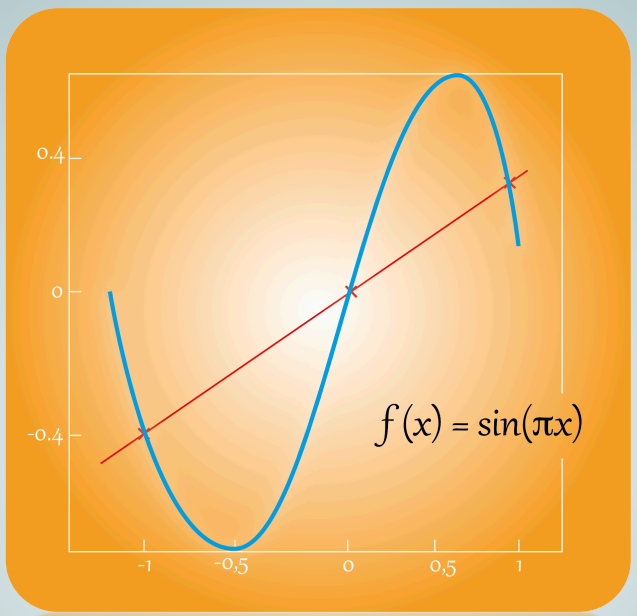
Métodos Numéricos

Capítulo 1. Preliminares matemáticos



Carlos Beltrán Álvarez

Departamento de Matemáticas, Estadística y Computación

FACULTAD DE CIENCIAS

UNIVERSIDAD DE CANTABRIA

APUNTES PARA LA ASIGNATURA:

Métodos Numéricos
Tercero de Grado en Física

Carlos Beltrán
Profesor Titular de Análisis Matemático

Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria

Índice general

Capítulo 1

Preliminares matemáticos

Una de las afirmaciones más famosas de la Historia, tanto por su lucidez como por su fundamental aportación al cambio de paradigma en el conocimiento humano, se la debemos al padre de la Ciencia moderna, Galileo Galilei, en su obra maestra "Il Saggiatore":

> La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' qua li è scritto. Egli è scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezi è impossibile a intenderne umanamente parola; senza questi è un aggirarsi vanamente per un oscuro laberinto.

Siguiendo el consejo de este sabio, para evitar perdernos en el oscuro laberinto en el que vivimos, comenzamos estos apuntes con un repaso de algunos resultados (más o menos) elementales de cálculo y análisis numérico.

1.1. Límites de funciones y sucesiones, continuidad y derivabilidad

Comenzamos recordando la definición de límite. Para ello es útil recordar que un número real x_0 se dice que es un *punto de acumulación* de un conjunto $X \subseteq \mathbb{R}^n$ si para cada $\epsilon > 0$ existe un $x \in X \setminus \{x_0\}$ tal que $\|x - x_0\| < \epsilon$. Esto es, si uno puede acercarse a x_0 arbitrariamente "pisando" solo en puntos de X.

Definición 1.1.1 *Dada una función $f : X \to \mathbb{R}^m$ con $X \subseteq \mathbb{R}^n$, y dado $x_0 \in \mathbb{R}^n$ un punto de acumulación de X, decimos que f tiene límite $L \in \mathbb{R}^m$ en x_0 cuando para todo $\epsilon > 0$ existe $\delta > 0$ tal que $0 < \|x - x_0\| < \delta$ junto con $x \in X$ implica $\|f(x) - L\| < \epsilon$. Escribimos esta propiedad de varias maneras equivalentes:*

$$\lim_{x \to x_0} f(x) = L, \quad f(x) \underset{x \to x_0}{\to} L, \quad f(x) \to L \text{ cuando } x \to x_0.$$

Definición 1.1.2 *Dada una función $f : X \to \mathbb{R}^m$ con $X \subseteq \mathbb{R}^n$, y dado $x_0 \in X$ un punto de acumulación de X, decimos que f es continua en x_0 si $\lim_{x \to x_0} f(x) = f(x_0)$. Decimos que la función es continua en $A \subseteq X$ si lo es en cada punto de A. Si f es continua en todo su dominio, decimos simplemente que f es continua.*

