ E(n-1), \quad n = 1, 2, \ldots,$$

   Así, se obtiene la ecuación recurrente siguiente: $E(n) = 1 - n\ E(n-1) \quad n = 1, 2, \cdots$ .Conocido un valor aproximado de $E(n-1)$, mediante la ecuación recurrente podemos calcular un valor

aproximado de $E(n)$ lo que nos permite obtener el siguiente algoritmo de cálculo para $E(n)$ para $n = 0, 1, \cdots, N$.

Algoritmo

Datos de entrada: N.

Datos de salida: $E(n)$.

1. $E(0) = 1 - e^{-1} = 0{,}6321205588\ldots$,

2. Para $n = 1, \ldots, N$

$$E(n) = 1 - nE(n-1)$$

Fin de bucle n.

3. Imprimir $E(n)$.

4. Fin.

Con precisión de 5 cifras decimales, los resultados de la aplicación del algoritmo precedente se muestran en la siguiente tabla:

	n	$E(n)$
	0	0,63212
	1	0,36788
	2	0,26424
	3	0,20728
	$\vdots$	
	12	0,05809
$*$	13	0,24478
$*$	14	$-2{,}42688$
$*$	15	37,40316
	$\vdots$	
$*$	20	$-69'478,033{,}14.$

Observe los valores señalados con $*$. El análisis de $E(n)$ muestra que $E(n) \geq 0$ y $E(n) \longrightarrow 0$ cuando $n \to \infty$. A partir de $n = 13$ los resultados son absurdos. Si se realizan los cálculos con un número mayor de cifras decimales, los resultados absurdos se obtienen para n>13.

Note que tomando $\widetilde{E} = rd(E(0)) = 0{,}63212$ como dato de entrada, el error de redondeo en cada iteración es amplificado por $-n$. Estos hechos demuestran que el algoritmo está mal condicionado. Se puede demostrar que el número de condicionamiento de este procedimiento es $C = -n$, y en consecuencia pequeños errores en los datos de entrada provocan grandes errores en los datos de salida, lo que muestra que el algoritmo es inestable numéricamente.

Segundo Algoritmo. Tomando en cuenta que $E(n) > 0, \quad n = 0, 1, \ldots,$ y, $E(n) \longrightarrow 0$ cuando $n \to \infty$, basta elegir n suficientemente grande para obtener $E(n+1) \simeq 0$, entonces

$$0 \simeq E(n+1) = 1 - (n+1)E(n),$$

de donde

$$E(n) = \frac{1}{n+1},$$

y para $n-1, n-2, \ldots, 1$, tenemos

$$E(n-1) = \frac{1 - E(n)}{n}.$$

Para N suficientemente grande, se establece el algoritmo siguiente.

Algoritmo

Datos de entrada: N.

Datos de salida: $E(n)$.

1. $E(N) = \dfrac{1}{N+1}$.

2. Para $n = N-1, \cdots, 1$

$$E(n) = \frac{1 - E(n)}{n}$$

Fin de bucle n.

3. Imprimir $E(n)$.

4. Fin.

Para $N = 20$, en la tabla siguiente se muestran los resultados de la aplicación del algoritmo precedente.

n	$E(n)$
20	0,04762
19	0,04762
18	0,05130
17	0,05277
$\vdots$	
5	0,14553
4	0,17089
3	0,20728
2	0,26424
1	0,36788.

Estos resultados son satisfactorios. Este algoritmo está bien condicionado y los pequños erreres en los datos de entrada provocan pequeños errores en los datos de salida, es decir que el algoritmo es numéricamente estable, no obstante el algoritmo presenta un inconveniente: el número de operaciones que se requiere para calcular $E(n_j)$ a partir de $E(N)$ con una precisión fijada debe ser grande.

Tercer algoritmo. Por la serie de Taylor de e^x se tiene $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad x \in [0,1]$, y por el teorema de la convergencia uniforme y la integración (véase el capítulo 3 donde se tratan las sucesiones y series de funciones), resulta

$$\begin{aligned} E(n) &= \int_0^1 x^n e^{x-1} dx = e^{-1} \int_0^1 x^n e^x dx = e^{-1} \int_0^1 x^n \left(\sum_{k=0}^{\infty} \frac{x^k}{k!} \right) dx \\ &= e^{-1} \sum_{k=0}^{\infty} \frac{1}{k!} \int_0^1 x^{n+k} dx = e^{-1} \sum_{k=0}^{\infty} \frac{1}{k!(n+k+1)}. \end{aligned}$$

Así, $E(n) = e^{-1} \sum_{k=0}^{\infty} \dfrac{1}{k!(n+k+1)} \qquad n = 0, 1, \cdots$. Vemos que $E(n)$ se representa como una serie numérica convergente. Lamentablemente la serie no puede ser evaluada en el computador, necesitamos transformarla en una suma finita $S_{m_0}(n)$. Para el efecto, como la serie $\displaystyle\sum_{k=0}^{\infty} \frac{1}{k!(n+k+1)}$ es convergente, entonces

$$\forall \varepsilon > 0, \ \exists m_0 \in \mathbb{Z}^+ \text{ tal que } \forall m \geq m_0 \Rightarrow \left| \sum_{k=0}^{\infty} \frac{1}{k!(n+k+1)} - \sum_{k=0}^{m} \frac{1}{k!(n+k+1)} \right| < \varepsilon,$$

o bien $\displaystyle\sum_{k=m_0+1}^{\infty} \frac{1}{k!(n+k+1)} < \varepsilon.$

Sea (a_k) una sucesión de números positivos tal que $\sum_{k=1}^{\infty} a_k = 1$. Determinemos m_0 tal que

$$\frac{\frac{1}{k!(n+k+1)}}{a_k} < \varepsilon \quad \text{si} \quad k \geq m_0.$$

Pongamos $a_k = \dfrac{1}{k(k+1)}$. Se tiene $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1$. Entonces $\frac{k(k+1)}{k!(n+k+1)} \leq \varepsilon$. Sea $\varepsilon = 10^{-6}$, determinemos m_0 tal que $\frac{1}{(k-1)!} < 10^{-6}$. Esta última desigualdad se verifica para todo $k \geq 11$, luego $m_0 = 11$. Así

$$S_{m_0}(n) = e^{-1} \sum_{k=0}^{11} \frac{1}{k!(n+k+1)}.$$

De este modo $E(n)$ es aproximado con la suma $S_{m_0}(n)$ con una precisión de 10^{-6}. Cada término de la suma está bien condicionada, escribimos $S_{m_o}(n)$ de la manera siguiente:

$$\begin{aligned} S_{mo}(n) &= e^{-1}\left(\frac{1}{n+1} + \frac{1}{1!(n+2)} + \frac{1}{2!(n+3)} + \frac{1}{3!(n+4)} + \ldots + \frac{1}{10!(n+11)} + + \frac{1}{11!(n+12)}\right) \\ &= e^{-1}\left(\frac{1}{n+1} + \frac{1}{n+2} + \frac{1}{2}\left(\frac{1}{n+3} + \frac{1}{3}\left(\frac{1}{n+4} + \ldots + \frac{1}{10}\left(\frac{1}{n+11} + \right.\right.\right.\right. \\ &\quad \left.\left.\left.\left. + \frac{1}{11 \times (n+12)}\right)\ldots\right)\right)\right). \end{aligned}$$

Tenemos así el siguiente algoritmo siguiente:

Algoritmo

Datos de entrada: n.

Datos de salida: $E(n)$.

1. $s = \dfrac{1}{11 \times (n+12)}$
2. Para $k = 1, \cdots, 12$

   $j = 12 - k$

   $s = \dfrac{1}{n+j} + \dfrac{1}{j}s$

   Fin de bucle k.
3. $E(n) = s$.

4-Imprimir E(n).

5. Fin.

En la tabla siguiente se muestran los resultados de la aplicación de este algoritmo.

n	$E(n) \simeq S_{mo}(n)$
1	0,367879
2	0,264241
3	0,207276
4	0,170893
$\vdots$	
100	0,009805
$\vdots$	
500	0,001992
$\vdots$	
1000	0,000998
$\vdots$	
1000000	0,000009999.

Para cada n, $S_{m_0}(n)$ requiere de 59 operaciones elementales. Este algoritmo reune todas las características: condicionamiento, estabilidad, convergencia. Además es fácil de programar y cada $S_{m_o}(n)$ es independiente del cálculo de $S_{m_o}(n-1)$ o de $S_{mo}(n+1)$. En consecuencia este es uno de los mejores algoritmos que puede construirse para aproximar $E(n)$, $\quad n = 1, 2, \ldots$.

6. Ejemplo de un método convergente.

   Consideremos como problema (P) el cálculo de $a^{1/n}$, donde $a \in \mathbb{R}$, $\quad n \in \mathbb{N}$ con $n \geq 2$.

   Notemos primeramente que la raíz $n-$ésima de a está bien definida para todo a si n es impar, y $a \geq 0$ si n es par.

   Supongamos $a \geq 0$, $\quad n \geq 2$. Definimos la función f de $\mathbb{R}^+$ en $\mathbb{R}$, por $\quad f(x) = x^n - a \quad x \in \mathbb{R}^+$. Tenemos la siguiente equivalencia: $f(x) = 0 \Leftrightarrow x = a^{1/n}$, es decir que la ecuación $f(x) = 0$ tiene una única solución $x = a^{1/n} \in \mathbb{R}^+$. Apliquemos el método de Newton cuya interpretación geométrica indicamos a continuación.

