CAPÍTULO 1. ERRORES DE REDONDEO Y ESTABILIDAD

INTRODUCCIÓN

Al momento de aplicar las Matemáticas a situaciones del mundo real nos encontramos a menudo con problemas que no pueden ser resueltos analíticamente o de manera exacta y cuya solución debe ser abordada con ayuda de algún procedimiento numérico. A continuación consideramos algunos problemas típicos, ya formulados matemáticamente, para los cuales estudiaremos técnicas numéricas de solución.

Problema 1.1 Encontrar el área de la región comprendida entre las gráficas de $y = 2\operatorname{sen} x$, $y = e^{-x}$ con $x \in [0, \pi]$. ♦

Problema 1.2 Encontrar las raíces de la ecuación polinómica

$$x^5 + 11x^4 - 21x^3 - 10x^2 - 21x - 5 = 0 \quad \blacklozenge$$

Problema 1.3 Resolver los siguientes sistemas de ecuaciones:

a) El sistema lineal $AX = b$ con

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix} \qquad b = \begin{pmatrix} 3 \\ -2 \\ 2 \\ -2 \\ 1 \end{pmatrix}$$

b) El sistema no-lineal

$$\begin{cases} x^2 + xy^3 = 9 \\ 3x^2y - y^3 = 4 \end{cases} \quad \blacklozenge$$

Problema 1.4 Dada la siguiente tabla de datos correspondiente a una cierta función $y = f(x)$,

x_k	–2	–1	0	1	2	3
$f(x_k)$	–5	1	1	1	7	25

TABLA 1.1

encontrar el polinomio de menor grado que pase a través de los puntos dados.

Cuál será una estimación para los valores $f(x)$ correspondientes a $x = -1.5$ y $x = 1.5$? ♦

Problema 1.5 Hallar el valor de cada una de las siguientes integrales:

a) $\int_0^1 \frac{\operatorname{sen} x}{x}\,dx$ **b)** $\int_0^1 e^{x^2}\,dx$

c) $\int_0^{\frac{\pi}{2}} \sqrt{1-\frac{\operatorname{sen}^2 x}{4}}\,dx$ (elíptica) **d)** $\int_2^3 \frac{1}{\ln x}\,dx$ ♦

Problema 1.6 Resolver el problema de valor inicial

$$\begin{cases} \frac{d^2\theta}{dt^2} + \frac{d\theta}{dt} + 16\operatorname{sen}\theta = 0 \\ \theta(0) = \frac{\pi}{4},\ \theta'(0) = 0 \end{cases}$$ ♦

En relación con los problemas anteriores, tenemos que:

En el problema 1.1, es necesario determinar los puntos de intersección de las gráficas de $y = 2\operatorname{sen} x$ y $y = e^{-x}$, para lo cual debemos resolver la ecuación $2\operatorname{sen} x = e^{-x}$ y no disponemos de un método algebraico para hacerlo.

En el problema 1.2, se trata de hallar los ceros de un polinomio de grado 5 y, como sabemos, sólo se conocen métodos algebraicos para encontrar raíces de ecuaciones polinómicas de grado menor o igual que 4.

En el problema 1.3, tenemos dos sistemas de ecuaciones: El de la parte a) es lineal y conocemos métodos de solución (por ejemplo, el método de eliminación Gaussiana), sin embargo, para sistemas de tamaño mayor, no sólo es conveniente sino necesario implementar tales métodos a través del computador (**método numérico**). En la parte b) tenemos un sistema no-lineal y no conocemos métodos algebraicos generales para resolverlo.

El problema 1.4 se puede resolver analíticamente (por **interpolación**), sin embargo para determinar los coeficientes de dichos polinomios existen técnicas que permiten encontrarlos rápidamente y que pueden implementarse en el computador.

El problema 1.5, corresponde a integrales definidas cuyo integrando tiene antiderivada que no es elemental.

Finalmente, en el problema 1.6, la ecuación diferencial ordinaria

$$\frac{d^2\theta}{dt^2} + \frac{d\theta}{dt} + 16\operatorname{sen}\theta = 0 \text{ (ecuación de movimiento de un péndulo)}$$

es no-lineal (por la presencia de $\operatorname{sen}\theta$) y no disponemos de un método analítico para resolverla.

Los problemas anteriores sirven como motivación para el estudio de cinco grandes temas en un primer curso de **métodos numéricos**: solución numérica de una ecuación no-lineal en una variable, solución numérica de sistemas de ecuaciones lineales y no-lineales, interpolación polinomial, integración numérica y solución numérica de problemas de valor inicial para ecuaciones diferenciales ordinarias.

Qué es un método numérico?

Un método numérico es un procedimiento mediante el cual se obtiene, casi siempre de manera aproximada, la solución de ciertos problemas realizando cálculos puramente aritméticos y lógicos (operaciones aritméticas elementales, cálculo de funciones, consulta de una tabla de valores, cálculo proposicional, etc.). Un tal procedimiento consiste de una lista **finita** de instrucciones precisas que especifican una secuencia de operaciones algebraicas y lógicas (**algoritmo**), que producen o bien una aproximación de la solución del problema (**solución numérica**) o bien un mensaje. La eficiencia en el cálculo de dicha aproximación depende, en parte, de la facilidad de implementación del algoritmo y de las características especiales y limitaciones de los instrumentos de cálculo (**los computadores**). En general, al emplear estos instrumentos de cálculo se introducen **errores** llamados **de redondeo**.

1.1 ARITMÉTICA FINITA

Siendo los computadores la herramienta básica en los métodos numéricos es conveniente indicar cómo son los números del computador y cómo se simula su aritmética.

La mayoría de los computadores usan sólo un subconjunto **finito**, relativamente pequeño, de los números reales para representar a "todos" los números reales; este conjunto, que sólo contiene números racionales y que describiremos más adelante, es llamado **conjunto de números de punto flotante o conjunto de números de máquina en punto flotante o** simplemente **conjunto de punto flotante**.

Cada número del computador se representa mediante un número finito de dígitos (**aritmética finita**), según se indica a continuación:

Un número del computador o de punto flotante, distinto de cero, se describe matemáticamente en la forma

$$\sigma \times (.a_1 a_2 ... a_t)_\beta \times \beta^e$$

forma en la cual los símbolos que allí aparecen, tienen el siguiente significado:

$\sigma = +1$ o $\sigma = -1$ es el signo del número.

β es un entero que denota la **base** del sistema numérico usado. Por lo general $\beta = 2$ (**Sistema Binario**), $\beta = 8$ (**Sistema Octal**) o $\beta = 16$ (**Sistema Hexadecimal**).

a_i, $i = 1,2,...,t$, es un entero con $0 \le a_i \le \beta - 1$. Los enteros $0,1,...,\beta - 1$ son llamados **dígitos** en la base β . Nosotros asumiremos en todo lo que sigue que $a_1 \neq 0$, en cuyo caso el número se dice que está en **forma normalizada**.

$(.a_1a_2...a_t)_\beta$ denota la suma $\frac{a_1}{\beta^1}+\frac{a_2}{\beta^2}+...+\frac{a_t}{\beta^t}$ y es llamada la **mantisa o fracción** del número de punto flotante.

El entero t indica el número de dígitos en la base β que se usan para representar el número de punto flotante, y es llamado **precisión**. Por lo general $t=6$ o $t=7$ con $\beta=10$ (precisión sencilla), $t=14$ o $t=15$ con $\beta=10$ (doble precisión). En algunos computadores se pueden hacer representaciones en precisión sencilla, doble precisión e incluso en precisión mayor.

e es un entero llamado el **exponente**, y es tal que $L\le e\le U$ para ciertos enteros L y U; es común encontrar $L=-U$ o $L=-U\pm1$. Un caso frecuente es $L=-63$ y $U=64$, para un total de 128 posibles exponentes.

El número cero requiere una representación especial.

De acuerdo con lo anterior un conjunto de punto flotante $\mathbf{F}$ queda caracterizado por cuatro parámetros:

a) La base β,
b) La precisión t,
c) Los enteros L y U tales que $L\le e\le U$, donde e es el exponente.

Cualesquiera sean los parámetros elegidos, los conjuntos de punto flotante correspondientes comparten las mismas características cualitativas, entre ellas la carencia de algunas de las propiedades algebraicas de que gozan los números reales.

Una de las características de todo conjunto de punto flotante $\mathbf{F}$ es que es **finito** y tiene

$$2(\beta-1)\beta^{t-1}(U-L+1)+1$$

números diferentes (incluyendo el cero), y donde los distintos de cero están en forma normalizada. En efecto:

a_1 puede tomar $\beta-1$ valores y a_i, $i=2,3,...,t$ puede tomar β valores, así que hay $(\beta-1)\underbrace{\beta\ ...\ \beta}_{t-1}=(\beta-1)\beta^{t-1}$ fracciones positivas distintas.

Ahora, considerando que el número de posibles exponentes es $U-L+1$, que el número de punto flotante puede ser positivo o negativo, y teniendo en cuenta que el número cero está también en el conjunto de punto flotante, concluímos que el conjunto $\mathbf{F}$ tiene

$$2(\beta-1)\beta^{t-1}(U-L+1)+1$$

números diferentes.

Lo anterior nos dice que se usan $2(\beta-1)\beta^{t-1}(U-L+1)+1$ números de punto flotante para "representar" el conjunto continuo de los números reales (que es infinito), lo que implica que muchos números reales tendrían que ser representados por un mismo número de punto flotante.

Como ejemplo, consideremos el conjunto de punto flotante $\mathbf{F}$ con parámetros $\beta=2$ (Binario), $t=3$, $L=-1$, $U=2$. Tal conjunto $\mathbf{F}$ tiene

$$2(2-1)2^{3-1}(2-(-1)+1)+1=33$$

números diferentes (incluyendo el cero).

