Apuntes de teórico - Métodos Numéricos

Instituto de Matemáticas y Estadística "Rafael Laguardia"
Facultad de Ingeniería, Universidad de la República
2016 - Montevideo, Uruguay

Nota importante: El presente material forma parte de una versión **en proceso de revisión** de un texto teórico para la asignatura "Métodos Numéricos". Por lo tanto, para preparar los exámenes se debe utilizar el material de las clases teóricas y la **bibliografía recomendada**.

Índice general

Capítulo 1

Errores y Representación

1.1. Aritmética de Punto Flotante y Errores

Consideremos un proceso o sistema real (**PR**) del cual se desea conocer el comportamiento de un determinado parámetro (**x**) conociendo información de base (**d**). Consideraremos dos formas de resolver problemas de este tipo:

- **Experimentación.** A través de la realización de experimentos es posible obtener valores reales del comportamiento que se desea estudiar. A pesar de esto, en muchos casos, la experimentación resulta costosa (ensayos de materiales, inspección de recursos naturales) en otros casos, se tardaría demasiado (políticas económicas, sistemas biológicos).
- **Modelamiento computacional.** utilizando herramientas de las ciencias exactas es posible formular problemas matemáticos que modelan o describen el comportamiento de dichos sistemas. A través del uso de *métodos numéricos* es posible resolver estos problemas y así, predecir el comportamiento de los sistemas, aunque obteniendo un error con respecto al proceso real en todos los casos.

En la Figura 1.1 se representan los dos métodos propuestos.

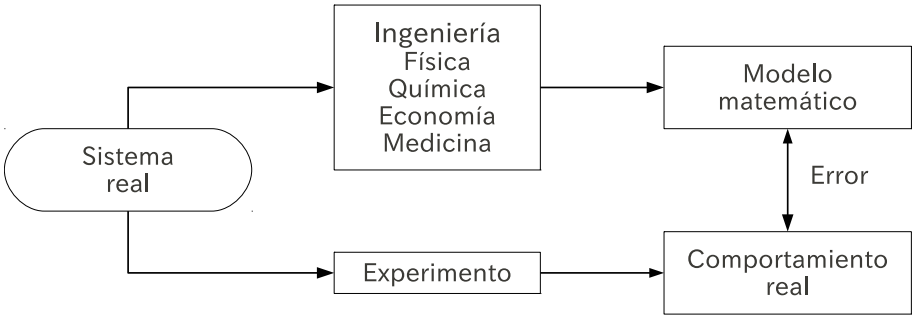

Figura 1.1: Simulación vs. Experimentación

Errores Al resolver problemas reales utilizando modelos computacionales siempre existen errores en la solución. Algunas de las fuentes de error mas habituales son las siguentes:

- Errores de medición en los datos de entrada del modelo (afectando el modelo y la calidad de las soluciones).
- Errores en el modelo matemático utilizado (considerando hipótesis que no se adapten a la realidad).
- Errores en el método numérico que resuelve el problema matemático (error de truncamiento, convergencia, etc).
- Errores en las operaciones de punto flotante (error computacional).
- Error Humano (errores en programación) y de Máquina (defectos de hardware).

En este capítulo nos centraremos en los errores de representación de punto flotante y de operaciones entre números representados en punto flotante.

1.2. Error de Operaciones de Máquina: Aritmética de Punto Flotante

Para resolver los problemas anteriormente mencionados, es necesario operar con números reales. Sea cual fuere el computador utilizado, la cantidad de números representables es finita, por lo tanto, siempre va a existir el error debido a la propia *representación* o almacenamiento del número. Al almacenar un número x podremos guardar algún número próximo a éste de los cuales la máquina pueda representar. Para poder trabajar en este tema, recordemos el Sistema Posicional de números reales.

Definición 1.2.1 (Sistema Posicional)**.** Sea una base $\beta \in \mathbb{N}$ con $\beta \geq 2$, y sea x un número real, entonces puede ser escrito como

$$x = d_n \cdot \beta^n + d_{n-1} \cdot \beta^{n-1} + \cdots + d_0 \cdot \beta^0 + d_{-1} \cdot \beta^{-1} + \ldots \qquad d_i \in \mathbb{N} \quad 0 \leq d_i \leq \beta - 1 \, i = 0 \ldots n$$

siendo esta espresión única (salvo excepciones como $0{,}999\cdots = 1$).

Particularmente utilizaremos los sistemas decimal y binario correspondientes a β igual a 10 y 2, respectivamente.

Sistema decimal Este es el sistema utilizado habitualmente para representar números. El valor de la base es 10, como se muestra a continuación:

$$5432,05 = 5 \cdot 10^3 + 4 \cdot 10^2 + 3 \cdot 10^1 + 2 \cdot 10^0 + 0 \cdot 10^{-1} + 5 \cdot 10^{-2}$$

La cantidad de bits almacenables de un computador es múltiplo de 2, por lo tanto es razonable también estudiar el sistema que utiliza base 2, el cual es llamado sistema binario.

suma (+)	producto (·)
$0+0=0$	$0\cdot 0=0$
$1+0=1$	$1\cdot 0=0$
$0+1=1$	$0\cdot 1=0$
$1+1=10$	$1\cdot 1=1$

Cuadro 1.1: Tabla de operaciones binarias

Sistema binario En este sistema de representación, el conjunto de dígitos se reduce a 0 y 1. Recordemos que se deben redefinir las operaciones suma y producto las cuales están dadas por la tabla 1.1.

Conversión de sistema de representación

A través de un ejemplo veremos cómo realizar la conversión entre distintos sistemas para representar el mismo número.

