# Índice General

# Capítulo 1

# Introducción y Conceptos Fundamentales

## 1.0 Introducción

El término Análisis Numérico se hizo de uso general cuando se fundó el Instituto de Análisis Numérico en la Universidad de California, Los Angeles en el año 1947. En sus comienzos estuvo asociado a todo lo que significa procesamiento de datos, pero hoy en día su significado es mucho más preciso: es la teoría de los métodos de resolución aproximada de problemas de la matemática.

Cuando decimos métodos de resolución aproximada, estamos hablando de procedimientos que nos permiten obtener la solución con una precisión arbitraria en un número finito de pasos. Es entonces natural que un tema central dentro del análisis numérico sea el análisis del error presente en estos métodos.

Introduzcamos algunos conceptos que mencionaremos frecuentemente durante el desarrollo del curso:

*Problema Numérico:* Es una descripción precisa de la relación funcional entre los datos de entrada (input) y los datos de salida (output). Tanto el conjunto de los datos de entrada como los de salida son finitos y la relación funcional entre ellos puede estar dada de manera explícita como implícita.

*Algoritmo:* Es un conjunto de instrucciones para efectuar operaciones

matemáticas y/o lógicas que transforman los datos de entrada de un problema numérico en los resultados o datos de salida.

*Método Numérico:* Es un procedimiento tanto para transformar un problema matemático en un problema numérico como para resolver un problema numérico.

**Ejemplo 1.0.1.** *Resolver el sistema de ecuaciones lineales* $Ax = b$, *siendo* $A$ *una matriz* $n \times n$, $x$ *y* $b$ *vectores* $n \times 1$. *Este es un problema numérico cuyos datos de entrada son los elementos de la matriz* $(a_{ij})$, $i, j = 1 : n$ *y los elementos del vector* $(b_i)$, $i = 1 : n$ *y cuyos datos de salida son los* $(x_i)$, $i = 1 : n$.

**Ejemplo 1.0.2.** *Resolver el problema de frontera:*

$$\frac{d^2y}{dx^2} = x^2 + y^2,\ y(0) = 0,\ y(5) = 1$$

*Este es un problema matemático, ya que los datos de salida es la función* $y(x)$ *que no puede ser especificada por un número finito de puntos. Este problema matemático puede ser aproximado por un problema numérico si se consideran como valores de salida los valores* $y(x)$ *para* $x = ih$, $i = 1 : n$, $h = 5/n$ *y*

$$\frac{d^2y}{dx^2}(ih) \cong \frac{y((i+1)h) - 2y(ih) + y((i-1)h)}{h^2}$$

*Entonces, llamando* $x_i = ih$, $y_i = y(ih)$, *se tiene*

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} \cong x_i^2 + y_i^2, \quad i = 1 : n-1$$

*de donde*

$$y_{i+1} - 2y_i + y_{i-1} \cong h^2(x_i^2 + y_i^2), \quad i = 1 : n-1$$

*es decir:*

$$\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-2} \\ y_{n-1} \end{bmatrix} = h^2 \begin{bmatrix} y_1^2 \\ y_1^2 \\ \vdots \\ y_{n-2}^2 \\ y_{n-1}^2 \end{bmatrix} + r$$

*con* $r = [h^2x_1^2, h^2x_2^2, \ldots, h^2x_{n-1}^2 + 1]^T$.

*Este es un sistema de ecuaciones no lineales en* $y_1, y_2, \ldots, y_{n-1}$. *Se trata ahora de buscar un algoritmo de resolución de este problema numérico.*

Entre los métodos numéricos llamados clásicos o fundamentales se encuentran los siguientes:

1. Métodos de resolución de ecuaciones lineales.
2. Métodos de resolución de ecuaciones no lineales.
3. Aproximación de funciones.
4. Métodos de integración numérica.
5. Métodos de resolución de ecuaciones diferenciales ordinarias.
6. Métodos de resolución de ecuaciones diferenciales en derivadas parciales.
7. Métodos de resolución de ecuaciones integrales.
8. Métodos de optimización de funciones.

## 1.0 Breve descripción de los pasos a seguir para la resolución de un problema de análisis numérico

Una ecuación de la física matemática puede expresarse como

$$Au = f \tag{1.1}$$

siendo $A$ una aplicación de un conjunto $X$ (espacio de funciones) en un conjunto $Y$ (espacio funcional); $u$ es la incógnita que se busca en $X$, $f$ es un dato y pertenece a $Y$.

**Ejemplo 1.0.3.**

$$Au = \int_a^b u(x)g(x,t)dx, \quad con\, g(x,t) : [a,b] \times [a,b] \longrightarrow \mathbb{R}\; y\, u(x) : [a,b] \longrightarrow \mathbb{R}$$

*Se busca* $u$ *en* $X = C[a,b]$ *tal que* $Au = f$, *es decir*

$$\int_a^b u(x)g(x,t)dx = f(t), \;\; t \in [a,b]$$

En general es imposible determinar $u$ explícitamente o bien su forma explícita es muy complicada. Las diferentes fases de la resolución de (1.1) son las siguientes:

a) *Estudio teórico de la ecuación:* esto es, el estudio de la existencia y unicidad de la solución de (1.1), y en el caso en que no exista unicidad caracterizar una solución o el conjunto de soluciones. Esta fase es propia del análisis clásico o del análisis funcional.

b) *La aproximación:* consiste en reemplazar $X$ e $Y$ por otros espacios más simples $X_n$ e $Y_n$ (en general son espacios de dimensión finita y $n$ es un parámetro que tenderá a infinito). El operador $A$ es reemplazado por un operador $A_n$ y la función $f$ por una función $f_n$ de $Y_n$. No se buscará $u$ en $X$, sino $u_n$ en $X_n$ tal que

$$A_n u_n = f_n \tag{1.2}$$

lo cual nos define un sistema de ecuaciones algebraicas, como en el ejemplo 1.2. Las preguntas que surgen ahora son:

(a) existencia y unicidad de la solución de (1.2)

(b) estabilidad y convergencia de la solución

Por estabilidad se entiende lo siguiente: $u_n$ es estable si para $n \to \infty$ la sucesión $u_n$ permanece acotada en un cierto sentido.

Por convergencia se entiende lo siguiente: $u_n$ es convergente si $u_n \to u$, $n \to \infty$, para una cierta noción de límite, siendo $u$ la solución del problema original.

c) *La resolución de* (1.2)*:* para $n$ fijo, implica la elección de un algoritmo numérico. Como se pueden presentar diferentes algoritmos, el criterio de elección puede contemplar parámetros tales como:

(a) **El costo:** que se mide en términos del número de operaciones elementales, esto es, suma y multiplicación, y de la ocupación de espacio en la memoria de un computador.

(b) **La estabilidad numérica:** este criterio amerita algunas precisiones. Los números reales no pueden ser introducidos en la memoria de un computador; son sus aproximaciones finitas las que son introducidas como datos y tratados durante la ejecución del algoritmo. Habrá así dos fuentes de errores: los datos cortados o redondeados y la pérdida de cifras significativas en las operaciones elementales.

Si el algoritmo comporta un gran número de operaciones aritméticas, los errores de redondeo pueden acumularse hasta falsear los resultados. Un algoritmo se dice *numéricamente estable* si no es muy sensible a esta acumulación de errores de redondeo.

(c) **La ejecución del algoritmo:** significa la escritura para la computadora de un programa de trabajo, que no es sino la codificación del algoritmo en un lenguaje apropiado para ella.

(d) **La ejecución del programa en la computadora:** las computadora digitales se han transformado en instrumentos indispensables en el quehacer científico actual. Una de las principales razones, es que, mediante la implementación de métodos numéricos, ellas son una herramienta universal para la resolución de una gran variedad de problemas. Es así que hoy en día, con el desarrollo alcanzado por las computadoras digitales, la mayoría de los científicos deben tener un buen conocimiento de los métodos numéricos y de su implementación en las mismas.

Sin embargo, la resolución de problemas científicos no es la única motivación para el estudio de métodos numéricos. El análisis de los métodos numéricos es una actividad matemática, cuyo tema central es la construcción de diferentes clases de aproximaciones.

## 1.1 Fuentes de error

El estudio del error es un objetivo central en el análisis numérico: esto no sólo nos permite analizar, examinar y entender los resultados finales, sino que nos provee de una herramienta que nos servirá para planificar los cálculos y tomar decisiones durante los mismos. Se distinguen varios tipos de errores:

*Errores en los datos de entrada:* pueden ser el resultado de mediciones físicas, o cuando un número real es aproximado por un número finito de decimales.

*Errores de redondeo:* surgen cuando se calcula con números cuya representación está restringida a un número finito de dígitos, que es lo que usualmente ocurre. Por ejemplo, si tenemos una calculadora que usa 8 dígitos, la división 1/6=0.166666666... será redondeada a 0.16666667 o bien cortada a 0.16666666. Si a éste le sumamos 1/3 redondeado a 0.33333333, los errores de redondeo de ambos números se propagan a la suma.

*Errores de aproximación:* como ya hemos dicho, muchos métodos numéricos no dan la solución de un problema dado (aun trabajando sin errores

de redondeo), sino que obtienen la solución de un problema más simple, que aproxima al anterior.

**Ejemplo 1.1.1.** *Los dos primeros términos de la serie de Taylor de* $\sqrt{x+1}$ *dan una buena aproximación cuando* $x$ *es pequeño:* $\sqrt{x+1} \cong 1 + x/2$. *El error cometido se conoce como error de truncamiento.*

**Ejemplo 1.1.2.** *Si evaluamos la integral* $\int_0^1 f(x)dx$ *usando la regla trapezoidal, se obtiene:*

$$\int_0^1 f(x)dx \cong \frac{h}{2}\sum_{i=0}^{n-1}(f_i + f_{i+1}),\ \ h = \frac{1}{n},\ f_i = f(ih)$$

*y el error de aproximación o error de truncamiento es la diferencia entre la suma y la integral.*

**Ejemplo 1.1.3.** *Dada la ecuación diferencial*

$$\frac{dy}{dx} = f(x,y),\ \ y(x_0) = y_0$$

*introducimos la aproximación*

$$\frac{dy}{dx} \cong \frac{y(x+h) - y(h)}{h}, \qquad h \text{ pequeño}$$

*y definimos una solución aproximada* $y_j = y(x_j) = y(x_0 + jh)$ *mediante*

$$\frac{y_{j+1} - y_j}{h} = f(x_j, y_j)$$

*esto es*

$$y_{j+1} = y_j + hf(x_j, y_j),\ \ j \geq 0 \text{ (método de Euler)}$$

*Luego el error global de discretización o truncamiento se define como*

$$y_{j+1} - y(x_{j+1})$$

## 1.1 Errores relativos y absolutos

Vamos a estudiar ahora el efecto de los errores de entrada y los de redondeo en los resultados de los cálculos.

**Definición 1.1.1.** *Sea $\widetilde{a}$ un valor aproximado de una cantidad cuyo valor exacto es $a$. Definimos:*

$\widetilde{a} - a$, *error absoluto*

$|\widetilde{a} - a|$, *magnitud del error absoluto*

$\dfrac{\widetilde{a} - a}{a}$, *error relativo,* $a \neq 0$

$\left|\dfrac{\widetilde{a} - a}{a}\right|$, *magnitud del error relativo,* $a \neq 0$

$m$ *es una cota del error absoluto si* $m > 0$ *y* $|\widetilde{a} - a| \leq m$

$m$ *es una cota del error relativo si* $m > 0$ *y* $\left|\dfrac{\widetilde{a} - a}{a}\right| \leq m \;\; a \neq 0$

*Si las cantidades que se comparan son vectores o matrices, el valor absoluto deberá ser sustituido, respectivamente, por una norma vectorial o matricial.*

**Ejemplo 1.1.4.** *Sean* $\pi = 3.141592653529..$ *y* $a = 35000/11141 = 3.1415492..$ *entonces* $|a - \pi| \leq 0.000044$ *y* $|\dfrac{a - \pi}{\pi}| \leq 0.000014$

En las mediciones es más importante el error relativo que el absoluto. Cuando al medir una longitud decimos que hemos cometido un error de un centímetro, no estamos dando ninguna información, a menos que especifiquemos la magnitud de la longitud que estamos midiendo.

**Definición 1.1.2.** *Decimos que $\widetilde{a}$ tiene $t$ decimales correctos, si $t$ es el mayor entero para el cual se verifica* $|\widetilde{a} - a| \leq 0.5 \;\; 10^{-t}$.

Todo número real no nulo $a$ admite una representación decimal de este tipo:
$a = \pm 0.a_1a_2a_3 \ldots a_t a_{t+1} a_{t+2} \cdots 10^e$ con $a_i$ dígitos de 0 a 9 y $e$ un exponente entero.

**Definición 1.1.3.** *Diremos que la representación decimal está normalizada si* $a_1 \neq 0$

En lugar de error relativo, se usa con mucha frecuencia el concepto de dígitos significativos.

**Definición 1.1.4.** *Supongamos que $\widetilde{a}$ y $a$ están dados en representación normalizada. Decimos que $a$ tiene $t$ dígitos significativos si $t$ es el mayor entero para el cual se verifica:*

$$|\widetilde{a} - a| \leq 0.5 \ 10^{-t} \ 10^{e}$$

**Ejemplo 1.1.5.** $a = 1/6$, $\widetilde{a} = 0.166$, $|\widetilde{a} - a| \cong 0.000666 \leq 0.005 = 0.5 \ 10^{-2}$ *tiene dos decimales correctos y dos dígitos significativos.*

**Ejemplo 1.1.6.** $a = 89.567$, $\widetilde{a} = 89.568$, $|\widetilde{a} - a| \cong 0.001 \leq 0.005 = 0.5 \ 10^{-2}$ *tiene dos decimales correctos y como el error en la forma normalizada es* $|\widetilde{a} - a| \leq 0.5 \ 10^{-4} \ 10^{2}$, *a tiene cuatro dígitos significativos.*

**Ejemplo 1.1.7.** $a = 0.00345$, $\widetilde{a} = 0.00339$, $|\widetilde{a} - a| \cong 0.00006 \leq 0.0005 = 0.5 \ 10^{-3}$ *tiene tres decimales correctos y un dígito significativo.*

El número de decimales correctos da una idea de la magnitud del error absoluto, mientras que él número de dígitos o cifras significativas da una idea de la magnitud del error relativo. En efecto, el error relativo de un número con $t$ dígitos significativos es:

$$|\frac{\widetilde{a} - a}{a}| \leq \frac{0.5 \ 10^{-t} \ 10^{e}}{0.1 \ 10^{e}} = 5 \ 10^{-t}$$

**Observaciones:** El hecho de que $a$ tenga $t$ dígitos significativos no implica que se deba expresarlo con precisamente esa cantidad de dígitos, ya que las cifras que ocupan posiciones posteriores a la $t$-ésima siguen aportando información acerca de $a$. Por ejemplo si $\widetilde{a} = 0.461$ y $t = 3$, entonces $|\widetilde{a} - a| \leq 0.5 \ 10^{-3} \Rightarrow 0.4605 \leq a \leq 0.4615$; mientras que si $\widetilde{a} = 0.4614$ y $t = 3$ entonces $|\widetilde{a} - a| \leq 0.5 \ 10^{-3} \Rightarrow 0.4609 \leq a \leq 0.4619$.

Dígitos significativos o $t$ posiciones correctas no son sinónimos de dígitos iguales o $t$ posiciones iguales, ya que puede suceder lo primero sin lo segundo, a pesar que la recíproca siempre es cierta. Por ejemplo, si $a = 0.4 \ 10^{1}$, $\widetilde{a} = 0.3999999 \ 10^{1}$ tiene 6 dígitos significativos aunque ninguno igual a los de $a$.

Dado un número $a = a_0.a_1a_2a_3 \cdots a_t a_{t+1} a_{t+2} \cdots$ hay dos maneras de llevarlo a un número con $t$ decimales:

a) Por **redondeo**:

$$a = \begin{cases} a_0.a_1a_2a_3 \cdots a_t & \text{si } a_{t+1} < 5 \\ a_0.a_1a_2a_3 \cdots a_t + 10^{-t} & \text{si } a_{t+1} \geq 5 \end{cases}$$

b) Por **cortada**: $a = a_0.a_1a_2a_3\cdots a_t$

Es fácil ver que:

$$|\widetilde{a} - a| \leq \begin{cases} 10^{-t} & \text{por cortada} \\ 0.5\,10^{-t} & \text{por redondeo} \end{cases}$$

## 1.1 Propagación de errores

Sea *op* una operación aritmética $(+ - */)$; sean $X = x + \varepsilon,\ Y = y + \eta$ Aquí $x$ es el valor exacto y $X$ el valor aproximado. Queremos examinar el efecto del error propagado al efectuar una operación *op*. Pongamos $r_x = \varepsilon/x$ y $r_y = \eta/y$.

*Multiplicación y División:*

$$\frac{XY - xy}{xy} = \frac{(x+\varepsilon)(y+\eta) - xy}{xy} = \frac{\varepsilon y + \eta x + \varepsilon\eta}{xy} = \frac{\varepsilon}{x} + \frac{\eta}{y} + \frac{\varepsilon}{x}\frac{\eta}{y}$$

Si $|r_x| \ll 1$ y $|r_y| \ll 1$ se tiene que:

$$r_{xy} \cong r_x + r_y$$

$$\frac{\frac{X}{Y} - \frac{x}{y}}{\frac{x}{y}} = \frac{\frac{x+\varepsilon}{y+\eta} - \frac{x}{y}}{\frac{x}{y}} = \frac{\frac{xy + \varepsilon y - xy - \eta x}{(y+\eta)y}}{\frac{x}{y}} = \frac{\frac{\varepsilon y - \eta x}{xy}}{\frac{y+\eta}{y}} = \frac{\frac{\varepsilon}{x} - \frac{\eta}{y}}{1 + \frac{\eta}{y}}$$

Resulta entonces $r_{x/y} = \dfrac{r_x - r_y}{1 + r_y}$ y si $|r_y| \ll 1$

$$r_{r/y} \cong r_x - r_y$$

*Suma y Diferencia:*

$$\begin{aligned} \frac{(X \pm Y) - (x \pm y)}{x \pm y} &= \frac{((x+\varepsilon) \pm (y+\eta)) - (x \pm y)}{x \pm y} = \frac{\varepsilon}{x \pm y} \pm \frac{\eta}{x \pm y} \\ &= \frac{x}{x \pm y}\frac{\varepsilon}{x} \pm \frac{y}{x \pm y}\frac{\eta}{y} = \frac{x}{x \pm y} r_x \pm \frac{y}{x \pm y} r_y \end{aligned}$$

Observamos así que tanto en la multiplicación como en la división los errores relativos de los operadores no se propagan fuertemente a los resultados.

Esto también es cierto para la suma si $x$ e $y$ tienen el mismo signo, ya que en este caso

$$\left|\frac{x}{x+y}\right| \leq 1 \text{ y } \left|\frac{y}{x+y}\right| \leq 1$$

y por lo tanto $|r_{x\pm y}| \leq \text{máx}\{|r_x|, |r_y|\}$, al ser

$$\left|\frac{x}{x+y}\right| + \left|\frac{y}{x+y}\right| = 1$$

Si los dos tienen signos opuestos, al menos uno de los factores $\left|\frac{x}{x+y}\right|$ ó $\left|\frac{y}{x+y}\right|$ es mayor que uno y uno de los errores relativos $r_x$ ó $r_y$ será amplificado. Esto es drástico si $x \cong -y$; en este caso ocurre lo que se conoce como el fenómeno de cancelación (catastrófica) de dígitos.

**Ejemplo 1.1.8.** *Para calcular $e^x$ usamos su desarrollo de Taylor. Si trabajamos con 5 dígitos significativos y queremos calcular $e^{-5.5}$ tendríamos:*

$$\begin{aligned} e^{-5.5} = \quad & +1.0000 \\ & -5.5000 \\ & +15.125 \\ & -27.730 \\ & +38.129 \\ & -41.942 \\ & +38.446 \\ & -30.208 \\ & \vdots \\ \hline & +0.0026363 \end{aligned}$$

*La suma se termina después de 25 términos pues los términos que siguen no la alteran. Sin embargo $e^{-5.5} = 0.00408677$, de manera que la suma que calculamos no tiene ningún dígito significativo. (referencia: Forsythe, Malcolm, Moler, Pág 15). El remedio para este mal es calcular la suma para $x = 5.5$ y luego tomar el recíproco: $e^{-5.5} = 1/e^{5.5} \cong 0.0040865$; de esta manera se obtienen 4 dígitos significativos.*

**Ejemplo 1.1.9.** *La ecuación cuadrática $ax^2 + bx + c = 0$ tiene dos raíces:*

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

*Si $a = 1$, $b = -10^5$, $c = 1$ y se trabaja con una calculadora con 8 dígitos significativos, los valores de las raíces calculadas son:*

$$x_1 = 100000.00 \qquad x_2 = 0$$

*mientras que los valores exactos redondeados a 11 dígitos significativos son:*

$$x_1 = 99999.999990 \qquad x_2 = 0.000010000000001$$

*Al calcular la segunda raíz se ha producido una cancelación de cifras significativas. Esta cancelación se puede evitar si calculamos las raíces así:*

$$x_1 = -\frac{b + sgn\, b\sqrt{b^2 - 4ac}}{2a} \qquad x_2 = \frac{c}{ax_1}$$

*Si aplicamos esto al ejemplo obtenemos:*

$$x_1 = 100000.00 \qquad x_2 = 0.00001$$

*siendo ambas respuestas aceptables (referencia: Forsythe, Malcolm, Moler, pág. 20)*

**Ejemplo 1.1.10.** *Si se quiere calcular la siguiente expresión para $|x| << 1$*

$$\frac{1}{1+2x} - \frac{1-x}{1+x}$$

*para evitar pérdida de cifras significativas es conveniente calcularla mediante*

$$\frac{2x^2}{(1+2x)(1+x)}$$

*y también porque se hacen menos operaciones.*

## 1.2 Los Símbolos o y O

Consideremos dos funciones $f, g : D \to \mathbb{R}, D \subset \mathbb{R}, D$ abierto y tomemos $a \in D$. Queremos comparar $f(x)$ y $g(x)$ cuando $x \to a$. Entonces:

**Definición 1.2.1.** *Decimos que*

*i) $f$ es de orden "O grande"con respecto a $g$ cuando $x \to a$ si existe una constante $K > 0$ tal que*

$$\left|\frac{f(x)}{g(x)}\right| \leq K$$

*para $x \neq a$ en un entorno de $a$. Pondremos $f = O(g)$ cuando $x \to a$.*

*ii) $f$ es de orden "o pequeña"con respecto a $g$ cuando $x \to a$ si*

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 0$$

*Pondremos $f = o(g)$ cuando $x \to a$.*

**Ejemplo 1.2.1.** *$f = O(1)$ y $f = o(1)$ para $x \to a$ significan, respectivamente, que $f$ se mantiene acotada y que $f(x)$ tiende a cero cuando $x \to a$.*

**Ejemplo 1.2.2.** *Sea $f(x) = e^x$. Su desarrollo de Taylor hasta la derivada segunda es:*

$$e^x = 1 + x + e^\zeta \frac{x^2}{2}, \qquad \zeta \in (0, x)$$

*de donde*

$$\left| \frac{e^x - (1+x)}{x^2} \right| = \frac{1}{2} e^\zeta$$

*y por lo tanto*

$$\left| \frac{e^x - (1+x)}{x^2} \right| \leq \sup_{0<|x|<1} \frac{1}{2} e^\zeta = K$$

*resulta $e^x - (1+x) = O(x^2)$, cuando $x \to 0$, lo cual significa que $e^x - (1+x)$ se comporta como $x^2$ cuando $x$ es pequeño.*

*Consideremos ahora*

$$\left| \frac{e^x - (1+x)}{x} \right| = \frac{1}{2} e^\zeta |x| \leq K|x|$$

*De aquí*

$$\lim_{x \to 0} \frac{e^x - (1+x)}{x} = 0$$

*de donde $e^x - (1+x) = o(x)$.*

**Propiedad 1.2.1.** *Si $f = o(g)$ cuando $x \to a$ entonces $f = O(g)$ cuando $x \to a$.*

**Demostración.-** Supongamos que $f = o(g)$ cuando $x \to a$, entonces $\lim_{x \to a} \frac{f(x)}{g(x)} = 0$, es decir dado $\varepsilon = K$, existe $\delta > 0$ tal que $\left| \frac{f(x)}{g(x)} \right| < K$ si $0 < |x - a| < \delta$.

El recíproco no es cierto. Basta tomar $f = g$

**Ejemplo 1.2.3.** *Sea* $f(x) = 2x^2 - x^3$. *Se tiene:*

*i) Para* $x \to 0$, $f(x) = O(x^n)$, $n \leq 2$ *y* $f(x) = o(x^n), n < 2$.

*ii) Para* $x \to \infty$, $f(x) = O(x^n)$, $n \geq 3$ *y* $f(x) = o(x^n), n > 3$.

**Nota:** Convendremos en indicar $O(g)$, $o(g)$ una función arbitraria $f$ tal que $f = O(g)$, $f = o(g)$ respectivamente. Con esta notación podemos escribir:

i) $O(1)+O(1) = O(1)$ para indicar que la suma de dos funciones acotadas es acotada.

ii) $O(1).o(1) = o(1)$

En particular $O(h^n)$ y $o(h^n)$ cuando $h \to 0$ indican una función $f$ tal que

$$\left|\frac{f(h)}{h^n}\right| < K, \text{ cuando } h \to 0 \text{ y } \lim_{h\to 0} \frac{f(h)}{h^n} = 0,$$

respectivamente.

**Propiedad 1.2.2.** *Si* $f = O(h^n)$ *cuando* $h \to 0$ *para algún* $n \in Z^+$ *entonces* $f = O(h^{n-1})$ *cuando* $h \to 0$.

**Demostración.-** Si $f = O(h^n)$ existen $K$ y $\delta > 0$ tal que $\left|\frac{f(h)}{h^n}\right| \leq K$ para $0 < |h| < \delta$. Por lo tanto $\left|\frac{f(h)}{h^{n-1}}\right| \leq hK$, es decir $f = o(h^{n-1})$. Por la propiedad 1.6.1, $f = O(h^{n-1})$.

## 1.3 Fórmula general de propagación de errores

Consideremos una función $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$, con $D$ abierto y $f$ con segundas derivadas parciales continuas. Sea $X = (X_1, X_2, \ldots, X_n)$ un valor aproximado de $x = (x_1, x_2, \ldots, x_n)$; introduzcamos la notación siguiente para los errores absolutos y relativos:

$$\varepsilon_i = X_i - x_i,\ r_i = \frac{\varepsilon_i}{x_i},\ i = 1 : n$$

El problema es saber cómo se han propagado los errores en los datos cuando se calcula $f(X)$, es decir, se trata de hallar una relación entre dichos errores y el error $f(X) - f(x)$.

**Teorema 1.3.1.** *Llamemos*

$$\varepsilon_y = Y - y,\ r_y = \frac{\varepsilon_y}{y}$$

*con* $Y = f(X),\ y = f(x)$. *Entonces:*

$$a)\ \varepsilon_y \cong \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}\varepsilon_i$$

$$b)\ |\varepsilon_y| \leq \sum_{i=1}^{n} \left|\frac{\partial f(x)}{\partial x_i}\right| |\varepsilon_i|$$

$$c)\ |r_y| \leq \sum_{i=1}^{n} \left|\frac{\partial f(x)}{\partial x_i}\frac{x_i}{y}\right| |r_i|$$

**Prueba:** Resulta de considerar el desarrollo de Taylor de $f$ alrededor de $x$ :

$$f(X) = f(x) + \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}(X_i - x_i) + O(||X - x||^2)$$

y despreciar los términos de segundo orden. Aquí $||.||$ indica una norma en $\mathbb{R}^n$.

**Ejemplo 1.3.1.** *Sea* $z = ln(xy)$, *con* $x$ *e* $y$ *reales no nulos del mismo signo. En este caso se tiene:*

$$\varepsilon_z \cong \frac{1}{x}\varepsilon_x + \frac{1}{y}\varepsilon_y = r_x + r_y$$

*con* $\varepsilon_x, \varepsilon_y$ *errores absolutos de los datos.*

## 1.4 Condición y Estabilidad

Consideremos el siguiente sistema de ecuaciones diferenciales:

$$\begin{cases} y_1^{'} &= y_2 \\ y_2^{'} &= y_1 \end{cases} \tag{1.3}$$

La solución está dada por:

$$\begin{cases} y_1 &= c_1 e^x + c_2 e^{-x} \\ y_2 &= c_1 e^x - c_2 e^{-x} \end{cases} \tag{1.4}$$

Si a (1.3) le añadimos las condiciones iniciales

$$y_1(0) = -y_2(0) = 1 \tag{1.5}$$

los coeficientes de $c_1$ de (1.4) resultan ser $c_1 = 0$ y $c_2 = 1$.

Supongamos que (1.3) sea resuelto numéricamente por cualquier método que calcule $y_1$, $y_2$ en una sucesión de puntos $x_1, x_2, x_3 \cdots$. El efecto del error de redondeo es equivalente a calcular (con exactitud) una solución con condiciones iniciales resultantes de perturbar (1.5). Pero la más pequeña perturbación causará una contribución al término $e^x$ que dominará $e^{-x}$ para $x$ grande y en este caso será imposible obtener una buena aproximación de la solución. Se dice entonces que este problema es *mal condicionado*. Un problema mal condicionado puede ser resuelto con una precisión razonable si los cálculos se realizan muy cuidadosamente (por ejemplo usando precisión múltiple) y esto independientemente del método numérico utilizado. En cambio un problema bien condicionado puede ser resuelto por cualquier método numérico que sea estable para ese problema. Vemos así que el concepto de *condición* está relacionado con la sensibilidad de la solución de un problema ante perturbaciones de los datos.

Consideremos un segundo ejemplo: se quiere calcular para $n = 0 : 8$ las siguientes integrales:

$$y_n = \int_0^1 \frac{x^n}{x+5} dx$$

En principio podemos resolverlo así:

$$y_n + 5y_{n-1} = \int_0^1 \left( \frac{x^n}{x+5} + 5\frac{x^{n-1}}{x+5} \right) dx = \int_0^1 x^{n-1} dx = \frac{1}{n}$$

y plantear la siguiente fórmula de recurrencia:

$$y_n = \frac{1}{n} - 5y_{n-1}, \ n = 1, 2, 3 \cdots \tag{1.6}$$

Calculamos el valor inicial usando una tabla de logaritmos de tres lugares:

$$y_0 = \int_0^1 \frac{x^0}{x+5} dx = ln(x+5)|_0^1 \cong 0.182 \pm 0.0005$$

y los subsiguientes usando (1.6)

$y_1 = 1 - 5 \times 0.182 \cong 0.090$
$y_2 = 1/2 - 5 \times 0.090 \cong 0.05$
$y_3 = 1/3 - 5 \times 0.050 \cong 0.083$
$y_4 = -0.165$ lo cual es evidentemente falso.

Lo que ha ocurrido es que el error $\varepsilon_0 = 0.0005$ se ha ido amplificando por un factor 5 hasta que $y_4$ destruye el resultado exacto. Si usáramos más decimales el absurdo ocurriría en etapas posteriores. Decimos entonces que este método es *numéricamente inestable*. La inestabilidad resulta cuando los errores de redondeo se propagan de manera creciente durante el cálculo, lo que produce una distorsión en los resultados. Vemos que el concepto de estabilidad o inestabilidad numérica está relacionado con el método numérico utilizado por el comportamiento de los errores de redondeo introducidos en la solución numérica.

En el caso del ejemplo podemos plantear la siguiente fórmula de recurrencia regresiva:

$$y_{n-1} = \frac{1}{5n} - \frac{y_n}{5}, \qquad n = N, N-1, \cdots, 1, 0$$

En este caso el error se divide entre cinco.

Veamos cómo elegimos $N$. Es fácil ver que la sucesión $\{y_n\}$ es decreciente y que

$$\frac{1}{6(n+1)} < y_n < \frac{1}{6n} < y_{n-1} < \frac{1}{6(n-1)}$$

Resulta así que la sucesión decrece como $\frac{1}{6n}$. Podemos pensar que $y_0 \cong y_{10} \cong$

$\frac{1}{60}$, lo cual nos permite calcular

$$
\begin{aligned}
y_8 &= 1/45 - y_9/5 \cong 0.019 \\
y_7 &\cong 0.021 \\
y_6 &\cong 0.025 \\
y_5 &\cong 0.028 \\
y_4 &\cong 0.034 \\
y_3 &\cong 0.043 \\
y_2 &\cong 0.058 \\
y_1 &\cong 0.088 \\
y_0 &\cong 0.182 \qquad\qquad \text{CORRECTO !}
\end{aligned}
$$

**Nota**: No cometer el error de creer que toda fórmula de recurrencia regresiva es numéricamente estable.

# 1.5 Sistemas Numéricos: aritmética de punto flotante

## 1.5 Sistemas de posición

Los números reales se representan mediante un sistema de posición con una base $B$, siendo $B$ un entero mayor o igual que 2. Así todo número real $a$ admite una representación en la forma:

$$a = \pm(a_n B^n + a_{n-1}B^{n-1} + \cdots + a_1 B + a_0 + a_{-1}B^{-1} a_{-2}B^{-2} + \cdots) = \pm \sum_{i=-\infty}^{n(a)} a_i B^i$$

donde los $a_i$ son tales que $0 \leq a_i \leq B-1$, $a_i$ enteros. Esta representación es única siempre que se excluyan las ambigüedades tipo 0.99999...=1 en el caso decimal.

Ejemplos clásicos:

| base | sistema numérico | dígitos |
|---|---|---|
| 2 | binario | 0,1 |
| 8 | octal | 0,1,2,3,4,5,6,7 |
| 10 | decimal | 0,1,2,3,4,5,6,7,8,9 |
| 16 | hexadecimal | 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F |

Una de las grandes ventajas del sistema de posición es que las operaciones aritméticas pueden realizarse de acuerdo a reglas muy sencillas. Cuanto más pequeña es la base más simples son las reglas.

**Ejercicio**: Describir la suma y el producto en el sistema binario.

## 1.5 Aritmética del computador

La aritmética realizada por un computador involucra sólo números con un número finito de dígitos; esto implica que los cálculos se realizan con valores aproximados de los números verdaderos. El error cometido al representar un número real por una aproximación finita es lo que ya vimos como error de redondeo. Veremos que sólo un *conjunto finito* de números pueden ser representados en la máquina.

La representación de un número en una computadora consta de tres partes: una parte para el signo, una parte exponencial, llamada *característica*, y una parte fraccionaria, llamada *mantisa*. Por ejemplo en una IBM 370, un número en precisión simple consta de 1 dígito binario (bit) para el signo, un exponente de 7 bits en base 16 y una mantisa de 24 bits. Como 24 bits corresponden a 6 ó 7 dígitos decimales, podemos suponer que este número tiene por lo menos, seis cifras decimales de precisión para el sistema de numeración. El exponente de siete bits da un rango de 0 a 127, pero debido a que se necesitan exponente negativos, automáticamente se le resta 64; el rango del exponente es entonces de -64 a 63.

Por ejemplo:

$$0 \qquad 1000100 \qquad 101100001000000001000000$$

El bit a la izquierda es 0, lo cual indica que el número es positivo. Los siguientes siete bits son equivalentes al número decimal:

$$1.2^6 + 1.2^2 = 68$$

de donde resulta que el exponente es $68 - 64 = 4$. La mantisa la da los 24 bits finales:

$$m = 1(1/2)^1 + 1.(1/2)^3 + 1.(1/2)^4 + 1.(1/2)^9 + 1.(1/2)^{18}$$

Luego el número de máquina dado representa al número decimal:

$$m \times 16^4 = 45184.25$$

Sin embargo el número de máquina inmediatamente anterior:

$$0 \quad 1000100 \quad 10110000100000000111111 \quad = \quad 45184.23828125$$

mientras que el inmediatamente posterior es:

$$0 \quad 1000100 \quad 10111000010000000001000001 \quad = \quad 45184.25390625$$

lo cual señala que el número de máquina dado debe representar no solamente a 45184.25, sino a un conjunto infinito de números reales que están entre este número y sus números de máquina más próximos.

Para asegurar la unicidad de la representación y obtener toda la precisión disponible se requiere que la representación sea normalizada; esto se logra cuando por lo menos uno de los cuatro primeros bits de la mantisa de un número de máquina, contados de izquierda a derecha, sea uno.

El sistema que estamos describiendo usa $15 \times 2^{28} + 1$ números de la forma:

$$\pm 0.d_1 d_2 \cdots d_{24} \times 16^{e_1 e_2 \cdots e_7}$$

para representar a todos los números reales; esto implica que el número de máquina más pequeño mayor que cero que puede representarse es

$$0\ 0000000\ 00010000000000000000000000 \ = \ 16^{-65} \cong 10^{-78}$$

mientras que el más grande es

$$0\ 1111111\ 11111111111111111111111111 \ = \ 16^{63} \cong 10^{76}$$

Los números con magnitud menor que la primera resultan en lo que se llama *underflow*, y generalmente se les da el valor cero, mientras que los mayores que la segunda resultan en *overflow* y causa que los cálculos se detengan. El mayor número positivo $a$ tal que $1 + a = 1$ se llama el *epsilon de máquina.*

Un conjunto $F$ de números de *punto flotante* en una computadora, está caracterizado por $F =< B, t, m, M >$, con $B$ =base, $t$ =precisión de la máquina, $m$ y $M$ =rango del exponente. Como los humanos seguimos pensando en términos de una representación decimal, el análisis que haremos a continuación será en base 10. Sea el número real $a$ no nulo con una representación decimal $a = \pm 10^e (0.d_1 d_2 d_3 \cdots d_k d_{k+1} d_{k+2} \cdots)$, $e$ =entero, $d_i$ =dígitos, $d_1 \neq 0$.