Los números de $\mathbf{F}$, distintos de cero, son de la forma

$$\pm(.a_1a_2a_3)_2\times 2^e$$

con $a_1=1$, a_2, $a_3=0$, 1 y $e=-1$, 0, 1, 2 ; así que las fracciones positivas distintas son:

$$(.100)_2=\frac{1}{2}+\frac{0}{2^2}+\frac{0}{2^3}=\frac{1}{2}=\frac{8}{16}$$

$$(.101)_2=\frac{1}{2}+\frac{0}{2^2}+\frac{1}{2^3}=\frac{5}{8}=\frac{10}{16}$$

$$(.110)_2=\frac{1}{2}+\frac{1}{2^2}+\frac{0}{2^3}=\frac{3}{4}=\frac{12}{16}$$

$$(.111)_2=\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}=\frac{7}{8}=\frac{14}{16}$$

Combinando estas mantisas con los exponentes, obtenemos todos los números positivos de $\mathbf{F}$ que aparecen en la TABLA 1.2 siguiente.

MANTISA	EXP. -1	EXP. 0	EXP. 1	EXP. 2
$(.100)_2=\frac{8}{16}$	$(.100)_2\times 2^{-1}=\frac{4}{16}$	$(.100)_2\times 2^{0}=\frac{8}{16}$	$(.100)_2\times 2^{1}=\frac{16}{16}$	$(.100)_2\times 2^{2}=\frac{32}{16}$
$(.101)_2=\frac{10}{16}$	$(.101)_2\times 2^{-1}=\frac{5}{16}$	$(.101)_2\times 2^{0}=\frac{10}{16}$	$(.101)_2\times 2^{1}=\frac{20}{16}$	$(.101)_2\times 2^{2}=\frac{40}{16}$
$(.110)_2=\frac{12}{16}$	$(.110)_2\times 2^{-1}=\frac{6}{16}$	$(.110)_2\times 2^{0}=\frac{12}{16}$	$(.110)_2\times 2^{1}=\frac{24}{16}$	$(.110)_2\times 2^{2}=\frac{48}{16}$
$(.111)_2=\frac{14}{16}$	$(.111)_2\times 2^{-1}=\frac{7}{16}$	$(.111)_2\times 2^{0}=\frac{14}{16}$	$(.111)_2\times 2^{1}=\frac{28}{16}$	$(.111)_2\times 2^{2}=\frac{56}{16}$

TABLA 1.2

Como estamos más familiarizados con los números decimales (en base $\beta = 10$), los 33 elementos de $\mathbf{F}$ en forma (racional) decimal son

$$0,\ \pm\frac{4}{16},\ \pm\frac{5}{16},\ \pm\frac{6}{16},\ \pm\frac{7}{16},\ \pm\frac{8}{16},\ \pm\frac{10}{16},\ \pm\frac{12}{16},\ \pm\frac{14}{16},$$

$$\pm\frac{16}{16},\ \pm\frac{20}{16},\ \pm\frac{24}{16},\ \pm\frac{28}{16},\ \pm\frac{32}{16},\ \pm\frac{40}{16},\ \pm\frac{48}{16},\ \pm\frac{56}{16}.$$

Una representación de los números positivos y el cero de $\mathbf{F}$ en la recta real se muestra en la FIGURA 1.1 siguiente.

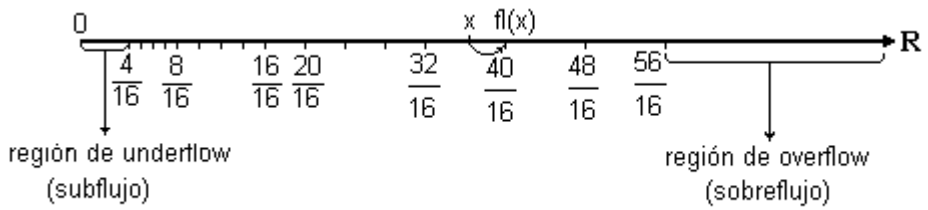

FIGURA 1.1

Algunos hechos que se pueden observar en un conjunto de punto flotante $\mathbf{F}$ son:

1. Todo número real x que entra en el computador o que es el resultado de un cálculo, es reemplazado (si es posible) por un número de punto flotante que notaremos fl(x). Existen reglas para escoger tal número (**reglas de redondeo**), por lo general es el número de punto flotante más cercano a x. La diferencia $\left| x - \text{fl}(x) \right|$ se llama **error (absoluto) de redondeo**.

2. Si observamos la distribución de los elementos de $\mathbf{F}$, en la recta real, vemos que no están igualmente espaciados (están más densamente distribuídos el la cercanía del cero), lo que implica que el error de redondeo puede depender del tamaño del número (entre más grande sea el número en valor absoluto, mayor puede ser el error de redondeo).

En el ejemplo, el número de punto flotante positivo más pequeño es $\frac{4}{16} = \frac{1}{4}$, y el número de punto flotante positivo más grande es $\frac{56}{16} = \frac{7}{2}$.

En general, en un conjunto de punto flotante $\mathbf{F}$ con parámetros β, t, L y U, se tiene que

$$\boxed{F_L = (.100\ldots0)_\beta \times \beta^L = \beta^{L-1}}$$

es el número de punto flotante positivo más pequeño (para el ejemplo, $F_L = 2^{-1-1} = \frac{1}{4}$), y

$$\boxed{F_U = (.\gamma\gamma\ldots\gamma)_\beta \times \beta^U = \left(1-\beta^{-t}\right)\beta^U} \quad \text{con } \gamma = \beta - 1$$

es el número de punto flotante positivo más grande (para el ejemplo, $F_U = \left(1 - 2^{-3}\right)2^2 = \frac{7}{2}$)

A la región $R_L = \left\{ x \in \mathbf{R} \,/\, 0 < |x| < F_L \right\}$ se le llama **región de underflow o subflujo**, y en algunos computadores si un número real cae en esta región, el número es redondeado a cero.

Por otra parte, a la región $R_U = \left\{ x \in \mathbf{R} \,/\, |x| > F_U \right\}$, se le llama **región de overflow o sobreflujo**, y en algunos computadores si un número real cae en esta región, el número es redondeado al número de punto flotante más cercano (F_U , $-F_U$) o se informa del fenómeno overflow.

Se define como **rango del conjunto $\mathbf{F}$**, al conjunto

$$R_F = \left\{ x \in \mathbf{R} / x = 0 \text{ o } F_L \le |x| \le F_U \right\}$$

De acuerdo con ésto, todo número de punto flotante, distinto de cero, fl(x), debe satisfacer

$$F_L \le \left| \, \text{fl}(x) \, \right| \le F_U$$

3. La combinación aritmética usual $+$, $-$, $\times$, $\div$ de dos números de punto flotante no siempre produce un número de punto flotante.

Supongamos que $\text{fl}(x), \text{fl}(y) \in \mathbf{F}$. Veamos, como ejemplo, que la suma usual $\text{fl}(x) + \text{fl}(y)$ no necesariamente será un número en $\mathbf{F}$. Para ello consideremos el conjunto de punto flotante $\mathbf{F}$ dado en el ejemplo: $\text{fl}(x) = \frac{28}{16} \in \mathbf{F}$, $\text{fl}(y) = \frac{5}{16} \in \mathbf{F}$, sin embargo $\text{fl}(x) + \text{fl}(y) = \frac{28}{16} + \frac{5}{16} = \frac{33}{16} \notin \mathbf{F}$. Luego la adición usual no es cerrada en el sentido matemático ordinario.

Una manera de simular la adición y las demás operaciones aritméticas entre números reales, pero realizadas por el computador es la siguiente:

Si x e y son números reales en el rango de $\mathbf{F}$, definimos las operaciones $\oplus$, $\ominus$, $\otimes$ y $\oslash$, a las que nos referiremos como **operaciones de punto flotante**, así

$$x \oplus y = \text{fl}\left(\text{fl}(x) + \text{fl}(y)\right)$$
$$x \ominus y = \text{fl}\left(\text{fl}(x) - \text{fl}(y)\right)$$
$$x \otimes y = \text{fl}\left(\text{fl}(x) \times \text{fl}(y)\right)$$
$$x \oslash y = \text{fl}\left(\text{fl}(x) \div \text{fl}(y)\right)$$

donde $+$, $-$, $\times$ y $\div$ son las operaciones aritméticas usuales.

Ilustraremos estas operaciones en el conjunto $\mathbf{F}$ del ejemplo, al tiempo que pondremos de manifiesto la carencia de ciertas propiedades para tales operaciones. Supondremos que fl(x) se escoge como el número de punto flotante más cercano a x y que cuando el número x equidista de dos números de punto flotante, se escoge fl(x) como el más cercano a la derecha si es positivo o el más cercano hacia la izquierda si es negativo:

Tomemos en $\mathbf{F}$, los números $\frac{28}{16}$ y $\frac{5}{16}$ y supongamos que $x,y \in \mathbf{R}$ son tales que $\text{fl}(x)=\frac{28}{16}$ y $\text{fl}(y)=\frac{5}{16}$. Entonces

$$x \oplus y = \text{fl}\left(\frac{28}{16}+\frac{5}{16}\right)=\text{fl}\left(\frac{33}{16}\right)=\frac{32}{16}$$

$$x \ominus y = \text{fl}\left(\frac{28}{16}-\frac{5}{16}\right)=\text{fl}\left(\frac{23}{16}\right)=\frac{24}{16}$$

$$x \otimes y = \text{fl}\left(\frac{28}{16}\times\frac{5}{16}\right)=\text{fl}\left(\frac{35}{64}\right)=\frac{32}{64}=\frac{8}{16}$$

$$x \oplus y = \text{fl}\left(\frac{28}{16}\div\frac{5}{16}\right)=\text{fl}\left(\frac{28}{5}\right)=\frac{56}{16} \quad \text{(fenómeno overflow)}$$