Definición 1.1.3 *Sea $\{x_k\}_{1 \leq k \leq \infty} \subseteq \mathbb{R}^n$ una sucesión infinita de vectores reales o complejos. Decimos que la sucesión converge a x si para todo $\epsilon > 0$ existe algún $k_0 \geq 1$ tal que $k \geq k_0$ implica $\|x_k - x\| < \epsilon$. Escribimos esta propiedad de varias maneras equivalentes:*

$$\lim_{k \to \infty} x_k = x, \quad x_k \underset{k \to \infty}{\to} x, \quad x_k \underset{k}{\to} x, \quad x_k \to x \text{ cuando } k \to \infty.$$

Las dos definiciones vistas de límite (para funciones y para sucesiones) se relacionan con el siguiente resultado.

Teorema 1.1.4 *Dada una función $f : X \to \mathbb{R}^m$ con $X \subseteq \mathbb{R}^n$, y dado un punto de acumulación $x_0 \in \mathbb{R}^n$ de X, son equivalentes:*

I) $\lim_{x\to x_0} f(x) = L$

II) *Para cualquier sucesión de vectores $\{x_k\}_{1\leq k\leq\infty}$ que converja a x_0, la sucesión $\{f(x_k)\}_{1\leq k\leq\infty}$ converge a L.*

El conjunto de las funciones continuas definidas en un conjunto X se denota por $C(X)$, aunque si X es un intervalo se abusa ligeramente de la notación y se denota $C[a, b]$ o $C(a, b]$, etc.

Definición 1.1.5 (Derivada Direccional, Diferencial, Matriz Jacobiana) *Sea $f : X \to \mathbb{R}^m$ una función definida en un conjunto $X \subseteq \mathbb{R}^n$ y sea $x_0 \in X$ tal que hay un entorno abierto que contiene a x_0 y está contenido en X. La función es diferenciable (o derivable) en x_0 si existe una aplicación lineal L tal que*

$$\lim_{v\to 0} \frac{\|f(x_0 + v) - f(x_0) - Lv\|}{\|v\|} = 0.$$

En este caso, la aplicación L se llama derivada o diferencial de f en x_0 y se denota por $Df(x_0)$ con lo que tenemos

$$\begin{array}{rccc} Df(x_0): & \mathbb{R}^n & \to & \mathbb{R}^m \\ & v & \mapsto & \lim_{t\to 0} \frac{f(x_0+tv)-f(x_0)}{t} \end{array}$$

Nótese que esta aplicación manda cada vector v a la derivada direccional de f en el punto x_0 en la dirección de v. Si f es diferenciable en x_0 entonces la aplicación $Df(x)$ en bases canónicas de $\mathbb{R}^n$ y $\mathbb{R}^m$ es simplemente la matriz Jacobiana $Jf(x_0)$ cuya entrada (i, j) es $\partial f_i/\partial x_j$.
Decimos que la función es derivable en $A \subseteq X$ si lo es en cada punto de A. Si f es derivable en todo su dominio, decimos simplemente que f es derivable. En dicho caso (que técnicamente solo es posible si el dominio de f es abierto) la función $x \to Df(x)$ está definida en el mismo dominio, se llama la derivada de f, y a veces la denotamos simplemente por f'. Nótese que f' va de X al conjunto de aplicaciones lineales de $\mathbb{R}^n$ a $\mathbb{R}^m$, esto es, eligiendo bases canónicas, f' va de X al conjunto de matrices $\mathbb{R}^{m\times n} \equiv \mathbb{R}^{mn}$. Si $m = n = 1$, la derivada usual es $f'(x_0) = Jf(x_0)$. Si $X = [a, b] \subseteq \mathbb{R}$, en los extremos a, b hablamos de derivadas laterales, no estrictamente de derivadas.