La representación de este número en forma de punto flotante con $B = 10$ es

$$fl(a) = \pm 10^e (0.d_1 d_2 \cdots d_t^{'})$$

con $m \leq e \leq M$; para el dígito $d_t^{'}$ hay dos manera de escogerlo:

$$d_t^{'} = d_t, \text{ por cortada}$$
$$d_t^{'} = \begin{cases} d_t, & \text{si } d_{t+1} < 5 \text{ (por redondeo)} \\ d_t + 1, & \text{caso contrario} \end{cases}$$

**Lema 1.5.1.** *El error relativo en la representación de punto flotante $fl(a)$ con $t$ dígitos, para el número $a$ es:*

$$\frac{|fl(a) - a|}{|a|} \leq s\, 10^{1-t},\ con\, s = \begin{cases} 1, por\ cortada \\ 0.5, por\ redondeo \end{cases}$$

**Prueba:** Para la cortada se tiene que:

$$\frac{|a - fl(a)|}{|a|} = \frac{10^e 10^{-t}(.d_{t+1}d_{t+2}\cdots)}{10^e(.d_1d_2\cdots)} = 10^{-t}\frac{(.d_{t+1}d_{t+2}\cdots)}{(.d_1d_2\cdots)}$$

pero como $d_1 \geq 1$ y $0.d_{t+1}d_{t+2}\cdots \leq 1$, esto implica que:

$$\frac{|a - fl(a)|}{|a|} \leq 10^{1-t}$$

De la misma forma se prueba en el caso del redondeo.

**Corolario 1.5.1.** *El error en la representación decimal flotante con $t$ dígitos es:*

$$|a - fl(a)| \leq |a|10^{1-t}s,\ con\, s = \begin{cases} 1, por\ cortada \\ 0.5, por\ redondeo \end{cases}$$

Mediante análogos razonamientos se tiene para el caso general:

**Teorema 1.5.1.** *Para una computadora con una aritmética de punto flotante de base $B$ y una mantisa con $t$ dígitos, todo número real en el rango de punto flotante puede representarse con un error relativo tal que*

$$\frac{|fl(a) - a|}{|a|} \leq s\, B^{1-t},\ con\, s = \begin{cases} 1, por\ cortada \\ 0.5, por\ redondeo \end{cases}$$

## 1.5 Operaciones con punto flotante

Sean $fl(x)$ y $fl(y)$ las representaciones en punto flotante de $x$ e $y$; los símbolos $\oplus \ominus \odot \oslash$ representan respectivamente las operaciones suma, diferencia, producto y cociente en la máquina. Suponemos una aritmética de $t$ dígitos dada

por

$$\begin{aligned} x \oplus y &= fl(fl(x) + fl(y)) \\ x \ominus y &= fl(fl(x) - fl(y)) \\ x \odot y &= fl(fl(x).fl(y)) \\ x \oslash y &= fl(fl(x)/fl(y)) \end{aligned}$$

Esto corresponde a efectuar la aritmética exacta en las representaciones del punto flotante de $x$ e $y$, luego la conversión del resultado exacto a su representación de punto flotante.

**Ejemplo 1.5.1.** *Sean* $t = 4$, $x = 1234.67$, $y = -0.9999\,10^{-1}$, *entonces:*

$$\begin{aligned} x \oplus y &= fl(fl(x) + fl(y)) = fl(0.1235\,10^4 - 0.000009999\,10^4) \\ &= fl(0.123490001\,10^4) = 0.1235\,10^4 = 1235 \end{aligned}$$

$$\begin{aligned} x \odot y &= fl(fl(x).fl(y)) = -fl(0.1235\,10^4\,.\,0.9999\,10^{-1}) \\ &= -fl(0.12348765\,10^3) = -0.1235\,10^3 = -123.5 \end{aligned}$$

Vemos que el resultado de una operación en punto flotante es igual al valor redondeado del resultado de la operación y por lo tanto:

$$\frac{|fl(x\,op\,y) - x\,op\,y|}{|x\,op\,y|} \leq = \begin{cases} 0.5\,10^{1-t}, \text{por redondeo} \\ 10^{1-t}, \text{por cortada} \end{cases}$$

En esta aritmética no siempre es cierto que:

$$\begin{aligned} (x \oplus y) \oplus z &= x \oplus (y \oplus z) \\ (x \odot y) \odot z &= x \odot (y \odot z) \end{aligned}$$

Es decir, la suma y el producto no tienen por qué ser asociativos.

**Ejemplo 1.5.2.** *Sean:*

*a)* $x = 0.12345$, $y = 0.42357\,10^4$, $z = -y$, $t = 5$. *Resulta:*

$$\begin{aligned} (x \oplus y) \oplus z &= 0.1 \\ x \oplus (y \oplus z) &= 0.12345 \end{aligned}$$

*b)* $x = 0.123$, $y = 0.456\,10^4$, $z = 0.231$, $t = 3$. *Entonces:*

$$\begin{aligned} (x \odot y) \odot z &= 0.130\,10^{-1} \\ x \odot (y \odot z) &= 0.129\,10^{-1} \end{aligned}$$

## 1.5 Propagación del error de redondeo en una suma extendida

Se quiere calcular $\sum_{i=1}^{n} x_i$ siendo $x_i = fl(\text{x}_i)$. Pongamos

$$fl(x_1 + x_2) = (x_1 + x_2)(1 + \varepsilon_2), \text{ con } |\varepsilon_2| \leq \begin{cases} 0.5\,10^{1-t}, \text{por redondeo} \\ 10^{1-t}, \text{por cortada} \end{cases}$$

Entonces:

$$fl(x_1 + x_2 + x_3) = fl(fl(x_1 + x_2) + x_3) = ((x_1 + x_2)(1 + \varepsilon_2) + x_3)(1 + \varepsilon_3)$$

en donde $|\varepsilon_3|$ tiene la misma cota que $|\varepsilon_2|$. En general:

$$\begin{aligned} fl(x_1 + x_2 + x_3 \cdots + x_n) &= (x_1 + x_2)(1 + \varepsilon_2)(1 + \varepsilon_3) \cdots (1 + \varepsilon_n) \\ &\quad + x_3(1 + \varepsilon_3)(1 + \varepsilon_4) \cdots (1 + \varepsilon_n) \\ &\quad + \cdots + x_n(1 + \varepsilon_n) \\ &\cong (x_1 + \cdots + x_n) + (x_1 + x_2)(\varepsilon_2 + \cdots + \varepsilon_n) + \\ &\quad + x_3(\varepsilon_3 + \cdots + \varepsilon_n) \cdots + x_n\varepsilon_n \end{aligned}$$

resultando

$$\begin{aligned} fl(x_1 + \cdots + x_n) - (x_1 + \cdots + x_n) &= (x_1 + x_2)(\varepsilon_2 + \cdots + \varepsilon_n) \\ &\quad + x_3(\varepsilon_3 + \cdots + \varepsilon_n) + \cdots + x_n\varepsilon_n \end{aligned}$$

Se deduce entonces que la mejor estrategia para la suma es ordenar los números de menor a mayor y luego sumarlos en este orden.

Por último, observamos que si op indica una de las cuatro operaciones aritméticas, entonces la versión calculada de $\text{x op y}$ con x e y y dos números exactos, está dada por

$$x\,op\,y$$

siendo *op* la operación op de la aritmética de punto flotante, es decir

$$x\,op\,y = fl(x \,\text{op}\, y) \quad \text{y} \quad x = fl(\text{x}) \quad y = fl(\text{y})$$

El error cometido es:

$$x\,op\,y - \text{x op y} = (x\,op\,y - x \,\text{op}\, y) + (x \,\text{op}\, y - \text{x op y})$$

El primer término corresponde al error de redondeo de punto flotante y el segundo al error propagado.

## 1.6 Desastres

Antes de finalizar este capítulo vamos a relatar algunos desastres ocurridos por errores en el uso de la aritmética del computador.

**La falla del misil Patriot**

El 25 de febrero de 1991 durante la guerra del golfo, un misil Patriot estadounidense disparado desde Dharan, Arabia Saudita, falló en interceptar un misil iraquí Scud. El Scud dio en el blanco, una barraca del ejército americano, matando a 28 soldados. La causa fue la imprecisión de un cálculo del tiempo debido a errores de aritmética del computador. Un reloj interno medía el tiempo en unidades de décimas de segundos. Para producir el tiempo en segundos se multiplicaba por 1/10. Este cálculo se realizaba usando un registro de punto fijo de 24 bits. En particular 1/10, que no tiene una expansión binaria finita se truncaba a 24 bits. Como la batería del Patriot tenía su tiempo de vida de alrededor de 100 horas, este error de truncamiento se propagaba de manera que conducía a un error significativo. Más específicamente, el número 1/10 es igual a $1/2^4 + 1/2^5 + 1/2^8 + 1/2^9 + 1/2^{12} + 1/2^{13} + \ldots$ o lo que es lo mismo, su expansión binaria es $0.0001100110011001100110011\ldots$. El registro del Patriot almacenaba sólo 0.00011001100110011001100 introduciendo un error de $0.0000000000000000000000110011\ldots$ binarios, alrededor de 0.000000095 decimales. Multiplicando por el número de décimas de un segundo en 100 horas el error es igual a $0.000000095 \times 100 \times 60 \times 60 \times 10 = 0.34$ segundos. Un Scud viaja cerca de 1,676 metros por segundo, es decir más de medio kilómetro en 0.34 segundos. Esto puso al Scud lejos de la predicción del Patriot de lo que se llama el rango de puerta para ser interceptado. Irónicamente, el hecho de que los cálculos del tiempo habían sido mejorados en algunas partes del código, pero no en todo, contribuyó al problema, ya que significó que las imprecisiones no se cancelaron.

**La explosión del Arianne 5**

El 4 de junio de 1996, el cohete Arianne 5, fabricado por la Agencia Espacial Europea, explotó 40 segundos después del despegue. Era su primer viaje después de 10 años de desarrollo a un costo de 7 billones de dólares. El cohete y su carga fueron valuados en 500 millones de dólares. La investigación determinó que la causa de la falla fue un error de software del sistema de referencia inercial. Específicamente, un número de punto flotante de 64 bits relacionado a la velocidad horizontal del cohete con respecto a la plataforma, fue convertido a un entero con signo de 16 bits. El número de 64 bits era mayor que 32,768, el mayor entero que se puede representar con un entero de 16 bits con signo, y por lo tanto la conversión falló.

# Capítulo 2

# Ecuaciones No Lineales.

## 2.0 Introducción.

***Problema general:*** Hallar $x \in \mathbb{R}^n$ tal que $F(x) = 0$, siendo $F : \mathbb{R}^n \to \mathbb{R}^m$ una función no lineal. De manera equivalente, hallar $x = (x_1, \ldots, x_n)$ tal que

$$\begin{array}{rcl} f_1(x_1, \ldots, x_n) & = & 0 \\ f_2(x_1, \ldots, x_n) & = & 0 \\ \vdots & \vdots & \vdots \\ f_m(x_1, \ldots, x_n) & = & 0 \end{array}$$

con $f_j : \mathbb{R}^n \to \mathbb{R}, \quad j = 1 : m$

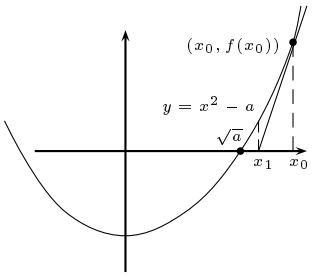
***Caso Particular 1:*** $n = m = 1$, es decir, hallar $x \in \mathbb{R}$ tal que $f(x) = 0$ con $f : \mathbb{R} \to \mathbb{R}$

***Caso Particular 2:*** Hallar una o más raíces del polinomio

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \qquad a_n \neq 0$$

**Ejemplo 2.0.1.** *Hallar la raíz positiva de un número $a > 0$.*

Definiendo $f(x) = x^2 - a$, planteamos la búsqueda de la raíz positiva de $f(x) = x^2 - a = 0$.




Sea $x_0$ una aproximación inicial a la solución $\sqrt{a}$. Tracemos la recta tangente a la gráfica de $f$ en el punto $(x_0, f(x_0))$. Sea $x_1$ el punto de intersección de la tangente con el eje $x$, es decir,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = x_0 - \frac{x_0^2 - a}{2x_0} = \frac{1}{2}\left(x_0 + \frac{a}{x_0}\right)$$

Esto nos da una mejor aproximación a $\sqrt{a}$. Repitiendo el proceso tendremos que

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right), \qquad n \geq 0 \tag{2.1}$$

Cada uno de estos pasos se llama una "*iteración*" (o aproximación sucesiva); $x_n$ se llama el "*iterado $n-$ésimo*" y generalmente se espera que el iterado $x_{n+1}$ mejore los resultados previos.

Analicemos la *convergencia* de la sucesión $\{x_n\}$ a $\sqrt{a}$.

$$x_{n+1} - \sqrt{a} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) - \sqrt{a} = \frac{x_n^2 + a - 2\sqrt{a}x_n}{2x_n} = \frac{1}{2x_n}(x_n - \sqrt{a})^2 \tag{2.2}$$

De esta expresión resulta que si $0 < x_0 < +\infty$, $x_n \geq \sqrt{a} > 0 \quad \forall n$. Además,

$$x_n - x_{n+1} = x_n - \frac{1}{2}x_n - \frac{a}{2x_n} = \frac{1}{2x_n}(x_n^2 - a);$$

Por ser $x_n > \sqrt{a}$, resulta que $\{x_n\}$ es una sucesión decreciente acotada inferiormente por $\sqrt{a}$. Así, existe $\alpha > 0$ tal que $\lim\limits_{n\to\infty} x_n = \alpha$

Tomando límite en (2.1) resulta

$$\alpha = \frac{1}{2}\left(\alpha + \frac{a}{\alpha}\right)$$

obteniéndose $\alpha = \sqrt{a}$.

Analicemos ahora la *velocidad de convergencia*; usando la fórmula (2.2) y poniendo $\varepsilon_n = x_n - \sqrt{a}$ (*error* en la $n-$ésima iteración) se ve que

$$\varepsilon_{n+1} = \frac{1}{2x_n}\varepsilon_n^2 \tag{2.3}$$

Siendo $x_n \geq \sqrt{a}$,

$$\varepsilon_{n+1} \leq \frac{1}{2\sqrt{a}}\varepsilon_n^2 \quad \forall n \tag{2.4}$$

o bien de (2.3)

$$\lim_{n\to\infty} \frac{|\varepsilon_{n+1}|}{|\varepsilon_n^2|} = \frac{1}{2\sqrt{a}} \tag{2.5}$$

Decimos entonces que los errores convergen *cuadráticamente* a cero. Veamos los errores relativos:

$$\frac{\varepsilon_{n+1}}{\sqrt{a}} \leq \frac{1}{2}\left(\frac{\varepsilon_n}{\sqrt{a}}\right)^2 \quad \forall n$$

es decir, $r_{n+1} \leq \dfrac{1}{2}r_n^2 \quad \forall n$

Sea $r_0 = 0.1$ entonces
$r_1 \leq 0.005$
$r_2 \leq 0.125 \cdot 10^{-4}$
$r_3 \leq 0.8 \cdot 10^{-10}$

En cada iteración por lo menos se duplica el número de cifras significativas.

Este ejemplo ilustra la construcción de un algoritmo o método iterativo para resolver una ecuación no lineal, dándose un análisis completo de la convergencia.

Resolver la ecuación $x^2 = a$ nos puede inducir el siguiente algoritmo

$$x_{n+1} = \frac{a}{x_n} \quad \forall n$$

pues $x^2 = a \quad \Leftrightarrow x = \dfrac{a}{x}$. Sin embargo para $x_0 \neq 0$ se tiene $x_1 = \dfrac{a}{x_0}$, $x_2 = x_0$, $x_3 = \dfrac{a}{x_0}$, $x_4 = x_0$, $\cdots$, es decir, la sucesión $\{x_n\}$ es oscilante.

**Definición 2.0.1.** *Sea $\{x_n\}$ una sucesión que converge a $\alpha$. Sea $\varepsilon_n = x_n - \alpha$. Si existe un número $p \geq 1$ y una constante $C > 0$ tal que*

$$\lim_{n\to\infty} \frac{|\varepsilon_{n+1}|}{|\varepsilon_n|^p} = C$$

*$p$ se llama el* ***orden de convergencia*** *de la sucesión.*

Para $p = 1, 2, 3$, la convergencia se dice lineal, cuadrática y cúbica, respectivamente. Si $1 < p < 2$ la convergencia es superlineal. En el caso lineal, la constante $C$ se llama la *tasa de convergencia* y necesariamente $0 < C < 1$.

También se puede definir orden de convergencia de la sucesión $\{x_n\}$ al número $p \geq 1$ tal que

$$|\varepsilon_{n+1}| \leq C|\varepsilon_n|^p \quad \text{para algún } C > 0, \quad \text{con } 0 < C < 1 \quad \text{para } p = 1$$

En el caso $p = 1$, podemos decir también que $\{x_n\}$ converge linealmente a $\alpha$ si $\exists\ 0 < C < 1$ tal que

$$|\varepsilon_n| \leq C^n|\varepsilon_0| \quad n \geq 0$$

**Definición 2.0.2.** *Decimos que $\alpha$ es una raíz de multiplicidad $p$ de una función $f$ si existe $h(x)$ continua en $\alpha$ tal que*

$$f(x) = (x - \alpha)^p h(x) \quad y\ \ h(\alpha) \neq 0$$

**Teorema 2.0.1.** *$f \in C^p[a, b]$ tiene un cero $\alpha$ de multiplicidad $p$ sii*

$$f^{(j)}(\alpha) = 0 \quad j = 0 : p - 1 \quad y \quad f^{(p)}(\alpha) \neq 0$$

**Prueba:**

Por definición

$$f(x) = (x - \alpha)^p h(x)$$

Aplicando la fórmula de Leibniz para la derivada de un producto:

$$\begin{aligned} f^{(j)}(x) &= \sum_{h=0}^{j} \binom{j}{k} [(x-\alpha)^p]^{(k)} h(x)^{(j-k)} \\ &= \sum_{k=0}^{j} \binom{j}{k} p(p-1)\cdots(p-k+1)(x-\alpha)^{(p-k)} h(x)^{(j-k)} \end{aligned}$$

Resulta entonces que si $j < p \quad f^{(j)}(\alpha) = 0$ y, si $j = p, \quad f^{(p)}(\alpha) = p!$ $h(\alpha) \neq 0$.

Recíprocamente,

$$\begin{aligned} f(x) &= f(\alpha) + f'(\alpha)(x-\alpha) + \frac{1}{2}f''(\alpha)(x-\alpha)^2 + \cdots \\ &\quad \cdots + \frac{1}{(p-1)!}f^{(p-1)}(\alpha)(x-\alpha)^{p-1} + \frac{f^{(p)}(\xi)}{p!}(x-\alpha)^p \qquad x < \xi < \alpha \end{aligned}$$

Aplicando la hipótesis se tiene que

$$f(x) = \frac{f^{(p)}(\xi)}{p!}(x-\alpha)^p$$

Siendo $f^{(p)}(x)$ continua en $\alpha$, $f(x) = (x-\alpha)^p h(x)$ con $h(x) = \dfrac{f^{(p)}(\xi)}{p!}$, $\xi = \xi(x)$.

□

## 2.1 Método de la bisección

Sea $f(x)$ continua en $[a,b]$ con $f(a) \cdot f(b) < 0$, entonces existe $\alpha$ tal que $f(\alpha) = 0$. Supongamos que $\alpha$ es raíz simple (usualmente $[a,b]$ se elige de manera que solo contenga una raíz de $f$, pero el algoritmo siempre será convergente a alguna raíz $\alpha$ de $[a,b]$).

**Algoritmo 1:** Bisección(f,a,b,raíz,$\varepsilon$)

1. Poner $c = a + (b-a)/2$
2. Si $c - a \leq \varepsilon$, entonces raíz= $c$. Stop
3. Si signo$(f(a)) \neq$ signo$(f(c))$, entonces $b = c$; de lo contrario $a = c$.
4. Regresar a 1.

Observar que el intervalo $[a,b]$ se bisecta en $[a,c]$ y $[c,b]$; de allí el nombre. La raíz $\alpha$ está en $[a,c]$ o en $[c,b]$ y en consecuencia

$$|c-\alpha| \leq b - c = c - a$$

lo cual justifica el paso 2. Observar también la forma en que se calcula $c$. Esto es para evitar que el punto eventualmente salga del intervalo $[a,b]$. Por ejemplo si se usa aproximación por cortada con precisión $t = 3$ y $a = 0.982$, $b = 0.987$, entonces $\dfrac{a+b}{2} = 0.980$ mientras que $a + \dfrac{b-a}{2} = 0.984$. Cuando el algoritmo se termina $c$ será una aproximación de la raíz con

$$|c-\alpha| \leq \varepsilon$$

Observar además, que, con el fin de evitar multiplicaciones innecesarias o problemas de underflow y overflow, para determinar si la función cambia de signo, usamos la función signo en 3.

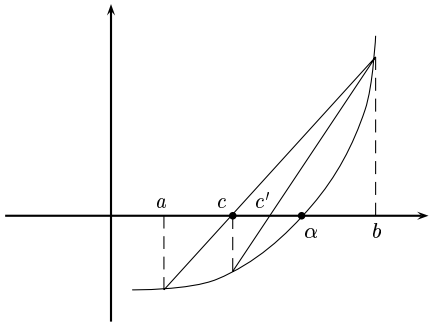
La demostración de la convergencia del método de la bisección queda como ejercicio. Analicemos la velocidad de convergencia; sea $c_n$ el valor $n-$ésimo de $c$ en el algoritmo. Luego, si $\lim_{n\to\infty} c_n = \alpha$ con $c_n = \dfrac{a_n + b_n}{2}$ resulta:

$$|c_n - \alpha| \leq \left(\frac{1}{2}\right)^n (b-a) \tag{2.6}$$

De la definición 2.0.1 y (2.6) sacamos que el método de la bisección es linealmente convergente con tasa igual a 0.5. Esta convergencia es lenta ya que en cada paso sólo se gana 1 bit. Siendo $10^{-1} \approx 2^{-3.3}$, se tiene que cada 3.3 pasos se gana un decimal. A pesar de ello, este método resulta muy apropiado para obtener buenas aproximaciones iniciales que serán usadas en la aplicación de otros métodos más eficientes. Notar que la convergencia no depende de la función $f$.

## 2.2 El Método de la Regula Falsi.

Supongamos $f(x)$ continua en $[a,b]$ con $f(a)\cdot f(b) < 0$.




El método consiste en aproximar la gráfica de $f$ por una recta que une $(a, f(a))$ y $(b, f(b))$. La raíz $c$ de la recta es usada como aproximación a una raíz $\alpha$ de $f(x)$:

$$c = b - f(b)\left(\frac{b-a}{f(b)-f(a)}\right)$$

**Algoritmo 2:** Regfalsi(f,a,b,raíz,$\varepsilon$)

1. $c = b - f(b)\left(\dfrac{b-a}{f(b)-f(a)}\right)$
2. Si signo$(f(a)) \neq$ signo$(f(c))$, entonces; $b = c$; si no $a = c$.

3. Si $|f(c)| \leq \varepsilon$, tomar raíz$= c$. Stop

4. De lo contrario regresar a 1.

Observar que la Regula Falsi (o método de la falsa posición) falla completamente en cuanto a dar un intervalo "*pequeño*" que contenga la raíz, ya que si $f$ es convexa y $f'(x) > 0$, el punto $c$ siempre queda a la izquierda de la raíz, y si $f$ es cóncava con $f'(x) > 0$, $c$ queda a la derecha de la raíz.

Mostraremos que la convergencia es lineal; en algunos casos puede ser más rápido que el método de la bisección, pero no siempre es así. Para demostrar la convergencia del método, introduzcamos las llamadas *diferencias divididas* de primero y segundo orden.

**Definición 2.2.1.** *Sean $a \neq b \neq c$,*

$$\textit{diferencia dividida de primer orden:} \quad f[a,b] = \frac{f(b) - f(a)}{b - a}$$

$$\textit{diferencia dividida de segundo orden:} \quad f[a,b,c] = \frac{f[b,c] - f[a,b]}{c - a}$$

**Propiedades:**

i) $f[a,b] = f[b,a]$.

ii) $f[x_1, x_2, x_3] = f[x_i, x_j, x_k]$ con $(i,j,k)$ una permutación cualquiera de $(1,2,3)$.

iii) Si $f$ es continua en $[a,b]$ y derivable en $(a,b)$, existe $\xi \in (a,b)$ tal que $f[a,b] = f'(\xi)$.

iv) Si $f$ es dos veces continuamente diferenciable en $x \in I = [\text{mín}(a,b,c), \text{máx}(a,b,c)]$, existe $\eta \in I$ tal que

$$f[a,b,c] = \frac{1}{2} f''(\eta)$$

Probemos iv). Supongamos $a < b < c$. Luego

$$\frac{f(c) - f(b)}{c - b} = f'(b) + \frac{1}{2} f''(\eta_1)(c - b) \tag{2.7}$$

y

$$\frac{f(a)-f(b)}{a-b} = f'(b) + \frac{1}{2}f''(\eta_2)(a-b) \tag{2.8}$$

Restando (2.7) y (2.8) y dividiendo por $(c-a)$, resulta

$$f[a,b,c] = \frac{1}{2}(\frac{c-b}{c-a}f''(\eta_1) + \frac{b-a}{c-a}f''(\eta_2)) = \frac{1}{2}(\alpha f''(\eta_1) + \beta f''(\eta_2))$$

Siendo $\alpha \geq 0,\ \beta \geq 0,\ \alpha+\beta = 1$ existe $\eta$ tal que $f[a,b,c] = \frac{1}{2}f''(\eta)$.

Regresemos al método de la Regula Falsi y supongamos que $f''$ es continua. Siendo

$$c = b - f(b)\left[\frac{b-a}{f(b)-f(a)}\right],$$

se tiene que

$$\begin{aligned}
c-\alpha &= b-\alpha-f(b)\left[\frac{b-a}{f(b)-f(a)}\right] = \frac{(b-\alpha)(f(b)-f(a)) - bf(b) + af(b)}{f(b)-f(a)} \\
&= \frac{-\alpha f(b) + \alpha f(a) - bf(a) + af(b)}{f(b)-f(a)} = \frac{(a-\alpha)f(b) - (b-\alpha)f(a)}{f(b)-f(a)} \\
&= (a-\alpha)(b-\alpha)\frac{1}{f(b)-f(a)}\left[\frac{f(b)}{b-\alpha} - \frac{f(a)}{a-\alpha}\right] \\
&= (a-\alpha)(b-\alpha)\frac{1}{\frac{f(b)-f(a)}{b-a}}\frac{\left[\frac{f(b)-f(\alpha)}{b-\alpha} - \frac{f(a)-f(\alpha)}{a-\alpha}\right]}{b-a} \\
&= (a-\alpha)(b-\alpha)\frac{f[a,\alpha,b]}{f[a,b]}
\end{aligned}$$

Usando las propiedades iii) y iv) de las diferencias divididas de primero y segundo orden, se tiene:

$$c-\alpha = \frac{1}{2}(a-\alpha)(b-\alpha)\frac{f''(\eta)}{f'(\xi)}, \quad \eta,\xi \in (a,b) \tag{2.9}$$

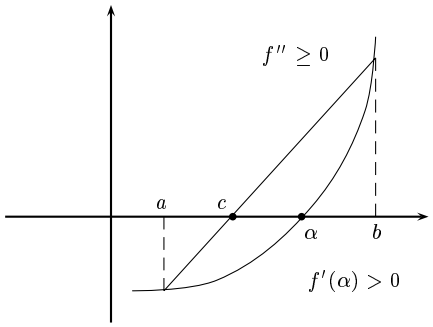
**Teorema 2.2.1.** *Sea $f$ dos veces continuamente diferenciable en $[a,b]$ con $\alpha$ una única raíz en $[a,b]$. Supongamos que $f(a)\cdot f(b) < 0,\ \ f'(\alpha) \neq 0,$ y $f''$*

*no cambia de signo en* $[a,b]$. *Si* $M = \left|\dfrac{\omega-\alpha}{2}\right| \max\limits_{x\in[a,b]} \left|\dfrac{f''(x)}{f'(x)}\right| < 1$, *con* $\omega = b$ *ó* $\omega = a$ *según el caso, entonces el método de la Regula Falsi converge. Esta convergencia es lineal.*

**Demostración:**

Supongamos que $f''(x) > 0$ en $[a,b]$, es decir, $f$ es convexa. Entonces, el segmento de recta que une $(x_1, f(x_1))$ y $(x_2, f(x_2))$ está por encima de la gráfica de $f$, $\forall x_1, x_2$ tales que $a \le x_1 \le x_2 \le b$. (Si $f'' < 0$, cambiar $f$ por $-f$ y hacer el mismo razonamiento).

**Caso 1:** $f'(\alpha) > 0$




Aquí resulta que $c$ siempre verifica que $a < c < \alpha$. En este caso $b-\alpha$ =constante. Sea $a_n = n-$ésimo valor de $a$ en el algoritmo. De (2.9) resulta

$$a_{n+1} - \alpha = \frac{1}{2}(a_n - \alpha)(b-\alpha)\frac{f''(\eta_n)}{f'(\xi_n)},$$

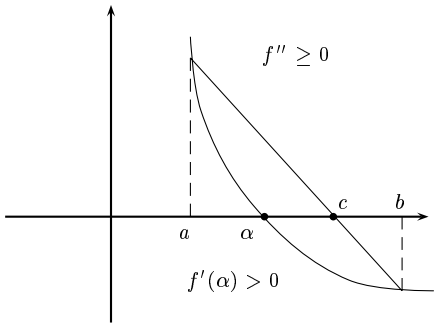
$$\eta_n, \xi_n \in (a_n, b)$$

Se tiene así que

$$|\xi_{n+1}| \le M|\xi_n| \quad \text{con} \quad M = \frac{b-\alpha}{2} \max_{x\in[a,b]} \left|\frac{f''(x)}{f'(x)}\right|$$

Por lo tanto $|\xi_{n+1}| \le M^n|\xi_0|$ y si $M < 1$ podemos asegurar que $\lim \xi_n = 0$ y que la convergencia es lineal.

**Caso 2:** $f'(\alpha) < 0$




En este caso $c$ siempre satisface

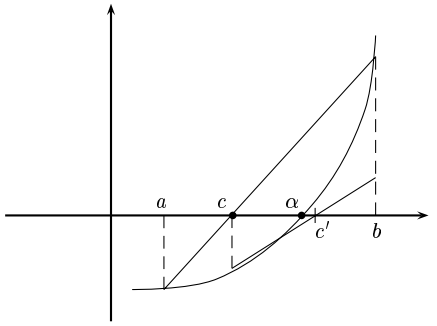
$$\alpha < c < b$$

y $\alpha - a$ =constante. Se obtienen las mismas conclusiones que en el caso 1, pero con

$$M = \left|\frac{a-\alpha}{2}\right| \max_{x\in[a,b]} \left|\frac{f''(x)}{f'(x)}\right|$$

## 2.3 Método de la Regula Falsi Modificado.

Reemplaza las secantes por rectas de menor pendiente hasta que $\alpha$ quede encerrada entre dos iterados sucesivos.




Este método es también conocido como el método de Illinois; se puede probar que el orden de convergencia es $\approx 1.442$

Referencia: Atkinson, N., (1973): *A New High Order Method of Regula Falsi Type for Computing a Root of an Equation*, BIT, vol. 13, pp 253-265.

**Algoritmo 3:** Regula falsi Modificado(f,a,b,raíz,$\varepsilon_1$,$\varepsilon_2$)

1. $c_{\text{viejo}} = a, \quad F_a = f(a) \quad F_b = f(b).$
2. $c_{\text{nuevo}} = b - F_b\left(\dfrac{b-a}{F_b - F_a}\right)$
3. Si signo$(f(a)) \neq$ signo$(f(c_{\text{nuevo}}))$, hacer $b = c_{\text{nuevo}}, \quad F_b = f(c_{\text{nuevo}})$ y si además signo$(f(c_{\text{viejo}}))$ = signo$(f(c_{\text{nuevo}}))$ hacer $F_a = \dfrac{F_a}{2}$; si signo$(f(a)) =$

signo$(f(c_{\text{nuevo}}))$ hacer $a = c_{\text{nuevo}}$, $F_a = f(c_{\text{nuevo}})$ y si además signo$(f(c_{\text{viejo}})) =$ signo$(f(c_{\text{nuevo}}))$ hacer $F_b = \dfrac{F_b}{2}$

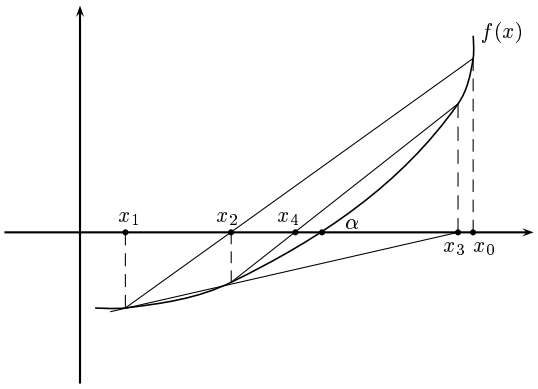
4. Si signo$(f(c_{\text{viejo}})) \neq$ signo$(f(c_{\text{nuevo}}))$ y $|c_{\text{viejo}} - c_{\text{nuevo}}| \leq \xi_1$ $|f(c_{\text{nuevo}})| \leq \xi_2$, hacer raíz z= $c_{\text{nuevo}}$. Stop

5. De lo contrario hacer $c_{\text{viejo}} = c_{\text{nuevo}}$ y regresar a 2.

## 2.4 El Método de la Secante.

La gráfica de $y = f(x)$ se aproxima por una recta secante en una vecindad de la raíz, pero esta secante se determina en dos iterados sucesivos, independientemente del cambio de signo de $f$. Tampoco la raíz queda encerrada entre los iterados

$$\begin{cases} x_0, x_1 \quad \text{iniciales} \\ \\ x_{n+1} = x_n - f(x_n)\dfrac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \quad n \geq 1 \end{cases} \tag{2.10}$$




Observar que a pesar de que en la fórmula intervienen dos iterados, solo hay que hacer una evaluación por cada iteración.

**Algoritmo 4:** Secante(f,a,b,raíz,$\varepsilon$,itmax,flag)

1. $c = b - f(b)\dfrac{b-a}{f(b)-f(a)}$

2. $a = b, \quad b = c$

3. Si $|f(c)| \leq \varepsilon$, entonces $c$ =raíz; flag= 1. Stop. Si iter $\geq$ itmax, entonces $c$ =raíz, flag= 2. Stop.

4. De lo contrario, regresar a 1.

Estudio de la convergencia: Vimos en (2.9) que

$$c - \alpha = \frac{1}{2}(a-\alpha)(b-\alpha)\frac{f''(\eta)}{f'(\xi)}, \quad \eta, \xi \in (a,b)$$

Aplicado a $\{x_n\}$ nos da:

$$x_{n+1} - \alpha = \frac{1}{2}(x_{n-1}-\alpha)(x_n-\alpha)\frac{f''(\eta_n)}{f'(\xi_n)}, \qquad \begin{array}{l} \eta_n \text{ entre } x_{n-1}, x_n, \alpha \\ \xi_n \text{ entre } x_{n-1}, x_n \end{array}$$

es decir,

$$\varepsilon_{n+1} = \frac{1}{2}\varepsilon_{n-1}\varepsilon_n\frac{f''(\eta_n)}{f'(\xi_n)}, \quad \eta_n \text{ entre } x_{n-1}, x_n, \alpha \text{ y } \xi_n \text{ entre } x_{n-1}, x_n \quad (2.11)$$

**Teorema 2.4.1.** *Sea $f$ dos veces continuamente diferenciable en un intervalo conteniendo a $\alpha$ y supongamos $f'(\alpha) \neq 0$. Si $x_0$ y $x_1$ están suficientemente próximos a $\alpha$, entonces la sucesión $\{x_n\}$ de iterados converge a $\alpha$. El orden de convergencia es*

$$p = \frac{1+\sqrt{5}}{2} \approx 1.62$$

**Demostración:**

Llamemos $M = \max\limits_{x \in I} \dfrac{|f''(x)|}{2|f'(x)|}$, siendo $I = [\alpha - \varepsilon, \alpha + \varepsilon]$, con $\varepsilon > 0$. De la igualdad (2.11) resulta que $\forall x_0, x_1 \in I$

$$|\varepsilon_2| \leq M|\varepsilon_0||\varepsilon_1|$$

Más aún, supongamos que $x_0, x_1$ se han escogido de manera que

$$\delta = \text{máx}\,\{M|\varepsilon_0|, M|\varepsilon_1|\} < 1$$

Con esta elección y siendo

$$M|\varepsilon_2| \leq M|\varepsilon_0|M|\varepsilon_1|$$

resulta que $M|\varepsilon_2| \leq \delta^2 < \delta < 1$ y en consecuencia, $|\varepsilon_2| < \dfrac{\delta}{M} = \text{máx}\{|\varepsilon_0|, |\varepsilon_1|\} < \varepsilon$, pues $x_0, x_1 \in I$.

La última desigualdad implica que $x_2 \in I$. Supongamos que $x_3, \ldots, x_n \in I \quad (M|\varepsilon_i| < \delta \quad i = 3, \ldots, n)$, entonces

$$|\varepsilon_{n+1}| \quad \leq \quad M|\varepsilon_{n-1}||\varepsilon_n|$$

$$M|\varepsilon_{n+1}| \quad \leq \quad M|\varepsilon_{n-1}|M|\varepsilon_n| < \delta^2 < \delta < 1 \quad \Rightarrow |\varepsilon_{n+1}| \leq \frac{1}{M} \leq \varepsilon$$

Se concluye que $x_n \in I \quad \forall n$.
Mostraremos la convergencia de $\{x_n\}$ hacia $\alpha$. Vemos que

$$\begin{array}{rcl} M|\varepsilon_0| & \leq & \delta \\ M|\varepsilon_1| & \leq & \delta \\ M|\varepsilon_2| & \leq & \delta^2 \\ M|\varepsilon_3| & \leq & M|\varepsilon_2|M|\varepsilon_1| \leq \delta^2\delta = \delta^3 \\ M|\varepsilon_4| & \leq & M|\varepsilon_3|M|\varepsilon_2| \leq \delta^3\delta^2 = \delta^5 \\ M|\varepsilon_5| & \leq & M|\varepsilon_4|M|\varepsilon_3| \leq \delta^5\delta^3 = \delta^8 \end{array}$$

Así si $M|\varepsilon_{n-1}| \leq \delta^{q_{n-1}}$ y $M|\varepsilon_n| \leq \delta^{q_n}$ resulta

$$M|\varepsilon_{n+1}| \leq M|\varepsilon_n|M|\varepsilon_{n-1}| \leq \delta^{q_n+q_{n-1}} = \delta^{q_{n+1}}$$

Se tiene que la sucesión $\{q_n\}$ verifica

$$\left\{ \begin{array}{l} q_0 = q_1 = 1 \\ q_{n+1} = q_n + q_{n-1} \end{array} \right.$$

cuya solución es

$$q_n = \frac{1}{\sqrt{5}} \left[ \left( \frac{1+\sqrt{5}}{2} \right)^{n+1} - \left( \frac{1-\sqrt{5}}{2} \right)^{n+1} \right] \quad n \geq 0$$

llamada la solución de Fibonacci. Los números $\dfrac{1+\sqrt{5}}{2} \approx 1.618$ y $\dfrac{1-\sqrt{5}}{2} \approx -0.618$ se conocen como los números de Fibonacci. Por lo tanto si $n$ es grande

$$q_n \approx \frac{1}{\sqrt{5}}(1.618)^{n+1} \quad \text{y} \quad q_n \to \infty \quad \text{si} \quad n \to \infty.$$

Como $|\varepsilon_n| \leq \dfrac{\delta^{q_n}}{M}$ con $\delta < 1$, se obtiene $\lim\limits_{n\to\infty} |\varepsilon_n| = 0$.