Ejemplo 1.2.1 (Conversión de sistema decimal a binario)**.** Obtendremos la representación binaria del número $176,524$.

$$(176,524)_{10} \longrightarrow (?)_2$$

parte entera: comenzamos por convertir la parte entera dividiendo por dos.

divisor	dividendo	cociente	resto
176	2	88	0
88	2	44	0
44	2	22	0
22	2	11	0
11	2	5	1
5	2	2	1
2	2	1	0

tomamos el último cociente, luego todos los restos ascendiendo hasta llegar al último resto. En este caso obtendríamos 10110000. Esto es equivalente a:

$$\begin{aligned}(176)_{10} &= 2\cdot 88 = 2^2\cdot 44 = 2^3\cdot 22 = 2^4\cdot 11\\ \ldots &= 2^4 + 2^4\cdot 10 = 2^4 + 2^5\cdot 5\\ \ldots &= 2^4 + 2^5 + 2^5\cdot 4 = 2^4 + 2^5 + 2^7\\ (176)_{10} &= (10110000)_2\end{aligned}$$

parte fraccional: ahora consideramos la parte fraccional y la multiplicamos por 2 reiteradamente, definiendo los dígitos binarios en función de que los resultados sean o no mayores que 1, como se describe en las ecuaciones siguientes:

$$(0,524)_{10} \longrightarrow (?)_2$$

$$
\begin{aligned}
0,524 \cdot 2 = 1,048 \geq 1 &\Rightarrow d_{-1} = 1 \\
0,048 \cdot 2 = 0,096 \leq 1 &\Rightarrow d_{-2} = 0 \\
0,096 \cdot 2 = 0,192 \leq 1 &\Rightarrow d_{-3} = 0 \\
0,192 \cdot 2 = 0,384 \leq 1 &\Rightarrow d_{-4} = 0 \\
0,384 \cdot 2 = 0,768 \leq 1 &\Rightarrow d_{-5} = 0 \\
0,768 \cdot 2 = 1,536 \geq 1 &\Rightarrow d_{-6} = 1 \\
&\vdots \\
(0,524)_{10} &= (0,100001...)_2
\end{aligned}
$$

Por lo tanto obtenemos la conversión del número:

$$(176,524)_{10} \longrightarrow (10110000,100001\ldots)_2$$

1.2.1. Representación de punto fijo

Los números reales con representación en **Punto fijo** en base β presentan un número fijo de decimales y están dados por la siguiente expresión:

$$\mathrm{PF}(x) = (-1)^s \cdot d_n\, d_{n-1} \ldots d_0\,,\, d_{-1}\, d_{-2} \ldots d_{-m}$$

donde:

- s: parámetro del signo, pudiendo valer 1 o 0.
- $0 \leq d_i \leq \beta - 1$.

Por tanto, si el sistema es binario, se utilizarán $n + m + 2$ bits para esta representación.

1.2.2. Representación de punto flotante

Los números reales con representación en **Punto flotante normalizado** en base β están dados por la siguiente expresión:

$$\mathrm{PF}(x) = (-1)^s \cdot 0, \underbrace{a_1\, a_2 \ldots a_p}_{\text{mantisa: } m} \cdot \beta^e = (-1)^s \cdot m \cdot \beta^{e-p}$$

donde:

- $0 \leq a_i \leq \beta - 1$.
- $L \leq e \leq U$.
- s: parámetro del signo, pudiendo valer 1 o 0.
- m: mantisa.

- e: exponente.

La normalización viene en el hecho de que $\beta^{-1} \leq m < 1$, evitando que los números tengan distintas representaciones posibles.

Si β es 2, m tendrá p bits asignados, e tendrá q bits asignados, y s tendrá 1 bit asignado, por lo que la suma de bits asignados será $N = 1 + q + p$. Al variar los bits asignados para cada uno de estos parámetros obtenemos distintos grados de precisión y rangos de números representables.

La Norma *IEEE 754* del año 1985, establece un estándar para la representación de números en punto flotante. En la misma se definen dos tipos de precisión: precisión simple (32-bits) y precisión doble (64-bits). En el cuadro 1.2 se describen ambos sistemas. La norma define también otras variantes de los mismos que no presentaremos aquí. Veamos cuál es el rango de

Precisión	N	s	p	q
simple	32	1	23	8
doble	64	1	52	11

Cuadro 1.2: tipos de precisión según *IEEE 754*

números representables normalizados para cada uno de estos sistemas de representación. Para ello analizamos cuál es el rango válido para exponente y mantisa. En el caso de precisión simple, por ejemplo, el exponente se almacena en 8 bits binarios, por lo que podemos almacenar

$$2^8 = 256 \text{ números}$$

de esta forma no es posible representar exponentes negativos, por lo que se resta 127 al número almacenado obteniendo un rango viable para representar números entre 0 y 1 fácilmente.

Para el caso del exponente tenemos:

$$E = e - (2^{q-1} - 1) = e - d \qquad d = 2^{q-1} - 1$$

se reservan $e = 00\ldots0$ y $e = 11\ldots1$
Rangos números normalizados para el exponente:

$$e_{min} = -2^{q-1} + 2 \qquad e_{max} = 2^{q-1} - 1$$

Precisión	e_{min}	e_{max}
simple	-126	127
doble	-1022	1023

Ahora vemos que para obtener el mínimo real normalizado:

$$Real_{min} = 1,0\ldots0 \cdot 2^{e_{min}}$$

Precisión	$\text{Real}_{min}(2)$	$\text{Real}_{min}(10)$
simple	$1 \cdot 2^{-126}$	$1,2 \cdot 10^{-38}$
doble	$1 \cdot 2^{-1022}$	$1,8 \cdot 10^{-308}$

y para obtener el máximo real normalizado:

$$Real_{max} = 1,11\ldots1 \cdot 2^{e_{max}}$$

Precisión	$\text{Real}_{max}(2)$	$\text{Real}_{max}(10)$
simple	$1,1\ldots1 \cdot 2^{127}$	$3,4 \cdot 10^{38}$
doble	$1,1\ldots1 \cdot 2^{1023}$	$1,8 \cdot 10^{308}$

En la figura 1.2 podemos ver un esquema de la distribución de los números reales representables normalizados próximos al número uno.