Si Df es derivable, podemos considerar su derivada $D(Df)$ que se denota $D^2 f$ y se llama segunda derivada, y siguiendo este mismo método las derivadas sucesivas. A veces se usa la notación:

$$f^{(0)} = f = D^0(f), \quad f^{(1)} = f' = D(f), \quad f^{(2)} = f'' = D^2(f), \dots$$

El conjunto de funciones k veces diferenciables (esto es, tales que está bien definida $D^k(f)$) en un conjunto X con dichas k derivadas continuas se denota $C^k(X, \mathbb{R}^m)$, de forma que se tiene $C(X, \mathbb{R}^m) \equiv C^0(X, \mathbb{R}^m) \supseteq C^1(X, \mathbb{R}^m) \supseteq C^2(X, \mathbb{R}^m) \supseteq \cdots$. La intersección de todos esos conjuntos (a veces llamado como el conjunto de funciones infinitamente diferenciables, *smooth functions* en inglés) se denota $C^\infty(X, \mathbb{R}^m)$. Si $m = 1$ es habitual escribir $C^k(X)$ en lugar de $C^k(X, \mathbb{R})$.
Por el momento, esta definición tiene sentido solo cuando X es un conjunto abierto. Sin embargo, frecuentemente utilizaremos la notación $f \in C^k[a, b]$ que no está incluida en la discusión anterior por no ser $[a, b]$ abierto. Extendemos pues la definición para cualquier $X \subseteq \mathbb{R}^n$: dado $X \subseteq \mathbb{R}^n$, definimos $C^k(X, \mathbb{R}^m)$ como el conjunto de funciones definidas en X que pueden extenderse a algún conjunto abierto $U \supseteq X$ de forma que la extensión está en $C^k(U, \mathbb{R}^m)$.

Teorema 1.1.6 *Supongamos que $f : X \to \mathbb{R}^m$ con X abierto es tal que existen todas las derivadas parciales de orden k y que son continuas con respecto a x. Entonces, $f \in C^k(X, \mathbb{R}^m)$.*

En virtud del Teorema 1.1.6, la gran mayoría de las funciones con las que trabajaremos son de tipo C^∞, esto es infinitamente diferenciables. Entran en esta categoría los polinomios, las funciones trigonométricas e hiperbólicas, la exponencial, el logaritmo (definido en $(0, \infty)$), la raíz cuadrada (que pertenece a $C^\infty(0, \infty)$ y a $C[0, \infty)$), etc.
Tres de los resultados básicos más importantes del cálculo son el Teorema de los Valores Intermedios (TVI), el Teorema de los Valores Extremos (TVE) y el Teorema del Valor Medio (TVM), que resulta imprescindible conocer y diferenciar. El primero y el tercero de ellos son válidos para funciones de varias variables (el segundo, no) y los dos primeros tienen corolarios bien conocidos.

Teorema 1.1.7 (Teorema de los Valores Intermedios) *Sea $f \in C(X)$ con $X \subseteq \mathbb{R}^n$ conexo por caminos (por ejemplo, X un intervalo). Si y_1, y_2 están en la imagen de f entonces todo número entre y_1 e y_2 está también en la imagen de f. Más generalmente, la imagen de conjunto conexo por una función continua es siempre un conjunto conexo.*

Corolario 1.1.8 (Teorema de Bolzano) *Sea $f \in C(X)$ con $X \subseteq \mathbb{R}^n$ conexo por caminos (por ejemplo, X un intervalo). Si $f(a)$ y $f(b)$ tienen distinto signo, entonces existe al menos un c en X tal que $f(c) = 0$.*

Teorema 1.1.9 (Teorema del Valor Medio) *Sea $f \in C[a, b]$ y f derivable en (a, b). Entonces, existe un número $c \in (a, b)$ tal que*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Corolario 1.1.10 (Teorema de Rolle) *Sea $f \in C[a, b]$ y f derivable en (a, b). Si $f(a) = f(b)$ entonces existe $c \in (a, b)$ tal que $f'(c) = 0$.*