Mediante un análisis más cuidadoso se puede probar que

$$\lim_{n\to\infty} \frac{|\varepsilon_{n+1}|}{|\varepsilon_n|^p} = \left| \frac{f''(\alpha)}{2f'(\alpha)} \right|^{\frac{\sqrt{5}-1}{2}} \quad \text{con} \quad p = \frac{1+\sqrt{5}}{2}$$

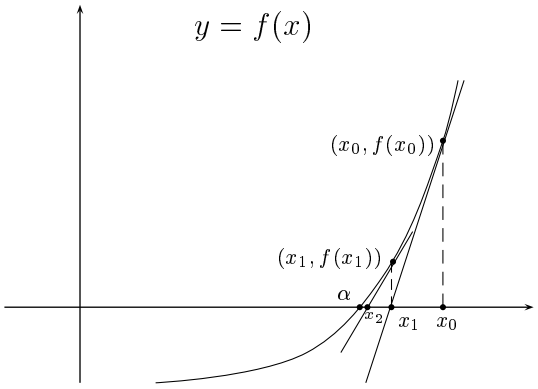
Ver Ostroroski A.M. (1973): *Solution of Equation in Euclidean and Banach Spaces.*

**Observaciones:**

1) Que $x_0, x_1$ estén suficientemente cerca de $\alpha$ significa que $M \text{ máx } \{|\varepsilon_0|, |\varepsilon_1|\} < 1 \quad \text{y} \quad |\varepsilon_0| \leq \varepsilon \quad |\varepsilon_1| \leq \varepsilon$.

2) Si $\left|\dfrac{f''(\alpha)}{2f'(\alpha)}\right|$ es grande, habrá que tomar $x_0, x_1$ muy próximos a $\alpha$. Es posible encontrar ejemplos en donde se muestre que si $x_0, x_1$ no están bien elegidos el método puede ser divergente.

## 2.5 El Método de Newton




Consiste en aproximar $y = f(x)$ por una recta tangente a la gráfica de $f$ y usar la raíz de la recta como aproximación a la raíz $\alpha$.
Si $x_0$ es una aproximación inicial de $\alpha$, la recta tangente a la gráfica de $f(x)$ en $(x_0, f(x_0))$ corta al eje $x$ en

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Entonces, el método de Newton es:

$$\begin{cases} x_0 \text{ inicial} \\ x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}, \quad n \geq 0 \end{cases} \tag{2.12}$$

En cada paso se necesitan una evaluación de la función y otra de su derivada; esto último puede resultar costoso.

Observar que siendo $f'(x_n) \approx \dfrac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} \quad n \geq 1$, el método de Newton no es sino una *"variante"* del método de la secante.

Estudiemos la convergencia, expandiendo $f(x)$ alrededor de $x_n$:

$$f(x) = f(x_n) + (x - x_n)f'(x_n) + \frac{(x - x_n)^2}{2}f''(\xi_n) \qquad \xi_n \in intervalo(x_n, x)$$

Para $x = \alpha$,

$$0 = f(x_n) + (\alpha - x_n)f'(x_n) + \frac{(\alpha - x_n)^2}{2}f''(\xi_n) \qquad \xi_n \in intervalo(x_n, x)$$

$$0 = \frac{f(x_n)}{f'(x_n)} + (\alpha - x_n) + (\alpha - x_n)^2\frac{f''(\xi_n)}{2f'(x_n)}$$

Usando (2.12) se obtiene

$$x_{n+1} - \alpha = (x_n - \alpha)^2\frac{f''(\xi_n)}{2f'(x_n)} \qquad \xi_n \in intervalo(x_n, \alpha)$$

o lo que es lo mismo

$$\varepsilon_{n+1} = \varepsilon_n^2\frac{f''(\xi_n)}{2f'(x_n)}, \qquad \xi_n \in int(x_n, \alpha) \tag{2.13}$$

**Teorema 2.5.1.** *Sea $f$ 2 veces continuamente diferenciable $\forall x$ en algún entorno de $\alpha$ con $f(\alpha) = 0$ y $f'(\alpha) \neq 0$. Si $x_0$ está suficientemente próximo a $\alpha$, los iterados $x_n$ de (2.12) convergen a $\alpha$. La convergencia es cuadrática.*

**Demostración:**

Tomemos $I = [\alpha - \varepsilon, \alpha + \varepsilon]$, $\varepsilon > 0$ $\quad M = \frac{1}{2}\max\limits_{x\in I}\left|\frac{f''(x)}{f'(x)}\right|$ y elijamos $x_0 \in I$, tal que $M|\varepsilon_0| < 1$. Usando (2.13) se tiene que

$$|\varepsilon_1| \leq M|\varepsilon_0|^2 = (M|\varepsilon_0|)|\varepsilon_0| < |\varepsilon_0| \leq \varepsilon \quad \Rightarrow \quad x_1 \in I$$

Además

$$M|\varepsilon_1| \leq (M|\varepsilon_0|)^2 < 1$$

Supongamos que $x_n \in I$ y $M|\varepsilon_n| < 1$ entonces

$$|\varepsilon_{n+1}| \leq M|\varepsilon_n|^2 \leq (M|\varepsilon_n|)|\varepsilon_n| < |\varepsilon_n| < \varepsilon \quad \Rightarrow \quad x_{n+1} \in I$$

y

$$M|\varepsilon_{n+1}| \leq (M|\varepsilon_n|)^2 < 1 \tag{2.14}$$

Concluimos que $x_n \in I$ y $M|\varepsilon_n| < 1 \quad \forall n$. Además, de (2.14) resulta

$$\begin{aligned} M|\varepsilon_{n+1}| &\leq (M|\varepsilon_0|)^{2^n} \\ |\varepsilon_{n+1}| &\leq \frac{1}{M}(M|\varepsilon_0|)^{2^n} \end{aligned}$$

Al ser $M|\varepsilon_0| < 1$ resulta $\lim_{n\to\infty} \varepsilon_n = 0$ y por lo tanto $x_n \to \alpha$.

Además, tomando límites en (2.13) se obtiene

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n^2} = \frac{f''(\alpha)}{2f'(\alpha)}$$

por la continuidad de $f''$ y que $\xi_n \to \alpha$ cuando $n \to \infty$. □

**Nota:** La convergencia resulta si

$$M|x_0 - \alpha| = M|\varepsilon_0| < 1,$$

es decir, si $|x_0 - \alpha| < \frac{1}{M}$, lo cual nos dice cuán cerca $x_0$ debe estar de $\alpha$.

**Algoritmo 5:** Newton(f,df,$x_0$,$\varepsilon$,raíz,itmax,flag)

1. iter=1
2. $denom = f'(x_0)$
3. Si $denom = 0$, entonces $flag = 2$, stop.
4. $x_1 = x_0 - f(x_0)/denom$
5. Si $|x_1 - x_0| \leq \varepsilon$ entonces $flag = 0$, raíz z= $x_1$, stop.
6. Si $iter = itmax$, hacer $flag = 1$, raíz= $x_1$, stop.
7. De lo contrario $iter = iter + 1$, $x_0 = x_1$. Regresar a 2.

Más adelante justificaremos el criterio de parada $|x_n - x_{n-1}| \leq \varepsilon$ que estamos usando.

## 2.5 Comparación entre el método de Newton y el de la Secante.

Vamos a calcular el tiempo necesario empleado por ambos métodos para alcanzar la raíz $\alpha$ dentro de una misma tolerancia $\varepsilon$. Denotemos $\{x_n\}$ la sucesión de iterados generados por el método de Newton e $\{y_n\}$ la sucesión de iterados generados por el método de la secante. Para simplificar el análisis, supongamos que $x_0 = y_0 \approx \alpha \quad y_1 \approx \alpha$.

$$\begin{aligned} x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)}, \quad n \geq 0 \\ \\ y_{n+1} &= y_n - f(y_n)\frac{y_n - y_{n-1}}{f(y_n) - f(y_{n-1})} \end{aligned}$$

con

$$\lim_{n\to\infty} \frac{|\varepsilon_{n+1}|}{|\varepsilon_n|^2} = \frac{1}{2}\left|\frac{f''(\alpha)}{f'(\alpha)}\right| = c \quad \text{para Newton}$$

y

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^{\frac{1+\sqrt{5}}{2}}} = \left|\frac{f''(\alpha)}{2f'(\alpha)}\right|^{\frac{\sqrt{5}-1}{2}} = d \quad \text{para la secante}$$

Notar que $d = c^{s-1}$ con $s = \dfrac{1+\sqrt{5}}{2}$.

Para $n$ grande,

$$\begin{aligned} |\varepsilon_{n+1}| &\approx c|\varepsilon_n|^2 \\ \\ |e_{n+1}| &\approx d|e_n|^s \quad \text{con } s = \frac{1+\sqrt{5}}{2} \end{aligned}$$

Inductivamente

$$c|\varepsilon_n| \approx (c|\varepsilon_0|)^{2^n}$$

o lo que es lo mismo

$$\boxed{|\varepsilon_n| \approx \frac{1}{c}(c|\varepsilon_0|)^{2^n} \quad n \geq 0, \quad \text{Newton}} \tag{2.15}$$

Análogamente para la secante:

$$|e_n| \approx d|e_{n-1}|^s \approx d^{1+s+\cdots+s^{n-1}}|e_0|^{s^n} = d^{\frac{s^n-1}{s-1}}|e_0|^{s^n} = c^{s^n-1}|e_0|^{s^n}$$

es decir,

$$\boxed{|e_n| \approx \frac{1}{c}(c|e_0|)^{s^n} \quad n \geq 0, \quad \text{Secante}} \tag{2.16}$$

¿Cuántos pasos debemos hacer en Newton para que $|x_n - \alpha| = |\varepsilon_n| \leq \varepsilon$?

De (2.15) obtenemos

$$(c|\varepsilon_0|)^{2^n} \leq c\varepsilon$$

$$2^n \ln(c|\varepsilon_0|) \leq \ln c\varepsilon \quad \Rightarrow \quad 2^n \geq \frac{\ln c\varepsilon}{\ln(c|\varepsilon_0|)}$$

es decir,

$$\boxed{n \geq \frac{k}{\ln 2}, \quad k = \ln\left(\frac{\ln c\varepsilon}{\ln(c|\varepsilon_0|)}\right), \quad \text{Newton}}$$

Sea $m$ el tiempo necesario para calcular $f(x)$ y $pm$ el tiempo necesario para calcular $f'(x)$. El tiempo mínimo para obtener la precisión $\varepsilon$ deseada con el método de Newton es

$$\boxed{T_N = (m + mp) \cdot n = m(p+1)\frac{k}{\ln 2}, \quad \text{Newton}}$$

Para el método de la secante se ve de (2.16) que:

$$|e_n| \leq \varepsilon \quad \Rightarrow \quad \frac{1}{c}[c(|\varepsilon_0|)^{s^n}] \leq \varepsilon$$

$$(c|e_0|)^{s^n} \leq c\varepsilon \quad \Rightarrow \quad s^n \ln(c|e_0|) \leq \ln c\varepsilon \quad \Rightarrow \quad s^n \geq \frac{\ln c\varepsilon}{\ln(c|\varepsilon_0|)}$$

$$\Rightarrow \boxed{n \geq \frac{k}{\ln s}, \quad k = \ln\left(\frac{\ln c\varepsilon}{\ln(c|\varepsilon_0|)}\right), \quad \text{Secante}}$$

pues además hemos supuesto que $|e_0| = |\varepsilon_0|$. El tiempo mínimo para la secante es

$$T_s = m \cdot n = m\frac{k}{\ln s}$$

Para comparar $T_s$ y $T_N$ hacemos

$$\frac{T_s}{T_N} = \frac{\ln 2}{(1+p)\ln s}$$

El método de la secante será más rápido que el de Newton si $T_s < T_N$, es decir, si $\dfrac{\ln 2}{\ln s} < 1 + p \quad \Rightarrow \quad p > \dfrac{\ln 2}{\ln s} - 1 \approx 0.44$.

**Conclusión:** Si el tiempo para calcular $f'(x)$ es más del 44% del tiempo necesario para calcular $f(x)$, el método de la secante es más eficiente.

## 2.6 El Método de Steffensen

Es una modificación del método de Newton, análogo al de la secante pero usando una aproximación diferente de la derivada de $f$.

$$\begin{cases} x_0 \\ \\ x_{n+1} = x_n - \dfrac{f(x_n)}{\dfrac{f(x_n + f(x_n)) - f(x_n)}{f(x_n)}}, \quad n \geq 0 \end{cases}$$

Se puede demostrar que

$$\lim \frac{\varepsilon_{n+1}}{\varepsilon_n^2} = \frac{1}{2}\frac{f''(\alpha)}{f'(\alpha)} + \frac{1}{2}f''(\alpha),$$

es decir que el método es cuadráticamente convergente. Observar que se hacen dos evaluaciones por paso.

## 2.7 Teoría General de los Métodos Iterativos

Los métodos de Newton, secante y Steffensen pueden considerarse casos particulares del siguiente método iterativo más general: sea $x_{n+1}$ determinado por evaluaciones de la función y/o de las derivadas en los puntos $x_n, x_{n-1}, \ldots, x_{n-m+1}$ y pongamos

$$x_{n+1} = g(x_n, x_{n-1}, \ldots, x_{n-m+1}), \quad n = m, m+1, \ldots$$

La función $g$ se llama *función de iteración*, y $\{x_n\}$ la *sucesión de iterados*. En los métodos particulares estudiados se tiene que

$$g(x) = x - \frac{f(x)}{f'(x)}, \quad m = 1, \quad \text{para Newton}$$

$$g(x,y) = x - f(x)\frac{x-y}{f(x)-f(y)}, \quad m = 2, \quad \text{para la secante}$$

$$g(x) = x - \frac{f(x)}{\dfrac{f(x+f(x))-f(x)}{f(x)}}, \quad m = 1, \quad \text{para Steffensen}$$

La teoría general de los métodos iterativos es simple cuando $m = 1$, es decir,

$$x_{n+1} = g(x_n), \qquad n = 0, 1, 2, \ldots \tag{2.17}$$

Sea $\{x_n\}$ una sucesión generada por (2.17) para un $x_0$ inicial dado. Supongamos que $x_n \to \alpha$ y que $g$ es continua. Tomando límites en (2.17):

$$\alpha = g(\alpha) \tag{2.18}$$

Si (2.18) es cierto decimos que $\alpha$ es un *punto fijo de* $g$. Para resolver el problema $f(x) = 0$, podemos construir una función $g$ tal que un punto fijo de $g$ sea una raíz de $f$.

**Ejemplo 2.7.1.** *La ecuación* $x^3 - x - 1 = 0$ *tiene una raíz entre 1.25 y 1.5. Vamos a asociarle distintas funciones de iteración a esta ecuación y estudiaremos el comportamiento de los iterados:*

*a)* $g_1(x) = x^3 - 1$ *b)* $g_2(x) = \sqrt[3]{x+1}$ *c)* $g_3(x) = \dfrac{1}{x^2-1}$

*d)* $g_4(x) = \dfrac{2x^3+1}{3x^2-1}$

*Tomemos para todos los casos,* $x_0 = \dfrac{5}{4} = 1.25$. *Vemos que*

| (a) | (b) | (c) | (d) |
|---|---|---|---|
| 1.250000000000 | 1.250000000000 | 1.250000000000 | 1.250000000000 |
| 0.953125000000 | 1.310370697104 | 1.777777777778 | 1.330508474576 |
| −0.134136199951 | 1.321987115986 | 0.462857142857 | 1.324748959227 |
| −1.002413448279 | 1.324199039542 | −1.272647938830 | 1.324717958140 |
| −2.007257833092 | 1.324619383172 | 1.161385910881 | 1.324717957245 |
| −9.087410436263 | 1.324699233154 | 0.623231105322 | 1.324717957245 |
| −751.447699656674 | 1.324714400655 | −1.163510106951 | 1.324717957245 |

Desarrollaremos una teoría que nos permita explicar estos comportamientos.

**Lema 2.7.1.** *Sea $g(x)$ continua en $[a,b]$, $g:[a,b]\to[a,b]$. Entonces $x=g(x)$ tiene un punto fijo en $[a,b]$.*

**Prueba:**

Sea $f(x)=x-g(x)$; si $f(a)=0$ entonces $a$ es un punto fijo, si no $f(a)=a-g(a)<0$. Si $f(b)=0$ entonces $b$ es un punto fijo; si no $f(b)=b-g(b)>0$. Por lo tanto, si $f(b)>0$ y $f(a)<0$, existe $\alpha\in(a,b)$ tal que $f(\alpha)=0$, o lo que es lo mismo $\alpha=g(\alpha)$.

□

**Lema 2.7.2.** *Sea $g(x)$ definida en $[a,b]$ y supongamos que $g:[a,b]\to[a,b]$ y que además existe $0<\lambda<1$ con*

$$|g(x)-g(y)|\leq\lambda|x-y|,\ \ \forall x,y\in[a,b]$$

*($g$ se llama una contracción). Luego $x=g(x)$ tiene una única solución $\alpha$ en $[a,b]$ y los iterados $x_{n+1}=g(x_n),\ \ n\geq 0$ convergen a $\alpha$, cualquiera sea $x_0\in[a,b]$.*

**Prueba:**

**(Unicidad)** Sean $\alpha$ y $\beta$ dos soluciones en $[a,b]$, $\alpha\neq\beta$, entonces $|\alpha-\beta|=|g(\alpha)-g(\beta)|<\lambda|\alpha-\beta|<|\alpha-\beta|$, absurdo.

**(Existencia)** Por el lema 2.7.1, siendo $g$ una contracción, ella es continua, y por lo tanto existe $\alpha\in[a,b]$ tal que $\alpha=g(\alpha)$.

**(Convergencia)** Por inducción es fácil ver que si $x_0\in[a,b]$, $x_n\in[a,b]\ \ \forall n$. Resulta así que:

$$|x_{n+1}-\alpha|=|g(x_n)-g(\alpha)|\leq\lambda|x_n-\alpha|\leq\lambda^{n+1}|x_0-\alpha|$$

y por lo tanto, cuando $n \to \infty$, $x_{n+1} \to \alpha$, pues $\lambda^{n+1} \to 0$.

□

**Teorema 2.7.1.** *Sea $g(x)$ continuamente diferenciable en $[a,b]$, $g : [a,b] \to [a,b]$ y $\lambda = \max\limits_{x\in[a,b]} |g'(x)| < 1$. Entonces*

*(1) $x = g(x)$ tiene un única solución en $[a,b]$.*

*(2) $\forall x_0$ en $[a,b]$ con $x_{n+1} = g(x_n)$, $n \geq 0$, los iterados $x_n$ convergen a $\alpha$.*

*(3) $\lim\limits_{n\to\infty} \dfrac{\varepsilon_{n+1}}{\varepsilon_n} = g'(\alpha)$, con $\varepsilon_n = x_n - \alpha$*

**Demostración:**

Si $g$ es diferenciable en $[a,b]$, existe $\xi \in (a,b)$ tal que

$$g(x) - g(y) = g'(\xi)(x - y)$$

Por lo tanto $|g(x) - g(y)| \leq \lambda|x - y|$, $\lambda < 1$; de aquí se tiene que (1) y (2) resultan de los lemas (2.7.1) y (2.7.2). Estudiemos la velocidad de convergencia:

$$x_{n+1} - \alpha = g(x_n) - g(\alpha) = g'(\xi_n)(x_n - \alpha), \quad \text{para } \xi_n \text{ entre } x_n \text{ y } \alpha$$

Como $x_n \to \alpha$ y $g'$ es continua, $g'(\xi_n) \to g'(\alpha)$, de donde

$$\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{x_n - \alpha} = g'(\alpha).$$

□

**Teorema 2.7.2.** *Sea $\alpha$ una solución de $x = g(x)$ y supongamos que $g(x)$ es continuamente diferenciable en algún entorno de $\alpha$ con $|g'(\alpha)| < 1$. Si $x_0$ se elige lo suficientemente próximo a $\alpha$ entonces los resultados 1),2) y 3) del Teorema 2.7.1 son válidos.*

**Demostración:**

Sea $I = [\alpha - \varepsilon, \alpha + \varepsilon]$ con $\varepsilon > 0$ elegido de manera que $\lambda = \max\limits_{x\in I} |g'(x)| < 1$. Entonces $g$ es una contracción si $g(I) \subseteq I$. En efecto, si $x \in I \;\Rightarrow\; |\alpha - x| \leq \varepsilon$ y siendo $|g(\alpha) - g(x)| \leq \lambda|\alpha - x| < |\alpha - x|$ resulta $|g(\alpha) - g(x)| \leq \varepsilon$, es decir $|\alpha - g(x)| \leq \varepsilon$ y por lo tanto $g(x) \in I$. Aplicando el teorema anterior obtenemos los resultados deseados.

□

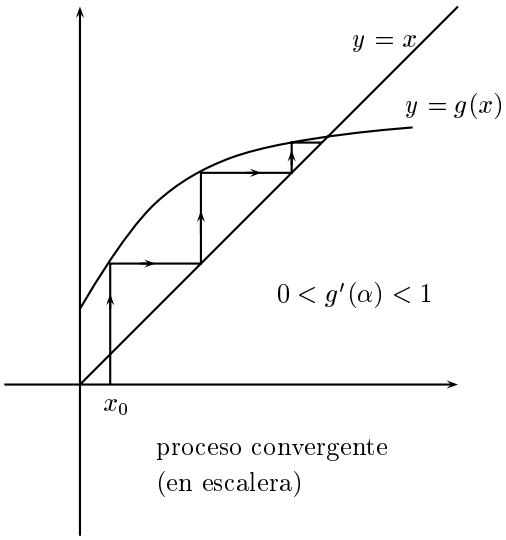
**Observaciones:**

1) Si $|g'(\alpha)| > 1$ y $x_n \approx \alpha$ se tiene que $x_{n+1} - \alpha = g'(\xi_n)(x_n - \alpha)$ implica $|x_{n+1} - \alpha| > |x_n - \alpha|$ y la convergencia no es posible.

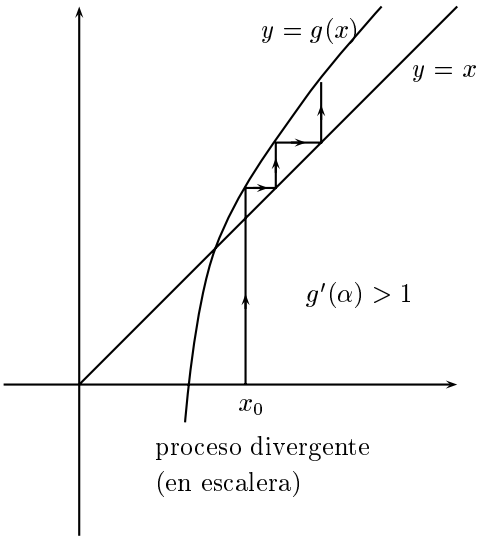
2) Sea $g(x) = \tanh x$. La raíz es $\alpha = 0$; la derivada es $g'(x) = \dfrac{1}{\cosh^2 x}$, es decir $g'(0) = 1$. En este caso $x_{n+1} = g(x_n)$ es convergente pues $|g'(x)| < 1 \ \ \forall x \neq \alpha = 0$.

3) Sea $x = g(x)$, $g(x) = \dfrac{a}{x}$; entonces $g'(x) = -\dfrac{a}{x^2}$, $\quad g'(\sqrt{a}) \equiv -1$ pero $x_{n+1} = \dfrac{a}{x_n}$ es divergente. Es fácil ver que si $0 < x_0 < \sqrt{a}$, $|g'(x_n)| > 1 \ \ \forall n$.

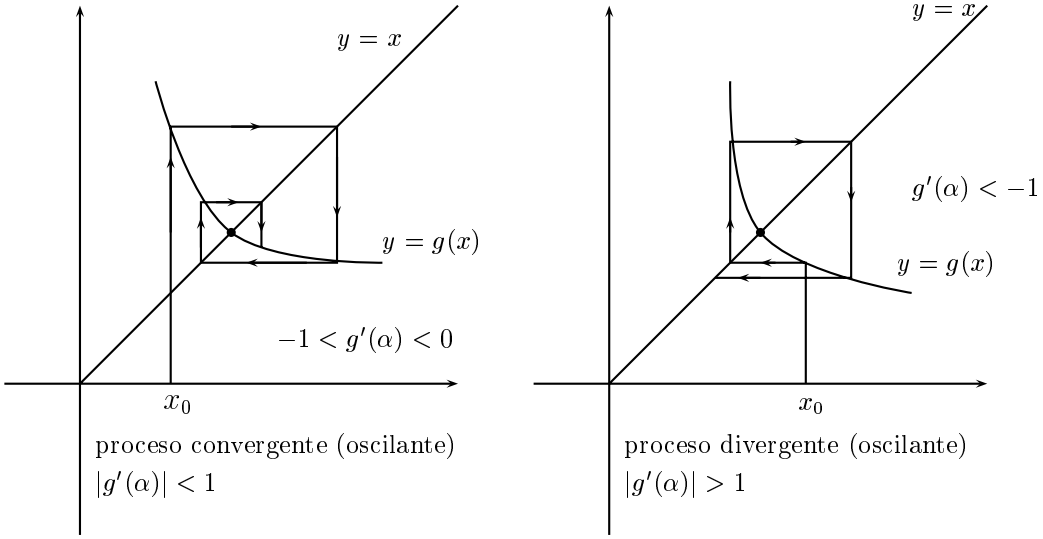
El proceso iterativo $x_{n+1} = g(x_n)$ puede representarse geométricamente considerando que el punto fijo $\alpha$ es el punto de intersección de las funciones $y = x$ e $y = g(x)$. Podemos visualizar cuatro casos: (a) $0 < g'(\alpha) < 1$, (b) $g'(\alpha) > 1$, (c) $-1 < g'(\alpha) < 0$, (d) $g'(\alpha) < -1$.




proceso convergente
(en escalera)




proceso divergente
(en escalera)




**Ejemplo 2.7.2.** *Para el ejemplo 2.7.1 con* $\alpha \approx 1.3$ *se tiene*

*a)* $g(x) = x^3 - 1 \quad g'(x) = 3x^2, \qquad g'(\alpha) \approx 4.2$

*b)* $g(x) = \sqrt[3]{x+1} \quad g'(x) = \dfrac{1}{3(x+1)^{2/3}}, \qquad g'(\alpha) \approx 0.19$

*c)* $g(x) = \dfrac{1}{x^2-1} \quad g'(x) = \dfrac{-2x}{(x^2-1)^2}, \qquad g'(\alpha) \approx -5.5$

*d)* $g(x) = \dfrac{2x^3+1}{3x^2-1} \quad g'(x) = \dfrac{6x(x^3-x-1)}{(3x^2-1)^2}, \quad g'(\alpha) = 0$

## 2.7 Métodos iterativos de orden superior.

**Teorema 2.7.3.** *Sea* $\alpha$ *una raíz de* $x = g(x)$*;* $g$ *una función* $p$ *veces continuamente diferenciable* $\forall x$ *cerca de* $\alpha$*,* $p \geq 2$*. Supongamos además que*

$$g'(\alpha) = g''(\alpha) = \cdots = g^{(p-1)}(\alpha) = 0, \quad g^{(p)}(\alpha) \neq 0$$

*Si* $x_0$ *se elige suficientemente próximo a* $\alpha$*, la iteración*

$$x_{n+1} = g(x_n) \quad n \geq 0$$

*tendrá al menos un orden de convergencia igual a* $p$ *y*

$$\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{(x_n - \alpha)^p} = \lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n^p} = \frac{g^{(p)}(\alpha)}{p!}$$

**Demostración:**

Siendo

$$x_{n+1} = g(x_n) = g(\alpha) \quad + \quad (x_n - \alpha)g'(\alpha) + \cdots$$

$$\cdots + \frac{(x_n - \alpha)^{p-1}}{(p-1)!} g^{(p-1)}(\alpha) + \frac{(x_n - \alpha)^p}{p!} g^{(p)}(\xi_n)$$

con $\xi_n$ entre $\alpha$ y $x_n$ y $g^{(j)}(\alpha) = 0 \quad j = 0, \ldots, (p-1)$, resulta

$$x_{n+1} - \alpha = \frac{(x_n - \alpha)^p}{p!} g^{(p)}(\xi_n)$$

Como $g'(\alpha) = 0 < 1, \quad x_n \to \alpha$ y por lo tanto,

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n^p} = \frac{g^{(p)}(\alpha)}{p!}$$

□

**Aplicación:** Método de Newton: En este caso la función de iteración es

$$g(x) = x - \frac{f(x)}{f'(x)} \quad \text{y} \quad g(\alpha) = \alpha \quad \Leftrightarrow \quad f(\alpha) = 0$$

Derivemos:

$$g'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2};$$

para $x = \alpha$ se tiene

$$g'(\alpha) = 1 - \frac{[f'(\alpha)]^2}{[f'(\alpha)]^2} = 0 \quad \forall \text{ función} f$$

La derivada segunda es

$$g''(x) = \frac{(f'(x)f''(x) + f(x)f''(x))(f'(x))^2 - 2f(x)[f''(x)]^2 f'(x)}{[f'(x)]^4}$$

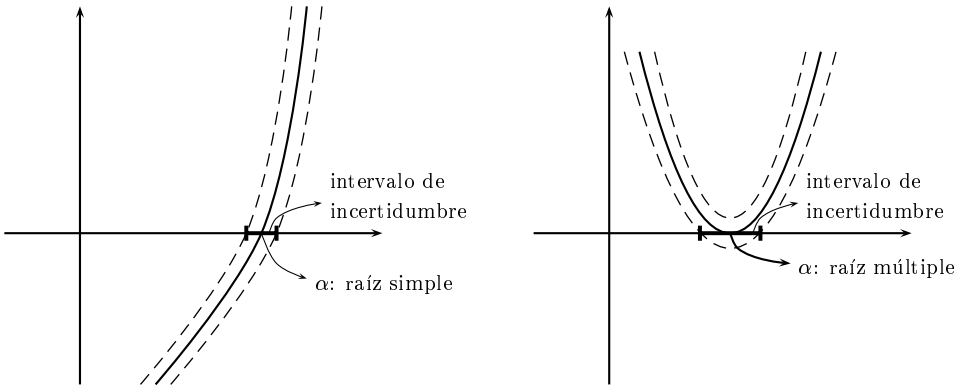
Así $g''(\alpha) = \dfrac{f''(\alpha)}{f'(\alpha)}$; si $f''(\alpha) = 0$ el método resultaría de tercer orden. Pero esto ya sería para algún caso particular y no para toda $f$. Por lo tanto

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n^2} = \frac{1}{2}\frac{f''(\alpha)}{f'(\alpha)}$$

y coincide con lo demostrado anteriormente.

## 2.8 Raíces Múltiples

Cuando se aproxima una raíz de $f(x)$ existe un intervalo de incertidumbre que se alarga cuando la raíz es múltiple.




La franja da la región de incertidumbre en la evaluación de $f(x)$ debido a errores de redondeo y a la aritmética finita.

En la discusión del método de Newton y de la secante supusimos que $f'(\alpha) \neq 0$, es decir que $\alpha$ es raíz simple. Si $f'(\alpha) = 0$ los métodos pierden su orden de convergencia. Veamos el caso del método de Newton.

Sea $k(x) = \dfrac{f(x)}{f'(x)}$; siendo $f(x) = (x-\alpha)^p h(x)$ con $h(\alpha) \neq 0$ y continuamente diferenciable, entonces

$$k(x) = \frac{(x-\alpha)h(x)}{ph(x) + (x-\alpha)h'(x)} \quad \text{y} \quad g(x) = x - \frac{(x-\alpha)h(x)}{ph(x) + (x-\alpha)h'(x)} = x - k(x)$$

Calculemos la derivada:

$$k'(x) = \frac{h(x)}{ph(x) + (x-\alpha)h'(x)} + (x-\alpha)\frac{d}{dx}\left[\frac{h(x)}{ph(x) + (x-\alpha)h'(x)}\right]$$

Luego, $k'(\alpha) = \dfrac{h(\alpha)}{ph(\alpha)} = \dfrac{1}{p}$ y $g'(\alpha) = 1 - \dfrac{1}{p} \neq 0$ para $p \neq 1$. Es decir, el método de Newton es lineal con tasa de convergencia $1 - \dfrac{1}{p}$. Para $p = 2$, la tasa es $\dfrac{1}{2}$, la misma que la del método de la bisección. Observar que $\alpha$ es raíz simple de $k(x)$.

## 2.8 Método de Newton para raíces múltiples.

Modifiquemos el método de Newton poniendo

$$x_{n+1} = x_n - p\frac{f(x_n)}{f'(x_n)} \qquad n \geq 0$$

siendo $p$ la multiplicidad de la raíz. Luego $g'(\alpha) = 0$ con $g(x) = x - pk(x)$ y la convergencia cuadrática se recupera. El inconveniente de este método es que requiere conocimiento previo de la multiplicidad $p$ de la raíz.

## 2.8 Métodos generales.

Si no se conoce a priori la multiplicidad $p$, podemos usar la función $k(x)$ definida anteriormente, la cual, como hemos visto, tiene a $\alpha$ como una raíz simple. Todos los métodos estudiados pueden aplicarse a $k(x)$.

**Newton:**

$$x_{n+1} = x_n - \frac{k(x_n)}{k'(x_n)}$$

con $k(x_n) = \dfrac{f(x_n)}{f'(x_n)}$ y $k'(x_n) = 1 - \dfrac{f''(x_n)}{f'(x_n)}k(x_n)$.

**Secante:**

$$x_{n+1} = x_n - k(x_n)\frac{x_n - x_{n-1}}{k(x_n) - k(x_{n-1})}, \quad n \geq 0$$

Observemos que en Newton se necesitan calcular $f, f', f''$ y en la secante $f$ y $f'$. El intervalo de incertidumbre para la raíz permanecerá sin modificaciones ya que $h(x)$ contiene la multiplicidad dentro de ella.

Siempre que se sepa la multiplicidad de una raíz, es conveniente aplicar diferenciación para hacerla simple (si $\alpha$ es raíz de multiplicidad $p$ de $f$ entonces $\alpha$ es raíz simple de $f^{(p-1)}(x)$).