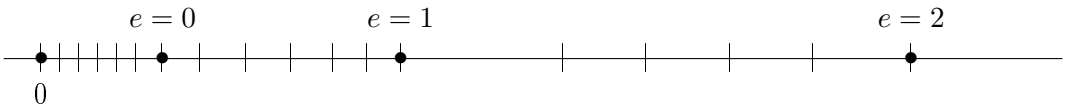

Figura 1.2: Distribución de números reales representables

Cero: se usa $e = 00\ldots0$ y mantisa nula $a_i = 0 \qquad i = 1\ldots p.$

$$0 = (s\ \underbrace{00\ldots0}_{m}\ \underbrace{00\ldots0}_{e} \rightarrow \begin{cases} s = 0 & \Rightarrow +0 \\ s = 1 & \Rightarrow -0 \end{cases}$$

Desnormalizados: se usa $e = 00\ldots0$ y mantisa no nula $\exists a_i \neq 0, \qquad i = 1\ldots p$. Son de la forma:

$$x_d = (-1)^s \cdot 0, a_1\, a_2\, \ldots\, a_p \cdot 2^{e_{min}}$$

$$Real_{min}(\text{desnormalizados}) = 0,0\ldots1 \cdot 2^{e_{min}}$$

Precisión	$\text{Real}_{min}(2)$(des.)	$\text{Real}_{min}(10)$(des.)
simple	$1 \cdot 2^{-127-23}$	$1,4 \cdot 10^{-45}$
doble	$1 \cdot 2^{-1022-52}$	$4,9 \cdot 10^{-324}$

Extiende el rango de representación próximo a cero pero con precisión limitada.

1.2.3. Aproximación de reales a punto flotante:

Redondeo y truncamiento: Como hemos visto, los sistemas de representación pueden representar una cantidad finita de números, por lo tanto al desear representar un número real que no esté incluido en el mismo, el computador deberá aproximarlo a otro. Para esta aproximación existen dos métodos habitualmente usados:

- **Redondeo:** aproxima el real al número representable más cercano. Si está equidistante, se aproxima al que tiene el dígito menos representativo igual a 0.
- **Truncamiento:** aproximo el real al número representable menor más próximo.

Límites de representación: Existen casos particulares de aproximación cuando el número a aproximar está fuera del rango abarcado por el sistema.

- **Overflow:** $x \in \mathbb{R}$ es mayor en magnitud que el mayor número representable por el sistema. Se considera como $x =$ **Inf** y es almacenado con mantisa nula, $s = 0$ y exponente $e = 11\ldots1$. Sucede lo mismo cuando si x es menor que el número mas negativo representable, y se almacena como **-Inf**.
- **Underflow:** $x \in \mathbb{R}$ es no nulo pero tiene menor magnitud que cualquier número representable (es muy próximo a cero).

1.2.4. Épsilon de máquina

Definición 1.2.2 (Épsilon de máquina)**.** Llamaremos *épsilon de máquina* (ε_{mach}) a la separación entre los números 1 y el siguiente número representable de un sistema de punto flotante.

En Octave y Matlab existe la función *eps*, la cual nos da el valor ε_{mach} para precisión doble ($2,22 \cdot 10^{-16}$) y simple ($1,19 \cdot 10^{-7}$) (ver help eps). Realice en Octave las siguientes operaciones:

```
>> a = 1 + 1.10 e-16 ;     [Enter]
>> a - 1                   [Enter]
   ans = 0                 [ a es exactamente 1 en PF]

>> a = 1 + 1.12 e-16 ;     [Enter]
>> a - 1                   [Enter]
   ans = 2.2204e-16        [ a es distinto a 1 en PF]

>> eps                     [Enter]
   ans = 2.2204e-16        [el epsilon de maquina]
```

podemos verificar de esta forma que Octave utiliza redondeo ya que para pasar de 1 al siguiente número debemos sumarle $\varepsilon_{mach}/2$.

1.3. Errores absolutos y relativos

Sea $\|.\|$ la norma euclideana en $\mathbb{R}^n$, consideremos las siguientes definiciones.

Definición 1.3.1 (Error absoluto)**.** Sean $\mathbf{x} \in \mathbb{R}^n$ un vector de valores incógnita o desconocido y $\bar{\mathbf{x}} \in \mathbb{R}^n$ una aproximación de $\mathbf{x}$. Definimos el error absoluto de $\mathbf{x}$ ($\Delta_{\mathbf{x}}$) como la norma de la diferencia entre estos valores:

$$\Delta_{\mathbf{x}} = \|\mathbf{x} - \bar{\mathbf{x}}\|$$

Definición 1.3.2 (Error relativo)**.** Sea $\mathbf{x} \in \mathbb{R}^n$ un vector de valores incógnita o desconocido y $\bar{\mathbf{x}} \in \mathbb{R}^n$ una aproximación de $\mathbf{x}$. Definimos el error relativo de $\mathbf{x}$ ($\delta_{\mathbf{x}}$) como la norma de la diferencia entre estos valores sobre la norma de $\mathbf{x}$:

$$\delta_{\mathbf{x}} = \frac{\|\mathbf{x} - \bar{\mathbf{x}}\|}{\|\mathbf{x}\|} \qquad \mathbf{x} \neq \vec{0}$$

1.4. Error de representación en Punto Flotante

Dado un valor real $x \in \mathbb{R}$ y su representación normalizada $PF(x)$ con p dígitos luego de la coma

$$x = 1, a_1 a_2 \dots a_p a_{p+1} \cdots \cdot 2^{exp} \qquad PF(x) = 1, a_1\, a_2 \dots\, a'_p \cdot 2^{exp}.$$

Para calcular el error de la aproximación de punto flotante, aplicaremos la definición de error relativo

$$\delta_x = \frac{|PF(x) - x|}{|x|} = \frac{|1, a_1\, a_2 \dots\, a_p\, a_{p+1} - 1, a_1\, a_2 \dots\, a'_p| \cdot 2^{exp}}{|1, a_1\, a_2 \dots\, a_p\, a_{p+1}| \cdot 2^{exp}}$$

simplificamos

$$\delta_x = \frac{|0, 0\, 0 \dots\, (a_p - a'_p)\, a_{p+1} \dots|}{1} \leq \varepsilon_{mach} < 2^{-p-1}$$

por lo tanto,

$$\frac{|PF(x) - x|}{|x|} \leq \varepsilon_{mach}.$$

Proposición 1.4.1. *Sea $x \in \mathbb{R}$ un real y $PF(x)$ su representación, entonces*

$$PF(x) = x(1 + \delta_x) \quad |\delta_x| \leq \varepsilon_{mach}$$

Demostración.

$$\frac{|PF(x) - x|}{|x|} = \frac{|x(1 + \delta_x) - x|}{|x|} = \frac{|x\delta_x|}{|x|} = |\delta_x| \leq \varepsilon_{mach}$$

□

1.4.1. Operaciones en punto flotante

La aritmética en punto flotante no es asociativa, ni tampoco distributiva. Las operaciones se hacen por etapas y en cada operación se aplica el redondeo correspondiente por lo tanto el resultado es alterado al cambiar los factores. Veamos un ejemplo.

