

# **ANÁLISIS NUMERICO I**

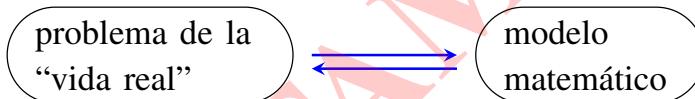
# Análisis Numérico / Análisis Numérico I

## Clase 1

En esta clase veremos algunas definiciones básicas, presentaremos algunos problemas que estudiaremos durante el curso y repasaremos algunos conceptos matemáticos que serán necesarios para las clases siguientes.

### Preliminares y definiciones básicas

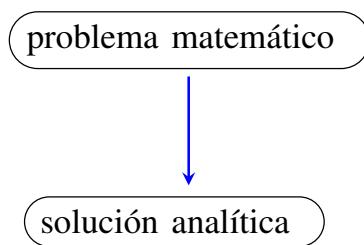
Los **modelos matemáticos** son la herramienta básica para resolver problemas científicos. Generalmente, tales problemas se originan en situaciones o “**problemas de la vida real**” de las más variadas áreas o disciplinas. Características propias del problema a estudiar permiten definir ecuaciones y/o inecuaciones que ayudan a formular el modelo matemático.



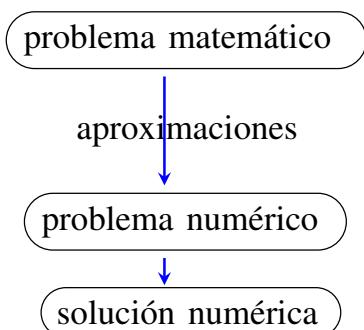
Es fácil imaginar que, frecuentemente, en problemas complejos se realizan algunas simplificaciones al construir el modelo. De allí que el modelo matemático no es necesariamente una descripción exacta de la realidad y que las respuestas que se obtienen deben ser chequeadas y comparadas con resultados experimentales.

Así, un modelo matemático es una buena descripción del problema a estudiar, pero en general no siempre da una respuesta directa al problema considerado. Una primera posibilidad podría ser realizar muchas más simplificaciones en el modelo de manera de poder obtener una solución analítica. La desventaja de esto es que el modelo matemático podría diferir mucho del problema real. En lugar de esto, se pueden realizar aproximaciones numéricas del problema, que si bien pueden introducir errores, es posible estudiar y estimar cómo estas aproximaciones afectan a la precisión de la solución.

modelo matemático 1



modelo matemático 2



**Definición 1** Un **problema numérico** es una clara y nada ambigua descripción de la conexión funcional entre datos de entrada (variables independientes del problema o *input*) y los datos de salida (resultados deseados o *output*). Estos datos, *input* y *output*, consisten en un conjunto finito de cantidades reales.

---

Para resolver un problema numérico se disponen, además de la teoría matemática, de 2 herramientas fundamentales: las **computadoras** y los **algoritmos**.

**Definición 2** *Un algoritmo para un problema numérico es una completa descripción de un número finito de pasos con operaciones bien definidas, y sin ambigüedades, a través de las cuales una lista de datos de entrada se convierte en una lista de datos de salida.*

El **objetivo del Análisis Numérico** es formular y estudiar métodos numéricos y algoritmos para obtener la solución de problemas provenientes de la vida real y/o de diversas áreas, modelizados matemáticamente.

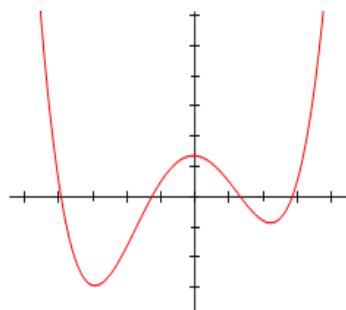
### Temario y algunos problemas:

1. **Teoría de errores.** Fuentes de error, aritmética finita, artimética del computador.

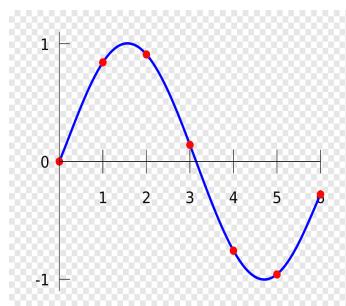


2. **Ecuaciones no lineales.**

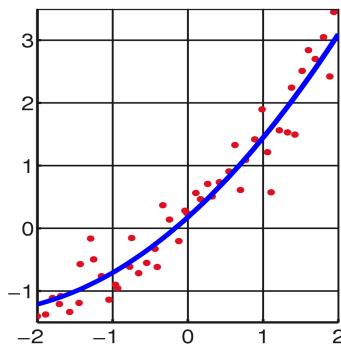
Dada  $f : \mathbb{R} \rightarrow \mathbb{R}$  hallar  $x_*$  solución de  $f(x) = 0$ .



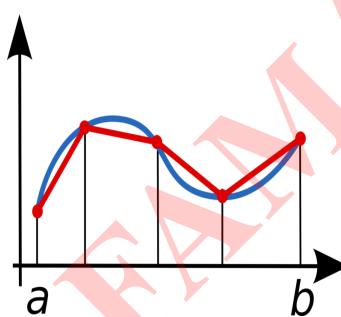
3. **Interpolación polinomial.**



4. **Teoría de mejor aproximación.** Método de cuadrados mínimos.

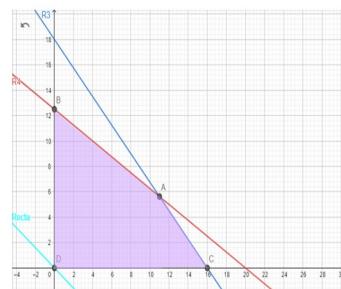


5. **Integración numérica.**



6. **Resolución numérica de sistemas lineales.** Resolver numéricamente sistemas lineales de la forma  $Ax = b$ .

7. **Programación lineal.**



## Preliminares matemáticos

**Teorema 1 (Valor intermedio para funciones continuas).** Sea  $f$  continua en  $[a, b]$ . Sea  $d$  entre  $f(a)$  y  $f(b)$  entonces existe  $c \in [a, b]$  tal que  $f(c) = d$ .

**Teorema 2 (Valor medio).** Sea  $f$  continua en  $[a, b]$  y derivable en  $(a, b)$ . Entonces para todo par  $x, c \in [a, b]$  se cumple que

$$\frac{f(x) - f(c)}{x - c} = f'(\xi), \quad \text{para algún } \xi \text{ entre } x \text{ y } c.$$

Esto dice que  $f(x) = f(c) + f'(\xi)(x - c)$ .

**Teorema 3 (Taylor).** Si  $f \in C^{(n)}[a, b]$  y existe  $f^{(n+1)}(a, b)$  entonces para todo par  $x, c \in [a, b]$  se tiene que

$$f(x) = \sum_{k=0}^n \frac{1}{k!} f^{(k)}(c)(x-c)^k + E_n(x),$$

donde

$$E_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x-c)^{n+1}, \quad \text{para algún } \xi \text{ entre } x \text{ y } c.$$

**Observación:** tomando  $y = c$ ,  $(x - c) = h$  y por lo tanto  $x = y + h$ , entonces

$$f(y+h) = f(y) + hf'(y) + \frac{h^2}{2}f''(y) + \cdots + \frac{h^n}{n!}f^{(n)}(y) + \frac{h^{n+1}}{(n+1)!}f^{(n+1)}(\xi),$$

para algún  $\xi$  entre  $y$  e  $(y+h)$ .

**Teorema 4 (Taylor con resto integral).** Si  $f \in C^{(n)}[a, b]$  y existe  $f^{(n+1)}(a, b)$  entonces para todo par  $x, c \in [a, b]$  se tiene que

$$f(x) = \sum_{k=0}^n \frac{1}{k!} f^{(k)}(c)(x-c)^k + R_n(x),$$

donde

$$R_n(x) = \frac{1}{n!} \int_c^x f^{(n+1)}(t)(x-t)^n dt.$$

## Órdenes de convergencia

Al resolver un problema numérico generalmente no es usual obtener una solución en forma directa o con una fórmula cerrada. Por el contrario, se suele obtener una sucesión de aproximaciones cuya precisión aumenta progresivamente. De allí que es muy importante el concepto de convergencia de sucesiones y velocidad de esta convergencia.

Consideremos, por ejemplo, la sucesión dada por  $x_n = 3^{-n}$  para  $n \in \mathbb{N}$ . Es claro que

$$x_n = 3^{-n} = \frac{1}{3^n} \rightarrow 0 \quad \text{cuando} \quad n \rightarrow \infty.$$

Luego

$$\frac{|x_{n+1} - 0|}{|x_n - 0|} = \frac{3^{-(n+1)}}{3^{-n}} = \frac{3^n}{3^{n+1}} \rightarrow \frac{1}{3} = C.$$

En este caso, cuando  $C$  es menor que 1, se dice la convergencia es **lineal**. En breve formalizaremos esta definición. Notar que en este ejemplo particular se tiene que

$$|x_{n+1} - 0| = \frac{1}{3}|x_n - 0| \quad \text{para todo } n.$$

Ahora consideremos la sucesión dada por  $x_n = \frac{1}{n!}$ , la que claramente converge a 0. Luego,

$$\frac{|x_{n+1} - 0|}{|x_n - 0|} = \frac{n!}{(n+1)!} = \frac{1}{n+1} \rightarrow 0, \quad \text{cuando } n \rightarrow \infty.$$

En este caso se dice que la convergencia es **superlineal**. Veamos formalmente estas definiciones.

---

**Definición 3** Sea  $\{x_n\}$  una sucesión de números reales que converge a  $x_*$ .

Se dice que la sucesión  $\{x_n\}$  tiene tasa de convergencia (al menos) **lineal** si existe una constante  $c$  tal que  $0 < c < 1$  y un  $N \in \mathbb{N}$  tal que

$$|x_{n+1} - x_*| \leq c|x_n - x_*| \quad \text{para todo } n \geq N.$$

Se dice que la tasa de convergencia es (al menos) **superlineal** si existe una sucesión  $\{\varepsilon_n\}$  que converge a 0 y un  $N \in \mathbb{N}$  tal que

$$|x_{n+1} - x_*| \leq \varepsilon_n |x_n - x_*| \quad \text{para todo } n \geq N.$$

Se dice que la tasa de convergencia es (al menos) **cuadrática** si existe una constante positiva  $c$  y un  $N \in \mathbb{N}$  tal que

$$|x_{n+1} - x_*| \leq c|x_n - x_*|^2 \quad \text{para todo } n \geq N.$$

## Notación $\mathcal{O}$ grande y $o$ chica

Veremos ahora una notación usual para comparar sucesiones y funciones.

Sean  $\{x_n\}$  y  $\{\alpha_n\}$  dos sucesiones distintas. Se dice que

$$x_n = \mathcal{O}(\alpha_n)$$

si existen una constante  $C > 0$  y  $r \in \mathbb{N}$  tal que  $|x_n| \leq C|\alpha_n|$  para todo  $n \geq r$ .

Se dice que

$$x_n = o(\alpha_n)$$

si existe una sucesión  $\{\varepsilon_n\}$  que converge a 0, con  $\varepsilon_n \geq 0$  y un  $r \in \mathbb{N}$  tal que  $|x_n| \leq \varepsilon_n |\alpha_n|$  para todo  $n \geq r$ . Intuitivamente, esto dice que  $\lim_{n \rightarrow \infty} (x_n / \alpha_n) = 0$ .

### Ejemplo 1:

$$\frac{n+1}{n^2} = \mathcal{O}\left(\frac{1}{n}\right).$$

Si

$$\frac{n+1}{n^2} \leq C \frac{1}{n} \Rightarrow \frac{n+1}{n} \leq C,$$

por lo tanto basta tomar  $C = 2$  y  $r = 1$ .

### Ejemplo 2:

$$\frac{1}{n \ln n} = o\left(\frac{1}{n}\right).$$

Si

$$\frac{1}{n \ln n} \leq \varepsilon_n \frac{1}{n},$$

basta tomar  $\varepsilon_n = 1/\ln n$ .

Esta notación también se puede usar para comparar funciones. Se dice que

$$f(x) = \mathcal{O}(g(x)) \quad \text{cuando } x \rightarrow \infty$$

si existen una constante  $C > 0$  y  $r \in \mathbb{R}$  tal que  $|f(x)| \leq C|g(x)|$  para todo  $x \geq r$ .

Análogamente, se dice que  $f(x) = o(g(x))$  cuando  $x \rightarrow \infty$  si  $\lim_{x \rightarrow \infty} (f(x)/g(x)) = 0$ .

---

**Ejemplo 3:**

$\sqrt{x^2 + 1} = \mathcal{O}(x)$  si  $x \rightarrow \infty$ , pues  $\frac{\sqrt{x^2 + 1}}{x} = \sqrt{\frac{x^2 + 1}{x^2}} = \sqrt{1 + \frac{1}{x^2}} \leq C$ , luego basta tomar  $C = 2$  y  $r = 1$ .

**Ejemplo 4:**

$\frac{1}{x^2} = o\left(\frac{1}{x}\right)$ , pues

$$\frac{\frac{1}{x^2}}{\frac{1}{x}} = \frac{x}{x^2} = \frac{1}{x} \rightarrow 0,$$

si  $x \rightarrow \infty$ .

Por último, consideremos el caso de comparación de dos funciones cuando  $x \rightarrow x_*$ .

Se dice que

$$f(x) = \mathcal{O}(g(x)) \quad \text{cuando } x \rightarrow x_*$$

si existen una constante  $C > 0$  y un entorno alrededor de  $x_*$  tal que  $|f(x)| \leq C|g(x)|$  para todo  $x$  en ese entorno.

Análogamente, se dice que  $f(x) = o(g(x))$  cuando  $x \rightarrow x_*$  si  $\lim_{x \rightarrow x_*} (f(x)/g(x)) = 0$ .

**Ejemplo 5:** (Ejercicio)

Ver que

$$\sin x = x - \frac{x^3}{6} + \mathcal{O}(x^5), \quad \text{si } x \rightarrow 0.$$

## Clase 2

### Algoritmo de Horner: evaluación de polinomios

El algoritmo de Horner es un algoritmo reconocido por su eficiencia para evaluar polinomios utilizando un número mínimo de operaciones (sumas y productos).

Para fijar ideas veamos un ejemplo. Consideremos

$$p(x) = 2 + 4x - 5x^2 + 2x^3 - 6x^4 + 8x^5 + 10x^6$$

Primero notemos que, dado  $x$ , el método usual para evaluar  $x^k$  requiere  $(k - 1)$  productos:

$$x^k = \underbrace{x \cdot \dots \cdot x}_k,$$

por ejemplo, si  $k = 2$ , se requiere 1 producto, si  $k = 3$ , se requieren 2 productos, etc.

**Método 1:** evaluar las potencias de la manera usual, multiplicar por la constante respectiva y sumar los monomios. Es fácil ver que se requieren:

$$0 + 1 + 2 + 3 + 4 + 5 + 6 = \sum_{i=1}^6 i = \frac{6(6+1)}{2} = 21 \text{ productos, y por lo tanto, se tienen}$$

# sumas	=	6
# productos	=	21

**Método 2:** la idea consiste en evaluar las potencias de  $x$  en forma sucesiva:

$$x^2 = x * x, \quad x^3 = x * x^2, \quad x^4 = x * x^3, \quad x^5 = x * x^4, \quad x^6 = x * x^5,$$

teniendo en cuenta que cada potencia de  $x$  se debe multiplicar por un coeficiente, se tienen  $0 + 1 + 2 + 2 + 2 + 2 + 2 = 11$  productos, y por lo tanto

# sumas	=	6
# productos	=	11

**Método 3:** conocido como método de Horner o multiplicación encajada. La idea consiste en reescribir convenientemente el polinomio  $p(x)$  de modo de reducir el número de productos:

$$p(x) = 2 + x(4 + x(-5 + x(2 + x(-6 + x(8 + x \cdot 10))))).$$

Es fácil ver que ahora se requieren:

# sumas	=	6
# productos	=	6

Si el grado de  $p(x)$  es  $n$ , se requieren  $n(n + 1)/2$  productos en el método 1,  $2n - 1$  en el método 2 y sólo  $n$  en el método 3.

Si  $p(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$ , con  $a_n \neq 0$ , la evaluación de  $p(x)$  en  $x = z$  se realiza con los siguientes pasos:

$$\begin{aligned} b_{n-1} &= a_n \\ b_{n-2} &= a_{n-1} + z * b_{n-1} \\ &\vdots \\ b_0 &= a_1 + z * b_1 \\ p(z) &= a_0 + z * b_0. \end{aligned}$$

---

En forma más compacta, podemos escribir:

### Algoritmo de Horner (multiplicación encajada)

Dados el polinomio  $p(x)$ , de grado  $n$ , con coeficientes  $a_i$ , para  $i = 0, \dots, n$ , con  $a_n \neq 0$  y un número real  $z$  en el que se desea evaluar  $p(x)$ .

```
input  $n; a_i, i = 0, \dots, n; z$ 
 $b_{n-1} \leftarrow a_n$ 
for  $k = n - 1$  to  $0$  step  $-1$ , do
     $b_{k-1} \leftarrow a_k + z * b_k$ 
end do
output  $b_i, i = -1, \dots, n - 1$ 
end
```

Notar que  $b_{-1} = p(z)$

## Fuentes de error y aritmética de la computadora

Una de las tareas más importantes en Análisis Numérico es estimar la precisión del resultado en un cálculo numérico.

Los resultados numéricos están influenciados por diferentes tipos de errores. Algunos de estos pueden eliminarse, en otros se puede reducir su influencia, y otros son inevitables y nada se puede hacer.

### Principales fuentes de error.

1. **Errores en los datos de entrada** (frecuentemente inevitables). Los datos de entrada pueden ser el resultado de mediciones experimentales que podrían estar afectadas a errores sistemáticos debido al equipamiento utilizado. También, se tienen los errores que se producen al representar un número real irracional con un número finito de dígitos.
2. **Errores de redondeo.** Aparecen cuando los cálculos se realizan usando un número finito de dígitos.
3. **Errores de truncamiento.** Aparecen cuando un proceso infinito es reemplazado por un proceso finito. Ejemplos: aproximar una serie por una suma parcial, aproximar una función por un polinomio (Taylor), aproximar una derivada por un cociente incremental, etc.
4. **Errores humanos.** Son frecuentes y a veces difíciles de detectar. Ejemplos: errores en la formulación del problema, en los cálculos “a mano”, al escribir un programa en la computadora, etc.

## Errores absolutos y relativos

Antes de analizar que ocurre al representar números en una computadoras y se realizan operaciones, veremos algunos conceptos básicos que sirven en un contexto más general.

**Definición 1** Cuando un número real  $r$  (valor exacto) es aproximado por otro número  $\tilde{r}$ , se define el **error** por  $r - \tilde{r}$ . Llamaremos, respectivamente, a

$$\begin{aligned}\textbf{Error absoluto: } \Delta r &= |r - \tilde{r}|, \\ \textbf{Error relativo: } \delta r &= \left| \frac{r - \tilde{r}}{r} \right| = \frac{\Delta r}{|r|}.\end{aligned}$$

También se llama **error (relativo) porcentual** al producto  $100 * \delta r$ .

En general, el error relativo y el error porcentual son más útiles que el error absoluto, porque da una idea del error cometido relativo a la magnitud de la cantidad que se está considerando.

En términos prácticos no se conocen exactamente los valores de los errores absolutos y relativos sino que se tienen cotas de ellos. Siempre se trata que las cotas sean lo más ajustadas posible. Así, por ejemplo, si  $r = \sqrt{2}$  y  $\tilde{r} = 1.414$ , entonces

$$\begin{aligned}\Delta r &= |r - \tilde{r}| = |\sqrt{2} - 1.414| = |0.0002135...| \leq 0.00022 \\ \delta r &= \frac{\Delta r}{|r|} = \frac{|0.0002135...|}{\sqrt{2}} \leq 0.00016\end{aligned}$$

y el error porcentual es 0.016 %.

Las siguientes notaciones son equivalentes:

$$\tilde{r} = 1.414, \quad \Delta r \leq 0.22 \cdot 10^{-3}$$

$$r = 1.414 \pm 0.22 \cdot 10^{-3}$$

$$1.41378 \leq r \leq 1.41422$$

## Redondeo y truncado

Existen dos maneras de escribir un número real reduciendo el número de dígitos: **redondeo** y **truncado**.

**Redondeo.** Veamos algunos ejemplos de redondeo a 3 dígitos (decimales):

$$\begin{aligned}0.774432 &\rightarrow 0.774 \\ 1.23767 &\rightarrow 1.238 \\ 0.3225 &\rightarrow 0.323\end{aligned}$$

En general, la aproximación por redondeo a  $n$  dígitos decimales  $\tilde{r}$  de un número  $r$  es un número de  $n$  dígitos decimales (después del punto decimal) que coinciden con los de  $r$  si el dígito  $(n+1)$  de  $r$  es 0, 1, 2, 3 o 4. Por otro lado, si el dígito  $(n+1)$  de  $r$  es 5, 6, 7, 8 o 9, entonces para definir  $\tilde{r}$  se suma una unidad al  $n$ -ésimo dígito de  $r$ . Así se cumple que:

$$|r - \tilde{r}| \leq \frac{1}{2} \cdot 10^{-n}. \tag{1}$$

Para fijar ideas veamos un ejemplo muy sencillo, con  $n = 1$ :

- si  $r = 0.11$  entonces  $\tilde{r} = 0.1$  y  $|r - \tilde{r}| = 0.01 \leq 0.05 = 5 \cdot 10^{-2} = \frac{1}{2} \cdot 10^{-1}$ ;
- si  $r = 0.16$  entonces  $\tilde{r} = 0.2$  y  $|r - \tilde{r}| = 0.04 \leq 0.05 = 5 \cdot 10^{-2} = \frac{1}{2} \cdot 10^{-1}$ .

**Truncado.** Veamos algunos ejemplos de truncado a 3 dígitos (decimales):

$$\begin{array}{ll} 0.774432 & \rightarrow 0.774 \\ 1.23767 & \rightarrow 1.237 \\ 0.3225 & \rightarrow 0.322 \end{array}$$

En general, la aproximación por truncado (o truncamiento) a  $n$  dígitos decimales  $\hat{r}$  de un número  $r$  es un número de  $n$  dígitos decimales (después del punto decimal) que coinciden con los  $n$  primeros dígitos de  $r$ . Así se cumple que:

$$|r - \hat{r}| \leq 10^{-n}, \quad (2)$$

Para fijar ideas veamos un ejemplo muy sencillo, con  $n = 1$ :

- si  $r = 0.11$  entonces  $\hat{r} = 0.1$  y  $|r - \hat{r}| = 0.01 \leq 0.1 = 10^{-1}$ ;
- si  $r = 0.19$  entonces  $\hat{r} = 0.1$  y  $|r - \hat{r}| = 0.09 \leq 0.1 = 10^{-1}$ .

En general, las computadoras trabajan con el sistema de redondeo y no el truncado.

Es importante observar que, si bien el error absoluto que introduce el redondeo depende de la magnitud del número, el error relativo, que es el más significativo, por ser independiente de la magnitud, está controlado en términos de la cantidad de dígitos con la que se trabaja. Así vamos a definir el número de dígitos significativos en términos del error relativo.

**Definición 2** *El número  $\tilde{r}$  aproxima a  $r$  con  $m$  dígitos significativos si*

$$\delta r = \frac{\Delta r}{|r|} \leq 5 \cdot 10^{-m}.$$

Esto dice que el error relativo es del orden de  $10^{-m}$ .

Así, en los ejemplos anteriores tenemos

$r$	$\tilde{r}$	$\Delta r$	$\delta r$	díg. signif.
0.774432	0.774	0.00043	0.00056 ( $< 5 \cdot 10^{-3}$ )	3
1.23767	1.238	0.00033	0.00027 ( $< 5 \cdot 10^{-4}$ )	4
0.3225	0.322	0.0005	0.0015 ( $< 5 \cdot 10^{-3}$ )	3

## Errores en las operaciones

Analizaremos los errores que se introducen en las operaciones básicas (suma, resta, multiplicación y división). Más específicamente, nos centraremos en las dos primeras.

Sean  $x_1, x_2 \in \mathbb{R}$ , y  $\bar{x}_1, \bar{x}_2$  aproximaciones de  $x_1$  y  $x_2$  respectivamente.

Sean  $y = x_1 + x_2$ ,  $\bar{y} = \bar{x}_1 + \bar{x}_2$ .

---

El error en la operación **suma** está dado por:

$$y - \bar{y} = (x_1 + x_2) - (\bar{x}_1 + \bar{x}_2) = (x_1 - \bar{x}_1) + (x_2 - \bar{x}_2),$$

por lo tanto el error absoluto

$$\begin{aligned}\Delta y &= |y - \bar{y}| \leq |x_1 - \bar{x}_1| + |x_2 - \bar{x}_2| \\ \Delta y &\leq \Delta x_1 + \Delta x_2\end{aligned}$$

y el error relativo

$$\delta y = \frac{\Delta y}{|y|} \leq \frac{\Delta x_1 + \Delta x_2}{|x_1 + x_2|}.$$

En general, si  $y = \sum_{i=1}^n x_i$  entonces  $\Delta y \leq \sum_{i=1}^n \Delta x_i$ .