$\uparrow$ Overflow, ya que $\frac{28}{5} > \frac{56}{16} = F_U$

Tomemos $\frac{6}{16} \in \mathbf{F}$ y supongamos que $z \in \mathbf{R}$ es tal que $\text{fl}(z)=\frac{6}{16}$, entonces

$$z \ominus y = \text{fl}\left(\frac{6}{16}-\frac{5}{16}\right)=\text{fl}\left(\frac{1}{16}\right)=0 \quad \text{(fenómeno underflow)}$$

$\uparrow$ Underflow, ya que $0 < \frac{1}{16} < \frac{4}{16} = F_L$

Como $1=\frac{16}{16}$, $\frac{7}{8}=\frac{14}{16}$, $\frac{5}{8}=\frac{10}{16} \in \mathbf{F}$, entonces existen $u,v,w \in \mathbf{R}$ tales que $\text{fl}(u)=1$, $\text{fl}(v)=\frac{7}{8}$ y $\text{fl}(w)=\frac{5}{8}$. Entonces

$$u \oplus (v \ominus w) = \text{fl}\left(1+\text{fl}\left(\frac{7}{8}-\frac{5}{8}\right)\right)=\text{fl}\left(1+\text{fl}\left(\frac{2}{8}\right)\right)$$

$$=\text{fl}\left(1+\frac{2}{8}\right)=\text{fl}\left(\frac{10}{8}\right)=\frac{10}{8}=\frac{20}{16}$$

$$(u \oplus v) \ominus w = fl\left(fl\left(1+\frac{7}{8}\right)-\frac{5}{8}\right) = fl\left(fl\left(\frac{15}{8}\right)-\frac{5}{8}\right)$$
$$= fl\left(\frac{16}{8}-\frac{5}{8}\right) = fl\left(\frac{11}{8}\right) = \frac{24}{16}$$

luego

$$u \oplus (v \ominus w) \neq (u \oplus v) \ominus w$$

Análogamente, como $3 \in \mathbf{F}$, entonces existe $r \in \mathbf{R}$ tal que $fl(r) = 3$ y se tiene que

$$r \otimes (y \otimes x) = fl\left(3 \times fl\left(\frac{5}{16} \times \frac{28}{16}\right)\right) = fl\left(3 \times fl\left(\frac{35}{64}\right)\right)$$
$$= fl\left(3 \times \frac{32}{64}\right) = fl\left(\frac{96}{64}\right) = fl\left(\frac{24}{16}\right) = \frac{24}{16}$$

$$(r \otimes y) \otimes x = fl\left(fl\left(3 \times \frac{5}{16}\right) \times \frac{28}{16}\right) = fl\left(fl\left(\frac{15}{16}\right) \times \left(\frac{28}{16}\right)\right)$$
$$= fl\left(\frac{16}{16} \times \frac{28}{16}\right) = fl\left(\frac{28}{16}\right) = \frac{28}{16}$$

Así que,

$$r \otimes (y \otimes x) \neq (r \otimes y) \otimes x$$

Finalmente, como $\frac{1}{4} \in \mathbf{F}$, existe $s \in \mathbf{R}$ tal que $fl(s) = \frac{1}{4}$ y

$$r \otimes (v \ominus s) = fl\left(3 \times fl\left(\frac{7}{8}-\frac{1}{4}\right)\right) = fl\left(3 \times fl\left(\frac{5}{8}\right)\right)$$
$$= fl\left(3 \times \frac{5}{8}\right) = fl\left(\frac{15}{8}\right) = \frac{32}{16}$$

Como

$$r \otimes v = fl\left(3 \times \frac{7}{8}\right) = fl\left(\frac{21}{8}\right) = \frac{20}{8}$$

y

$$r \otimes s = fl\left(3 \times \frac{1}{4}\right) = fl\left(\frac{3}{4}\right) = \frac{3}{4}$$

entonces

$$(r \otimes v) \ominus (r \otimes s) = fl\left(\frac{20}{8}-\frac{3}{4}\right) = fl\left(\frac{14}{8}\right) = \frac{28}{16}$$

Así que

$$r \otimes (v \ominus s) \neq (r \otimes v) \ominus (r \otimes s)$$

1.2 ERRORES DE REDONDEO

Sabemos que todo número real $x \neq 0$ puede escribirse en la forma decimal normalizada siguiente

$$x = \pm(.a_1a_2 \ ... \ a_ta_{t+1} \ ... \ ) \times 10^n \text{, n algún entero.}$$

Para simplificar el análisis de los errores de redondeo, supongamos que nuestro conjunto de punto flotante $\mathbf{F}$ es de t-dígitos (precisión t) en base 10 (decimal); en tal caso la forma de punto flotante (normalizada) de x, $fl(x)$, se obtiene finalizando la mantisa de x después de t-dígitos. Se acostumbran dos formas para hacerlo:

i. Cortando o truncando el número x: En este caso

$$fl(x) = \pm(.a_1a_2...a_t) \times 10^n \text{, (no importa como sea } a_{t+1})$$

ii. Redondeando el número x: En este caso

$$fl(x) = \begin{cases} \pm(.a_1a_2...a_t) \times 10^n, & \text{si } \ 0 \le a_{t+1} < 5 \\ \pm(.a_1a_2...a_t) \times 10^n \pm 1.0 \times 10^{n-t}, & \text{si } \ a_{t+1} \ge 5 \end{cases}$$

El error $|x - fl(x)|$ que resulta al reemplazar un número x por su representante de punto flotante, fl(x), se seguirá denominando **error de redondeo**, independientemente de que se use el método de cortado o de redondeo.

Ejemplo 1.1 Supongamos $t = 5$ y usemos las reglas de redondeo y cortado para encontrar el representante de punto flotante decimal en cada uno de los siguientes casos:

a) $e = 2.718281828...$ (irracional)

$= (.2718281828...) \times 10^1$ (forma decimal normalizada)

Entonces

$$fl(e) = \begin{cases} (.27182) \times 10^1, & \text{cortando} \\ (.27183) \times 10^1, & \text{redondeando} \quad (\text{ya que } a_6 = 8 > 5) \end{cases}$$

b) $\pi = 3.141592653...$ (irracional)

$= (.3141592653...) \times 10^1$ (forma decimal normalizada)

Entonces

$$\text{fl}(\pi) = \begin{cases} (.31415) \times 10^1, & \text{cortando} \\ (.31416) \times 10^1, & \text{redondeando} \end{cases}$$

c) $x = -123456789$ (racional)

$= -(.123456789) \times 10^9$ (forma decimal normalizada)

Entonces

$$\text{fl}(x) = \begin{cases} -(.12345) \times 10^9, & \text{cortando} \\ -(.12346) \times 10^9, & \text{redondeando} \end{cases}$$

d) $y = .0000213475$ (racional)

$= (.213475) \times 10^{-4}$ (forma decimal normalizada)

Entonces

$$\text{fl}(y) = \begin{cases} (.21347) \times 10^{-4}, & \text{cortando} \\ (.21348) \times 10^{-4}, & \text{redondeando} \end{cases}$$

Qué pasa si se redondea el número y antes de normalizarlo?

e) $z = \frac{2}{3} = .6666666...$ (racional, periódico)

$= (.6666666...) \times 10^0$ (forma decimal normalizada)

Entonces

$$\text{fl}(z) = \begin{cases} (.66666) \times 10^0, & \text{cortando} \\ (.66667) \times 10^0, & \text{redondeando} \end{cases}$$ ♦

Cómo medir los errores de redondeo?

Hay varias formas acostumbradas para medir errores de aproximación; algunas de ellas se dan en la siguiente definición.

Definición 1.1 Sea x^* una aproximación de un número real x. El **error de** x^* **con respecto** a x es $\in = x - x^*$; el **error absoluto de** x^* **con respecto a** x es $E = \left| x - x^* \right|$ y el **error relativo de** x^* **con respecto a** x, $x \neq 0$, es $Er = \dfrac{\left| x - x^* \right|}{\left| x \right|}$. También se define el **error porcentual de** x^* **con respecto a** x, como $Er \times 100$ y se expresa en porcentaje (%). ∇

Un caso particular de aproximación de un número x es cuando $x^* = fl(x)$, y se tiene

$$E = \left| x - fl(x) \right| \quad y \quad Er = \frac{\left| x - fl(x) \right|}{\left| x \right|}, \quad x \neq 0$$

Ya vimos que el error de redondeo puede depender del tamaño del número, pues los números de punto flotante no están distribuidos de manera uniforme en la recta real; desde este punto de vista el error relativo es una mejor medida del error de redondeo que el error absoluto.

Estimemos la menor cota superior para el error relativo cuando un número real $x \neq 0$ es aproximado por su representante de punto flotante, $fl(x)$, en una aritmética decimal de t-dígitos.

Sea

$$x = (.a_1a_2...a_ta_{t+1}...) \times 10^n \text{, n algún entero,}$$

un número real positivo cualquiera en forma decimal normalizada.