   Sea $x_0 \in \mathbb{R}^+$ una aproximación de $a^{1/n}$. La ecuación de la recta tangente L_1 a la gráfica de f en el punto $(x_0,\ f(x_0))$ viene dada por: $\quad y - f(x_0) = f'(x_0)(x - x_0)$. Esta recta corta al eje X en el punto $(x_1, 0)$, en tal caso tenemos $y = 0$, y

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

   Repetimos nuevamente el proceso descrito previamente. La ecuación de la recta tangente L_2 a la gráfica de f en el punto $(x_1,\ f(x_1))$ es: $\ y - f(x_1) = f'(x_1)(x - x_1)$, que corta al eje X en el punto $(x_2, 0)$. Obtenemos entonces

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

   Continuando con este procedimiento k veces, deducimos que

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad k = 0, 1, 2, \ldots, .f'(x_k) \neq 0.$$

   Este último esquema numérico se conoce con el nombre de método de Newton. En el capítulo resolución numérica de ecuaciones no lineales se aborda con más detalle este método.

   En la siguiente figura se ilustran las rectas tangentes L_1, L_2, L_3 a la gráfica de f en los puntos $(x_0,\ f(x_0))$, $(x_1,\ f(x_1))$ y $(x_2,\ f(x_2))$. Note que la gráfica de la función f corta al eje X en $x = a^{1/n}$.

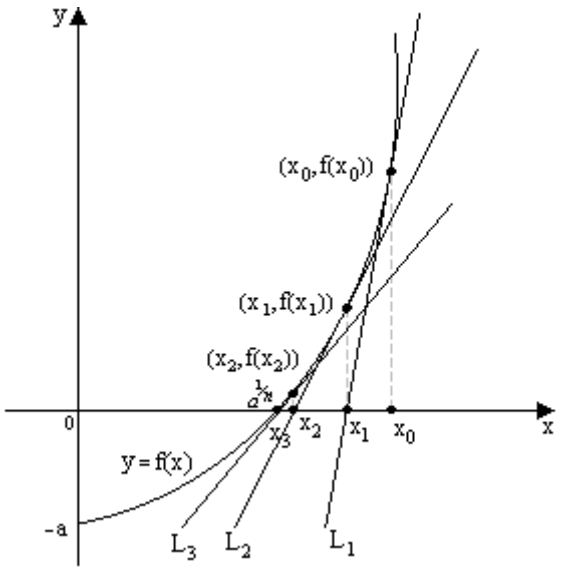


Figura 10

De acuerdo a la definición de la función f y al esquema de Newton arriba obtenido, tenemos

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^n - a}{n\, x_k^{n-1}} = \frac{(n-1)x_k^n + a}{n\, x_k^{n-1}} = \frac{1}{n}\left((n-1)\, x_k + \frac{a}{x_k^{n-1}}\right), \quad k = 0, 1, \ldots,$$

al que nos referiremos como esquema numérico para aproximar $a^{1/n}$.

Si fijamos el número entero m, número de iteraciones a realizar, el procedimiento de cálculo de $a^{1/n}$ es el siguiente:

$$\begin{cases} x_0 > 0 \ \text{ dado}, \\ x_{k+1} = \dfrac{1}{n}\left((n-1)x_k + \dfrac{a}{x_k^{n-1}}\right) \qquad k = 0, 1, \ldots, m. \end{cases}$$

Este algoritmo requiere de la lectura de los siguientes datos de entrada: n, a, m. Como resultado obtendremos el valor aproximado de $a^{1/n}$. Los datos de salida son : a, n, $a^{1/n}$. Se puede probar que dada una aproximación inicial $x_0 > 0$ apropiada de $a^{1/n}$, la sucesión (x_k) generada por el esquema numérico para aproximar $a^{1/n}$, converge efectivamente al valor de $a^{1/n}$. En estas condiciones proponemos un primer algoritmo que es un tanto incompleto como se verá más adelante.

Algoritmo 1.

Datos de entrada: a, n, m.

Datos de salida: $a^{1/n}$.

1. $x = x_0$.
2. Para $k = 0, 1, \ldots, m$

$$x_k = \frac{1}{n}\left((n-1)x + \frac{a}{x^{k-1}}\right)$$
$$x = x_k$$

Fin de bucle k.

3. Imprimir $x = a^{1/n}$.
4. Fin.

Así por ejemplo si $a = 2, \quad n = 2$, el procedimiento para calcular $\sqrt{2}$ es el siguiente:

$$\begin{cases} x_0 > 0 \quad \text{dado}, \\ x_{k+1} = \dfrac{1}{2}\left(x_k + \dfrac{2}{x_k}\right) \qquad k = 0, 1, \ldots \end{cases}$$

Si fijamos el número de iteraciones $m = 5$ y el punto inicial $x_0 = 2$, la aplicación del algoritmo de cálculo de $\sqrt{2}$ nos da los siguientes resultados:

$$\begin{aligned} x_1 &= \frac{1}{2}\left(x_0 + \frac{2}{x_0}\right) = 1{,}5, \qquad x_2 = \frac{1}{2}\left(x_1 + \frac{2}{x_1}\right) = 1{,}4166667, \\ x_3 &= \frac{1}{2}\left(x_2 + \frac{2}{x_2}\right) = 1{,}414215687, \qquad x_4 = \frac{1}{2}\left(x_3 + \frac{2}{x_3}\right) = 1{,}414213563, \\ x_5 &= \frac{1}{2}\left(x_4 + \frac{2}{x_4}\right) = 1{,}414213563. \end{aligned}$$

El valor exacto es $\sqrt{2}$ y el obtenido en una calculadora de bolsillo es 1,414213562 . . . . En este ejemplo vemos las siguientes características: el procedimiento de cálculo descrito en el algoritmo 1 tiene una estructura bien definida, el número de repeticiones de concluye en m pasos, el procedimiento permite aproximar $\sqrt{2}$ con una precisión de 10^{-10}. El algoritmo 1 presenta un inconveniente que es la selección del punto inicial x_0 del que se ha dicho debe ser una aproximación inicial apropiada de $a^{1/n}$. No obstante, para a suficiente grande queda la duda de como elegir dicho punto. Un procedimiento de selección es el siguiente:

a) Para $0 < a < 1$, obtenemos el siguiente esquema numérico:

$$\begin{cases} x_0 = 1 \\ x_{k+1} = \dfrac{1}{n}\left((n-1)x_k + \dfrac{a}{x_k^{n-1}}\right) \end{cases} \quad k = 0, 1, \ldots, m.$$

b) Supongamos que $a > 1$. Sea $j \in \mathbb{N}$ el más pequeño entero tal que $0 < 10^{-jn}a < 1$. Entonces

$$f(x) = x^n - a = 10^{jn}\left(\frac{x^n}{10^{jn}} - \frac{a}{10^{jn}}\right) = 10^{jn}\left(\left(\frac{x}{10^j}\right)^n - \frac{a}{10^{jn}}\right).$$

Ponemos $b = a \times 10^{-jn}$, $t = 10^{-j} \times x$ y definimos $g(t) = t^n - b$. Resulta

$$g(t) = 0 \Longleftrightarrow t = b^{1/n} \Longleftrightarrow 10^{-j}\, x = 10^{-j}a^{1/n} \Longleftrightarrow x = a^{1/n}.$$

Así, $g(t) = 0 \Leftrightarrow x = a^{1/n}$, que muestra que la raíz de la ecuación $g(t) = 0$ es la misma de $f(x) = 0$.

El algoritmo descrito precedentemente puede ser aplicado a condición de reemplazar a por b. Así por ejemplo, sea $a = 36254932{,}65$ y $n = 4$. Resulta que $a = 0{,}3625493265 \times 10^8$, $j = 2$ y $b = 0{,}3625493265$. Entonces

$$\begin{cases} t_0 = 1, \\ t_{k+1} = \dfrac{1}{4}\left(3t_k + \dfrac{0{,}3625493265}{t_k^3}\right) \end{cases} \quad k = 0, 1, \ldots, 5.$$

Tenemos

$$\begin{aligned} t_1 &= 0{,}8406373318, \qquad t_2 = 0{,}783052196, \qquad t_3 = 0{,}7760600163, \\ t_4 &= 0{,}7759643795, \qquad t_5 = 0{,}775964362. \end{aligned}$$

Luego $x = 0{,}775964362 \times 10^2$, con lo cual $a^{1/4}$ se aproxima por 77,5964362 con una precisión $\varepsilon = 10^{-8}$. El valor obtenido de una calculadora de bolsillo es $(36'254,932{,}65)^{1/4} = 77{,}59643619\ldots$.

c) Finalmente, si $a < 0$ y n es impar ponemos $c = -a$ y aplicamos los resultados descritos precedentemente en *a*) y *b*).

Para elaborar un algoritmo completo de cálculo de $a^{1/n}$ introducimos dos variables *indi* e $info$ y que toman los valores 0 y 1. La variable *indi* lo utilizamos para la paridad de n, esto es, n par entonces $indi = 0$, n impar entonces $indi = 1$. La variable $info$ es utilizada para el signo de a, así: $a > 0$ entonces ínfo= 0, $a < 0$ entonces $info = 1$.

Algoritmo 2

Datos de entrada: a, n, m.

Datos de salida: $S = a^{1/n}$.