El Teorema de Rolle tiene una versión más fuerte que no es habitual encontrar en textos básicos de cálculo, y que se obitene aplicando el resultado básico a f, a f', y así sucesivamente hasta llegar a $f^{(k)}$:

Corolario 1.1.11 (Teorema de Rolle Generalizado) *Sea $f \in C[a, b]$ y f k veces derivable en (a, b). Si f alcanza el mismo valor en $k + 1$ puntos distintos de $[a, b]$ entonces existe $c \in (a, b)$ tal que $f^{(k)}(c) = 0$.*

Teorema 1.1.12 (Teorema de los Valores Extremos) *Sea $f \in C(X)$ con $X \subseteq \mathbb{R}^n$ compacto (esto es, cerrado y acotado). Entonces, existen $c_1, c_2 \in X$ tales que $f(c_1) \leq f(x) \leq f(c_2)$ para todo $x \in X$.*

En las condiciones del Teorema 1.1.12, es un resultado elemental estudiado en los primeros cursos de cálculo que si f es derivable en (a, b) entonces c_1 y c_2 solo pueden ser puntos en los que la derivada se anula, o bien los extremos del intervalo. Esto se demuestra fácilmente utilizando los teoremas 1.1.9 y 1.1.12.
De forma similar, si $f : X \to \mathbb{R}$ con $X \subseteq \mathbb{R}^n$ es diferenciable y si f alcanza su máximo o su mínimo en $x \in X$ entonces o bien x está en la frontera de X o bien todas las derivadas parciales se anulan en x, esto es, el gradiente de f en x es 0.

1.2. La integral de una función de una variable real

No trataremos de exponer aquí toda la teoría que lleva a la definición de la integral de una variable real (cuestión que merecería todo un curso tratándolo con rigor, y que solo llegó a ser entendida correctamente a partir de mediados del Siglo XX con la aparición de la Integral de Lebesgue). En su lugar consideraremos una definición simplificada válida únicamente para funciones continuas en $[a, b]$:

Definición 1.2.1 *Dada $f \in C[a,b]$, definimos su integral como*

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \frac{b-a}{n+1} \sum_{i=0}^{n} f\left(a + i\frac{b-a}{n}\right).$$

Esto es, dado n tomamos la media ponderada de los valores de f en los $n+1$ puntos que se obtienen al dividir $[a,b]$ en n partes iguales, y tomamos entonces el límite de esa cantidad.

Pertenece al ámbito de los cursos de integración ver que la definición que hemos dado coincide con la definición de la integral definida (de Riemann y de Lebesgue) para el caso de $f \in C[a,b]$.
Un resultado que no se suele contar en los libros de texto básicos pero que es muy útil conocer es el Teorema del Valor Medio Ponderado para Integrales (TVMPI):

Teorema 1.2.2 (Teorema del Valor Medio Ponderado para Integrales) *Sean $f, g \in C[a,b]$.Supongamos que g no cambia se signo en $[a,b]$. Entonces, existe $c \in (a,b)$ tal que*

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx.$$

Cuando $g \equiv 1$ obtenemos el conocido resultado que nos dice que si $f \in C[a,b]$ entonces f alcanza su valor promedio de f, esto es que existe $c \in (a,b)$ tal que

$$f(c) = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

Recordemos asímismo otro de los pilares fundamentales del análisis matemático:

Teorema 1.2.3 (Teorema Fundamental del Cálculo) *Sea $f : [a,b] \to \mathbb{R}$ continua. Entonces,*

$$F(x) = \int_a^x f(t)\,dt$$

es una primitiva de f, en el sentido de que $F'(x) = f(x)$ para todo $x \in (a,b)$, y de la misma manera en a, b utilizando las derivadas laterales. Es más, sea G cualquier primitiva de f en ese mismo sentido. Entonces, $F(x)$ y $G(x)$ difieren a lo más en una constante, y se tiene

$$\int_a^x f(t)\,dt = G(x) - G(a), \quad \textit{esto es} \quad G(x) = G(a) + \int_a^x f(t)\,dt.$$

La última afirmación del Teorema Fundamental del Cálculo suele llamarse Regla de Barrow.