Si fl(x) se obtiene por **redondeo**, tenemos:

a) Si $0 \leq a_{t+1} < 5$, entonces

$$fl(x) = (.a_1a_2...a_t) \times 10^n$$

y entonces

$$Er = \frac{\left| (.a_1a_2...a_ta_{t+1}...) \times 10^n - (.a_1a_2...a_t) \times 10^n \right|}{\left| (.a_1a_2...a_ta_{t+1}...) \times 10^n \right|}$$

$$= \frac{\left| (.a_{t+1}a_{t+2}...) \times 10^{n-t} \right|}{\left| (.a_1a_2...a_ta_{t+1}...) \times 10^n \right|}$$

$$= \frac{\left| (.a_{t+1}a_{t+2}...) \right|}{\left| (.a_1a_2...a_ta_{t+1}...) \right|} \times 10^{-t}$$

$$< \frac{.5}{.1} \times 10^{-t} = 5 \times 10^{-t}$$

b) Si $5 \le a_{t+1} \le 9$, entonces

$$fl(x) = (.a_1a_2...a_t) \times 10^n + .10 \times 10^{n-t}$$

así que

$$Er = \frac{\left| (.a_1a_2...a_ta_{t+1}...) \times 10^n - \left[(.a_1a_2...a_t) \times 10^n + .10 \times 10^{n-t} \right] \right|}{\left| (.a_1a_2...a_ta_{t+1}...) \times 10^n \right|}$$

$$= \frac{\left| (.a_{t+1}a_{t+2}...) \times 10^{n-t} - .10 \times 10^{n-t} \right|}{\left| (.a_1a_2...a_ta_{t+1}...) \times 10^n \right|}$$

$$= \frac{\left| (.a_{t+1}a_{t+2}...) - .10 \right|}{\left| (.a_1a_2...a_ta_{t+1}...) \right|} \times 10^{-t}$$

$$\le \frac{.5}{\left| (.a_1a_2...a_ta_{t+1}...) \right|} \times 10^{-t}, \text{ ya que } .a_{t+1}a_{t+2}... \ge .5$$

$$< \frac{.5}{.1} \times 10^{-t} = 5 \times 10^{-t}$$

ya que $.a_1a_2...a_ta_{t+1}... \ge .10\ ...\ 05 > .100...00$

$\uparrow$

posición $t+1$

De **a)** y **b)** se tiene que si $x \neq 0$, $x \in R_F$ y $fl(x)$ se obtiene por redondeo, entonces

$$\boxed{Er = \frac{\left| x - fl(x) \right|}{\left| x \right|} < 5 \times 10^{-t}}$$

y 5×10^{-t} es la menor cota superior para el error relativo.

Observe, en el trabajo anterior, que $E = \left| x - fl(x) \right| \le 5 \times 10^{n-(t+1)}$.

Se puede verificar que si $fl(x)$ se obtiene por cortado, entonces

$$Er = \frac{\left| x - fl(x) \right|}{\left| x \right|} < 10 \times 10^{-t} = 10^{-t+1}, \text{ y}$$

$$E = \left| x - fl(x) \right| \le 10 \times 10^{n-(t+1)}$$

Ejemplo 1.2 Encuentre el error absoluto y el error relativo de x^* con respecto a x, en cada uno de los siguientes casos:

a) $x = (.50) \times 10^2$, $x^* = (.51) \times 10^2$. Entonces

$$E = \left| (.5)\times 10^2 - (.51)\times 10^2 \right| = \left| -(.01)\times 10^2 \right| = (.1)\times 10^1 = 1.0$$

$$Er = \frac{(.1)\times 10^1}{(.5)\times 10^2} = \frac{.1}{(.5)\times 10^1} = \frac{1}{50} = (.2)\times 10^{-1} = .02 \equiv 2\%$$

b) $x = (.50)\times 10^{-3}$, $x^* = (.51)\times 10^{-3}$. Entonces

$$E = (.01)\times 10^{-3} = (.1)\times 10^{-4} = .00001$$

$$Er = \frac{(.1)\times 10^{-4}}{(.5)\times 10^{-3}} = \frac{(.1)\times 10^{-1}}{.5} = \frac{1}{50} = .02 \equiv 2\%$$

c) $x = (.50)\times 10^6$, $x^* = (.51)\times 10^6$. Entonces

$$E = (.01)\times 10^6 = (.1)\times 10^5 = 10000$$

$$Er = \frac{(.1)\times 10^5}{(.5)\times 10^6} = \frac{.1}{(.5)\times 10^1} = \frac{1}{50} = .02 \equiv 2\% \quad \blacklozenge$$

Este ejemplo nos muestra que el error relativo es **invariante al cambio de escala** y se usa como una medida de precisión o cercanía.

Teniendo en cuenta la menor cota superior para el error relativo usando redondeo, se define el concepto de cifras significativas.

Definición 1.2 Se dice que el número x^* aproxima con sus **primeros t-dígitos o cifras significativas** al número $x \neq 0$, si **t** es el **mayor entero no negativo** para el cual

$$\boxed{Er = \frac{\left| x - x^* \right|}{\left| x \right|} < 5\times 10^{-t}}$$

Los t-dígitos significativos, a que se refiere esta definición, son los primeros t-dígitos en la mantisa de x^* cuando x^* se escribe en forma decimal normalizada. ∇

De acuerdo con la definición anterior, si $x^* = fl(x)$ en una aritmética de punto flotante decimal con redondeo a t-dígitos, entonces $fl(x)$ aproxima a x con t cifras significativas, es decir, todos los dígitos en la mantisa de $fl(x)$ son significativos con respecto a x.

También se define el concepto de **cifras decimales exactas**, como sigue:

Definición 1.3 Se dice que el número x^* aproxima con sus **primeras k-cifras decimales exactas** al número x, si **k** es el **mayor entero no negativo** tal que

$$\boxed{E = \left| x - x^* \right| \leq 5 \times 10^{-(k+1)}}$$

Las k cifras decimales exactas, a que se refiere esta definición, son las primeras k cifras contadas a partir del punto decimal en x^*, cuando x^* se escribe en forma decimal. ∇

Los dos conceptos anteriores pueden aparecer definidos de manera distinta en otros textos. Aquí se usarán las definiciones dadas.

Ejemplo 1.3 Si $x = .003451$ y $x^* = .003348$, entonces

$$.00005 < \left| x - x^* \right| = .000103 < .0005 = 5 \times 10^{-4} < 5 \times 10^{-3} < 5 \times 10^{-2} < 5 \times 10^{-1}$$

así que $k = 3$ es el mayor entero no negativo tal que $\left| .003451 - .003348 \right| \leq 5 \times 10^{-(k+1)}$. Luego $.003348$ aproxima a $.003451$ con sus tres primeras cifras decimales exactas, que son en este caso 0, 0 y 3.

Observe que si $y = 28.003451$ y $y^* = 28.003348$, entonces

$$.00005 < \left| y - y^* \right| = .000103 < .0005 = 5 \times 10^{-4} < 5 \times 10^{-3} < 5 \times 10^{-2} < 5 \times 10^{-1}$$

y nuevamente, y^* aproxima a y con sus primeras tres cifras decimales exactas, que son, por supuesto, 0, 0 y 3.

Ahora, el error relativo de x^* con respecto a x es

$$.005 < Er = \frac{.000103}{.003451} = .029... < .05 = 5 \times 10^{-2} < 5 \times 10^{-1}$$

así que $t = 2$ es el mayor entero no negativo que satisface

$$\frac{\left| .003451 - .003348 \right|}{\left| .003451 \right|} < 5 \times 10^{-t}$$

y por tanto x^* aproxima a x con sus primeros **2**-dígitos significativos que son 3 y 3 (Por qué?). Con cuántas cifras significativas aproxima y^* a y? ♦

Ejemplo 1.4 Con cuántas cifras significativas aproxima $.333$ a $\frac{1}{3}$?

Como

$$\frac{\left| \frac{1}{3} - .333 \right|}{\left| \frac{1}{3} \right|} = \left| \frac{\frac{1}{3} - .333}{\frac{1}{3}} \right| = \left| 1 - .999 \right| = .001$$

y $.0005 < .001 < .005 = 5\times10^{-3} < 5\times10^{-2} < 5\times10^{-1}$, entonces $t = 3$ es el mayor entero no negativo tal que

$$\frac{\left|\frac{1}{3} - .333\right|}{\left|\frac{1}{3}\right|} < 5\times10^{-t}$$

Por lo tanto .333 aproxima a $\frac{1}{3}$ con **3** cifras significativas. Observe que .333 es el número en aritmética de punto flotante decimal con redondeo a tres dígitos que representa a $\frac{1}{3}$. ♦

Ejemplo 1.5 Dónde debe estar x^* para que aproxime a 1000 con 4 cifras significativas?

De acuerdo con la definición 1.2, x^* debe ser tal que

i) $\left|\frac{1000 - x^*}{1000}\right| < 5\times10^{-4}$, y

ii) $\left|\frac{1000 - x^*}{1000}\right| \geq 5\times10^{-5}$

La desigualdad **i)** tiene como solución $999.5 < x^* < 1000.5$ y la desigualdad **ii)** tiene como solución $x^* \leq 999.95$ o $x^* \geq 1000.05$. Interceptando las dos soluciones se obtiene que x^* debe estar en

$$(999.5,\ 999.95] \cup [1000.05,\ 1000.5) \quad \blacklozenge$$

1.3 PÉRDIDA DE CIFRAS SIGNIFICATIVAS

Sean $x = .43574628$ y $y = .43574781$. Si usamos aritmética (de punto flotante) decimal con redondeo a **6** dígitos, entonces los representantes de x y y son

$$fl(x) = .435746\ ,\ fl(y) = .435748$$

Se sabe que fl(x) y fl(y) aproximan a x e y, respectivamente, con todas sus seis cifras significativas (**Verifíquelo**).

Ahora,

$$x - y = -1.53\times10^{-6} = -.153\times10^{-5}$$

y

$$x \ominus y = fl\big(fl(x) - fl(y)\big) = fl(.435746 - .435748)$$
$$= fl\left(-2.0\times10^{-6}\right) = fl\left(-.2\times10^{-5}\right) = -.2\times10^{-5}$$

por tanto el error relativo de $x \ominus y$ con respecto a $x - y$ es

$$\frac{\left| -.2 \times 10^{-5} - \left(-.153 \times 10^{-5} \right) \right|}{\left| -.153 \times 10^{-5} \right|} = \frac{.047}{.153} = .307\ldots < .5 = 5 \times 10^{-1}$$

Luego $x \ominus y$ aproxima al valor exacto $x - y$ con únicamente una cifra significativa (1), así que hubo pérdida de **5** cifras significativas ($fl(x)$, $fl(y)$ tenían cada uno **6** cifras significativas con respecto a x e y, respectivamente); lo anterior sugiere que en un coputador debe evitarse la resta de números "casi iguales". Como **ejercicio**, revise en el mismo ejemplo, qué pasa con las operaciones $\oplus$, $\otimes$ y $\oslash$.