# 2.9 Estimación de la tasa de convergencia en los métodos iterativos.

Si $x_{n+1} = g(x_n)$, hemos visto que

$$\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{x_n - \alpha} = g'(\alpha)$$

Como $\alpha$ es desconocido, queremos dar una estimación de $g'(\alpha)$. En efecto,

$$\frac{x_{n+2}-x_{n+1}}{x_{n+1}-x_n}=\frac{(x_{n+2}-\alpha)-(x_{n+1}-\alpha)}{(x_{n+1}-\alpha)-(x_n-\alpha)}=\frac{\frac{(x_{n+2}-\alpha)}{(x_{n+1}-\alpha)}-1}{1-\frac{x_n-\alpha}{x_{n+1}-\alpha}}\approx\frac{g'(\alpha)-1}{1-\frac{1}{g'(\alpha)}}$$

para $n$ suficientemente grande. Luego

$$g'(\alpha)\approx\frac{x_{n+2}-x_{n+1}}{x_{n+1}-x_n}$$

Luego la tasa de convergencia es $\lambda=|g'(\alpha)|$

## 2.10 Aceleración de la Convergencia:

Estudiaremos el llamado **Método de Aitken o extrapolación de Aitken o $\triangle^2$-Aitken.**

Hemos visto que

$$\lim_{n\to\infty}\frac{\varepsilon_{n+1}}{\varepsilon_n}=g'(\alpha)$$

Mostraremos cómo obtener una mejora significativa de la convergencia del método de iteraciones mediante un uso adecuado de esta información. Siendo

$$\frac{x_{n+1}-\alpha}{x_n-\alpha}\approx g'(\alpha)\quad\text{para } n \text{ suficientemente grande}$$

resulta

$$\frac{x_{n+1}-\alpha}{x_n-\alpha}\approx\frac{x_{n+2}-\alpha}{x_{n+1}-\alpha}$$

lo cual implica

$$(x_{n+1}-\alpha)^2\approx(x_n-\alpha)(x_{n+2}-\alpha)$$

Despejando $\alpha$:

$$\alpha\approx x_n'=\frac{x_nx_{n+2}-x_{n+1}^2}{x_{n+2}-2x_{n+1}+x_n}=x_n-\frac{(x_{n+1}-x_n)^2}{(x_{n+2}-x_{n+1})-(x_{n+1}-x_n)}\quad(2.19)$$

conocido como el método de Aitken. También se llama la fórmula de extrapolación de Aitken pues $x_n'$ se encuentra realmente extrapolando los errores a cero. Esto es, si ponemos

$$\delta_n=g(x_n)-x_n=x_{n+1}-x_n=\varepsilon_{n+1}-\varepsilon_n$$

de igual forma se tiene

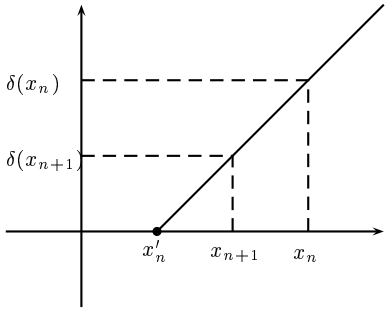
$$\delta_{n+1}=\varepsilon_{n+2}-\varepsilon_{n+1}$$

Considerando la recta que pasa por $(x_n, \delta_n), (x_{n+1}, \delta_{n+1})$ se ve que $x'_n$ es la intersección de dicha recta con el eje $x$. Es decir $x'_n$ es un valor extrapolado de la recta. También se lo conoce como el método $\triangle^2$ de Aitken pues introduciendo la notación de diferencias divididas

$$\begin{aligned} \triangle x_n &= x_{n+1} - x_n \\ \triangle^2 x_n &= \triangle(\triangle x_n) = \triangle(x_{n+1} - x_n) = x_{n+2} - 2x_{n+1} - x_n \end{aligned}$$

el método se escribe

$$x'_n = x_n - \frac{(\triangle x_n)^2}{\triangle^2 x_n}$$




**Teorema 2.10.1.** *(Teorema de la aceleración de Aitken). Sea $\{x_n\}$ una sucesión que converge a $\alpha$. Entonces la nueva sucesión*

$$x'_n = x_n - \frac{(x_{n+1} - x_n)}{(x_{n+2} - x_{n+1}) - (x_{n+1} - x_n)} \qquad n \geq 0$$

*converge a $\alpha$ más rápido. En efecto,*

$$\lim_{n\to\infty} \frac{x'_n - \alpha}{x_n - \alpha} = 0$$

**Demostración:**

Pongamos $\varepsilon_n = x_n - \alpha$. Entonces de la primera expresión para $x'_n$ en (2.19)

$$x'_n = \frac{(\alpha + \varepsilon_n)(\alpha + \varepsilon_{n+2}) - (\alpha + \varepsilon_{n+1})^2}{(\alpha + \varepsilon_{n+2}) - 2(\alpha + \varepsilon_{n+1}) + (\alpha + \varepsilon_n)} = \alpha + \frac{\varepsilon_n \varepsilon_{n+2} - \varepsilon_{n+1}^2}{\varepsilon_{n+2} - 2\varepsilon_{n+1} + \varepsilon_n}$$

Siendo

$$\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{x_n - \alpha} = \eta$$

se tiene que

$$x_{n+1} - \alpha = \eta(x_n - \alpha) + \delta_n(x_n - \alpha)$$

con $\lim_{n\to\infty}\delta_n = 0$ y $|\eta| < 1$. Luego $\varepsilon_{n+1} = (\eta+\delta_n)\varepsilon_n$ y $\varepsilon_{n+2} = (\eta+\delta_{n+1})\varepsilon_{n+1} = (\eta+\delta_{n+1})(\eta+\delta_n)\varepsilon_n$. Luego

$$x'_n - \alpha = \frac{\varepsilon_n^2(\eta+\delta_{n+1})(\eta+\delta_n) - (\eta+\delta_n)^2\varepsilon_n^2}{(\eta+\delta_{n+1})(\eta+\delta_n)\varepsilon_n - 2(\eta+\delta_n)\varepsilon_n + \varepsilon_n} =$$
$$= \varepsilon_n \frac{(\eta+\delta_{n+1})(\eta+\delta_n) - (\eta+\delta_n)^2}{(\eta+\delta_{n+1})(\eta+\delta_{n+1}) - 2(\eta+\delta_n) + 1}$$

Luego

$$\lim_{n\to\infty}\frac{x_{n+1}-\alpha}{x_n-\alpha} = \lim_{n\to\infty}\frac{(\eta+\delta_{n+1})(\eta+\delta_n) - (\eta+\delta_n)^2}{(\eta+\delta_{n+1})(\eta+\delta_{n+1}) - 2(\eta+\delta_n)+1} = 0$$

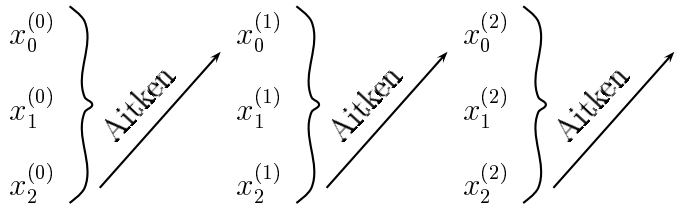
□

## 2.10 El Algoritmo Modificado de Aitken:

A partir de $x_0$ dado calculamos $x_1 = g(x_0), x_2 = g(x_1)$ y aplicamos la fórmula de Aitken

$$x'_0 = x_0 - \frac{(x_1-x_0)^2}{x_2 - 2x_1 + x_0}$$

$x'_0$ se usa como un nuevo valor inicial para dos iteraciones más: $x_0^{(1)} = x'_0$, $x_1^{(1)} = g(x_0^{(1)})$, $x_2^{(1)} = g(x_1^{(1)})$. Nuevamente aplicamos la fórmula de Aitken a $x_0^{(1)}, x_1^{(1)}$ y $x_2^{(1)}$ para obtener un valor acelerado $x_0^{(2)}$ que a su vez se usa para comenzar una nueva iteración. Si $x_2^{(k)} - 2x_1^{(k)} + x_0^{(k)} = 0$, hacemos $x_0^{(k+1)} = x_0^{(k)}$. Esquemáticamente




Se puede demostrar que este método es cuadráticamente convergente bajo la hipótesis que $x = g(x)$ tiene un punto fijo $x = \alpha$, $g \in C^{(3)}$ en un entorno de $\alpha$ y $g'(\alpha) \neq 1$. De hecho

$$x_{n+1} = x_n - \frac{(g(x_n)-x_n)^2}{g(g(x_n)) - 2g(x_n) + x_n}$$

Poniendo $f(x) = g(x) - x$, si $\alpha$ es punto fijo de $g$ entonces $f(\alpha) = 0$ y

$$x_{n+1} = x_n - \frac{f(x_n)}{\frac{f(x_n + f(x_n)) - f(x_n)}{f(x_n)}}$$

que es el método de Steffensen estudiado anteriormente.

**Algoritmo 6:** Algoritmo de Steffensen o Aitken Modificado $(g, x_0, \varepsilon$,raíz,itmax,flag)

1. iter:=1
2. Mientras $i \leq itmax$ hacer pasos 3-6
3. Tomar $x_1 = g(x_0),\ x_2 = g(x_1)$ y calcular raíz z $= x_0 - \dfrac{(x_1 - x_0)^2}{(x_2 - x_1) + (x_0 - x_1)}$
4. Si $|$raíz z $- x_0| \leq \varepsilon$ entonces flag=0, stop.
5. iter:=iter+1
6. hacer $x_0$ :=raíz
7. flag:=1 (itmax es alcanzado), stop.

**Nota:** Si usáramos la estimación del error

$$|x_1 - x_2| \leq (1 - \lambda)\varepsilon/\lambda \tag{2.20}$$

con $\lambda$ tal que $\lambda \approx \left|\dfrac{x_2 - x_1}{x_1 - x_0}\right|$, obtendríamos un mejor criterio de parada, posiblemente deteniéndonos unas iteraciones anteriores. Este último criterio se justificará en la siguiente sección.

## 2.11 Criterios de Parada

Para el problema $f(x) = 0$. La pregunta natural que surge es ¿Cuándo parar un proceso iterativo? La respuesta depende de

1) La precisión deseada por el usuario, señalado con una cierta tolerancia $\varepsilon$.

2) La mayor precisión que se puede esperar de la raíz, basada en la precisión con que se calcula $f(x)$.

**Criterio 1:** $|f(x_n)| \leq \varepsilon$

Puede no ser un buen criterio. Si es cierto, entonces siendo

$$f(x_n) = f(x_n) - f(\alpha) = f'(\xi_n)(x_n - \alpha) \qquad \xi_n \in int(x_n, \alpha)$$

resulta

$$x_n - \alpha = \frac{f(x_n)}{f'(\xi_n)}$$

y por lo tanto

$$\begin{array}{lcll} |x_n - \alpha| & \approx & |f(x_n)| \leq \varepsilon & \text{si} \quad |f'(\alpha)| \approx 1 \\ |x_n - \alpha| & > & \varepsilon & \text{si} \quad |f'(\alpha)| \ll 1 \\ |x_n - \alpha| & \ll & \varepsilon & \text{si} \quad |f'(\alpha)| \gg 1 \end{array}$$

resultando, en este último caso, en cálculos innecesarios.

**Criterio 2:** $|x_{n+1} - x_n| \leq \varepsilon$

En el caso particular del método de Newton funciona:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad \Rightarrow \quad x_{n+1} - x_n = -\frac{f(x_n)}{f'(x_n)} \approx -\frac{f(x_n)}{f'(\xi_n)} = -(x_n - \alpha)$$

Resulta que:

$$\text{si} \quad |x_{n+1} - x_n| \leq \varepsilon \qquad \Rightarrow \qquad |x_n - \alpha| \leq \varepsilon$$

Para métodos linealmente convergentes podría ser impreciso. Sea $x_{n+1} = g(x_n)$ y supongamos $g'(\alpha) \approx 1$. Para $x_n \approx \alpha$,

$$x_{n+1} - \alpha = g(x_n) - g(\alpha) = g'(\xi_n)(x_n - \alpha) \approx g'(\alpha)(x_n - \alpha)$$

y

$$\begin{array}{rcl} x_{n+1} - x_n & = & (x_{n+1} - \alpha) - (x_n - \alpha) \approx \\ & \approx & g'(\alpha)(x_n - \alpha) - (x_n - \alpha) = (x_n - \alpha)(g'(\alpha) - 1) \end{array}$$

de donde

$$x_n - \alpha \approx \frac{x_{n+1} - x_n}{g'(\alpha) - 1}$$

Luego, si $g'(\alpha) \approx 1$ entonces $|x_n - \alpha| \gg |x_{n+1} - x_n|$. Notar que si $-1 < g'(\alpha) < 0$, el criterio es bueno.

**Criterio 3:** Sea $\lambda = |g'(\alpha)|$ (calculado por ejemplo mediante $\lambda = \left|\dfrac{x_{n+2} - x_{n+1}}{x_{n+1} - x_n}\right|$), Entonces, siendo

$$x_{n+1} - \alpha \approx g'(\alpha)(x_n - \alpha) \approx \frac{g'(\alpha)}{g'(\alpha) - 1}(x_{n+1} - x_n)$$

resulta

$$|x_{n+1} - \alpha| \leq \frac{\lambda}{1 - \lambda}|x_{n+1} - x_n|.$$

Por lo tanto si

$$|x_{n+1} - x_n| \leq \frac{1 - \lambda}{\lambda}\varepsilon \quad \Rightarrow \quad |x_{n+1} - \alpha| \leq \varepsilon.$$

que es el criterio dado en (2.20).

## 2.12 Raíces de Polinomios

Se busca resolver la ecuación

$$p(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n = 0 \qquad a_0 \neq 0$$

Por el teorema fundamental del álgebra esta ecuación posee exactamente nraíces $\alpha_1,\ \alpha_2, \ldots \alpha_n$ y $p(x) = a_0(x - \alpha_1)(x - \alpha_2)\ldots(x - \alpha_n)$

Si el problema consiste en buscar una raíz particular para la cual se conoce una aproximación inicial, lo mejor es modificar algunos de los métodos iterativos estudiados para explotar las ventajas que ofrecen la forma especial de los polinomios. Así por ejemplo, el método de Newton o el de la secante pueden ser utilizados, ya sea para hallar raíces reales como complejas, usando para ello aritmética compleja.

### 2.12.1 Regla de Horner. Deflación

Para evaluar $p(x)$ en $x = z$ utilizamos el esquema de Horner

$$\begin{cases} b_0 &= a_0 \\ b_i &= b_{i-1}z + a_i \quad i = 1, 2, \ldots n \end{cases}$$

y entonces $b_n = p(z)$. Consideremos el polinomio

$$q(x) = b_0 x^{n-1} + b_1 x^{n-2} + \cdots + b_{n-1}$$

entonces

$$b_n + (x - z)q(x) = b_n + (x - z)(b_0 x^{n-1} + b_1 x^{n-2} + \cdots + b_{n-1}) = p(x)$$

Luego $q(x)$ es el cociente y $b_n$ es el resto de la división de $p(x)$ entre $x - z$. En particular si $z$ es raíz de $p(x)$ entonces $p(z) = 0$, $b_n = 0$ y $p(x) = (x - z)q(x)$.

Para hallar otras raíces de $p(x)$ las buscamos en $q(x)$. Este proceso se llama **deflación**. Por deflaciones sucesivas se pueden obtener los ceros de un polinomio manipulando polinomios de grado inferior. Esto ahorra operaciones aritméticas y además las iteraciones no pueden converger a una misma raíz simple más de una vez. Si se tiene muy poca información sobre las raíces, se puede proceder de la siguiente manera: se elige más o menos aleatoriamente una aproximación inicial; la probablilidad de obtener convergencia a alguna raíz es grande. Se usa esta raíz para factorizar el polinomio y se continúa de la misma forma hasta que todas las raíces hayan sido calculadas. Así, para calcular $q(z)$ se hace

$$\begin{cases} c_0 &= b_0 \\ c_i &= c_{i-1} z + b_i \quad i = 1, 2, \ldots n \end{cases}$$

y entonces $c_{n-1} = q(z)$. Si $c_{n-1} = 0$, $z$ es raíz de $q(x)$ y se escribe:

$$q(x) = (x - z)h(x)$$

con

$$h(x) = c_0 x^{n-2} + c_1 x^{n-3} + \cdots + c_{n-2}$$

y el proceso continúa.

### 2.12.2 Método de Newton-Raphson aplicado a polinomios.

**Algoritmo:** PoliNewton($a, n, x_0, itmax$,raíz,$b, ier, \varepsilon$)

**Nota:** $a$: vector de coeficientes del polinomio $p$.
itmax: máximo número de iteraciones.
$b$: vector de coeficientes del polinomio obtenido por deflación.
ier: indicador del error.

1. $itnum = 1$
2. $b_0 = c = a_0$
3. Para $i = 1, \ldots, n-1$
   $b_i = x_0 b_{i-1} + a_i$
   $c = x_0 c + b_i$
4. $b_n = a_n + x_0 b_{n-1}$
5. $x_1 = x_0 - (b_n/c)$
6. Si $|x_1 - x_0| \leq \varepsilon$ entonces hacer $ier = 0$, raíz= $x_1$; stop.
7. Si $itnum = itmax$, hacer $ier = 1$, raíz= $x_1$; stop.
8. Sino hacer $itnum = itnum + 1$, $x_0 = x_1$; ir a 3.

### 2.12.3 Estrategia de Wilkinson

Todo lo dicho en el punto (2.12.1) es cierto si $z$ es tal que $p(z) = 0$; pero si $z$ no es **raíz exacta**, o hay errores de redondeo en el cálculo de $b_0, b_1, \ldots, b_n$, puede ocurrir que estos errores se propagen de tal forma que el efecto sea que los ceros calculados se desvíen más y más de los ceros de $p(z)$:

La estrategia propuesta por Wilkinson en *"Rounding errors in algebraic process"* (1963) es la siguiente:

i) Las raíces se determinan en orden creciente de magnitud.

ii) Toda raíz obtenida usando Newton sobre un polinomio reducido $q$ (deflación) debería ser inmediatamente refinada aplicando Newton nuevamente al polinomio original $p$, usando la raíz calculada como aproximación inicial. Sólo después que esto se haya realizado seguir con el siguiente paso en el proceso de deflación. Con esto los errores que resultan en la deflación son despreciables.

### 2.12.4 Ecuaciones Algebraicas mal condicionadas

Hemos visto que las raíces múltiples son *mal condicionadas*, esto es, son sensibles a pequeñas perturbaciones tales como los errores de redondeo. Si $p(x)$ tiene raíces que están muy cerca unas de otras, estas raíces serán sensibles a

perturbaciones en los coeficientes de $p(x)$. Lo sorprendente es que raíces mal condicionadas aunque aparezcan más separadas.

El siguiente ejemplo, conocido como el pérfido polinomio, proporcionado por Wilkinson (1963) nos muestra esta situación:

$$P(x) = (x-1)(x-2)\cdots(x-20) = \prod_{j=1}^{20}(x-j) = x^{20} - 210x^{19} + \cdots$$

los ceros son $x = 1, 2, \ldots, 20$ y están bien separados. Expandiendo el producto tenemos que

$$\begin{aligned} P(x) = \quad & x^{20} - 210x^{19} + 2061x^{18} - 1256850x^{17} + 53327946x^{16} \\ & -1672280820x^{15} + 40171771630x^{14} - 756111184500x^{13} \\ & +11310276995381x^{12} - 135585182899530x^{11} \\ & +1307535010540395x^{10} - 10142299865511450x^{9} \\ & +63030812099294896x^{8} - 311333643161390640x^{7} \\ & +1206647803780373360x^{6} - 3599979517947607200x^{5} \\ & +8037811822645051776x^{4} - 12870931245150988800x^{3} \\ & +13803759753640704000x^{2} - 8752948036761600000x \\ & +2432902008176640000 \end{aligned}$$

Las raíces obtenidas con Matlab son:

| | |
|---|---|
| 19.99963598445644 | 9.99444861924099 |
| 19.00345824088232 | 9.00129414216583 |
| 17.98431945441163 | 7.99978647800271 |
| 17.04035699826114 | 7.00002489935984 |
| 15.91981890268148 | 5.99999789179362 |
| 15.10210083979336 | 5.00000014903298 |
| 13.89957813405348 | 3.99999999068689 |
| 13.07791101181363 | 3.00000000034467 |
| 11.95920497957842 | 1.99999999999590 |
| 11.01806328344453 | 1.00000000000008 |

Wilkinson calculó, usando una aritmética de punto flotante con base 2 y precisión 90, las raíces de la ecuación

$$p(x) + 2^{-23}x^{19} = 0 \tag{2.21}$$

es decir con el coeficiente $-210$ cambiado por $-210 + 2^{-23}$. En este caso aparecen 5 pares de raíces complejas. A continuación mostramos las raíces

de la ecuación (2.21) calculadas con Matlab:

| | |
|---|---|
| 20.84688448609406 | |
| $19.50239411919753 + 1.94031477683351i$ | 8.91842091728162 |
| $19.50239411919753 - 1.94031477683351i$ | 8.00706527446574 |
| $16.73061875292583 + 2.81257505050394i$ | 6.00000876711597 |
| $16.73061875292583 - 2.81257505050394i$ | 6.99970419052263 |
| $13.99206818842974 + 2.51860596955270i$ | 4.99999961141216 |
| $13.99206818842974 - 2.51860596955270i$ | 4.00000002122367 |
| $11.79321723281362 + 1.65143203229712i$ | 2.99999999954785 |
| $11.79321723281362 - 1.65143203229712i$ | 1.99999999998745 |
| $10.09566013241218 + 0.64124056256306i$ | 1.00000000000037 |
| $10.09566013241218 - 0.64124056256306i$ | |

Así, un pequeño cambio en el coeficiente de $x^{19}$ produjo 10 ceros complejos. La razón no es un problema de redondeo, ni del algoritmo elegido, sino que es una cuestión de sensibilidad del problema a pequeñas perturbaciones.

### 2.12.5 El Método de Muller

Este método generaliza el método de la secante y si bien se aplica a funciones en general, vamos a desarrollarlo para polinomios; es este último caso en donde su utilidad es mayor y de allí su popularidad.

La idea del método es la siguiente: sean $x_{n-2}, x_{n-1}, x_n$, tres aproximaciones distintas a la raíz $\alpha$; la nueva aproximación $x_{n+1}$ se obtiene como la raíz del polinomio cuadrático que pasa por los puntos $(x_{n-2}, p(x_{n-2}))$, $(x_{n-1}, p(x_{n-1})), (x_n, p(x_n))$. Llamemos $q(x)$ a dicho polinomio cuadrático; usando la *fórmula de interpolación de Newton* (que estudiaremos en el próximo capítulo), $q(x)$ puede expresarse así:

$$q(x) = p(x_n) + (x - x_n)p[x_n, x_{n-1}] + (x - x_n)(x - x_{n-1})p[x_n, x_{n-1}, x_{n-2}] \tag{2.22}$$

con

$$p[x_n, x_{n-1}] = \frac{p(x_{n-1}) - p(x_n)}{x_{n-1} - x_n} \qquad p[x_n, x_{n-1}, x_{n-2}] = \frac{p[x_{n-1}, x_{n-2}] - p[x_{n-1}, x_n]}{x_{n-2} - x_n}$$

**Ejercicio 1:** Mostrar que $q(x_i) = p(x_i), \quad i = n, n-1, n-2$

**Ejercicio 2:** Hallar la parábola que pasa por los puntos $(1, 3), (2, 7), (5, -2)$

Para hallar los ceros de $q(x)$ dado en (2.22) lo reescribimos así

$$q(x) = p(x_n) + w(x - x_n) + p[x_n, x_{n-1}, x_{n-2}](x - x_n)^2 \tag{2.23}$$

con

$$\begin{aligned} w &= p[x_n, x_{n-1}] + (x_n - x_{n-1})p[x_n, x_{n-1}, x_{n-2}] = \\ &= p[x_n, x_{n-1}] + (x_n - x_{n-1})\frac{p[x_n, x_{n-2}] - p[x_{n-2}, x_{n-1}]}{(x_n - x_{n-1})} = \\ &= p[x_n, x_{n-1}] + p[x_n, x_{n-2}] - p[x_{n-2}, x_{n-1}] \end{aligned} \tag{2.24}$$

Usando (2.24), se desea hallar una raíz de $q(x)$, mejor dicho el valor más pequeño de $(x - x_n)$ que satisface $q(x) = 0$.

$$\begin{aligned} x - x_n &= \frac{-w \pm \sqrt{w^2 - 4p[x_n, x_{n-1}, x_{n-2}]p(x_n)}}{2p[x_n, x_{n-1}, x_{n-2}]} = \\ &= -\frac{\left(w - sign(w)\sqrt{w^2 - 4p[x_n, x_{n-1}, x_{n-2}]p(x_n)}\right)}{2p[x_n, x_{n-1}, x_{n-2}]} \end{aligned} \tag{2.25}$$

con el signo escogido de manera de minimizar la magnitud del numerador. Racionalizando (2.25) para evitar pérdida de cifras significativas (fenómeno de cancelación) resulta

$$x_{n+1} = x_n - \frac{2p(x_n)}{w + sign(w)\sqrt{w^2 - 4p[x_n, x_{n-1}, x_{n-2}]p(x_n)}} \tag{2.26}$$

con el signo escogido de manera de maximizar la magnitud del denominador. Puede suceder que la raíz cuadrada en (2.26) se haga imaginaria. Esta posibilidad se considera una ventaja del método ya que nos conduce a ceros complejos.

La convergencia no es tan rápida como la del método de Newton pero es superior a la de la secante. Se puede probar que el orden de convergencia para una raíz simple es

$$p \approx 1.84$$

# Capítulo 3

# Algebra Lineal Numérica

## 3.0 Métodos Directos

**Problema:** Resolver el sistema de ecuaciones

$$Ax = b \tag{3.1}$$

con $A$ una matriz $n \times n$ y $x, b$ vectores $n \times 1$.

Estudiaremos una primera clase de métodos para la resolución de sistemas lineales: los métodos directos. Ellos se aplican a sistemas lineales densos, esto es, en donde los coeficientes de la matriz son casi todos no nulos. Sin embargo sólo pueden aplicarse si el orden de la matriz no es muy grande, debido a que, como normalmente la matriz se almacena en la memoria principal del computador, las limitaciones de la memoria imponen limitaciones en el orden de la matriz.

Por método directo entendemos un método que permite obtener la solución del sistema efectuando un número finito de operaciones elementales sobre los números reales. El estudio de un método directo lleva a examinar cuidadosamente los dos puntos esenciales siguientes:

- costo del método, expresado en el número de operaciones elementales.
- propagación de los errores de redondeo y estabilidad del método.

Estos dos criterios sirven para clasificar los diferentes métodos y por lo tanto recomendar algunos.

Cuando A es no singular, parece natural buscar fórmulas explícitas para resolver numéricamente (3.1); en particular están las bien conocidas fórmulas

o Regla de Cramer:

$$x_i = \frac{det\ A_i}{det\ A}, \qquad i = 1 : n$$

siendo $A_i$ la matriz $A$ con la i-ésima columna reemplazada por el vector $b$. Pero este método es imposible utilizarlo, al menos en una computadora secuencial, no sólo para sistemas grandes, sino aún para los pequeños. Observar que para calcular la solución hay que evaluar $n+1$ determinantes y efectuar $n$ divisiones. Además el cálculo de un determinante exige a priori $(n-1)n!$ multiplicaciones y $n! - 1$ sumas. Se tiene así un total de

$$(n^2 + n)n! - 1$$

operaciones elementales. Y por ejemplo si $n = 10$ tendríamos que realizar $4\ 10^8$ operaciones, lo cual es enorme aún para una computadora electrónica. Veremos que organizando el cálculo en forma adecuada, se puede reducir notablemente el número de operaciones a aproximadamente $2n^3/3$ operaciones, que en el caso de nuestro ejemplo, representaría aproximadamente un total de 666 operaciones elementales.

## 3.1 Sistemas Triangulares

Los métodos directos que vamos a estudiar se basan en la siguiente observación. Un sistema triangular superior

$$Ux = b$$

tiene la forma siguiente:

$$\begin{array}{rcl} u_{11}x_1 + u_{12}x_2 + u_{13}x_3 + \cdots + u_{1n}x_n & = & b_1 \\ u_{22}x_2 + u_{23}x_3 + \cdots + u_{2n}x_n & = & b_2 \\ \cdots\cdots\cdots\cdots\cdots & \cdots & \cdots \\ u_{ii}x_i + \cdots + u_{in}x_n & = & b_i \\ \cdots\cdots\cdots\cdots\cdots & \cdots & \cdots \\ u_{nn}x_n & = & b_n \end{array}$$

Si $u_{11}u_{22}\ldots u_{nn} \neq 0$, es decir si $\det(U) \neq 0$, el sistema se resuelve mediante sustitución regresiva:

$$\begin{array}{rcl} x_n & = & b_n/u_{nn} \\ x_i & = & (b_i - \sum_{j=i+1}^{n} u_{ij}x_j)/u_{ii} \end{array}$$

para $i = n-1, \ldots, 1$

De la misma forma se tiene que si $L$ es una matriz triangular inferior, el sistema

$$Lx = b$$

$$\begin{array}{lcl} l_{11}x_1 & = & b_1 \\ l_{21}x_1 + l_{22}x_2 & = & b_2 \\ \cdots\cdots\cdots\cdots & \cdots & \cdots \\ l_{i1}x_1 + l_{i2}x_2 + \cdots + l_{ii}x_i & = & b_i \\ \cdots\cdots\cdots\cdots & \cdots & \cdots \\ l_{n1}x_1 + l_{n2}x_2 + l_{n3}x_3 + \cdots + l_{nn}x_n & = & b_n \end{array}$$

se resuelve por sustitución progresiva siempre que $\det(L) \neq 0$,

$$\begin{aligned} x_1 &= b_1/l_{11} \\ x_i &= (b_i - \textstyle\sum_{j=1}^{i-1} l_{ij}x_j)/l_{ii} \end{aligned}$$

para $i = 2, \ldots, n$

Contemos el número de operaciones necesarias para resolver estos sistemas triangulares:

1. $n$ divisiones
2. $1 + 2 + 3 + \cdots + (n-1) = n(n-1)/2$ adiciones e igual número de multiplicaciones

Por lo tanto el total es igual a $n^2$ operaciones.

**Ejercicio 3.1.1.** *La suma, el producto por un escalar y el producto de matrices triangulares es una matriz triangular del mismo tipo. Si existe, la inversa de una matriz triangular es una matriz triangular del mismo tipo.*

La simplicidad de los sistemas triangulares sugiere la idea de triangularizar la matriz $A$ del sistema lineal general

$$Ax = b$$

es decir, buscar una matriz $M$ inversible y fácilmente calculable de manera que $MA$ sea triangular. Entonces será suficiente resolver el sistema

$$MAx = Mb$$

equivalente al sistema original, en el sentido que tienen las mismas soluciones. Por lo general $MA$ será triangular superior.

## 3.2 El Método de Eliminación Gaussiana

El primer método directo que vamos a estudiar es el método clásico de eliminación de Gauss. Supongamos que $A = (a_{ij})$ es no singular. Entonces,

$$Ax = b$$

tiene una única solución. Esta solución la denotaremos

$$x = A^{-1}b$$

El sistema puede escribirse como

$$\begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n & = & b_2 \\ \cdots\cdots\cdots\cdots\cdots & \cdots & \cdots \\ a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \cdots + a_{in}x_n & = & b_i \\ \cdots\cdots\cdots\cdots\cdots & \cdots & \cdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n & = & b_n \end{array}$$

Supongamos que $a_{11} \neq 0$; esto nos permite eliminar $x_1$ de las últimas $n-1$ ecuaciones, restando de la i-ésima ecuación el múltiplo

$$m_{i1} = a_{i1}/a_{11} \qquad i = 2:n$$

de la primera ecuación, es decir, hacemos

$$\text{fila } i - m_{i1}\text{fila } 1 = \text{fila } i - (a_{i1}/a_{11})\text{fila } 1 =$$

$$(a_{i1}, a_{i2}, \ldots, a_{in}) - (a_{i1}/a_{11})(a_{11}, a_{12}, \ldots, a_{1n}) =$$

$$(0,\ a_{i2} - (a_{i1}/a_{11})a_{12},\ a_{i3} - (a_{i1}/a_{11})a_{13},\ \ldots,\ a_{in} - (a_{i1}/a_{11})a_{1n}) =$$

$$(0,\ a_{i2} - m_{i1}a_{12},\ a_{i3} - m_{i1}a_{13},\ \ldots,\ a_{in} - m_{i1}a_{1n})$$

para $i = 2:n$.

La últimas $n-1$ ecuaciones del sistema original se transforman en

$$\begin{array}{rcl}
a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \cdots + a_{2n}^{(2)}x_n & = & b_2^{(2)} \\
\cdots\cdots\cdots\cdots\cdots & \cdots & \cdots \\
a_{i2}^{(2)}x_2 + a_{i3}^{(2)}x_3 + \cdots + a_{in}^{(2)}x_n & = & b_i^{(2)} \\
\cdots\cdots\cdots\cdots\cdots & \cdots & \cdots \\
a_{n2}^{(2)}x_2 + a_{n3}^{(2)}x_3 + \cdots + a_{nn}^{(2)}x_n & = & b_n^{(2)}
\end{array}$$

en donde los nuevos coeficientes están dados por

$$\begin{array}{rcll}
a_{ij}^{(2)} & = & a_{ij} - m_{i1}a_{1j} & i,j = 2:n \\
b_i^{(2)} & = & b_i - m_{i1}b_1 & i = 2:n \\
m_{i1} & = & a_{i1}/a_{11} & i = 2:n
\end{array}$$

El sistema es un sistema de $n-1$ ecuaciones con $(n-1)$ incógnitas $x_1, x_2, \ldots, x_n$. Si $a_{22}^{(2)} \neq 0$ de una manera similar podemos proceder con la eliminación de $x_2$ de las últimas $n-2$ ecuaciones, poniendo

$$m_{i2} = a_{i2}^{(2)}/a_{22}^{(2)} \qquad i = 3:n$$

Los coeficientes del nuevo sistema están dados por

$$a_{ij}^{(3)} = a_{ij}^{(2)} - m_{i2}a_{2j}^{(2)} \qquad i,j = 3:n$$

$$b_i^{(3)} = b_i^{(2)} - m_{i2}b_2^{(2)} \qquad i = 3:n$$

Los elementos $a_{11}$, $a_{22}^{(2)}$, $a_{33}^{(3)}$ ... se llaman pivotes. Si son no nulos, podemos continuar la eliminación hasta cubrir $n-1$ pasos, obteniendo en este último paso la ecuación

$$a_{nn}^{(n)}x_n = b_n^{(n)}$$

Juntando las primeras n ecuaciones de estos n sistemas obtenemos

$$\begin{array}{rcl}
a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \cdots + a_{1n}^{(1)}x_n & = & b_1^{(1)} \\
a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \cdots + a_{2n}^{(2)}x_n & = & b_2^{(2)} \\
\cdots\cdots\cdots\cdots\cdots & \cdots & \cdots \\
a_{ii}^{(i)}x_i + \cdots + a_{in}^{(i)}x_n & = & b_i^{(i)} \\
\cdots\cdots\cdots\cdots\cdots & \cdots & \cdots \\
a_{nn}^{(n)}x_n & = & b_n^{(n)}
\end{array}$$

donde $a_{ij}^{(1)} = a_{ij}$, para todo $i,j$, $\quad b_1^{(1)} = b_1$

En el paso $k$, si $a_{kk}^{(k)} \neq 0$ podemos eliminar $x_k$, con

$$m_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)} \qquad i = k+1, \ldots, n$$

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)} \qquad i, j = k+1 : n$$

$$b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)} \qquad i = k+1 : n$$

Como $b$ es transformado de la misma manera que las columnas de A, la descripción se simplifica considerando a $b$ como la columna $n+1$ de A, poniendo:

$$a_{i,n+1}^{(k)} = b_i^{(k)}, \qquad i, k = 1 : n$$

obteniéndose

$$m_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)} \qquad i = k+1, \ldots, n$$

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)} \qquad i = k+1, \ldots, n, \quad j = k+1 : n+1$$

**Observaciones:**

1. Número de operaciones:

   (a) **Modificación de $A$:**
   Número de divisiones en el paso $k$: $n-k$
   Número de productos en el paso $k$: $n-k$ por cada $i$= $(n-k)(n-k)$
   Número de adiciones en el paso $k$: $n-k$ por cada $i$=$(n-k)(n-k)$
   Total:
   $$\sum_{k=1}^{n-1}(n-k) + 2\sum_{k=1}^{n-1}(n-k)^2 = 2n^3/3 + O(n^2)$$

   (b) **Modificación de $b$:**
   Número total de productos: $\sum_{k=1}^{n-1}(n-k)$
   Número total de adiciones: $\sum_{k=1}^{n-1}(n-k)$
   Total:
   $$2\sum_{k=1}^{n-1}(n-k) = n^2 - n = O(n^2)$$

   (c) **Solución de $Ux = \tilde{b}$:** $n^2$

   (d) **Solución de $Ax = b \approx$** $2/3n^2$

2. Resolución simultánea de $p$ sistemas con la misma matriz de coeficientes $Ax = b^{(j)} \qquad j = 1 : p$. En este caso se le agrega a la matriz $A$ las $p$ columnas y se aplica el método anterior. Después hay que resolver $p$ sistemas triangulares con la misma matriz $U$

3. Cálculo de la inversa. Este es un caso particular del anterior: en efecto, eligiendo $b^{(j)} = e^{(j)}$, siendo este último el vector canónico, se aplica la eliminación gaussiana a la matriz

$$[A|I]$$

en donde $I$ es la matriz identidad y luego se resuelven $n$ sistemas de ecuaciones.

4. Cálculo del determinante:

$$det(A) = a_{11}^{(1)} a_{22}^{(2)} \dots a_{nn}^{(n)}$$

## 3.3 Estrategia del pivote

Para llevar a cabo el proceso de eliminación gaussiana hemos supuesto que $a_{kk}^{(k)} \neq 0$. Sin embargo es fácil construir ejemplos en donde esto no ocurre.

**Ejemplo 3.3.1.**

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_1 + x_2 + 2x_3 &= 2 \\ x_1 + 2x_2 + 2x_3 &= 1 \end{aligned}$$

El primer paso en el método de Gauss nos conduce al sistema

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_3 &= 1 \\ x_2 + x_3 &= 0 \end{aligned}$$

lo cual nos da $a_{22}^{(2)} = 0$ y no podríamos continuar el proceso. Consideremos el mismo sistema pero con las filas 1 y 2 intercambiadas. La eliminación en el primer paso nos da:

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + x_3 &= 0 \\ x_3 &= 1 \end{aligned}$$

En el caso general, si $a_{kk}^{(k)} = 0$ y si la matriz es no singular, entonces existirá $i$ con $i = k+1, \dots, n$ tal que $a_{ik}^{(k)} \neq 0$. Supongamos que $a_{rk}^{(k)} \neq 0$, entonces

intercambiamos las filas $r$ y $k$ y seguimos el proceso. Sin embargo éste no es el único caso en donde se necesita el intercambio de filas. Veamos el siguiente ejemplo:

**Ejemplo 3.3.2.**

$$\begin{aligned} 0.0001\ x_1 + x_2 &= 1.0 \\ x_1 + x_2 &= 2 \end{aligned}$$

*con solución* $x_1 = 1.0001,\ x_2 = 0.9999$

Supongamos que trabajamos con una aritmética de punto flotante de precisión t=3; realizando la eliminación gaussiana, se tiene

$$\begin{aligned} 0.0001\ x_1 + x_2 &= 1.0 \\ -10^4\ x_2 &= -10^4 \end{aligned}$$

ya que $1 - 10^4 = (0.00001 - 0.1)10^5 = -0.99999\ 10^5 = -0.1\ 10^5 = -10^4$ y la solución de este sistema resulta:

$$x_2 = 1 \ \text{(buena)} \qquad x_1 = 0 \ \text{(mala)}$$

Pero si intercambiamos filas obtenemos

$$\begin{aligned} x_1 + x_2 &= 2.0 \\ x_2 &= 1 \end{aligned}$$

con $x_1 = x_2 = 1$, lo cual es razonable.

**Estrategia del pivote parcial:** elegir $r$ como el menor entero tal que

$$|a_{rk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|$$

y luego intercambiar filas $k$ y $r$.

**Estrategia del pivote total:** elegir $r$ y $s$ como los menores enteros tales que

$$|a_{rs}^{(k)}| = \max_{k \le i,j \le n} |a_{ij}^{(k)}|$$

y luego intercambiar filas $k$ y $r$ y columnas $k$ y $s$.

Se puede demostrar que el pivote total controla los errores de redondeo. Los resultados teóricos del pivote parcial no son tan buenos como los del total, pero en casi todos los problemas prácticos, el comportamiento del error es

casi el mismo que cuando se hace pivote total. Como este último requiere más tiempo, el pivote parcial es más utilizado.

Observaciones: para describir el algoritmo de la eliminación gaussiana hemos trabajado con la hipótesis que $A$ es no singular, pero usualmente no se conoce a priori si $A$ es o no singular. Sin embargo el mismo método nos da la respuesta. Por ejemplo, apliquemos eliminación gaussiana con pivote parcial.

1. Si el algoritmo se termina exitosamente, es decir si y solo si todos los pivotes son no nulos, entonces,

$$det(A) = (-1)^m a_{11}^{(1)} a_{22}^{(2)} \dots a_{nn}^{(n)}$$

   siendo $m$ el número de intercambios de filas realizados.

2. Si para algún $k$, $a_{kk}^{(k)} = 0$ y además $a_{ik}^{(k)} = 0, \forall i$, entonces $det(A) = 0$ y el algoritmo se termina.