1. $\begin{cases} \text{Si } n \text{ par, hacer } indi = 0, \\ \text{Si } n \text{ impar, hacer } indi = 1. \end{cases}$
2. $\begin{cases} \text{Si } a > 0, \text{ hacer } info = 0, \\ \text{Si } a < 0, \text{ hacer } info = 1. \end{cases}$
3. Si $indi = 0$ e $info = 1$, Imprimir, "Error". Continuar en 11.
4. Si $indi = 1$ e $info = 1$. Hacer $c = a$. Continuar en 6.
5. Poner $c = a$.
6. Determinar $j \in \mathbb{N}$ el más pequeño tal que $b = 10^{-jn}c < 1$.
7. Poner $x = 1$.
8. Para $k = 1, \ldots, m$
   $x_k = \dfrac{1}{n}\left((n-1)x + \dfrac{b}{x^{n-1}}\right)$
   $x = x_k$.
   Fin bucle k
9. Si $indi = 1$ e $info = 1$. Hacer $x = -x_k$. Poner $S = 10^j x$.
10. Imprimir resultados: S.
11. Fin.

7. Ejemplo de un método numérico impracticable.

   Sean $a_1, a_2, b_1, b_2, c_1, c_2 \in \mathbb{R}$. Supongamos que $a_1,\ a_2,\ b_1,\ b_2$ son no nulos y la matriz $A = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix}$ es invertible. Consideramos el sistema de ecuaciones lineales: $(x, y) \in \mathbb{R}^2$ tal que $\begin{cases} a_1x + b_1y = c_1, \\ a_2x + b_2y = c_2. \end{cases}$ Puesto que A es invertible, este sistema de ecuaciones tiene solución única $(x, y) \in \mathbb{R}^2$. Calculemos esta solución. Para el efecto se disponen de dos métodos. El primero es el conocido método de Cramer cuya solución se calculan como se muestra a continuación

$$x = \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}, \qquad y = \frac{\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}},$$

   donde $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$ denota el determinante de la matriz real $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$. El segundo método que consideramos es el de eliminación gaussiana que se indica a continuación: se calcula $k = -\dfrac{a_2}{a_1}$, y se obtiene el sistema de ecuaciones lineales triangular superior siguiente: $\begin{cases} a_1x + b_1y = c_1 \\ (b_2 + kb_1)\, y = c_2 + kc_1, \end{cases}$ cuya solución se calcula como sigue: $\begin{cases} y = \dfrac{c_2 + kc_1}{b_2 + kb_1}, \\ x = \dfrac{1}{a_1}\,(c_1 - b_1y)\,. \end{cases}$ Contabilicemos el número de operaciones que se realizan con cada método. Con el método de Cramer, el cálculo del determinante $\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix} = c_1b_2 - c_2b_1$ implica tres operaciones elementales. Como se deben calcular tres determinantes y dos cocientes, resultan 11 operaciones elementales. Con el método de eliminación gaussiana se tienen las siguientes operaciones elementales: en el cálculo de k se realiza un cociente, el cálculo de y implica 5 operaciones elementales y el de x implica 3 operaciones elementales. En total se requieren de 9 operaciones elementales.

   A juzgar por el número de operaciones elementales, el método de eliminación gaussiana realiza 2 operaciones elementales menos que en el de Cramer. A más de esta razón, el método de eliminación gaussiana es mucho más estable numéricamente. En conclusión, para resolver numéricamente un sistema lineal de dos ecuaciones con dos incógnitas, se debe aplicar el método de eliminación gaussiana.

En lo sucesivo consideraremos sistemas de ecuaciones lineales $A\widehat{x} = \widehat{b}$, donde $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ con $A \neq 0$ y $\overrightarrow{b} \in \mathbb{R}^n$.

Llamamos método directo de resolución del sistema de ecuaciones lineales, un método que conduce a la solución del problema al cabo de un número finito de pasos, o bien en un número finito de operaciones aritméticas (suma, resta, multiplicación y división) que es función de la dimensión del sistema. Para cada método directo estudiado, se debe estimar:

i) El número de operaciones elementales necesarias en la ejecución del algoritmo, es decir que se debe determinar una función N_{oper}: $\mathbb{Z}^+ \to \mathbb{R}$ que a cada $n \in \mathbb{Z}^+$ asocie $N_{oper}\ (n)$.

ii) Precisión del método. Esta precisión depende sobre todo del condicionamiento de la matriz y de la estabilidad del método, es decir que pequeños errores en los datos de entrada provocan pequeños errores en los datos de salida, o lo que es lo mismo, es insensible a la propagación de errores de redondeo.

Supongamos que para resolver el sistema de ecuaciones lineales utilizamos la regla de Cramer:

$$x_i = \frac{\Delta_i}{\det(A)} \qquad i = 1, \ldots, n,$$

donde Δ_i es el determinante de la matriz obtenida al reemplazar la columna i-ésima de A por $\overrightarrow{b}$, $\det(A) \neq 0$.

Estimemos el número de operaciones elementales que se requieren para el cálculo de un determinante de una matriz C de $n \times n$. Para el efecto, determinemos el número de operaciones elementales que se requieren para calcular determinantes de orden 2 y 3.

Para calcular el determinante de una matriz C de 2×2 se efectúan las siguientes operaciones:

$$\text{productos}: \ 2 = 2! \times 1, \qquad \text{adiciones}: \ 1 = 2! - 1,$$

luego $N_{oper}(2) = 3$ operaciones elementales. Si C es una matriz real de 3×3, $C = (C_{ij})_{3x3}$, entonces

$$\det(C) = C_{11} \begin{vmatrix} C_{22} & C_{23} \\ c_{32} & C_{33} \end{vmatrix} - C_{12} \begin{vmatrix} C_{21} & C_{23} \\ C_{31} & C_{32} \end{vmatrix} + C_{13} \begin{vmatrix} C_{21} & C_{22} \\ C_{31} & C_{32} \end{vmatrix},$$

y el número de operaciones elementales se obtiene del modo siguiente: el cálculo de cada determinante de 2×2 requiere de 3 operaciones elementales, de la descomposición precedente, se obtiene

$$\text{multiplicaciones}: 9 = 3 \times 2 + 3, \qquad \text{sumas}: \ 5 = 3 \times 1 + 2,$$

con lo que $N_{oper}(3) = 14$ operaciones elementales.

En general, si C es una matriz de $n \times n$, se tiene

$$\text{multiplicaciones}: \sum_{j=1}^{n-1} \prod_{k=1}^{j} (n + 1 - k), \qquad \text{sumas}: \ n! - 1.$$

El número de operaciones elementales aplicando el método de Cramer es:

$$\begin{aligned} \text{multiplicaciones} &: (n+1) \sum_{j=1}^{n-1} \prod_{k=1}^{j} (n + 1 - k), \\ \text{sumas} &: (n+1)(n! - 1), \\ \text{divisiones} &: n, \end{aligned}$$

con lo cual

$$Noper(n) = n + (n+1)(n! - 1) + (n+1) \sum_{j=1}^{n-1} \prod_{k=1}^{j} (n + 1 - k) = (n+1)! + (n+1) \sum_{j=1}^{n-1} \prod_{k=1}^{j} (n + 1 - k).$$

Así, $Noper(5) = 330$ operaciones elementales, $Noper(6)$=1961 operaciones elementales. Note el tiempo que se requeriría para resolver un sistema de ecuaciones lineales de 5×5 usando una calculadora de bosillo: aproximadamente medio minuto por operación implica aproximadamente 165 minutos el tiempo requerido para resolver dicho sistema de ecuaciones, ¿cuánto tarda usted en resolver un tal sistema? Si despreciamos los $n - 2$ términos del sumatorio, tenemos $N = 2(n + 1)!$ y para $n = 20$, se obtiene $N \simeq 1{,}021818893 \times 10^{20} < Noper(20)$ que muestra que este método es impracticable. Con otros métodos, un sistema de ecuaciones lineales de 5×5 y con el uso de una calculadora de bolsillo y con el tiempo estimado de medio minuto por operación, se requerirá aproximadamente una hora; un sistema de ecuaciones lineales de 20×20 y con el uso de los computadores actuales requerirá de fracciones de segundo.

Por otra parte, las operaciones sumas y restas alternadas incrementan los errores de redondeo, que a su vez deterioran la calidad de la solución. Más aún, cuando n es demasiado grande, a causa de los errores de redondeo, puede provocarse un overflow lo que a su vez provocará una detención en la ejecución del programa. Por estas razones, el cálculo del determinante mediante este procedimiento definitivamente es impracticable, pués es mal condicionado e inestable numéricamente. Consecuentenemente, para el cálculo del determinante de una matriz debe aplicarse otros métodos y algoritmos que son relativamente económicos y fáciles de programarse e implementarse en un PC.

En conclusión, si se utiliza la regla de Cramer para hallar la solución del sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, del punto de vista numérico es impracticable.

Si A es una matriz invertible, la solución del sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$ tiene una única solución

$$\overrightarrow{x} = A^{-1}\overrightarrow{b},$$

donde $A^{-1} = \frac{1}{\det(A)}(A^D)^t$ y $A^D = [(-1)^{i+j} \text{ menor } (a_{ij})] \;\; i = 1, ..., n, \;\; j = 1, ..., n.$

El cálculo de la matriz A^D implica el cálculo de n^2 determinantes de matrices de $(n-1) \times (n-1)$. Adicionamos a esto el cálculo de $\det(A)$ y a continuación el producto de A^{-1}por $\overrightarrow{b}$. Mediante un razonamiento similar al precedente se puede mostrar que el número de operaciones elementales $Noper(n)$ es muy grande, con lo cual este método es igualmente impracticable. Más aún, si se toma en consideración los errores de redondeo, estos pueden ser muy grandes lo que conducirá a resultados completamente distorsionados. En definitiva, se trata de un método mal condicionado e inestable numéricamente, por lo tanto inutilizable del punto de vista numérico. Más adelante se tratan métodos directos de resolución de sistemas de ecuaciones que son fáciles de aplicarse con un número de operaciones $Noper(n)$ muy razonable.