Ejemplo 1.4.1 (Operaciones en Octave)**.** Realice las siguientes operaciones en Octave:

```
 >> (1 + 1.1e-16) + 1.1e-16
ans = 1

 >> 1 + (1.1e-16 + 1.1e-16)
ans = 1.0000
```

podemos concluir que en el primer caso, en el paréntesis se obtiene 1 como resultado y luego 1 nuevamente. En el segundo caso dentro del paréntesis se obtiene el épsilon de máquina por lo tanto al ser sumado a 1 se obtiene el NPF siguiente a 1.

Al analizar el costo de la ejecución de algoritmos debemos contar la cantidad de operaciones que se realizan, por lo tanto definiremos una unidad de conteo de las mismas.

Definición 1.4.1 (*flop*)**.** Denotaremos por *flop* a una simple operación de punto flotante (suma, resta, producto o división).

En el caso de un producto escalar de dos vectores de n elementos, la cantidad de *flops* es igual a $2n-1$.

1.4.2. Error al aproximar con números representables

Sean las operaciones $+, \times, -$ y $/$ para denominador no nulo, se cumple: $x \in \mathbb{R}, x = PF(x)$ $y \in \mathbb{R}, y = PF(y)$

$$PF(x \circ y) = (x \circ y)\,(1 + \delta_{x+y}) \qquad |\delta_{x+y}| \leq \varepsilon_{mach}$$

siendo $\circ$ alguna de las operaciones consideradas. Esto se debe a que las máquinas operan con más precisión que la utilizada en la representación.

1.4.3. Error al aproximar con números reales

Sean x e y dos números reales con su respectivas representaciones y errores de representación

$$x \in \mathbb{R}, \quad PF(x) = x(1+\delta_x), \quad |\delta_x| \leq \varepsilon_{mach}$$

$$y \in \mathbb{R}, \quad PF(y) = y(1+\delta_y), \quad |\delta_y| \leq \varepsilon_{mach}$$

calcularemos cual es el error cometido al operar.

Suma

$$\delta_+ = \frac{|x+y-(PF(x)+PF(y))|}{|x+y|} = \frac{|x\,\delta_x + y\,\delta_y|}{|x+y|} = \frac{x\,|\delta_x| + y\,|\delta_y|}{x+y}$$

por lo tanto

$$\delta_+ \leq \frac{x+y}{x+y}\,\varepsilon_{mach} \leq \varepsilon_{mach}$$

vemos que el error al sumar está acotado y su cota es igual a la de la representación de punto flotante.

Resta

$$\delta_- = \frac{|x-y-(PF(x)-PF(y))|}{|x-y|} = \frac{|x\,\delta_x - y\,\delta_y|}{|x-y|} = \frac{x\,|\delta_x| + y\,|\delta_y|}{x-y}$$

por lo tanto

$$\delta_- \leq \frac{x+y}{|x-y|}\,\varepsilon_{mach}$$

este error no está acotado si $x \approx y$

$$\delta_- \leq \frac{2x}{|x-y|}\,\varepsilon_{mach} \leq \frac{2\varepsilon_{mach}}{|1-y/x|}$$

Por lo tanto al realizar resta de números muy próximos podemos obtener un error grande y no detectarlo. Este fenómeno lleva el nombre de cancelación catastrófica.

Ejemplo 1.4.2 (Cancelación catastrófica)**.** Al querer calcular la solución de la ecuación

$$x^2 - 56x + 1 = 0 \quad \Rightarrow \quad r_{1,2} = 28 \pm \sqrt{783}$$

Consideremos que tenemos 5 cifras de precisión. La primer raíz se puede calcular sin error considerable.

$$r_1 = 28 + \sqrt{783} \approx 28 + 27{,}982 = 55{,}982 \pm 0{,}005 \quad (5 \text{ cifras})$$

al redondear la raíz, al calcular la segunda raíz se obtiene como resta de dos valores muy próximos

$$r_2 = 28 - \sqrt{783} = 28 - 27{,}98213 \cdots \approx 0{,}018 \pm 0{,}005 \quad (2 \text{ cifras})$$

Una forma de evitar esto es reescribir la ecuación para evitar el redondeo de la raíz y la resta

$$x^2 - 56x + 1 = (x - r_1)(x - r_2) = x^2 - (r_1 + r_2)x + r_1\, r_2$$

utilizamos el valor r_1 calculado y despejamos r_2

$$r_2 = \frac{1}{r_1} = \frac{1}{55{,}982} = 0{,}17862 \quad (5 \text{ cifras})$$

Observación 1.4.1. Conviene reescribir fórmulas para evitar problemas numéricos como cancelación catastrófica, overflow, etc.

1.5. Cálculo de derivadas por cocientes incrementales

Sea f una función real $f(x) : \mathbb{R} \longrightarrow \mathbb{R}$, de clase C^2. Recordamos la definición de la derivada primera

$$\frac{df}{dx}(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

No es posible calcular límites numéricamente, por lo tanto, es necesario estimarla. Veremos a continuación diferentes maneras de realizar esta aproximación. Para calcular numéricamente la misma, se realizan aproximaciones como por ejemplo el cociente incremental $\Delta_{f(x),h}$ para un paso $h \in \mathbb{R}^+$ pequeño.

También analizaremos los errores cometidos al llevar este problema a la computadora. Las fuentes de errores que analizaremos son dos:

- Error debido a no trabajar con precisión infinita, al que llamaremos **error de redondeo**.
- Error debido al truncamiento de la serie infinita en el desarrollo de Taylor del que se despeja la derivada, al que llamaremos **error de truncamiento**.