Ejemplo 1.6 Encontrar las raíces de la ecuación cuadrática

$$x^2 - 400.2x + 80 = 0$$

usando la fórmula usual y aritmética decimal con redondeo a **4** dígitos.

De acuerdo con la fórmula usual, las raíces son

$$x_1 = \frac{400.2 + \sqrt{(400.2)^2 - 320}}{2} \quad \text{y} \quad x_2 = \frac{400.2 - \sqrt{(400.2)^2 - 320}}{2}$$

Si hacemos los cálculos para x_1 y x_2, usando aritmética decimal con redondeo a **4** dígitos, obtenemos

$$x_1^* = \frac{400.2 + \sqrt{160200 - 320}}{2} = \frac{400.2 + \sqrt{159900}}{2} = \frac{400.2 + 399.9}{2} = \frac{800.1}{2} = 400.1$$

$$x_2^* = \frac{400.2 - \sqrt{160200 - 320}}{2} = \frac{400.2 - \sqrt{159900}}{2} = \frac{400.2 - 399.9}{2} = \frac{.3}{2} = .1500$$

Como las raíces exactas de la ecuación son $x_1 = 400.0$ y $x_2 = .2$, entonces x_1^* es una aproximación precisa (a **4** dígitos) de x_1, mientras que x_2^* es una aproximación muy pobre de x_2 (únicamente tiene una cifra significativa con respecto a x_2) .

La deficiencia en la estimación de x_2 se debe a que los números 400.2 y $\sqrt{(400.2)^2 - 320}$ son números muy cercanos entre sí (en una aritmética finita con redondeo a **4** dígitos). En este caso se consigue una aproximación más exacta para x_2, aumentando la precisión de la aritmética o "racionalizando el numerador".

Si racionalizamos el numerador, es decir, si hacemos

$$x_2 = \frac{400.2 - \sqrt{(400.2)^2 - 320}}{2} \times \frac{400.2 + \sqrt{(400.2)^2 - 320}}{400.2 + \sqrt{(400.2)^2 - 320}} = \frac{160}{400.2 + \sqrt{(400.2)^2 - 320}}$$

$$= 80\left(\frac{2}{400.2 + \sqrt{(400.2)^2 - 320}}\right) = \frac{c}{x_1}$$, donde c es el término constante en la ecuación $x^2 + bx + c = 0$, obtenemos

$$x_2^* = \frac{80}{x_1^*} = \frac{80}{400.1} = .2000$$

que coincide con el valor exacto de x_2, en este caso.

Cómo resolvería la ecuación $x^2 + 400.2x + 80 = 0$, usando aritmética decimal con redondeo a cuatro dígitos, si quiere intentar evitar la pérdida de cifras significativas en el cálculo de las raíces? ♦

Ejercicio 1.1 Elabore un programa de computador que resuelva la ecuación cuadrática general $ax^2 + bx + c = 0$ (aún en el caso de raíces complejas), usando aritmética finita y que intente evitar la pérdida de cifras significativas en el cálculo de las raíces. ♦

Ejemplo 1.7 Recordemos que para todo $x \in \mathbf{R}$

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ \ldots \ + \frac{x^n}{n!} + \ \ldots$$

Si usamos aritmética de computador para estimar e^x, a partir de la serie, sólo podremos tomar un número finito de términos; digamos que tomamos los primeros $n+1$ términos (para un cierto n), entonces

$$e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!}$$

El polinomio

$$p_n(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!}$$

se llama **polinomio de Taylor de grado n** para la función $f(x) = e^x$ en el punto $a = 0$ o también **polinomio de Maclaurin**.

Se sabe que

$$e^x = p_n(x) + R_n(x;0)$$

con

$$R_n(x;0) = e^{\xi} \frac{x^{n+1}}{(n+1)!} \text{ para algún } \xi \text{ entre 0 y x}$$

o también

$$R_n(x;0) = \frac{1}{n!}\int_0^x (x-t)^n e^t dt$$

Observe que $R_n(x;0)$ no es otra cosa que el **residuo en la serie de Taylor** cuando se toman los primeros $n+1$ términos. A $R_n(x;0)$ se le llamará **error de truncamiento o de fórmula** al aproximar la función e^x mediante el polinomio $p_n(x)$.

El error de truncamiento o de fórmula ocurre cuando un proceso matemático se interrumpe antes de su terminación.

Supongamos que queremos estimar e^{-5} y e^5 a partir del polinomio de Taylor, es decir,

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

$$e^5 = 1 + 5 + \frac{5^2}{2!} + \ldots + \frac{5^n}{n!} = p_n(5)$$

Cuál es la aproximación que se obtiene para e^{-5} y e^5, si se trabaja en una aritmética (de punto flotante) decimal con redondeo a **4** dígitos?

Las aproximaciones correspondientes a e^{-5} y e^5 aparecen en la TABLA 1.3.

De acuerdo con los resultados de la TABLA 1.3, en una aritmética decimal con redondeo a **4** dígitos, $e^{-5} \approx .9993 \times 10^{-2}$ (la suma $\sum_{k=0}^{n} \frac{(-5)^k}{k!}$ se estabilizó en $n = 22$) y $e^5 \approx 148.4$ (la suma $\sum_{k=0}^{n} \frac{5^k}{k!}$ se estabilizó en $n = 14$).

El valor exacto de e^{-5} es $6.737946999... \times 10^{-3}$ y el de e^5 es $148.4131591...$. Se observa que para e^5 todos los cuatro dígitos obtenidos en la aproximación son significativos, mientras que para e^{-5} sólo hay un dígito significativo.

A qué se debe el problema en el cálculo de e^{-5}? Se debe, entre otros, a la suma alternada (hay que evitarlas) y al hecho de que hay términos relativamente grandes con respecto al número pequeño e^{-5}, los cuales al ser sumados producen pérdida de cifras significativas.

Una forma más adecuada de calcular e^{-5} es aumentando la precisión de la aritmética o calculando $\frac{1}{e^5}$: para la aritmética de punto flotante decimal con redondeo a cuatro dígitos

$$\frac{1}{e^5} = \frac{1}{148.4} = 6.739 \times 10^{-3}$$

que es una mejor aproximación de e^{-5}.

Con cuántas cifras significativas aproxima 6.739×10^{-3} al valor exacto e^{-5}? ♦

Grado n	Término	Suma ($\sum_{k=0}^{n} \frac{(-5)^k}{k!}$)	Suma ($\sum_{k=0}^{n} \frac{5^k}{k!}$)
0	1.000	1.000	1.000
1	−5.000	−4.000	6.000
2	12.50	8.500	18.50
3	−20.83	−12.33	39.33
4	26.04	13.71	65.37
5	−26.04	−12.33	91.41
6	21.70	9.370	113.1
7	−15.50	−6.130	128.6
8	9.688	3.558	138.3
9	−5.382	−1.824	143.7
10	2.691	.8670	146.4
11	−1.223	−.3560	147.6
12	.5097	.1537	148.1
13	−.1960	$-.4230 \times 10^{-1}$	148.3
14	$.7001 \times 10^{-1}$	$.2771 \times 10^{-1}$	148.4
15	$-.2333 \times 10^{-1}$	$.4380 \times 10^{-2}$	148.4
16	$.7294 \times 10^{-2}$	$.1167 \times 10^{-1}$	
17	$-.2145 \times 10^{-2}$	$.9525 \times 10^{-2}$	
18	$.5959 \times 10^{-3}$	$.1012 \times 10^{-1}$	

19	$-.1568\times10^{-3}$	$.9963\times10^{-2}$	
20	$.3920\times10^{-4}$	$.1000\times10^{-1}$	
21	$-.9333\times10^{-5}$	$.9991\times10^{-2}$	
22	$.2121\times10^{-5}$	$.9993\times10^{-2}$	
23	$-.4611\times10^{-6}$	$.9993\times10^{-2}$	

TABLA 1.3

1.4 ESTABILIDAD DE UN ALGORITMO

Los ejemplos **1.6** y **1.7** anteriores, muestran como un algoritmo mal concebido puede conducir a una respuesta defectuosa de un problema perfectamente bien planteado. La deficiencia fue corregida cambiando el algoritmo.

Cuando al aplicar un algoritmo para resolver un problema, el efecto acumulativo de los errores, incluyendo errores de redondeo, es limitado de modo que se genera un resultado útil, el **algoritmo** se dice **estable**; en caso contrario, es decir, cuando los errores crecen de manera incontrolada de modo que se genera una respuesta defectuosa al problema, el algoritmo se dice **inestable**.

Ejemplo 1.8 Supongamos que queremos calcular

$$I_n = \int_0^1 x^n e^{x-1}dx, \quad n = 1,2,3,...$$

Una forma de proceder para estos cálculos es como se indica a continuación:

Usando integración por partes con $u = x^n$ y $dv = e^{x-1}dx$, tenemos que

$$I_n = \int_0^1 x^n e^{x-1}dx = x^n e^{x-1}\Big]_0^1 - \int_0^1 nx^{n-1}e^{x-1}dx = 1 - n\underbrace{\int_0^1 x^{n-1}e^{x-1}dx}_{I_{n-1}}$$

es decir, $I_n = 1 - nI_{n-1}$, $n = 2,3,4,...$. Luego $I_n = 1 - nI_{n-1}$, $n = 2,3,4,...$ con $I_1 = \int_0^1 xe^{x-1}dx = \frac{1}{e}$ (irracional).