1.12. Ejercicios

1. Sean T un triángulo cuyos vértices son $\overrightarrow{u}_1 = (x_1, y_1)$, $\overrightarrow{u}_2 = (x_2, y_2)$, $\overrightarrow{u}_3 = (x_3, y_3) \in \mathbb{R}^2$ que suponemos son no colineales y distintos, y, un punto dado $\overrightarrow{x} = (a, b) \in \mathbb{R}^2$. Se considera el siguiente problema: determinar si $(a, b) \in T$ o $(a, b) \notin T$. Fundamentar matemáticamente la solución del problema y elaborar un algoritmo numérico. Determine el número de operaciones elementales, asignaciones, comparciones. Realice comprobaciones de su algoritmo.

2. Se considera un triángulo T cuyos vértices son $\overrightarrow{u}_1 (x_1, y_1)$, $\overrightarrow{u}_2 = (x_2, y_2)$, $\overrightarrow{u}_3 = (x_3, y_3) \in \mathbb{R}^2$. Suponemos que $\overrightarrow{u}_1$, $\overrightarrow{u}_2$, $\overrightarrow{u}_3$ son distintos y no colineales. Elaborar un algoritmo que permita calcular su perímetro y su área. Recuerde que si $\overrightarrow{x}, \overrightarrow{y} \in \mathbb{R}^2$, la métrica ecuclídea d está definida como $d(\overrightarrow{x}, \overrightarrow{y}) = \|\overrightarrow{x} - \overrightarrow{y}\|$ y $\|\overrightarrow{x}\| = \left(\overrightarrow{x}^T \overrightarrow{x}\right)^{\frac{1}{2}}$. Determine el número de operaciones elementales. Realice comprobaciones de su algoritmo.

3. Sean $\overrightarrow{u}_1 = (x_1, y_1)$, $\overrightarrow{u}_2 = (x_2, y_2)$, $\overrightarrow{u}_2 = (x_3, y_3) \in \mathbb{R}^2$ los vértices de un triángulo T. Supongamos que $\overrightarrow{u}_1$, $\overrightarrow{u}_2$, $\overrightarrow{u}_3$ son distintos y no colineales. Elaborar un algoritmo que permita calcular los ángulos interiores del triángulo T y determinar si T es un triángulo rectángulo, isósceles o escaleno. Calcule el número de operaciones elementales y de comprobaciones.

4. Se consideran $\overrightarrow{u}_1 = (x_1, y_1)$, $\overrightarrow{u}_2 = (x_2, y_2)$, $\overrightarrow{u}_3 = (x_3, y_3)$, $\overrightarrow{u}_4 = (x_4, y_4)$ puntos de $\mathbb{R}^2$ dados. Suponemos que dichos puntos son distintos y al menos tres de ellos no son colineales. Elabore un algoritmo que permita identificar si el cuadrilátero es un paralelogramo y en este caso identificar si es un rectángulo. Además, se debe calcular el área de dicho cuadrilátero. Determine el número de operaciones elementales, asignaciones y comprobaciones. Realice pruebas para verificar su algoritmo.

5. Sean $a, b, c, d \in \mathbb{R}$ y $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ una matriz invertible. Obviamente $ad - bc \neq 0$ y $A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

   Ponemos $A^{-1} = \begin{bmatrix} p & r \\ q & s \end{bmatrix}$. Note que $\begin{bmatrix} p & r \\ q & s \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} p \\ q \end{bmatrix}$, $\begin{bmatrix} p & r \\ q & s \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} r \\ s \end{bmatrix}$, o sea $\begin{bmatrix} p \\ q \end{bmatrix}$ es la solución del sistema de ecuaciones $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ y $\begin{bmatrix} r \\ s \end{bmatrix}$ es la solución del sistema $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$de determinan las columnas de A^{-1}. Elabore un algoritmo que resuelva los dos sistemas de ecuaciones lineales de modo que el número de opreaciones elementales sea el más pequeño posible y escriba A^{-1}. Compruebe con las siguientes matrices:

a) $\begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}$. **b)** $\begin{bmatrix} 3 & -2 \\ 0 & 8 \end{bmatrix}$. **c)** $\begin{bmatrix} 1 & 2 \\ 5 & 2 \end{bmatrix}$. **d)** $\begin{bmatrix} 1 & \sqrt{2} \\ 5\sqrt{3} & 2\sqrt{5} \end{bmatrix}$.

6. Sean A, B, C matrices reales de 2×2. En cada item se define una matriz D, elabore un algoritmo para calcular la matriz D.

   a) $D = A\,(B + C)$. **b)** $D = AB - C$. **c)** $D = (A - B)\,C + I$ con I la matriz identidad.

   d) $D = C\,(B - A)\,C$. **e)** $D = B\left(I + A + A^2 + A^3 + A^4\right)C$. **f)** $D = B\left(I - A + A^2 - A^3 + A^4\right)C$.

   Compruebe cada algoritmo con las siguientes matrices:

   $A = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}$, $C = \begin{bmatrix} 3 & 5 \\ -2 & 1 \end{bmatrix}$.

7. **a)** Sea A una matriz real no nula de $m \times m$. Se define $A^{-1} = A$ y $A^{n+1} = A^n A$ para $n \in \mathbb{N}$. Elabore un algoritmo que permita calcular A^n.

   b) Verifique su algoritmo con $n = 3$ y la matriz siguiente $A = \begin{bmatrix} 1 & \frac{1}{2} \\ -\frac{1}{3} & 1 \end{bmatrix}$.

   c) Si $A = \begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ \frac{1}{2} & 2 & 0 \\ 1 & 1 & 3 \end{bmatrix}$, aplique su algoritmo y calcule A^3.

8. Aplique el método de eliminación gaussiana con pivoting parcial para hallar, si existe, la solución de cada uno de los sistemas de ecuaciones lineales que se proponen. En caso de calcular la solución, compruebe. De no ser posible, indique si el sistema de ecuaciones tiene infinitas soluciones o ninguna solución.

   a) $\begin{cases} 3x + y - z = 5 \\ x + 2y = 8 \\ y - 2z = -5. \end{cases}$ **b)** $\begin{cases} x + 2y + 3z = -2 \\ -x + y + z = -1 \\ 2x + 3y = -3. \end{cases}$ **c)** $\begin{cases} 4x + y - z = -5 \\ 8x + 2z = -6 \\ x + y = -1. \end{cases}$

   d) $\begin{cases} 0{,}2x + 0{,}3y + 0{,}4z = 0{,}9 \\ -0{,}1x + 0{,}1y + 0{,}2z = 0{,}2 \\ 1{,}1x + 0{,}2y - 2z = -0{,}7. \end{cases}$ **e)** $\begin{cases} y + 2z = -1 \\ 0{,}2x + 1{,}1y + 0{,}3z = 1{,}2 \\ 0{,}3x - y - 2z = -1. \end{cases}$ **f)** $\begin{cases} 2x + 3y - 2z = 66 \\ y + 4z = 90 \\ y + 5z = 45. \end{cases}$

   g) $\begin{cases} 0{,}3x + 0{,}2y + 0{,}5z = 1 \\ 0{,}1x - 0{,}1y = 0 \\ 0{,}2x + 1{,}1y + 0{,}3z = 1{,}1. \end{cases}$ **h)** $\begin{cases} x + 2y + 3z = 2 \\ 1{,}1x + 0{,}5y + 1{,}6z = 2{,}2 \\ -x - 2y - 3z = -2. \end{cases}$ **i)** $\begin{cases} 50x + 20y + 8z = 20{,}6 \\ 30x + 15y + 16z = 15{,}2 \\ 25x + 32y + 40z = 21{,}9. \end{cases}$

   j) $\begin{cases} \frac{1}{2}x + \frac{1}{3}y + \frac{1}{6}z = 11 \\ \frac{1}{6}x + \frac{1}{2}y + z = 21 \\ x + \frac{1}{3}y + \frac{1}{4}z = \frac{37}{2}. \end{cases}$ **k)** $\begin{cases} \frac{1}{4}x + \frac{1}{5}y + \frac{9}{20}z = 27 \\ \frac{1}{5}x + \frac{1}{4}y + \frac{9}{20}z = 27 \\ x + \frac{1}{2}y + \frac{3}{2}z = 90. \end{cases}$

   l) $\begin{cases} \sqrt{2}x + \sqrt{3}y + z = 5 + \sqrt{6} \\ x + \sqrt{2}y + \sqrt{3}z = 4\sqrt{2} + \sqrt{6} \\ -x + \sqrt{3}y + \sqrt{2}z = 3 - \sqrt{2} + 2\sqrt{3}. \end{cases}$ **m)** $\begin{cases} 0{,}8x + 1{,}5y + 2{,}3z = 2{,}4 \\ 1{,}2x + 0{,}8y + 2z = 3{,}6 \\ 1{,}2x - 0{,}4y + 0{,}8z = 3{,}0. \end{cases}$

9. Sean $a, b, c \in \mathbb{R}$ con $a \neq 0$. Se considera la ecuación: hallar $x \in \mathbb{R}$ tal que $ax^4 + bx + c = 0$. Elaborar un algoritmo que permita identificar la existencia de raíces reales y como calcularlas. Verifique su algortimo en los siguientes casos:

   a) $t^4 - 9t^2 + 20 = 0$. **b)** $3t^4 + 7t^2 - 40 = 0$. **c)** $2t^4 + 9t^2 + 4 = 0$.