El caso de la **resta** es similar. Sean  $y = x_1 - x_2$ ,  $\bar{y} = \bar{x}_1 - \bar{x}_2$ . Entonces el error absoluto se obtiene haciendo

$$\begin{aligned}\Delta y &= |y - \bar{y}| = |(x_1 - x_2) - (\bar{x}_1 - \bar{x}_2)| = |(x_1 - \bar{x}_1) - (x_2 - \bar{x}_2)| \\ \Delta y &\leq \Delta x_1 + \Delta x_2\end{aligned}$$

y el error relativo

$$\delta y = \frac{\Delta y}{|y|} \leq \frac{\Delta x_1 + \Delta x_2}{|x_1 - x_2|}.$$

Es posible estimar cotas de errores para la **multiplicación** y la **división** definiendo  $y = x_1 * x_2$ ,  $\bar{y} = \bar{x}_1 * \bar{x}_2$ ,  $z = x_1 / x_2$ ,  $\bar{z} = \bar{x}_1 / \bar{x}_2$ . Se puede deducir que:

$$\Delta y \lesssim |x_2| \Delta x_1 + |x_1| \Delta x_2 \quad \delta y = \frac{\Delta y}{|y|} \lesssim \frac{\Delta x_1}{|x_1|} + \frac{\Delta x_2}{|x_2|}$$

y que

$$\Delta z \lesssim \frac{1}{|x_2|} \Delta x_1 + \frac{|x_1|}{|x_2|^2} \Delta x_2 \quad \delta z = \frac{\Delta z}{|z|} \lesssim \frac{\Delta x_1}{|x_1|} + \frac{\Delta x_2}{|x_2|}$$

# Clase 3

## Cancelación de dígitos significativos

Recordemos la definición de dígitos significativos: el número  $\bar{a}$  aproxima al número real  $a$  con  $r$  **dígitos significativos** si

$$\frac{\Delta a}{|a|} \leq 5 \cdot 10^{-r} = \frac{1}{2} \cdot 10^{1-r}.$$

Un efecto no deseable en algoritmos numéricos es la gran cancelación de dígitos significativos que se produce en la resta de números próximos. Para fijar ideas veamos un ejemplo.

Sean  $x_1 = 10.123455 \pm 0.5 \cdot 10^{-6}$  y  $x_2 = 10.123789 \pm 0.5 \cdot 10^{-6}$ .

$x_1$  y  $x_2$  tienen error absoluto menor o igual a  $0.5 \cdot 10^{-6}$  y error relativo menor a  $0.5 \cdot 10^{-7} = 5 \cdot 10^{-8}$ , esto significa que ambos tienen 8 dígitos significativos.

Ahora bien, la resta  $y = x_1 - x_2 = -0.000334 \pm 10^{-6}$ , tiene un error absoluto pequeño, sin embargo el error relativo

$$\frac{\Delta y}{|y|} \leq \frac{10^{-6}}{0.000334} < 3 \cdot 10^{-3} < 5 \cdot 10^{-3},$$

por lo tanto la resta  $y$  tiene **sólo 3 dígitos significativos**.

Por lo tanto, es recomendable evitar restas de números próximos, siempre que sea posible.

## Representación de números en una computadora

El ser humano está acostumbrado a utilizar un sistema de numeración decimal, el cual es un sistema posicional con base  $\beta = 10$ . La mayoría de las computadoras usa internamente la base  $\beta$  igual a 2 o 16.

**Definición 1** sea  $\beta \in \mathbb{N}, \beta \geq 2$ , todo número real  $r$  puede ser escrito en la forma:

$$(\pm d_n d_{n-1} \dots d_2 d_1 d_0. d_{-1} d_{-2} \dots)_{\beta}$$

donde  $d_n, d_{n-1}, \dots, d_0, d_{-1}, \dots$  son números naturales entre 0 y  $(\beta - 1)$ . El valor del número  $r$  es:

$$\pm d_n \beta^n + d_{n-1} \beta^{n-1} + \dots + d_2 \beta^2 + d_1 \beta^1 + d_0 \beta^0 + d_{-1} \beta^{-1} + d_{-2} \beta^{-2} + \dots$$

### Ejemplos:

$$1. (760)_8 = 7 \cdot 8^2 + 6 \cdot 8^1 + 0 \cdot 8^0 = (496)_{10}$$

$$2. (101.101)_2 = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} = (5.625)_{10}$$

$$3. (0.333\dots)_{10} = 3 \cdot 10^{-1} + 3 \cdot 10^{-2} + \dots = \frac{1}{3}$$

---

$$4. (0.1)_{10} = (0.0001100110011\dots)_2$$

**Notar que en el último ejemplo  $(0.1)_{10}$  no tiene representación binaria finita!**

**Observaciones:**

1. la mayoría de los números reales no pueden ser representados exactamente en cualquier base;
2. aparecen errores de representación cuando un número es convertido de un sistema de numeración a otro;
3. aparecen errores debido a que la computadora usa aritmética finita.

## ¿cómo se representan los números en una computadora?

Básicamente, existen dos sistemas de representación de números en una computadora:

- sistema de punto fijo,
- sistema de punto flotante.

### Sistema de punto fijo

El primero de ellos es el utilizado por las primeras computadoras (aproximadamente en 1940–1950) y donde los números se representan utilizando una cantidad fija de números enteros y de números fraccionarios. Por ejemplo, si usáramos la base  $\beta$ ,  $(s+1)$  dígitos para la parte entera y  $t$  para la parte fraccionaria, tendríamos:

$$\pm d_s d_{s-1} \dots d_2 d_1 d_0 . d_{-1} d_{-2} \dots d_{-t},$$

donde cada  $d_i \in \{0, \dots, \beta - 1\}$ . En sistemas contables, aún hoy en día, suele usarse este sistema donde la cantidad de dígitos fraccionarios es  $t = 2$  para representar los centavos.

La principal desventaja de este sistema es que no es posible representar simultáneamente números reales muy pequeños y muy grandes, sin que la cantidad de dígitos  $s$  y  $t$  sean demasiados grandes. Por ejemplo si  $s = 3$  y  $t = 3$ , el número más grande y el más pequeño que se pueden representar en este sistema son 999.999 y 000.001, respectivamente. La manera de solucionar este problema es usar la notación científica y esto da origen al otro sistema.

### Sistema de punto flotante

**Definición 2** Un sistema de punto flotante  $(\beta, t, L, U)$  es el conjunto de números normalizados en punto flotante en el sistema de numeración con base  $\beta$ , y  $t$  dígitos para la parte fraccionaria, es decir, números de la forma:

$$x = m \beta^e$$

donde

$$m = \pm 0. d_{-1} d_{-2} \dots d_{-t}$$

con  $d_{-i} \in \{0, \dots, \beta - 1\}$  para  $i = 1, \dots, t$ , con  $d_{-1} \neq 0$  y  $L \leq e \leq U$ . Además,  $\beta$ ,  $e$  y  $m$  se denominan base, exponente y mantisa, respectivamente. Es decir,  $1/\beta \leq |m| < 1$ .

### Observaciones:

1. aunque el sistema de punto flotante permite representar magnitudes de órdenes muy variados, a diferencia del sistema de punto fijo, también puede ocurrir *overflow* si  $e > U$  o *underflow* si  $e < L$ ;
2. **el cero no puede representarse en este sistema de números normalizados.**

## Errores de redondeo en aritmética de punto flotante

Al representar números en un sistema de punto flotante  $(\beta, t, L, U)$  se producen errores de redondeo debido a la precisión limitada. Asumiendo redondeo, estimaremos una cota de los errores absoluto y relativo.

Supongamos que podemos escribir un número real (exacto) en la forma:

$$x = m\beta^e, \quad \frac{1}{\beta} \leq |m| < 1,$$

donde el exponente  $e$  es tal que  $L \leq e \leq U$ .

Escribimos ahora su representante en el sistema de punto flotante:

$$fl(x) = x_r = m_r\beta^e, \quad \frac{1}{\beta} \leq |m| < 1,$$

donde  $m_r$  es la mantisa que se obtiene redondeando a  $t$  dígitos la parte fraccionaria de  $m$ . Entonces, es claro que

$$|m_r - m| \leq \frac{1}{2}\beta^{-t},$$

y por lo tanto, una cota del error del absoluto de representación en  $x$  es

$$|x_r - x| \leq \frac{1}{2}\beta^{-t}\beta^e.$$

Para el error relativo, tenemos lo siguiente:

$$\frac{|x_r - x|}{|x|} \leq \frac{\frac{1}{2}\beta^{-t}\beta^e}{|m|\beta^e} = \frac{1}{2|m|}\beta^{-t} \leq \frac{1}{2}\beta^{1-t},$$

pues si  $|m| \geq 1/\beta$  entonces  $\frac{1}{|m|} \leq \beta$ .

Luego el error relativo debido al redondeo en la representación en el sistema de punto flotante está acotado por:

$$\frac{|x_r - x|}{|x|} \leq \frac{1}{2}\beta^{1-t} = \mu,$$

donde  $\mu$  se llama unidad de redondeo.

**Notar que el error absoluto de representación en punto flotante depende del orden de la magnitud, en cambio el error relativo no.**

## ¿Cómo se realiza la suma en aritmética de punto flotante?

Para fijar ideas veremos dos ejemplos con el sistema de punto flotante dado por  $(\beta, t, L, U) = (10, 4, -9, 9)$ . Sean

$$x = m_x \beta^{e_x}, \quad y = m_y \beta^{e_y},$$

con  $x \geq y$ . Queremos calcular  $z = fl(x + y)$

**Ejemplo 1:** sean  $x = 0.1234 \cdot 10^0$ ,  $y = 0.4567 \cdot 10^{-2}$ , entonces

$$\begin{aligned} x + y &= 0.1234 \cdot 10^0 + 0.4567 \cdot 10^{-2} \\ &= 0.1234 \cdot 10^0 + 0.004567 \cdot 10^0 \\ &= (0.1234 + 0.004567) \cdot 10^0 = 0.12797 \cdot 10^0, \end{aligned}$$

por lo tanto,  $fl(x + y) = 0.1280 \cdot 10^0$ .

**Ejemplo 2:** sean  $x = 0.1234 \cdot 10^0$ ,  $y = 0.5678 \cdot 10^{-5}$ , entonces

$$\begin{aligned} x + y &= 0.1234 \cdot 10^0 + 0.5678 \cdot 10^{-5} \\ &= 0.1234 \cdot 10^0 + 0.000005678 \cdot 10^0 \\ &= (0.1234 + 0.000005678) \cdot 10^0 = 0.123405678 \cdot 10^0, \end{aligned}$$

por lo tanto,  $fl(x + y) = 0.1234 \cdot 10^0 = x$ .

Esto ocurre porque el orden de magnitud de  $x$  es con respecto a  $y$  es muy grande.

**Observación:** algunas propiedades o axiomas de la aritmética infinita dejan de valer en aritmética de punto flotante. Veamos con un ejemplo que la propiedad asociativa  $((a + b) + c = a + (b + c))$  no es válida en un sistema de punto flotante, es decir:  $fl(fl(a + b) + c) \neq fl(a + fl(b + c))$ .

Dado el sistema de punto flotante dado por  $(\beta, t, L, U) = (10, 4, -9, 9)$  consideremos los números  $a = 0.9876 \cdot 10^4$ ,  $b = -0.9880 \cdot 10^4$  y  $c = 0.3456 \cdot 10^1$ . Entonces, por un lado,

$$\begin{aligned} fl(fl(a + b) + c) &= fl(fl(0.9876 \cdot 10^4 - 0.9880 \cdot 10^4) + c) \\ &= fl(fl(-0.0004 \cdot 10^4) + c) \\ &= fl(-0.4000 \cdot 10^1 + 0.3456 \cdot 10^1) \\ &= fl(-0.0544 \cdot 10^1) \\ &= -0.5440 \cdot 10^0 \end{aligned}$$

Por otro lado,

$$\begin{aligned} fl(a + fl(b + c)) &= fl(a + fl(-0.9880 \cdot 10^4 + 0.0003456 \cdot 10^4)) \\ &= fl(a - fl(0.9876544 \cdot 10^4)) \\ &= fl(0.9876 \cdot 10^4 - 0.9877 \cdot 10^4) \\ &= fl(-0.0001 \cdot 10^4) \\ &= -0.1000 \cdot 10^1 \end{aligned}$$

---

### **Observaciones de implementación:**

1. dado que en una implementación o programa se realizan muchas operaciones, cada una con su correspondiente error, es conveniente prestar atención en las operaciones que se realizan;

2. si  $x$  e  $y$  son números reales y en una programa se tiene una sentencia del tipo

```
if x == y then . . .
```

es más conveniente reemplazarla por una sentencia del tipo

```
if (abs(x-y)) < epsilon then . . .
```

para algún valor de  $\text{epsilon}$  dado por el usuario, puesto que es casi imposible que se verifique la primera sentencia.

# Clase 4 - Solución de ecuaciones no lineales

## El problema

Dada  $f : \mathbb{R} \rightarrow \mathbb{R}$  no lineal, se desea encontrar una solución  $r$  de la ecuación

$$f(x) = 0.$$

**Idea:** comenzando con algún  $x_0 \in \mathbb{R}$ , generar una sucesión  $\{x_k\}$  a través de un algoritmo numérico iterativo, y se espera que tal sucesión converja a  $r$  donde  $f(r) = 0$ .

## Método de bisección

Este método se basa fuertemente en el teorema del valor intermedio: si  $f$  es continua en  $[a, b]$  y si  $f(a)f(b) < 0$ , entonces  $f$  debe tener una raíz en  $(a, b)$ .

### Idea del método de bisección

Si  $f(a)f(b) < 0$ , se calculan  $c = \frac{a+b}{2}$  y  $f(c)$ . Sean  $x_0 = c$ : una aproximación a la raíz  $r$  de  $f$  y  $|e_0| = |x_0 - r| \leq \frac{b-a}{2}$ : error de aproximación inicial. Se tienen 3 posibilidades:

1. si  $f(a)f(c) < 0$ , entonces hay una raíz en el intervalo  $[a, c]$ . Reasignamos  $b \leftarrow c$  y se repite el procedimiento en el nuevo intervalo  $[a, b]$ .

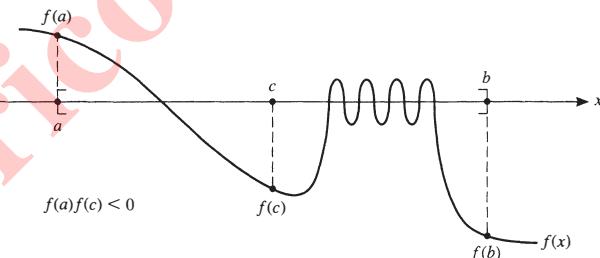


Figura 1: Caso:  $f(a)f(c) < 0$

2. si  $f(a)f(c) > 0$ , entonces hay una raíz en el intervalo  $[c, b]$ . Reasignamos  $a \leftarrow c$  y se repite el procedimiento en el nuevo intervalo  $[a, b]$ .

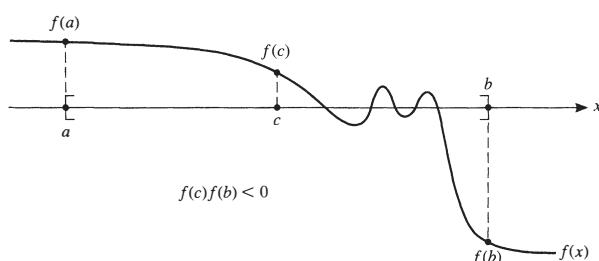


Figura 2: Caso:  $f(a)f(c) > 0$

3. si  $f(a)f(c) = 0$ , entonces  $f(c) = 0$  y  $x_0 = c$  es la raíz buscada.

Este caso es casi imposible en la práctica debido a los errores de redondeo. Por lo tanto, el criterio de parada no dependerá de que  $f(c) = 0$  sino de que  $|f(c)| < TOL$ , donde  $TOL$  es una tolerancia dada por el usuario.

Veamos algunos comentarios de implementación antes de dar el algoritmo.

- Por cuestiones numéricas, en vez de calcular el punto medio haciendo  $c \leftarrow (a + b)/2$ , es más conveniente calcular  $c \leftarrow a + (b - a)/2$ .
- Para determinar el cambio de signo de la función, en vez de analizar si  $f(a)f(c) < 0$ , es más conveniente usar la función  $sign$ , y analizar si  $sign(f(a)) \neq sign(f(b))$ .
- Se utilizan 3 criterios de parada en el algoritmos:
  1. el número máximo de pasos permitidos ( $M$ );
  2. el error en la variable es suficientemente pequeño ( $\delta$ );
  3. el valor de  $|f(c)|$  es suficientemente pequeño ( $\varepsilon$ ).

Se pueden construir ejemplos patológicos y simples donde uno de los criterios vale y el otro no. (Ver Figuras (3 y 4)). Para que el algoritmo sea más robusto se utilizan los tres criterios de parada. De todos modos, siempre es conveniente revisar que se cumplan las hipótesis para que se pueda aplicar el método y el algoritmo y analizar los resultados obtenidos.

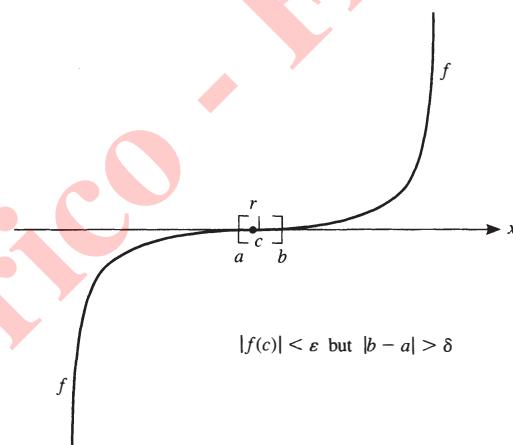


Figura 3: Ejemplo 1

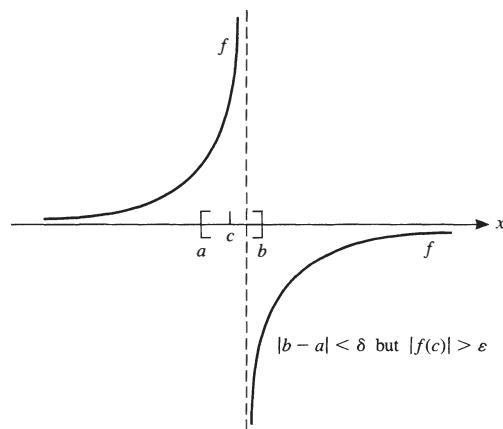


Figura 4: Ejemplo 2

---

## Algoritmo de bisección

Dados los siguientes datos de entrada y parámetros algorítmicos:  $a$  y  $b$  extremos izquierdo y derecho del intervalo,  $M$  el máximo número de pasos (iteraciones) permitidas,  $\delta$  la tolerancia para el error  $e$  (en la variable  $x$ ) y  $\varepsilon$  la tolerancia para los valores funcionales.

```
input  $a, b, M, \delta, \varepsilon$ 
 $u \leftarrow f(a)$ 
 $v \leftarrow f(b)$ 
 $e \leftarrow b - a$ 
output  $a, b, u, v$ 
if  $\text{sign}(u) = \text{sign}(v)$  then STOP (1)
for  $k = 1, 2, \dots, M$  do
     $e \leftarrow e/2$ 
     $c \leftarrow a + e$ 
     $w \leftarrow f(c)$ 
    output  $k, c, w, e$ 
    if  $|e| < \delta$  or  $|w| < \varepsilon$  then STOP (2)
    if  $\text{sign}(w) \neq \text{sign}(u)$  then
         $b \leftarrow c$ 
         $v \leftarrow w$ 
    else
         $a \leftarrow c$ 
         $u \leftarrow w$ 
    endif
end
```

### Observaciones:

1. En el algoritmo aparecen dos paradas (STOP). La primera detecta que los signos en los extremos del intervalo inicial son iguales y por lo tanto no puede continuar el algoritmo. Notar que el algoritmo no puede chequear continuidad (ver Figura (4)). La segunda parada detecta que alguno de los criterios de parada adoptados se cumple.
2. Se podría usar otro criterio de parada basado en que el error relativo sea pequeño en lugar de considerar el error absoluto.
3. El algoritmo *encuentra* una raíz de  $f$  en  $[a, b]$ , aunque pueden existir varias raíces. Para localizar alguna en particular se debe elegir el intervalo inicial cuidadosamente.

El siguiente teorema da un resultado de convergencia del método. Por un lado muestra que el método es **global**, en el sentido que si se cumplen las hipótesis del teorema del valor intermedio, el algoritmo encuentra una solución independiente del tamaño del intervalo. Por otro lado, de la demostración se puede deducir que la sucesión generada por este algoritmo **converge linealmente**.

**Teorema 1.** Si  $[a_0, b_0], [a_1, b_1], \dots, [a_n, b_n], \dots$  denotan los sucesivos intervalos en el método de bisección, entonces existen los límites:  $\lim_{n \rightarrow \infty} a_n$  y  $\lim_{n \rightarrow \infty} b_n$ , son iguales y representan una raíz de  $f$ . Si  $c_n = \frac{1}{2}(a_n + b_n)$  y  $r = \lim_{n \rightarrow \infty} c_n$ , entonces  $|r - c_n| \leq \frac{1}{2^{n+1}}(b_0 - a_0)$ .

*Demostración.* Si  $[a_0, b_0], [a_1, b_1], \dots, [a_n, b_n], \dots$  son los intervalos generados en el método de bisección, se tiene que

$$a_0 \leq a_1 \leq a_2 \leq \dots \leq b_0$$

$$b_0 \geq b_1 \geq b_2 \geq \dots \geq a_0.$$

Luego  $\{a_n\}$  es una sucesión creciente y acotada superiormente, entonces  $\{a_n\}$  es convergente. Análogamente,  $\{b_n\}$  es una sucesión decreciente y acotada inferiormente, por lo tanto también es convergente. Además,

$$b_{n+1} - a_{n+1} = \frac{1}{2}(b_n - a_n), \quad n \geq 0,$$

y aplicando esta ecuación repetidamente, se obtiene que

$$b_n - a_n = \frac{1}{2^n}(b_0 - a_0), \quad n \geq 0.$$

Luego

$$\lim_{n \rightarrow \infty} b_n - \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} (b_n - a_n) = \lim_{n \rightarrow \infty} \frac{1}{2^n}(b_0 - a_0) = (b_0 - a_0) \lim_{n \rightarrow \infty} \frac{1}{2^n} = 0.$$

Sea  $r = \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n$ . Luego tomando límite cuando  $n$  tiende a  $\infty$  en la desigualdad  $f(a_n)f(b_n) < 0$ , se obtiene que  $(f(r))^2 \leq 0$ . De allí que  $(f(r))^2 = 0$ , y en consecuencia  $f(r) = 0$ , es decir,  $r$  es una raíz de  $f$ .

Por último sea  $c_n = \frac{1}{2}(a_n + b_n)$ . Luego,

$$|r - c_n| \leq \frac{1}{2}|b_n - a_n| \leq \frac{1}{2^{n+1}}(b_0 - a_0).$$

□

**Ejemplo:** Supongamos que se aplica el método de bisección para determinar una raíz  $r$  de una función  $f$  en el intervalo  $[50, 63]$ .

- ¿Cuántos pasos se deberían realizar con el algoritmo de bisección si se desea obtener una raíz en el intervalo  $[50, 63]$  con una precisión absoluta de  $10^{-6}$ ?
- ¿Cuántos pasos se deberían realizar con el algoritmo de bisección si se desea obtener una raíz en el intervalo  $[50, 63]$  con una precisión relativa de  $10^{-6}$ ?

---

Para responder a), y usando el Teorema anterior se tiene que

$$|r - c_n| \leq \frac{13}{2^{n+1}} \leq 10^{-6},$$

y por lo tanto  $13 \cdot 10^6 \leq 2^{n+1}$ . Luego tomando logaritmo natural a ambos lados, se tiene que

$$(n+1) \ln 2 \geq \ln(13) + 6 \ln(10),$$

de donde se deduce que  $n \geq 23$ .

Para el caso b), también se usa el teorema anterior y que  $50 \leq r \leq 63$ , entonces

$$\frac{|r - c_n|}{|r|} \leq \frac{1}{2^{n+1}} \frac{13}{|r|} \leq \frac{13}{50 \cdot 2^{n+1}} \leq 10^{-6}.$$

Luego  $2^{n+1} \geq \frac{13}{50} \cdot 10^6$ , y aplicando logaritmo natural a ambos miembros de la desigualdad se deduce que  $n \geq 17$ .

# Clase 5 - Solución de ecuaciones no lineales (2)

## El problema

Recordemos el problema que comenzamos a estudiar en la clase anterior: dada  $f : \mathbb{R} \rightarrow \mathbb{R}$  no lineal, se desea hallar una solución  $r$  de la ecuación

$$f(x) = 0. \quad (1)$$

Al igual que antes, estudiaremos otro método iterativo para resolver este problema, el cual genere una sucesión de aproximaciones que se espera que converja a la solución buscada  $r$ .

## Idea del método de Newton

Dado que en general la función  $f$  es no lineal, resolver la ecuación (1) es un **problema difícil**. La idea del método de Newton consiste en reemplazar la resolución de este problema difícil por la resolución de una sucesión de **problemas fáciles**, cuyas soluciones convergen a la solución del problema difícil, bajo adecuadas hipótesis.