1.3. El teorema de Taylor

Los polinomios de Taylor son una de las herramientas fundamentales para el estudio de los métodos numéricos. Recordemos el siguiente resultado.

Teorema 1.3.1 (Teorema de Taylor con Resto de Lagrange y de Cauchy) *Supongamos que $f \in C^n[a,b]$, que f^{n+1} existe en (a,b) y que $x_0 \in [a,b]$. Entonces, para cada $x \in [a,b]$ existe $\zeta_x \in (x_0, x)$ tal que*

$$f(x) = P_n(x) + R_n(x),$$

donde

$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

es el n–ésimo polinomio de Taylor para f respecto a x_0 y $R_n(x)$ (llamado error de truncamiento) es

$$R_n(x) = \frac{f^{(n+1)}(\zeta_x)}{(n+1)!}(x - x_0)^{n+1},$$

que se suele llamar resto de Lagrange. Es más, si $f \in C^{n+1}[a, b]$, podemos escribir $f(x) = P_n(x) + S_n(x)$ donde

$$S_n(x) = \int_{x_0}^{x} \frac{f^{n+1}(t)}{n!}(t - x_0)^n \, dt,$$

que se suele llamar resto de Cauchy.

Si dejamos tender n a infinito en el teorema 1.3.1, suponiendo que f es de tipo C^∞, obtenemos una expresión llamada serie de Taylor:

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$

Para muchas de las funciones habituales y para todos los x_0 en su dominio, se tiene que esta serie formal coincide punto a punto con la función original, en un intervalo abierto que contiene a x_0. Sin embargo, no siempre es cierta esta igualdad. Cuando se cumple decimos que tenemos una función analítica (o C^ω), una propiedad más restrictiva que ser C^∞, esto es $C^\omega(X) \subsetneq C^\infty(X)$.
En el caso particular $x_0 = 0$ los polinomios de Taylor se suelen llamar polinomios de Maclaurin y las series de Taylor, las series de Maclaurin.
El siguiente resultado será útil en varias ocasiones.

Teorema 1.3.2 *Sean g, u dos funciones continuas en x_0 de forma que u es derivable en $x = x_0$ y $u(x_0) = 0$. Entonces, $f(x) = g(x)u(x)$ es derivable en $x = x_0$ y $f'(x_0) = g(x_0)u'(x_0)$.*

Demostración. Usamos la definición de la derivada como límite:

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \to x_0} \frac{g(x)u(x)}{x - x_0} = g(x_0)u'(x_0),$$

luego el límite existe y vale lo que afirma el teorema.

□

1.4. Algunos comentarios genéricos sobre aritmética computacional

Estamos acostumbrados a operar utilizando las reglas habituales que hacen conmutativos a la suma y al producto, que permiten el uso de las propiedades asociativa y distributiva, que proporcionan igualdades muy útiles como $(\sqrt{2})^2 = 2$ ó $\log(2/3) = \log 2 - \log 3$, y desigualdades también útiles como $a > b, c > d \Rightarrow a + c > b + d$ teniendo muchas de estas propiedades un efecto importantísimo en nuestra capacidad de comprensión del mundo. Estas propiedades son ciertas para los números reales (muchas de ellas, también para los complejos), pero lamentablemente pueden fallar cuando utilizamos números en la calculadora o el ordenador. Por ejemplo, si