1.5.1. Diferencia hacia adelante

Esta aproximación consiste en aproximar la derivada de la función como el cociente incremental. Utiliza el propio punto x y $x+h$.

$$\Delta_{f(x),h} = \frac{f(x+h) - f(x)}{h}$$

A continuación calcularemos una cota superior para el error absoluto entre el valor real de la derivada y la representación en punto flotante de el cociente incremental.

$$\begin{aligned}
\text{Error}_{\text{absoluto}} &= \left| f'(x) - \frac{PF(f(x+h)) - PF(f(x))}{h} \right| \\
\dots &= \left| f'(x) + \Delta_{f(x),h} - \Delta_{f(x),h} - \frac{PF(f(x+h)) - PF(f(x))}{h} \right| \\
\dots &\leq \left| f'(x) + \Delta_{f(x),h} \right| + \left| \Delta_{f(x),h} - \frac{PF(f(x+h)) - PF(f(x))}{h} \right| \\
\text{Error}_{\text{absoluto}} &\leq \text{Error}_{\text{trunc}} + \text{Error}_{\text{PF}}
\end{aligned}$$

Este error tiene dos componentes, una debida a el truncamiento de los términos del desarrollo de Taylor, y el error de punto flotante de la propia representación.

Error de truncamiento

Planteamos Taylor próximo al punto x:

$$f(x+h) = f(x) + f'(x)\,h + f''(c)\,\frac{h^2}{2} \qquad c \in [x, x+h]$$

entonces podemos despejar el cociente incremental de paso h

$$\Delta_{f(x),h} - f'(x) = \frac{f(x+h) - f(x)}{h} - f'(x) = f''(c)\frac{h}{2}$$

por lo tanto obtenemos una buena aproximación para el error de truncamiento

$$\text{Error}_{\text{trunc.}} = \frac{|f''(c)|}{2}\,h \approx \frac{|f''(x)|}{2}\,h$$

Se concluye por tanto que el error de truncamiento es de orden h. Esto significa que menor será el error cuanto menor sea el paso h.

Error de punto flotante

$$PF(f(x+h)) = f(x+h)\,(1+\delta_h) \qquad |\delta_h| \leq \varepsilon_{mach}$$

$$PF(f(x)) = f(x)\,(1+\delta_f) \qquad |\delta_f| \leq \varepsilon_{mach}$$

$$\begin{aligned}
\text{Error}_{\text{PF}} &= \left| \frac{PF(f(x+h)) - PF(f(x))}{h} - \frac{f(x+h) - f(x)}{h} \right| \\
\dots &= \frac{|f(x+h)\,\delta_{f(x+h)} - f(x)\,\delta_{f(x)}|}{h} \\
\dots &\leq \frac{|f(x+h)|\,|\delta_{f(x+h)}| + |f(x)|\,|\delta_{f(x)}|}{h} \\
\dots &\leq \frac{|f(x+h)| + |f(x)|}{h}\,\varepsilon_{mach} \\
\text{Error}_{\text{PF}} &\leq \frac{2|f(x)|\,\varepsilon_{mach}}{h}
\end{aligned}$$

Al contrario de lo visto para el error de truncamiento, este error es inversamente proporcional al paso. Esto significa que un paso demasiado pequeño incrementa el error relacionado a la representación en punto flotante.

Es así que este análisis combina restricciones contrapuestas, ya que por un lado se requiere un paso pequeño para minimizar el error de truncamiento, pero por otro un paso demasiado pequeño complica el trabajo en punto flotante.

Por tanto, el error total combinando ambas fuentes de error toma la forma:

$$\text{Error}_{\text{total}} \leq \frac{2|f(x)|\,\varepsilon_{mach}}{h} + \frac{|f''(x)|}{2}\,h$$

Paso óptimo Dado que h puede ser elegido, es importante buscar el valor de h que minimice el error total. Para ello utilizamos la expresión del error obtenida y la derivamos para encontrar un mínimo

$$\frac{d\text{Error}}{dh} = 0$$

utilizando la expresión del error total obtenemos

$$\frac{-2|f(x)|\,\varepsilon_{mach}}{h^2} + \frac{|f''(x)|}{2} = 0$$

por lo tanto

$$h_{opt} = 2\sqrt{\frac{|f(x)|}{|f''(x)|}}\,\sqrt{\varepsilon_{mach}}$$

por ejemplo si $f''(x) \approx O(f(x))$ entonces $h_{opt} \approx \sqrt{\varepsilon_{mach}} = 10^{-8}$

1.5.2. Diferencia centrada

En esta aproximación se utilizan los puntos $x + h$ y $x - h$ y podemos ver que el error de truncamiento es de orden h^2

$$\Delta_{f(x),h} = \frac{f(x+h) - f(x-h)}{2\,h} \qquad E_{trunc} = O(h^2)$$

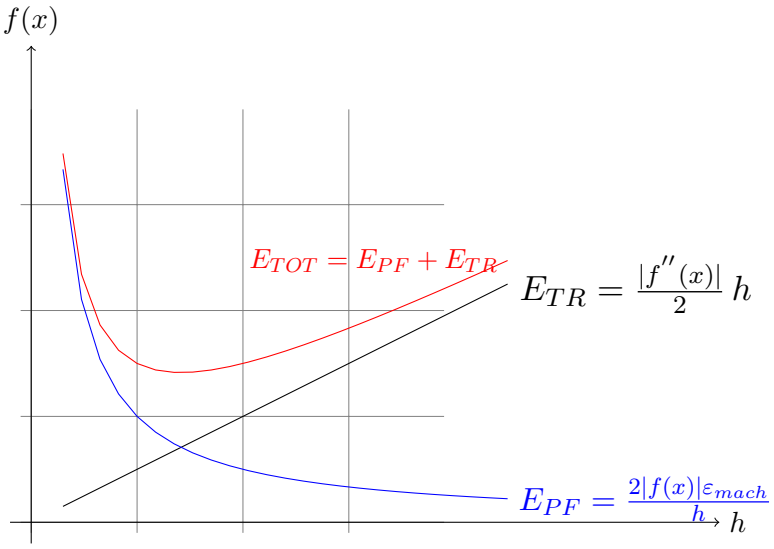

Figura 1.3: h óptimo

Aplicamos el desarrollo de Taylor en el punto x tomando como paso h y $-h$:

$$f(x+h) = f(x) + f'(x)\,h + f''(x)\,\frac{h^2}{2!} + f(c)'''\,\frac{h^3}{3!} \qquad c \in [x, x+h]$$

$$f(x-h) = f(x) - f'(x)\,h + f''(x)\,\frac{h^2}{2!} - f(d)'''\,\frac{h^3}{3!} \qquad d \in [x-h, x]$$

luego restamos miembro a miembro, obteniendo

$$f(x+h) - f(x-h) = 2h\,f'(x) + (f'''(c) + f'''(d))\,\frac{h^3}{3!}$$

por lo tanto el error de truncamiento será:

$$E_{trunc} = |\Delta f_{x,h} - f'(x)| = |f'''(c) + f'''(d)|\,\frac{h^3}{3!}\,\frac{1}{2\,h} \approx \frac{|f'''(x)|}{3!}\,h^2$$