Supongamos que  $r$  es una raíz de  $f$  y que  $x$  es una aproximación de  $r$  a una distancia  $h$ , es decir,  $r = x + h$  o equivalentemente  $h = r - x$ . Supongamos también que  $f''$  existe y es continua en un entorno de  $x$  que contiene a  $r$ , entonces haciendo el desarrollo de Taylor de  $f$  centrado en  $x$ , tenemos que

$$0 = f(r) = f(x + h) = f(x) + f'(x)h + \mathcal{O}(h^2).$$

Cuando la aproximación  $x$  es próxima a  $r$ , el valor de  $h$  es pequeño y por lo tanto  $h^2$  es más pequeño aún y así podríamos despreciar el término  $\mathcal{O}(h^2)$

$$0 = f(x) + hf'(x),$$

para despejar  $h$

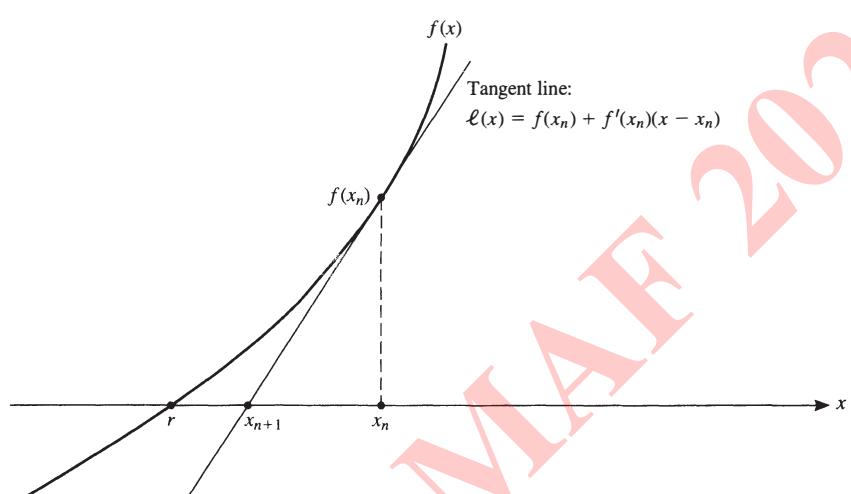
$$h = -\frac{f(x)}{f'(x)}.$$

Ahora bien, es claro que no conocemos exactamente  $h$  pues si tuviéramos  $x$  y  $h$  tendríamos la solución  $r$ . Sin embargo, si la aproximación  $x$  está cerca de  $r$ , entonces la nueva aproximación  $(x + h) = x - \frac{f(x)}{f'(x)}$  debería estar aún más cerca de  $r$ .

Entonces, comenzando con una aproximación  $x_0$  de  $r$ , la iteración del método de Newton consiste en calcular

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

Gráficamente, dado el punto  $(x_n, f(x_n))$  la idea consiste en aproximar el gráfico de la función  $f$  por la recta tangente a  $f$  que pasa por  $(x_n, f(x_n))$ . Tal recta tangente está dada por  $l_n(x) = f'(x_n)(x - x_n) + f(x_n)$ . Ver Figura (1).



**Figura 1:** Interpretación geométrica del método de Newton.

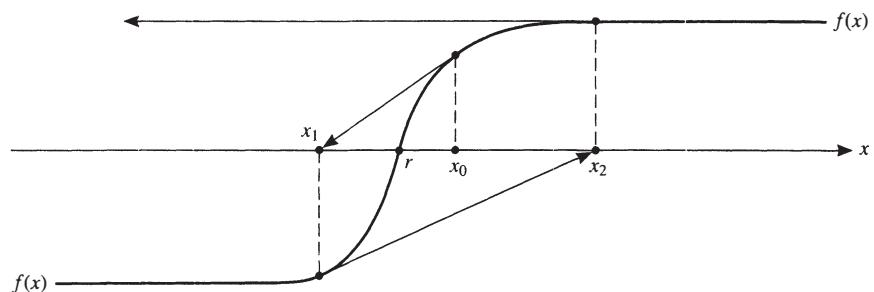
Así, en lugar de buscar la raíz de  $f$  (**problema difícil**), se calcula la raíz de  $l_n$ , es decir, se busca  $x_{n+1}$  solución de  $l_n(x) = 0$ . Es decir:

$$l(x_{n+1}) = f'(x_n)(x_{n+1} - x_n) + f(x_n) = 0,$$

y por lo tanto

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

Existen ejemplos donde esta idea podría fallar como se puede ver en la Figura (2). Mirando el gráfico se puede deducir que la aproximación inicial  $x_0$  debe estar suficientemente cerca de la raíz  $r$  para que el método converja.



**Figura 2:** Ejemplo donde el método de Newton podría diverger.

---

## Algoritmo del método de Newton

Dados los siguientes datos de entrada y parámetros algorítmicos: la aproximación inicial  $x_0$ , el número máximo de iteraciones permitido  $M$ ,  $\delta$  la tolerancia para el error  $e$  (en la variable  $x$ ) y  $\varepsilon$  la tolerancia para los valores funcionales.

```
input  $x_0, M, \delta, \varepsilon$ 
 $v \leftarrow f(x_0)$ 
output  $0, x_0, v$ 
if  $|v| < \varepsilon$  then STOP (1)
for  $k = 1, 2, \dots, M$  do
     $x_1 \leftarrow x_0 - v/f'(x_0)$ 
     $v \leftarrow f(x_1)$ 
output  $k, x_1, v$ 
if  $|x_1 - x_0| < \delta$  or  $|v| < \varepsilon$  then STOP (2)
     $x_0 \leftarrow x_1$ 
end
```

### Observaciones:

1. En el algoritmo aparecen dos paradas (STOP). La primera detecta que el punto inicial satisface la tolerancia del valor funcional. La segunda detecta que alguno de los criterios de parada considerados se cumple. Además el algoritmo podría parar por la cantidad máxima de iteraciones permitida.
2. Se podría usar otro criterio de parada basado en que el error relativo sea pequeño en lugar de considerar el error absoluto.
3. El algoritmo requiere subprogramas o procedimientos externos que calculen  $f(x)$  y  $f'(x)$ .
4. sería conveniente controlar en el programa que  $f'(x_0) \neq 0$  en cada iteración.

## Análisis de errores

El siguiente resultado muestra que bajo ciertas hipótesis la sucesión generada por el método de Newton converge **local** y **cuadráticamente**, es decir que si se comienza con una aproximación suficientemente próxima a la solución el método converge cuadráticamente.

**Teorema 1.** Si  $f''$  es continua en un entorno de una raíz  $r$  de  $f$  y si  $f'(r) \neq 0$  entonces existe  $\delta > 0$  tal que si el punto inicial  $x_0$  satisface  $|r - x_0| \leq \delta$  luego todos los puntos de la sucesión  $\{x_n\}$  generados por el algoritmo del método de Newton satisfacen que  $|r - x_n| \leq \delta$  para todo  $n$ , la sucesión  $\{x_n\}$  converge a  $r$  y la convergencia es cuadrática, i.e., existen una

constante  $c = c(\delta)$  y un natural  $N$  tal que

$$|r - x_{n+1}| \leq c|r - x_n|^2, \quad \text{para } n \geq N.$$

*Demostración.* Consideremos el error en la aproximación  $x_n$  en la iteración  $n$ :

$$e_n = r - x_n. \tag{2}$$

El error en la iteración  $(n + 1)$  es:

$$e_{n+1} = r - x_{n+1} = r - \left( x_n - \frac{f(x_n)}{f'(x_n)} \right) = r - x_n + \frac{f(x_n)}{f'(x_n)} = e_n + \frac{f(x_n)}{f'(x_n)},$$

Luego

$$e_{n+1} = \frac{e_n f'(x_n) + f(x_n)}{f'(x_n)} \tag{3}$$

Si escribimos el desarrollo de Taylor de  $f$  alrededor de  $x_n$  tenemos que

$$f(x_n + h) = f(x_n) + f'(x_n)h + \frac{1}{2}f''(\xi_n)h^2,$$

para algún  $\xi_n$  entre  $x_n$  y  $x_n + h$ .

Tomando  $h = e_n = r - x_n$ , se tiene que  $x_n + h = r$  y luego

$$0 = f(r) = f(x_n) + f'(x_n)h + \frac{1}{2}f''(\xi_n)h^2,$$

para algún  $\xi_n$  entre  $x_n$  y  $r$ . Luego

$$e_n f'(x_n) + f(x_n) = -\frac{1}{2}f''(\xi_n)h^2. \tag{4}$$

Usando (3) y (4) obtenemos que

$$e_{n+1} = -\frac{1}{2} \frac{f''(\xi_n)}{f'(x_n)} e_n^2. \tag{5}$$

Para acotar esta expresión definimos la siguiente función para  $\delta > 0$

$$c(\delta) = \frac{1}{2} \frac{\max_{|x-r| \leq \delta} |f''(x)|}{\min_{|x-r| \leq \delta} |f'(x)|}.$$

Como  $f'$  y  $f''$  son funciones continuas entonces  $|f'|$  y  $|f''|$  alcanzan sus valores extremos en un intervalo cerrado y acotado alrededor de  $r$ . Luego, dado  $\delta > 0$ , para todo par  $x$  y  $\xi$  tal que  $|\xi - r| \leq \delta$  y  $|x - r| \leq \delta$  existe una constante  $c = c(\delta)$  tal que

$$\frac{1}{2} \left| \frac{f''(\xi)}{f'(x)} \right| \leq c(\delta).$$

Ahora notemos que si  $\delta \rightarrow 0$  entonces  $c(\delta) \rightarrow \frac{1}{2} \left| \frac{f''(r)}{f'(r)} \right|$ , que es un número finito porque  $f'(r) \neq 0$  y por lo tanto  $\delta c(\delta) \rightarrow 0$  cuando  $\delta \rightarrow 0$ . Entonces elegimos  $\delta$  suficientemente pequeño tal que  $\rho = \delta c(\delta) < 1$ .

Supongamos que la aproximación inicial  $x_0$  es tal que  $e_0 = |x_0 - r| \leq \delta$ , y como  $\xi_0$  está entre  $x_0$  y  $r$ , entonces  $|\xi_0 - r| \leq \delta$ . Luego por la definición de  $c(\delta)$  tenemos que

$$\frac{1}{2} \left| \frac{f''(\xi_0)}{f'(x_0)} \right| \leq c(\delta).$$

Finalmente por (5)

$$|x_1 - r| = |e_1| = \frac{1}{2} \left| \frac{f''(\xi_0)}{f'(x_0)} \right| |e_0|^2 \leq c(\delta) |e_0|^2 = c(\delta) |e_0| |e_0| \leq c(\delta) \delta |e_0| = \rho |e_0| < |e_0| \leq \delta.$$

Esto dice que  $x_1$  está a una distancia de  $r$  menor que  $\delta$  y además que si la sucesión converge, lo hace cuadráticamente. Repitiendo este argumento obtenemos que

$$\begin{aligned} |e_1| &\leq \rho |e_0| \\ |e_2| &\leq \rho |e_1| \leq \rho^2 |e_0| \\ |e_3| &\leq \rho |e_2| \leq \rho^3 |e_0| \\ &\vdots \end{aligned}$$

En general, tenemos que

$$|e_n| \leq \rho^n |e_0|.$$

Como  $0 < \rho < 1$ , entonces  $\lim_{n \rightarrow \infty} \rho^n = 0$  y por lo tanto  $\lim_{n \rightarrow \infty} e_n = 0$  y así  $\lim_{n \rightarrow \infty} x_n = r$ .

□

**Observación:** el método de Newton tiene convergencia cuadrática local para determinar raíces simples, bajo adecuadas hipótesis.

El siguiente resultado que sólo enunciaremos aunque su demostración puede consultarse en la Bibliografía muestra que bajo hipótesis de convexidad en todo el dominio el método de Newton converge independientemente del punto inicial.

**Teorema 2.** Si  $f''$  es continua en  $\mathbb{R}$ ,  $f$  es creciente y convexa en  $\mathbb{R}$  y tiene una raíz, entonces esa raíz es única y la iteración de Newton convergerá a esa raíz independientemente del punto inicial  $x_0$ .

**Ejercicio:** hallar una función que cumpla las hipótesis del teorema. Analizar la utilidad práctica del teorema.

**Aplicación del método de Newton:** Sea  $R > 0$  y  $x_* = \sqrt{R}$ , entonces  $x_*$  es una raíz de  $x^2 - R = 0$ . Si se aplica el método de Newton a  $f(x) = x^2 - R$ , se obtiene la siguiente fórmula de iteración:

$$x_{n+1} = \frac{1}{2} \left( x_n + \frac{R}{x_n} \right).$$

Por ejemplo si  $R = 17$  y se comienza con  $x_0 = 4$ , se obtienen

$$\begin{aligned} x_1 &= 4.12 \\ x_2 &= 4.123106 \\ x_3 &= 4.1231056256177 \\ x_4 &= 4.123105625617660549821409856, \end{aligned}$$

que tiene 28 dígitos correctos.

## Clase 6 - Solución de ecuaciones no lineales (3)

### Método de la secante

Continuamos con el mismo problema de resolver una ecuación no lineal. Hasta ahora vimos el método de bisección y el método de Newton. Recordemos que la iteración del método de Newton está dada por:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad \text{para } n \geq 0. \quad (1)$$

Si bien este método tiene convergencia cuadrática local, tiene como desventaja que requiere la evaluación de la derivada de la función  $f$  en cada iteración. Uno de los métodos más conocidos que evita esto es el **método de la secante**.

La idea del método de la secante consiste en reemplazar  $f'(x_n)$  en la iteración de Newton (1) por una aproximación dada por el cociente incremental, dado por la pendiente de la recta secante que pasa por los puntos  $(x_n, f(x_n))$  y  $(x_n + h, f(x_n + h))$ :

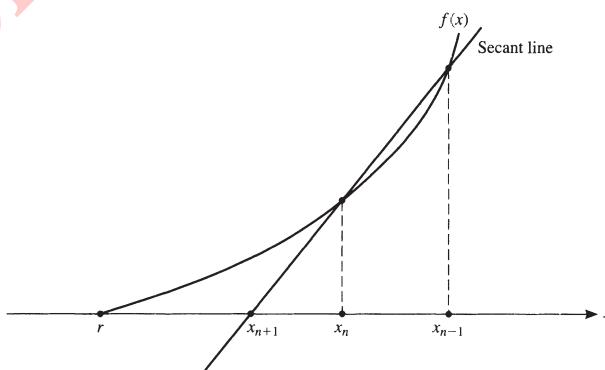
$$f'(x_n) \approx a_n = \frac{f(x_n + h) - f(x_n)}{h},$$

para algún  $h$  suficientemente pequeño.

Para evitar evaluar  $f$  en un punto adicional  $(x_n + h)$  se elige  $h = x_{n-1} - x_n$ , entonces:

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

Ver Figura 1.



**Figura 1:** Interpretación gráfica del método secante.

Así, la iteración del método secante consiste en:

$$x_{n+1} = x_n - \frac{f(x_n)}{\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}} \quad \text{para } n \geq 1,$$

es decir,

$$x_{n+1} = x_n - f(x_n) \left[ \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \right] \quad \text{para } n \geq 1. \quad (2)$$

---

## Algoritmo del método de la secante

Dados los siguientes datos de entrada y parámetros algorítmicos:  $a$  la penúltima aproximación de  $r$ ,  $b$  la última aproximación de  $r$ , el número máximo de iteraciones permitido  $M$ ,  $\delta$  la tolerancia para el error  $e$  (en la variable  $x$ ) y  $\varepsilon$  la tolerancia para los valores funcionales.

```
input  $a, b, M, \delta, \varepsilon$ 
 $fa \leftarrow f(a)$ 
 $fb \leftarrow f(b)$ 
output  $0, a, fa$ 
output  $1, b, fb$ 
for  $k = 2, 3, \dots, M$  do
    if  $|fa| < |fb|$  do
         $a \leftrightarrow b; fa \leftrightarrow fb$ 
    endif
     $s \leftarrow (b - a)/(fb - fa)$ 
     $b \leftarrow a$ 
     $fb \leftarrow fa$ 
     $a \leftarrow a - fa * s$ 
     $fa \leftarrow f(a)$ 
    output  $k, a, fa$ 
    if  $|b - a| < \delta$  or  $|fa| < \varepsilon$  then STOP
end do
```

### Observaciones:

- 1. En el algoritmo los puntos  $a$  y  $b$  pueden intercambiarse para lograr que  $|f(b)| \leq |f(a)|$ . Así, para el par  $\{x_{n-1}, x_n\}$  se satisface que  $|f(x_n)| \leq |f(x_{n-1})|$ , y para el par siguiente  $\{x_n, x_{n+1}\}$  se tiene que  $|f(x_{n+1})| \leq |f(x_n)|$ . Esto garantiza que la sucesión  $\{|f(x_n)|\}$  es no creciente.
- 2. El algoritmo se detiene por el número máximo de iteraciones permitidas, por satisfacer la tolerancia para los valores funcionales, o por la tolerancia en la diferencia de dos aproximaciones sucesivas.
- 3. En cuanto al análisis de errores, es posible probar que:

$$e_{n+1} \approx ce_n^\alpha = (ce_n^{\alpha-1})e_n^1,$$

donde  $\alpha = (1 + \sqrt{5})/2 = 1.618334\dots$  Como  $1 < \alpha < 2$  se dice que el método de la secante tiene **convergencia superlineal**. Además, por recurrencia

$$e_{n+1} \approx ce_n^\alpha \approx c^{1+\alpha} e_{n-1}^{\alpha^2}$$

donde  $\alpha^2 = (1 + \sqrt{5})^2/4 = (3 + \sqrt{5})/2 = 2.618334\dots$ , esto dice que **dos iteraciones de método de la secante** es mejor que **una iteración del método de Newton**.

## Iteración de punto fijo

En esta sección veremos que es un punto fijo de una función dada, como encontrarlos numéricamente y cuál es la conexión con el problema de determinar raíces de funciones.

**Definición 1.** *un punto fijo de una función  $g$  es un número  $p$ , en el dominio de  $g$ , tal que  $g(p) = p$ .*

Por un lado, si  $p$  es una raíz de una función  $f$ , esto es,  $f(p) = 0$ , entonces es posible definir diferentes funciones  $g$  con un punto fijo en  $p$ , por ejemplo:  $g(x) = x - f(x)$ , o  $g(x) = x + 3f(x)$ .

Por otro lado, si  $g$  tiene un punto fijo en  $p$ , esto es,  $g(p) = p$ , entonces la función  $f(x) = x - g(x)$  tiene una raíz en  $p$ .

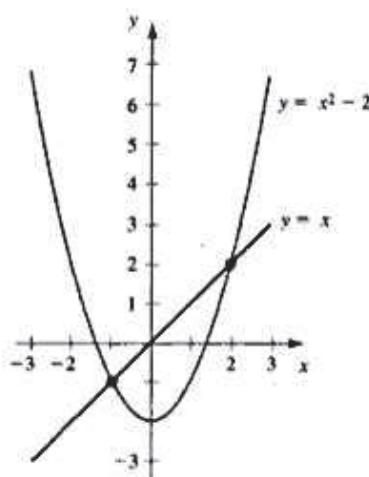
Aunque estamos interesados en el problema de determinar soluciones de una ecuación no lineal, o equivalentemente, encontrar raíces de funciones no lineales veremos que la forma de la iteración de punto es muy fácil de estudiar y analizar. Además algunas opciones de punto fijo dan origen a técnicas matemáticas y computacionales muy poderosas para determinar raíces.

**Ejemplo 1:** se busca determinar los posibles puntos fijos de la función  $g(x) = x^2 - 2$  en el intervalo  $[-2, 3]$ .

Para esto se plantea la ecuación  $g(x) = x$ , es decir, se debe resolver la ecuación cuadrática  $x^2 - 2 = x$ , o sea,  $x^2 - x - 2 = 0$ . Luego es fácil verificar que  $x = -1$  y  $x = 2$  son puntos fijos ( $g(p) = p$ ) pues

$$g(-1) = (-1)^2 - 2 = -1 \quad \text{y} \quad g(2) = (2)^2 - 2 = 2.$$

Ver Figura 2.



**Figura 2:** Puntos fijos de  $g$ .

A continuación veremos un resultado que da condiciones suficientes para la existencia y unicidad del punto fijo.

## Teorema 1.

1. Si  $g \in C[a, b]$  (es decir,  $g$  es una función continua en el intervalo  $[a, b]$ ) y  $g(x) \in [a, b]$  para todo  $x \in [a, b]$  entonces existe  $p \in [a, b]$  tal que  $g(p) = p$ . (**EXISTENCIA**)
2. Si además existe  $g'(x)$  para todo  $x \in (a, b)$  y existe una constante positiva  $k < 1$  tal que  $|g'(x)| \leq k$  para todo  $x \in (a, b)$ , entonces el punto fijo en  $(a, b)$  es único. (Ver Figura 3). (**UNICIDAD**)

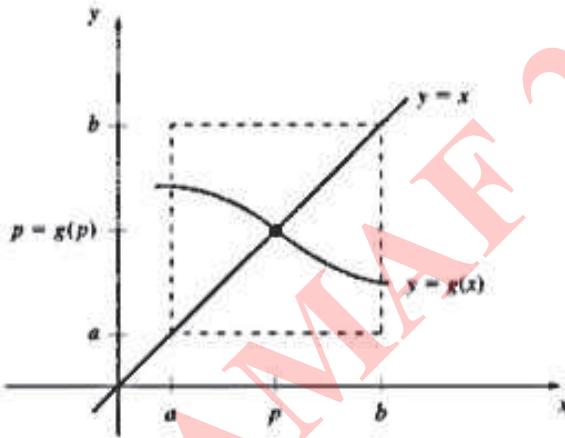


Figura 3: Unidadad del punto fijo.

*Demostración.*

1. Si  $g(a) = a$  o  $g(b) = b$ , el punto fijo está en uno de los extremos del intervalo y ya estaría probado.

Supongamos que esto **no** es cierto, entonces  $g(a) > a$  y  $g(b) < b$ . Sea  $h(x) = g(x) - x$  una función definida en  $[a, b]$ , que además es continua en  $[a, b]$  pues  $g$  y  $x$  lo son y resta de funciones continuas es una función continua. Además, por lo anterior, tenemos que

$$h(a) = g(a) - a > 0 \quad \text{y} \quad h(b) = g(b) - b < 0.$$

Entonces, por el Teorema del Valor Intermedio, existe  $p \in (a, b)$  tal que  $h(p) = 0$ , esto es,  $g(p) = p$  y por lo tanto  $p$  es un punto fijo de  $g$ .

2. Ahora supongamos que existen dos puntos fijos distintos  $p$  y  $q$  en  $[a, b]$ , es decir,  $p, q \in [a, b]$ ,  $p \neq q$  tal que  $g(p) = p$  y  $g(q) = q$ . Por el Teorema del Valor Medio existe  $\xi$  entre  $p$  y  $q$ , y por lo tanto en  $[a, b]$  tal que

$$g(p) - g(q) = g'(\xi)(p - q).$$

Luego usando la hipótesis que  $|g'(x)| \leq k < 1$  tenemos que

$$|p - q| = |g(p) - g(q)| = |g'(\xi)||p - q| \leq k|p - q| < |p - q|,$$

lo que es una contradicción que provino de suponer que habrían dos puntos fijos distintos en  $[a, b]$ , y por lo tanto el punto fijo es único.

□

Analicemos la existencia y unicidad de punto fijo en los siguientes ejemplos.

**Ejemplo 1:** considerar  $g(x) = \frac{(x^2 - 1)}{3} = \frac{x^2}{3} - \frac{1}{3}$  en el intervalo  $[-1, 1]$ .

Es fácil ver  $g$  tiene un mínimo absoluto en  $x = 0$  y  $g(0) = -1/3$ . Además tiene máximos absolutos en  $x = -1$  y  $x = 1$  donde  $g(-1) = 0$  y  $g(1) = 0$ . Además, claramente  $g$  es continua en  $[-1, 1]$  y  $|g'(x)| = \left| \frac{2}{3}x \right| \leq \frac{2}{3}$  para todo  $x \in [-1, 1]$ . Por lo tanto existe un único punto fijo  $p$  de  $g$  en el intervalo  $[-1, 1]$ . Para determinar  $p$ , planteamos

$$g(p) = p, \Rightarrow \frac{p^2 - 1}{3} = p, \Rightarrow p^2 - 3p - 1 = 0,$$

cuyas raíces son  $p_1 = \frac{(3 - \sqrt{13})}{2} = -0,302776\dots$  y  $p_2 = \frac{(3 + \sqrt{13})}{2} = 3,302776\dots$  El único punto fijo es claramente  $p_1$  pues este punto está en  $[-1, 1]$  en cambio  $p_2$  no. Ver Figura 4.

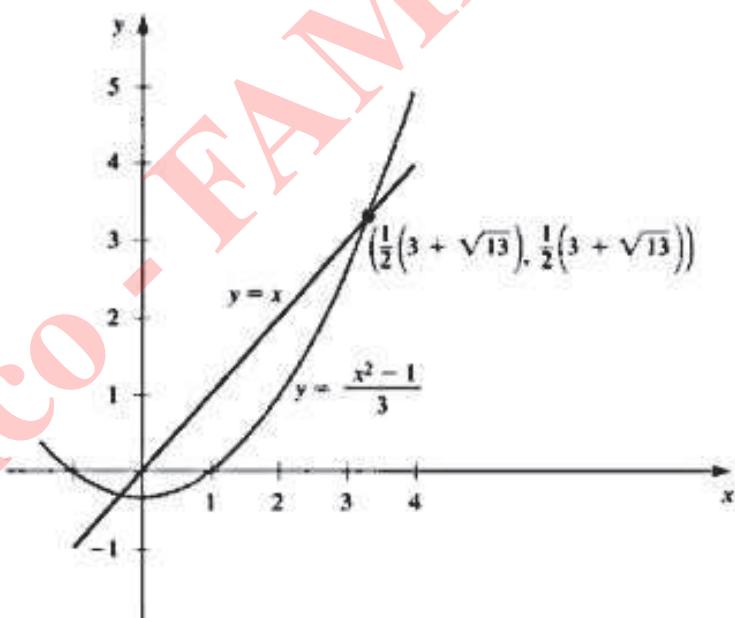


Figura 4: Puntos fijos del Ejemplo 1.