3. Los multiplicadores $m_{ik}$ tienen magnitudes menores o iguales que 1, lo cual controla los errores de redondeo.

Vamos a mostrar el siguiente ejemplo debido a J. Wilkinson, en Error Analysis of Direct Method of Matrix Inversion, J. Assoc. Comp. Mach. 8 (1961), 281-330, en donde se muestra que la estrategia del pivote parcial no es suficiente para asegurar la estabilidad numérica de la eliminación gaussiana.
Para n impar definimos

$$A = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 1 \\ 1 & 1 & 0 & & \vdots & -1 \\ -1 & 1 & 1 & & & 1 \\ 1 & -1 & 1 & & & -1 \\ -1 & 1 & -1 & & \vdots & \vdots \\ \vdots & \vdots & \vdots & & 0 & (-1)^n \\ (-1)^n & (-1)^{n-1} & (-1)^{n-2} & & 1 & 1 \end{pmatrix}$$

y consideramos el sistema

$$Ax = e_1$$

cuya solución es la primera columna de $A^{-1}$. Es fácil ver que

$$x = (1/2, 0, \dots, 0, 1/2)^T$$

Definimos la matriz

$$B = A + 12e_n e_n^T$$

$B$ es igual a $A$ salvo que $B(n,n) = 32$.
La solución de $By = e_1$ es

$$y = (12, 0, \ldots, 0, 12)^T - 11 + 12^n \left(-12^{n+1}, 12^n, -12^{n-1}, \ldots, 12^3, 12^{n+1}\right)^T .$$

Entonces

$$||x - y||_\infty = 12^3 1 + 12^n \approx 12^3 \quad \text{para } n \text{ grande.}$$

Supongamos que se hace eliminación gaussiana con pivote parcial sobre $Ax = e_1$ y $By = e_1$. Es fácil ver que no se requieren intercambios, que los elementos en las matrices reducidas quedan iguales, salvo en la columna en donde se están introduciendo los ceros y en la última. Los elementos en la última columnas van a ser

$$(1, -2, 2^2, -2^3, \ldots, 2^{n-1})^T \quad \text{para } A$$

$$(1, -2, 2^2, -2^3, \ldots, 2^{n-1} + 1/2)^T \quad \text{para } B$$

El lado derecho queda

$$(1, -1, 2, -2^2, \ldots, -2^{n-3}, 2^{n-2})^T$$

Si este cálculo se hace con una aritmética de punto flotante en un sistema binario con $t = 27 = n$, entonces el número $2^{n-1} + 1/2$ tiene una representación

$$\underbrace{10\ldots0}_{n}.1 = 0.100\ldots01 \ \times 10^n$$

Como $t = n$, éste será truncado a $0.1 \times 10^n$ y por lo tanto ambos sistemas, en esta máquina, tienen la misma solución. Se puede probar además, que no se cometen errores de redondeo durante el cálculo. Por lo tanto el error en $y$, solución de $By = e_1$ es

$$||x - y||_\infty ||y||_\infty \simeq 2^{-3} 2^{-1} = 14 = 0.25$$

**Ejercicio 3.3.1.** *verificar que se puede obtener una buena solución de*

$$Bx = e_1$$

*con pivoteo completo.*

## 3.4 Descomposición LU

Hemos visto que la eliminación gaussiana nos permite tratar simultáneamente varios sistemas de ecuaciones lineales con la misma matriz de coeficientes. Cuando los segundos miembros de varios sistemas de ecuaciones no están disponibles de una sola vez, sino que éstos pueden estar en función de soluciones previas, es conveniente hallar una descomposición de la matriz $A$ en una matriz triangular inferior inversible y otra superior

$$A = LU$$

Entonces $Ax = b$ si y solo si

$$Ly = b \qquad Ux = y$$

Conocidos $U$ y $L$ podemos resolver $Ax = b$ con $n^2$ operaciones.

**Teorema 3.4.1.** *Teorema de la Descomposición LU. Sean A no singular y*

$$L = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ m_{21} & 1 & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ m_{i1} & \dots & m_{i,i-1} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ m_{n1} & m_{n2} & \dots & \dots & m_{n,n-1} & 1 \end{pmatrix}$$

$$U = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & \dots & \dots & a_{1n}^{(1)} \\ \dots & a_{22}^{(2)} & \dots & \dots & \dots & a_{2n}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & a_{jj}^{(j)} & \dots & a_{jn}^{(j)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & a_{nn}^{(n)} \end{pmatrix}$$

*con* $m_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)} \quad a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}$, $i, j = k+1 : n, \quad k = 1 : n-1$. *Supongamos además que no hemos realizado pivote parcial ni total, entonces*

$$A = LU$$

**Prueba:** La demostración consiste en verificar que

$$(LU)_{ij} = a_{ij}$$

Sea $U_{ij} = u_{ij} = a_{ij}^{(i)}, \quad j \geq i$. Entonces $(LU)_{ij}$ = fila $i$ de $L\times$ columna $j$ de $U = (m_{i1}, m_{i2} \ldots, m_{i,i-1}, 1, \ldots, 0) \times (u_{1j}, u_{2j}, \ldots, u_{jj}, \ldots, 0)^T$. Luego, si $i \leq j$

$$\begin{aligned}(LU)_{ij} &= m_{i1}u_{1j} + m_{i2}u_{2j} + \cdots + m_{i,i-1}u_{i-1,j} + u_{ij} \\ &= \textstyle\sum_{k=1}^{i-1} m_{ik}u_{kj} + u_{ij} \\ &= \textstyle\sum_{k=1}^{i-1} m_{ik}a_{kj}^{(k)} + a_{ij}^{(i)} \\ &= \textstyle\sum_{k=1}^{i-1}(a_{ij}^{(k)} - a_{ij}^{(k+1)}) + a_{ij}^{(i)} = a_{ij}^{(1)} = a_{ij}\end{aligned}$$

Si $i > j$ la demostración es análoga.

**Teorema 3.4.2.** *La descomposición LU de una matriz A no singular es única.*

**Prueba:** Razonemos por el absurdo. Sean dos descomposiciones $LU$ tales que $A = L_1U_1 = L_2U_2$. Las matrices $L_i,\ U_i,\ i = 1, 2$ son no singulares y por lo tanto $L_2^{-1}L_1 = U_2U_1^{-1}$. Como la inversa de una matriz triangular inferior (superior) es triangular inferior (superior) y el producto de dos triangulares inferiores (superiores) es triangular inferior (superior), se tiene que el primer miembro de la última igualdad es una matriz triangular inferior y el segundo miembro es una matriz triangular superior. Además $L_1$ y $L_2$ tienen diagonales unitarias, lo que implica que

$$L_2^{-1}L_1 = U_2U_1^{-1} = I$$

de donde $L_1 = L_2$ y $U_1 = U_2$.

**Corolario 3.4.1.** $det(A) = del(LU) = det(L)det(U) = a_{11}^{(1)}a_{22}^{(2)} \ldots a_{nn}^{(n)}$

Notar que la estrategia del pivote total o parcial modifica el resultado del teorema de descomposición $LU$. Realizando intercambios de filas obtendríamos

$$LU = A'$$

siendo $A'$ la matriz que resulta de aplicar a $A$ los intercambios de filas realizados durante la eliminación y en el mismo orden. Formalmente, si $P$ es una matriz de permutación, esto es una matriz que resulta de permutar las filas de la matriz identidad, entonces los intercambios pueden representarse mediante $PA$ y la descomposición se escribe

$$LU = PA$$

Resumiendo, la implementación de la eliminación gaussiana con pivote parcial contempla los siguientes pasos.

1. Ir almacenando las filas de los pivotes en un vector $(p_1, p_2 \ldots, p_{n-1})$
2. Ir almacenando los multiplicadores $m_{ik}$ en la parte estrictamente triangular inferior de la matriz $A$.
3. Intercambiar filas $k$ y $p_k$, $k = 1 : n-1$ del vector $b$, obteniendo un vector $b'$.
4. Resolver $Ly = b'$ y $Ux = y$

Para el cálculo del determinante se tiene

1. Aplicar eliminación gaussiana con pivote parcial
2. Calcular $det(A) = (-1)^m a_{11}^{(1)} a_{22}^{(2)} \ldots a_{nn}^{(n)}$ donde $m$ es el número total de intercambios realizados.

## 3.5 Descomposición de Cholesky

Los métodos discutidos anteriormente para resolver sistemas de ecuaciones lineales pueden fallar si no se aplica pivote parcial o total. Sin embargo existe una clase importante de matrices para las cuales no es necesario realizar pivoteo para obtener su factorización. Estas matrices son las llamadas matrices simétricas definidas positivas.

**Definición 3.5.1.** *$A$ es simétrica si $A^T = A$, siendo $A^T$ la traspuesta de $A$.*

**Definición 3.5.2.** *$A$ es definida positiva si*

$$x^T Ax > 0$$

*para todo $x \neq 0$*

**Propiedades:**

1. Los elementos diagonales de una matriz definida positiva son positivos.
2. Si $A$ es definida positiva $A^T$ también lo es.

3. La inversa de una matriz definida positiva siempre existe.

4. La inversa de una matriz definida positiva es definida positiva.

5. Todas las submatrices principales de una matriz definida positiva son definidas positivas.

6. El determinante de una matriz definida positiva es positivo.

7. Todos los menores principales de una matriz definida positiva son positivos.

8. Si existe, la inversa de una matriz simétrica es simétrica.

9. Todas las submatrices principales de una matriz simétrica son simétricas.

10. Una matriz simétrica es definida positiva si y solo si todos los menores principales son positivos (Teorema de Sylvester)

**Teorema 3.5.1.** *(Cholesky) Sea $A$ una matriz simétrica definida positiva de orden $n$. Entonces existe una única matriz triangular inferior $L$ con $l_{ii} > 0$ tal que*

$$A = LL^T$$

**Prueba:** El teorema se prueba por inducción sobre n. Para $n = 1$ el teorema es trivial, pues una matriz definida positiva $1 \times 1$ es un escalar positivo.

$$A = \alpha = l_{11} l_{11}$$

con $l_{11} = +\sqrt{\alpha}$.

Supongamos que el teorema es cierto para matrices simétricas definidas positivas de orden $n-1$ y consideremos una matriz $A$ de orden $n$. Esta matriz puede particionarse de la siguiente manera:

$$A = \begin{pmatrix} A_{n-1} & b \\ b^T & a_{nn} \end{pmatrix}$$

con $A_{n-1}$ de orden $n-1$, $b\ \varepsilon R^{n-1}$ y $A_{n-1}$ simétrica y definida positiva. Por la hipótesis inductiva existe una única matriz $L_{n-1}$ de orden $n-1$ triangular inferior con elementos diagonales positivos que satisface

$$A_{n-1} = L_{n-1} L_{n-1}^T$$

Definamos

$$L = \begin{pmatrix} L_{n-1} & 0 \\ c^T & \alpha \end{pmatrix}$$

y tratemos de hallar $\alpha > 0$ y $c \; \varepsilon \; R^n$ tales que

$$LL^T = \begin{pmatrix} L_{n-1} & 0 \\ c^T & \alpha \end{pmatrix} \begin{pmatrix} L_{n-1}^T & c \\ 0 & \alpha \end{pmatrix} = \begin{pmatrix} A_{n-1} & b \\ b^T & a_{nn} \end{pmatrix} = A$$

Necesariamente se debe cumplir

1. $L_{n-1}L_{n-1}^T = A_{n-1}$
2. $L_{n-1}c = b$
3. $c^T c + \alpha^2 = a_{nn}$ con $\alpha > 0$

La primera igualdad se satisface por la hipótesis inductiva.La segunda es cierta para el único $c = L_{n-1}^{-1}b$ ya que $det(L_{n-1}) > 0$. En cuanto a la tercera se tiene que

$$\alpha^2 = a_{nn} - c^T c$$

Como

$$det(A) = det \begin{pmatrix} L_{n-1} & 0 \\ c^T & \alpha \end{pmatrix} det \begin{pmatrix} L_{n-1}^T & c \\ 0 & \alpha \end{pmatrix} = (det(L_{n-1}))^2 \alpha^2$$

y $det(A) > 0$, resulta $\alpha^2 > 0$. Tomando $\alpha = +\sqrt{a_{nn} - c^T c}$ se tiene que $\alpha > 0$ y $\alpha$ único.

Nota: Observar que no se pide $l_{ii} = 1$. Para la unicidad es necesario que sean positivos (o negativos o de signo alternados, pero se necesita imponer una condición extra).

**Ejemplo 3.5.1.** *Consideremos la matriz de Hilbert de orden 3*

$$A = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$$

*Para hallar su descomposición de Cholesky planteamos*

$$\begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{pmatrix} = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$$

Por identificación de elementos es fácil ver que

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & \sqrt{3}/6 & 0 \\ 1/3 & \sqrt{3}/6 & \sqrt{5}/30 \end{pmatrix}$$

y que el algoritmo puede expresarse así

**Algoritmo de Cholesky**

Para $i = 1 : n$

$l_{ii} := \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}$

Para $j = i + 1 : n$

$l_{ji} := (a_{ij} - \sum_{k=1}^{i-1} l_{ik} l_{jk})/l_{ii}$

El método requiere sólo $n(n+1)/2$ lugares de memoria, en vez de los $n^2$ lugares usuales. El número de operaciones es $n^3/3$, la mitad de las requeridas en la descomposición $LU$. Durante el transcurso del cálculo se deben determinar $n$ raíces cuadradas. El teorema de la descomposición de Cholesky asegura que los radicandos son positivos. Además de la ecuación

$$A = LL^T$$

sacamos que

$$a_{ii} = l_{i1}^2 + l_{i2}^2 + \cdots + l_{ii}^2$$

con $l_{ii} > 0$ lo cual nos dice que

$$|l_{ij}| \leq \sqrt{a_{ii}} \qquad j = 1 : i$$

es decir los elementos de $L$ están acotados por las raíces cuadradas de los elementos diagonales de $A$ y por lo tanto los elementos de $L$ no pueden crecer mucho.

Las raíces cuadradas del método de Cholesky pueden evitarse hallando una matriz inferior $\tilde{L}$ con diagonal unitaria y una matriz diagonal $D$ inversible tal que

$$A = \tilde{L} D^{-1} \tilde{L}^T$$

En este caso el sistema $Ax = b$ se resuelve así:

$$\begin{aligned} \tilde{L} y &= b \\ \tilde{L}^T x &= Dy \end{aligned}$$

La última factorización se puede hacer con aproximadamente el mismo número de operaciones que el método de Cholesky: $n^3/3$ y sin raíces cuadradas.

**Ejemplo 3.5.2.** *Consideremos el ejemplo de la matriz de Hilbert de orden 3. Para evitar las raíces cuadradas procedemos así:*

$$\begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} \delta_1 & 0 & 0 \\ 0 & \delta_2 & 0 \\ 0 & 0 & \delta_3 \end{pmatrix} \begin{pmatrix} 1 & l_{21} & l_{31} \\ 0 & 1 & l_{32} \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$$

*De aquí se tiene*

$$\begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} \delta_1 & \delta_1 l_{21} & \delta_1 l_{31} \\ 0 & \delta_2 & \delta_2 l_{32} \\ 0 & 0 & \delta_3 \end{pmatrix} = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$$

*y por lo tanto*

$$\begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/3 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/12 & 0 \\ 0 & 0 & 1/180 \end{pmatrix} \begin{pmatrix} 1 & 1/2 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$$

## 3.6 Matrices tridiagonales

Una matriz tridiagonal es de la forma

$$A = \begin{pmatrix} a_1 & c_1 & 0 & \dots & \dots & 0 \\ b_2 & a_2 & c_2 & 0 & \dots & 0 \\ 0 & b_3 & a_3 & c_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & b_n & a_n \end{pmatrix}$$

**Teorema 3.6.1.** *Si la descomposición $LU$ existe y se hace sin pivote entonces*

$$A = LU = \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & 0 \\ \beta_2 & 1 & 0 & \dots & \dots & \dots \\ 0 & \beta_3 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \beta_{n-1} & 1 & 0 \\ \dots & \dots & \dots & \dots & \beta_n & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 & c_1 & 0 & \dots & \dots & 0 \\ 0 & \alpha_2 & c_2 & \dots & \dots & \dots \\ 0 & 0 & \alpha_3 & c_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & c_{n-1} \\ \dots & \dots & \dots & \dots & 0 & \alpha_n \end{pmatrix}$$

**Prueba:** Igualando los elementos hallamos que $\alpha_1, \alpha_2, \ldots, \alpha_n$ y $\beta_2, \ldots, \beta_n$ quedan unívocamente determinados por

$$\begin{aligned} \alpha_1 &= a_1 \\ \beta_k &= b_k/\alpha_{k-1} \\ \alpha_k &= a_k - \beta_k c_{k-1} \end{aligned}$$

para $k = 2 : n$

La solución del sistema $Ax = f$ se halla mediante sustitución progresiva y regresiva:

$$Ax = f \Leftrightarrow LUx = f \Leftrightarrow Ly = f \quad \text{y} \quad Ux = y$$

esto es

$$\begin{cases} y_1 = f_1 \\ y_i = f_i - \beta_i y_{i-1} \quad i = 2 : n \end{cases} \qquad \begin{cases} x_n = y_n/\alpha_n \\ x_i = (y_i - c_i x_{i+1})/\alpha_i \quad i = n-1 : 1 \end{cases}$$

Es fácil ver que el número total de operaciones es $O(n)$

**Teorema 3.6.2.** *Si los coeficientes $a_i, b_i, c_i$ satisfacen*

1. $|a_1| > |c_1| > 0$
2. $|a_i| \geq |b_i| + |c_i|$ , $b_i, c_i \neq 0$, $i = 2 : n-1$
3. $|a_n| > |b_n| > 0$

*entonces*

1. *$A$ es no singular*
2. $|\beta_i| \leq 1$, $i = 1 : n-1$
3. $|a_i| - |b_i| < |\alpha_i| < |a_i| + |b_i|$ $i = 2 : n-1$

Observaciones: Las condiciones 1 y 2 de la hipótesis dicen que la matriz es diagonal dominante. La condición 2 de la tesis asegura que los multiplicadores de la factorización $LU$ son en magnitud menores que 1, lo cual asegura la estabilidad y evita el pivoteo. La condición 3 dice que los pivotes están acotados.

**Ejemplo 3.6.1.** *Sea*

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

*Entonces*

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 0 & 2/7 & 1 & 0 \\ 0 & 0 & 7/26 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 7/2 & 1 & 0 \\ 0 & 0 & 26/7 & 1 \\ 0 & 0 & 0 & 45/26 \end{pmatrix}$$

Como las hipótesis del teorema 3.6.2 son satisfechas, los $\beta_i$ efectivamente verifican que sus magnitudes son menores o iguales que 1 y los $\alpha_i$ están acotados por los elementos de su fila correspondiente en la matriz $A$.

## 3.7 Normas vectoriales y matriciales

**Definición 3.7.1.** *Una norma vectorial sobre $\mathbb{R}^n$ (o $\mathbb{C}^n$) es una función $||.||$ a valores reales (o complejos) satisfaciendo*

1. $||x|| \geq 0 \quad \forall\, x \in \mathbb{R}^n \quad y \quad \|x\| = 0 \iff x = 0$
2. $\|\alpha x\| = |\alpha| \|x\| \quad \forall\, x \in \mathbb{R}^n \quad y \quad \forall\, \alpha \in \mathbb{R}$ *(o* $\mathbb{C}$*)*
3. $\|x + y\| \leq \|x\| + \|y\| \quad \forall\, x, y \in \mathbb{R}^n$ *(Desigualdad Triangular)*

**Ejemplo 3.7.1.** *Sea* $x = (x_1, x_2, \ldots, x_n)^T$

1. $\|x\|_2 = \left( \sum_{i=1}^{n} |x_i|^2 \right)^{1/2} = (x^T x)^{1/2}$ *(norma $l_2$ o euclidiana)*
2. $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ *(norma $l_1$)*
3. $\|x\|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p}$, $p \in [1, \infty]$ *(norma $l_p$)*
4. $\|x\|_\infty = \max\limits_{1 \leq i \leq n} |x_i|$ *(norma $l_\infty$ o uniforme)*

**Ejercicio 3.7.1.** *Probar que* $\|x\|_\infty = \lim_{p\to\infty} \|x\|_p$

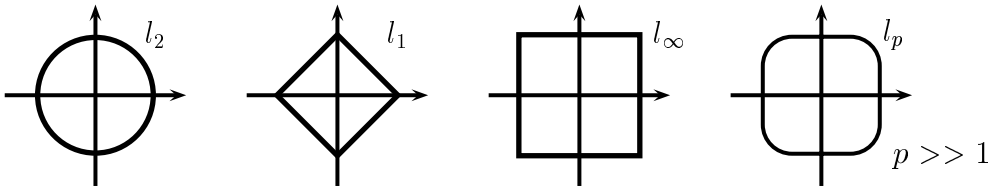
**Definición 3.7.2.** *Una aplicación* $(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *(o de* $\mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$*) es un producto escalar si*

1. $(x,x) \geq 0 \quad \forall\, x \quad y \quad (x,x) = 0 \quad \Longleftrightarrow \quad x = 0$
2. $(x,y) = (y,x) \quad \forall\, x,y \in \mathbb{R}^n \quad (\,(x,y) = \overline{(y,x)}, \quad si \quad x,y \in \mathbb{C}^n)$
3. $(\alpha x, y) = \alpha(x,y) \quad \forall\, x,y \in \mathbb{R}^n \quad y \quad \alpha \in \mathbb{R}$
4. $(x+z,y) = (x,y) + (z,y) \quad \forall\, x,y,z$

**Ejercicio 3.7.2.** *Dado un producto escalar, mostrar que* $\|x\| = (x,x)^{1/2}$ *define una norma. Para ello se requiere de la desigualdad de Cauchy-Schwarz*

$$|(x,y)|^{1/2} \leq \|x\|\|y\|$$

Al conjunto $\{x : \|x\| \leq 1\}$ lo llamamos la bola unitaria y la superficie $\{x : \|x\| = 1\}$ la esfera unitaria.




**Teorema 3.7.1.** *Sea* $\|.\|$ *una norma en* $\mathbb{R}^n$ *(o* $\mathbb{C}^n$*). Entonces* $\|.\|$ *es una función continua de sus componentes. (Atkinson, página 143)*

**Teorema 3.7.2.** *Sean* $\|.\|$ *y* $N$ *dos normas sobre* $\mathbb{R}^n$ *o* $\mathbb{C}^n$*. Entonces existen constantes* $c_1$ *y* $c_2$*, positivas tales que*

$$c_1\|x\| \leq N(x) \leq c_2\|x\| \quad \forall\, x.$$

*(Atkinson, página 144)*

**Definición 3.7.3.** *Una sucesión de vectores* $\{x^{(1)}, x^{(2)}, \ldots, x^{(k)}, \ldots\}$ *en* $\mathbb{R}^n$ *o* $\mathbb{C}^n$ *converge a un vector* $x$ *si y solo si* $\|x - x^{(k)}\| \to 0$ *cuando* $k \to \infty$

Observar que no se especifica la norma. En los espacios de dimensión finita no importa cuál se elija ya que si $\|.\|$ y $N$ son dos normas, aplicando el teorema de equivalencias de normas se tiene,

$$c_1\|x - x^{(k)}\| \leq N(x - x^{(k)}) \leq c_2\|x - x^{(k)}\|$$

y

$$\|x - x^{(k)}\| \to 0 \quad \Longleftrightarrow \quad N(x - x^{(k)}) \to 0.$$

**Definición 3.7.4.** *Una norma matricial sobre el espacio $\mathbb{R}^{n,n}$ de matrices reales de orden $n$ es una aplicación $\|.\|$ de $\mathbb{R}^{n,n}$ en $\mathbb{R}_+$ tal que*

1. $\|A\| \geq 0 \quad \forall\ A$ *en* $\mathbb{R}^{n,n} \quad$ *y* $\quad \|A\| = 0 \quad \Longleftrightarrow \quad A = 0.$
2. $\|\lambda\| = |\lambda|\|A\|$
3. $\|A + B\| \leq \|A\| + \|B\|$
4. $\|AB\| \leq \|A\|B\|$

Observar que el conjunto de matrices de orden $n$ puede considerarse, mediante algún criterio para ordenar sus elementos, un espacio vectorial equivalente al espacio vectorial $\mathbb{R}^{n^2}$ pero con una operación adicional: el producto de matrices. Esta operación hace que no todas las normas vectoriales sean normas matriciales. En efecto, $N(A) = \max\limits_{i,j} |a_{ij}|$ no es una norma matricial según la definición 3.7.4 pues no satisface (4).

Existe una forma natural y más geométrica de definir la norma de una matriz a partir de una norma vectorial. Si $x \in \mathbb{R}^n$ y $\|.\|$ es una norma vectorial sobre $\mathbb{R}^n$ entonces $\|.\|$ es la longitud de $x$ y $\|Ax\|$ es la de $Ax$.

**Definición 3.7.5.** *Dada una norma vectorial en $\mathbb{R}^n$, se define la norma matricial subordinada a o inducida por dicha norma matricial mediante*

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

Observar que como $\|.\|$ es una función continua de sus componentes y $\{x/\|x\| = 1\}$ es un compacto, la norma matricial subordinada está bien definida. Es fácil ver que 3.7.5 verifica la definición de norma matricial. Por otro lado, es importante destacar que si $I$ es la matriz identidad, entonces

$$\|I\| = \max_{x \neq 0} \frac{\|Ix\|}{\|x\|} = 1$$

Además se verifica que

$$\|Ax\| \leq \|A\|\|x\|, \quad \forall\ x$$

No toda norma matricial es subordinada a una norma vectorial. Consideremos la norma de Fröbenius definida por

$$F(A) = (\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2)^{1/2}$$

Es fácil ver que es una norma matricial. Sin embargo $F(I) = \sqrt{n}$ si $n > 1$ y no puede ser en consecuencia una norma matricial subordinada a una norma vectorial. Más sorprendente, es compatible con la norma vectorial euclidiana en el sentido que,

$$\|Ax\|_2 \leq F(A)\|x\|_2 \quad \forall\, x$$

### Interpretación geométrica de la norma matricial subordinada

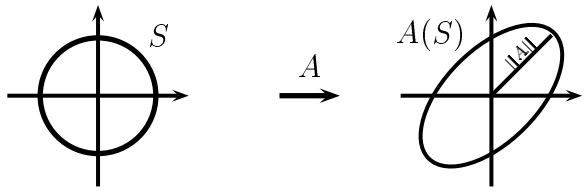
Sean

$$\|A\| = \max_{\|x\|=1} \|Ax\| \qquad \text{y} \qquad S = \{x : \ \|x\| = 1\}$$

La norma $\|A\|$ puede escribirse entonces

$$\|A\| = \max_{x\in S} \|Ax\| = \max_{z\in A(S)} \|z\|$$

con $A(S)$ =imagen de $S$ bajo $A$. Luego $\|A\|$ mide el efecto de $A$ sobre la esfera unitaria. Si $\|A\| > 1$, $\|A\|$ es el mayor estiramiento que ha sufrido.




### Cálculo de la norma matricial $\|A\|_1$

Vamos a mostrar que

$$\text{si} \quad \|x\|_1 = \sum_{i=1}^{n} |x_i| \quad \text{entonces} \quad \|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$$

conocida como la norma columna. Se tiene

$$\|Ax\|_1 \ = \ \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{ij}x_j| \leq \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}||x_j| = \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}|$$

$$\leq \ (\max_{j} \sum_{i=1}^{n} |a_{ij}|) \sum_{j=1}^{n} |x_j| = S\|x\|_1$$

siendo $S = \max\limits_{1\le j\le n} \sum\limits_{i=1}^{n} |a_{ij}|$. Por lo tanto como $\|Ax\|_1 \le \|A\|_1\|x\|_1$, resulta $S \ge \|A\|_1$.

Sea $k$ el índice donde se alcanza dicho máximo. Sea $x$ el $k$-ésimo vector canónico. Entonces $\|x\|_1 = 1$, $Ax = (a_{1k}, a_{2k}, \ldots, a_{nk})^T$ y $\|Ax\|_1 = \sum\limits_{i=1}^{n} |a_{ik}| = S\|x_1\|$. Por lo tanto $\|A\|_1 = S$.

**Cálculo de la norma matricial uniforme $\|A\|_\infty$**

Veremos que

$$\text{si} \quad \|x\|_\infty = \max_{1\le i\le n} |x_i| \quad \text{entonces} \quad \|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$$

conocida como la norma fila.

$$\begin{aligned} \|Ax\|_\infty &= \max_{1\le i\le n} |\sum_{j=1}^{n} a_{ij}x_j| \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}||x_j| = \\ &\le (\max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|)\|x\|_\infty = S\|x\|_\infty \end{aligned}$$

siendo $S = \max\limits_{1\le i\le n} \sum\limits_{j=1}^{n} |a_{ij}|$. Por lo tanto como $\|Ax\|_\infty \le \|A\|_\infty\|x\|_\infty$, resulta $S \ge \|A\|_\infty$.

Sea $k$ el índice donde se alcanza este máximo y sea $x$ tal que

$$x_j = \begin{cases} \dfrac{a_{kj}}{|a_{kj}|} & \text{si, } a_{kj} \ne 0 \\ 0 & \text{si no} \end{cases}$$

Vemos que $\|x\|_\infty = 1$. Además, para este $x$, se tiene

$$\|Ax\|_\infty = \max_{1\le i\le n} \left|\sum_{j=1}^{n} a_{ij}x_j\right| \ge \left|\sum_{j=1}^{n} a_{kj}x_j\right| = \sum_{j=1}^{n} |a_{ij}| = S = S\|x\|_\infty$$

Por lo tanto, $\|A\|_\infty \ge S$. En consecuencia $S = \|A\|_\infty$.

**Cálculo de la norma matricial $\|A\|_2$**

Para realizar este cálculo necesitamos recordar algunas definiciones y resultados del álgebra lineal.

**Definición 3.7.6.** *Sea $A$ una matriz de orden $n$. Un escalar $\lambda$ (real o complejo) es un autovalor o valor propio de $A$ si existe $x \neq 0$ tal que $Ax = \lambda x$. El vector $x$ se llama un autovalor o vector propio. El espectro de $A$ es el conjunto de todos los autovalores de $A$ y lo denotaremos $\sigma(A)$. El radio espectral de $A$ se define como*

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$$

Observe que si $\lambda$ es un autovalor entonces $(\lambda I - A)x = 0$ con $x \neq 0$ y esto es cierto si y solo si $\det(\lambda I - A) = 0$. Este último se llama el polinomio característico y los autovalores son las raíces del mismo. En consecuencia una matriz de orden $n$ tiene exactamente $n$ autovalores.

# Resultados

1. Los autovalores de una matriz y su traspuesta son los mismos.

   En efecto,

   $$\det(\lambda I - A^T) = \det(\lambda I - A)^T = \det(\lambda I - A)$$

2. Los autovalores de una matriz real y simétrica son reales.

   En efecto,

   $$\bar{\lambda}\bar{x}^T x = (\overline{Ax})^T x = \bar{x}^T \bar{A}^T x = \bar{x}^T A^T x = \bar{x}^T A x = \lambda \bar{x}^T x$$

   Luego

   $$\lambda x^T x = \bar{\lambda} x^T x \implies \lambda = \bar{\lambda}$$

3. Los autovalores de $B^T B$, $B$ real, son reales y no negativos. Esto es cierto pues $B^T B$ es simétrica y

   $$B^T Bx = \lambda x \implies \bar{x}^T B^T Bx = \lambda \bar{x}^T x \implies \|Bx\|_2^2 = \lambda \|x\|_2^2 \implies \lambda \geq 0$$

   Observar que $B$ es real, pero $\lambda$ puede ser nulo. Para que $\lambda > 0$ se necesita que $B$ sea no singular.

4. Si $B = B^T$ con autovalores $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ entonces

   $$\lambda_1 \|x\|_2^2 \leq x^T Bx \leq \lambda_n \|x\|_2^2 \qquad \forall\, x \in \mathbb{R}^n$$

   (ver Atkinson)

Pasemos ahora al cálculo de $\|A\|_2$. Mostraremos que

$$\|A\|_2 = (\rho(A^T A))^{1/2}$$

Sea $r^2 = \rho(A^T A)$. Por el resultado 3 los autovalores de $A^T A$ son no negativos y $r^2$ es su máximo autovalor. Por el resultado 4

$$\|Ax\|_2^2 \leq r^2 \|x\|_2^2 \qquad \forall\, x \in \mathbb{R}^n$$

y por lo tanto

$$\|A\|_2 \leq r$$

Sea $u \neq 0$ tal que

$$A^T A u = r^2 u$$

Entonces

$$u^T A^T A u = r^2 u^T u \qquad \text{y} \qquad \|Au\|_2^2 = r^2 \|u\|_2^2$$

Es así que $\|A\|_2 \geq r$ y en consecuencia $\|A\|_2 = r$

**Ejemplo 3.7.2.** *Sea* $A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & -5 \\ 0 & 3 & -1 \end{pmatrix}$. *Se tiene que* $A^T A = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 13 & -17 \\ 0 & -17 & 26 \end{pmatrix}$ *y*

$$\begin{aligned} p(\lambda) &= (\lambda - 4)(\lambda - 13)(\lambda - 26) - (\lambda - 4)289 \\ &= (\lambda - 4)(\lambda^2 - 39\lambda + 338 - 289) \\ &= (\lambda - 4)(\lambda^2 - 39\lambda + 49) \end{aligned}$$

*Los autovalores de* $A^T A$ *son:*

$$\lambda_1 = 4 \qquad \lambda_2 = (39 + 5\sqrt{53})/2 \qquad \lambda_3 = (39 - 5\sqrt{53})/2$$

*y en consecuencia*

$$\|A\|_2 = (\lambda_2)^{1/2} \approx 37.003$$

## 3.8 Análisis del error

En general, cuando resolvemos el sistema

$$Ax = b, \qquad A \in \mathbb{R}^{n \times n}, \quad x, b \in \mathbb{R}^n$$

obtenemos una solución $\tilde{x}$ calculada que difiere de la solución exacta $x = A^{-1}b$. Los errores puede tener diferentes orígenes. Por ejemplo, los coeficientes de $A$ y $b$ pueden provenir de medidas o cálculos aproximados. La representación de estos elementos en una computadora implican automáticamente un error en aquéllos coeficientes que no pueden ser representados exactamente con el número de cifras significativas del computador. Además, la aplicación de un algoritmo que utiliza las operaciones elementales, nos da resultados aproximados, pues el resultado de cada operación elemental es un número afectado por errores de redondeo. Se trata entonces de buscar una estimación del error

$$\tilde{x} - x = \tilde{x} - A^{-1}b$$

La primera idea que surge es la de relacionar este error con el residuo o vector residual:

$$r = b - A\tilde{x}$$

Intuitivamente parecería razonable decir que $\|r\|$ pequeño implica $\|\tilde{x} - x\|$ pequeño y $\tilde{x}$ es una buena aproximación. Sin embargo, en el siguiente ejemplo

$$A = \begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix} \qquad b = \begin{pmatrix} 3 \\ 3.0001 \end{pmatrix}$$

el vector residual correspondiente a $\tilde{x} = (3, 0)^T$ es $r = (0, 0.0002)^T$ y por lo tanto $\|r\|_\infty$ es pequeño; podríamos concluir erróneamente que $\|\tilde{x} - x\|_\infty$ es pequeño. La solución exacta es $x = (1, 1)^T$ y en consecuencia $\|\tilde{x} - x\|_\infty = 2$. Aquí aprendimos que si $\|r\|$ es pequeño, no podemos asegurar que $\tilde{x}$ es una buena aproximación de $x$.

En este ejemplo notemos que la solución del sistema representa la intersección de las rectas

$$\begin{aligned} \mathrm{l}_1 : &\qquad x_1 + 2x_2 = 3 \\ \mathrm{l}_2 : &\qquad 1.0001x_1 + 2x_2 = 3.0001 \end{aligned}$$

que son casi paralelas. Si las rectas no fuesen casi paralelas se esperaría que un residuo pequeño implicara una buena aproximación.

Analicemos ahora el siguiente sistema lineal

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \qquad \text{con solución} \qquad x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

y también el sistema perturbado

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \end{pmatrix} = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix} \qquad \text{con solución} \qquad \tilde{x} = \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix}$$

Podemos observar que un error de 1/200 en el segundo miembro produce un error relativo del orden de 10/1 en el resultado, lo cual significa una amplificación de los errores relativos de los datos del orden de 2000 (pues $2000 \cdot \dfrac{1}{200} = \dfrac{10}{1}$)

Consideremos ahora el sistema perturbado (esta vez en la matriz)

$$\begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.98 & 9 \\ 6.99 & 4.99 & 9 & 9.980 \end{pmatrix} \begin{pmatrix} \tilde{\tilde{x}}_1 \\ \tilde{\tilde{x}}_2 \\ \tilde{\tilde{x}}_3 \\ \tilde{\tilde{x}}_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \quad \text{con solución} \quad \tilde{\tilde{x}} = \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}$$

Aquí nuevamente pequeños cambios en los datos (los elementos de la matriz) alteran los resultados.

La matriz del sistema es simétrica, su determinante vale 1 y la inversa

$$A^{-1} = \begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}$$

es inofensiva.

Vamos entonces a realizar un estudio del error para averiguar qué ocurre en estos ejemplos.

**Teorema 3.8.1.** *Si $\tilde{x}$ es una aproximación de la solución de $Ax = b$ y $A$ es una matriz no singular, entonces para cualquier norma subordinada,*

$$\|x - \tilde{x}\| \leq \|r\|\|A^{-1}\| \qquad y \qquad \frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|}, \quad si \quad x \neq 0, b \neq 0$$

**Prueba:** Como $r = b - A\tilde{x} = Ax - A\tilde{x}$ y $A$ es no singular resulta $x - \tilde{x} = A^{-1}r$ de donde $\|x - \tilde{x}\| = \|A^{-1}r\| \leq \|A^{-1}\|\|r\|$. Como $b = Ax$, $\|b\| \leq \|A\|\|x\|$ y

por lo tanto $\|x\| \geq \dfrac{\|b\|}{\|A\|}$. Resulta

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|}$$

**Nota 1:** Observar que se tiene en términos del error relativo

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|},$$

donde

$$A(x + \delta x) = b + \delta b \qquad \text{y} \qquad \tilde{x} = x + \delta x$$

En efecto, con $A\tilde{x} = b + \delta b \implies b - A\tilde{x} = -\delta b = r$

**Nota 2:** Las desigualdades del teorema implican que las cantidades $\|A^{-1}\|$ y $\|A\|\|A^{-1}\|$ pueden ser usadas para explicar una conexión entre el residuo y el error absoluto y relativo

**Definición 3.8.1.** *El número de condición $\kappa(A)$ de la matriz no singular $A$ se define como*

$$\kappa(A) = \|A\|\|A^{-1}\|$$

Con esta notación las desigualdades del teorema se escriben

$$\|x - \tilde{x}\| \leq \kappa(A)\frac{\|r\|}{\|A\|} \qquad\qquad \frac{\|x - \tilde{x}\|}{\|x\|} \leq \kappa(A)\frac{\|r\|}{\|b\|}$$

**Nota 3:** Observar que si $A$ es no singular,

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = \kappa(A)$$

Si $\kappa(A)$ está cerca de uno se espera que $A$ tenga un buen comportamiento; decimos en este caso que $A$ está **bien condicionada** y si $\kappa(A)$ es significativamente mayor que uno, decimos que $A$ está **mal condicionada**.