Usando aritmética (de punto flotante) decimal con redondeo a **6** dígitos y la fórmula de recurrencia $I_n = 1 - nI_{n-1}$, obtenemos

$$I_1 \approx .367879 = I_1^*,\ I_2 \approx .264242 = I_2^*,\ I_3 \approx .207274 = I_3^*,\ I_4 \approx .170904 = I_4^*,$$
$$I_5 \approx .145480 = I_5^*,\ I_6 \approx .127120 = I_6^*,\ I_7 \approx .110160 = I_7^*,\ I_8 \approx .118720 = I_8^*,$$
$$I_9 \approx -.0684800 = I_9^*$$

Es claro que el valor $-.0684800\ (\approx I_9)$ es **incorrecto**, pues x^9e^{x-1} es continua y positiva sobre el intervalo $(0,1)$. Qué causó este resultado? Observe que únicamente hay error de redondeo en el cálculo de I_1, donde $\frac{1}{e}$ fue redondeado a **6** dígitos significativos. Como la fórmula de recurrencia obtenida en la integración por partes es exacta para la aritmética real, entonces no hay error de fórmula y así el error en I_9 es debido en su totalidad al error de redondeo en I_1. El error inicial fue $\in\ \approx 4.412 \times 10^{-7}$.

Al calcular I_2, tenemos

$$I_2 = 1 - 2I_1 = 1 - 2\left(I_1^* + \in\right) = 1 - 2I_1^* - 2 \in = I_2^* - 2 \in$$

entonces $I_2 - I_2^* = -2 \in$.

Ahora,

$$I_3 = 1 - 3I_2 = 1 - 3\left(I_2^* - 2 \in\right) = 1 - 3I_2^* + (-2)(-3) \in = I_3^* + (-2)(-3) \in$$

así que $I_3 - I_3^* = (-2)(-3)\in$.

Al llegar al cálculo de I_9, obtenemos

$$I_9 = I_9^* + (-2)(-3)...(-9)\in$$

es decir,

$$I_9 - I_9^* = (-2)(-3)...(-9)\in = 9!\in$$

De donde

$$I_9 - I_9^* \approx 362880\left(4.412\times 10^{-7}\right) \approx .160102656$$

El valor de I_9, con por lo menos **4** cifras decimales exactas, es

$$I_9 = -.0684800 + .160102656 = .091622656$$

Observe que el error absoluto, debido a los cálculos, crece a medida que n aumenta, y es mucho más grande que el valor real (en valor absoluto) que se está aproximando (se puede ver que si $\in$ es el error inicial, entonces el error después de **n** pasos es $\in_n = I_n - I_n^* = (-1)^{n-1} n!\in$, y $\lim_{n\to\infty} |\in_n| = \lim_{n\to\infty} \left|(-1)^{n-1} n!\in\right| = +\infty$; mientras que $0 < I_n \le \frac{1}{n+1}$). En conclusión, el **algoritmo** dado por la fórmula de recurrencia

$$I_n = 1 - nI_{n-1}, \quad n = 2,3,... \quad \text{con} \quad I_1 = \frac{1}{e}$$

es **inestable**.

Cómo podemos escoger un algoritmo diferente el cual evite esta inestabilidad?

Si reescribimos la relación de recurrencia como

$$I_{n-1} = \frac{1 - I_n}{n},\ n = N,...,3,2$$

entonces en cada paso del cálculo el error en I_n es dividido por **n**. Así que, si comenzamos con un valor para algún I_n con $n >> 1$, y trabajamos hacia atrás, cualquier error inicial o errores de redondeo que ocurran estarán decreciendo en cada paso. Este es un ejemplo de **algoritmo estable**.

Para obtener un valor inicial, notemos que

$$I_n = \int_0^1 x^n e^{x-1} dx \le \int_0^1 x^n dx = \frac{x^{n+1}}{n+1}\Big]_0^1 = \frac{1}{n+1}$$

Por lo tanto $I_n \to 0$ cuando $n \to +\infty$.

Por ejemplo, si aproximamos I_{20} por 0 y usamos el valor 0 como un valor inicial, entonces cometemos un error inicial $\in$ tal que $0 \le \in \le \frac{1}{21}$; este error es multiplicado por $\frac{1}{20}$ al calcular I_{19}, así que el error en el cálculo de I_{19}, que es $\frac{1}{20}\in$, es tal que

$$0 \le \frac{1}{20}\in \le \frac{1}{20}\frac{1}{21}$$

Procediendo de la manera anterior, el error en el cálculo de I_{15} es tal que

$$0 \le \frac{1}{16}\frac{1}{17}\ldots\frac{1}{20}\in \le \frac{1}{16}\frac{1}{17}\ldots\frac{1}{20}\frac{1}{21} \approx 2.56\times 10^{-8} < 5\times 10^{-8}$$

lo que garantiza una precisión de por lo menos 7 cifras decimales exactas de precisión para los valores calculados de $I_{15},\ldots,I_9$.

Haciendo los cálculos para $I_{20},\ldots,I_9$, obtenemos

$I_{20} \approx .0000000000$, $I_{19} \approx .05000000000$, $I_{18} \approx .0500000000$,
$I_{17} \approx .0527777778$, $I_{16} \approx .05571895425$, $I_{15} \approx .05901756536$,
$I_{14} \approx .06273216231$, $I_{13} \approx .06694770269$, $I_{12} \approx .07177325364$,
$I_{11} \approx .07735222886$, $I_{10} \approx .0838770701$, $I_9 \approx .09161229299$ ♦

Ejemplo 1.9 La sucesión $\{p_n\}_n$ con $p_n = \left(\frac{1}{3}\right)^n$, $n = 0,1,\ldots$ se puede generar de varias maneras; dos de ellas son:

i) $x_0 = 1$, $x_1 = \frac{1}{3}$, $x_n = \left(\frac{5}{6}\right)x_{n-1} - \left(\frac{1}{6}\right)x_{n-2}$, $n = 2,3,\ldots$

ii) $y_0 = 1$, $y_1 = \frac{1}{3}$, $y_n = \left(\frac{5}{3}\right)y_{n-1} - \left(\frac{4}{9}\right)y_{n-2}$, $n = 2,3,\ldots$

Veamos que, efectivamente, la sucesión definida en **i)** es igual a la sucesión $\left\{\left(\frac{1}{3}\right)^n\right\}_n$, $n = 0,1,\ldots$. En efecto:

$$x_0 = 1 = \left(\frac{1}{3}\right)^0;\ x_1 = \frac{1}{3} = \left(\frac{1}{3}\right)^1, \text{ y}$$

$$x_2 = \left(\frac{5}{6}\right)x_1 - \left(\frac{1}{6}\right)x_0 = \left(\frac{5}{6}\right)\frac{1}{3} - \left(\frac{1}{6}\right)1 = \frac{5}{18} - \frac{1}{6} = \frac{1}{9} = \left(\frac{1}{3}\right)^2.$$

Supongamos que $x_k = \left(\frac{5}{6}\right)x_{k-1} - \left(\frac{1}{6}\right)x_{k-2} = \left(\frac{1}{3}\right)^k$ para $2 \le k < n$, y veamos que $x_n = \left(\frac{5}{6}\right)x_{n-1} - \left(\frac{1}{6}\right)x_{n-2} = \left(\frac{1}{3}\right)^n$:

$$x_n = \left(\frac{5}{6}\right)\left(\frac{1}{3}\right)^{n-1} - \left(\frac{1}{6}\right)\left(\frac{1}{3}\right)^{n-2} = \left(\frac{1}{3}\right)^{n-2}\left(\frac{5}{6}\frac{1}{3} - \frac{1}{6}\right) = \left(\frac{1}{3}\right)^{n-2}\left(\frac{1}{3}\right)^2 = \left(\frac{1}{3}\right)^n$$

Luego

$$x_n = \left(\frac{5}{6}\right)x_{n-1} - \left(\frac{1}{6}\right)x_{n-2} = \left(\frac{1}{3}\right)^n, \text{ para todo } n = 0,1,...$$

Análogamente, se puede verificar que la sucesión definida en **ii)** es igual a la sucesión $\{p_n\}_n$ con $p_n = \left(\frac{1}{3}\right)^n$, $n = 0,1,...$.

Si usamos aritmética decimal con redondeo a 7 dígitos para calcular los primeros términos de las sucesiónes $\{p_n\}_n$, $\{x_n\}_n$ y $\{y_n\}_n$, se obtienen los resultados que se muestran en la TABLA 1.4 siguiente.

n	p_n^*	x_n^*	y_n^*
0	1.000000	1.000000	1.000000
1	.3333333	.3333333	.3333333
2	.1111111	.1111111	.1111111
3	$.3703704 \times 10^{-1}$	$.3703704 \times 10^{-1}$	$.3703706 \times 10^{-1}$
4	$.1234568 \times 10^{-1}$	$.1234568 \times 10^{-1}$	$.1234571 \times 10^{-1}$
5	$.4115226 \times 10^{-2}$	$.4115227 \times 10^{-2}$	$.4115268 \times 10^{-2}$
6	$.1371742 \times 10^{-2}$	$.1371743 \times 10^{-2}$	$.1371797 \times 10^{-2}$
7	$.4572474 \times 10^{-3}$	$.4572475 \times 10^{-3}$	$.4573210 \times 10^{-3}$
8	$.1524158 \times 10^{-3}$	$.1524159 \times 10^{-3}$	$.1525139 \times 10^{-3}$
9	$.5080526 \times 10^{-4}$	$.5080529 \times 10^{-4}$	$.5093613 \times 10^{-4}$
10	$.1693509 \times 10^{-4}$	$.1693510 \times 10^{-4}$	$.1710958 \times 10^{-4}$
11	$.5645029 \times 10^{-5}$	$.5645036 \times 10^{-5}$	$.5877683 \times 10^{-5}$
12	$.1881676 \times 10^{-5}$	$.1881680 \times 10^{-5}$	$.2191882 \times 10^{-5}$
13	$.6272255 \times 10^{-6}$	$.6272272 \times 10^{-6}$	$.1040833 \times 10^{-5}$
14	$.2090752 \times 10^{-6}$	$.2090760 \times 10^{-6}$	$.7605514 \times 10^{-6}$
15	$.6969172 \times 10^{-7}$	$.6969214 \times 10^{-7}$	$.8049934 \times 10^{-6}$
16	$.2323057 \times 10^{-7}$	$.2323078 \times 10^{-7}$	$.1003633 \times 10^{-5}$
17	$.7743524 \times 10^{-8}$	$.7743628 \times 10^{-8}$	$.1314947 \times 10^{-5}$
18	$.2581175 \times 10^{-8}$	$.2581227 \times 10^{-8}$	$.1745519 \times 10^{-5}$
19	$.8603916 \times 10^{-9}$	$.8604175 \times 10^{-9}$	$.2324777 \times 10^{-5}$
20	$.2867972 \times 10^{-9}$	$.2868101 \times 10^{-9}$	$.3098842 \times 10^{-5}$