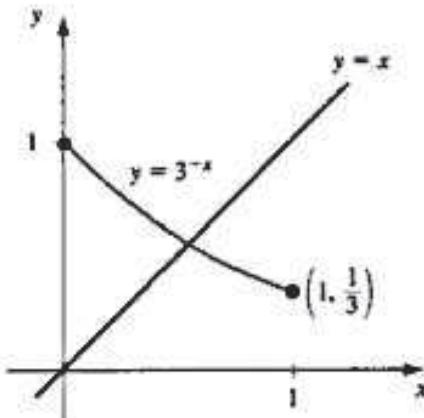
Notar en el gráfico que  $p_2$  también es un punto fijo de  $g$  en el intervalo  $[3, 4]$ . Sin embargo,  $g(4) = 5$  y  $g'(4) = \frac{8}{3} > 1$ , por lo que no se satisfacen las hipótesis del teorema anterior en el intervalo  $[3, 4]$ . Esto dice que tales hipótesis son condiciones suficientes para garantizar la existencia y unicidad de un punto fijo, pero no son necesarias.

**Ejemplo 2:** considerar  $g(x) = 3^{-x} = e^{-(\ln 3)x}$  en el intervalo  $[0, 1]$ .

Su derivada es  $g'(x) = -(\ln 3)e^{-(\ln 3)x} = -(\ln 3)3^{-x} < 0$ , por lo tanto  $g$  es decreciente en el intervalo  $[0, 1]$ . Entonces,

$$g(0) = 1 \geq g(x) \geq \frac{1}{3} = g(1),$$

y por el teorema anterior, existe un punto fijo en  $[0, 1]$ . Por otro lado,  $g'(0) = -\ln 3 = -1.0986\dots$  y por lo tanto no se puede usar el teorema anterior para garantizar unicidad pues  $|g'(x)| \not< 1$  en el intervalo  $(0, 1)$ . Ver Figura 5.



**Figura 5:** Puntos fijos del Ejemplo 1.

### Idea del algoritmo de punto fijo

Para calcular aproximadamente el punto fijo de una función  $g$  primero se inicia con una aproximación inicial  $p_0$  y calculando  $p_n = g(p_{n-1})$  para  $n \geq 1$  se obtiene una sucesión de aproximaciones  $\{p_n\}$ . Si la función  $g$  es continua y la sucesión converge entonces lo hace a un punto fijo  $p$  de  $g$  pues

$$p = \lim_{n \rightarrow \infty} p_n = \lim_{n \rightarrow \infty} g(p_{n-1}) = g(\lim_{n \rightarrow \infty} p_{n-1}) = g(p).$$

### Algoritmo del método de punto fijo

Dados los siguientes datos de entrada y parámetros algorítmicos:  $p_0$  una aproximación inicial, el número máximo de iteraciones permitido  $M$  y  $\delta$  la tolerancia para el error  $e$  (en la variable  $x$ )

```

input  $p_0, M, \delta$ 
output  $0, p_0$ 
 $i \leftarrow 1$ 
while  $i \leq M$  then do
     $p \leftarrow g(p_0)$ 
    output  $i, p$ 
    if  $|p - p_0| < \delta$  then STOP
     $i \leftarrow i + 1$ 
     $p_0 \leftarrow p$ 
end while
```

### Observaciones:

1. El algoritmo es muy fácil de implementar.
2. los criterios de parada utilizados son la distancia entre dos iteraciones sucesivas y el número máximo de iteraciones.

---

## Teorema 2.

Sea  $g \in C[a, b]$  tal que  $g(x) \in [a, b]$  para todo  $x \in [a, b]$ . Supongamos que existe  $g'(x)$  para todo  $x \in (a, b)$  y existe una constante positiva  $0 < k < 1$  tal que  $|g'(x)| \leq k$  para todo  $x \in (a, b)$ , entonces para cualquier  $p_0 \in [a, b]$  la sucesión definida por  $p_n = g(p_{n-1})$ , para  $n \geq 1$ , converge al único punto fijo  $p$  en  $(a, b)$ .

*Demostración.* Por el teorema anterior, se sabe que bajo estas hipótesis existe un único punto fijo  $p \in [a, b]$ . Como  $g(x) \in [a, b]$  para todo  $x \in [a, b]$ , la sucesión de aproximaciones  $\{p_n\}$  está bien definida para todo  $n$ , es decir,  $p_n \in [a, b]$  para todo  $n$ .

Para probar la convergencia se usará el Teorema del Valor Medio en lo siguiente

$$|p_n - p| = |g(p_{n-1}) - g(p)| = |g'(\xi_n)| |p_{n-1} - p| \leq k |p_{n-1} - p|,$$

luego, por recurrencia, se tiene que

$$|p_n - p| \leq k |p_{n-1} - p| \leq k^2 |p_{n-2} - p| \leq \cdots \leq k^n |p_0 - p|.$$

Como  $0 < k < 1$  entonces  $\lim_{n \rightarrow \infty} k^n = 0$ , luego

$$\lim_{n \rightarrow \infty} |p_n - p| \leq \lim_{n \rightarrow \infty} k^n |p_0 - p| = |p_0 - p| \lim_{n \rightarrow \infty} k^n = 0,$$

y por lo tanto, la sucesión  $\{p_n\}$  converge al punto fijo  $p$ .  $\square$

**Corolario 1.** Si  $g$  es una función que satisface las hipótesis del teorema anterior, se tienen las siguientes cotas de error

$$\begin{aligned} |p_n - p| &\leq k^n \max\{p_0 - a, b - p_0\} \\ |p_n - p| &\leq \frac{k^n}{1-k} |p_1 - p_0| \quad \text{para todo } n \geq 1 \end{aligned}$$

*Demostración.* La demostración puede consultarse en el libro de Burden-Faires.  $\square$

### Análisis de error en métodos de punto fijo

Supongamos que  $p$  es un punto fijo de una función  $g$  y que  $\{p_n\}$  es la sucesión, que converge a  $p$ , definida por  $p_{n+1} = g(p_n)$ .

Sea  $p_n$  una aproximación del punto fijo  $p$ , es decir  $p_n = p + h$ , y consideremos el desarrollo de Taylor de  $g$  centrado en  $p$ :

$$\begin{aligned} p_{n+1} = g(p_n) = g(p+h) &= g(p) + g'(p)(p_n - p) + \frac{g''(p)}{2}(p_n - p)^2 + \dots \\ &\quad + \frac{g^{(r-1)}(p)}{(r-1)!}(p_n - p)^{r-1} + \frac{g^{(r)}(\xi_n)}{r!}(p_n - p)^r, \end{aligned} \tag{3}$$

para algún  $\xi_n$  entre  $p_n$  y  $p$ .

Supongamos ahora que  $g'(p) = g''(p) = \dots = g^{(r-1)}(p) = 0$  pero  $g^{(r)}(p) \neq 0$ , entonces

$$e_{n+1} = p_{n+1} - p = g(p_n) - g(p) = \frac{g^{(r)}(\xi_n)}{r!}(p_n - p)^r = \frac{g^{(r)}(\xi_n)}{r!}e_n^r,$$

esto es,

$$e_{n+1} = \frac{g^{(r)}(\xi_n)}{r!}e_n^r,$$

y tomando límite se obtiene

$$\lim_{n \rightarrow \infty} \frac{|e_{n+1}|}{|e_n|^r} = \frac{|g^{(r)}(p)|}{r!} = C,$$

por lo tanto el método tiene orden de convergencia (al menos)  $r$ .

**Conclusión:** si las derivadas de la función de iteración de punto fijo se anulan en el punto fijo  $p$  hasta el orden  $(r-1)$  entonces el método tiene orden de convergencia (al menos)  $r$ .

Usando este resultado se obtienen tres corolarios interesantes que relacionan el método de Newton con el método de punto fijo para una función  $f$  que satisface las hipótesis del teorema de convergencia de Newton.

**Corolario 2.** Si  $f$  es una función que tiene una raíz simple  $p$ , entonces el método de Newton es un método de punto fijo y tiene orden de convergencia (al menos) 2.

*Demostración.* Sea  $g(x) = x - \frac{f(x)}{f'(x)}$ , la función de iteración del método de Newton. Es claro que si  $p$  es una solución de  $f(x) = 0$ , entonces  $p$  es un punto fijo de  $g$  pues

$$g(p) = p - \frac{f(p)}{f'(p)} = p,$$

ya que  $f(p) = 0$  y  $f'(p) \neq 0$ .

Ahora calculemos  $g'(x)$  y luego evaluemos en  $p$ :

$$g'(x) = 1 - \frac{(f'(x))^2 - f''(x)f(x)}{(f'(x))^2} = 1 - 1 + \frac{f''(x)f(x)}{(f'(x))^2} = \frac{f''(x)f(x)}{(f'(x))^2},$$

entonces

$$g'(p) = \frac{f''(p)f(p)}{(f'(p))^2} = 0,$$

y por lo tanto el método tiene orden de convergencia (al menos) 2.  $\square$

**Corolario 3.** Si  $p$  es una raíz de multiplicidad  $r \geq 2$  de  $f$ , entonces el método de Newton tiene orden 1.

*Demostración.* Ya vimos que si  $g(x) = x - \frac{f(x)}{f'(x)}$  entonces  $g'(x) = \frac{f(x)f''(x)}{(f'(x))^2}$ .

Ahora, supongamos que  $p$  es una raíz de multiplicidad  $r$  de  $f$ , esto es,  $f(x) = (x-p)^r h(x)$ , con  $h$  una función tal que  $h(p) \neq 0$  y  $r \geq 2$ .

La derivada primera de  $f$  es

$$f'(x) = r(x-p)^{r-1}h(x) + (x-p)^r h'(x) = (x-p)^{r-1} [rh(x) + (x-p)h'(x)],$$

y la derivada segunda de  $f$  es

$$\begin{aligned} f''(x) &= r(r-1)(x-p)^{r-2}h(x) + 2r(x-p)^{r-1}h'(x) + (x-p)^r h''(x), \\ &= (x-p)^{r-2} [r(r-1)h(x) + 2r(x-p)h'(x) + (x-p)^2 h''(x)]. \end{aligned}$$

Luego

$$g'(x) = \frac{(x-p)^r h(x)(x-p)^{r-2} [r(r-1)h(x) + 2r(x-p)h'(x) + (x-p)^2 h''(x)]}{(x-p)^{2r-2} [rh(x) + (x-p)h'(x)]^2},$$

entonces

$$g'(p) = \frac{h(p)r(r-1)h(p)}{r^2(h(p))^2} = \frac{r-1}{r} \neq 0,$$

pues  $r \geq 2$ . □

Por último, modificando levemente la función de iteración del método de Newton se puede recuperar la convergencia cuadrática aún en casos de raíces múltiples.

**Corolario 4.** Si  $p$  es una raíz de multiplicidad  $r \geq 2$  de  $f$ , entonces la siguiente modificación del método de Newton recupera la convergencia cuadrática:

$$x_{n+1} = x_n - r \frac{f(x_n)}{f'(x_n)}, \quad \text{esto es,} \quad g(x) = x - r \frac{f(x)}{f'(x)}.$$

*Demostración.* Usando las expresiones de  $f, f'$  y  $f''$  obtenidas en el corolario anterior, se tiene que

$$\begin{aligned} g'(x) &= 1 - r \frac{(f'(x))^2 - f''(x)f(x)}{(f'(x))^2} = 1 - r + r \frac{f(x)f''(x)}{(f'(x))^2} \\ &= 1 - r + r \frac{(x-p)^r h(x)(x-p)^{r-2} [r(r-1)h(x) + 2r(x-p)h'(x) + (x-p)^2 h''(x)]}{(x-p)^{2r-2} [rh(x) + (x-p)h'(x)]^2} \\ &= 1 - r + r \frac{h(x) [r(r-1)h(x) + 2r(x-p)h'(x) + (x-p)^2 h''(x)]}{[rh(x) + (x-p)h'(x)]^2}. \end{aligned}$$

Evaluando en el punto fijo  $p$  se obtiene

$$g'(p) = 1 - r + r \frac{h(p)r(r-1)h(p)}{r^2(h(p))^2} = 1 - r + (r-1) = 0,$$

y por lo tanto el método de Newton modificado tiene convergencia (al menos) cuadrática. □

# Clase 7 - Interpolación polinomial

## El problema

Dada una tabla de  $(n + 1)$  puntos:  $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$  donde  $x_0, x_1, \dots, x_n$  son distintos, se desea determinar un polinomio  $p$ , con el **menor grado posible**, tal que

$$p(x_i) = y_i \quad \text{para } i = 0, \dots, n.$$

En este caso se dice que tal polinomio  $p$  **interpola** el conjunto de puntos dados. Ver Figura 1.

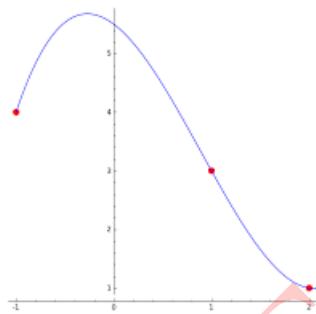


Figura 1: Interpolación polinomial.

A continuación veremos un resultado de existencia y unicidad del polinomio interpolante.

**Teorema 1.** *Dados  $x_0, x_1, \dots, x_n$  números reales distintos con valores asociados  $y_0, y_1, \dots, y_n$  entonces existe un único polinomio  $p_n$  de grado menor o igual a  $n$  tal que  $p_n(x_i) = y_i$ , para  $i = 0, \dots, n$ .*

*Demostración.*

**Existencia:** probaremos la existencia del polinomio interpolante  $p$  por inducción.

Para  $n = 0$ , es el caso obvio pues el polinomio constante  $p_0(x) = y_0$  satisface que tiene grado menor o igual a 0 y que  $p_0(x_0) = y_0$ .

Ahora supongamos, por hipótesis inductiva que se tiene un polinomio  $p_{k-1}$  de grado  $\leq k-1$  con  $p_{k-1}(x_i) = y_i$  para  $i = 0, \dots, k-1$ . Vamos a construir el polinomio  $p_k$  de grado  $\leq k$ , tal que  $p_k(x_i) = y_i$  para  $i = 0, \dots, k$ , de la forma

$$p_k(x) = p_{k-1}(x) + c(x - x_0)(x - x_1) \dots (x - x_{k-1}),$$

donde  $c$  es una constante a determinar. Notar que este polinomio tiene grado  $\leq k$ . Además  $p_k$  interpola los primeros  $k$  puntos que interpola  $p_{k-1}$  pues

$$p_k(x_i) = p_{k-1}(x_i) + c(x_i - x_0)(x_i - x_1) \dots (x_i - x_{k-1}) = y_i, \quad i = 0, \dots, k-1.$$

El coeficiente  $c$  se determina usando la condición de interpolación  $p_k(x_k) = y_k$ , es decir:

$$p_k(x_k) = p_{k-1}(x_k) + c(x_k - x_0)(x_k - x_1) \dots (x_k - x_{k-1}) = y_k,$$

de donde se deduce que

$$c = \frac{y_k - p_{k-1}(x_k)}{(x_k - x_0)(x_k - x_1) \dots (x_k - x_{k-1})}.$$

El coeficiente  $c$  está bien definido porque los números  $x_0, x_1, \dots, x_n$  son distintos y el denominador nunca se anula. Esto prueba la existencia del polinomio interpolante  $p_k$ .

**Unicidad:** supongamos que existen dos polinomios interpolantes  $p_n$  y  $q_n$  de grado  $\leq n$ , esto es,  $p_n(x_i) = y_i$  y  $q_n(x_i) = y_i$  para  $i = 0, \dots, n$ .

Sea  $h = p_n - q_n$ . Claramente  $h$  es un polinomio de grado  $\leq n$ . Además,  $h(x_i) = 0$  para  $i = 0, \dots, n$ , por lo tanto  $h$  es un polinomio de grado  $\leq n$  y tiene  $(n+1)$  raíces reales. Luego, por el teorema fundamental del Álgebra,  $h(x) = 0$  para todo  $x$  y por lo tanto  $p_n = q_n$ .

□

### Forma de Newton del polinomio interpolante

Si bien el polinomio interpolante es único, puede escribirse de diferentes formas. La forma de Newton se obtiene inductivamente, como se hizo en la prueba de la existencia del teorema anterior, agregando un nuevo término al polinomio interpolante de un grado menor.

**Si  $n = 0$ :** vimos que es suficiente definir el polinomio constante  $p_0(x) = c_0 = y_0$ .

**Si  $n = 1$ :** dados los puntos  $(x_0, y_0), (x_1, y_1)$ , se construye  $p_1$  tal que  $p_1(x) = c_0 + c_1(x - x_0)$  y  $p_1(x_0) = c_0 = y_0$ . Usando que  $p_1(x_1) = y_1$ , entonces  $y_1 = c_0 + c_1(x_1 - x_0)$  y por lo tanto,  $c_1 = \frac{y_1 - c_0}{x_1 - x_0}$ .

**Si  $n = k$ :** tenemos que

$$p_k(x) = p_{k-1}(x) + c_k(x - x_0)(x - x_1) \dots (x - x_{k-1}),$$

y por recurrencia obtenemos que

$$p_k(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \dots + c_k(x - x_0) \dots (x - x_{k-1}).$$

La forma de Newton compacta del polinomio interpolante resulta en

$$p_k(x) = \sum_{i=0}^k c_i \prod_{j=0}^{i-1} (x - x_j).$$

Aquí se adopta la convención que  $\prod_{j=0}^m (x - x_j) = 1$  si  $m < 0$ .

Para evaluar  $p_k(x)$ , una vez calculados los coeficientes  $c_k$ , es conveniente usar el algoritmo de Horner de multiplicación encajada.

### Forma de Lagrange del polinomio interpolante

Veremos otra forma alternativa de expresar el polinomio interpolante  $p_n$ , de grado  $\leq n$ , asociado a los puntos  $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$  donde  $x_0, x_1, \dots, x_n$  son distintos.

Primero se definen los polinomios básicos de Lagrange asociado a los puntos distintos  $x_0, x_1, \dots, x_n$ :

$$l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{(x - x_j)}{(x_i - x_j)} \quad \text{para } i = 0, \dots, n.$$

Así, por ejemplo, para  $i = 0$ , se tiene  $l_0(x) = \prod_{j=1}^n \frac{(x - x_j)}{(x_0 - x_j)}$ .

Es claro que el grado de  $l_i(x)$  es igual  $n$  para  $i = 0, \dots, n$  y que

$$l_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j. \end{cases}$$

Ahora definimos la **forma de Lagrange** del polinomio interpolante por

$$p_n(x) = \sum_{i=0}^n y_i l_i(x).$$

Es claro que  $p_n$  es un polinomio de grado  $\leq n$  y que  $p_n(x_i) = y_i$ , para  $i = 0, \dots, n$ .

**Ejercicio:** probar que  $\sum_{i=0}^n l_i(x) = 1$ .

Sea  $p(x) = \sum_{i=0}^n l_i(x)$ . Es claro que  $p$  tiene grado  $\leq n$ . Sea  $h(x) = p(x) - 1$ , un polinomio de grado  $\leq n$ . Además,  $h(x_k) = 0$  para  $k = 0, \dots, n$ , es decir,  $h$  es polinomio de grado  $\leq n$  y tiene  $n+1$  raíces. Luego  $h(x) = 0$  para todo  $x$ , y por lo tanto,  $p(x) = 1$  para todo  $x$ .

Esto también puede ser generalizado a  $\sum_{i=0}^n x_i^m l_i(x) = x^m$ , si  $m \leq n$ .

## Error en el polinomio interpolante

Ahora veremos un resultado sobre el error que se comete al reemplazar una función por un polinomio que la interpola en algunos puntos dados. Primero veremos dos observaciones muy simples que serán útiles en el teorema siguiente.

**Observación 1:** si  $p$  es un polinomio de grado igual a 0 entonces  $p'(x) \equiv 0$ ;  
 si  $p$  es un polinomio de grado igual a 1 entonces  $p''(x) \equiv 0$ ;  
 si  $p$  es un polinomio de grado igual a 2 entonces  $p'''(x) \equiv 0$ ; y en general,  
 si  $p$  es un polinomio de grado igual a  $n$  entonces  $p^{(n+1)}(x) \equiv 0$ .

**Observación 2:** si  $f$  es una función continua en  $[a, b]$  y derivable en  $(a, b)$ . Si además,  $f(a) = f(b)$  entonces existe  $\alpha \in (a, b)$  tal que  $f'(\alpha) = 0$  (Teorema de Rolle). En particular, si  $f(a) = f(b) = 0$  entonces existe  $\alpha \in (a, b)$  tal que  $f'(\alpha) = 0$ . Más aún, si  $f(a) = f(b) = f(c) = 0$  entonces existen  $\alpha \in (a, b)$  y  $\beta \in (b, c)$  tal que  $f'(\alpha) = f'(\beta) = 0$ .

**Teorema 2.** Sea  $f$  una función en  $C^{n+1}[a, b]$  y  $p$  un polinomio de grado  $\leq n$  que interpola a  $f$  en  $(n+1)$  puntos distintos  $x_0, x_1, \dots, x_n$  en  $[a, b]$ . Entonces para cada  $x \in [a, b]$  existe un  $\xi = \xi_x \in (a, b)$  tal que

$$f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) \prod_{i=0}^n (x - x_i).$$

*Demostración.* Si  $x = x_i$  para  $i = 0, 1, \dots, n$ , la igualdad es trivialmente cierta, pues  $p$  interpola a  $f$  en esos puntos, por hipótesis.

Ahora supongamos que  $x \neq x_0, x_1, \dots, x_n$ , y definimos

$$\begin{aligned} w(t) &= \prod_{i=0}^n (t - x_i) && (\text{polinomio en } t) \\ c &= \frac{f(x) - p(x)}{w(x)} && (\text{constante, pues } x \text{ está fija}) \\ \varphi(t) &= f(t) - p(t) - cw(t) && (\text{función en } t), \end{aligned}$$

la constante  $c$  está bien definida pues  $w(x) \neq 0$  si  $x \neq x_i$  para  $i = 0, \dots, n$ .

Notar que si  $t = x_0, x_1, \dots, x_n$  entonces  $\varphi(t) = 0$ , pues  $p$  interpola a  $f$  en esos puntos y porque  $w$  se anula allí.

Además, por las definiciones de  $w, c$  y  $\varphi$  es fácil ver que si  $t = x$  entonces  $\varphi(t) = 0$ .

Luego,  $\varphi(t)$  tiene (al menos)  $(n+2)$  raíces:  $x_0, x_1, \dots, x_n, x$ , y por el Teorema de Rolle (y la Observación 2) tenemos que:

$\varphi'(t)$  tiene (al menos)  $(n+1)$  raíces en  $(a, b)$ .

$\varphi''(t)$  tiene (al menos)  $n$  raíces en  $(a, b)$ .

$\varphi'''(t)$  tiene (al menos)  $(n-1)$  raíces en  $(a, b)$ .

⋮

$\varphi^{(n+1)}(t)$  tiene (al menos) 1 raíz en  $(a, b)$ . Sea  $\xi = \xi_x \in (a, b)$  tal raíz de  $\varphi^{(n+1)}(t)$ .

Luego,

$$0 = \varphi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - p^{(n+1)}(\xi) - cw^{(n+1)}(\xi). \quad (1)$$

Como  $\prod_{i=0}^n (t - x_i) = t^{n+1} + \text{ términos de orden menor}$ , entonces  $w^{(n+1)}(\xi) = (n+1)!$ . Además, como  $p$  es un polinomio de grado menor o igual que  $n$  entonces, por la Observación 1,  $p^{(n+1)}(x) = 0$  para todo  $x$ , en particular para  $x = \xi$ .

Finalmente de (1) y de la definición de  $c$ ,

$$0 = f^{(n+1)}(\xi) - c(n+1)! = f^{(n+1)}(\xi) - \frac{f(x) - p(x)}{w(x)}(n+1)!,$$

y por lo tanto,

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^n (x - x_i).$$

□

**Ejemplo:** dar una estimación del error que se comete al aproximar la función  $f(x) = \sin x$  por un polinomio interpolante de grado 9, que interpola a  $f$  en 10 puntos en el intervalo  $[0, 1]$ .

Como el polinomio interpolante tiene grado  $n = 9$  es claro que se requieren 10 puntos. Por el teorema anterior, para poder acotar la expresión del teorema se necesita una cota de  $f^{(10)}(\xi)$ . Como  $f(x) = \sin x$ , es fácil deducir que  $|f^{(10)}(\xi)| \leq 1$ .

Además, si bien no se conocen los puntos de interpolación, sabemos que todos ellos pertenecen al intervalo  $[0, 1]$ , por lo tanto  $\prod_{i=0}^9 |x - x_i| \leq 1$ .