**Ejemplo 3.8.1.** *Apliquemos estos resultados a algunos casos.*

1. *La matriz considerada al comienzo es tal que $\|A\|_\infty = 30.001$ con una inversa*

$$A^{-1} = \begin{pmatrix} -10.000 & 10.000 \\ 5000.5 & -5000 \end{pmatrix}$$

*Es decir*

$$\|A^{-1}\|_\infty = 20.000 \qquad y \qquad \kappa(A) = 60.002$$

2. *Para la segunda matriz se tiene* $\|A\|_\infty = 33$, $\|A^{-1}\|_\infty = 136$, $\kappa(A) = 4.488$

3. *La matriz de Hilbert de orden* $n$, *es una matriz invertible y está dada por:*

$$H_n = \begin{pmatrix} 1 & 1/2 & 1/3 & \cdots & 1/n \\ 1/2 & 1/3 & \cdots & \cdots & 1/(n+1) \\ 1/3 & 1/4 & \cdots & \cdots & 1/(n+2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1/n & 1/(n+1) & \cdots & \cdots & 1/(2n-1) \end{pmatrix}$$

*La inversa* $H_n^{-1}$ *se conoce explícitamente; si* $H_n^{-1} = (h_{ij}) \quad 1 \leq i, j \leq n$, *entonces*

$$h_{ij} = \frac{(-1)^{i+j}(n+i-1)!(n+j-1)!}{(i+j-1)[(i-1)!(j-1)!]^2(n-i)!(n-j)!} \quad 1 \leq i \leq n, \;\; 1 \leq j \leq n$$

*Para* $n = 6$, *por ejemplo,* $\kappa_\infty(H_6) = 1.5 \cdot 10^7$

*La matriz de Hilbert es un ejemplo clásico para comprobar la eficiencia de los métodos de resolución de ecuaciones en tratar problemas mal condicionados.*

Hasta ahora el análisis del error lo hemos hecho suponiendo que $A$ se puede representar exactamente. Supongamos que se tenga

$$(A + \delta A)x = b + \delta b$$

en vez de $Ax = b$. Normalmente si $\|\delta A\|$ y $\|b\|$ son pequeños, $\tilde{x}$ debe ser tal que $\|x - \tilde{x}\|$ es pequeño. Sin embargo cuando las matrices son mal condicionadas, los errores de redondeo pueden hacer que $\|x - \tilde{x}\|$ sea grande.

**Teorema 3.8.2.** *Supongamos que* $A$ *es no singular y que* $\|\delta A\| \leq \dfrac{1}{\|A\|}$. *Entonces la solución* $\tilde{x}$ *de* $(A + \delta A)x = b + \delta b$ *aproxima a la solución* $x$ *de* $Ax = b$ *con una estimación del error relativo dada por*

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\dfrac{\|\delta A\|}{\|A\|}} \left( \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right)$$

**Prueba:** Ver W. Kahan, Numerical linear algebra, *Canadian Math. Bull.*, 1966, pp. 756-801.

Es decir, si $\kappa(A)$ no es muy grande, entonces cambios pequeños en $A$ y $b$ producen cambios pequeños en $x$.

## 3.9 El método de refinamiento iterativo

Supongamos que $Ax = b$ ha sido resuelto por descomposición $LU$ y que esta descomposición, así como los índices de la fila pivote se han almacenado. Sea $x^{(0)}$ la solución calculada y $r^{(0)} = b - x^{(0)}$ el residuo. Si $\delta x^{(0)} = x - x^{(0)}$, entonces

$$A\delta x^{(0)} = A(x - x^{(0)}) = Ax - Ax^{(0)} = b - Ax^{(0)} = r^{(0)}.$$

Es decir $\delta x^{(0)}$ es la solución de $A\delta x^{(0)} = r^{(0)}$ y por lo tanto $LU\delta x^{(0)} = r^{(0)}$.

Definamos una nueva solución aproximada

$$x^{(1)} = x^{(0)} + \delta x^{(0)}$$

y repetimos el proceso para $s = 1, 2, 3, \cdots$

$$\begin{cases} r^{(s)} &= b - Ax^{(s)} \\ LU\delta x^{(s)} &= r^{(s)} \\ x^{(s+1)} &= x^{(s)} + \delta x^{(s)} \end{cases}$$

El cálculo de $r^{(s)}$ lleva $n^2$ operaciones y el de $\delta x^{(s)}$ $n^2$ operaciones más. Es decir, para calcular $x^{(1)}, x^{(2)}, \cdots$ (los valores mejorados) se requieren $2n^2$ operaciones por paso, un número muy inferior al que se refiere para calcular $x^{(0)}$. Este método se conoce como *refinamiento iterativo o conexión residual.*

Es muy importante obtener valores previos de $r^{(0)}$, puesto que $x^{(0)}$ resuelve aproximadamente $Ax = b$, $r^{(0)}$ encerrará pérdidas de cifras significativas y por lo tanto $r^{(0)}$ debe calcularse en doble precisión.

## 3.10 Métodos Iterativos

Cuando el sistema lineal es muy grande, los métodos directos, basados en la eliminación gaussiana deben ser sustituidos por otros métodos más rápidos y con menor requerimiento de memoria como por ejemplo los métodos iterativos.

Una técnica iterativa para resolver $Ax = b$ empieza con una aproximación inicial $x^{(0)}$, y genera una sucesión de vectores $\{x^{(k)}\}_{k=0}^{\infty}$ que converge a la solución $x$.

Estas técnicas son muy eficientes para sistemas grandes con un gran porcentaje alto de ceros. El campo de mayor aplicación es el de las ecuaciones en derivadas parciales, en donde surgen sistemas de orden muy grande.

Comenzaremos definiendo y analizando algunos métodos iterativos clásicos. Si en $Ax = b$, los coeficientes $a_{ii}$ son todos no nulos, podemos despejar $x_i$ de la i-ésima ecuación

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad i = 1:n$$

obteniéndose

$$x_i = \left( \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j + b_i \right) / a_{ii} \quad i = 1:n \tag{3.2}$$

### 3.10.1 Método de Jacobi

Calculamos una sucesión de aproximaciones $x^{(1)}, x^{(2)}, \cdots$

$$x_i^{(k+1)} = \left( -\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^{(k)} + b_i \right) / a_{ii} \quad i = 1:n \tag{3.3}$$

con $x^{(0)}$ dado.

Observar que si la sucesión de vectores $x^{(k)} \to x$, es decir si $\|x^{(k)} - x\| \to 0$ cuando $k \to \infty$, tomando límites en (3.3) resultaría

$$x_i = \left( -\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j + b_i \right) / a_{ii} \quad i = 1:n$$

que es la solución de $Ax = b$.

### 3.10.2 Método de Gauss-Seidel

Observar que en Jacobi los valores mejorados no se usan si no en la iteración siguiente. Si dichos valores se incorporan inmediatamente obtenemos el método de Gauss-Seidel:

$$x_i^{(k+1)} = \left( -\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i \right) / a_{ii} \quad i = 1:n \tag{3.4}$$

Podemos entonces resumir ambos algoritmos de la siguiente forma:

**Algoritmo de Jacobi:** Jacobi$(A, b, x_0, tol, itmax)$

**E**ntrada:

$A$ : matriz de coeficientes del sistema $Ax = b$
$b$ : vector del $2^{\text{do}}$ miembro de $Ax = b$
$x_0$ : aproximación inicial
$tol$ : tolerancia en el error
$itmax$ : número máximo de iteraciones

**S**alida

$x_0$ : valor calculado de la solución

1. $k := 1$
2. mientras $k \leq itmax$ hacer:
3. para $i := 1 : n$

   $x_i := (-\sum_{j=1, j\neq i}^{n} a_{ij} x_{0_j}^{(k)} + b_i)/a_{ii}$
4. si $\|x - x_0\| \leq tol$ entonces $x_0 = x$. Stop
5. si no $k := k + 1$, $x_0 := x$

**Algoritmo de Gauss-Seidel:** Gauss-Seidel $(A, b, x_0, tol, itmax)$.

Sustituir 3. en Jacobi por

$$x_i = (-\sum_{j=1}^{i-1} a_{ij} x_j - \sum_{j=i+1}^{n} a_{ij} x_{0_j} + b_i)/a_{ii}$$

Los pasos 3. de ambos algoritmos requieren que $a_{ii} \neq 0$, si este no es el caso, se puede efectuar un reordenamiento de las ecuaciones para que ningún $a_{ii}$ resulte cero. Se sugiere que las ecuaciones sean arregladas de manera que $a_{ii}$ sea lo más grande posible. El criterio de parada utilizado es iterar hasta que

$$\|x^{(k)} - x^{(k-1)}\| < tol$$

Otro criterio es hacer que el error relativo sea menor que una cierta tolerancia

$$\frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^{(k)}\|} < tol$$

## 3.11 Estudio de la convergencia de los Métodos Iterativos

Mostraremos que los métodos de Jacobi y de Gauss-Seidel pueden escribirse como

$$x^{(k+1)} = Bx^{(k)} + c \qquad\qquad k = 0, 1, 2, \cdots \tag{3.5}$$

$x \in \mathbb{R}$, $B \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$. Esta es la forma general de los *métodos iterativos estacionarios*, llamados así porque $B$ y $c$ son constantes en todas las iteraciones. La matriz $B$ se llama la *matriz de iteración*.

Para obtener (3.5), escribamos $A$ así

$$A = D(L + I + U)$$

donde $D = diag(a_{11}, \cdots, a_{nn})$,

$$L_{ij} = \begin{cases} \dfrac{a_{ij}}{a_{ii}} & \text{si } j < i \\ 0 & \text{si } j \geq i \end{cases} \qquad \text{y} \qquad \begin{cases} \dfrac{a_{ij}}{a_{ii}} & \text{si } j > i \\ 0 & \text{si } j \leq i \end{cases}$$

El método de Jacobi se escribe ahora

$$x^{(k+1)} = -(L + U)x^{(k)} + D^{-1}b$$

y el de Gauss-Seidel

$$x^{(k+1)} = -Lx^{(k+1)} - Ux^{(k)} + D^{-1}b$$

de donde

$$(I + L)x^{(k+1)} = -Ux^{(k)} + D^{-1}b$$

y por lo tanto

$$x^{(k+1)} = -(I + L)^{-1}Ux^{(k)} + (I + L)^{-1}D^{-1}b$$

Entonces la matriz de iteración en cada método es:

| Jacobi | Gauss-Seidel |
|---|---|
| $B_J = -(I + L)$ | $B_{GS} = -(I + L)^{-1}U$ |

Si $B$ es $B_J$ ó $B_{GS}$ la ecuación vectorial

$$x = Bx + c$$

tiene una única solución pues $(I-B)^{-1}$ existe. En efecto, para Jacobi:

$$I - B_J = I + L + U = D^{-1}A$$

y para Gauss-Seidel:

$$I - B_{GS} = I + (I+L)^{-1}U = (I+L)^{-1}(I+L+U) = (I+L)^{-1}D^{-1}A$$

Estudiemos ahora la convergencia de la sucesión $\{x^{(k)}\}$. Esto requerirá de los siguientes lemas:

**Lema 3.11.1.** *Si $\rho(B) < 1$ entonces $(I-B)^{-1}$ existe y además*

$$(I-B)^{-1} = I + B + B^2 + B^3 + \cdots = \sum_{k=0}^{+\infty} B^k$$

**Prueba:** ver Burden-Faires, pag. 469

**Lema 3.11.2.** *Las siguientes afirmaciones son equivalentes:*

*i)* $\lim_{k\to\infty} (A^k)_{ij} = 0 \qquad i,j = 1,2,\cdots,n$

*ii)* $\lim_{k\to\infty} \|A^k\| = 0$ *para alguna norma inducida*

*iii)* $\rho(A) < 1$

*iv* $\lim_{k\to\infty} A^k x = 0 \qquad \forall\ x$

**Prueba:** Isaacson-Keller, página. 14.

**Lema 3.11.3.** *$\rho(A) < \|A\|$, para cualquier norma inducida*

**Prueba:** Sea $\lambda$ tal que $\rho(A) = |\lambda|$ y $x$ el autovector asociado a él. Entonces $Ax = \lambda x \implies \|Ax\| = |\lambda|\|x\|$, de donde

$$\rho(A) = |\lambda| = \frac{\|Ax\|}{\|x\|} \leq \max_{z\neq 0} \frac{\|Az\|}{\|z\|} = \|A\|$$

□

**Teorema 3.11.1.** *Para cualquier $x^{(0)} \in \mathbb{R}^n$, la sucesión $\{x^k\}_{k=0}^{\infty}$ determinada por*

$$x^{(k)} = Bx^{(k-1)} + c \qquad \text{para} \quad k \geq 1,\ c \neq 0$$

*converge a la solución única $x = Bx + c$ si y solo si $\rho(B) < 1$*

**Demostración:**

$$\begin{aligned}
x^{(1)} &= Bx^{(0)} + c \\
x^{(2)} &= Bx^{(1)} + c = B^2x^{(0)} + (B+I)c \\
x^{(3)} &= Bx^{(2)} + c = B^3x^{(0)} + (B^2+B+I)c \\
\vdots &= \qquad \vdots \\
x^{(k)} &= Bx^{(k-1)} + c = B^kx^{(0)} + (B^{k-1}+\cdots+B^2+B+I)c
\end{aligned}$$

Por lo tanto,

$$\lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} B^k x^{(0)} + \lim_{k\to\infty} \left(\sum_{j=0}^{k-1} B^j\right) c$$

Si $\rho(B) < 1$, aplicando los lemas 1 y 2:

$$\lim_{k\to\infty} x^{(k)} = 0 + (I-B)^{-1}c$$

Por lo tanto $\{x^k\}$ converge a $x = (I-B)^{-1}c$ y ésta es la única solución de $x = Bx + c$.

Recíprocamente, sea $\{x^k\}$ tal que $x^{(k)} \to x \quad \forall\, x^{(0)}$. Entonces, $x$ verifica $x = Bx + c$ y

$$x - x^{(k)} = Bx + c - Bx^{(k-1)} - c = B(x - x^{(k-1)}) = \cdots = B^k(x - x^{(0)})$$

Por lo tanto, para todo $x^{(0)}$

$$\lim_{k\to\infty} B^k(x - x^{(0)}) = \lim_{k\to\infty} (x - x^{(k)}) = 0$$

Siendo $x_0$ un valor arbitrario, $z = x^{(0)} - x$ también lo es y entonces

$$\lim_{k\to\infty} B^k z = 0 \quad \forall\, z$$

y por lo tanto, por el lema 2 iv) se tiene que $\rho(B) < 1$ □

**Corolario 3.11.1.** *Una condición suficiente para que un método iterativo estacionario $x^{(k)} = Bx^{(k-1)} + c$ converja para toda aproximación inicial $x^{(0)}$ es que $\|B\| < 1$ para alguna norma matricial inducida.*

**Prueba:** Por el lema 3, $\rho(B) \le \|B\|$, para toda norma matricial inducida. Si $\|B\| \le 1 \implies \rho(B) < 1$ y el teorema anterior se aplica. □

### 3.11.1 Estimación del error

Sabemos que $x^{(k)} - x = B(x^{(k)} - x)$ y esto se puede escribir

$$x^{(k)} - x = -B(x^{(k)} - x^{(k-1)}) + B(x^{(k)} - x)$$

Supongamos que $\beta = \|B\| < 1$, entonces

$$\begin{aligned} \|x^{(k)} - x\| &\leq \|B\|\|x^{(k)} - x^{(k-1)}\| + \|B\|\|x^{(k)} - x\| \\ &= \beta\|x^{(k)} - x^{(k-1)}\| + \beta\|x^{(k)} - x\| \end{aligned}$$

obteniéndose

$$(1-\beta)\|x^{(k)} - x\| \leq \beta\|x^{(k)} - x^{(k-1)}\|$$

es decir

$$\|x^{(k)} - x\| \leq \frac{\beta}{(1-\beta)}\|x^{(k)} - x^{(k-1)}\|$$

Las iteraciones alcanzan una precisión $\varepsilon$ cuando

$$\frac{\beta}{(1-\beta)}\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon$$

y el criterio de parada resulta en

$$\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon\frac{(1-\beta)}{\beta}$$

Observar que $x^{(k)} - x = B(x^{(k-1)} - x)$ dice que la convergencia es lineal pues

$$\|\varepsilon^{(k)}\| \leq \|B\|\|\varepsilon^{(k-1)}\|$$

Por lo tanto, estos métodos iterativos son buenos candidatos para el proceso de aceleración de Aitken:

$$\begin{aligned} \tilde{x}_i^{(k)} &= x_i^{(k+2)} - \frac{\left(x_i^{(k+2)} - x_i^{(k+1)}\right)^2}{\left(x_i^{(k+2)} - x_i^{(k+1)}\right) - \left(x_i^{(k+1)} - x_i^{(k)}\right)} \\ &= x_i^{(k+2)} - \frac{\left(\Delta x_i^{(k+1)}\right)^2}{\Delta^2 x_i^{(k)}} \qquad i = 1, 2, \cdots, n \end{aligned}$$

**Nota:** Para obtener una estimación de $\beta = \|B\|$, se procede análogamente al caso de los métodos iterativos para ecuaciones no lineales

$$x^{(k+1)} - x^{(k)} = B(x^{(k)} - x^{(k-1)}) \qquad k \geq 1$$

Luego

$$\|x^{(k+1)} - x^{(k)}\| = \|B\|\|x^{(k)} - x^{(k-1)}\| = \beta\|x^{(k)} - x^{(k-1)}\|$$

y finalmente

$$\frac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k)} - x^{(k-1)}\|} \approx \beta \qquad \text{para } k \text{ grande}$$

**Definición 3.11.1.** *$A$ es fuertemente diagonal dominante si*

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \qquad \forall\ i$$

**Ejemplo 3.11.1.**

$$A = \begin{pmatrix} 5 & -1 & 0 & 0 \\ 5 & 10 & -2 & 0 \\ 0 & -1 & 2 & 1/2 \\ -3 & -1 & -5 & -10 \end{pmatrix}$$

**Teorema 3.11.2.** *Si $A$ es fuertemente diagonal dominante los métodos iterativos de Gauss-Seidel y Jacobi son convergentes*

**Demostración:**

Método de Jacobi: Observar que

$$(B_J)_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}} & \text{si, } i \neq j \\ 0 & \text{si, } i = j \end{cases}$$

Como

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

calculando la norma del máximo

$$\|B_J\|_\infty = \max_i \sum_{j=1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1$$

Método de Gauss-Seidel:

$$\|B_{GS}\|_\infty = \max_{x\neq 0} \frac{\|B_{GS}x\|_\infty}{\|x\|_\infty}.$$

Sea $y = B_{GS}x, \quad x \neq 0$ y $k$ tal que

$$|y_k| = \max_i |y_i| = \|y\|_\infty = \|B_{GS}x\|_\infty.$$

Como $y = -(I+L)^{-1}Ux$ se tiene que $(I+L)y = -Ux$ lo cual implica que $y = -Ly - Ux$, obteniéndose

$$y_k = -\sum_{j=1}^{k-1} \frac{a_{kj}}{a_{kk}} y_j - \sum_{j=k+1}^{n} \frac{a_{kj}}{a_{kk}} x_j$$

$$\begin{aligned}
\|y\|_\infty = |y_k| &\leq \sum_{j=1}^{k-1} \left|\frac{a_{kj}}{a_{kk}}\right| |y_j| + \sum_{j=k+1}^{n} \left|\frac{a_{kj}}{a_{kk}}\right| |x_j| \\
&\leq \|y\|_\infty \sum_{j=1}^{k-1} \left|\frac{a_{kj}}{a_{kk}}\right| + \|x\|_\infty \sum_{j=k+1}^{n} \left|\frac{a_{kj}}{a_{kk}}\right| \\
&\leq \|y\|_\infty s_k + \|x\|_\infty r_k,
\end{aligned}$$

$$\text{con} \quad s_k = \sum_{j=1}^{k-1} \left|\frac{a_{kj}}{a_{kk}}\right| \quad \text{y} \quad r_k = \sum_{j=k+1}^{n} \left|\frac{a_{kj}}{a_{kk}}\right|$$

Entonces, $\|y\|_\infty \leq \dfrac{r_k}{1-s_k}\|x\|_\infty$, es decir

$$\frac{\|B_{GS}x\|_\infty}{\|x\|_\infty} \leq \frac{r_k}{1-s_k}$$

Por lo tanto

$$\|B_{GS}\|_\infty = \max_{x\neq 0} \frac{\|B_{GS}x\|_\infty}{\|x\|_\infty} \leq \max_{1\leq i\leq n} \frac{r_i}{1-s_i}$$

$$\text{Como } r_i + s_i = \sum_{j \neq i}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1 \qquad \Longrightarrow \qquad \frac{r_i}{1 - s_i} < 1, \quad \text{resulta}$$

$$\|B_{GS}\|_\infty < 1 \qquad \square$$

En general no se conoce cuál de las dos técnicas, la de Jacobi o la de Gauss-Seidel, deba usarse, salvo en un caso especial

**Teorema 3.11.3.** *Si $a_{ij} \leq 0$ para $i \neq j$ y $a_{ii} > 0 \ \forall\, i$, entonces una y sólo una de las siguientes condiciones se cumple:*

*1)* $\rho_J = \rho_{GS} = 0$

*2)* $\rho_J = \rho_{GS} = 1$

*3)* $0 < \rho_{GS} < \rho_J < 1$

*4)* $1 < \rho_J < \rho_{GS}$

*en donde $\rho_J = \rho(B_J)$ y $\rho_{GS} = \rho(B_{GS})$*

Este resultado es debido a Stein y Rosenberg. Nos dice que las iteraciones de Jacobi y Gauss-Seidel convergen ambas o divergen ambas; pero cuando convergen el método de Gauss-Seidel es más rápido que el de Jacobi, excepto para el caso trivial 1). Por ello el método de Jacobi es raras veces usado.

## 3.12 Aceleración de los procesos iterativos estacionarios: Métodos de sobre-relajación sucesiva (S.O.R.)

El método de Gauss-Seidel (3.4) puede escribirse

$$x_i^{(k+1)} = x_i^{(k)} + r_i^{(k)} \quad i = 1 : n$$

con

$$r_i^{(k)} = \left( -\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} + b_i \right) / a_{ii} \quad i = 1 : n$$

Definamos la siguiente modificación

$$x_i^{(k+1)} = x_i^{(k)} + \omega r_i^{(k)} \qquad i = 1 : n$$

Para ciertos valores $\omega$, tendremos una convergencia significativamente más rápida. Estos métodos se conocen como *métodos de relajación.* Para $0 < \omega < 1$, se llaman *de subrelajación* y se pueden emplear para obtener la convergencia de algunos sistemas que no son convergentes por el método de Gauss-Seidel. Para $1 < \omega$ los procedimientos se llaman método de sobre-relajación y se pueden usar para acelerar la convergencia de sistemas que son convergentes por el método de Gauss-Seidel. Estos métodos se abrevian S.O.R (Successive Over-Relaxation).

Estos métodos pueden expresarse en la forma matricial siguiente

$$x^{(k+1)} = Bx^{(k)} + c$$

En efecto,

$$x^{(k+1)} = x^{(k)} + \omega(-Lx^{(k+1)} - Ux^{(k)} - Ix^{(k)}) + \omega D^{-1}b$$

de donde

$$\begin{aligned}(I + \omega L)x^{(k+1)} &= (1-\omega)x^{(k)} - \omega U x^{(k)} + \omega D^{-1}b \\ &= \big((1-\omega)I - \omega U\big)x^{(k)} + \omega D^{-1}b\end{aligned}$$

Como $(I + \omega L)$ es invertible

$$x^{(k+1)} = (I + \omega L)^{-1}[(1-\omega)I - \omega U]x^{(k)} + (I + \omega L)^{-1}\omega D^{-1}b$$

Resultando

$$\boxed{B_\omega = (I + \omega L)^{-1}[(1-\omega)I - \omega U]}$$

¿Cómo se escoge el valor apropiado de $\omega$?. Aún cuando no se conoce una respuesta completa, los siguientes resultados pueden usarse.

**Teorema 3.12.1.** *Si $a_{ii} \neq 0 \ \ \forall\ i$, entonces $\rho(B_\omega) \geq |\omega - 1|$. Esto nos dice que para que $\rho(B_\omega) < 1$ necesariamente $0 < \omega < 2$.*

**Demostración:**

$$\begin{aligned}\det B_\omega &= \det\{(I + \omega L)^{-1}[(1-\omega)I - \omega U]\} \\ &= \det(I + \omega L)^{-1} \det[(1-\omega)I - \omega U] \\ &= 1 \cdot (1-\omega)^n\end{aligned} \tag{3.6}$$

Por otro lado,

$$\det(B_\omega - \lambda I) = (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n), \qquad \text{con} \quad \lambda_1, \lambda_2, \cdots, \lambda_n$$

los valores propios de $B_\omega$. Entonces, para $\lambda = 0$

$$detB_\omega = \lambda_1 \cdot \lambda_2 \cdot \cdots \cdot \lambda_n \tag{3.7}$$

De (3.6) y (3.7) resulta

$$|\lambda_1||\lambda_2|\cdots|\lambda_n| = |1-\omega|^n = |\omega - 1|^n$$

Siendo

$$\rho(B_\omega) = \max_{1\leq i\leq n} |\lambda_i| \quad \Longrightarrow \quad \rho(B_\omega)^n \geq |\omega - 1|^n \quad \Longleftrightarrow \quad \rho(B_\omega) \geq |\omega - 1|$$

Como un método iterativo estacionario es convergente si y solo si el radio espectral de la matriz de iteración es menor que uno, se tiene

$$\rho(B_\omega) < 1 \quad \Longrightarrow \quad |\omega - 1| < 1 \quad \Longleftrightarrow \quad 0 < \omega < 2$$

**Teorema 3.12.2.** *Si $A$ es definida positiva y tridiagonal, entonces $\rho(B_{B_{GS}}) = [\rho(B_J)]^2 < 1$ y la elección óptima de $\omega$ para el método SOR es*

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(B_J)]^2}}$$

*Para este valor de $\omega$, $\rho(B_\omega) = \omega - 1$*

**Ejemplo 3.12.1.** *Para la matriz*

$$A = \begin{pmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}$$

*que es diagonal dominante y definida positiva se tiene*

$$B_J = \begin{pmatrix} 0 & -0.75 & 0 \\ -0.75 & 0 & 0.25 \\ 0 & 0.25 & 0 \end{pmatrix}$$

*y por lo tanto*

$$\det(B_J - \lambda I) = -\lambda(\lambda^2 - 0.625),$$

*de donde*

$$\rho(B_J) = \sqrt{0.625} = 0.791 \qquad y \qquad \rho(B_{B_{GS}}) = 0.625$$

*El valor optimal de* $\omega$ *es*

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(B_J)]^2}} = \frac{2}{1 + \sqrt{1 - 0.625}} \approx 1.24$$

*y el radio espectral correspondiente,*

$$\rho(B_\omega) = 0.24,$$

*lo cual nos da*

$$\rho(B_\omega) < \rho(B_{B_{GS}}) < \rho(B_J)$$
$$0.24 < 0.625 < 0.791$$

# Capítulo 4

# Interpolación Polinomial

## 4.0 Aproximación e Interpolación

Son muchas las razones para aproximar o interpolar funciones. El tipo de aproximación depende de la aplicación buscada y de la facilidad o dificultad para obtenerla. En cualquier caso, las funciones aproximantes más simples parecieran ser los polinomios. Existen otros tipos, no menos importantes, tales como las funciones racionales, las trigonométricas, las polinomiales a trozos, etc. Estudiaremos la aproximación de funciones continuas en un intervalo cerrado y acotado. En general, un polinomio $p_n(x)$ de grado $\leq n$ se dice una aproximación a una función $f(x)$ en $[a,b]$ si alguna medida de la diferencia

$$p_n(x) - f(x)$$

en $[a,b]$, es pequeña. Esto llega a ser preciso cuando la medida de esta diferencia y la magnitud de lo pequeño son especificadas. Sea $V = C[a,b]$ el espacio vectorial de las funciones continuas sobre $[a,b]$.

**Definición 4.0.1.** *Una* norma *es una función de* $V$ *en* $\mathbb{R}^+$ *tal que*

*(i)* $\|f\| \geq 0 \quad \forall f \in V$

*(ii)* $\|f\| = 0 \;\Leftrightarrow\; f = 0 \quad \forall f \in V$

*(iii)* $\|\lambda f\| = |\lambda|\|f\| \quad \forall \lambda \in \mathbb{R}$

*(iv)* $\|f + g\| \leq \|f\| + \|g\| \quad \forall f, g \in V$

**Ejemplo 4.0.1.**

*1)* $\|f\|_\infty = \max\limits_{x\in[a,b]} |f(x)|$ *norma infinita*

*2)* $\|f\|_2 = \left[\int_a^b |f(x)|^2 dx\right]^{1/2}$ *norma* $L_2$

*3)* $\|f\|_{2,\infty} = \left[\int_a^b |f(x)|^2 w(x) dx\right]^{1/2}$, *siendo* $w(x) > 0$ *y continua en* $(a,b)$

*4)* $\|f\|_p = \left[\int_a^b |f(x)|^p dx\right]^{1/p}$ $p \geq 1$ *norma* $L_p$

Una medida de la diferencia o error en la aproximación de $f(x)$ por $p_n(x)$ se denota

$$e(f(x) - p_n(x))$$

lo cual requiere las propiedades i), iii) y iv) de una norma, pero no necesariamente la ii). Es decir

$$e(f) = 0$$

no necesariamente implica que $f \equiv 0$. Tal medida se la conoce como una *seminorma.*

**Ejemplo 4.0.2.** *Sean* $x_0, x_1, \ldots, x_n$ *puntos de* $[a,b]$ *y* $f(x_i)$ $i = 0 : n$, *los valores de* $f$ *en dichos puntos. Entonces*

*a)* $s(f) = \sum\limits_{i=0}^{m} |f(x_i)|^2$

*b)* $s(f) = \max\limits_{1\leq i\leq m} |f(x_i)|$

*son seminormas en* $[a,b]$.

Dada una norma o seminorma en $C[a,b]$ nos preguntamos lo siguiente:

a) ¿ Existe un polinomio de un grado máximo especificado que minimice el error ?

b) Si tal polinomio existe ¿ es único ?

c) Si tal polinomio existe ¿ cómo puede determinarse ?

Veamos que para normas y seminormas distintas, las respuestas son diferentes.

## 4.1 Aproximación por polinomios de Taylor

**Problema 1:** Entre todos los polinomios de grado $\leq n$, hallar aquel que hace

$$e(f(x) - p(x)) = \sum_{k=0}^{m} |p^{(k))}(x_0) - f^{(k)}(x_0)| = \min$$

Aquí $x_0$ y $m$ son fijos. La solución está dada por el polinomio de Taylor de grado $m$, si $m \leq n$ y de grado $n$ si $m > n$:

$$P(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

con $n = m$ si $m \leq n$. El mínimo es

$$\begin{cases} 0 & \text{si} \quad n = m \\ \\ \displaystyle\sum_{k=n+1}^{m} |f^{(k)}(x_0)| & \text{si} \quad n < m \end{cases}$$

Si queremos conocer el error cometido al sustituir $f(x)$ por $P(x)$, éste está dado por

$$P(x) - f(x) = R(x) = -\frac{f^{(n+1)}(\xi_x)}{(n+1)!}(x - x_0)^{n+1}, \quad \xi_x \in int(x, x_0)$$

**Ejercicio:** Calcular el polinomio de Taylor de tercer grado que aproxima $\sqrt{x+1}$ en $x = 1$. Usando la fórmula del resto, calcular el error cometido en $x = 1$.

## 4.2 Interpolación Polinomial

Sean $x_0, x_1, \ldots, x_n$ números reales o complejos *distintos* y $n + 1$ valores $y_0, y_1, \ldots, y_n$, con $y_i = f(x_i) \;\; i = 0 : n$.

**Problema 2:** Hallar un polinomio $p_n(x)$ de grado $\leq n$ tal que

$$e(f(x) - p_n(x)) = \sum_{i=0}^{n} |f(x_i) - p_n(x_i)| = \min$$

Esto equivale a hallar un polinomio $p_n(x)$ de grado $\leq n$ tal que

$$p_n(x_i) = f(x_i) \qquad i = 0 : n$$

Escribiendo $p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$, el problema de interpolación es lo mismo que resolver el sistema lineal de $n+1$ ecuaciones con $n+1$ incógnitas

$$\begin{array}{rcl} p_n(x_0) & = & a_0 + a_1 x_0 + \cdots + a_n x_0^n \quad = \quad y_0 \\ p_n(x_1) & = & a_0 + a_1 x_1 + \cdots + a_n x_1^n \quad = \quad y_1 \\ \vdots & & \\ p_n(x_n) & = & a_0 + a_1 x_n + \cdots + a_n x_n^n \quad = \quad y_n \end{array}$$

**Teorema 4.2.1.** *Dados $n+1$ puntos distintos $x_0, x_1, \ldots, x_n$ y $n+1$ ordenadas $y_0, y_1, \ldots, y_n$ existe un único polinomio de grado $\leq n$ que interpola a $y_i$ en $x_i$, $i = 0 : n$*

**Demostración:**

Vamos a dar tres demostraciones distintas.

PRUEBA 1: La matriz de coeficientes del sistema es

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}$$

llamada matriz de Vandermonde. Sea

$$V_n(x) = det \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^n \\ 1 & x & x^2 & \cdots & x^n \end{pmatrix}$$

Por lo tanto

$$V_n(x_n) = \text{determinante de la matriz de Vandermonde}$$

Vamos a calcular $V_n(x)$ desarrollando el determinante por la última fila

$$V_n(x) = V_{n-1}(x_{n-1})\ x^n + p_{n-1}(x)$$

siendo $p_{n-1}(x)$ un polinomio de grado a lo sumo $n-1$. Resulta entonces que $V_n(x)$ es un polinomio de grado $n$ con

$$V_n(x_0) = V_n(x_1) = \cdots = V_n(x_{n-1}) = 0$$

Entonces

$$V_n(x) = V_{n-1}(x_{n-1})(x - x_0)\cdots(x - x_{n-1})$$

es decir

$$V_n(x) = V_{n-1}(x_{n-1})\prod_{j=0}^{n-1}(x - x_j)$$

Luego, $V_1(x_1) = \underbrace{V_0(x_0)}_{\text{es igual a } 1}(x_1 - x_0)$

$$V_2(x_2) = V_1(x_1)(x_2 - x_1)(x_2 - x_0) = (x_1 - x_0)(x_2 - x_1)(x_2 - x_0)$$

Es fácil ver que

$$V_n(x_n) = \prod_{n\geq i>j\geq 0}(x_i - x_j)$$

Entonces $V_n(x_n) \neq 0 \quad \Leftrightarrow \quad x_i \neq x_j \ \forall i, j$ y existe un único $p_n(x)$ que resuelve el problema 2.

PRUEBA 2: El sistema lineal se escribe

$$Xa = y$$

$Xa = y$ tiene una única solución si y solo si $Xb = 0$ tiene como única solución la trivial. Supongamos $Xb = 0$ para algún $b = (b_0, b_1, \ldots, b_n)'$. Sea $q(x) = b_0 + b_1x + \cdots + b_nx^n$. Del sistema $Xb = 0$ se tiene que $q(x_i) = 0 \quad i = 0 : n$ con grado $q \leq n$. Necesariamente

$$q(x) \equiv 0$$

y $b = 0$.

PRUEBA 3: (Constructiva) Construyamos un polinomio $l_i(x)$ de grado $\leq n$ tal que

$$l_i(x_j) = \delta_{ij} \qquad 0 \leq j \leq n$$

Luego $l_i(x) = c(x - x_0)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n)$ y $c$ debe ser tal que $l_i(x_i) = 1$, es decir

$$c = \frac{1}{\prod_{j\neq i}(x_i - x_j)}$$

En consecuencia

$$l_i(x) = \frac{\prod_{j\neq i}(x - x_j)}{\prod_{j\neq i}(x_i - x_j)} \tag{4.1}$$

Consideremos ahora

$$p(x) = y_0 l_0(x) + y_1 l_1(x) + \cdots + y_n l_n(x) \tag{4.2}$$

Luego $p(x_i) = y_i l_i(x_i) = y_i \quad i = 0 : n$ y grado $p \leq n$ dado que grado $l_i(x) \leq n \ \ \forall i = 0 : n$.

UNICIDAD: Sea $q(x)$ otro polinomio tal que $q(x_i) = y_i \ \ i = 0 : n$. Pongamos $r(x) = q(x) - p(x)$. Entonces $r(x_i) = 0 \quad i = 0 : n$, grado de $r \leq n$. Necesariamente $r(x) \equiv 0$ y $p(x) \equiv q(x)$.

□

La fórmula (4.2) se llama la *forma de Lagrange* para el polinomio de interpolación y los $l_i(x)$ se llaman los polinomios de Lagrange.