1.5.3. Aproximación de derivada segunda

Partimos de la definición de derivada segunda

$$f''(x) = \lim_{h \to 0} \frac{f'(x+h) - f'(x)}{h}$$

y aproximamos las derivadas utilizando cocientes incrementales.

Fórmula hacia adelante usando la aproximación de diferencia hacia adelante para cada derivada, obtenemos

$$\Delta^2 f_x = \frac{\Delta f_{x+h} - \Delta f_x}{h} = \frac{f(x+2h) - f(x+h) - (f(x+h) - f(x))}{h^2}$$

obteniendo una primer fórmula

$$\Delta^2 f_x = \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2}.$$

Fórmula centrada si aproximamos la derivada $f'(x)$ utilizando diferencia hacia atrás, obtenemos la siguiente expresión

$$\Delta^2 f_x = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

Error de truncamiento
Calculemos el error de truncamiento para la fórmula centrada. Aplicamos Taylor con paso h y $-h$ y sumamos miembro a miembro:

$$\begin{aligned} f(x+h) &= f(x) + f'(x)h + f''(x)\,\tfrac{h^2}{2} + f'''(x)\,\tfrac{h^3}{3!} + f^{iv}(c)\,\tfrac{h^4}{4!} \\ &+ \\ f(x-h) &= f(x) - f'(x)h + f''(x)\,\tfrac{h^2}{2} - f'''(x)\,\tfrac{h^3}{3!} + f^{iv}(d)\,\tfrac{h^4}{4!} \\ f(x+h) + f(x-h) &= 2\,f(x) + f''(x)h^2 + f^{iv}(x)\,\tfrac{h^4}{4!} \end{aligned}$$

por lo tanto el error de truncamiento está dado por la siguiente expresión

$$\left| \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - f''(x) \right| \cong \frac{f^{iv}(x)}{4!}\,h^2$$

1.6. Extrapolación de Richardson

El método de la diferencia centrada visto en la sección 1.5.2 ilustra un abordaje interesante para reducir el error de truncamiento de la derivada.

Una generalización de lo allí realizado podría ser considerar la fórmula de la diferencia centrada pero ahora evaluada en $\frac{h}{10}$:

$$\Delta_{f(x),\frac{h}{10}} = \frac{f(x+\frac{h}{10}) - f(x-\frac{h}{10})}{2\frac{h}{10}} = f'(x) + \frac{f'''(x)}{3!}\frac{h^2}{10^2} + \frac{f^v(x)}{5!}\frac{h^4}{10^4} + \dots$$

Entonces, combinando convenientemente la expresión anterior con la de la diferencia centrada llegamos a:

$$\Delta_{f(x),h} - 100\Delta_{f(x),\frac{h}{10}} = (1-100)f'(x) + o(h^4)$$

Es decir que logramos reducir el orden de aproximación de la derivada de una función a orden h^4 (nótese que con el primer planteamiento, aproximando f' con el cociente incremental simple el orden era h):

$$\frac{\Delta_{f(x),h} - 100\Delta_{f(x),\frac{h}{10}}}{1-100} = f'(x) + o(h^4)$$

Esto bien puede generalizarse a cualquier aproximación que admita una expresión del error de truncamiento como la expansión de una serie de potencias. De esta manera, combinando cuidadosamente estas técnicas es posible mejorar el orden del error.

Sea $x \in \mathbb{R}$ un valor que se desea estimar a partir de cierta formulación $T(h)$ y que verifica la siguiente expresión:

$$T(h) = x + a_0\, h^{p_1} + O(h^{p_2}) \qquad 1 \leq p_1 < p_2$$

Se observa que es posible despejar x y obtener una aproximación de orden p_1, debido al término $a_0\, h^{p_1}$. En el caso que se desee obtener una nueva aproximación de x de mayor orden (p_2) podemos aplicar el siguiente razonamiento.

$$T(h) = x + a_0\, h^{p_1} + O(h^{p_2})$$
$$T(q\, h) = x + a_0\, q^{p_1}\, h^{p_1} + O(h^{p_2})$$

multiplico la primera por q y resto miembro a miembro

$$T(q\, h) - T(h)\, q^{p_1} = (1 - q^{p_1})\, x + O(h^{p_2})$$

logrando eliminar el término $a_o\, h^{p_1}$ y obteniendo una aproximación de x de orden p_2:

$$R(h) = \frac{T(q\, h) - T(h)\, q^{p_1}}{1 - q^{p_1}} = x + O(h^{p_2})$$

podemos reescribir la fórmula para reducir errores numéricos debido a operaciones, obteniendo la expresión general de la aproximación de Richardson:

$$R(h) = T(h) + \underbrace{\frac{T(q\, h) - T(h)}{1 - q^{p_1}}}_{\text{corrección de Richardson}} .$$

Solicitamos al lector reconocer la analogía de esta expresión general con lo que realizamos en el ejemplo de la diferencia centrada donde $T(h)$ jugaría el rol de $\Delta_{f(x),h}$; x que es el valor a estimar era en ese caso la derivada de f; $p_1 = 2$ y $p_2 = 4$; el valor de q elegido fue $1/10$; y se encontró una mejor estimación $R(h) = \frac{\Delta_{f(x),h} - 100\Delta_{f(x),\frac{h}{10}}}{1-100}$.

1.7. Propagación de errores

Dada una función $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$, $z = f(\mathbf{x})$. Es útil poder estimar cual será el error o la variación en el valor de z al tener un error o variación conocido para los valores x_i.