10. En cada item se define una función u y una partición uniforme $\tau(m)$ del intervalo $[a, b]$ que se indica y $m = 10$. Calcule $I(u) = \int_a^b u(x)\,dx$ y una aproximación $I(v_m)$ de $I(u)$ calculada con la regla del rectángulo. Compare los resultados.

   a) $u(x) = x \quad x \in [0, 10]$. **b)** $u(x) = -3x + 2 \quad x \in [-1, 2]$. **c)** $u(x) = 2x^2 + 5 \quad x \in [-1, 2]$.

   d) $u(x) = x^3 - x^2 + 1 \quad x \in [0, 1]$. **e)** $u(x) = \dfrac{1}{1+x} \quad x \in [0, 1]$. **f)** $u(x) = e^{-x} \quad x \in [0, 4]$

   g) $u(x) = \operatorname{sen}(x) \quad x \in \left[0, \dfrac{\pi}{2}\right]$. **h)** $u(x) = \cos^2(x) \quad x \in [0, \pi]$. **i)** $u(x) = \ln(x) \quad x \in [1, e]$.

   j) $u(x) = \sqrt{1 + x^2} \quad x \in [0, 2]$.

11. En cada item se define una función real φ. Elabore un algoritmo de cálculo de $\varphi(x)$ de modo que el número de operaciones elementales sea el más pequeño posible, contabilice dicho número.

   a) $\varphi(x) = 10 - \dfrac{1}{x^2} - \dfrac{1}{6x^4} - \dfrac{1}{10x^6} - \dfrac{1}{14x^8} - \dfrac{1}{18x^{10}} \quad x > 1.$

   b) $\varphi(x) = 1 + \dfrac{3}{\sqrt{1+x}} - \dfrac{5}{1+x} + \dfrac{7}{(1+x)^{\frac{3}{2}}} - \dfrac{9}{(1+x)^2} + \dfrac{11}{(1+x)^{\frac{5}{2}}} \quad x \geq 0.$

   c) $\varphi(x) = \dfrac{4}{(x^2-3)^{\frac{1}{2}}} + \dfrac{9}{5(x^2-3)^{\frac{3}{2}}} + \dfrac{14}{9(x^2-3)^{\frac{5}{2}}} + \dfrac{19}{14(x^2-3)^{\frac{7}{2}}} + \dfrac{24}{19(x^2-3)^{\frac{9}{2}}} \quad x > \sqrt{3}.$

   d) $\varphi(x) = 2\left(x^2 + \dfrac{4}{3}x^4 + \dfrac{16}{9}x^6 + \dfrac{256}{81}x^8\right)^{\frac{1}{2}} \quad x \in \mathbb{R}.$

   e) $\varphi(x) = 1 + \dfrac{1}{2}\operatorname{sen}(x) - \dfrac{1}{4}\operatorname{sen}^2(x) + \dfrac{1}{8}\operatorname{sen}^3(x) - \dfrac{1}{16}\operatorname{sen}^4(x) \quad x \in \left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right]$.

   f) $\varphi(x) = \dfrac{1}{3} + \dfrac{2}{9}\cos^2(x)\operatorname{sen}(x) - \dfrac{3}{16}\cos^3(x)\operatorname{sen}^2(x) + \dfrac{4}{21}\cos^4(x)\operatorname{sen}^3(x) - \dfrac{5}{26}\cos^5(x)\operatorname{sen}^4(x) \quad x \in \mathbb{R}.$

12. Para $n \in \mathbb{Z}^+$ con $3 \leq n \leq 19$, se define

$$f_n(x) = \sum_{k=0}^{n} \frac{(-1)^k}{(2k+1)!} x^{k+\frac{1}{2}} \quad x \in \left[0, \pi^2\right].$$

   a) Para cada n impar, elabore un algoritmo para calcular valores aproximados de $f_n(x)$ de modo que el número de operaciones elementales sea el más pequeño posible y se eviten sumas y restas alternadas.

   b) Para $n = 19$, $x = \left(\dfrac{\pi}{6}\right)^2$ y una aproximación de $\pi \simeq 3{,}1415926536$, se tiene $f_{19}\left(\dfrac{\pi}{6}\right) = 0{,}5$. Aplique el algoritmo desarrollado en la parte a) precedente y calcule $f_n(x)$ para n, x y la aproximación de π que se indica, obtendrá una aproximación de 0,5

   i) $n = 5$, $\pi \simeq 3{,}1415$, $x = \left(\dfrac{\pi}{6}\right)^2$. **ii)** $n = 7$, $\pi \simeq 3{,}141593$, $x = \left(\dfrac{\pi}{6}\right)^2$.

   iii) $n = 9$, $\pi \simeq 3{,}14159265$, $x = \left(\dfrac{\pi}{6}\right)^2$. **vi)** $n = 11$, $\pi \simeq 3{,}141593$, $x = \left(\dfrac{\pi}{6}\right)^2$.

13. La función real g se define sobre $[0, 10]$ como sigue:

$$g(x) = \sum_{k=0}^{15} \frac{(-1)^k}{k!10^k} x^{k+2} \quad x \in [0, 10]$$

   a) Elabore un algoritmo para calcular valores aproximados de $g(x)$ de modo que se eviten los cálculos directos de $k!$, 10^k, x^{k+2} y sumas y restas alternadas; y, en lo posible que el número de operaciones elementales sea el más pequeño posible.

b) Contabilice el número de operaciones elementales de su algoritmo.

c) Aplique su algoritmo para calcular $g(1)$ y compruebe que obtendrá una aproximación de 0,904837418.

d) Aplique su algoritmo para calcular $g(10)$ y obtendrá una aproximación de 36,78794412.

14. Considerar la función h definida como

$$h(x) = \sum_{k=0}^{5} \frac{x^{2^k}}{(8k)!5^{3k^2}} = x + \frac{x^2}{8!5^3} + \frac{x^4}{16!5^{12}} + \frac{x^8}{24!5^{27}} + \frac{x^{16}}{32!5^{48}} + \frac{x^{32}}{40!5^{75}} \quad x \in \mathbb{R}.$$

a) Con la calculadora de bolsillo calcule $(8k)!$, 5^{3k^2}, $(8k)!5^{3k^2}$ y $\dfrac{1}{(8k)!5^{3k^2}}$ para $k = 0, 1, \ldots, 5$ y analice las dificultades de cálculo y los resultados que obtiene.

b) Utilice el desarrollo de $h(x)$ para calcular $h(20)$ y explique las dificultades de cálculo que se presentan.

c) Sean $x \in \mathbb{R}$ y $u(x) = \dfrac{1}{25 \times \cdots \times 32 \times 5^3} \left(\dfrac{x}{25}\right)^8$. Note que $u(x) = \dfrac{1}{25 \times \cdots \times 32 \times 5^3} \left(\dfrac{x}{25}\right)^8 = \dfrac{1}{125} \times \dfrac{\frac{x}{25}}{25} \times \dfrac{\frac{x}{25}}{26} \times \cdots \times \dfrac{\frac{x}{25}}{32}$.
Calcule $u(20)$.

d) A partir de la escritura de $h(x)$ siguiente:

$$\begin{aligned} h(x) &= x + \frac{x^2}{8!5^3}\left(1 + \frac{1}{9 \times \cdots \times 16 \times 5}\left(\frac{x}{5^4}\right)^2\left(1 + \frac{1}{17 \times \cdots \times 24 \times 5^3}\left(\frac{x}{5^3}\right)^4\right.\right. \\ &\quad \left.\left.\left(1 + \frac{1}{25 \times \cdots \times 32 \times 5^5}\left(\frac{x}{5^2}\right)^8\left(1 + \frac{1}{17 \times \cdots \times 24 \times 5^3}\left(\frac{x}{5}\right)^{16}\right)\right)\right)\right) \end{aligned}$$

y de la observación en la parte c) precedente, exprese $h(x)$ en forma más conveniente y calcule $h(20)$. Explique las dificultades o bondades de cálculo con la nueva escritura de h.

15. Elaborar un logaritmo que permita calcular los valores de $P_m(x)$, $Q_m(x)$ $\quad m = 0, 1, 2, \ldots,$ $x \in [-1, 1]$, si

$$P_m(x) = 1 + \frac{(m!)^2}{(2m)!} \sum_{k=1}^{m} (-1)^k \frac{(2m+2k)!}{(m-k)!(m+k)!(2k)!}\, x^{2k} \quad x \in [-1, 1],$$

$$Q_m(x) = x + \frac{(m!)^2}{(2m+1)!} \sum_{k=1}^{m} (-1)^k \frac{(2m+2k+1)!}{(m-k)!(m+k)!(2k+1)!}\, x^{2k+1} \quad x \in [-1, 1],$$

Los polinomios P_m y Q_m son conocidos como polinomios de Legendre.

16. Se define $v(x) = \sum_{k=0}^{15} \dfrac{1}{k!} \dfrac{1}{(x^2+1)^k} \quad x \in [0, 10]$.

a) Utilice directamente la escritura de $v(x)$ para calcular $v(3)$ y determine el número de operaciones elementales que realiza. Indique las posibles dificultades de cálculo de $v(3)$.

b) Elabore un algoritmo para calcular valores aproximados de $v(3)$ de modo que el número de operaciónes elementales sea el más pequeño posible y contabilice el total de dichas operaciones elementales en el cálculo de $v(x)$ $\quad x \in [0, 10]$. Compare su resultado con el siguiente: $v(3) \simeq 1{,}105170918$.

c) Aplique su algoritmo y calcule $v(10)$ y compruebe su resultado con $v(10) \simeq 1{,}009950167$.