Entonces

$$|\operatorname{sen}x - p(x)| = \left| \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^n (x - x_i) \right| = \frac{|f^{(n+1)}(\xi)|}{(n+1)!} \left| \prod_{i=0}^n (x - x_i) \right| \leq \frac{1}{10!} < 2.8 \cdot 10^{-7}.$$

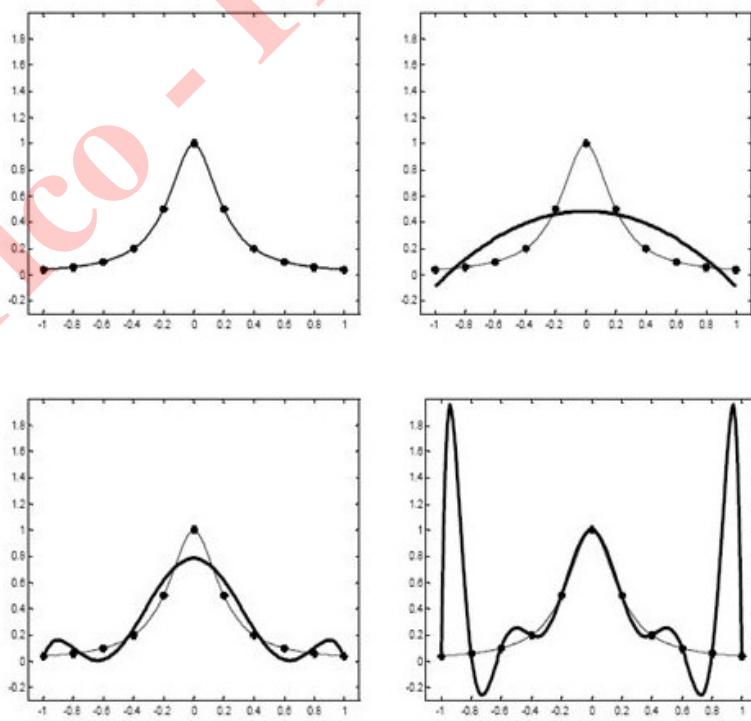
### Convergencia uniforme de los polinomios de interpolación

El teorema anterior da una expresión, para cada punto  $x$ , del error que se comete al interpolar una función  $f$  por un polinomio interpolante  $p$ .

Ahora bien, si se construyen polinomios interpolantes  $p_n$  utilizando cada vez más puntos, o equivalentemente de grado cada vez más alto, sería natural esperar que estos polinomios convergieran uniformemente a la función  $f$  en el intervalo  $[a, b]$ . Es decir, esperaríamos

$$\|f - p_n\|_\infty = \max_{a \leq x \leq b} |f(x) - p_n(x)| \rightarrow 0 \quad \text{si } n \rightarrow \infty.$$

Lamentablemente, para la mayoría de las funciones continuas no ocurre esto debido al comportamiento inestable de los polinomios de grado alto. Ver Figura 2.



**Figura 2:** Ejemplo de no convergencia uniforme de los polinomios interpolantes.

Como puede verse en la figura, a medida que se aumenta el grado del polinomio, y por lo tanto la cantidad de puntos de interpolación, el gráfico del polinomio tiende a comportarse muy diferente al gráfico de la función.

## Clase 8 - Interpolación polinomial (2)

### Repaso:

- **El problema:** Dada una tabla de  $(n+1)$  puntos:  $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$ , con  $x_0, x_1, \dots, x_n$  distintos, se desea determinar un polinomio  $p$ , con el **menor grado posible**, tal que

$$p(x_i) = y_i \quad \text{para } i = 0, \dots, n.$$

En este caso se dice que tal polinomio  $p$  **interpola** el conjunto de puntos dados.

- Existencia y unicidad del polinomio interpolante.
- Forma de Lagrange.
- Forma de Newton.
- Error en el polinomio interpolante.
- Convergencia de los polinomios de interpolación.

### Diferencias divididas

Recordemos la forma de Newton del polinomio interpolante basado en los puntos distintos  $x_0, x_1, \dots, x_n$ :

$$\begin{aligned} p_n(x) &= c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \dots + c_n(x - x_0) \dots (x - x_{n-1}) \\ &= \sum_{i=0}^n c_i \prod_{j=0}^{i-1} (x - x_j). \end{aligned}$$

Notemos que el coeficiente  $c_0$  se obtiene simplemente de  $c_0 = y_0$  (o de  $f(x_0)$ ). El coeficiente  $c_1$  se calcula con  $y_1$  (o de  $f(x_1)$ ), despejando en

$$y_1 = p_1(x_1) = c_0 + c_1(x_1 - x_0), \quad \text{esto es} \quad c_1 = \frac{y_1 - y_0}{x_1 - x_0},$$

este coeficiente depende de  $x_0, x_1, y_0, y_1$ . En general, para calcular el coeficiente  $c_k$  se requieren conocer  $x_0, \dots, x_k, y_0, \dots, y_k$ , o si estamos interpolando a una función  $f$  se requieren:  $x_0, \dots, x_k, f(x_0), \dots, f(x_k)$ . Este coeficiente se denota por

$$c_k = f[x_0, x_1, \dots, x_k],$$

para  $k = 0, \dots, n$  y se denomina **diferencias divididas**.

Ahora la forma de Newton compacta del polinomio interpolante resulta

$$p_k(x) = \sum_{i=0}^k f[x_0, x_1, \dots, x_i] \prod_{j=0}^{i-1} (x - x_j),$$

donde si  $k = 0$ :

$$f[x_0] = f(x_0),$$

si  $k = 1$ :

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Veremos a continuación un resultado general para determinar las diferencias divididas asociadas a un polinomio que interpola una función  $f$ .

**Teorema 1.** Dados  $x_0, x_1, \dots, x_n$  números reales distintos, las diferencias divididas satisfacen la siguiente ecuación

$$f[x_0, x_1, \dots, x_n] = \frac{f[x_1, x_2, \dots, x_n] - f[x_0, x_1, \dots, x_{n-1}]}{x_n - x_0}.$$

*Demostración.*

Sea  $p_{n-1}$  el polinomio de grado  $\leq n-1$  que interpola a  $f$  en los  $n$  puntos  $x_0, x_1, \dots, x_{n-1}$ .

Sea  $q$  el polinomio de grado  $\leq n-1$  que interpola a  $f$  en los  $n$  puntos  $x_1, x_2, \dots, x_n$ .

Sea  $p_n$  el polinomio de grado  $\leq n$  que interpola a  $f$  en los  $n+1$  puntos  $x_0, x_1, \dots, x_n$ .

Se afirma que

$$p_n(x) = q(x) + \frac{(x - x_n)}{(x_n - x_0)} [q(x) - p_{n-1}(x)]. \quad (1)$$

Es claro que a ambos lados de la igualdad se tienen polinomios de grado  $\leq n$ . Además, para  $i = 0$ ,

$$p_n(x_0) = f(x_0) \quad \text{y} \quad q(x_0) + \frac{(x_0 - x_n)}{(x_n - x_0)} [q(x_0) - p_{n-1}(x_0)] = p_{n-1}(x_0) = f(x_0).$$

Para  $i = 1, \dots, n-1$ ,

$$p_n(x_i) = f(x_i) \quad \text{y} \quad q(x_i) + \frac{(x_i - x_n)}{(x_n - x_0)} [q(x_i) - p_{n-1}(x_i)] = f(x_i),$$

pues  $q(x_i) = p_{n-1}(x_i) = f(x_i)$  para  $i = 1, \dots, n-1$ .

Para  $i = n$ ,

$$p_n(x_n) = f(x_n) \quad \text{y} \quad q(x_n) + \frac{(x_n - x_n)}{(x_n - x_0)} [q(x_n) - p_{n-1}(x_n)] = q(x_n) = f(x_n).$$

Por lo tanto, a ambos lados lados de (1) se tiene un polinomio de grado  $\leq n$  y ambos interpolan a  $f$  en los mismos  $n+1$  puntos distintos. Luego por unicidad del polinomio interpolante, la igualdad (1) es cierta. Como ambos polinomios son iguales, los coeficientes de cada potencia de  $x$  deben coincidir. En particular considerando el coeficiente de  $x^n$  a ambos lados de (1) obtenemos

$$\begin{aligned} f[x_0, x_1, \dots, x_n] &= \frac{1}{x_n - x_0} (f[x_1, x_2, \dots, x_n] - f[x_0, x_1, \dots, x_{n-1}]) \\ &= \frac{f[x_1, x_2, \dots, x_n] - f[x_0, x_1, \dots, x_{n-1}]}{x_n - x_0}. \end{aligned}$$

□

Luego

$$\begin{aligned}
 f[x_0] &= f(x_0) \\
 f[x_0, x_1] &= \frac{f[x_1] - f[x_0]}{x_1 - x_0} \\
 f[x_0, x_1, x_2] &= \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} \\
 &\vdots \quad = \quad \vdots \\
 f[x_i, x_{i+1}, \dots, x_{i+j}] &= \frac{f[x_{i+1}, x_{i+2}, \dots, x_{i+j}] - f[x_i, x_{i+1}, \dots, x_{i+j-1}]}{x_{i+j} - x_i}
 \end{aligned} \tag{2}$$

### Tabla de diferencias divididas

Dados 4 puntos distintos (no necesariamente ordenados) se puede construir la tabla de diferencias divididas de la siguiente manera:

$x_0$	$f[x_0]$	$f[x_0, x_1]$	$f[x_0, x_1, x_2]$	$f[x_0, x_1, x_2, x_3]$
$x_1$	$f[x_1]$	$f[x_1, x_2]$	$f[x_1, x_2, x_3]$	
$x_2$	$f[x_2]$	$f[x_2, x_3]$		
$x_3$	$f[x_3]$			

**Ejemplo:** Dados los siguientes valores:

$x$	3	1	5	6
$f(x)$	1	-3	2	4

a) obtener la tabla de diferencias divididas

3	1	2	-3/8	7/40
1	-3	5/4	3/20	
5	2	2		
6	4			

b) obtener el polinomio interpolante en la forma de Newton:

$$p(x) = 1 + 2(x-3) - \frac{3}{8}(x-3)(x-1) + \frac{7}{40}(x-3)(x-1)(x-5).$$

## Algoritmo para calcular los coeficientes de la tabla de diferencias divididas

Para obtener algorítmicamente los coeficientes de la tabla de diferencias divididas se puede pensar la misma como en un arreglo matricial de la siguiente forma:

$x_0$	$c_{00}$	$c_{01}$	$c_{02}$	$c_{03}$	$\dots$	$c_{0,n-1}$	$c_{0,n}$
$x_1$	$c_{10}$	$c_{11}$	$c_{12}$	$c_{13}$	$\dots$	$c_{1,n-1}$	
$x_2$	$c_{20}$	$c_{21}$	$c_{22}$	$c_{23}$	$\ddots$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$		
$\vdots$	$\vdots$	$\vdots$	$\ddots$				
$x_{n-1}$	$c_{n-1,0}$	$c_{n-1,1}$					
$x_n$	$c_{n,0}$						

donde  $c_{ij} = f[x_i, x_{i+1}, \dots, x_{i+j}]$ .

Dados los datos de entrada  $x_0, x_1, \dots, x_n$  y  $f(x_0), f(x_1), \dots, f(x_n)$  se pueden calcular estos coeficientes a partir de la fórmula (2) de la siguiente manera:

```
for j = 1 to n do
    for i = 0 to n - j do
         $c_{i,j} \leftarrow (c_{i+1,j-1} - c_{i,j-1}) / (x_{i+j} - x_i)$ 
    end do
end do
```

Este algoritmo puede ser optimizado de modo de almacenar estos coeficientes en un vector en vez de una matriz para ahorrar espacio de almacenamiento.

## Propiedades de las diferencias divididas

Veremos a continuación algunos resultados interesantes acerca de las diferencias divididas. El primero es un resultado sobre la permutación de los puntos de interpolación, el segundo sobre el error de interpolación y el tercero y último relaciona las diferencias divididas con las derivadas de orden superior.

**Teorema 2.** Sean  $x_0, x_1, \dots, x_n$  números reales distintos y  $z_0, z_1, \dots, z_n$  un reordenamiento de  $x_0, x_1, \dots, x_n$ . Entonces  $f[z_0, z_1, \dots, z_n] = f[x_0, x_1, \dots, x_n]$ .

*Demostración.* Primero es importante notar que el polinomio interpolante de grado  $\leq n$  basado en los puntos  $x_0, x_1, \dots, x_n$  coincide con el polinomio interpolante de grado  $\leq n$  basado en los puntos  $z_0, z_1, \dots, z_n$ , por unicidad del polinomio interpolante.

Luego, el coeficiente de  $x^n$  en el polinomio de grado  $\leq n$  que interpola a  $f$  en  $z_0, z_1, \dots, z_n$  es  $f[z_0, z_1, \dots, z_n]$  y el coeficiente de  $x^n$  en el polinomio de grado  $\leq n$  que interpola a  $f$  en  $x_0, x_1, \dots, x_n$  es  $f[x_0, x_1, \dots, x_n]$ , y deben ser iguales pues estos dos polinomios son iguales.

□

**Teorema 3.** Sea  $p$  el polinomio de grado  $\leq n$  que interpola a  $f$  en los  $n + 1$  nodos distintos  $x_0, x_1, \dots, x_n$ . Si  $t$  es un número real distinto de los nodos, entonces

$$f(t) - p(t) = f[x_0, x_1, \dots, x_n, t] \prod_{j=0}^n (t - x_j).$$

*Demostración.* Sea  $q$  el polinomio de grado  $\leq n + 1$  que interpola a  $f$  en  $x_0, x_1, \dots, x_n, t$ . Por la manera en que se genera el polinomio interpolante en la forma de Newton se sabe que

$$q(x) = p(x) + f[x_0, x_1, \dots, x_n, t] \prod_{j=0}^n (x - x_j).$$

Como  $q$  interpola a  $f$  en el punto  $t$ , se tiene que  $q(t) = f(t)$ , y entonces:

$$f(t) = p(t) + f[x_0, x_1, \dots, x_n, t] \prod_{j=0}^n (t - x_j),$$

por lo tanto,

$$f(t) - p(t) = f[x_0, x_1, \dots, x_n, t] \prod_{j=0}^n (t - x_j).$$

□

**Teorema 4.** Si  $f$  es una función  $n$  veces continuamente diferenciable en  $[a, b]$  y  $x_0, x_1, \dots, x_n$  son  $n + 1$  nodos distintos en  $[a, b]$ , entonces existe un punto  $\xi \in (a, b)$  tal que

$$f[x_0, x_1, \dots, x_n] = \frac{1}{n!} f^{(n)}(\xi).$$

*Demostración.* Sea  $p$  el polinomio de grado  $\leq n - 1$  que interpola a  $f$  en  $x_0, x_1, \dots, x_{n-1}$ . Por el teorema del error en el polinomio interpolante de la clase anterior, aplicado a  $x = x_n$ , se sabe que

$$f(x_n) - p(x_n) = \frac{1}{n!} f^{(n)}(\xi) \prod_{j=0}^{n-1} (x_n - x_j).$$

Ahora, por el Teorema 3 se obtiene

$$f(x_n) - p(x_n) = f[x_0, x_1, \dots, x_n] \prod_{j=0}^{n-1} (x_n - x_j),$$

y por lo tanto

$$f[x_0, x_1, \dots, x_n] = \frac{1}{n!} f^{(n)}(\xi).$$

□

## Interpolación de Hermite

Comenzaremos analizando a las diferencias divididas como función de sus argumentos. Hasta ahora hemos definido a las diferencias divididas para puntos distintos  $x_0, x_1, \dots, x_n$ . Consideremos ahora el caso simple de 2 puntos  $x_0$  y  $x$ :

$$f[x_0, x] = \frac{f[x] - f[x_0]}{x - x_0} = \frac{f(x) - f(x_0)}{x - x_0}.$$

Ahora tomando límite para  $x$  que tiende a  $x_0$  se tiene que

$$\lim_{x \rightarrow x_0} f[x_0, x] = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0).$$

Luego es posible extender la definición de diferencias divididas para números repetidos de la siguiente manera

$$f[x_0, x_0] = \lim_{x \rightarrow x_0} f[x_0, x] = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0).$$

Con esta generalización se pueden construir un polinomio interpolante de grado 3, con sólo 2 puntos de interpolación y agregando 2 condiciones de interpolación de la derivada en esos mismos puntos, esto es:

$$\begin{aligned} p(x_0) &= f(x_0), & p(x_1) &= f(x_1), \\ p'(x_0) &= f'(x_0), & p'(x_1) &= f'(x_1). \end{aligned}$$

Así se obtiene la siguiente tabla de diferencia dividida

$x_0$	$f[x_0]$	$f'(x_0)$	$f[x_0, x_0, x_1]$	$f[x_0, x_0, x_1, x_1]$
$x_0$	$f[x_0]$	$f[x_0, x_1]$	$f[x_0, x_1, x_1]$	
$x_1$	$f[x_1]$	$f'(x_1)$		
$x_1$	$f[x_1]$			

Ahora, el polinomio interpolante basado en esta tabla está dado por

$$p(x) = f[x_0] + f'(x_0)(x - x_0) + f[x_0, x_0, x_1](x - x_0)^2 + f[x_0, x_0, x_1, x_1](x - x_0)^2(x - x_1).$$

En general, el polinomio interpolante que usa las derivadas en un punto se llama **forma de Hermite**.

El siguiente resultado es una generalización del Teorema 4.

**Teorema 5.** Se  $f$  una función definida en  $[a, b]$ ,  $n$  veces continuamente diferenciable en  $[a, b]$ . Sean  $x_0, x_1, \dots, x_n \in [a, b]$  puntos distintos y  $z \in (a, b)$ . Entonces

$$\lim_{(x_0, x_1, \dots, x_n) \rightarrow (z, z, \dots, z)} f[x_0, x_1, \dots, x_n] = \frac{f^{(n)}(z)}{n!}.$$

---

*Demostración.* Por el Teorema 4 se sabe que existe  $\xi \in (a, b)$  tal que

$$f[x_0, x_1, \dots, x_n] = \frac{1}{n!} f^{(n)}(\xi).$$

Como  $x_0 \rightarrow z, x_1 \rightarrow z, \dots, x_n \rightarrow z$  entonces  $\xi \rightarrow z$ .

Como  $f^{(n)}$  es continua entonces

$$\lim_{(x_0, x_1, \dots, x_n) \rightarrow (z, z, \dots, z)} f[x_0, x_1, \dots, x_n] = \lim_{\xi \rightarrow z} \frac{1}{n!} f^{(n)}(\xi) = \frac{f^{(n)}(z)}{n!}.$$

□

El siguiente corolario es una consecuencia directa del teorema anterior.

**Corolario 1.** Si  $f$  es  $n$  veces continuamente diferenciable en un entorno del punto  $x_0$ , entonces

$$f[x_0, x_0, \dots, x_0] = \frac{f^{(n)}(x_0)}{n!}.$$

*Demostración.* Trivial, basta tomar  $z = x_0$ . □

**Ejemplo:** determinar el polinomio de grado 4 que interpola los siguientes datos:

$$\begin{array}{lll} p(1) = 2, & p'(1) = 3, \\ p(2) = 6, & p'(2) = 7, & p''(2) = 8. \end{array}$$

La tabla de diferencias divididas resulta en

1	2	3	1	2	-1
1	2	4	3	1	
2	6	7	4		
2	6	7			
2	6				

y el polinomio interpolante de Hermite es

$$p(x) = 2 + 3(x-1) + (x-1)^2 + 2(x-1)^2(x-2) - (x-1)^2(x-2)^2.$$

## Clase 9 - Interpolación polinomial (3)

### Repaso

**El problema:** Dada una tabla de  $(n+1)$  puntos:  $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$  donde  $x_0, x_1, \dots, x_n$  son distintos, se desea determinar un polinomio  $p$ , con el **menor grado posible**, tal que:

$$p(x_i) = y_i \quad \text{para } i = 0, \dots, n.$$

En este caso se dice que tal polinomio  $p$  **interpola** el conjunto de puntos dados.

- Existencia y unicidad del polinomio interpolante.
- Forma de Lagrange.
- Forma de Newton.
- Error en el polinomio interpolante.
- Convergencia de los polinomios de interpolación.
- Diferencias divididas.
- Polinomios de Hermite.

### Splines

Antes de introducir el concepto de splines vamos a considerar el caso simple de interpolación lineal que será muy útil en lo que sigue.

Sea  $f$  una función definida en el intervalo  $[x_0, x_1]$  2 veces continuamente diferenciable. El polinomio de grado  $\leq 1$  que interpola a  $f$  en los puntos  $x_0, x_1$  es:

$$p(x) = f[x_0] + f[x_0, x_1](x - x_0),$$

y el error está dado por

$$e(x) = f[x_0, x_1, x](x - x_0)(x - x_1) = \frac{f''(\xi)}{2!}(x - x_0)(x - x_1),$$

para  $x, \xi \in (x_0, x_1)$ .

Sea  $M > 0$  una constante tal que  $|f''(x)| \leq M$  para todo  $x \in [x_0, x_1]$ .

Sea  $\varphi(x) = (x - x_0)(x - x_1)$ , una función cuadrática, cuyo gráfico es una parábola con las ramas hacia arriba, sus raíces son  $x_0$  y  $x_1$  y su mínimo se alcanza en  $x_m = (x_0 + x_1)/2$ .

Por lo tanto

$$\begin{aligned} |\varphi(x)| \leq |(x_m - x_0)(x_m - x_1)| &= \left| \left( \frac{x_0 + x_1}{2} - x_0 \right) \left( \frac{x_0 + x_1}{2} - x_1 \right) \right| \\ &= \left| \frac{(x_1 - x_0)}{2} \frac{(x_0 - x_1)}{2} \right| = \frac{|x_1 - x_0|^2}{4}. \end{aligned}$$

Por lo tanto,

$$|e(x)| \leq \frac{M}{8} |x_1 - x_0|^2. \tag{1}$$

Supongamos que se desea interpolar una función  $f$  por un polinomio interpolante  $p_n$ . Usar pocos puntos de interpolación podría generar un polinomio que no aproxime bien a la función. Por otro lado, como se comentó anteriormente, y contrariamente a lo que podría esperarse, aumentar la cantidad de puntos de interpolación no mejora la convergencia uniforme del polinomio interpolante  $p_n$  a la función  $f$ . Esto es conocido como fenómeno de Runge.

Una idea que trata de conciliar estos conceptos opuestos es aplicar interpolación con polinomios de grado bajo por subintervalos. Esto es conocido como **interpolación polinomial por partes** o **interpolación segmentaria** o simplemente **splines**.

**Definición 1.** Una función **spline** está formada por polinomios definidos en subintervalos, los cuales se unen entre sí obedeciendo ciertas condiciones de continuidad.

Más formalmente, dados  $n + 1$  puntos tales que  $x_0 < x_1 < \dots < x_n$ , que denominaremos **nodos**, y un entero  $k \geq 0$ , un **spline de grado k** es una función  $S$  definida en  $[x_0, x_n]$  que satisface:

- $S$  es un polinomio de grado  $\leq k$  en cada subintervalo  $[x_i, x_{i+1})$ , para  $i = 0, \dots, n - 1$ ;
- las derivadas  $S^{(i)}$  son continuas en  $[x_0, x_n]$ , para  $i = 0, \dots, k - 1$ .

Veremos con un poco más de detalles los splines lineales y cúbicos, esto es, de grado 1 y 3.

### Splines lineales

Dados los  $n + 1$  nodos tales que  $x_0 < x_1 < \dots < x_n$ , un **spline lineal** ( $k = 1$ ) es una función  $S$  definida en  $[x_0, x_n]$  que satisface:

- $S$  es un polinomio de grado  $\leq 1$  (recta) en cada subintervalo  $[x_i, x_{i+1})$ , para  $i = 0, \dots, n - 1$ ;
- la función  $S$  es continua en  $[x_0, x_n]$ .

Es decir,

$$S(x) = \begin{cases} S_0(x) = a_0x + b_0, & x \in [x_0, x_1) \\ S_1(x) = a_1x + b_1, & x \in [x_1, x_2) \\ \vdots & \vdots \\ S_{n-1}(x) = a_{n-1}x + b_{n-1}, & x \in [x_{n-1}, x_n] \end{cases},$$

donde los  $2n$  coeficientes  $a_i, b_i$ , para  $i = 0, \dots, n - 1$  son las incógnitas a ser determinadas. Para eso, se deben tener  $2n$  condiciones.

Notar que la segunda condición significa que los polinomios de grado  $\leq 1$  se pegan bien en los  $(n - 1)$  nodos interiores  $x_1, \dots, x_{n-1}$ . Las  $(n + 1)$  condiciones faltantes corresponden a las  $(n + 1)$  condiciones de interpolación  $S(x_i) = f(x_i)$  para  $i = 0, \dots, n$ . Ver Figura 1.

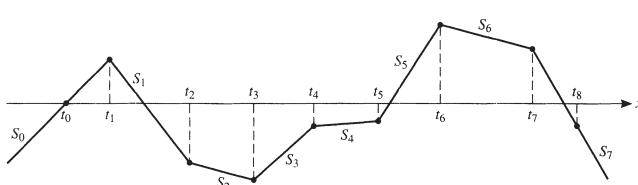


Figura 1: Gráfico de spline lineal ( $k = 1$ )

Dado un  $i$  fijo, se pueden determinar los coeficientes  $a_i, b_i$ , para  $i = 0, \dots, n-1$  de la siguiente manera:

$$\begin{aligned} a_i x_i + b_i &= S_i(x_i) = f(x_i) \\ a_i x_{i+1} + b_i &= \lim_{x \rightarrow x_{i+1}} S_i(x) = S_{i+1}(x_{i+1}) = f(x_{i+1}) \end{aligned}$$

Restando la segunda ecuación menos la primera obtenemos  $a_i(x_{i+1} - x_i) = f(x_{i+1}) - f(x_i)$ , y por lo tanto

$$a_i = \frac{f(x_{i+1}) - f(x_i)}{(x_{i+1} - x_i)}, \quad b_i = f(x_i) - a_i x_i.$$

**Observación:** supongamos que  $f$  es 2 veces continuamente diferenciable en  $[a, b]$  y  $x_k = a + kh$ ,  $k = 0, \dots, n$ , con  $h = (b - a)/n$ .