Fórmula 1

$x_i \in \mathbb{R} \quad i = 1, \ldots, n$ valores desconocidos exactos
$\bar{x}_i \in \mathbb{R} \quad i = 1, \ldots, n$ valores conocidos aproximados
$\bar{x}_i = x_i + \varepsilon_i \quad i = 1, \ldots, n$

De esta forma los errores absolutos en cada componente del vector $\mathbf{x}$ son:

$$\Delta_{x_i} = |\varepsilon_i| \in \mathbb{R}^+ \qquad i = 1, \ldots, n$$

Ahora queremos evaluar la función $f(\mathbf{x})$:

$$z = f(x_1, x_2, \ldots, x_n)$$

pero dado que cada componente tiene errores logramos evaluar

$$\bar{z} = f(x_1 + \varepsilon_1, \ldots, x_n + \varepsilon_n)$$

por lo que deseamos saber cual es el error que estamos comentiendo en z

$$\Delta_z = |f(x_1 + \varepsilon_1, \ldots, x_n + \varepsilon_n) - f(x_1, \ldots, x_n)|$$

Aplicando Taylor:

$$f(x_1 + \varepsilon_1, \ldots, x_n + \varepsilon_n) \approx f(x_1, \ldots, x_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) \cdot \varepsilon_i$$

por lo tanto aplicando la definición del error de z

$$\Delta_z \approx \left| \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) \cdot \varepsilon_i \right| \leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) \right| \cdot |\varepsilon_i|$$

por lo tanto obtenemos que el error absoluto de z está acotado superiormente por la suma de los errores absolutos de x_i multiplicados por la derivada parcial de f correspondiente:

$$\Delta_z \leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \cdot \Delta_{x_i}$$

Fórmula 2 En este caso consideramos el error de cada componente ε_i como variables aleatorias independientes con esperanza cero y varianza finita:

$$\mathrm{E}(\varepsilon_i) = 0 \qquad \mathrm{Var}(\varepsilon_i) = \Delta_{x_i}^2 < \infty$$

$$\Delta_z^2 = \mathrm{Var}(z) = \mathrm{Var}(f(x_1 + \varepsilon_1, \ldots, x_n + \varepsilon_n))$$

recordando Taylor:

$$f(x_1 + \varepsilon_1, \ldots, x_n + \varepsilon_n) \approx f(x_1, \ldots, x_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) \cdot \varepsilon_i$$

elevamos al cuadrado ambos lados de la igualdad y despreciamos algunos términos, obteniendo:

$$\mathrm{Var}(f(x_1+\varepsilon_1,\ldots,x_n+\varepsilon_n)) \approx \sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2 \cdot \mathrm{Var}(\varepsilon_i)$$

por lo tanto obtenemos:

$$\Delta_z = \sqrt{\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2 \cdot \Delta x_1^2}$$

1.8. Ejercicios

Ejercicio 1. Encuentre experimentalmente los siguientes valores de su calculadora, con dos cifras de precisión:

(A) El valor $\varepsilon_{\text{mach}}$ definido como el mínimo x tal que la representación en punto flotante de $1 + x$ es mayor que 1.

(B) El mayor número representable.

(C) El menor número positivo representable.

Ejercicio 2. Sea $P_n = (x_n, y_n)$, $n \in \mathbb{N}$, la sucesión generada, a partir de valores iniciales x_0, y_0, por la fórmula de recurrencia $\begin{cases} x_{n+1} = \{2x_n + y_n\} \\ y_{n+1} = \{x_n + y_n\} \end{cases}$ donde $\{u\}$ es la parte decimal de u.

(A) Muestre que si $x_0 = y_0 = \frac{1}{2}$ entonces P_n es periódica con $P_{n+3} = P_n$.

(B) Analice lo que sucede si $x_0 = y_0 = \frac{1}{3}$.

(C) Implemente un programa que calcule y grafique los primeros 100 puntos P_n de cada una de las cucesiones anteriores. Explique el resultado obtenido.

Ejercicio 3. *Errores relativo y absoluto.*

(A) Al determinar una constante C, se obtuvo el valor 92.34 con un error relativo de un 0.1 %. ¿En qué intervalo se encuentra C? ¿Cuál es el error absoluto?

(B) ¿Cuántos dígitos del número $\sqrt{22}$ deben darse para determinarlo con un error relativo no exceda el 0.1 %?

(C) En una medición se obtiene el valor $v = 17261$. Se sabe que el error relativo es del 1 %. ¿Cómo debería escribirse v para reflejar este hecho?

Ejercicio 4. *Representación interna de números.* Una computadora tiene un sistema de punto flotante decimal con 5 dígitos de precisión y 2 dígitos para el exponente. ¿Cuántos números diferentes pueden representarse con dicha arquitectura? ¿Cuáles son la menor y la mayor separación entre números representables consecutivos? Estime el valor ε_{match} (ver Ejercicio 1).

Ejercicio 5. *Cancelación catastrófica y desborde.*

(A) Se desea calcular numéricamente $\lim_{n\to\infty} \int_n^{n+1} \log(x)\, dx$. ¿Cómo puede reescribirse dicha integral para evitar efectos de cancelación catastrófica?

(B) Reescriba la expresión $\frac{e^x}{e^x+1}$ para poder evaluarla en valores grandes de x evitando efectos de desborde.

(C) Comente los inconvenientes que pueden surgir al implementar un programa para calcular la derivada de $\cos(x)$ utilizando el cociente incremental $\frac{\cos(x+h)-\cos(x)}{h}$. ¿Cómo reescribiría usted dicho cociente?

Ejercicio 6. *Errores en operaciones.*

(A) El diámetro interior de un tanque de agua esférico es de $1{,}5 \pm 0{,}05$ m. Calcule su volumen (con el error correspondiente) aproximando $\pi \simeq 3{,}1416$.

(B) Un campo rectangular mide aproximdamente 2000 por 3000 metros. ¿Con qué error deberán medirse los lados para obtener el área con un error inferior a un metro cuadrado?