TABLA 1.4

Si comparamos los valores calculados $p_{20}^* = .2867972 \times 10^{-9}$, $x_{20}^* = .286810 \times 10^{-9}$ y $y_{20}^* = .3098842 \times 10^{-5}$ con el valor exacto $\left(\frac{1}{3}\right)^{20} = \frac{1}{3486784401} = 2.8679719... \times 10^{-10}$, se puede ver que $p_{20}^* = .2867972 \times 10^{-9}$ aproxima al valor exacto con todas sus 7 cifras significativas, $x_{20}^* = .286810 \times 10^{-9}$ aproxima al valor exacto con cinco cifras significativas (de siete), mientras que $y_{20}^* = .3098842 \times 10^{-5}$ aproxima al valor exacto con ninguna cifra significativa.

Qué puede decirse de la estabilidad numérica de las fórmulas que definen las sucesiones $\{p_n\}_n$, $\{x_n\}_n$ y $\{y_n\}_n$?

Observamos que la fórmula para calcular y_n produce rápidamente pérdida de cifras significativas, mientras que la fórmulas para calcular p_n y x_n no, así que el **algoritmo** para calcular y_n **es inestable**, mientras que los **algoritmos** para calcular p_n y x_n son **estables**.

Si calculamos más términos de la sucesiones $\{p_n\}_n$, $\{x_n\}_n$ y $\{y_n\}_n$, se obtienen los resultados que se muestran en la TABLA 1.5 siguiente.

n	p_n^*	x_n^*	y_n^*
30	$.4856936 \times 10^{-14}$	$.4869553 \times 10^{-14}$	$.5502329 \times 10^{-4}$
40	$.8225264 \times 10^{-19}$	$.9457497 \times 10^{-19}$	$.9770887 \times 10^{-3}$
50	$.1392956 \times 10^{-23}$	$.1342649 \times 10^{-23}$	$.1735087 \times 10^{-1}$
60	$.2358983 \times 10^{-28}$	$.1177509 \times 10^{-25}$	$.3081121 \times 10^{0}$
70	$.3994957 \times 10^{-33}$	$.1147647 \times 10^{-28}$	$.5471369 \times 10^{1}$
80	$.6765496 \times 10^{-38}$	$.1120711 \times 10^{-31}$	$.9715907 \times 10^{2}$
90	$.1149065 \times 10^{-42}$	$.1094443 \times 10^{-34}$	$.1725324 \times 10^{4}$
100	0	$.1068791 \times 10^{-37}$	$.3063784 \times 10^{5}$

TABLA 1.5

Observe, en los cálculos de la tabla anterior, que $p_n^* \to 0$ y $x_n^* \to 0$, mientras que $y_n^* \to \infty$ (cuando $n \to \infty$), y es claro que $\lim_{n\to\infty} \left(\frac{1}{3}\right)^n = 0$.

Otra forma de estudiar la estabilidad numérica de las fórmulas definidas en este ejemplo es como sigue:

Las sucesiones definidas en **i)** y **ii)** pueden verse como ecuaciones en diferencias con condición inicial:

i)
$$\begin{cases} x_n = \left(\frac{5}{6}\right) x_{n-1} - \left(\frac{1}{6}\right) x_{n-2}, \quad n = 2,3,\ldots & (1.1.a) \\ x_0 = 1, \quad x_1 = \frac{1}{3} & (1.1.b) \end{cases}$$

ii)
$$\begin{cases} y_n = \left(\frac{5}{3}\right)y_{n-1} - \left(\frac{4}{9}\right)y_{n-2}, \quad n = 2,3,\ldots & (1.2.a) \\ y_0 = 1, \;\; y_1 = \frac{1}{3} & (1.2.b) \end{cases}$$

Se puede probar que la **solución general de la ecuación en diferencias (1.1.a)**, es

$$x_n = c_1\left(\frac{1}{3}\right)^n + c_2\left(\frac{1}{2}\right)^n$$

con c_1 y c_2 constantes arbitrarias.

Nótese que la solución general anterior es el conjunto de todas las combinaciones lineales de las soluciones particulares $\left(\frac{1}{3}\right)^n$ y $\left(\frac{1}{2}\right)^n$, de la ecuación (1.1.a). Tales soluciones particulares pueden obtenerse buscando soluciones de la forma $x_n = \lambda^n$ con $\lambda \neq 0$, para la ecuación mencionada.

Para que se satisfagan las condiciones iniciales exactas (1.1.b), $x_0 = 1$ y $x_1 = \frac{1}{3}$, deben escogerse $c_1 = 1$ y $c_2 = 0$, es decir, **la solución de la ecuación en diferencias** (1.1.a) **que satisface la condición inicial** (1.1.b) es la sucesión

$$\{p_n\}_n \;\text{ con }\; p_n = \left(\frac{1}{3}\right)^n \;,\; n = 0,1,2,\ldots$$

Si las condiciones iniciales son cambiadas por $x_0 = 1.000000$ y $x_1 = .3333333$ (redondeando las condiciones iniciales (1.1.b) a siete dígitos), entonces los valores de las constantes son ahora $c_1 = 1.0000002$ y $c_2 = -.2 \times 10^{-6}$, así que la solución de la ecuación en diferencias (1.1.a) con las nuevas condiciones es

$$x_n = 1.0000002\left(\frac{1}{3}\right)^n - .2 \times 10^{-6}\left(\frac{1}{2}\right)^n$$

y entonces al calcular $p_n = \left(\frac{1}{3}\right)^n$, mediante esta última fórmula, el error es tan solo

$$\in_n = 1.0000002\left(\frac{1}{3}\right)^n - .2 \times 10^{-6}\left(\frac{1}{2}\right)^n - \left(\frac{1}{3}\right)^n = .2 \times 10^{-6}\left(\left(\frac{1}{3}\right)^n - \left(\frac{1}{2}\right)^n\right)$$

($\in_n \to 0$ cuando $n \to \infty$ y observe que $p_n \to 0$ cuando $n \to \infty$)

En este caso el **algoritmo** se considera **estable**.

En cuanto a la ecuación en diferencias (1.2.a), tenemos que su solución general es

$$y_n = c_1\left(\frac{1}{3}\right)^n + c_2\left(\frac{4}{3}\right)^n$$

con c_1 y c_2 constantes arbitrarias.

Para que se satisfagan las condiciones iniciales (1.2.b), $y_0 = 1$ y $y_1 = \frac{1}{3}$, deben escogerse $c_1 = 1$ y $c_2 = 0$, es decir, la solución de la ecuación en diferencias (1.2.a) con condición inicial (1.2.b) es la sucesión

$$\{p_n\}_n \text{ con } p_n = \left(\frac{1}{3}\right)^n,\ n = 0,1,2,...$$

Si las condiciones iniciales son cambiadas por $y_0 = 1.000000$ y $y_1 = .3333333$ (redondeando las condiciones iniciales (1.2.b) a siete dígitos), entonces los valores de las constantes son ahora $c_1 = \frac{3.0000001}{3}$ y $c_2 = \frac{1.0 \times 10^{-7}}{3}$, es decir, la solución de la ecuación en diferencias (1.2.a) que satisface las nuevas condiciones, es

$$y_n = \frac{3.0000001}{3}\left(\frac{1}{3}\right)^n + \frac{1.0 \times 10^{-7}}{3}\left(\frac{4}{3}\right)^n$$

El error al calcular $p_n = \left(\frac{1}{3}\right)^n$, mediante esta última fórmula, es

$$\in_n = \frac{3.0000001}{3}\left(\frac{1}{3}\right)^n + \frac{1.0 \times 10^{-7}}{3}\left(\frac{4}{3}\right)^n - \left(\frac{1}{3}\right)^n = \frac{1.0 \times 10^{-7}}{3}\left(\left(\frac{1}{3}\right)^n + \left(\frac{4}{3}\right)^n\right)$$

($\in_n \to +\infty$ cuando $n \to \infty$, mientras que $p_n \to 0$ cuando $n \to \infty$)

En este caso el **algoritmo** definido por la fórmula **ii)** es **inestable**. ♦

1.5 CONDICIONAMIENTO DE UN PROBLEMA

Para ciertos problemas "buenas" respuestas no pueden ser obtenidas por cualquier algoritmo, porque el problema es sensible a errores pequeños cometidos en la representación de los datos o en la aritmética. Hay que distinguir entre algoritmos inestables y problemas sensibles a cambios pequeños en los datos.

Un **problema** se dice **bien condicionado** si pequeños cambios en los datos inducen sólo un cambio pequeño en el resultado, es decir, problemas "cercanos" tienen respuesta "cercana". El buen condicionamiento es algo inherente al problema.

Veamos un ejemplo.