Si  $S$  es un spline lineal, en cada intervalo  $[x_k, x_{k+1}]$  se tiene un polinomio de grado  $\leq 1$ . Entonces el error de interpolación para cada  $x \in [a, b]$  está dado por:

$$|e(x)| \leq \frac{M}{8} h^2,$$

donde  $|f''(x)| \leq M$  para todo  $x \in [a, b] = [x_0, x_n]$ .

### Splines cúbicos

Dados los  $n+1$  nodos tales que  $x_0 < x_1 < \dots < x_n$ , un **spline cúbico** ( $k = 3$ ) es una función  $S$  definida en  $[x_0, x_n]$  que satisface:

- $S$  es un polinomio de grado  $\leq 3$  en cada subintervalo  $[x_i, x_{i+1}]$ , para  $i = 0, \dots, n-1$ ;
- las funciones  $S, S'$  y  $S''$  son continuas en  $[x_0, x_n]$ .

Es decir,

$$S(x) = \begin{cases} S_0(x) = a_0 x^3 + b_0 x^2 + c_0 x + d_0, & x \in [x_0, x_1) \\ S_1(x) = a_1 x^3 + b_1 x^2 + c_1 x + d_1, & x \in [x_1, x_2) \\ \vdots & \vdots \\ S_{n-1}(x) = a_{n-1} x^3 + b_{n-1} x^2 + c_{n-1} x + d_{n-1}, & x \in [x_{n-1}, x_n] \end{cases},$$

donde los  $4n$  coeficientes  $a_i, b_i, c_i, d_i$ , para  $i = 0, \dots, n-1$  son las incógnitas a ser determinadas. Para eso, se deben tener  $4n$  condiciones.

$$\begin{aligned} S(x_i) &= f(x_i), & i = 0, \dots, n & ((n+1) \text{ condiciones de interpolación}) \\ S_i(x_{i+1}) &= \lim_{x \rightarrow x_{i+1}} S_i(x) = S_{i+1}(x_{i+1}), & i = 0, \dots, n-2 & ((n-1) \text{ condiciones de continuidad de } S) \\ S'_i(x_{i+1}) &= \lim_{x \rightarrow x_{i+1}} S'_i(x) = S'_{i+1}(x_{i+1}), & i = 0, \dots, n-2 & ((n-1) \text{ condiciones de continuidad de } S') \\ S''_i(x_{i+1}) &= \lim_{x \rightarrow x_{i+1}} S''_i(x) = S''_{i+1}(x_{i+1}), & i = 0, \dots, n-2 & ((n-1) \text{ condiciones de continuidad de } S'') \end{aligned}$$

Esto da un total de  $(n+1) + 3(n-1) = 4n-2$  condiciones. Para poder determinar una única solución se deben imponer dos condiciones adicionales:

$$S''(x_0) = S''_0(x_0) = 0 \quad \text{y} \quad S''(x_n) = S''_{n-1}(x_n) = 0.$$

---

En este caso, se denomina **spline con condiciones naturales** o simplemente **spline natural**.

Otras veces se suele utilizar

$$S'(x_0) = S'_0(x_0) = f'(x_0) \quad \text{y} \quad S'(x_n) = S'_{n-1}(x_n) = f'(x_n).$$

En este caso, se denomina **spline con condiciones correctas**.

Estas condiciones suelen estar asociadas a características del problema y son indicadas en el problema o proporcionadas por quien presenta el problema.

## Clase 10 - Aproximación de funciones

El estudio de aproximación de funciones o, mejor dicho, de la teoría de aproximación comprende dos tipos de problemas: i) la búsqueda de parámetros óptimos de un modelo funcional propuesto que represente un conjunto de datos dados; y ii) dada una función de manera explícita, se desea encontrar un tipo o modelo más sencillo para representarla, por ejemplo un polinomio, que permita estimar valores funcionales de una manera más simple.

### Aproximación discreta por cuadrados mínimos

Consideremos el problema de estimar una función desconocida a través de un conjunto de datos experimentales:  $(x_i, y_i)$ ,  $i = 1, \dots, m$ .

Por ejemplo, consideremos los siguientes datos

$x_i$	1	2	3	4	5	6	7	8	9	10
$y_i$	1.3	3.5	4.2	5.0	7.0	8.8	10.1	12.5	13.0	15.6

Podemos representarlos gráficamente

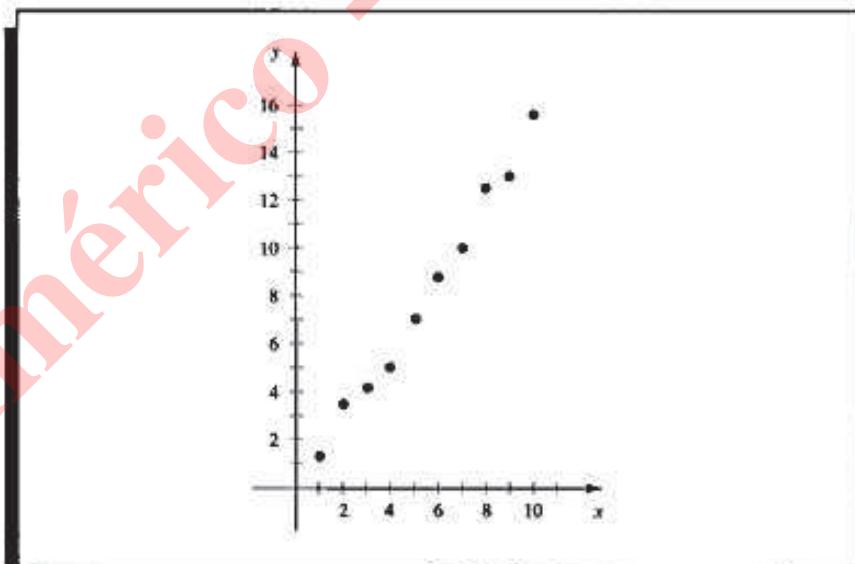
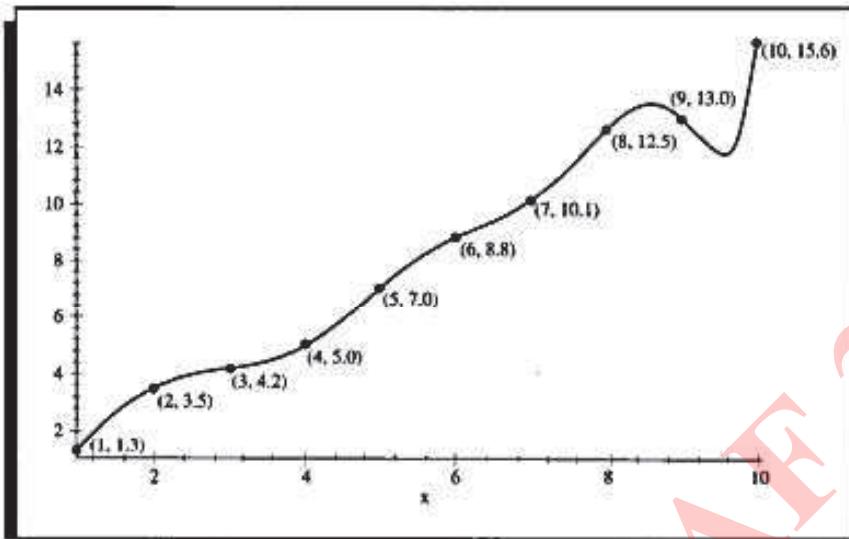


Figura 1: Datos medidos

Si se construyera un polinomio interpolante usando los  $m$  puntos se obtendría un polinomio de grado alto y ya se mencionó que esto no es conveniente. Para el caso particular de los datos del ejemplo dado donde se tienen 10 puntos se obtendría un polinomio interpolante de grado 9. Ver Figura (2).

Sin embargo, no es razonable exigir que la función pase exactamente por todos esos puntos. La Figura (1) sugiere que la relación entre  $x$  e  $y$  es lineal (polinomio de grado 1). El motivo que no exista una recta que se ajuste a estos datos, es decir que pase por todos los puntos, es que estos datos tienen errores.



**Figura 2:** Polinomio de grado 9

Sería mejor pensar en un modelo lineal y encontrar la *mejor* recta (en cierto sentido), aún cuando no coincide con los datos en ningún punto. El problema ahora será determinar cuál es la mejor recta

$$y(x) = a_1x + a_0$$

que *mejor ajusta* a esos datos, esto es, determinar los coeficientes  $a_1$  y  $a_0$  óptimos.

Una idea posible es determinar  $a_0$  y  $a_1$  tales que minimicen la función

$$E_\infty(a_0, a_1) = \max_{1 \leq i \leq 10} |y_i - (a_1x_i + a_0)|.$$

Esto es conocido como el *problema minimax*.

Otra alternativa es determinar  $a_0$  y  $a_1$  tales que minimicen la función

$$E_1(a_0, a_1) = \sum_{i=1}^{10} |y_i - (a_1x_i + a_0)|.$$

Esa función es conocida como *desviación absoluta*.

Por último, el **método de cuadrados mínimos** para ajustar a una recta con  $m$  datos consiste en determinar  $a_0$  y  $a_1$  tales que minimicen la función

$$E(a_0, a_1) = E_2(a_0, a_1) = \sum_{i=1}^m [y_i - (a_1x_i + a_0)]^2$$

con respecto a las variables  $a_0$  y  $a_1$ . En el ejemplo anterior  $m = 10$ .

Una condición necesaria para tener un mínimo es que las derivadas parciales de  $E$  con respecto a  $a_0$  y  $a_1$  deben ser cero, esto es,

$$\begin{aligned} \frac{\partial}{\partial a_0} E(a_0, a_1) &= \frac{\partial}{\partial a_0} \sum_{i=1}^m [y_i - (a_1x_i + a_0)]^2 = 2 \sum_{i=1}^m (y_i - a_1x_i - a_0)(-1) = 0, \\ \frac{\partial}{\partial a_1} E(a_0, a_1) &= \frac{\partial}{\partial a_1} \sum_{i=1}^m [y_i - (a_1x_i + a_0)]^2 = 2 \sum_{i=1}^m (y_i - a_1x_i - a_0)(-x_i) = 0. \end{aligned}$$

Luego

$$\begin{aligned} (-2) \sum_{i=1}^m (y_i - a_1 x_i - a_0) = 0 &\Rightarrow \sum_{i=1}^m y_i - a_1 \sum_{i=1}^m x_i - m a_0 = 0, \\ (-2) \sum_{i=1}^m (y_i - a_1 x_i - a_0) x_i = 0 &\Rightarrow \sum_{i=1}^m x_i y_i - a_1 \sum_{i=1}^m x_i^2 - a_0 \sum_{i=1}^m x_i = 0. \end{aligned}$$

Reordenando estas ecuaciones se puede obtener el siguiente sistema lineal de dos ecuaciones con las 2 incógnitas  $a_0$  y  $a_1$ :

$$\begin{cases} a_0 m + a_1 \sum_{i=1}^m x_i = \sum_{i=1}^m y_i \\ a_0 \sum_{i=1}^m x_i + a_1 \sum_{i=1}^m x_i^2 = \sum_{i=1}^m x_i y_i. \end{cases}$$

Estas ecuaciones son conocidas como **ecuaciones normales**.

Resolviendo este sistema se obtienen los coeficientes buscados

$$a_0 = \frac{\left( \sum_{i=1}^m x_i^2 \right) \left( \sum_{i=1}^m y_i \right) - \left( \sum_{i=1}^m x_i y_i \right) \left( \sum_{i=1}^m x_i \right)}{m \left( \sum_{i=1}^m x_i^2 \right) - \left( \sum_{i=1}^m x_i \right)^2}, \quad a_1 = \frac{m \left( \sum_{i=1}^m x_i y_i \right) - \left( \sum_{i=1}^m x_i \right) \left( \sum_{i=1}^m y_i \right)}{m \left( \sum_{i=1}^m x_i^2 \right) - \left( \sum_{i=1}^m x_i \right)^2}.$$

Para los datos del problema inicial se obtienen  $a_0 = -0.360$  y  $a_1 = 1.538$ , y el gráfico de la aproximación lineal puede verse en la Figura (3).

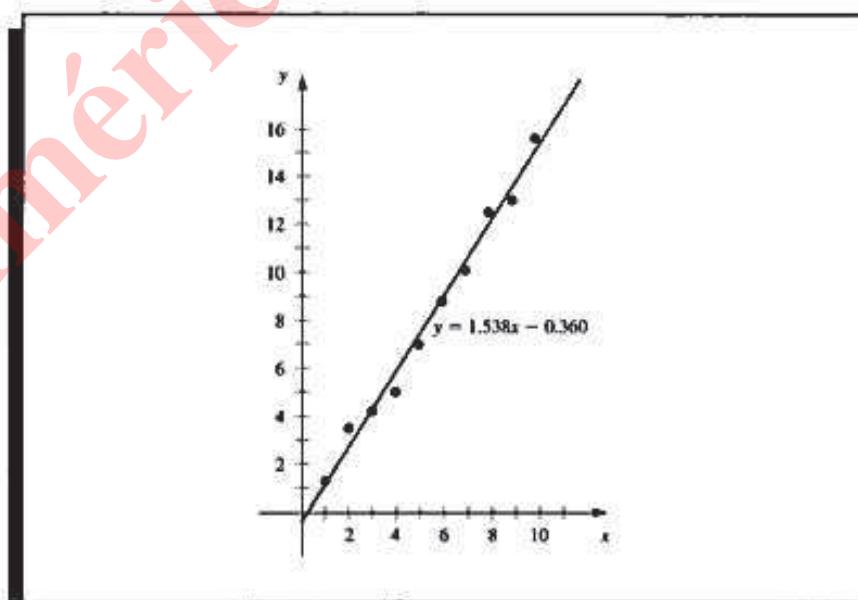


Figura 3: Modelo lineal

Ahora consideremos el caso general donde se tienen los puntos  $\{(x_i, y_i)\}$  para  $i = 1, \dots, m$  y se propone un modelo polinomial de grado  $n$ , con  $n < m - 1$

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x^1 + a_0 = \sum_{j=0}^n a_j x_i^j.$$

Al igual que en el caso de dos coeficientes, se deben determinar  $a_0, a_1, \dots, a_n$  que minimizan

$$E(a_0, \dots, a_n) = \sum_{i=1}^m [y_i - P_n(x_i)]^2.$$

Podemos reescribir esta función convenientemente

$$\begin{aligned} E(a_0, \dots, a_n) &= \sum_{i=1}^m [y_i - P_n(x_i)]^2 \\ &= \sum_{i=1}^m y_i^2 - 2 \sum_{i=1}^m P_n(x_i) y_i + \sum_{i=1}^m (P_n(x_i))^2 \\ &= \sum_{i=1}^m y_i^2 - 2 \sum_{i=1}^m \left( \sum_{j=0}^n a_j x_i^j \right) y_i + \sum_{i=1}^m \left( \sum_{j=0}^n a_j x_i^j \right)^2 \\ &= \sum_{i=1}^m y_i^2 - 2 \sum_{i=1}^m y_i \left( \sum_{j=0}^n a_j x_i^j \right) + \sum_{i=1}^m \left( \sum_{j=0}^n a_j x_i^j \right) \left( \sum_{k=0}^n a_k x_i^k \right) \\ &= \sum_{i=1}^m y_i^2 - 2 \sum_{j=0}^n a_j \left( \sum_{i=1}^m y_i x_i^j \right) + \sum_{j=0}^n \sum_{k=0}^n a_j a_k \left( \sum_{i=1}^m x_i^{j+k} \right) \end{aligned}$$

Para minimizar  $E$  se deben calcular las derivadas parciales  $\partial E / \partial a_j$  para  $j = 0, \dots, n$  e igualar a cero:

$$\frac{\partial E}{\partial a_j} = -2 \sum_{i=1}^m y_i x_i^j + 2 \sum_{k=0}^n a_k \left( \sum_{i=1}^m x_i^{j+k} \right) = 0, \quad \text{para } j = 0, \dots, n.$$

Así tenemos las  $n+1$  **ecuaciones normales** en las  $n+1$  incógnitas  $a_0, \dots, a_n$ :

$$\sum_{k=0}^n a_k \sum_{i=1}^m x_i^{j+k} = \sum_{i=1}^m y_i x_i^j, \quad \text{para } j = 0, \dots, n.$$

Es decir, se obtiene el siguiente sistema lineal

$$\left\{ \begin{array}{lcllllll} a_0 \sum_{i=1}^m x_i^0 & + & a_1 \sum_{i=1}^m x_i^1 & + & a_2 \sum_{i=1}^m x_i^2 & + & \dots & + & a_n \sum_{i=1}^m x_i^n = \sum_{i=1}^m y_i x_i^0, \\ a_0 \sum_{i=1}^m x_i^1 & + & a_1 \sum_{i=1}^m x_i^2 & + & a_2 \sum_{i=1}^m x_i^3 & + & \dots & + & a_n \sum_{i=1}^m x_i^{n+1} = \sum_{i=1}^m y_i x_i^1, \\ \vdots & & \vdots & & \vdots & & & & \vdots \\ a_0 \sum_{i=1}^m x_i^n & + & a_1 \sum_{i=1}^m x_i^{n+1} & + & a_2 \sum_{i=1}^m x_i^{n+2} & + & \dots & + & a_n \sum_{i=1}^m x_i^{2n} = \sum_{i=1}^m y_i x_i^n. \end{array} \right.$$

Este sistema lineal puede expresarse matricialmente, y es posible probar que si las  $x_i$  son todas distintas entonces las ecuaciones normales admiten una única solución.

**Observación:** a veces el modelo propuesto no es polinomial. Por ejemplo i)  $y = b e^{ax}$  o ii)  $y = bx^a$ , donde  $a$  y  $b$  son los coeficientes a determinar.

Lamentablemente, si se repitiera el procedimiento anterior cambiando el polinomio por los modelos i) o ii) no se obtiene un sistema lineal, sino un sistema no lineal el cual no se puede resolver por métodos sencillos.

---

Un método más simple consiste en aplicar logaritmo natural en los modelos *i*) o *ii*):

$$\ln y = \ln b + ax \quad \text{o} \quad \ln y = \ln b + a \ln x,$$

de esta manera es posible usar nuevamente un modelo lineal.

**Ejemplo:** supongamos que se conocen los siguientes datos

$x_i$	1.00	1.25	1.50	1.75	2.00
$y_i$	5.10	5.79	6.53	7.45	8.46

y se sabe que corresponden a un modelo de la forma  $y = be^{ax}$ . Aplicando logaritmo se obtiene  $\ln y = \ln b + ax$ , un modelo lineal  $\tilde{y} = \tilde{b} + ax$ . Antes de aplicar las fórmulas de cuadrados mínimos para el modelo lineal conviene reescribir la tabla anterior:

$x_i$	1.00	1.25	1.50	1.75	2.00
$\tilde{y}_i$	1.629	1.756	1.876	2.008	2.135

Así se obtienen  $a = 0.5056$  y  $\tilde{b} = 1.122$  y por lo tanto  $b = e^{\tilde{b}} = e^{1.122} = 3.071$ .

## Clase 11 - Aproximación de funciones (2)

Supongamos ahora que  $f \in C[a, b]$  y se desea determinar el mejor polinomio (en el sentido de cuadrados mínimos)  $P_n(x)$  de grado  $\leq n$  que minimice la siguiente medida del error entre la función  $f$  y  $P_n$  en el intervalo  $[a, b]$ :

$$E = E(a_0, \dots, a_n) = \int_a^b [f(x) - P_n(x)]^2 dx = \int_a^b [f(x) - \sum_{k=0}^n a_k x^k]^2 dx,$$

es decir, se deben determinar los coeficientes  $a_0, \dots, a_n$  que definen el polinomio  $P_n(x)$  de manera que  $E$  sea mínima. Ver Figura 1.

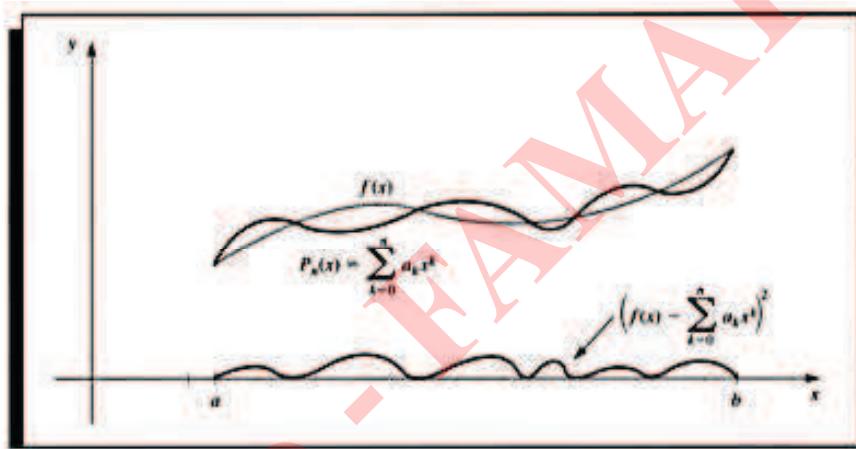


Figura 1: Gráficos de  $f$  y  $P_n$ .

Al igual que antes, una condición necesaria para encontrar un minimizador en  $a_0, \dots, a_n$  es que  $\partial E / \partial a_j = 0$  para todo  $j = 0, \dots, n$ . Antes de calcular estas derivadas, vamos a reescribir convenientemente la expresión  $E$ :

$$\begin{aligned} E = E(a_0, \dots, a_n) &= \int_a^b [f(x) - \sum_{k=0}^n a_k x^k]^2 dx \\ &= \int_a^b [f(x)]^2 dx - 2 \sum_{k=0}^n a_k \int_a^b x^k f(x) dx + \int_a^b [\sum_{k=0}^n a_k x^k]^2 dx, \end{aligned}$$

luego para  $j = 0, \dots, n$ ,

$$\frac{\partial E}{\partial a_j} = -2 \int_a^b x^j f(x) dx + 2 \sum_{k=0}^n a_k \int_a^b x^{k+j} dx = 0.$$

Así, se obtienen las **ecuaciones normales**:

$$\sum_{k=0}^n a_k \int_a^b x^{k+j} dx = \int_a^b x^j f(x) dx \quad \text{para } j = 0, \dots, n.$$

**Ejemplo:** determinar el polinomio de aproximación de cuadrados mínimos de grado  $\leq 2$  para la función  $f(x) = \sin \pi x$  en el intervalo  $[0, 1]$ .

Las ecuaciones normales para el polinomio  $P_2(x) = a_2x^2 + a_1x + a_0$ , están dadas por:

$$\begin{aligned} a_0 \int_0^1 1 dx + a_1 \int_0^1 x dx + a_2 \int_0^1 x^2 dx &= \int_0^1 \sin \pi x dx \\ a_0 \int_0^1 x dx + a_1 \int_0^1 x^2 dx + a_2 \int_0^1 x^3 dx &= \int_0^1 x \sin \pi x dx \\ a_0 \int_0^1 x^2 dx + a_1 \int_0^1 x^3 dx + a_2 \int_0^1 x^4 dx &= \int_0^1 x^2 \sin \pi x dx \end{aligned}$$

Calculando las integrales este sistema lineal puede escribirse matricialmente

$$\begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 2/\pi \\ 1/\pi \\ (\pi^2 - 4)/\pi^3 \end{bmatrix}$$

de donde se obtiene que

$$a_0 = \frac{12\pi^2 - 120}{\pi^3} \approx -0.050465 \quad a_1 = -a_2 = \frac{720 - 60\pi^2}{\pi^3} \approx 4.12251$$

y el polinomio de grado  $\leq 2$  de mejor aproximación por cuadrados mínimos está dado por  $P_2(x) = -4.12251x^2 + 4.12251x - 0.050465$ . Ver Figura 2.

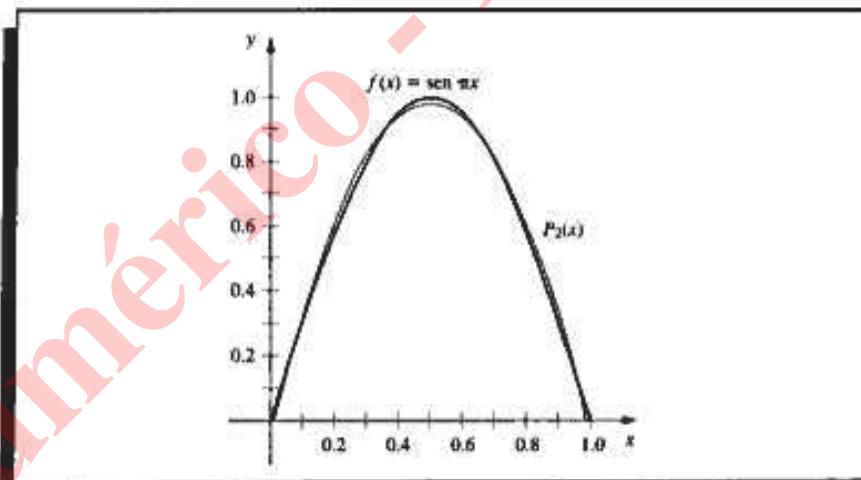


Figura 2: Gráficos de  $\sin(x)$  y  $P_2$ .