**Ejemplo 4.2.1.**

$$\begin{aligned}
p_0(x) &= y_0 \\
p_1(x) &= y_0 \frac{(x - x_1)}{(x_0 - x_1)} + y_1 \frac{(x - x_0)}{(x_1 - x_0)} \\
p_2(x) &= y_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + y_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}
\end{aligned}$$

**Ejemplo 4.2.2.** *El polinomio de grado* $\leq 2$ *que pasa por* $(0, 1), (-1, 3), (2, 1)$ *es*

$$\begin{aligned}
p(x) &= 1\frac{(x + 1)(x - 2)}{(0 + 1)(0 - 2)} + 3\frac{(x - 0)(x - 2)}{(-1 - 0)(-1 - 2)} + 1\frac{(x - 0)(x + 1)}{(2 - 0)(2 + 1)} \\
&= -\frac{1}{2}(x^2 - x - 2) + (x^2 - 2x) + \frac{1}{6}(x^+ x) \\
&= \frac{2}{3}x^2 - \frac{4}{3}x + 1
\end{aligned}$$

Si $f(x)$ es conocida en $x_i = 0 : n$, el polinomio que interpola $f$ en esos puntos es

$$p_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x)$$

**Teorema 4.2.2.** *Sean $x_0, x_1, \ldots, x_n$, $n+1$ números reales distintos y $f$ una función a valores reales $n+1$ veces continuamente diferenciable en el intervalo $I_t = int(t, x_0, x_1, \ldots, x_n)$ siendo $t$ un número real arbitrario. Entonces existe $\xi$ en $I_t$ tal que*

$$f(t) - P_n(t) = \frac{(t-x_0)\cdots(t-x_n)}{(n+1)!} f^{(n+1)}(\xi)$$

*siendo $P_n(x)$ el polinomio de interpolación de $f$ en $x_i$, $i = 0:n$*

**Demostración:**

Si $t = x_i$ para algún $i = 0:n$ el resultado es trivial. Supongamos $t \neq x_i$ $\forall i$. Pongamos

$$\begin{aligned} E(z) &= f(z) - P_n(z) \\ G(z) &= E(z) - \frac{\psi(z)}{\psi(t)} E(t) \qquad \forall\, z \in I_t \end{aligned}$$

con $\psi(z) = (z-x_0)\cdots(z-x_n)$.

$G(z)$ es $n+1$ veces continuamente diferenciable en $I_t$.

$$\text{Además} \qquad G(x_i) = E(x_i) - \frac{\psi(x_i)}{\psi(t)} E(t) = 0 \qquad i = 0:n$$

$$G(t) = E(t) - \frac{\psi(t)}{\psi(t)} E(t) = 0$$

Entonces $G$ tiene $n+2$ ceros distintos en $I_t$. Usando el teorema de Rolle generalizado, $\exists \xi \in I_t$ tal que $G^{(n+1)}(\xi) = 0$. Puesto que $E^{(n+1)}(t) = f^{(n+1)}(t)$ y $\psi^{(n+1)}(t) = (n+1)!$ se tiene

$$G^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{(n+1)!}{\psi(t)} E(t) = 0$$

por lo tanto

$$E(t) = \frac{\psi(t)}{(n+1)!} f^{(n+1)}(\xi)$$

□

**Teorema de Rolle Generalizado:** *Sea $f \in C[a,b]$, $n$ veces continuamente diferenciable en $(a,b)$. Si $f$ se anula en $n+1$ puntos distintos $x_0, \ldots, x_n$ de $[a,b]$, entonces existe un número $\xi \in (a,b)$ tal que $f^{(n)}(\xi) = 0$.*

**Ejemplo 4.2.3.** *Se desea preparar una tabla para la función* $y = e^x$, $x \in [0,1]$ *con un paso* $h$, *es decir,* $x_j - x_{j-1} = h$, *con* $x_{j-1}, x_j$ *valores de la tabla. ¿Cómo debe tomarse* $h$ *para que la interpolación lineal dé un error absoluto máximo de* $0.5\ 10^{-6}$*?*

Sea $x \in [0,1]$ y supongamos que $x \in [x_j, x_{j+1}]$. Usando la fórmula del error para la interpolación lineal

$$\begin{aligned} |f(x) - P_1(x)| &= \left| \frac{f^{(2)}(\xi)}{2!}(x - x_j)(x - x_{j+1}) \right| = \frac{e^\xi}{2!}|x - x_j||x - x_{j+1}| \\ &= \frac{e^\xi}{2!}|x - jh||x - (j+1)h| \\ &\leq \frac{1}{2} \max_{[0,1]} e^\xi \max_{x_j \leq x \leq x_{j+1}} |(x - jh)(x - (j+1)h)| \\ &= \frac{1}{2} e \max_{x_j \leq x \leq x_{j+1}} |(x - jh)(x - (j+1)h)| \end{aligned}$$

Si $q(x) = (x - jh)(x - (j+1)h)$, entonces $q(x)$ alcanza su valor mínimo en $x = \left(j + \frac{1}{2}\right) h$ y vale

$$\left(\frac{1}{2}h\right)\left(-\frac{1}{2}h\right) = -\frac{h^2}{4},$$

es decir, $\max\limits_{x_{j-1} \leq x \leq x_j} |q(x)| = \dfrac{h^2}{4}$.

En consecuencia el error de interpolación está acotado por

$$|f(x) - P_1(x)| \leq e\frac{h^2}{8}$$

Queremos que el error sea menor o igual que $0.5 \cdot 10^{-6}$; es suficiente elegir $h$ tal que $e\frac{h^2}{8} \leq 0.5 \cdot 10^{-6} \quad \Rightarrow \quad h < 0.00141$. Entonces es natural tomar $h = 0.001$.

## Análisis del error de redondeo en la interpolación lineal

Sean $f(x_0) = f_0 + \varepsilon_0, f(x_1) = f_1 + \varepsilon_1$, con $f_0, f_1$ elementos de una tabla y $\varepsilon_0, \varepsilon_1$ errores de redondeo. Supongamos que

$$\max\{|\varepsilon_0||\varepsilon_1|\} \leq \varepsilon$$

para $\varepsilon$ conocido. En una tabla de seis cifras, $\varepsilon = 5 \cdot 10^{-7}$.

Queremos acotar

$$\varepsilon(x) = f(x) - \frac{(x_1 - x)f_0 + (x - x_0)f_1}{x_1 - x_0} \qquad x_0 \leq x \leq x_1$$

usando

$$\begin{aligned} f_i &= f(x_i) - \varepsilon_i \quad i = 0, 1 \\ \varepsilon(x) &= f(x) - \frac{(x_1 - x)f(x_0) + (x - x_0)f(x_1)}{x_1 - x_0} + \frac{(x_1 - x)\varepsilon_0 + (x - x_0)\varepsilon_1}{x_1 - x_0} = \\ &= E(x) + R(x) \end{aligned}$$

$E(x)$ es el error teórico de interpolación y $R(x)$ el error de propagación de los errores de redondeo. Como $R(x)$ es una recta

$$\max_{[x_0, x_1]} |R(x)| = \max\{|\varepsilon_0|, |\varepsilon_1|\} \leq \varepsilon$$

con $x_1 = x_0 + h$, $x_0 \leq x \leq x_1$. Por lo tanto

$$|\varepsilon(x)| \leq \frac{h^2}{8} \max_{[x_0, x_1]} |f''(t)| + \max\{|\varepsilon_0|, |\varepsilon_1|\}$$

Para el caso del ejemplo anterior

$$|\varepsilon(x)| \leq 0.34 \cdot 10^{-6} + 5 \cdot 10^{-6} \approx 5.34 \cdot 10^{-6}$$

## 4.3 Diferencias Divididas. Forma de Newton para el polinomio de interpolación

Hay otras formas más apropiadas que la de Lagrange, para el polinomio de interpolación. La desventaja de la fórmula de Lagrange es que no se puede pasar de un polinomio de menor grado a otro de mayor grado utilizando la información que ya se tiene. Es decir, si le agregamos un nuevo punto al conjunto $\{x_i\}_{i=0}^n$, debemos calcular todo nuevamente. Sea $p_{n-1}$ el polinomio de interpolación en $x_0, \ldots, x_{n-1}$ y $p_n$ el polinomio de interpolación en $x_0, \ldots, x_{n-1}, x_n$. Quisiéramos tener una relación así:

$$p_n(x) = p_{n-1}(x) + c(x)$$

en donde $c(x)$ representa una corrección. En general $c(x)$ es un polinomio de grado $n$, ya que usualmente grado$(p_{n-1}) = n - 1$ y grado$(p_n) = n$. Sabemos también que

$$c(x_i) = p_n(x_i) - p_{n-1}(x_i) = 0 \quad \text{para } i = 0 : n - 1$$

por lo tanto $c(x) = a_n(x - x_0)\cdots(x - x_{n-1})$.

Siendo $p_n(x_n) = f(x_n) = y_n$

$$f(x_n) = p_{n-1}(x_n) + a_n \prod_{j=0}^{n-1}(x_n - x_j),$$

de donde

$$a_n = \frac{f(x_n) - p_{n-1}(x_n)}{\prod_{j=0}^{n-1}(x_n - x_j)}$$

Por razones que ya explicaremos, $a_n$ se llama la *diferencia dividida de orden n* de $f$ y se escribe

$$a_n = f[x_0, x_1, \ldots, x_n]$$

por lo tanto

$$p_n(x) = p_{n-1}(x) + f[x_0, x_1, \ldots, x_n](x - x_0)\cdots(x - x_{n-1})$$

### 4.3.1 Propiedades

1) $f[x_0, x_1, \ldots, x_n] = f[x_{i_0}, x_{i_1}, \ldots, x_{i_n}]$ siendo $(i_0, i_1, \ldots, i_n)$ una permutación de $(0, 1, 2, \ldots, n)$.

2) $f[x_0, x_1, \ldots, x_n] = \dfrac{f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}$

**Prueba:**

1) Sea $\psi_n(x) = (x - x_0)\cdots(x - x_n)$, luego

$$\psi_n'(x_i) = (x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)$$

Como

$$p_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i) = \sum_{i=0}^{n} \frac{\prod_{j\neq i}(x - x_j)}{\prod_{j\neq i}(x_i - x_j)} f(x_i),$$

se tiene

$$p_n(x) = \sum_{i=0}^{n} \frac{\psi_n(x)}{\psi_n'(x_i)} \frac{f(x_i)}{(x - x_i)}$$

Resulta entonces que:

$$\text{coef.}(p_n(x)) = a_n = f[x_0, x_1, \ldots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{\psi_n'(x_i)} = \sum_{j=0}^{n} \frac{f(x_{i_j})}{\psi_n'(x_{i_j})}$$

con $(i_0, \ldots, i_n)$ permutación de $(0, 1, \ldots, n)$, mostrándose que

$$f[x_0, x_1, \ldots, x_n] = f[x_{i_0}, x_{i_1}, \ldots, x_{i_n}]$$

2) Para $n = 1$ es cierto, pues

$$f[x_0, x_1] = \sum_{i=0}^{1} \frac{f(x_i)}{\psi_1'(x_i)} = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0} = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

Supongamos que la fórmula es válida para $n$ puntos; siendo

$$f[x_1, \ldots, x_n] = \sum_{i=1}^{n} \frac{f(x_i)}{(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}$$

y

$$f[x_0, \ldots, x_{n-1}] = \sum_{i=0}^{n-1} \frac{f(x_i)}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_{n-1})}$$

resulta

$$\frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0} = \frac{1}{x_n - x_0} \left\{ - \frac{f(x_0)}{(x_0 - x_1) \cdots (x_0 - x_{n-1})} + \right.$$

$$+ \sum_{i=1}^{n-1} \frac{f(x_i) \left( \frac{1}{x_i - x_n} - \frac{1}{x_i - x_0} \right)}{(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_{n-1})} +$$

$$\left. + \frac{f(x_n)}{(x_n - x_0) \cdots (x_n - x_{n-1})} \right\} =$$

$$= \frac{f(x_0)}{(x_0 - x_1) \cdots (x_0 - x_n)} + \sum_{i=1}^{n-1} \frac{f(x_i)}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} +$$

$$+ \frac{f(x_n)}{(x_n - x_0) \cdots (x_n - x_{n-1})} = \sum_{i=0}^{n} \frac{f(x_i)}{\prod_{j \neq i} (x_i - x_j)} = f[x_0, x_1, \ldots, x_n]$$

□

Regresemos a la fórmula; se tiene

$$\begin{aligned} p_0(x) &= f(x_0) \\ p_1(x) &= f(x_0) + (x - x_0)f[x_0, x_1] \\ p_2(x) &= f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \end{aligned}$$

y así siguiendo

$$\begin{aligned} p_n(x) \quad = \quad & f(x_0) + (x - x_0)f[x_0, x_n] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + \\ & \cdots + (x - x_0)\cdots(x - x_{n-1})f[x_0, x_1, \ldots, x_n] \end{aligned} \tag{4.3}$$

Esta se conoce como la *forma de Newton* de diferencias divididas para el polinomio de interpolación.

**Tabla de diferencias divididas de $f(x)$.** Si una tabla de valores $(x_i, f(x_i))$ es dada, a partir de ellos podemos construir una tabla de diferencias divididas

| $x_i$ | $f(x_i)$ | $f[x_i, x_{i+1}]$ | $f[x_i, x_{i+1}, x_{i+2}]$ | $f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$ |
|---|---|---|---|---|
| $x_0$ | $f_0$ | $f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0}$ | | |
| $x_1$ | $f_1$ | $f[x_1, x_2] = \frac{f_2 - f_1}{x_2 - x_1}$ | $f[x_0, x_1, x_2]$ | $f[x_0, x_1, x_2, x_3]$ |
| $x_2$ | $f_2$ | $f[x_3, x_2] = \frac{f_3 - f_2}{x_3 - x_2}$ | $f[x_1, x_2, x_3]$ | $f[x_1, x_2, x_3, x_4]$ |
| $x_3$ | $f_3$ | $f[x_3, x_4] = \frac{f_4 - f_3}{x_4 - x_3}$ | $f[x_2, x_3, x_4]$ | $f[x_2, x_3, x_4, x_5]$ |
| $x_4$ | $f_4$ | $f[x_4, x_5] = \frac{f_5 - f_4}{x_5 - x_4}$ | $f[x_3, x_4, x_5]$ | |
| $x_5$ | $f_5$ | | | |
| $\vdots$ | $\vdots$ | | | |

**Ejemplo 4.3.1.**

| $x$ | $f$ | $f[\cdot,\cdot]$ | $f[\cdot,\cdot,\cdot]$ | $f[\cdot,\cdot,\cdot,\cdot]$ | $f[\cdot,\cdot,\cdot,\cdot,\cdot]$ |
|---|---|---|---|---|---|
| *3.0* | *1* | | | | |
| | | *10* | | | |
| *3.1* | *2* | | *-300* | | |
| | | *-50* | | *3000* | |
| *3.2* | *-3* | | *600* | | $-\frac{70000}{4}$ |
| | | *70* | | *-4000* | |
| *3.3* | *4* | | *-600* | | |
| | | *-50* | | | |
| *3.4* | *-1* | | | | |

**Algoritmo para las diferencias divididas**: Difdiv$(d, x, n)$

A la entrada $d$ y $x$ son vectores con elementos $f(x_i)$ y $x_i$ $i = 0 : n$ respectivamente. A la salida, $d_i$ contiene $f[x_0, x_1, \ldots, x_i]$.

Paso 1: Para $i := 1 : n$
Para $j := n : -1 : i$
$d_j := (d_j - d_{j-1})/(x_j - x_{j-i})$
Paso 2: stop.

**Algoritmo para el polinomio de interpolación de Newton**: Interp$(d, x, n, t, p)$

A la entrada $d$ y $x$ son vectores conteniendo $f[x_0, x_1, \ldots, x_i]$, $x_i, i = 0 : n$, respectivamente. A la salida, $p$ contiene el valor $p_n(t)$ del polinomio de interpolación de grado $n$, que interpola $f$ en $x$.

Paso 1: $p := d_n$
Paso 2: Para $i = n - 1 : -1 : 0$
$p := d_i + (t - x_i)p$
Paso 3: stop.

### 4.3.2 Otras Propiedades de las Diferencias Divididas

**Teorema 4.3.1.** *Sea $p$ el polinomio de grado a lo sumo $n$ que interpola una función $f$ en un conjunto de $n+1$ nodos distintos, $x_0, x_1, \ldots, x_n$. Si $t$ es otro punto distinto de los nodos, entonces*

$$f(t) - p(t) = f[x_0, x_1, \ldots, x_n, t] \prod_{j=0}^{n} (t - x_j)$$

**Prueba:**

Sea $q$ el polinomio de grado a lo sumo $n+1$ que interpola $f$ en los nodos $x_0, x_1, \ldots, x_n, t$. Sabemos que $q$ se obtiene de $p$ añadiéndole un término, esto es

$$q(x) = p(x) + f[x_0, x_1, \ldots, x_n, t] \prod_{j=0}^{n} (x - x_j)$$

Como $q(t) = f(t)$, obtenemos de una vez, haciendo $x = t$,

$$f(t) = p(t) + f[x_0, x_1, \ldots, x_n, t] \prod_{j=0}^{n} (t - x_j)$$

□

**Teorema 4.3.2.** *Supongamos que $f \in C^{(n)}(I)$, $x_0, x_1, \ldots, x_n$ son números distintos en $I$, siendo*

$$I = [\min_{i=0:n} x_i, \max_{i=1:n} x_i]$$

*. Entonces existe un número $\xi$ en $I$ tal que*

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}$$

**Demostración:**

Sea $P(x)$ el polinomio de interpolación de $f$ en $x_0, x_1, \ldots, x_{n-1}$. Por el teorema anterior

$$f(x_n) - P(x_n) = f[x_0, x_1, \ldots, x_n] \prod_{j=0}^{n-1} (x_n - x_j) \tag{4.4}$$

Por el teorema del error de interpolación existe $\xi$ en $(a, b)$ tal que

$$f(x_n) - P(x_n) = \prod_{j=0}^{n-1} (x_n - x_j) \frac{f^{(n)}(\xi)}{n!} \tag{4.5}$$

Comparando las fórmulas 4.4 y 4.5 se obtiene la tesis.

□

**Corolario 4.3.1.** *Sean $P(x)$ un polinomio de grado $n$ y $x_0, \ldots, x_k$, $k+1$ puntos distintos, entonces*

$$P[x_0, x_1, \ldots, x_k] = \begin{cases} a_n & si \ \ k = n \\ 0 & si \ \ k > n \end{cases}$$

**Prueba:**

Por el teorema anterior

$$P[x_0, x_1, \ldots, x_k] = \frac{P^{(k)}(\xi)}{k!} \quad \text{con } \min(x_0, \ldots, x_k) \leq \xi \leq \max(x_0, \ldots, x_k)$$

Como $P^{(n)}(x) = a_n n!$ y $P^{(k)}(x) = 0$ si $k > n$, se tiene

$$\begin{aligned} P[x_0, x_1, \ldots, x_n] &= \frac{a_n n!}{n!} = a_n \quad \text{y} \\ P[x_0, x_1, \ldots, x_k] &= 0 \qquad\qquad \text{si } k > n \end{aligned}$$

□

Este teorema nos dice que la diferencia dividida de orden $n$ de un polinomio de grado $n$ es igual al coeficiente de mayor grado y la diferencia dividida de un orden mayor que $n$ es cero.

## Extensión de la definición de las diferencias divididas

Una fórmula de las diferencias divididas, llamada la fórmula de **Hermite-Genocchi**, se necesita en muchas situaciones, como lo veremos más adelante. Sean $x_i,\ i = 0 : n$ distintos y $f$ una función $n$ veces continuamente diferenciable, entonces

$$f[x_0, x_1, \ldots, x_n] = \int_{\tau_n} \cdots \int f^{(n)}(t_0 x_0 + \cdots + t_n x_n) dt_1 \ldots dt_n$$

siendo $\tau_n$ el **simplex** $n$-dimensional, es decir

$$\tau_n = \left\{ (t_1, \ldots, t_n)/t_1 \geq 0, \ldots t_n \geq 0 \quad \sum_{i=1}^{n} t_i \leq 1 \right\}$$

y

$$t_0 = 1 - \sum_{i=1}^{n} t_i \qquad \left( \therefore t_0 \geq 0 \quad \text{y} \quad \sum_{i=0}^{n} t_i = 1 \right)$$

La demostración de este resultado se hace por inducción y puede verse en Atkinson o Kincaid-Cheney. La fórmula también nos permite observar que $f[x_0, x_1, \ldots, x_n]$ es una función continua de las variables $x_0, x_1, \ldots, x_n$ sin

importar si son o no diferentes y por lo tanto, por ejemplo

$$f[\underbrace{x_0,\ldots,x_0}_{n+1}] = \int_{\tau_n}\cdots\int f^{(n)}(x_0)dt_1\ldots dt_n = f^{(n)}(x_0)\int\int_{\tau_n} dt_1\ldots dt_n$$

$$= f^{(n)}(x_0)\ \mathrm{Vol}(\tau_n) = \frac{f^{(n)}(x_0)}{n!} \tag{4.6}$$

Esto motiva la extensión de la definición de diferencias divididas al caso cuando alguno o todos los nodos coinciden.

**Definición 4.3.1.** *Sean $x_0 \leq x_1 \leq \cdots \leq x_n$. Entonces las diferencias divididas se definen como*

$$f[x_0,x_1,\ldots,x_n] = \begin{cases} \dfrac{f[x_1,x_2,\ldots,x_n]-f[x_0,x_1,\ldots,x_{n-1}]}{x_n-x_0} & si\ x_n \neq x_0 \\ f^{(n)}(x_0)/n! & si\ x_n = x_0 \end{cases}$$

Para entender mejor esta definición veamos el siguiente ejemplo.

**Ejemplo 4.3.2.** *Supongamos que los puntos son $x_0, x_0, x_0, x_1, x_1, x_2$*

| $x_i$ | $f(x_i)$ | $f[x_i, x_{i+1}]$ | $f[x_i, x_{i+1}, x_{i+2}]$ |
|---|---|---|---|
| $x_0$ | $f_0$ | | |
| | | $f[x_0,x_0] = f_0'$ | |
| $x_0$ | $f_0$ | | $f[x_0,x_0,x_0] = \dfrac{f_0''}{2}$ |
| | | $f[x_0,x_0] = f_0'$ | |
| $x_0$ | $f_0$ | | $f[x_0,x_0,x_1] = \dfrac{f[x_0,x_1]-f_0'}{x_1-x_0}$ |
| | | $f[x_0,x_1] = \dfrac{f(x_1)-f(x_0)}{x_1-x_0}$ | |
| $x_1$ | $f_1$ | | $f[x_0,x_1,x_1] = \dfrac{f_1'-f[x_0,x_1]}{x_1-x_0}$ |
| | | $f[x_1,x_1] = f_1'$ | |
| $x_1$ | $f_1$ | | $f[x_1,x_1,x_2] = \dfrac{f[x_1,x_2]-f_1'}{x_2-x_1}$ |
| | | $f[x_2,x_1] = \dfrac{f(x_2)-f(x_1)}{x_2-x_1}$ | |
| $x_2$ | $f_2$ | | |

Ahora podemos calcular derivadas,

$$\begin{aligned}
\frac{d}{dx}f[x_0,\ldots,x_n,x] &= \lim_{h\to 0}\frac{f[x_0,\ldots,x_n,x+h]-f[x_0,\ldots,x_n,x]}{h}\\
&= \lim_{h\to 0}\frac{f[x_0,\ldots,x_n,x+h]-f[x,x_0,\ldots,x_n]}{h}\\
&= \lim_{h\to 0} f[x,x_0,\ldots,x_n,x+h]\\
&= f[x,x_0,\ldots,x_n,x] = f[x_0,\ldots,x_n,x,x] \qquad (4.7)
\end{aligned}$$

## 4.4 Diferencias divididas con puntos igualmente espaciados

### 4.4.1 Operador de diferencias progresivas

Para $h>0$ definimos

$$\triangle f(x) = f(x+h) - f(x)$$

como la diferencia progresiva de $f$ en $x$ y a $\triangle$ le llamamos el operador de diferencias progresivas.

Las diferencias progresivas de $f$ en $x$ de orden $k$ se definen como

$$\triangle^{k+1} f(x) = \triangle^k f(x+h) - \triangle^k f(x) \quad k\geq 0$$

con $\triangle^0 f(x) = f(x)$.

Por ejemplo

$$\begin{aligned}
\triangle^2 f(x) &= \triangle f(x+h) - \triangle f(x) = f(x+2h) - f(x+h) - f(x+h) + \\
&\quad + f(x) = f(x+2h) - 2f(x+h) + f(x)\\
\triangle^3 f(x) &= f(x+3h) - 3f(x+2h) + 3f(x+h) - f(x)\\
&\vdots\\
\triangle^k f(x) &= \sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j} f(x+jh)
\end{aligned}$$

Cuando los nodos $x_0, x_1, \ldots, x_n$ son igualmente espaciados, es decir $x_i = x_0 + ih,\ i = 0:n$, se tiene

**Lema 4.4.1.** *Para* $k \geq 0$

$$f[x_0, x_1, \ldots, x_k] = \frac{1}{k!h^k} \triangle^k f_0$$

**Prueba:**

Para $k = 0$ es cierto pues

$$f[x_0] = \frac{1}{0!h^0} \triangle^0 f_0 = f(x_0)$$

$k = 1$

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{\triangle f(x_0)}{h} = \frac{\triangle f_0}{h}$$

Supongamos que la fórmula es cierta para $k \leq r$; tomemos $k = r + 1$

$$f[x_0, x_1, \ldots, x_{r+1}] = \frac{f[x_1, x_2, \ldots, x_{r+1}] - f[x_0, x_1, \ldots, x_r]}{x_{r+1} - x_0}$$

por la hipótesis inductiva

$$= \frac{\dfrac{\triangle^r f_1}{r!h^r} - \dfrac{\triangle^r f_0}{r!h^r}}{(r+1)h}$$

$$= \frac{\triangle^r f(x_0 + h) - \triangle^r f(x_0)}{(r+1)!h^{r+1}} = \frac{\triangle^{r+1} f_0}{(r+1)!h^{r+1}}$$

**Nota:** Es claro que si $p$ es un polinomio de grado $n$

$$\triangle^n p(x_0) = \text{cte.} \qquad \triangle^k p(x_0) = 0 \ \text{ si } k > n$$

Usando el lema anterior vamos a modificar la fórmula de interpolación de Newton a una fórmula que contiene diferencias progresivas en lugar de diferencias divididas.

Sea $x = x_0 + sh$, entonces $x - x_i = x_0 + sh - (x_0 + ih) = (s - i)h$.

$$\begin{aligned} P_n(x) &= P_n(x_0 + sh) = f[x_0] + shf[x_0, x_1] + s(s-1)h^2 f[x_0, x_1, x_2] + \\ &\quad \cdots + s(s-1)(s-2)\cdots(s-n+1)h^n f[x_0, x_1, \ldots, x_n] = \\ &= \sum_{k=0}^{n} s(s-1)\cdots(s-k+1)h^k f[x_0, x_1, \ldots, x_k] \end{aligned}$$

Poniendo $\binom{s}{k} = \dfrac{s(s-1)\cdots(s-k+1)}{k!}$ y siendo $f[x_0, x_1, \ldots, x_k] = \dfrac{\triangle^k f_0}{k! h^k}$ podemos expresar a $P_n(x)$ así:

$$\boxed{P_n(x) = P_n(x_0 + sh) = \sum_{k=0}^{n} \binom{s}{k} \triangle^k f_0}$$ Fórmula de Newton de diferencias progresivas

### 4.4.2 Operador de diferencias regresivas

Definamos

$$\begin{array}{rcll} \nabla f(x) & = & f(x) - f(x-h) & \\ \nabla^{(k+1)} f(x) & = & \nabla^k f(x) - \nabla^k f(x-h) & k \geq 1 \end{array}$$

Es así que:

$$\begin{array}{rcl} \nabla^2 f(x) & = & \nabla f(x) - \nabla f(x-h) = f(x) - f(x-h) - f(x-h) + f(x-2h) \\ & = & f(x) - 2f(x-h) + f(x-2h) \\ \nabla^3 f(x) & = & f(x) - 3f(x-h) + 3f(x-2h) - f(x-3h) \end{array}$$

En general se ve que

$$\nabla^k f(x) = \sum_{j=0}^{k} (-1)^j \binom{k}{j} f(x - jh)$$

Si los nodos interpolantes se reordenan así: $x_n, x_{n-1}, \ldots, x_0$, la fórmula de Newton para el polinomio de interpolación es:

$$\begin{array}{rcl} P_n(x) & = & f[x_n] + f[x_{n-1}, x_n](x - x_n) + f[x_{n-2}, x_{n-1}, x_n](x - x_n)(x - x_{n-1}) + \\ & & \cdots + f[x_0, \ldots, x_n](x - x_n)(x - x_{n-1}) \cdots (x - x_1) \end{array}$$

Usando espacios iguales con $x = x_n + sh$ y $x_i = x_n + (i - n)h$ se tiene $x - x_i = (s + n - i)h$, lo cual produce:

$$\begin{array}{rcl} P_n(x) & = & f[x_n] + f[x_{n-1}, x_n] sh + f[x_{n-2}, x_{n-1}, x_n] s(s+1)h^2 + \cdots \\ & & + f[x_0, \ldots, x_n] s(s+1) \cdots (s+n-1)h^n \end{array}$$

**Lema 4.4.2.**

$$f[x_{n-k}, \ldots, x_n] = \frac{1}{h^k k!} \nabla^k f(x_n)$$

**Prueba:** Ejercicio.

Extendiendo la notación del coeficiente binomial para incluir números negativos

$$\binom{-s}{k} = \frac{-s(-s-1)\cdots(-s-k+1)}{k!} = (-1)^k \frac{s(s+1)\cdots(s+k-1)}{k!}$$

$$\begin{aligned} P_n(x) &= f(x_n) + (-1)\binom{-s}{1}\nabla f(x_n) + (-1)^2\binom{-s}{2}\nabla^2 f(x_n) + \cdots \\ &\quad +(-1)^n\binom{-s}{n}\nabla^n f(x_n) \end{aligned}$$

es decir

$$\boxed{P_n(x) = \sum_{k=0}^{n}(-1)^k\binom{-s}{k}\nabla^k f(x_k)}$$ Fórmula de diferencias regresivas (de Newton)

Si en el ejemplo de la función de Bessel queremos calcular con los mismos datos J(5.7), es preferible usar la fórmula de diferencias regresivas, haciendo uso máximo de los puntos más cercanos a 5.7.

Cuando $x$ es un valor cerca del centro de la tabla, las fórmulas anteriores para el polinomio de interpolación no son muy apropiadas. En estas circunstancias se dispone de una gran variedad de fórmulas basadas en las llamadas *"diferencias centrales"*.

Se define, en base a nodos no tabulados,

$$\delta f(x_i) = f\left(x_i + \frac{h}{2}\right) - f\left(x_i - \frac{h}{2}\right)$$

Pero

$$\begin{aligned} \delta^2 f(x_i) &= \delta\left(\delta f(x_i)\right) = \delta(f(x_{i+1/2})) - f(x_{i-1/2}) = \\ &= f(x_{i+1}) - 2f(x_i) + f(x_{i-1}) \end{aligned}$$

Se ve que

$$\delta^{2k} f(x_i) = \delta^{2k-2} f(x_{i+1}) - 2\delta^{2k-2} f(x_i) + \delta^{2k-2} f(x_{i-1})$$

Por ejemplo si $k = 2$

$$\begin{aligned}\delta^4 f(x_i) &= \delta^2 f(x_{i+1}) - 2\delta^2 f(x_i) + \delta^2 f(x_{i-1}) = \\ &= f(x_{i+2}) - 2f(x_{i+1}) + f(x_i) - 2f(x_{i+1}) + 4f(x_i) - 2f(x_{i-1}) + \\ &\quad + f(x_i) - 2f(x_{i-1}) + f(x_{i-2}) \\ &= f(x_{i+2}) - 4f(x_{i+1}) + 6f(x_i) - 4f(x_{i-1}) + f(x_{i-2})\end{aligned}$$

La correspondiente tabla de diferencias centrales para $x_0, x_1, x_2, x_{-1}, x_{-2}, x_3$ es:

$$\begin{array}{llllll}
x_{-2} & f_{-2} & & & & \\
 & & \delta f_{-3/2} & & & \\
x_{-1} & f_{-1} & & \delta^2 f_{-1} & & \\
 & & \delta f_{-1/2} & & \delta^3 f_{-1/2} & \\
x_0 & f_0 & & \delta^2 f_0 & & \delta^4 f_0 \\
 & & \delta f_{1/2} & & \delta^3 f_{1/2} & \\
x_1 & f_1 & & \delta^2 f_1 & & \delta^4 f_1 \\
 & & \delta f_{3/2} & & \delta^3 f_{3/2} & \\
x_2 & f_2 & & \delta^2 f_2 & & \\
 & & \delta f_{5/2} & & & \\
x_3 & f_3 & & & &
\end{array}$$

**Fórmula de Everett para el polinomio de interpolación**

Se toma un número par de nodos y se pone $n = 2m + 1$, donde

$$x_{-m} < \cdots < x_{-1} < x_0 < x_1 < \cdots < x_m < x_{m+1}.$$

Denotando $s = \dfrac{x - x_0}{h}$, $t = 1 - s$. Se tiene

$$\begin{aligned}P_n(x) &= tf_0 + \frac{t(t^2-1)}{3!}\delta^2 f_0 + \cdots + \frac{t(t^2-1)(t^2-4)\cdots(t^2-m^2)}{(2m+1)!}\delta^{2m} f_0 + \\ &\quad sf_1 + \frac{s(s^2-1)}{3!}\delta^2 f_1 + \cdots + \frac{s(s^2-1)(s^2-4)\cdots(s^2-m^2)}{(2m+1)!}\delta^{2m} f_1\end{aligned}$$

Observar que si $m = 1$, se obtiene un polinomio cúbico.

## Interpolación equidistante y el fenómeno de Runge

Sea $f(x)$ continua en $[a, b]$, $x_0, x_1, \ldots, x_n \in [a, b]$. Si la sucesión de puntos $\left\{(x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)})\right\}$ cubre $[a, b]$,¿es posible decir que la sucesión de poli-

nomios de interpolación $\{P_n(x)\}$ converge a $f(x)$ en $[a,b]$ uniformemente?. Es decir, es cierto que

$$\|f(x) - P_n(x)\|_\infty = \max_{[a,b]} |f(x) - P_n(x)| \xrightarrow[n\to\infty]{} 0$$

Esta pregunta no está completamente resuelta. Sólo hay respuestas parciales.

**Primer resultado:** Interpolación con nodos igualmente espaciados.

Aquí $x_0^{(n)} = a \;\; x_j^{(n)} = x_0^{(n)} + jh_n \quad h_n = \dfrac{b-a}{n} \quad j = 0 : n$ y la respuesta es negativa. El ejemplo lo da la función

$$g(x) = \frac{1}{1+25x^2} \quad \text{en } [-1,1] \quad \text{con } x_i = -1 + \frac{2i}{n} \quad i = 0 : n$$

Runge (1901) descubrió que cuando el grado del polinomio $n \to \infty$, $P_n(x)$ diverge en los intervalos $0.726 \leq |x| \leq 1$. Este fenómeno se conoce como el fenómeno de Runge (ver Isaacson-Keller). Las figuras 4.4.2 y 4.4.2 ilustran esta situación en lo casos $n = 5$ y $n = 10$. Notar que la interpolación es buena en la porción central del intervalo. Tal comportamiento es típico de la interpolación equidistante con polinomios de alto grado y puede ser explicado teóricamente. (Isaacson-Keller, Analysis of Numerical Methods, página 275).

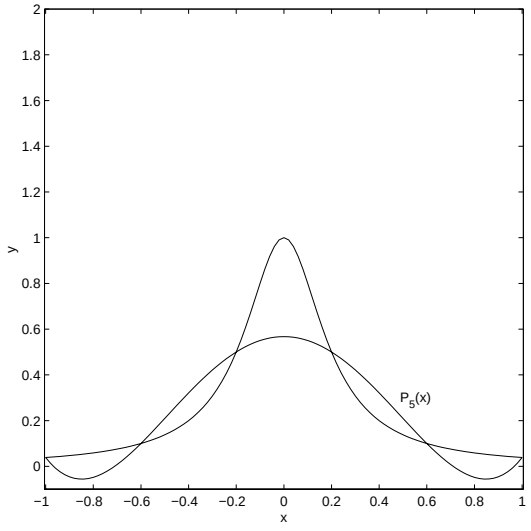



Figura 4.1: La función de Runge y el polinomio de interpolación de grado 5 con nodos $x_i = -1 + 2i/5, \;\; i = 0 : 5$

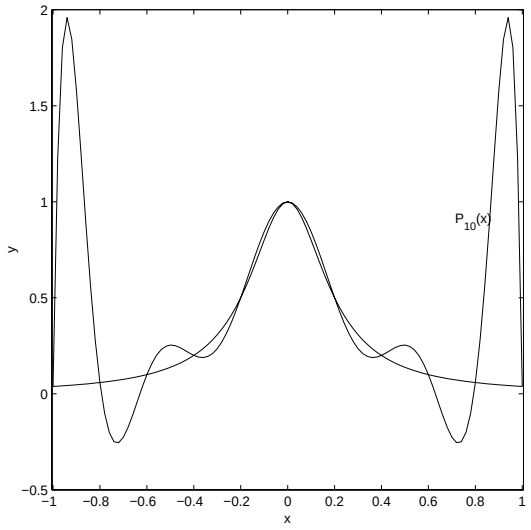



Figura 4.2: La función de Runge y el polinomio de interpolación de grado 10 con nodos $x_i = -1 + 2i/10, \quad i = 0:10$

**Segundo resultado:** Interpolación con abscisas de Tchebycheff.

El ejemplo de Runge responde en parte a la pregunta si los polinomios de interpolación $P_n$ convergen a $f(x)$ en $[a,b]$ si $x_0, x_1, \ldots, x_n$ cubren $[a,b]$. Una respuesta afirmativa se obtiene en el caso en que $f$ y sus derivadas primera y segunda son continuas en $[-1,1]$ y los puntos $x_j$ son exactamente:

$$x_j = \cos\left(\frac{2j+1}{n+1}\frac{\pi}{2}\right) \qquad j = 0:n$$

y $P_n(x)$ es el polinomio de interpolación de grado $n$ de $f$ en $[-1,1]$. Entonces

$$|P_n(x) - f(x)| \leq O\left(\frac{1}{\sqrt{n}}\right)$$

Los nodos $x_j$ se llaman las *abscisas de Chebyshev*, esto es, son los ceros del polinomio de Chebyshev de grado $n+1$.

En el caso de la función de Runge, si usáramos estos nodos obtendríamos una mejor aproximación, pero aún así esta apoximación no resulta muy buena. Este tipo de función se presta más a una aproximación por polinomios a trozos (splines).

Con datos equidistantes se puede obtener una mejor aproximación con el método de mínimos cuadrados ajustando un polinomio de menor grado (en el caso de Runge, $n = 6$).