17. Considere la función θ definida como

$$\theta(x) = \frac{1}{2} \sum_{k=0}^{15} \frac{(-1)^k x^k}{(2k+1)!4^k} \quad x \geq 0.$$

a) Utilice sin modificaciones $\theta(x)$ y calcule $\theta\left(\left(\frac{\pi}{3}\right)^2\right) \simeq 0{,}477464829$. ¿Qué dificultades de cálculo se presentan?

b) Elabore un algoritmo que facilite el cálculo de $\theta(x)$ y contabilice el número de operaciones elementales que se realizan. Calcule $\theta\left(\left(\frac{\pi}{3}\right)^2\right)$ y compare con el valor dado en i) precedente.

18. Se da la función f definida como $f(x) = \sum_{k=0}^{11} \frac{x^{2k}}{(2k)!} \quad x \in [0,2]$.

a) Calcule $f\left(\frac{1}{2}\right)$ y compare con $f(0{,}5) \simeq 1{,}127625966$. Contabilice el número de operaciones elementales que realiza.

b) Mejore aún la escritura de $f(x)$ siguiente:

$$f(x) = 1 + \frac{x^2}{2}\left(1 + \frac{x^2}{3\times 4}\left(1 + \frac{x^2}{5\times 6}\left(1 + \cdots + \frac{x^2}{19\times 20}\left(1 + \frac{x^2}{21\times 22}\right)\cdots\right)\right)\right)$$

y calcule $f\left(\frac{1}{2}\right)$. Contabilice el número de operaciones elementales que realiza.

19. Dada la función $g(x) = \sum_{k=0}^{15} \frac{x^{2k}}{(2k+1)!} \quad x \geq 0$. Mediante la elaboración de un algoritmo que facilite el cálculo de $g(x)$, Calcule $g\left(\frac{1}{2}\right)$ y compare con $g\left(\frac{1}{2}\right) \simeq 1{,}042190611$. ¿Cuántas operaciones se requieren para calcular $g(x)$ con y sin el algoritmo?

20. Aplique el esquema de Hörner para calcular $p(x)$ en x que se indica.
a) $p(x) = x^8 + x^7 + x^5 + x^3 + x^2 + x + 1, \quad x = 2.$
b) $p(x) = 5 - x^2 + 10x^3 + 7x^4 - 2x^5, \quad x = -3.$
c) $p(x) = 0{,}5 - 0{,}2x + 0{,}5x^2 + 3{,}25x^3 + 2{,}5x^4, \quad x = 0{,}8.$
d) $p(x) = 3 - 2{,}2x + 1{,}1x^3 - 2{,}8x^4 + 5{,}6x^5, \quad x = 1{,}5.$

21. Considere la función u definida como $u(x) = x^2 \quad x \in [0,2]$. En cada literal se da el número de puntos m de una partición uniforme $\tau(m-1)$ de $[0,2]$. Trace la gráfica de u y de su interpolante v_m utilizada en la regla del rectángulo. Calcule $I(u) = \int_0^2 x^2 dx$ e $I(v_m) = \sum_{j=1}^{m} hu\left(x_{j-1} + \frac{1}{2}jh\right)$.
a) $m = 2$. **b)** $m = 5$. **c)** $m = 9$. **d)** $m = 11$.
Compare los resultados. Para el efecto, calcule $|I(u) - I(v_m)|$ y concluya.

22. Se define la función f definida como $f(x) = \text{sen}(x) \quad x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ y $\pi \simeq 3{,}1415926536$. Se define una partición uniforme $\tau(m) = \left\{-\frac{\pi}{2} + ih \mid i = 0, 1, \ldots, m\right\}$ de $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ con $m \in \mathbb{Z}^+$ que en cada literal se define. Trace la gráfica de f y la de su interpolante f_m utilizada en la regla del rectángulo. Calcular $I(f) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \text{sen}(x)\, dx$, $I(f_m) = \sum_{i=1}^{m} hf\left(-\frac{\pi}{2} + \frac{i}{2}h\right)$.
a) $m = 3$. **b)** $m = 5$. **c)** $m = 7$. **d)** $m = 9$.
Para cada m dado en a), b), c), d) calcule $|I(f) - I(f_m)|$ y concluya.

23. Se define la función g como $g(x) = e^x - \ln(x) \quad x \in [1,2]$.
a) Calcule $I(g) = \int_1^2 g(x)\, dx$.
b) Se define una partición $\tau(m) = \{1 + ih \mid i = 0, 1, \ldots, m\}$ con $m \in \mathbb{Z}^+$ que se da en cada caso. Aplique la regla del rectángulo para aproximar $I(g)$, para $m = 3$, $m = 6$, $m = 9$, $m = 12$.
c) Calcule $|I(g) - I(g_m)|$ con $I(g_m)$ calculado en la parte b) precedente. Concluya.

24. Considere la función real f definida como $f(x) = e^x \quad x \in \mathbb{R}$. Se sabe que $f'(-1) = e^{-1}$. Calcule aproximaciones y_0' de la derivada $f'(1)$ para cada h que se indica y calcule $|f'(-1) - y_0'|$. Analice los resultados.

a) $h = -0{,}05$. **b)** $h = -0{,}0005$. **c)** $h = -0{,}000005$. **d)** $h = 0{,}005$. **e)** $h = 0{,}00005$.

f) $h = 0{,}0000005$.

25. Se define la función u como $u(x) = \sqrt{1+x^2} \quad x \in \mathbb{R}$. Tenemos $u'(x) = \dfrac{x}{\sqrt{1+x^2}} \quad x \in \mathbb{R}$. Calcule aproximaciones de la derivada $u'(0) = 0$ para cada h que se indica y estime $|u'(0) - u_0'|$.

a) $h = -0{,}02$. **b)** $h = -0{,}0002$. **c)** $h = -0{,}00002$. **d)** $h = 0{,}002$. **e)** $h = 0{,}00002$.

f) $h = 0{,}0000002$.

26. Considere aproximaciones $v'(x) = \operatorname{sen}(3x) \quad x \in \mathbb{R}$. Se sabe que $v'(x) = 3\cos(3x) \quad x \in \mathbb{R}$. Calcule aproximaciones v_0' de la derivada $v'\left(\dfrac{\pi}{6}\right) = 0$ con $\pi \simeq 3{,}1416926536$, para cada h que se indica.

a) $h = -0{,}04$. **b)** $h = -0{,}0004$. **c)** $h = -0{,}000004$. **d)** $h = 0{,}004$. **e)** $h = 0{,}00004$.

f) $h = 0{,}0000004$.

27. En cada item se define una función v y se dan un punto x_0 y varios valores de h. Calcule aproximaciones v_0' de la derivada $v'(x_0)$ y estime $|v'(x_0) - v_0'|$.

a) $v(x) = 2x^2 - 3 \quad x \in \mathbb{R},\ x_0 = -1,\ h = -0{,}003,\ h = -0{,}0003,\ h = 0{,}00003,\ h = 0{,}000003.$

b) $v(x) = \dfrac{1}{1+x} \quad x > -1,\ x_0 = 0,\ h = -0{,}05,\ h = -0{,}00005,\ h = 0{,}0005,\ h = 0{,}000005.$

c) $v(x) = \cos(x^2) \quad x \in \mathbb{R},\ x_0 = \sqrt{\dfrac{\pi}{2}},\ h = -0{,}001,\ h = -0{,}0001,\ h = 0{,}00001,\ h = 0{,}0000001.$

d) $v(x) = \ln(1+2x) \quad x > -\dfrac{1}{2},\ x_0 = 0,\ h = -0{,}015,\ h = -0{,}000015,\ h = 0{,}00015,\ h = 0{,}0000015.$

e) $v(x) = \left(1+2x^2\right)^{\frac{1}{3}} \quad x \in \mathbb{R},\ x_0 = 2,\ h = -0{,}025,\ h = -0{,}00025,\ h = 0{,}000025,\ h = 0{,}0000025.$

28. La solución del problema de valor inicial siguiente: $\begin{cases} y'(x) + 2y(x) = e^x & 0 < x < 1, \\ y(0) = \dfrac{1}{3}, \end{cases}$ es $y(x) = \dfrac{1}{3}e^x$. Para $m = 10$ y una partición uniforme del intervalo $[0,1]$, aplique el método de Euler explícito y calcule aproximaciones $y_j \quad j = 1, \ldots, 10$ de $y(x_j)$. Trace la gráfica de la función $y(x)$ y represente los puntos $(x_j, y_j) \quad j = 0, 1, \ldots, 10$. Calcule $|y(x_j) - y_j| \quad j = 0, 1, \ldots, 10$ y dé una solución.

29. La solución del problema de valor inicial $\begin{cases} y'(x) = \dfrac{2-y}{1-x^2}x & |x| < 1, \\ y(0) = 1, \end{cases}$ es $y(x) = 2 - \sqrt{1-x^2} \quad |x| < 1$. Para $m = 8$ y una partición uniforme del intervalo $[0, 0{,}9]$, aplique el método de Euler explícito y calcule aproximaciones $y_j \quad j = 0, 1, \ldots, 8$ de $y(x_j)$. Trace las gráficas de la fución $y(x)$ y de los valores calculados $(x_j, y_j) \quad j = 0, 1, \ldots, 8$. Calcule $|y(x_j) - y_j| \quad j = 0, 1, \ldots, 8$ y compare los resultados.

30. Considere el problema de valor inicial $\begin{cases} y'(x) = \dfrac{y(x)}{x} - 1 & 1 < x < 3, \\ y(1) = 2. \end{cases}$ cuya solución es $y(x) = x(2 - \ln(x)) \quad x > 0$. Para $m = 10$ y una partición uniforme del intervalo $[1, 3]$, aplique el método de Euler explícito y calcule aproximaciones $y_j \quad j = 0, 1, \ldots, 10$ de y(x_j). Trace las gráficas de la función $y(x) \quad x \in [1, 3]$ y de los valores calculados $(x_j, y_j) \quad j = 0, 1, \ldots, 10$. Calcule $|y(x_j) - y_j| \quad j = 0, 1, \ldots, 10$ y compare los resultados.