Los coeficientes de la matriz de coeficientes pueden calcularse usando la siguiente fórmula para el caso general

$$a_{jk} = \int_a^b x^{j+k} dx = \frac{b^{j+k+1} - a^{j+k+1}}{j+k+1}.$$

Esta matriz es llamada **matriz de Hilbert** y es conocida por ser mal condicionada. El concepto de *condicionamiento de una matriz* se estudia en álgebra lineal numérica y está asociado a la sensibilidad de la matriz frente a perturbaciones en los datos. Se dice que una matriz es mal condicionada cuando pequeñas perturbaciones en los datos producen grandes perturbaciones en las soluciones. Esta es una característica de la matriz de coeficientes y no del método numérico que se utilice para resolver el sistema lineal. Independientemente de esto, sería deseable que la matriz sea lo más simple posible, por ejemplo una matriz diagonal.

Para lograr esto, en lugar de proponer un polinomio de la forma  $P_n(x) = a_0 + a_1 x + \cdots + a_n x^n$ , como una combinación lineal de los polinomios  $\{x^j\}_{j=0}^n$ , consideraremos otra forma de expresar el mismo polinomio de cuadrados mínimos. A continuación veremos algunas definiciones básicas y resultados que serán necesarios.

**Definición 1.** *El conjunto de funciones  $\{\phi_0, \dots, \phi_n\}$  es **linealmente independiente** en el intervalo  $[a, b]$ , si siempre que*

$$c_0 \phi_0(x) + c_1 \phi_1(x) + \cdots + c_n \phi_n(x) = 0 \quad \text{para cualquier } x \in [a, b],$$

*se tiene que  $c_0 = c_1 = \dots = c_n = 0$ . En caso contrario se dice que ese conjunto de funciones es **linealmente dependiente**.*

**Teorema 1.** *Si  $\phi_j(x)$  es un polinomio en  $x$  de grado igual a  $j$  para  $j = 0, \dots, n$ , entonces  $\{\phi_0, \dots, \phi_n\}$  es un conjunto linealmente independiente para cualquier intervalo  $[a, b]$ .*

*Demostración.* Sean  $c_0, \dots, c_n$  números reales tales que

$$P(x) = c_0 \phi_0(x) + c_1 \phi_1(x) + \cdots + c_n \phi_n(x) = \sum_{j=0}^n c_j \phi_j(x) = 0,$$

para cualquier  $x \in [a, b]$ .

Como  $P(x)$  se anula en todo el intervalo  $[a, b]$ , los coeficientes de todas las potencias de  $x$  son iguales a cero. Como  $c_n \phi_n(x)$  es el único término que incluye  $x^n$  entonces el coeficiente  $c_n = 0$  y por lo tanto,  $\sum_{j=0}^{n-1} c_j \phi_j(x)$ .

Repitiendo esta misma idea, se tiene que el único término que incluye a  $x^{n-1}$  es  $c_{n-1} \phi_{n-1}(x)$  y de aquí se concluye que  $c_{n-1} = 0$ . De igual forma se obtiene que  $c_{n-2} = \cdots = c_1 = c_0 = 0$ , y, en consecuencia,  $\{\phi_0, \dots, \phi_n\}$  es un conjunto linealmente independiente.

□

El siguiente resultado para conjuntos de polinomios es análogo a otro que se utiliza en aplicaciones de álgebra lineal. No veremos la demostración, pero, en cambio, haremos un ejemplo.

**Teorema 2.** *Si  $\{\phi_0, \dots, \phi_n\}$  es un conjunto de polinomios linealmente independiente para cualquier intervalo  $[a, b]$  en el espacio de polinomios de grado  $\leq n$ , entonces todo polinomio de grado  $\leq n$  puede escribirse, de manera única, como combinación lineal de  $\{\phi_0, \dots, \phi_n\}$ .*

**Ejemplo:** Sean  $\phi_0(x) = 2$ ,  $\phi_1(x) = x - 3$  y  $\phi_2(x) = x^2 + 2x + 7$ . Por el Teorema 1,  $\{\phi_0, \phi_1, \phi_2\}$  es un conjunto linealmente independiente en cualquier intervalo  $[a, b]$ . Sea  $Q(x) = a_0 + a_1 x + a_2 x^2$  un polinomio cuadrático arbitrario. Aplicaremos el teorema anterior mostrando que existen coeficientes  $c_0, c_1$  y  $c_2$  tales que  $Q(x) = c_0 \phi_0(x) + c_1 \phi_1(x) + c_2 \phi_2(x)$ .

Notar que  $1 = \frac{1}{2} \phi_0(x)$ ,  $x = \phi_1(x) + 3 = \phi_1(x) + \frac{3}{2} \phi_0(x)$ , y que

$$x^2 = \phi_2(x) - 2x - 7 = \phi_2(x) - 2 \left( \phi_1(x) + \frac{3}{2} \phi_0(x) \right) - 7 \left( \frac{1}{2} \phi_0(x) \right) = \phi_2(x) - 2\phi_1(x) - \frac{13}{2} \phi_0(x).$$

---

Luego

$$\begin{aligned} Q(x) &= a_0 + a_1x + a_2x^2 \\ &= a_0 \left( \frac{1}{2}\phi_0(x) \right) + a_1 \left( \phi_1(x) + \frac{3}{2}\phi_0(x) \right) + a_2 \left( \phi_2(x) - 2\phi_1(x) - \frac{13}{2}\phi_0(x) \right) \\ &= \left( \frac{1}{2}a_0 + \frac{3}{2}a_1 - \frac{13}{2}a_2 \right) \phi_0(x) + (a_1 - 2a_2)\phi_1(x) + a_2\phi_2(x), \end{aligned}$$

y por lo tanto cualquier polinomio cuadrático puede escribirse como una combinación lineal de  $\{\phi_0, \phi_1, \phi_2\}$ .

## Clase 12 - Aproximación de funciones (3)

### Mejor aproximación de funciones con pesos

Ahora se reformulará el problema de mejor aproximación de funciones por cuadrados mínimos con funciones de peso.

**Definición 1.** Una función (integrable)  $\omega$  se llama **función de peso** en el intervalo  $I$  si  $\omega(x) \geq 0$  para todo  $x \in I$ , pero  $\omega(x) \neq 0$  para todo  $x$  en cualquier subintervalo de  $I$ , es decir  $\omega$  no puede ser constantemente cero en un subintervalo de  $I$ .

Las funciones de peso se utilizarán en la definición de la medida del error y permiten dar más o menos importancia a las aproximaciones en ciertas partes del intervalo. Por ejemplo, la función de peso definida por

$$\omega(x) = \frac{1}{\sqrt{1-x^2}} \quad \text{para } x \in (-1, 1)$$

pone menos énfasis cerca del origen y por el contrario mucho más cerca de los extremos del intervalo. Ver Figura (1).

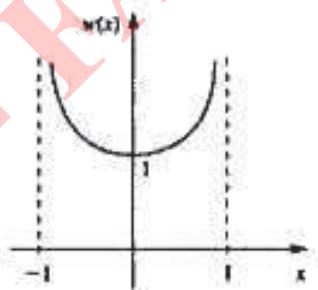


Figura 1: Función de peso  $w(x) = \frac{1}{\sqrt{1-x^2}}$

**El problema reformulado:** Dados  $\{\phi_0, \phi_1, \dots, \phi_n\}$  un conjunto de funciones linealmente independientes en  $[a, b]$ , una función de peso  $\omega$  definida en  $[a, b]$  y  $f$  una función continua en  $[a, b]$ , se desean determinar los coeficientes  $a_0, a_1, \dots, a_n$  de la combinación lineal que definen

$$P(x) = \sum_{k=0}^n a_k \phi_k(x),$$

tal que minimizan la medida del error

$$E = E(a_0, \dots, a_n) = \int_a^b \omega(x)[f(x) - P(x)]^2 dx = \int_a^b \omega(x)[f(x) - \sum_{k=0}^n a_k \phi_k(x)]^2 dx.$$

Notar que esto generaliza lo que vimos en la clase anterior tomando  $\omega(x) \equiv 1$  y  $\phi_k(x) = x^k$  para  $k = 0, \dots, n$ .

Nuevamente, una condición necesaria para encontrar un minimizador en  $(a_0, \dots, a_n)$  es que  $\partial E / \partial a_j = 0$  para todo  $j = 0, \dots, n$ , esto es,

$$\frac{\partial E}{\partial a_j} = 2 \int_a^b \omega(x)[f(x) - \sum_{k=0}^n a_k \phi_k(x)] \phi_j(x) dx = 0, \quad \text{para } j = 0, \dots, n.$$

Así, obtenemos las **ecuaciones normales** para el caso general

$$\sum_{k=0}^n a_k \int_a^b \omega(x) \phi_k(x) \phi_j(x) dx = \int_a^b \omega(x) f(x) \phi_j(x) dx, \quad \text{para } j = 0, \dots, n.$$

Si fuera posible elegir funciones  $\{\phi_0, \dots, \phi_n\}$  tal que

$$\int_a^b \omega(x) \phi_k(x) \phi_j(x) dx = \begin{cases} 0 & \text{si } j \neq k \\ \alpha_j > 0 & \text{si } j = k \end{cases} \quad (1)$$

las ecuaciones normales se podrían reducir a

$$a_j \int_a^b \omega(x) (\phi_j(x))^2 dx = \int_a^b \omega(x) f(x) \phi_j(x) dx, \quad \text{para } j = 0, \dots, n.$$

Ahora, usando (1) se tiene

$$a_j \alpha_j = \int_a^b \omega(x) f(x) \phi_j(x) dx, \quad \text{para } j = 0, \dots, n,$$

y por lo tanto

$$a_j = \frac{1}{\alpha_j} \int_a^b \omega(x) f(x) \phi_j(x) dx, \quad \text{para } j = 0, \dots, n.$$

**Definición 2.** El conjunto  $\{\phi_0, \dots, \phi_n\}$  es un **conjunto ortogonal de funciones en el intervalo  $[a, b]$** , con respecto a la función de peso  $\omega$ , si

$$\int_a^b \omega(x) \phi_k(x) \phi_j(x) dx = \begin{cases} 0 & \text{si } j \neq k \\ \alpha_j & \text{si } j = k. \end{cases}$$

Si  $\alpha_j = 1$  para todo  $j = 0, \dots, n$  se dice que el conjunto es **ortonormal**.

**Lema 1.** Si  $\{\phi_0, \phi_1, \dots, \phi_n\}$  es un conjunto ortogonal de funciones en el intervalo  $I$  con respecto a una función de peso  $\omega$  definida en  $I$  entonces son linealmente independientes.

*Demostración.* Supongamos que

$$\sum_{j=0}^n c_j \phi_j(x) = 0, \quad \text{para todo } x \in I.$$

Luego

$$0 = \int_a^b 0 \phi_k(x) \omega(x) dx = \int_a^b \sum_{j=0}^n c_j \phi_j(x) \phi_k(x) \omega(x) dx = \sum_{j=0}^n c_j \int_a^b \phi_j(x) \phi_k(x) \omega(x) dx = c_k \alpha_k,$$

como  $\alpha_k > 0$ , entonces  $c_k = 0$  para todo  $k = 0, \dots, n$ , y por lo tanto las funciones son linealmente independientes.  $\square$

**Lema 2.** Sea el conjunto de funciones polinomiales  $\{\phi_0, \phi_1, \dots, \phi_n\}$  es un conjunto ortogonal en el intervalo  $[a, b]$  con respecto a una función de peso  $\omega$ , con grado de  $\phi_k$  igual a  $k$  y  $Q_k(x)$  es un polinomio de grado  $k$  menor estricto que  $n$  entonces

$$\int_a^b \omega(x) \phi_n Q_k(x) dx = 0.$$

*Demostración.* Como  $Q_k(x)$  tiene grado  $k$  se sabe que existen coeficientes  $c_0, \dots, c_k$  tales que

$$Q_k(x) = \sum_{j=0}^k c_j \phi_j(x).$$

Luego

$$\int_a^b \omega(x) \phi_n(x) Q_k(x) dx = \sum_{j=0}^k c_j \int_a^b \omega(x) \phi_n(x) \phi_j(x) dx = 0,$$

pues  $\phi_n$  es ortogonal a  $\phi_j$  para cada  $j = 0, \dots, k$ . □

En resumen, con lo anterior se ha probado el siguiente teorema

**Teorema 1.** Si  $\{\phi_0, \phi_1, \dots, \phi_n\}$  es un conjunto ortogonal de funciones en el intervalo  $[a, b]$  con respecto a una función de peso  $\omega$  definida en  $[a, b]$ , entonces la aproximación por cuadrados mínimos a una función continua  $f$  respecto al peso  $\omega$  está dada por

$$P(x) = \sum_{k=0}^n a_k \phi_k(x),$$

donde para cada  $k = 0, \dots, n$ ,

$$a_k = \frac{\int_a^b \omega(x) f(x) \phi_k(x) dx}{\int_a^b \omega(x) (\phi_k(x))^2 dx} = \frac{1}{\alpha_k} \int_a^b \omega(x) f(x) \phi_k(x) dx.$$

El resultado siguiente da una relación de recurrencia que permite generar un conjunto de funciones ortogonales.

**Teorema 2.** El conjunto de funciones polinomiales  $\{\phi_0, \phi_1, \dots, \phi_n\}$  que se define a continuación es un conjunto ortogonal en el intervalo  $[a, b]$  con respecto a una función de peso  $\omega$

$$\phi_0(x) = 1, \quad \phi_1(x) = x - B_1 \quad \text{para cada } x \in [a, b],$$

donde

$$B_1 = \frac{\int_a^b x \omega(x) (\phi_0(x))^2 dx}{\int_a^b \omega(x) (\phi_0(x))^2 dx},$$

y para  $k \geq 2$

$$\phi_k(x) = (x - B_k) \phi_{k-1}(x) - C_k \phi_{k-2}(x) \quad \text{para cada } x \in [a, b],$$

donde

$$B_k = \frac{\int_a^b x \omega(x) (\phi_{k-1}(x))^2 dx}{\int_a^b \omega(x) (\phi_{k-1}(x))^2 dx} \quad \text{y} \quad C_k = \frac{\int_a^b x \omega(x) \phi_{k-1}(x) \phi_{k-2}(x) dx}{\int_a^b \omega(x) (\phi_{k-2}(x))^2 dx}.$$

*Demostración.* La prueba se hará por inducción en  $k$ .

Si  $k = 1$ , entonces,

$$\int_a^b \omega(x) \phi_1(x) \phi_0(x) dx = \int_a^b x \omega(x) \phi_0(x) dx - B_1 \int_a^b \omega(x) \phi_0(x) dx = 0,$$

pues  $\phi_0(x) = (\phi_0(x))^2$  y por la definición de  $B_1$ .

Ahora supongamos, por hipótesis inductiva, que  $\{\phi_0, \phi_1, \dots, \phi_{k-1}\}$  son ortogonales, y veámos que  $\phi_k$  es ortogonal a todas las funciones anteriores.

$$\begin{aligned} \int_a^b \omega(x) \phi_k(x) \phi_{k-1}(x) dx &= \int_a^b \omega(x) ((x - B_k) \phi_{k-1}(x) - C_k \phi_{k-2}(x)) \phi_{k-1}(x) dx \\ &= \int_a^b x \omega(x) (\phi_{k-1}(x))^2 dx - B_k \int_a^b \omega(x) (\phi_{k-1}(x))^2 dx \\ &\quad - C_k \int_a^b \omega(x) \phi_{k-2}(x) \phi_{k-1}(x) dx = 0, \end{aligned}$$

pues los dos primeros términos suman cero por la definición de  $B_k$  y el último término es cero por la hipótesis inductiva.

Además,

$$\begin{aligned} \int_a^b \omega(x) \phi_k(x) \phi_{k-2}(x) dx &= \int_a^b \omega(x) ((x - B_k) \phi_{k-1}(x) - C_k \phi_{k-2}(x)) \phi_{k-2}(x) dx \\ &= \int_a^b x \omega(x) \phi_{k-1}(x) \phi_{k-2}(x) dx - B_k \int_a^b \omega(x) \phi_{k-1}(x) \phi_{k-2}(x) dx \\ &\quad - C_k \int_a^b \omega(x) (\phi_{k-2}(x))^2 dx = 0, \end{aligned}$$

pues el primero y el tercer término suman cero por la definición de  $C_k$  y el segundo término es cero por la hipótesis inductiva.

Por último para  $0 \leq i \leq k-3$ , reemplazando  $k$  por  $i+1$  en  $\phi_k(x) = (x - B_k) \phi_{k-1}(x) - C_k \phi_{k-2}(x)$ , se obtiene que  $\phi_{i+1}(x) = (x - B_{i+1}) \phi_i(x) - C_{i+1} \phi_{i-1}(x)$  y por lo tanto

$$x \phi_i(x) = \phi_{i+1}(x) + B_{i+1} \phi_i(x) + C_{i+1} \phi_{i-1}(x). \quad (2)$$

Luego

$$\begin{aligned} \int_a^b \omega(x) \phi_k(x) \phi_i(x) dx &= \int_a^b \omega(x) ((x - B_k) \phi_{k-1}(x) - C_k \phi_{k-2}(x)) \phi_i(x) dx \\ &= \int_a^b x \omega(x) \phi_{k-1}(x) \phi_i(x) dx - B_k \int_a^b \omega(x) \phi_{k-1}(x) \phi_i(x) dx \\ &\quad - C_k \int_a^b \omega(x) \phi_{k-2}(x) \phi_i(x) dx. \end{aligned}$$

Los dos últimos términos son iguales a cero por la hipótesis inductiva. Usamos (2) para analizar el primer término:

$$\begin{aligned} \int_a^b x \omega(x) \phi_{k-1}(x) \phi_i(x) dx &= \int_a^b \omega(x) x \phi_i(x) \phi_{k-1}(x) dx \\ &= \int_a^b \omega(x) (\phi_{i+1}(x) + B_{i+1} \phi_i(x) + C_{i+1} \phi_{i-1}(x)) \phi_{k-1}(x) dx \\ &= \int_a^b \omega(x) \phi_{i+1}(x) \phi_{k-1}(x) dx + B_{i+1} \int_a^b \omega(x) \phi_i(x) \phi_{k-1}(x) dx \\ &= +C_{i+1} \int_a^b \omega(x) \phi_{i-1}(x) \phi_{k-1}(x) dx, \end{aligned}$$

y estos tres términos son iguales a cero por la hipótesis inductiva si  $0 \leq i \leq k-3$ . Por lo tanto queda demostrada la ortogonalidad en todos los casos.  $\square$

---

## Ejemplos:

- **Polinomios de Legendre:**  $I = [-1, 1]$ ,  $\omega(x) = 1$  para todo  $x \in I$ ,

$$\phi_0(x) = 1, \quad \phi_1(x) = x, \quad \phi_2(x) = x^2 - \frac{1}{3}, \quad \phi_3(x) = x^3 - \frac{3}{5}x, \quad \phi_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{25}, \dots$$

- **Polinomios de Chebyshev:**  $I = (-1, 1)$ ,  $\omega(x) = \frac{1}{\sqrt{1-x^2}}$  para todo  $x \in I$ ,

$$\phi_{k+1}(x) = \cos(k \arccos(x)), \quad \text{para todo } x \in I.$$

- **Polinomios de Laguerre:**  $I = [0, +\infty)$ ,  $\omega(x) = e^{-x}$  para todo  $x \in I$ ,

$$\phi_{k+1}(x) = \frac{e^x}{k!} \frac{d^k}{dx^k} (e^{-x} x^k), \quad \text{para todo } x \in I.$$

- **Polinomios de Hermite:**  $I = (-\infty, +\infty)$ ,  $\omega(x) = e^{-x^2}$  para todo  $x \in I$ ,

$$\phi_{k+1}(x) = (-1)^k e^{x^2} \frac{d^k}{dx^k} e^{x^2}, \quad \text{para todo } x \in I.$$

## Clase 13 - Integración numérica

La integración numérica es una herramienta de gran utilidad en las siguientes aplicaciones:

- Se desea calcular la integral definida de  $f$  en el intervalo  $[a, b]$ , pero  $f$  es muy complicada o no tiene primitiva, como por ejemplo

$$\int_0^1 e^{-x^2} dx$$

- no se tiene la función  $f$  en forma explícita sino sólo se conocen algunos valores funcionales, por ejemplo una lista de pares ordenados  $(x_i, y_i)$  para  $i = 0, \dots, n$  representados en un gráfico.

En ambos casos se desea estimar aproximadamente el valor de

$$\int_a^b f(x) dx.$$

Los métodos básicos que veremos son también conocidos como **cuadratura numérica** y tienen la forma

$$\sum_{i=0}^n a_i f(x_i).$$

Los métodos de cuadratura numérica se basan en interpolación numérica. Consideremos el conjunto de nodos distintos  $\{x_0, \dots, x_n\}$  en el intervalo  $[a, b]$ . Sea  $P_n$  el polinomio, que interpola a  $f$  en esos puntos, en la forma de Lagrange y  $e_n$  el término del error correspondiente

$$P_n(x) = \sum_{i=0}^n f(x_i) L_i(x), \quad e_n(x) = \frac{f^{n+1}(\xi_x)}{(n+1)!} \prod_{i=0}^n (x - x_i),$$

para algún  $\xi_x \in (x_0, x_n)$ . Es decir  $f(x) = P_n(x) + e_n(x)$ . Luego, integrando

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^b P_n(x) dx + \int_a^b e_n(x) dx \\ &= \int_a^b \sum_{i=0}^n f(x_i) L_i(x) dx + \int_a^b \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{i=0}^n (x - x_i) dx \\ &= \sum_{i=0}^n a_i f(x_i) + \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi_x) \prod_{i=0}^n (x - x_i) dx \end{aligned}$$

para algún  $\xi_x \in (x_0, x_n)$  y donde  $a_i = \int_a^b L_i(x) dx$  para  $i = 0, \dots, n$ .

Por lo tanto las fórmulas de cuadratura numérica están dadas por

$$\int_a^b f(x) dx \approx \sum_{i=0}^n a_i f(x_i),$$

con un error de integración numérica dado por

$$E_n(f) = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi_x) \prod_{i=0}^n (x - x_i) dx.$$

Cuando los puntos de interpolación estén igualmente espaciados, estas reglas o fórmulas se llaman cuadratura de Newton–Cotes. Inicialmente veremos las reglas simples y luego las reglas compuestas.

## Reglas simples

Las variantes que veremos a continuación dependen de la cantidad de puntos de interpolación que se utilicen.

### Regla del trapecio

Para integrar  $\int_a^b f(x) dx$  vamos a considerar dos puntos  $x_0 = a$ ,  $x_1 = b$  y  $h = b - a = x_1 - x_0$ . Ver Figura 1. De esta manera se aproxima a la función  $f$  por el polinomio interpolante

$$P_1(x) = \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1), \quad (1)$$

y su correspondiente error es

$$e_1(x) = \frac{f''(\xi_x)}{2!} (x - x_0)(x - x_1). \quad (2)$$

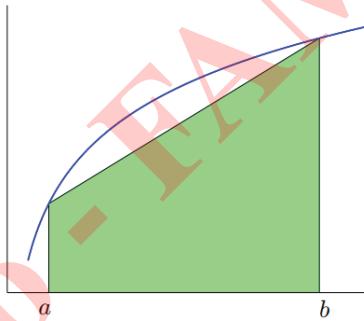


Figura 1: Regla del trapecio

Luego por (1),

$$\begin{aligned} \int_a^b f(x) dx &\approx \int_a^b \left( \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1) \right) dx \\ &= f(x_0) \int_a^b \frac{x - x_1}{x_0 - x_1} dx + f(x_1) \int_a^b \frac{x - x_0}{x_1 - x_0} dx. \end{aligned}$$

Es muy fácil verificar que ambas integrales son iguales a  $\frac{(x_1 - x_0)}{2} = \frac{b - a}{2} = \frac{h}{2}$ , y por lo tanto,

$$\int_a^b f(x) dx \approx I_T(f; a, b) = \frac{(b - a)}{2} [f(a) + f(b)] = \frac{h}{2} [f(a) + f(b)].$$

Para poder expresar el error de la integral numérica de una manera más simple se utiliza el siguiente teorema de Análisis Matemático, cuya demostración no se incluirá pero puede consultarse en libros de Cálculo.

**Teorema 1.** Supongamos que  $f \in C[a, b]$ , que  $g$  es una función integrable en  $[a, b]$  y que  $g$  no cambia de signo en  $[a, b]$ . Entonces existe  $c \in (a, b)$  tal que

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

En particular, si  $g(x) \equiv 1$ , entonces  $\int_a^b f(x) dx = f(c)(b - a)$ , esto es,  $f(c) = \frac{1}{b-a} \int_a^b f(x) dx$ .