dividimos 1 por 2, el resultado lo dividimos por 2 y así sucesivamente, en un número finito de pasos el ordenador nos dará el resultado 0.
Este problema es inevitable, dado que por muy potente que sea un ordenador solo contendrá una cantidad finita de información y es imposible poder representar en él de forma general un número con una cantidad de cifras arbitrariamente grande en su expansión decimal o binaria. Por ejemplo, no existe ni existirá jamás un ordenador en el que podamos escribir el número π con toda su precisión. ¿Qué podemos esperar entonces?
El problema de cómo representar números en ordenadores y calculadoras de una forma lo más eficaz posible (esto es, de forma que las dificultades expresadas más arriba pasen desapercibidas en la mayoría de aplicaciones) ha sido discutido durante decenas de años, y cuando los ordenadores adquirieron una importancia fundamental en el desarrollo de la ciencia y en el mantenimiento de la paz mundial se hizo necesario alcanzar un estándar que pudiese ser utilizado por todos a la vez, de modo que, aunque nadie pudiese representar por ejemplo π con exactitud, sí que estuviese garantizado que todo el mundo lo representase de la misma manera, evitando así muchos errores en la comunicación y en el cálculo de resultados.
La importancia de los ordenadores es anterior al establecimiento de este estándar (¿Habría observado Yuri Gagarin nuestro planeta desde ahí arriba, existiría el seguro mundial contra ambiciones llamado "destrucción mútua total y asegurada", sin los ordenadores primigenios? Estos sucesos son anteriores al estándar de representación actual.) pero la aritmética computacional es mucho mejor comprendida desde que en 1985 el Institute for Electrical and Electronic Engineers, IEEE, de los Estados Unidos, publicó el estándar que se utiliza hoy en día. Hay, esencialmente, una única variación de un procesador a otro, a saber, el número de bits. Antiguamente los procesadores de 32 bits eran los más habituales, a fecha de hoy (julio de 2017) son los de 64 bits los que tienen la mayoría de los PCs de sobremesa, pero existen máquinas potentes con mayor precisión usadas en los cálculos de las grandes compañías y de los gobiernos más poderosos del planeta. Tarde o temprano el estándar en los ordenadores personales pasará a ser el de 128 bits y seguramente más aún en el futuro, pero por mucho que crezca siempre será un número finito.
Dado que hacer una exposición general es demasiado abstracto, nos centraremos en el estándar IEEE de 64 bits, que nos permitirá comprobar las afirmaciones que hagamos usando nuestros ordenadores.
Los llamados "números en doble precisión" o "números en coma flotante de 64 bits" o "números en coma flotante largos", se representan por 64 cifras binarias (1's ó 0's): la cadena

$$\underbrace{s\ (1\ bit)}_{\text{signo}}\ \underbrace{c\ (11\ bits)}_{\text{característica}}\ \underbrace{m\ (52\ bits)}_{\text{mantisa}}$$

representa el número

$$(-1)^s\ 2^{c-1023}\ 1.m,$$

donde hemos de entender que tanto m como c están escritos en binario. El número 1023 se escribe en binario como 01111111111 (el primer 0 lo podemos quitar, pero así tiene 11 cifras). Por lo tanto el número

0 01111111111 0000000000000000000000000000000000000000000000000000

representa a:

$$(-1)^0\ 2^0\ 1{,}0 = 1,$$

y es fácil generar más ejemplos usando esta regla. Algunas excepciones a la regla son:

- Cuando la característica es 00000000000 la fórmula para representar el número se convierte en

  $$(-1)^s\ 2^{-1023}\ 0.m,$$

  de forma que puede codificar el número 0 (que puede tener signo, nótese).

- Cuando la característica es 11111111111, si la mantisa es todo ceros el número es $+\infty$ ó $-\infty$ dependiendo de si $s = 0$ o $s = 1$. Si la mantisa no es todo ceros, el número es NaN ("not a number"), por ejemplo el procesador devuelve este valor si le preguntamos cuánto vale $0/0$, $\infty - \infty$, u otras indeterminadas.