**Tercer resultado (Faber, 1914):**

Para todo conjunto de nodos prescrito

$$a \leq x_0^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)} \leq b$$

existe una función $f \ \varepsilon \ C[a,b]$ tal que los polinomios de interpolación de $f$ usando estos nodos no converge uniformemente a $f$.

**Cuarto resultado**

Si $f$ es una función continua en $[a,b]$, entonces existe un sistema de nodos

$$a \leq x_0^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)} \leq b$$

tales que los polinomios de interpolación $P_n$ de $f$ en dichos nodos verifican

$$\|f(x) - P_n(x)\|_\infty = \max_{[a,b]} |f(x) - P_n(x)| \xrightarrow[n\to\infty]{} 0$$

## 4.5 Interpolación y nodos de Chebyshev

Los polinomios de Chebyshev se definen por recurrencia

$$\begin{array}{ll} T_0(x) = 1 & T_1(x) = x \\ T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) & n \geq 1 \end{array}$$

**Teorema 4.5.1.**

$$T_n(x) = \cos(n \ \arccos(x)), \ \ n \geq 0 \ \ x\varepsilon[-1,1]$$

**Prueba:**

Siendo

$$\begin{array}{rcl} \cos(n+1)\theta & = & \cos\theta\cos n\theta - \sin\theta\sin n\theta \\ \cos(n-1)\theta & = & \cos\theta\cos n\theta + \sin\theta\sin n\theta \end{array}$$

resulta

$$\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta$$

Definamos $f_n = \cos(n \ \arccos(x))$ con $x = \cos\theta$. Entonces se tiene

$$\begin{array}{ll} f_0(x) = 1 & f_1(x) = x \\ f_{n+1}(x) = 2xf_n(x) - f_{n-1}(x) & n \geq 1 \end{array}$$

Por lo tanto $f_n(x) = T_n(x)$ para $n \geq 0$ □

Veamos algunas propiedades

1. $|T_n(x)| \leq 1$ para $-1 \leq x \leq 1$
2. $T_n(\cos \frac{j\pi}{n}) = (-1)^j$ para $j = 0 : n$
3. $T_n(\cos \frac{2j-1}{2n}\pi) = 0$ para $j = 1 : n$
4. El coeficiente de mayor grado de $T_n(x)$ es $2^{n-1}$ y por lo tanto $\frac{T_n(x)}{2^{n-1}}$ es mónico.

**Teorema 4.5.2.** *Si $p$ es un polinomio mónico de grado $n$, entonces*

$$||p||_\infty = \max_{-1 \leq x \leq 1} |p(x)| \geq ||\frac{T_n(x)}{2^{n-1}}|| = \frac{1}{2^{n-1}}$$

**Prueba:**

El teorema afirma que entre todos los polinomios mónicos de grado $n$, $\frac{T_n(x)}{2^{n-1}}$ es el que tiene la norma mínima. La prueba es por el absurdo. Supongamos que existe un polinomio $p$ mónico de grado $n$ tal que

$$|p(x)| < \frac{1}{2^{n-1}} \qquad |x| \leq 1$$

Sea $q(x) = \frac{T_n(x)}{2^{n-1}}$ y $x_i = cos\frac{i\pi}{n}$ para $i = 0 : n$. Entonces

$$(-1)^i p(x_i) \leq |p(x_i)| < \frac{1}{2^{n-1}} = (-1)^i q(x_i)$$

En consecuencia

$$(-1)^i [q(x_i) - p(x_i)] > 0 \qquad i = 0 : n$$

Luego $q - p$ oscila $n + 1$ veces en el intervalo $[-1, 1]$ y por lo tanto tiene $n$ raíces. Pero siendo $q$ y $p$ mónicos, $q - p$ es un polinomio de grado $n - 1$. Absurdo. □

**Error de interpolación: eligiendo los nodos**

Supongamos que $[a, b] = [-1, 1]$. Si $x, \xi \; \varepsilon \; [-1, 1]$, siendo

$$|f(x) - p_n(x)| = \frac{|\prod_{i=0}^{n}(x - x_i)|}{(n+1)!} |f^{(n+1)}(\xi)|$$

resulta

$$\max_{[-1,1]} |f(x) - p_n(x)| \leq \frac{1}{(n+1)!} \max_{[-1,1]} |\prod_{i=0}^{n}(x - x_i)| \max_{[-1,1]} |f^{(n+1)}(x)|$$

Por el teorema anterior $\max\limits_{[-1,1]} |\prod\limits_{i=0}^{n}(x - x_i)| \geq 2^{-n}$, y $2^{-n}$ se alcanza si $x_i$ son los ceros del polinomio $T_{n+1}(x)$, es decir

$$x_i = \cos \frac{2i+1}{2n+2}\pi \qquad i = 0 : n$$

Luego

$$\max_{[-1,1]} |f(x) - p_n(x)| \leq \frac{1}{2^n(n+1)!} \max_{[-1,1]} |f^{(n+1)}(x)|$$

esto es

$$||f(x) - p_n(x)||_\infty \leq \frac{1}{2^n(n+1)!} ||f^{(n+1)}(x)||_\infty$$

□

## 4.6 Funciones Splines

Las funciones splines o trazadores son funciones polinómicas a trozos diferenciables hasta un cierto orden en un intervalo dado.

**Definición 4.6.1.** *Sea $a = x_1 < x_2 < \cdots < x_m = b$ una subdivisión de $[a, b]$. Una función spline o trazador de grado $p$ con nodos en los puntos $a = x_1 < x_2 < \cdots < x_m = b$ es una función $s(x)$ con las siguientes propiedades:*

1. *sobre cada subintervalo $[x_i x_{i+1}]$, $i = 1 : m-1$, $s(x)$ es un polinomio de grado $p$.*
2. *$s(x)$ y sus primeras $(p-1)$ derivadas son continuas en $[a, b]$*

Ejemplos:

1. $p = 0$, la función $s(x)$ es constante a trozos.
2. $p = 1$, entonces $s(x)$ es una función continua y lineal a trozos.
3. $p = 3$, entonces $s(x)$ es un polinomio cúbico sobre cada sub-intervalo $[x_i, x_{i+1}]$ con $s(x)$, $s'(x)$, $s''(x)$ continuas.

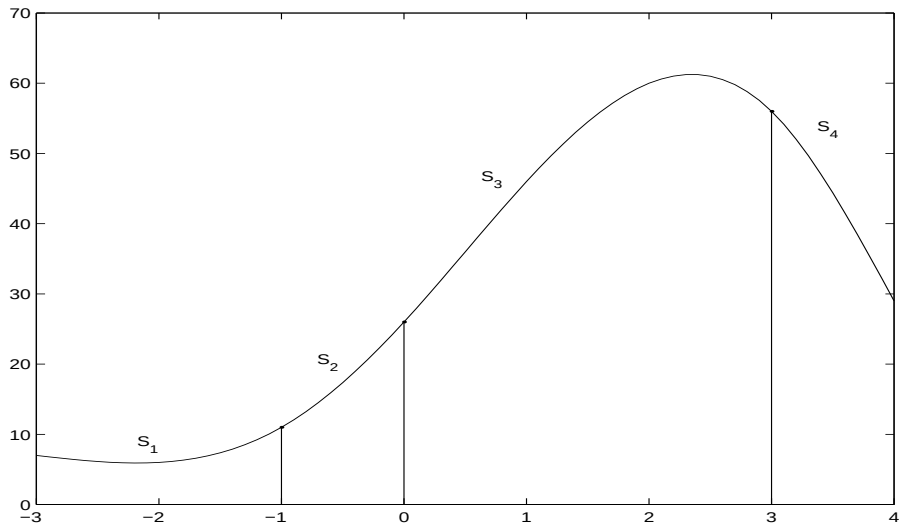



Figura 4.3: Una función spline cúbica

Observar que la derivada de una spline de grado $p$ es otra spline de grado $p-1$. Las funciones splines cúbicas son las más usadas entre las splines, debido a que son suficientemente suaves para ajustar datos y no tienen el comportamiento oscilatorio que caracteriza a la interpolación con polinomios de alto grado.

## Interpolación con Splines

Sean $a = x_1 < x_2 < \cdots < x_m = b$, $y(x_i) = y_i$, $i = 1 : m$. Una spline de interpolación o un trazador de interpolación es una función spline o trazador $s(x)$ tal que

$$s(x_i) = y_i, \qquad i = 1 : m$$

**Interpolación con spline lineal**: el problema de interpolación con splines lineales tiene una única solución, ya que toda spline lineal $s(x)$ puede expresarse así:

$$s(x) = \sum_{i=1}^{m} y_i \Phi_i(x)$$

con

$$\Phi_i(x) = \begin{cases} \frac{x - x_i}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i] \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, & x \in [x_i, x_{i+1}] \\ 0, & \text{si no} \end{cases}$$

y $\Phi_i(x)$ linealmente independientes.

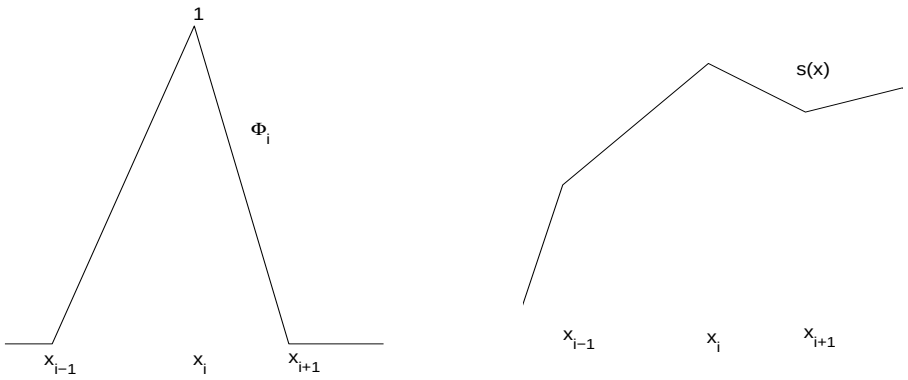



Figura 4.4: La función $\Phi_i$ y la spline lineal $s(x)$

**Interpolación con spline cúbica**. En este caso el problema tiene infinitas soluciones. En efecto si

$$s(x) = a_i + b_i x + c_i x^2 + d_i x^3$$

en $[x_i, x_{i+1}]$ para $i = 1 : m-1$, entonces hay $4(m-1) = 4m-4$ coeficientes $a_i, b_i, c_i, d_i$ a determinar, con las condiciones $s(x_i) = y_i$, $i = 1 : m$, $s^{(j)}(x_{i+0}) = s^{(j)}(x_{i-0})$, para $i = 2 : m-1$ y $j = 0, 1, 2$, lo que da un total de $m+3(m-2) = 4m-6$ condiciones. Luego hay dos grados de libertad en la elección de $s(x)$.

**Teorema 4.6.1.** *Toda función spline cúbica $s(x)$ de interpolación es, para $x \in [x_i, x_{i+1})$, un polinomio cúbico de interpolación de la forma*

$$q_i(x) = t y_{i+1} + \bar{t} y_i + h_i^2[(t^3 - t)k_{i+1} + (\bar{t}^3 - \bar{t})k_i]$$

*para $i = 1 : m-1$, en donde $h_i = x_{i+1} - x_i$, $t = \dfrac{x - x_i}{h_i}$, $\bar{t} = 1 - t$ y $k_1, k_2, \ldots, k_m$ satisfacen el sistema tridiagonal de ecuaciones*

$$h_{i-1}k_{i-1} + 2(h_{i-1} + h_i)k_i + h_i k_{i+1} = d_i - d_{i-1}$$

*con $i = 2 : m-1$ y $d_i = \dfrac{y_{i+1} - yi}{h_i}$.*

Como este sistema tiene $m$ incógnitas y $m-2$ ecuaciones, se necesitan dos condiciones adicionales para que la función spline $s(x)$ quede unívocamente determinada.

**Demostración:**

Notar que si $x = x_i$ entonces $t = 0$ y $\bar{t} = 1$; si $x \to x_{i+1}$ entonces $t \to 1$ y $\bar{t} \to 0$.

Es fácil ver que $q_i(x_i) = y_i$ y que $q_i(x) \to y_{i+1}$ cuando $x \to x_i$; esto dice también que $q_{i-1}(x) \to y_i$ cuando $x \to x_i$ y la continuidad de $s(x)$ se verifica. Además

$$q_i'(x) = \frac{y_{i+1}}{h_i} - \frac{y_i}{h_i} + h_i[(3t^2 - 1)k_{i+1} - (3\bar{t}^2 - 1)k_i] \quad \text{en} \quad [x_i, x_{i+1})$$

y por lo tanto

$$q_i'(x_i) = \frac{y_{i+1} - y_i}{h_i} - h_i(k_{i+1} + 2k_i) = d_i - h_i(k_{i+1} + 2k_i) \quad \text{siendo} \quad d_i = \frac{y_{i+1} - y_i}{h_i}$$

Por otro lado

$$q_i'(x_{i+1}) = \frac{y_{i+1} - y_i}{h_i} + h_i(2k_{i+1} + k_i) = d_i + h_i(2k_{i+1} + k_i)$$

y entonces para $i = i - 1$

$$q_{i-1}'(x_i) = d_{i-1} + h_{i-1}(2k_i + k_{i-1})$$

Pedir que $q'(x)$ sea continua implica que

$$q_{i-1}'(x_i) = q_i'(x_i) \qquad i = 2 : m - 1$$

esto es

$$d_{i-1} + h_{i-1}(2k_i + k_{i-1}) = d_i - h_i(k_{i+1} + 2k_i)$$

es decir

$$h_{i-1}k_{i-1} + 2(h_{i-1} + h_i)k_i + h_i k_{i+1} = d_i - d_{i-1} \quad \text{para} \quad i = 2 : m - 1$$

Por último

$$q_i''(x) = 6tk_{i+1} + 6\bar{t}k_i$$

lo que implica

$$q_i''(x_i) = 6k_i \qquad q_i''(x_{i+1}) = 6k_{i+1} \qquad q_{i-1}''(x_i) = 6k_i$$

y por lo tanto

$$q''_{i-1}(x_i) = q''_i(x_i)$$

De esta forma la continuidad de la derivada segunda es satisfecha. □

Faltan entonces las condiciones adicionales mencionadas en el teorema.

**Spline natural.** Este es el caso cuando se imponen las condiciones

$$k_1 = k_m = 0$$

es decir que $s''(x_1) = s''(x_m) = 0$. Con esta elección el sistema de ecuaciones se expresa así

$$\begin{pmatrix} 2(h_2+h_1) & h_1 & 0 & \dots & \dots & 0 \\ h_2 & 2(h_3+h_2) & h_3 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & h_{m-3} & 2(h_{m-3}+h_{m-2}) & h_{m-2} \\ 0 & \dots & \dots & 0 & h_{m-2} & 2(h_{m-2}+h_{m-1}) \end{pmatrix}$$

$$\begin{pmatrix} k_2 \\ k_3 \\ k_4 \\ \vdots \\ k_{m-1} \end{pmatrix} = \begin{pmatrix} d_2 - d_1 \\ d_3 - d_2 \\ d_4 - d_3 \\ \vdots \\ d_{m-1} - d_{m-2} \end{pmatrix}$$

Este sistema de ecuaciones cuya matriz es tridiagonal y simétrica es fuertemente diagonal dominante. Por lo tanto es inversible y la descomposición $LU$ puede hacerse sin intercambios de filas.

**Ejemplo 4.6.1.** *Construir la spline cúbica natural que interpola los valores (-3,7), (-1,11), (0,26), (3,56) y (4,29).*

Procedemos primero a calcular $h_i$ y $d_i$ para montar el sistema de ecuaciones. Se tiene $h_1 = 2,\ h_2 = 1,\ h_3 = 3,\ h_4 = 1$ y $d_1 = 2,\ d_2 = 15,\ d_3 = 10,\ d_4 = -27$. Como $k_1 = k_5 = 0$ el sistema queda

$$\begin{pmatrix} 6 & 1 & 0 \\ 1 & 8 & 3 \\ 0 & 3 & 8 \end{pmatrix} \begin{pmatrix} k_2 \\ k_3 \\ k_4 \end{pmatrix} = \begin{pmatrix} 13 \\ -15 \\ -37 \end{pmatrix}$$

de donde $k_2 = 2,\ k_3 = 1,\ k_4 = -5$ y la spline resulta ser

$$s(x) = \begin{cases} -2x+1 & \text{en} \quad (-\infty, -3] \\ x^3+9x^2+25x+28 & \text{en} \quad [-3,-1] \\ -x^3+3x^2+19x+26 & \text{en} \quad [-1,0] \\ -2x^3+3x^2+19x+26 & \text{en} \quad [0,3] \\ 5x^3-60x^2+208x-163 & \text{en} \quad [3,4] \\ 32x+157 & \text{en} \quad [4,\infty) \end{cases}$$

**Teorema 4.6.2.** *Sea $f$ en $C^{(2)}[a,b]$ y $a = x_1 < \cdots < x_m = b$. Si $s(x)$ es la spline cúbica natural que interpola $f$ en los nodos dados entonces*

$$\int_a^b [s''(x)]^2 dx \leq \int_a^b [f''(x)]^2 dx$$

**Demostración:** Sea $g = f - s$. Entonces $g(x_i) = 0, \quad i = 1:m$ y

$$\int_a^b [f''(x)]^2 dx = \int_a^b [s''(x)]^2 dx + 2\int_a^b s''(x)g''(x)dx + \int_a^b [g''(x)]^2 dx$$

Sólo hay que probar que $\int_a^b s''(x)g''(x)dx \geq 0$. Dividiendo el intervalo de integración, integrando por partes y usando las condiciones de la spline natural se tiene,

$$\int_a^b s''(x)g''(x)dx = \sum_{i=1}^{m-1} \int_{x_i}^{x_{i+1}} s''(x)g''(x)dx =$$

$$= \sum_{i=1}^{m-1} (s''(x_{i+1})g'(x_{i+1}) - s''(x_i)g'(x_i)) - \int_{x_i}^{x_{i+1}} s'''(x)g'(x)dx =$$

$$= -\sum_{i=1}^{m-1} c_i(g(x_{i+1}) - g(x_i)) = 0$$

□

Recordemos que la curvatura de una curva descrita por $y = f(x)$ es la cantidad

$$\frac{|f''(x)|}{(1+[f'(x)]^2)^{\frac{3}{2}}}$$

Podemos ver a $|f''(x)|$ como una aproximación a la curvatura. En la interpolación con spline cúbica natural estamos hallando una curva con curvatura minimal.

**Teorema 4.6.3.** *Sea*

$$h = \min_{i=1:m-1} h_i$$

*con* $h_i = x_{i+1} - x_i$ *y supongamos* $f$ *en* $C^{(4)}[a,b]$ *con* $x_1 = a$ *y* $x_m = b$ *y* $f''(a) = f''(b) = 0$. *Sea* $s(x)$ *la spline cúbica de interpolación de* $f$ *en los nodos dados, entonces existe* $M > 0$ *tal que*

$$\max_{[a,b]} |f^{(j)}(x) - s^{(j)}(x)| \leq M\ h^{4-j}, \qquad j = 0, 1, 2$$

**Condiciones en las derivadas de la frontera** En este caso se pide

$$s'(x_1) = y'_1 \qquad s'(x_m) = y'_m$$

lo cual da lugar al siguiente sistema:

$$\begin{pmatrix} 2h_1 & h_1 & 0 & \dots & \dots & 0 \\ h_1 & 2(h_1+h_2) & h_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & h_{m-2} & 2(h_{m-2}+h_{m-1}) & h_{m-1} \\ 0 & \dots & \dots & 0 & h_{m-1} & 2h_{m-1} \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ \vdots \\ k_m \end{pmatrix}$$

$$= \begin{pmatrix} d_1 - y'_1 \\ d_2 - d_1 \\ d_3 - d_2 \\ \vdots \\ y'_m - d_{m-1} \end{pmatrix}$$

Observar que esta matriz tridiagonal es también fuertemente diagonal dominante y por lo tanto es no singular y la descomposición $LU$ puede hacerse sin intercambios de filas.

**Teorema 4.6.4.** *Sea*

$$h = \min_{i=1:m-1} h_i$$

*con* $h_i = x_{i+1} - x_i$ *y supongamos* $f$ *en* $C^{(2)}[a,b]$ *con* $x_1 = a$ *y* $x_m = b$ *y* $f'(a) = y'_1 f'(b) = y'_m$. *Sea* $s(x)$ *la spline cúbica de interpolación de* $f$ *en los nodos dados, entonces existe* $M > 0$ *tal que*

$$\max_{[a,b]} |f^{(j)}(x) - s^{(j)}(x)| \leq M\ h^{4-j}, \qquad j = 0, 1, 2$$

**Condiciones de frontera periódicas** En este caso se pide

$$s'(x_1) = s'(x_m) \qquad\qquad s''(x_1) = s''(x_m)$$

de lo cual resulta un sistema de ecuaciones lineales pero con una matriz que deja de ser tridiagonal. Queda como ejercicio la construcción de dicha matriz.

Dependiendo del tipo de aplicación, se podría desear guardar los $k_i$ y usarlos directamente para evaluar la spline. Si la spline va a ser evaluada muchas veces, es mejor reordenar los términos. Si expresamos la spline como

$$s(x) = y_i + \beta_i(x - x_i) + \gamma_i(x - x_i)^2 + \delta_i(x - x_i)^3$$

en $[x_i, x_{i+1}]$ con $i = 1 : m-1$, es preferible calcular y guardar los coeficientes $\beta_i, \gamma_i$ y $\delta_i$.

Estos coeficientes verifican $s(x_i) = y_i$, $s'(x_i) = \beta_i$, $s''(x_i) = 2\gamma_i$ y $s'''(x_i) = 6\delta_i$. Recordando que

$$\begin{array}{rcccl} s'(x_i) & = & q'(x_i) & = & d_i - h_i(k_{i+1} + 2k_i) \\ s''(x_i) & = & q''(x_i) & = & 6k_i \\ s'''(x_i) & = & q'''(x_i) & = & 6\dfrac{k_{i+1} - k_i}{h_i} \end{array}$$

resulta

$$\begin{array}{rcl} \beta_i & = & d_i - h_i(k_{i+1} + 2k_i) \\ \gamma_i & = & 6k_i \\ \delta_i & = & 6\dfrac{k_{i+1} - k_i}{h_i} \end{array}$$

para $i = 1 : m - 1$.

# Capítulo 5

# Integración Numérica

## 5.0 Cuadratura Numérica

El problema de la cuadratura numérica consiste en calcular una aproximación de

$$I(f) = \int_a^b f(x)dx \qquad -\infty \leq a < x < b \leq \infty$$

La idea es buscar una aproximación $f_n(x)$ de $f(x)$, de manera que

$$I(f_n) = \int_a^b f_n(x)dx = I_n(f) \tag{5.1}$$

sea fácil de evaluar. $I_n(f)$ será entonces una aproximación de $I(f)$. Por ejemplo si $f_n(x)$ es elegida de manera que $\|f_n(x) - f(x)\|_\infty \to 0$, entonces

$$E_n(f) = I_n(f) - I(f) = \int_a^b (f_n(x) - f(x))dx$$

y por lo tanto

$$|E_n(f)| \leq \|f_n(x) - f(x)\|_\infty (b-a)$$

también tiende a cero cuando $n \to \infty$.

Muchos métodos de integración pueden verse dentro de este enfoque, si bien existen otros tipos que se estudian mejor desde otro punto de vista. La mayoría de las integrales numéricas tendrán la forma

$$I_n(f) = \sum_{j=0}^{n} A_j f(x_j) \tag{5.2}$$

con $x_0, x_1, \ldots, x_n$ puntos generalmente en $[a,b]$ y $A_j, j = 0 : n$ constantes llamadas *pesos* correspondientes a dichos nodos. La fórmula (5.2) también se la conoce como *cuadratura numérica*.

Para integrandos con algún tipo de mal comportamiento (por ejemplo, un valor funcional infinito en algún punto del intervalo de integración), es útil considerar al integrando de la forma

$$\int_a^b \underbrace{w(x)}_{\text{malo}} \underbrace{f(x)}_{\text{bueno}} \, dx$$

La mayoría de las fórmulas de integración numérica se basan en definir $f_n(x)$ mediante interpolación polinomial o polinomial a trozos. Comenzaremos la descripción de los métodos de cuadratura con los tres más populares: las reglas del rectángulo, trapecio y Simpson. Antes de ellos recordemos un resultado que aplicaremos en diferentes oportunidades:

**Teorema 5.0.1.** *(Valor medio para integrales) Sea $w(x) \geq 0$ integrable en $[a,b]$, con*

$$\int_a^b w(x)dx < \infty$$

*y sea $f(x)$ continua en $[a,b]$. Entonces existe al menos un punto $\xi$ en $[a,b]$ para el cual*

$$\int_a^b w(x)f(x)dx = f(\xi)\int_a^b w(x)dx$$

Observar que en el caso $w(x) = 1$, el teorema dice que

$$\int_a^b f(x)dx = f(\xi)(b-a)$$

### 5.0.1 La regla del rectángulo

Esta regla se obtiene al aproximar $f(x)$ por un polinomio de interpolación de grado cero que pasa por el punto medio $(\frac{a+b}{2}, f(\frac{a+b}{2}))$ e integrar dicho polinomio. Se tiene:

$$I_0(f) = \int_a^b f(\frac{a+b}{2})dx = (b-a)f(\frac{a+b}{2})$$

que no es más que el área del rectángulo de base $b-a$ y altura $f(\frac{a+b}{2})$.

Estimaremos el error suponiendo que $f$ es dos veces continuamente diferenciable en $[a,b]$. Siendo el error de interpolación

$$E(f) = (x - \frac{a+b}{2})f'(\xi) = (x - \frac{a+b}{2})f[a,x], \qquad \xi \ \ en \ \ (a,b)$$

resulta

$$E_0(f) = \int_a^b (f(x) - p_0(x))dx = \int_a^b (x - \frac{a+b}{2})f[a,x]dx$$

Quisiéramos aplicar el teorema del valor medio para integrales, pero el factor $(x - \frac{a+b}{2})$ cambia de signo en $[a,b]$. Entonces definimos

$$w(x) = \int_a^x (t - \frac{a+b}{2})dt$$

Se tiene así que

$$w(x) = \frac{1}{2}(x^2 - (a+b)x + ab)$$

y por lo tanto $w(x) \leq 0$ en $[a,b]$, con $w(a) = w(b) = 0$. Además

$$\int_a^b w(x)dx = -\frac{1}{12}(b-a)^3$$

Regresando a $E_0(f)$ e integrando por partes, se tiene

$$E_0(f) = \int_a^b w'(x)f[a,x]dx = w(x)f[a,x]|_a^b - \int_a^b w(x)f[a,x,x]dx$$

Usando $w(a) = w(b) = 0$ y aplicando el teorema del valor medio para integrales, resulta

$$E_0(f) = \frac{(b-a)^3}{12}f[a,\xi,\xi] = \frac{(b-a)^3}{24}f"(\eta) \qquad \eta\varepsilon(a,b)$$

Si $b-a$ no es suficientemente pequeño, la regla del rectángulo no da una buena aproximación. En este caso lo que se hace es tomar la integral como suma de integrales sobre subintervalos pequeñas y luego aplicar la regla del rectángulo en cada uno de ellos, dando lugar a lo que se conoce como la *regla compuesta del rectángulo*. Sea $n \geq 1$, $h = \frac{b-a}{2}$, $x_j = a + jh$, $j = 0:n$, entonces:

$$I_{0c} = \int_a^b f(x)dx = \sum_{j=1}^n \int_{x_{j-1}}^{x_j} f(x)dx = \sum_{j=1}^n hf(x_{j-1/2}) + \frac{h^3}{24}f"(\eta_j)$$

en donde $x_{j-1/2} = x_j - h/2$ es el punto medio del intervalo $[x_{j-1}, x_j]$ y $\eta_j$ está en dicho intervalo. Luego

$$I_{0c}(f) = h \sum_{j=1}^{n} f(x_{j-1/2})$$

y

$$E_{0c}(f) = \frac{h^3}{24} \sum_{j=1}^{n} f"(\eta_j) = \frac{h^2 nh}{24} (\frac{1}{n} \sum_{j=1}^{n} f"(\eta_j)) = \frac{h^2}{24}(b-a)f"(\xi), \qquad \xi \varepsilon [a,b]$$

### 5.0.2 La regla del trapecio

Se basa en aproximar $f(x)$ por un polinomio de interpolación lineal por los puntos $(a, f(a))$ y $(b, f(b))$. Se tiene:

$$I_1(f) = \int_a^b \frac{x-b}{a-b} f(a) + \frac{x-a}{b-a} f(b) dx = \frac{(b-a)}{2}(f(a) + f(b))$$

## 5.1 Polinomios Ortogonales

Para poder desarrollar un método de integración importante conocido como la cuadratura gaussiana, necesitamos introducir la noción de polinomios ortogonales y estudiar algunas propiedades de los mismos.

Sea $w(x) \geq 0$ integrable en $(a, b)$ tal que $\int_a^b w(x)g(x)dx = 0$ para $g(x)$ continua implique $g(x) = 0$ en $(a, b)$.

**Definición 5.1.1.** *Definimos el producto escalar de dos funciones $f$ y $g$ de $C[a, b]$ mediante*

$$(f, g) = \int_a^b w(x)f(x)g(x)dx \tag{5.3}$$

*y la norma como*

$$\|f\| = (\int_a^b w(x)f(x)^2 dx)^{1/2}$$

**Definición 5.1.2.** *Decimos que $f$ y $g$ son ortogonales respecto a (5.3) en $C[a, b]$ si*

$$(f, g) = \int_a^b w(x)f(x)g(x)dx = 0$$

**Teorema 5.1.1.** *(Gram-Schmidt) Existe una sucesión de polinomios ortogonales* $\{\varphi_n(x)\}_{n\geq 0}$ *con grado*$(\varphi_n) = n$ *para todo* $n$ *y* $(\varphi_n, \varphi_m) = 0$ *para todo* $m \neq n$, $n, m \geq 0$. *Además se puede construir una única sucesión de polinomios ortonormales si pedimos*

1. $(\varphi_n, \varphi_n) = 1$ *(ó* $\|\varphi_n\| = 1$*)*
2. *el coeficiente de* $x^n > 0$ *en* $\varphi_n(x)$

**Proposición 5.1.1.** *Si* $\{\varphi_n(x)\}_{n\geq 0}$ *es una sucesión de polinomios ortogonales, entonces es linealmente independiente. Además* $(\varphi_n, q) = 0$ *para todo polinomio* $q(x)$ *de grado menor que* $n$.

**Teorema 5.1.2.** *Fórmula de recurrencia. Sea* $\{\varphi_n(x)\}_{n\geq 0}$ *una familia de polinomios ortogonales en* $(a, b)$ *con función de peso* $w(x) \geq 0$. *Entonces para* $n \geq 1$

$$\varphi_{n+1}(x) = (a_n x + b_n)\varphi_n(x) - c_n \varphi_{n-1}(x) \tag{5.4}$$

*en donde* $a_n = \dfrac{A_{n+1}}{A_n}$, $b_n = a_n \dfrac{(x\varphi_n, \varphi_n)}{\|\varphi_n\|^2}$, $c_n = a_n \dfrac{a_n \|\varphi_n\|^2}{a_{n-1}\|\varphi_{n-1}\|^2}$ *y* $A_n$ *es el coeficiente de* $x^n$ *en* $\varphi_n(x)$.

Veamos ahora algunos ejemplos de polinomios ortogonales.

### 5.1.1 Polinomios de Legendre

Se definen mediante la fórmula

$$P_0(x) = 1 \qquad P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2 - 1)^n, \quad n = 1, 2, \ldots \tag{5.5}$$

Puesto que $(x^2 - 1)^n$ es un polinomio de grado $2n$, $P_n(x)$ es un polinomio de grado $n$. El coeficiente de $x^n$ es el mismo que el del polinomio

$$\frac{1}{2^n n!}\frac{d^n}{dx^n}x^{2n} = \frac{1}{2^n n!}2n(2n-1)\ldots(n+1)x^n$$

Por lo tanto $a_n$ de la fórmula de recurrencia (5.4) está dado por $a_n = \dfrac{A_{n+1}}{A_n} = \dfrac{2n+1}{n+1}$.

**Propiedades:**

1. Si $w(x) = 1$ en $[-1, 1]$,

$$(P_n, P_m) = \int_{-1}^{1} P_n P_m dx = \begin{cases} 0 & si \quad n \neq m \\ \dfrac{1}{2n+1} & si \quad n = m \end{cases}$$

Esto se demuestra probando primero que

$$\int_{-1}^{1} x^r P_n(x) dx = (-1)^r r! \int_{-1}^{1} \frac{d^{n-r}}{dx^{n-r}}(x^2-1)^n = 0$$

si $0 \leq r < n$

Luego hay que mostrar que si $r = n$

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

2. $P_n(-x) = (-1)^n P_n(x)$

3. $P_{n+1}(x) = \dfrac{2n+1}{n+1} x P_n(x) - \dfrac{n}{n+1} P_{n-1}(x)$

En efecto, ya vimos que $a_n = \dfrac{2n+1}{n+1}$. Por otro lado

$$\begin{aligned} \int_{-1}^{1} x P_n^2(x) dx &= \int_{-1}^{0} x P_n^2(x) dx + \int_{0}^{1} x P_n^2(x) dx \\ &= \int_{1}^{0} x P_n^2(-x) dx + \int_{0}^{1} x P_n^2(x) dx \\ &= -\int_{0}^{1} x P_n^2(x) dx + \int_{0}^{1} x P_n^2(x) dx = 0 \end{aligned}$$

En consecuencia $b_n = 0$ y $c_n = \dfrac{n}{n+1}$

4.

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \tfrac{1}{2}(3x^2-1) \\ P_3(x) &= \tfrac{1}{2}(5x^3-3x) \end{aligned}$$

### 5.1.2 Polinomios de Chebyshev

Los polinomios de Chebyshev introducidos en el capítulo 4, y que están dados por $T_n = cos(n\ arcosx)$, tienen la propiedad de ser ortogonales en $[-1, 1]$ con respecto a la función de peso $w(x) = \dfrac{1}{\sqrt{1-x^2}}$. Más aún,

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} T_n(x)T_m(x)dx = \begin{cases} 0 & si \quad n \neq m \\ \pi & si \quad n = m = 0 \\ \pi/2 & si \quad n = m > 0 \end{cases}$$

Esto se demuestra usando el hecho que

$$\int_0^{\pi} \cos nx \cos mx dx = 0, \quad si\ n \neq m$$

### 5.1.3 Polinomios de Laguerre

Están definidos como

$$L_n(x) = \frac{1}{n!e^{-x}} \frac{d^n}{dx^n}(x^n e^{-x}), \quad n \geq 0 \tag{5.6}$$

y resultan ortogonales con respecto a $w(x) = e^{-x}$ en $[0, \infty)$, es decir

$$\int_0^{\infty} e^{-x} L_n(x)L_m(x)dx = \begin{cases} 0 & si \quad n \neq m \\ 1 & si \quad n = m \end{cases}$$

De (5.6) resulta que

$$L_0(x) = 1 \qquad\qquad L_1(x) = 1 - x$$

Se verifica además que,

$$L_{n+1}(x) = \frac{1}{n+1}(2n + 1 - x)L_n(x) - \frac{n}{n+1}L_{n-1}(x)$$

### 5.1.4 Raíces de los polinomios ortogonales

Vamos a mostrar que las raíces de un polinomio ortogonal son simples y se encuentran en el intervalo de ortogonalidad

**Teorema 5.1.3.** *Sea $\varphi_n(x)$, $n \geq 0$ una familia de polinomios ortogonales en $[a, b]$ con función de peso $w(x) \geq 0$. Entonces el polinomio $\varphi_n(x)$ tiene exactamente $n$ raíces distintas en $(a, b)$.*

**Prueba:**

Sean $x_1, x_2, \ldots, x_r$ los ceros de $\varphi_n(x)$ para los cuales se cumplen

1. $a < x_i < b$
2. $\varphi_n(x)$ cambia de signo en $x_i$

   Es decir

$$\varphi_n(x) = h(x)(x - x_1)^{m_1}(x - x_2)^{m_2} \ldots (x - x_r)^{m_r}$$

   con $m_i$ impar para que cambie de signo y $h(x)$ no cambia de signo en $(a, b)$.

   Supongamos $r < n$ y definamos

$$p(x) = (x - x_1)(x - x_2) \ldots (x - x_r)$$

   Entonces

$$\varphi_n(x)p(x) = h(x)(x - x_1)^{m_1+1}(x - x_2)^{m_2+1} \ldots (x - x_r)^{m_r+1}$$

   no cambia de signo en $(a, b)$ y por lo tanto

$$\int_a^b w(x)p(x)\varphi_n(x)dx \neq 0$$

   contradiciendo la ortogonalidad de $\varphi_n(x)$.□

### 5.1.5 Cuadratura Gaussiana

Hemos visto que si $n+1$ puntos distintos $x_0, x_1, \ldots, x_n$ del intervalo $[a, b]$ son especificados, podemos hallar coeficientes $A_{0,n}, A_{1,n}, \ldots, A_{n,n}$, tales que

$$\int_a^b f(x)w(x)dx = \sum_{j=0}^{n} A_{j,n} f(x_j) \tag{5.7}$$

donde $w(x)$ es una función de peso definida anteriormente, de manera que (5.7) sea exacta para todo polinomio de grado $n$. Sin embargo, es posible elegir los nodos $x_j$ de manera que (5.7) sea exacta para polinomios de grado superior.

**Teorema 5.1.4.** *Si $x_0, x_1, \ldots, x_n$ se eligen como los ceros del polinomio $\varphi_{n+1}(x)$ de grado $n+1$ de la familia de polinomios ortogonales asociados a $w(x)$ entonces (5.7) es exacta para todo polinomio de grado $2n+1$ si los $A_{j,n}$ están dados por*

$$A_{j,n} = \int_a^b l_{j,n}(x) w(x) dx$$

*con* $l_{j,n}(x) = \dfrac{\prod_{i \neq j}(x - x_i)}{\prod_{i \neq j}(x_j - x_i)}$

**Teorema 5.1.5.** *El error de la cuadratura gaussiana está dado por*

$$E(f) = \sum_{j=0}^{n} A_{j,n} f(x_j) - \int_a^b f(x) w(x) dx = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b (\phi_{n+1}(x))^2 w(x) dx$$

*con* $\phi(x) = \prod_{i=0}^{n}(x - x_i)$.