Consideremos el siguiente sistema de ecuaciones lineales

$$\begin{cases} x + y = 2 \\ 10.05x + 10y = 21 \end{cases}$$

La solución exacta (única) de este sistema es $x = 20$ e $y = -18$. En este caso, el punto $(20,-18)$ es la intersección de las rectas casi paralelas:

$$L_1: x + y = 2 \text{, con pendiente } m_1 = -1.0$$

$$L_2: 10.05x + 10y = 21, \text{ con pendiente } m_2 = -1.005$$

Ahora cambiamos el coeficiente 10.05 por 10.1 (un cambio relativo de $\approx .5\%$) y consideramos el **sistema perturbado**

$$\begin{cases} x + y = 2 \\ 10.1x + 10y = 21 \end{cases}$$

La solución exacta de este sistema perturbado es $x = 10$, $y = -8$.

Se observa que un cambio pequeño en uno de los datos del problema (coeficientes y términos independientes del sistema) ha producido un gran cambio en la solución (de más de 100%). Este **problema** se dice que está **mal condicionado**. ♦

TALLER 1.

1. Convertir los siguientes números binarios a la forma decimal (equivalente decimal):

$$(.1100011)_2 \; ; \quad (.1111111)_2 \; ; \quad (1010)_2 \; ; \quad (100101)_2 \; ; \quad (1000001)_2 \; ; \quad (101.01)_2$$

2. Para los siguientes números x y x^*, con cuántas cifras decimales exactas y con cuántas cifras significativas aproxima x^* a x ?

a) $x = 451.023$, $x^* = 451.01$

b) $x = -.045113$, $x^* = -.04518$

c) $x = 23.4213$, $x^* = 23.4604$

3. Un paralelipípedo rectangular tiene lados de **3**, **4** y **5** centímetros, medidos solamente al centímetro más cercano. Determine el intervalo más pequeño en el cual debe estar el área lateral de este paralelipípedo y el intervalo más pequeño en el cual debe estar su volumen.

4. Sean (x_0, y_0) y (x_1, y_1), con $y_0 \neq y_1$, puntos dados de una cierta línea recta. Verifique que la abscisa del punto de intersección de dicha recta con el eje x, se puede calcular con cualquiera de las dos siguientes fórmulas

$$x = \frac{x_0 y_1 - x_1 y_0}{y_1 - y_0} \quad , \quad x = x_0 - \frac{(x_1 - x_0) y_0}{y_1 - y_0}$$

Use los datos $(x_0, y_0) = (1.31,\ 3.24)$, $(x_1, y_1) = (1.93,\ 4.76)$ y aritmética decimal con redondeo a tres dígitos para calcular dicha abscisa, utilizando las dos fórmulas. Cuál fórmula da el mejor resultado y por qué?

5. Considere el sistema de ecuaciones lineales

$$\begin{cases} 31.69x + 14.31y = 45.00 \\ 13.11x + 5.89y = 19.00 \end{cases}$$

Un método para resolver este sistema es multiplicar la primera ecuación por **13.11**, la segunda ecuación por **31.69** y restar las ecuaciones resultantes para obtener el valor de y; luego se multiplica la primera ecuación por **5.89**, la segunda ecuación por **14.31** y restamos las ecuaciones resultantes para obtener el valor de x.

Efectúe las operaciones indicadas usando aritmética decimal con corte a cuatro dígitos y compare los resultados obtenidos con la solución exacta del sistema. Si hay alguna diferencia en los resultados, puede explicar a qué se debe tal diferencia?

6. a) Escriba un programa que le produzca un error overflow en su computador.

b) Escriba un programa para determinar experimentalmente (no teóricamente) el número de punto flotante más pequeño y el más grande de su computador.

7. Calcule $\ln 2$ a partir de la serie de Maclaurin para la función $f(x) = \ln(x+1)$. Determine el menor número de términos en dicha serie que deben tomarse para conseguir $\ln 2$ con un error menor que 10^{-8}. Haga lo mismo para $\ln 1.5$ y $\ln 1.1$, y analice los resultados.

8. La aproximación $\operatorname{sen} x \approx x$ se usa a menudo para $|x|$ pequeño. Estime, con la ayuda del teorema de Taylor, el error de truncamiento al usar esta fórmula. Para qué rango de valores de x da esta aproximación resultados con una precisión de por lo menos seis cifras decimales exactas?

9. Sea $f(x) = e^{-x}$. Encuentre el polinomio de Taylor de tercer grado para f alrededor de $a = 1.0$, y úselo para aproximar $e^{-.99}$. Cuántas cifras decimales exactas se esperan en la aproximación calculada?

10. Discuta los problemas que se pueden presentar al evaluar las siguientes funciones y plantee alternativas que permitan evitarlos:

a) $f(x) = \ln(x+1) - \ln x$

b) $\operatorname{senh} x = \dfrac{e^x - e^{-x}}{2}$

c) $f(x) = \dfrac{1-\cos x}{x^2}$

d) $f(x) = \sqrt[3]{1+x} - 1$

11. Use aritmética decimal con redondeo a cuatro dígitos y una fórmula que intente evitar la pérdida de cifras significativas, para encontrar las raíces de cada una de las siguientes ecuaciones cuadráticas

a) $x^2 - 19.96x + .1995 = 0$

b) $x^2 + 40x - 1 = 0$

12. Considere la ecuación en diferencias

$$x_n = x_{n-1} + x_{n-2}, \quad n = 2,3,\ldots \qquad \textbf{(1)}$$

a) Verifique que la sucesión

$$x_n = \left(\frac{1+\sqrt{5}}{2}\right)^n, \quad n = 0,1,\ldots \qquad \textbf{(2)}$$

es solución de la ecuación en diferencias (1), y satisface las condiciones iniciales $x_0 = 1$ y $x_1 = \dfrac{1+\sqrt{5}}{2}$.

Utilice aritmética finita para calcular x_n, $n = 0,1,\ldots,20$ usando la fórmula (1) con las condiciones iniciales anteriores, y también usando la fórmula (2). Explique los resultados y concluya acerca de la estabilidad numérica de la fórmula (1) .

b) Verifique que la sucesión

$$x_n = \left(\frac{1-\sqrt{5}}{2}\right)^n, \quad n = 0,1,\ldots \qquad \textbf{(3)}$$

es solución de la ecuación en diferencias (1), y satisface las condiciones iniciales $x_0 = 1$ y $x_1 = \dfrac{1-\sqrt{5}}{2}$.

Utilice aritmética finita para calcular x_n, $n = 0,1,\ldots,20$ usando la fórmula (1) con las condiciones iniciales anteriores, y también usando la fórmula (3). Explique los resultados y concluya acerca de la estabilidad numérica de la fórmula (1) .

13. Considere la ecuación en diferencias

$$x_n = 2\left(x_{n-1} + x_{n-2}\right), \quad n = 2,3,\ldots$$

a) Verifique que si se dan las condiciones iniciales $x_0 = 1$ y $x_1 = 1-\sqrt{3}$, entonces $x_n = \left(1-\sqrt{3},\right)^n$, $n = 0,1,...$ es solución de la ecuación en diferencias dada y satisface las condiciones iniciales dadas.

b) Utilice aritmética finita para calcular x_n, $n = 0,1,\ldots,20$ usando tanto la fórmula $x_n = \left(1-\sqrt{3}\right)^n$, como la fórmula $x_n = 2\left(x_{n-1} + x_{n-2}\right)$, con las condiciones iniciales dadas en **a)**. Explique los resultados y concluya acerca de la estabilidad numérica de la fórmula $x_n = 2\left(x_{n-1} + x_{n-2}\right)$.

14. Las funciones de Bessel J_n satisfacen la siguiente fórmula de recurrencia

$$J_n(x) = 2(n-1)x^{-1}J_{n-1}(x) - J_{n-2}(x), \quad n = 2,3,... \qquad \textbf{(4)}$$

Empiece con $J_0(1) = .7651976866$ y $J_1(1) = .4400505857$ y use la fórmula de recurrencia anterior para calcular $J_n(1)$, $n = 2,3,...,20$. Se puede creer en los resultados obtenidos? Explique.

Nota: Se sabe que las funciones de Bessel J_n pueden definirse mediante la fórmula

$$J_n(x) = \frac{1}{\pi}\int_0^{\pi} \cos\left(x\,\mathrm{sen}\,\theta - n\theta\right) d\theta$$

15. Las funciones de Bessel Y_n satisfacen la misma fórmula de recurrencia **(4)** que las funciones de Bessel J_n . Empiece con $Y_0(1) = .0882569642$ y $Y_1(1) = -.7812128213$ y use la fórmula de recurrencia (4) para calcular $Y_n(1)$, $n = 2,3,...,20$. Decida si los resultados son confiables o no.

16. Escriba un programa de computador que calcule el valor de $S_N = \sum_{k=1}^{N} \frac{1}{k}$ para varios valores de N. Encuentre el valor de N tal que $S_n = S_N$ para todo $n \geq N$. Le parece

extraño que tal valor exista? Recuerde que la serie armónica $\sum_{k=1}^{\infty}\frac{1}{k}$ es divergente. Explique.

17. Para cualquier entero positivo N y una constante fija $r \neq 1$, se tiene la siguiente fórmula para la suma geométrica

$$G_N \equiv 1 + r + r^2 + \ldots + r^N = \frac{1 - r^{N+1}}{1 - r} \equiv Q_N$$

Escriba un programa de computador que calcule G_N y Q_N para valores arbitrarios de r y N. Si r se escoge muy cerca de **1**, los valores de G_N y Q_N pueden diferir. Cómo explica ésto? Cuál de los dos cree que es una mejor aproximación del valor exacto de la suma? Explique.

18. Defina una sucesión $\{x_n\}_n$, $n = 0,1,\ldots$ mediante la fórmula de recurrencia

$$x_{n+1} = x_n + \frac{1}{x_n}, \quad n = 0,1,\ldots \quad \text{donde} \quad x_0 > 0$$

Qué puede decir acerca de la existencia de $\lim_{n\to\infty} x_n$?