Si bien el detalle técnico de estas consideraciones es muy farragoso, podemos extraer algunas verdades prácticas que son las que más nos interesan en este curso:

- En nuestros ordenadores se puede representar un subconjunto finito de los números reales, siendo el cardinal de este conjunto menor que 2^{64}.
- Todos los números (señalamos que $\pm\infty$ y NaN no son números propiamente dichos) representables en el estándar IEEE de 64 bits son racionales. En particular, no podemos escribir exáctamente $\sqrt{2}, \pi, e, 1/3$, ni ningún otro número con infinitas cifras binarias.
- Algunos números que se pueden escribir fácilmente en notación decimal, por ejemplo $1/5$, no tienen representación finita en binario, luego no se pueden representar exactamente en nuestros ordenadores con el estándar de IEEE (obviamente, podemos representarlo de otras maneras).
- El número más grande que podemos escribir es $2^{1023}(2 - 2^{-52})$, y el más pequeño pero mayor que 0 es 2^{-1074}.
- Al escribir un número en el ordenador, el procesador toma el número representable en coma flotante más cercano al que le damos, tomando en caso de que haya dos a la misma distancia el más cercano a 0.
- Al operar con números, los procesadores supuestamente devuelven el número representable más cercano al resultado verdadero de la operación (operación que recibe como input números en coma flotante).
- Estas aproximaciones (en el input y en la operación) hacen que en cada operación computacional se genere un (normalmente muy pequeño) error de redondeo.
- El número más grande tal que al añadirlo a 1 sigue siendo 1 se llama "unidad de redondeo", vale $\epsilon = 2^{-53}$, y nos da la precisión relativa esperada en los cálculos. Por ejemplo, al sumar dos números a, b podemos esperar que el resultado sea $(a + b)(1 + \delta)$ donde $0 \leq |\delta| \leq \epsilon$. Lo mismo para la resta, el producto y la división.
- Hay unas pocas operaciones no recomendables, que deberíamos evitar en la medida de lo posible por favorecer la aparición de errores de redondeo (¡en ocasiones, serán inevitables! Lo veremos en el capítulo 3): restar dos números muy parecidos entre sí, dividir por un número muy pequeño, multiplicar un número muy pequeño por otro muy grande, o sacar la raíz cuadrada de un número muy pequeño.
- Las propiedad conmutativa de la suma y el producto sí se mantiene, pero no así la asociativa ni la distributiva. Probemos por ejemplo a escribir $(2^{53} + 1) - 2^{53}$ frente a $(2^{53} - 2^{53}) + 1$. ¡Nótese que estamos usando una de las operaciones poco recomendables!

1.5. Exercises. Computer arithmetic: representation and errors

Exercise 1.1

Let M,m,δ be respectively the largest number, the smallest nonzero and the smallest but greater than 1 IEEE 64 bits number. Check the result of computing the following operations:

$M+1, \quad M+M, \quad 3*M, \quad m/2, \quad m-m/2, \quad (m+m)/3, \quad 1+\delta/2, \quad (2+\delta)/2, M-M+1, \quad M+1-M, \quad , m+1-m.$

Do the usual associative and conmutative rules apply to this representation of numbers?

Exercise 1.2

Assume that we attempt to compute the harmonic series $\sum_{n\geq 1} \frac{1}{n}$ using Matlab by simply adding one term after another. What would be the result? Try to deduce a theoretic estimate. How long would it take to check the computation with Matlab?

Exercise 1.3

Same question as above but with $\sum_{n\geq 1} \frac{1}{2^n}$ and $\sum_{n\geq 1} \frac{1}{n!}$.

Exercise 1.4

Write down a program that causes an overflow error, and a program that causes an underflow error.

Exercise 1.5

Write down a program that computes exactly $n!$ where n is a positive integer. Which is the largest n that the program can work for?

Exercise 1.6

Same question as before but with n^2 and n^n.

Exercise 1.7

Write down a program that finds the greatest integer n with the property that all integers up to n are represented exactly in IEEE 64 bits (of course you can figure out this number from the theory!). Which is the largest integer actually representable in this notation?

Exercise 1.8

Write down a program that writes the number "99999..." until it becomes infinity. Which is the largest non–infinity number obtained this way?

Exercise 1.9

Same as above but with "0.000....001" (until it becomes equal to 0).