Ejercicio 7. *Cálculo de la derivada con el cociente incremental.* Dada una función $f\colon I \to \mathbb{R}$ de clase C^∞, donde $I \subseteq \mathbb{R}$ es un intervalo, se desea calcular la derivada $f'(x)$ en un punto $x \in I$ usando la *diferencia hacia adelante* en x, es dicir, empleando la fórumla

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

Al aproximar numéricamente la derivada por la diferencia hacia adelante, esto es, evaluando en un h pequeño no nulo, se cometen dos tipos de errores. En primer lugar está el *error de truncamiento*, que proviene de tomar un h pequeño fijo en lugar del el límite cuando $h \to 0$, y en segundo lugar el *error de redondeo* que son los errores numéricos de la máquina, tanto en la representación como en las operaciones.

(A) Calcular la derivada de la función $f(x) = \sqrt{x}$ en el punto $x = 1$, con la diferencia hacia adelante y usando $h = 1{,}5^k$ con $k = 0, 1, \ldots, 100$. Graficar, usando escala logarítmica, el error absoluto cometido en función de h. Explicar el comportamiento observado.

(B) Usando los resultados vistos en clase sobre los errores de truncamento y de redondeo, estimar el valor de h óptimo para el calculo anterior. Coteje este valor de h con el resultado obtenido en la parte anterior.

(C) Repetir las partes (A) y (B) para la función $\tan(x)$ y el punto $x = 1{,}57$.

(D) Repetir las partes (A) y (B) para el cálculo de la derivada segunda de $f(x) = \sqrt{x}$ en el punto $x = 1$, usando la *discretización* siguiente:

$$f''(x) \simeq \frac{f(x-h) - 2f(x) + f(x+h)}{h^2}.$$

Ejercicio 8. *Extrapolación de Richardson.* Considere las aproximaciones realizadas de las diferentes derivadas en el ejercicio enterior.

(A) A partir del vector de aproximaciones correspondientes a los diferentes valores de h, use extrapolación de Richardson para hallar un nuevo vector de aproximaciones. (El nuevo vector tendrá una entrada menos.)

(B) Calcule y grafique el error cometido, comparándolo con el correspondiente a las aproximaciones originales.

(C) Repita el procedimiento, extrapolando el último vector hallado.

Ejercicio 9. Se desea hallar las cuatro raíces de polinomio

$$P_4(x) = x^4 - 12x^3 + 54x^2 - 108x + 80{,}99999999999999.$$

(A) Resuelva el problema usando el comando `roots`. ¿Qué sucedió con el vector de coeficientes del polinomio?

(B) Observando que $P_4(x) = (x-3)^4 - 10^{-14}$, resuelva analíticamete el problema.

(C) Considere la ecuación $(x-3)^4 = 0$, con solución exacta $x = 3$ y la correspondiente solución del problema "perturbado" de las partes (A) y (B). Halle el error (variación) relativo de la solución.

(D) Halle la diferencia relativa en los coeficientes de la ecuación de las partes (A)-(B) y la de la parte (C). Extraiga conclusiones sobre el número de condición del problema, definido como la razón entre el error relativo en la solución y el error relativo en los datos de entrada.

Ejercicio 10. Se desea calcular los valores de la función exponencial a partir de su desarrollo en serie

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

(A) Use el programa dado para efectuar la suma anterior hasta $n = 100$, para un rango de valores de x:

(B) Investigue qué sucede con el error relativo en los resultados numéricos obtenidos. Use la función `exp` y grafique con `semilogy`. ¿Dónde se dan los peores resultados? Justifique.

(C) Piense una solución para hacer el cálculo en los valores anteriores con mejor precisión.

```
x=-20:20;
sum=ones(size(x));
t=x; n=1;
while n<100
   sum=sum+t;
   n=n+1;
   t=t.*x/n;
end
```

Ejercicio 11. (Ver Ejercicio 7) Se quiere obtener numéricamente la derivada de una función f mediante la siguiente discretización por diferencia centrada:

$$f'(x) \simeq \Delta f = \frac{f(x+h) - f(x-h)}{2h}.$$

(A) Halle una cota para el error de truncamiento debido a la discretización usada.

(B) Estime el error de redondeo debido al uso de aritmética de punto flotante.

(C) Estime el error total y el h óptimo.

(D) Usando lo anterior, estime $f'(x)$ y su error para $f(x) = e^x$ en $x = 0$. Compare Δf con $f'(x)$ para diferentes valores de h (p.ej. $h = 10^{-n}$, $n = 1, 2, \ldots$) y obtenga una gráfica experimental que verifique el h óptimo obtenido.

(E) Repita lo anterior para $f(x) = \text{sen}(x)$ en $x = 0$ y explique los resultados.

Ejercicio 12. En versiones anteriores de *Octave* se generaba un error al calcular $\text{arcsenh}(x)$ para valores negativos grandes. El objetivo de este ejercicio es analizar el problema y proponer una solución.

(A) Calcule $\text{arcsenh}(-10^{30})$ de dos formas: usando la fórmula $\text{arcsenh}(x) = \log(x + \sqrt{x^2 + 1})$ y utilizando la función `asinh`.

(B) Explique el resultado obtenido y proponga una forma de solucionarlo.

Ejercicio 13. *Propagación del error de redondeo.*

(A) Suponga que se conoce una cantidad $x > 0$ con error absoluto δx pequeño en relación a x. Si $y = \sqrt{x}$ estime el error absoluto δy en base a x y δx. Estime también el error relativo R_y en y en base x y al error relativo en x: $R_x = \delta x/x$.

(B) Si las cantidades $x_1, x_2 > 0$ se conocen con error δx, halle una cota al error absoluto en la cantidad $z = \sqrt{x_1} - \sqrt{x_2}$.

(C) Si $x_1 = 9 \times 10^{14} + 1$ y $x_2 = 9 \times 10^{14} - 1$, calcule en la computadora el valor z, llamando z_1 al resultado obtenido. Halle el error $\delta z = z - z_1$ cometido tomando como verdadero valor de z el resultado del cálculo $z = \frac{x_1 - x_2}{\sqrt{x_1} + \sqrt{x_2}}$.

(D) Compare δz con la cota obtenida en (B) suponiendo que el error se debe sólo a la propagación del error cometido al almacenar x_1 y x_2 en punto flotante.

(E) ¿Qué relación hay entre el error relativo inicial (en x_1 y x_2) y el final (en z)?