31. Considere el problema de valor inicial $\begin{cases} y'(x) = \dfrac{x^2 + y^2(x)}{3xy(x)} & 1 < x < 1{,}5, \\ y(1) = 2, \end{cases}$ la solución es $y(x) = \sqrt{x^2 + 3x} \quad x > 0$. Para $m = 5$ y una partición uniforme del intervalo $[1, 1{,}5]$, proceda como en el ejercicio precedente.

32. Representar en base 10 los siguientes números:

    a) $(4746)_8$ **b)** $(7412{,}352)_8$ **c)** $(AB98.C31)_{16}$ **d)** $(100{,}001)_3$ **e)** $(1{,}4142)_5$ **f)** $(111{,}0101)_2$

    g) $(0{,}1110111110111011111\ldots)_2$ **h)** $(235{,}3333\ldots)_6$.

33. En cada caso, representar los siguientes números en base 10 a la base b que se indica con 8 cifras de precisión para la parte fraccionaria.

    a) 3,14159, $b = 2$. **b)** 2,718281, $b = 4$. **c)** $\sqrt{2}$, $b = 8$. **d)** $\sqrt{5}$, $b = 5$. **e)** $\dfrac{27}{7}$, $b = 3$.

    f) 1726,00011, $b = 2$. **g)** 135,26 $b = 4$. **h)** 135,42, $b = 8$.

34. Sean $a, b \in \mathbb{N}$ tales que $a \neq b$, $1 < a \leq 10$, $1 < b \leq 10$. Elaborar algoritmos que permitan convertir números positivos en base a a base b y recíprocamente; y, obtener su equivalente en base 10. Sugerencia: considérese $M = (a_n \ldots a_0, a_{-1} \ldots a_{-m})_a$ y $M = (b_p \ldots b_0, b_{-1} \ldots b_{-q})_b$, donde $a_i, a_{-j} \in \{0, 1, \ldots, a-1\}$, $i = 0, 1, \ldots, n$, $j = 1, \ldots, m$, $b_k, b_{-l} \in \{0, 1, \ldots, b-1\}, k = 0, 1, \ldots, p$, $l = 1, \ldots, q$.

35. Sean $a, b, c \in \mathbb{R}^+$.

    a) Considerar las expresiones $u = (a - b)c$ y $v = ac - bc$ con $a \approx b$. Demostrar que u presenta un error relativo menor que v. Verifique con $a = 0{,}6392$, $b = 0{,}6375$ y $c = 0{,}9364$.

    b) Considerar la matriz $A = \begin{bmatrix} a & b \\ c & a \end{bmatrix}$. Estudie el condicionamiento de $\det(A)$.

    c) Si $a = \dfrac{1}{\sqrt{3}}$, $b = 1$, $c = \frac{1}{3}$, $a_1 = rd(a)$, $b_1 = rd(b)$, $c_1 = rd(c)$, estudie la existencia de soluciones de los sistemas de ecuaciones $\begin{cases} ax + by = 1 \\ cx + ay = 0 \end{cases}$ y $\begin{cases} a_1x + b_1y = 1 \\ c_1x + a_1y = 0. \end{cases}$

    d) Sean $a = \dfrac{1}{1500}$, $b = \dfrac{1}{701}$ y $c = \dfrac{7}{22500}$. Si a, b, c se redondean con 8 cifras decimales, estudie la existencia de soluciones de los sistemas de ecuaciones del literal c).

    e) De b), c) y d), ¿qué conclusiones puede obtener?

36. Determinar el número de operaciones elementales para calcular $\det(A)$ si este se calcula usando el método de menores y cofactores cuando A es una matriz de 3×3, de 4×4 y de 5×5. Generalice los resultados.

37. Determinar el número de operaciones elementales que se requieren para calcular A^D la matriz adjunta de A cuando A es una matriz de 3×3, de 4×4 y de 5×5.

38. Usando la aritmética de punto flotante con 3 dígitos, evaluar $f(x) = x^4 - x^3 + 6x^2 - 3x + 0{,}145$ en $x = 4{,}71$.

    a) Aplicar el esquema de Hörner para calcular $f(4{,}71)$.

    b) Determinar el valor exacto de $f(4{,}71)$ y, en cada caso, calcular el error relativo.

    c) Calcular con 3 cifras de precisión $f(-0{,}101)$ directamente y con el esquema de Hörner. Calcular el error relativo.

    d) Calcular con 3 cifras de precisión $f(-0{,}10001)$ directamente y con el esquema de Hörner. Calcular el error relativo.

39. Sea f la función real definida por $f(x) = 2 + \frac{3-x}{x^2-1} \quad |x| \neq 1$.

    a) Calcular f con 2 y 3 cifras en aritmética de punto flotante en $x = 0{,}85$, $x = 0{,}95$, $x = 0{,}99$.

    b) Puesto que $f(x) = \dfrac{x}{x-1} + \dfrac{x-1}{x+1}$, calcule $f(x)$ para los puntos x del literal a).

c) Calcule el valor exacto de $f(x)$ para los puntos x dados en a).

d) Calcule el error relativo de $f(x)$ para los resultados de a) y b).

40. Sea $f(x) = \dfrac{1+x-e^x}{x^2} \quad x \neq 0.$

a) Calcular $\lim_{x\to 0} f(x)$. **b)** Calcular $f(0{,}5 \times 10^{-10})$.

c) Hallar un algoritmo para aproximar $f(x)$ con $|x| \in]0, 10^{-5}]$, y aplique en los puntos $x = 0{,}5\times 10^{-5}$ y $x = 0{,}1 \times 10^{-5}$.

41. Sean $x > 1$ y $n \in \mathbb{N}$. Construya algoritmos que permitan aproximar $\dfrac{x^n}{n!}$ en los siguientes casos:

a) $1 < x < 10$ y $20 < n < 50$. **b)** $x \geq 10$ y $n > 50$.

42. Hallar $\lim_{x\to 0} f(x)$ para las funciones f que se dan a continuación. En cada caso elabore algoritmos que se adapten a la estabilidad numérica en un entorno de cero.

a) $f(x) = \sqrt{x^2+1} - 1$. **b)** $f(x) = \sqrt{x^2+1} - x$. **c)** $f(x) = -x + \operatorname{sen}(x)$. **d)** $f(x) = \dfrac{1}{x+1} - 1$.

e) $f(x) = 1 - \cos(x)$. **f)** $f(x) = \dfrac{e - e^{\cos(x)}}{x^2}, x \neq 0$. **g)** $f(x) = \dfrac{e^x - e^{-x}}{\operatorname{sen}(x)}, x \neq k\pi,\ k \in \mathbb{Z}$.

h) $f(x) = \dfrac{e^x - e^{\operatorname{sen}(x)}}{x^3}, x \neq 0$. **i)** $f(x) = \dfrac{1-\cos(x)}{x},\ x \neq 0$. **j)** $f(x) = \dfrac{e^x - (1+x)}{x^2},\ x \neq 0$.

43. Determinar los números de condicionamiento de las funciones siguientes:

a) $f(x) = \cos(x) \quad x \in \mathbb{R}$. **b)** $f(x) = \tan(x) \quad x \in \left]-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right[$. **c)** $f(x) = \ln(x) \quad x > 0$.

d) $f(x) = 0{,}5 + \dfrac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2} dt \quad x \geq 0$.

e) En los incisos a), b) y c) determinar el más grande subconjunto de $\mathbb{R}$ en el que f está bien condicionado.

f) Pruebe que la función del inciso d) está bien condicionada para todo $x \geq 0$.

44. Sea $\varphi : \mathbb{R}^3 \longrightarrow \mathbb{R}$ la función definida por $\varphi(x, y, z) = \frac{xy}{z}$ con $z \neq 0$.

a) Pruebe que $\varepsilon_\varphi = \varepsilon_x + \varepsilon_y - \varepsilon_z$,donde ε_x, ε_y, ε_z son los errores relativos de x, y, z respectivamente.

b) Determine el error acumulado.

45. Si $\varphi(x, y) = x^y \quad x, y \in \mathbb{R}^+$ y $\varepsilon_x = \varepsilon_y$, pruebe que el error relativo de φ viene dado por

$$\varepsilon_\varphi = \varepsilon_x + (y \ln(x))\varepsilon_y.$$

¿Qué número de condicionamiento influye en el cálculo de φ?.

46. Sea $A = (a_1, \ldots, a_n) \in \mathbb{R}^n$, donde a_i, $i = 1, \ldots, n$ son números de máquina. Sea $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ la función definida por $F(x_1, \ldots, x_n) = \sum_{i=1}^n a_i x_i$. Si $|\varepsilon_{x_i}| \leq eps$, $i = 1, \ldots, n$, pruebe que el error relativo de $F(x_1, \ldots, x_n)$ verifica

$$|\varepsilon_F| \leq eps \text{ si } \begin{cases} a_i > 0, & x_i > 0, \\ a_i < 0, & x_i < 0, \end{cases} \quad i = 1, \ldots n.$$

47. Sea $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ la función definida por $F(a_1, \ldots, a_n) = \frac{1}{n} \sum_{i=1}^n a_i$. Supongamos que $a_i > 0$ $\forall i = 1, \ldots, n$. Proponer un algoritmo de cálculo de $z = F(a_1, \ldots, a_n)$ y estudiar la propagación de los errores. Si el error en cada operación es $\varepsilon_i = \varepsilon$, ¿cuál es el error acumulado?.

48. Sean $(a_1,\ldots,a_n) \in \mathbb{R}^n$ y $\vec{z} = \vec{\varphi}(a_1,\ldots a_n)$, donde $\vec{\varphi}(a_1,\ldots a_n) = \begin{bmatrix} \varphi_1(a_1,\ldots,a_n) \\ \vdots \\ \varphi_n(a_1,\ldots,_n) \end{bmatrix}$, con $\varphi_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1,\ldots,n$ funciones de clase C^1. Supóngase que $\varphi_i(a_1,\ldots a_n) \neq 0$, $i = 1,\ldots,n$. Muestre que los números de condicionamiento de φ_i viene dados por

$$C_i = \frac{a_j}{\varphi_i(a_1,\ldots,a_n)} \frac{\partial \varphi_i}{\partial x_j}(a_1,\ldots,a_n), \quad i,j = 1,\ldots,n.$$

49. Considerar la ecuación $x^2 + 2px - q = 0$, donde $p > 0, q > 0$ y $p >> q$. Sea $x = -p + \sqrt{p^2+q}$ una raíz de la ecuación. Calcule ε_x y demuestre que

$$eps \leq \varepsilon_x = \frac{\Delta x}{x} \leq 3eps.$$

Nota: La notación $p >> q$ significa q es muy pequeño comparado con p o que p es muy grande comparado con q.

50. Se desea calcular $E(x) = \sqrt{1+x} - 1$ para $x = 0{,}0009$ y $x = 0{,}001$ con 4 cifras decimales.

a) Calcule $E(x)$ en dichos puntos. **b)** Utilizando $E(x)$, construya otra expresión que se adapte a la estabilidad numérica y aplique para los puntos dados x. Compare los resultados.

51. Considerar el sistema de ecuaciones lineales $\begin{cases} 0{,}002x + y = 0{,}2 \\ x + y = 1. \end{cases}$ Utilice el método de eliminación gaussiana para determinar:

a) La solución exacta del sistema. **b)** La solución aproximada con 5 cifras decimales. **c)** Intercambie la primera ecuación con la segunda y proceda como en los incisos a) y b). Compare los resultados.

d) El sistema de ecuaciones propuesto es equivalente al siguiente: $\begin{cases} x + 500y = 100 \\ x + y = 1. \end{cases}$ Usando la aritmética de punto flotante con 5 dígitos de precisión, resuelva el sistema de ecuaciones y compare con los resultados precedentes.

52. Sean $a, b \in \mathbb{R}^+$, $m, n \in \mathbb{N}$ tales que $0 \leq m \leq n$. Se desea calcular

$$F(m) = \sum_{k=0}^{m} \binom{n}{k} a^k b^{n-k} \quad m = 0,1,\ldots,n,$$

donde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

a) Pruebe que $\binom{n}{m+1} = \frac{n-m}{m+1}\binom{n}{m} \quad m = 0{,}1.\ldots,n-1.$

b) Sea $\varphi(k) = \binom{n}{k} a^k b^{n-k}$. Entonces $F(m) = \sum_{k=0}^{m} \varphi(k)$. Elabore un algoritmo que se adpate a la estabilidad numérica y aplique para $a = 0{,}1$, $b = 0{,}5$, $m = 4, n = 10$.

53. Sea $I(n) = \int_0^1 \frac{x^n}{x+5} dx$, $n = 0,1,2,\ldots$

a) Muestre que $I(n) + 5I(n-1) = \frac{1}{n}$.

b) Considerar el algoritmo: $I(n) = \frac{1}{n} - 5I(n-1) \quad n = 1,2,\ldots,25$. Calcule $I(0)$ e $I(n)$ para $n = 1,2,\ldots,25$.

c) Muestre que $\lim_{n\to\infty} I(n) = 0$.

d) Considere el siguiente algoritmo: $\begin{cases} I(n) = \frac{1}{5n}, \\ I(n-1) = \frac{1}{5}\left(\frac{1}{n} - I(n)\right), n = 25,24,\ldots,1. \end{cases}$ Calcule $I(n)$ e $I(n-1), n = 25,\ldots,1$. Compare con los resultados anteriores.

54. Sea $I(n) = \int_0^1 x^{\frac{n+4}{2}}\, e^x dx \quad n = -4, -3, -2, \ldots$

a) Muestre que $I(n) = e - \left(\frac{n}{2} + 2\right) I(n-2) \quad n = -2, -1, 0, \ldots$.

b) Calcule $I(-4)$.

c) Use el cambio de variable $x = t^2$ y muestre que $I(-3) = e - \int_o^1 e^{t^2} dt$. Utilice el polinomio de Taylor de $e^\alpha \quad \alpha \in \mathbb{R}$ para aproximar $\int_0^1 e^{t^2} dt$ y muestre que

$$\int_0^1 e^{t^2} dt = 1 + \frac{1}{3} + \frac{1}{2!}\frac{1}{5} + \cdots \frac{1}{k!}\frac{1}{2k+1} + E_{k+1},$$

donde E_{k+1} es el error cometido y k es tal que $\dfrac{1}{k!}\,\dfrac{1}{2k+1} < 10^{-6}$.

d) Utilizando el algoritmo dado en a), elabore un programa para calcular $I(n) \quad n = -4, -3, -2, \ldots, 25$.

e) Note que $\lim_{n\to\infty} I(n) = 0$. Establezca el siguiente algoritmo $\begin{cases} I(n) = \dfrac{2e}{n+4}, \\ I(n-1) = \dfrac{2e}{n+3}, \\ I(n-2) = \dfrac{2(e - I(n))}{n+4}. \end{cases}$

f) Elabore un programa para el cálculo de $I(n) \quad n = 25, 24, \ldots, -2$. Compare con los resultados dados en d).

g) Para $\varepsilon = 10^{-6}$, deduzca el siguiente algoritmo

$$I(n) = 2\left[\frac{1}{0!(n+6)} + \frac{1}{1!(n+8)} + \frac{1}{2!(n+10)} + \frac{1}{3!(n+12)} + \cdots + \frac{1}{8!(n+22)}\right].$$

Elabore un programa para el cálculo de $I(n)$, $n = -4, -3, \ldots, 25$. Compare los resultados con los otros algoritmos. ¿Qué concluye?.

h) ¿Por qué no es práctico utilizar la regla de los trapecios para cada n?.

1.13. Lecturas complementarias y bibliografía

1. Tom M. Apostol, Calculus, Volumen 1, Segunda Edición, Editorial Reverté, Barcelona, 1977.
2. N. Bakhvalov, Métodos Numéricos, Editorial Paraninfo, Madrid, 1980.
3. R. M. Barbolla, M. García, J. Margalef, E. Outerelo, J. L. Pinilla. J. M. Sánchez, Introducción al Análisis Real, Editorial Alambra Universidad, Madrid, 1981.
4. G. Birkhoff, S. Maclane, Algebra Moderna, Cuarta Edición, Editorial Vicens-Vives, Barcelona. 1974.
5. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.
6. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.
7. S. D. Conte, Carl de Boor, Análisis Numérico, Segunda Edición, Editorial Mc Graw-Hill, México, 1981.
8. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.
9. B. P. Demidovich, I. A. Maron, E. S. Schuwalowa, Métodos Numéricos de Análisis, Editorial Paraninfo, Madrid, 1980.

10. Francis G. Florey, Fundamentos de Algebra Lineal y Aplicaciones, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1980.

11. Ferruccio Fontanella, Aldo Pasquali, Calcolo Numerico. Metodi e Algoritmi, Volumi I, Pitagora Editrice Bologna, 1983.

12. Waltson Fulks, Cálculo Avanzado, Editorial Limusa, México, 1973.

13. Curtis F. Gerald, Patrick O. Wheatley, Análisis Numérico con Aplicaciones, Sexta Edición, Editorial Pearson Educación de México, México, 2000.

14. Gene H. Golub, Charles F. Van Loan, Matrix Computations, Second Edition, The Johns Hopkins University Press, Baltimore, 1989.

15. Günther Hämmerlin, Karl-Heinz Hoffmann, Numerical Mathematics, Editorial Springer-Verlag, New York, 1991.

16. Nicholas J. Higham, Accuracy and Stability of Numerical Algorithms, Editorial Society for Industrial and Applied Mathematics, Philadelphia, 1996.

17. Kenneth Hoffman, Ray Kunze, Algebra Lineal, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1987.

18. Robert W. Hornbeck, Numerical Methods, Quantum Publishers, Inc., New York, 1975.

19. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.

20. Rodolfo Luthe, Antonio Olivera, Fernando Schutz, Métodos Numéricos, Editorial Limusa, México, 1986.

21. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

22. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

23. Antonio Nieves, Federico C. Dominguez, Métodos Numéricos Aplicados a la Ingeniería, Tercera Reimpresión, Compañía Editorial Continental, S. A. De C. V., México, 1998.

24. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

25. A. A. Samarski, Introducción a los Métodos Numéricos, Editorial Mir, Moscú, 1986.

26. Michelle Schatzman, Analyse Numérique, Inter Editions, París, 1991.

27. Francis Scheid, Theory and Problems of Numerical Analysis, Schaum's Outline Series, Editorial McGraw-Hill, New York, 1968.

28. Michael Spivak, Calculus, Segunda Edición, Editorial Reverté, Barcelona, 1996.

29. J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Editorial Springer-Verlag, 1980.

30. E. A. Volkov, Métodos Numéricos, Editorial Mir, Moscú, 1990.