Como  $g(x) = (x - x_0)(x - x_1) = (x - a)(x - b)$  no cambia de signo en  $[a, b]$  y aplicando el teorema anterior se sabe que existe  $\xi$  independiente de  $x$  tal que

$$\begin{aligned} \int_a^b f''(\xi_x)(x-a)(x-b) dx &= f''(\xi) \int_a^b (x-a)(x-b) dx \\ &= f''(\xi) \int_a^b (x^2 - (a+b)x + ab) dx \\ &= f''(\xi) \left[ \frac{x^3}{3} - \frac{(a+b)}{2}x^2 + abx \right]_a^b. \end{aligned} \quad (3)$$

Luego,

$$\begin{aligned} \left[ \frac{x^3}{3} - \frac{(a+b)}{2}x^2 + abx \right]_a^b &= \frac{b^3}{3} - \frac{ab^2}{2} - \frac{b^3}{2} + ab^2 - \frac{a^3}{3} + \frac{a^3}{2} + \frac{a^2b}{2} - a^2b \\ &= \frac{1}{6}(2b^3 - 3ab^2 - 3b^3 + 6ab^2 - 2a^3 + 3a^3 + 3a^2b - 6a^2b) \\ &= \frac{1}{6}(-b^3 + 3ab^2 + a^3 - 3a^2b) \\ &= \frac{(a-b)^3}{6} = -\frac{(b-a)^3}{6} = -\frac{h^3}{6}. \end{aligned} \quad (4)$$

Entonces, por (2), (3) y (4), el error de la fórmula del trapecio está dado por:

$$\begin{aligned} E_T &= E_1(x) = \int_a^b e_1 dx = \frac{1}{2!} \int_a^b f''(\xi_x)(x-a)(x-b) dx \\ &= \frac{1}{2!} f''(\xi) \left( -\frac{h^3}{6} \right) = -\frac{h^3}{12} f''(\xi) = -\frac{(b-a)^3}{12} f''(\xi), \end{aligned}$$

para algún  $\xi \in (a, b)$ .

Resumiendo, la regla del trapecio simple para integración numérica en el intervalo  $[a, b]$  está dada por

$$\int_a^b f(x) dx \approx I_T(f; a, b) = \frac{(b-a)}{2} [f(a) + f(b)] = \frac{h}{2} [f(a) + f(b)],$$

y su correspondiente error es

$$E_T = -\frac{(b-a)^3}{12} f''(\xi) = -\frac{h^3}{12} f''(\xi),$$

para algún  $\xi \in (a, b)$ .

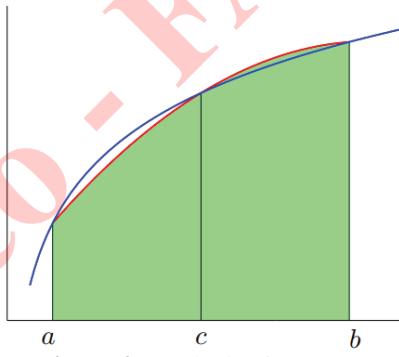
**Observación:** como el término del error tiene  $f''$ , se dice que la regla del trapecio será **exacta** cuando se aplica a una función tal que  $f''(x) \equiv 0$ , esto es, cualquier polinomio de grado menor o igual que 1. Se puede verificar esta conclusión estudiando el comportamiento de la regla del trapecio cuando se integran polinomios.

- Si  $f(x) \equiv 1$ , por un lado  $\int_a^b 1 dx = b - a$ ,
- y por otro,  $I_T(f; a, b) = \frac{(b-a)}{2} (1 + 1) = b - a$ . Luego, la regla del trapecio integra exactamente cualquier polinomio de grado 0.

- Si  $f(x) \equiv x$ , por un lado  $\int_a^b x dx = \frac{b^2 - a^2}{2}$ ,  
y por otro,  $I_T(f; a, b) = \frac{(b-a)}{2}(a+b) = \frac{b^2 - a^2}{2}$ . Luego, la regla del trapecio integra exactamente cualquier polinomio de grado 1.
- Si  $f(x) \equiv x^2$ , entonces  $\int_a^b x^2 dx = \frac{b^3 - a^3}{3}$ ,  
y en cambio,  $I_T(f; a, b) = \frac{(b-a)}{2}(a^2 + b^2) = \frac{b^3 + a^2b - ab^2 - a^3}{2}$ . Por lo tanto la regla del trapecio no integra exactamente a polinomios de grado 2.

### Regla de Simpson

En este caso, para integrar  $\int_a^b f(x) dx$  vamos a considerar tres puntos de interpolación  $x_0 = a$ ,  $x_1 = \frac{a+b}{2}$  y  $x_2 = b$ . Si llamamos  $h = \frac{b-a}{2}$  entonces,  $x_1 = x_0 + h$  y  $x_2 = a + 2h = b$ . Ver Figura 2. Con tres puntos ( $n = 2$ ) se construye un polinomio interpolante de grado 2, de manera análoga a la aplicada en la regla del trapecio.



**Figura 2:** Regla de Simpson

La deducción completa de la regla de Simpson es un ejercicio del práctico. El cálculo del error es un poco más complicado y no será demostrado por falta de tiempo. En resumen, la regla de Simpson está dada por:

$$\int_a^b f(x) dx \approx I_S(f; a, b) = \frac{h}{3} [f(a) + 4f(a+h) + f(a+2h)] = \frac{(b-a)/2}{3} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right],$$

y su correspondiente error es

$$E_S = -\frac{h^5}{90} f^{(4)}(\xi) = -\frac{((b-a)/2)^5}{90} f^{(4)}(\xi),$$

para algún  $\xi \in (a, b)$ .

**Observación:** como el término del error tiene una derivada de orden 4, la regla de Simpson integrará exactamente cuando  $f^{(4)} \equiv 0$ , esto es, para polinomios de grado menor o igual a 3.

Por último veremos 2 reglas más sencillas que utilizan un único punto de interpolación.

## Regla del rectángulo

Esta regla utiliza sólo uno de los extremos de integración para construir el polinomio interpolante, por ejemplo se considera sólo  $x_0 = a$  y por lo tanto el polinomio interpolante será una constante ( $n = 0$ ). Ver Figura 3.

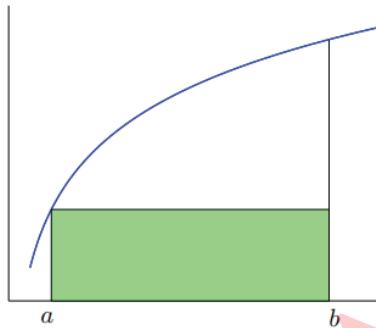


Figura 3: Regla del rectángulo

En este caso, la regla del rectángulo está dada por:

$$\int_a^b f(x) dx \approx I_R(f; a, b) = f(a)(b - a),$$

y su correspondiente error es

$$E_R = \frac{(b-a)^2}{2} f'(\xi),$$

para algún  $\xi \in (a, b)$ .

**Observación:** también es posible tomar  $x_0 = b$ , haciendo los cambios correspondientes en la fórmula de  $I_R$ . Además, como el término del error tiene una derivada de orden 1, la regla del rectángulo integrará exactamente a polinomios de grado 0.

## Regla del punto medio

Esta regla también utiliza sólo un punto pero en este caso es el punto medio del intervalo  $x_0 = (a+b)/2$ . Así, el polinomio interpolante será una constante ( $n = 0$ ). Ver Figura 4.

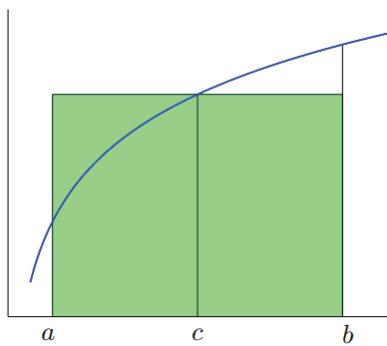


Figura 4: Regla del punto medio

En este caso, la regla del punto medio está dada por:

$$\int_a^b f(x) dx \approx I_{PM}(f; a, b) = f\left(\frac{a+b}{2}\right)(b - a),$$

y su correspondiente error es

$$E_{PM} = \frac{(b-a)^3}{24} f''(\xi),$$

para algún  $\xi \in (a, b)$ .

**Observación:** como el término del error tiene una derivada de orden 2, la regla del punto medio integrará exactamente a polinomios de grado 1.

La siguiente definición es de fundamental importancia en el estudio de las reglas de integración numérica.

**Definición 1.** La **precisión o grado de exactitud** de una fórmula o regla de cuadratura es el mayor entero no negativo  $n$  tal que la fórmula de integración es exacta para  $x^k$ , para todo  $k = 0, \dots, n$ .

Así los métodos considerados en esta clase tienen la siguiente precisión.

Regla de cuadratura	Precisión
Rectángulo	0
Punto medio	1
Trapecio	1
Simpson	3

En la siguiente tabla se resumen las reglas simples de integración numérica para estimar  $\int_a^b f(x) dx$ :

Regla	Puntos	Fórmula	Error	Precisión
Rectángulo	1	$f(a)(b-a)$	$\frac{(b-a)^2}{2} f'(\xi)$	0
Punto medio	1	$f\left(\frac{a+b}{2}\right)(b-a)$	$\frac{(b-a)^3}{24} f''(\xi)$	1
Trapecio	2	$\frac{(b-a)}{2} [f(a) + f(b)]$	$-\frac{(b-a)^3}{12} f''(\xi)$	1
Simpson	3	$\frac{(b-a)/2}{3} [f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)]$	$-\frac{((b-a)/2)^5}{90} f^{(4)}(\xi)$	3

## Clase 14 - Integración numérica (2)

Las fórmulas de Newton-Cotes y las reglas simples de integración numérica en general no son eficientes cuando se desea calcular una integral definida en un intervalo grande. Por un lado, se requerirían más puntos de interpolación y los valores de los coeficientes son más difíciles de calcular. Por otro lado, se sabe que los polinomios interpolantes suelen tener un comportamiento oscilatorio cuando se utilizan muchos puntos de interpolación. Las reglas compuestas que veremos a continuación se basan en particionar el intervalo de integración y usar las reglas simples que ya vimos.

### Reglas compuestas

Para entender mejor en qué consisten las reglas compuestas se analizará un ejemplo aplicando la Regla de Simpson.

Sea deseado estimar el valor de  $\int_0^4 e^x dx$ . Esta integral es fácil de calcular con métodos analíticos:

$$\int_0^4 e^x dx = [e^x]_0^4 = e^4 - e^0 = e^4 - 1 = 53.59815\dots$$

Si aplicamos la Regla de Simpson

$$\int_0^4 e^x dx \approx \frac{2}{3}(e^0 + 4e^2 + e^4) = 56.76958\dots$$

esta estimación tendría un error aproximado de  $-3.17143\dots$

Ahora, si dividimos el intervalo en dos:  $[0,4] = [0,2] \cup [2,4]$ , y aplicamos la Regla de Simpson en cada subintervalo se tiene

$$\begin{aligned} \int_0^4 e^x dx &= \int_0^2 e^x dx + \int_2^4 e^x dx \\ &\approx \frac{1}{3}(e^0 + 4e^1 + e^2) + \frac{1}{3}(e^2 + 4e^3 + e^4) \\ &\approx \frac{1}{3}(e^0 + 4e^1 + 2e^2 + 4e^3 + e^4) = 53.8635\dots, \end{aligned}$$

con una estimación del error de  $-0.26570\dots$

Si dividimos otra vez el intervalo  $[0,4] = [0,1] \cup [1,2] \cup [2,3] \cup [3,4]$ , y aplicamos la Regla de Simpson en cada subintervalo se tiene

$$\begin{aligned} \int_0^4 e^x dx &= \int_0^1 e^x dx + \int_1^2 e^x dx + \int_2^3 e^x dx + \int_3^4 e^x dx \\ &\approx \frac{1}{6}(e^0 + 4e^{1/2} + e) + \frac{1}{6}(e + 4e^{3/2} + e^2) \\ &\quad + \frac{1}{6}(e^2 + 4e^{5/2} + e^3) + \frac{1}{6}(e^3 + 4e^{7/2} + e^4) \\ &= \frac{1}{6}(e^0 + 4e^{1/2} + 2e + 4e^{3/2} + 2e^2 + 4e^{5/2} + 2e^3 + 4e^{7/2} + e^4) \\ &= 53.61622\dots, \end{aligned}$$

con una estimación del error de  $-0.01807\dots$

A continuación se generalizará esta idea para la Regla de Simpson.

## Regla compuesta de Simpson

Consideremos  $n$  par y subdividimos el intervalo  $[a, b]$  en  $n$  subintervalos iguales y aplicaremos la regla de Simpson en cada subintervalo. Como  $n$  es par, se tiene una **cantidad impar de puntos** igualmente espaciados  $x_j = a + jh$ , para  $j = 0, \dots, n$ , con  $h = (b - a)/n$ .

Luego,

$$\begin{aligned}\int_a^b f(x) dx &= \sum_{j=1}^{n/2} \int_{x_{2j-2}}^{x_{2j}} f(x) dx \\ &= \sum_{j=1}^{n/2} \left\{ \frac{h}{3} [f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})] - \frac{h^5}{90} f^{(4)}(\xi_j) \right\},\end{aligned}$$

para algún  $\xi_j \in (x_{2j-2}, x_{2j})$ , con  $j = 1, \dots, (n/2)$ , y  $f \in C^4[a, b]$ .

Notar que para  $j = 1, \dots, (n/2) - 1$ , el término  $f(x_{2j})$  aparece en los subintervalos  $[x_{2j-2}, x_{2j}]$  y  $[x_{2j}, x_{2j+2}]$ , entonces:

$$\int_a^b f(x) dx = \frac{h}{3} \left\{ f(x_0) + 2 \sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4 \sum_{j=1}^{n/2} f(x_{2j-1}) + f(x_n) \right\} - \frac{h^5}{90} \sum_{j=1}^{n/2} f^{(4)}(\xi_j),$$

para algún  $\xi_j \in (x_{2j-2}, x_{2j})$ , con  $j = 1, \dots, (n/2)$ , y  $f \in C^4[a, b]$ .

Consideremos el último término, correspondiente al error.

Como  $f^{(4)}$  es continua en  $[a, b]$ , entonces por el Teorema de valores extremos para funciones continuas, se tiene que

$$\begin{aligned}\min_{x \in [a, b]} f^{(4)}(x) &\leq f^{(4)}(\xi_j) \leq \max_{x \in [a, b]} f^{(4)}(x), \quad \text{para } j = 1, \dots, n/2 \\ \frac{n}{2} \min_{x \in [a, b]} f^{(4)}(x) &\leq \sum_{j=1}^{n/2} f^{(4)}(\xi_j) \leq \frac{n}{2} \max_{x \in [a, b]} f^{(4)}(x) \\ \min_{x \in [a, b]} f^{(4)}(x) &\leq \frac{2}{n} \sum_{j=1}^{n/2} f^{(4)}(\xi_j) \leq \max_{x \in [a, b]} f^{(4)}(x).\end{aligned}$$

Por el teorema del Valor Intermedio para funciones continuas, existe  $\mu \in (a, b)$  tal que

$$f^{(4)}(\mu) = \frac{2}{n} \sum_{j=1}^{n/2} f^{(4)}(\xi_j),$$

y por lo tanto,

$$\sum_{j=1}^{n/2} f^{(4)}(\xi_j) = \frac{n}{2} f^{(4)}(\mu). \tag{1}$$

Usando (1) y que  $h = (b - a)/n$ , el término del error en la regla compuesta de Simpson puede ser reformulado independientemente de  $\xi_j$ :

$$E(f) = -\frac{h^5}{90} \sum_{j=1}^{n/2} f^{(4)}(\xi_j) = -\frac{h^5}{180} n f^{(4)}(\mu) = -\frac{(b - a)}{180} h^4 f^{(4)}(\mu).$$

Los resultados desarrollados hasta aquí se resumen en el siguiente teorema:

---

**Teorema 1.** Sean  $f \in C^4[a, b]$ ,  $n$  par,  $h = (b - a)/n$  y  $x_j = a + jh$ , para  $j = 0, \dots, n$ . Entonces existe  $\mu \in (a, b)$  tal que la **regla compuesta de Simpson** para  $n$  subintervalos está dada por:

$$\int_a^b f(x) dx = \frac{h}{3} \left\{ f(x_0) + 2 \sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4 \sum_{j=1}^{n/2} f(x_{2j-1}) + f(x_n) \right\} - \frac{(b-a)}{180} h^4 f^{(4)}(\mu).$$

A continuación se presenta el pseudocódigo de la regla compuesta de Simpson.

### Algoritmo de la regla compuesta de Simpson

Dados los siguientes datos de entrada:  $a$ : extremo inferior de integración,  $b$ : extremo superior de integración y  $n$  un entero positivo par correspondiente al número de subintervalos en que se particiona el  $[a, b]$ .

```
input a, b, n
h ← (b - a)/n
sx0 ← f(a) + f(b)
sx1 ← 0      (suma de f(x_{2j-1}))
sx2 ← 0      (suma de f(x_{2j}))
x ← a
for j = 1, 2, ..., n - 1 do
    x ← x + h
    if j es par then
        sx2 ← sx2 + f(x)
    else
        sx1 ← sx1 + f(x)
    endif
endfor
sx ← (sx0 + 2sx2 + 4sx1) * h/3
output sx
end
```

### Regla compuesta del Trapecio

La deducción de esta regla se hace manera análoga a lo que se hizo con la regla compuesta de Simpson. En este caso comenzaremos enunciando el siguiente teorema.

**Teorema 2.** Sean  $f \in C^2[a, b]$ ,  $n$  un número entero positivo,  $h = (b - a)/n$  y  $x_j = a + jh$ , para  $j = 0, \dots, n$ . Entonces existe  $\mu \in (a, b)$  tal que la **regla compuesta del Trapecio** para  $n$  subintervalos está dada por:

$$\int_a^b f(x) dx = \frac{h}{2} \left\{ f(a) + 2 \sum_{j=1}^{n-1} f(x_j) + f(b) \right\} - \frac{(b-a)}{12} h^2 f''(\mu).$$

*Demostración.* Se comienza partiendo el intervalo  $[a, b]$  en  $n$  subintervalos y luego se aplica la regla simple del Trapecio en cada subintervalo (Figura (1)):

$$\begin{aligned} \int_a^b f(x) dx &= \sum_{j=1}^n \int_{x_{j-1}}^{x_j} f(x) dx \\ &= \sum_{j=1}^n \left\{ \frac{h}{2} [f(x_{j-1}) + f(x_j)] - \frac{h^3}{12} f''(\xi_j) \right\} \end{aligned}$$

para algún  $\xi_j \in (x_{j-1}, x_j)$ , con  $j = 1, \dots, n$ , y  $f \in C^2[a, b]$ .

Notar que los valores  $f(x_j)$  para  $j = 1, \dots, n-1$ , aparecen dos veces en la última expresión, entonces se puede escribir

$$\int_a^b f(x) dx = \frac{h}{2} \left\{ f(x_0) + 2 \sum_{j=1}^{n-1} f(x_j) + f(x_n) \right\} - \frac{h^3}{12} \sum_{j=1}^n f''(\xi_j),$$

para algún  $\xi_j \in (x_{j-1}, x_j)$ , con  $j = 1, \dots, n$ .

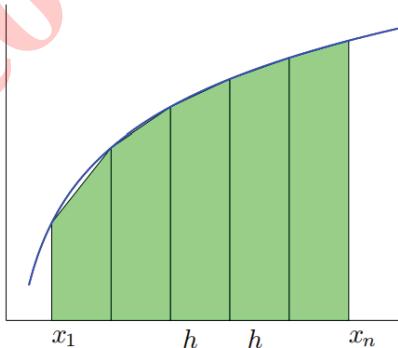


Figura 1: Regla compuesta del Trapecio

Ahora, consideramos el término del error

$$E(f) = -\frac{h^3}{12} \sum_{j=1}^n f''(\xi_j),$$

para algún  $\xi_j \in (x_{j-1}, x_j)$ , con  $j = 1, \dots, n$ . Como  $f''$  es continua en  $[a, b]$ , entonces por el Teorema de valores extremos para funciones continuas, se tiene que

$$\begin{aligned} \min_{x \in [a,b]} f''(x) &\leq f''(\xi_j) \leq \max_{x \in [a,b]} f''(x), \quad \text{para } j = 1, \dots, n/2 \\ n \min_{x \in [a,b]} f''(x) &\leq \sum_{j=1}^n f''(\xi_j) \leq n \max_{x \in [a,b]} f''(x) \\ \min_{x \in [a,b]} f''(x) &\leq \frac{1}{n} \sum_{j=1}^n f''(\xi_j) \leq \max_{x \in [a,b]} f''(x). \end{aligned}$$

Por el teorema del Valor Intermedio para funciones continuas, existe  $\mu \in (a, b)$  tal que

$$f''(\mu) = \frac{1}{n} \sum_{j=1}^n f''(\xi_j),$$

y por lo tanto,

$$\sum_{j=1}^n f''(\xi_j) = nf''(\mu). \quad (2)$$

Luego Usando (2) y que  $h = (b - a)/n$ , el término del error en la regla compuesta del Trapecio puede ser reformulado independientemente de  $\xi_j$ :

$$E(f) = -\frac{h^3}{12} \sum_{j=1}^n f''(\xi_j) = -\frac{h^3}{12} nf''(\mu) = -\frac{(b-a)}{12} h^2 f''(\mu).$$

□

El pseudocódigo es muy fácil de deducir a partir la fórmula de la regla compuesta del Trapecio.

### Reglas compuestas del Punto Medio y del Rectángulo

Las deducciones de las reglas compuestas del Punto Medio y del Rectángulo son análogas a las dos anteriores por lo que sólo enunciamos los teoremas siguientes. Los respectivos pseudocódigos se deducen fácilmente a partir de estos teoremas.

**Teorema 3.** Sean  $f \in C^2[a, b]$ ,  $n$  un número par,  $h = (b - a)/(n + 2)$  y  $x_j = a + (j + 1)h$ , para  $j = -1, 0, \dots, n + 1$ . Entonces existe  $\mu \in (a, b)$  tal que la **regla compuesta del Punto Medio para  $n + 2$  subintervalos** está dada por:

$$\int_a^b f(x) dx = 2h \sum_{j=0}^{n/2} f(x_{2j}) + \frac{(b-a)}{6} h^2 f''(\mu).$$

Ver Figura (2).

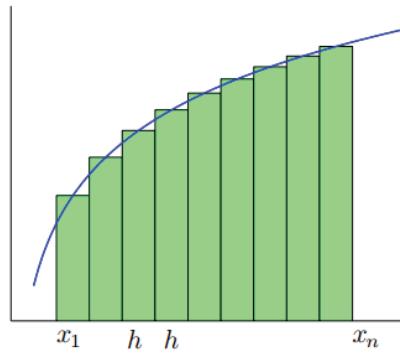


Figura 2: Regla compuesta del Punto Medio

**Teorema 4.** Sean  $f \in C^1[a, b]$ ,  $n$  un número entero positivo,  $h = (b - a)/n$  y  $x_j = a + jh$ , para  $j = 0, \dots, n$ . Entonces existe  $\mu \in (a, b)$  tal que la **regla compuesta del Rectángulo para  $n$  subintervalos** está dada por:

$$\int_a^b f(x) dx = h \sum_{j=0}^{n-1} f(x_j) + \frac{(b-a)}{2} h f'(\mu).$$

**Ejemplo:** si bien es posible calcular exactamente la integral

$$\int_0^\pi \sin x dx = [-\cos x]_0^\pi = [\cos x]_0^\pi = 2,$$

se desea determinar cuál es el entero  $n$  para estimar esta integral numéricamente con un error menor que 0.00002, utilizando las diferentes reglas compuestas de integración.

- **Regla compuesta del Rectángulo.**

Usamos la expresión del error para despejar  $n$  y el hecho que  $h = (b - a)/n$ :

$$E_R(f) = \frac{(b-a)}{2} h f'(\mu),$$

luego,

$$|E_R(f)| = \left| \frac{(\pi-0)}{2} \frac{\pi-0}{n} \cos(\mu) \right| = \frac{\pi^2}{2n} < 0.00002,$$

de aquí resulta que  $n > 246741$ .

- **Regla compuesta del Punto Medio.**

Usamos la expresión del error para despejar  $n$  y el hecho que  $h = (b - a)/(n + 2)$ :

$$E_{PM}(f) = \frac{(b-a)}{6} h^2 f''(\mu),$$

luego,

$$|E_{PM}(f)| = \left| \frac{(\pi-0)}{6} \left( \frac{\pi-0}{n+2} \right)^2 (-\sin(\mu)) \right| = \frac{\pi^3}{6(n+2)^2} < 0.00002,$$

de aquí resulta que  $n > 507$ .

- **Regla compuesta del Trapecio.**

Usamos la expresión del error para despejar  $n$  y el hecho que  $h = (b - a)/n$ :

$$E_T(f) = -\frac{(b-a)}{12} h^2 f''(\mu),$$

luego,

$$|E_T(f)| = \left| -\frac{(\pi-0)}{12} \left( \frac{\pi-0}{n} \right)^2 (-\sin(\mu)) \right| = \frac{\pi^3}{12n^3} < 0.00002,$$

de aquí resulta que  $n > 360$ .

- **Regla compuesta de Simpson.** Usamos la expresión del error para despejar  $n$  y el hecho que  $h = (b - a)/n$ :

$$E_S(f) = -\frac{(b-a)}{180} h^4 f^{(4)}(\mu),$$

luego,

$$|E_S(f)| = \left| -\frac{(\pi-0)}{180} \left( \frac{\pi-0}{n} \right)^4 \sin(\mu) \right| = \frac{\pi^5}{180n^4} < 0.00002,$$

de aquí resulta que  $n > 18$ .

En la siguiente tabla se resumen las reglas compuestas de integración numérica para estimar  $\int_a^b f(x) dx$ :

Regla	Fórmula	Error
Rectángulo	$h \sum_{j=0}^{n-1} f(x_j)$	$\frac{(b-a)}{2} h f'(\mu)$
Punto medio	$2h \sum_{j=0}^{n/2} f(x_{2j})$	$\frac{(b-a)}{6} h^2 f''(\mu)$
Trapecio	$\frac{h}{2} \left\{ f(a) + 2 \sum_{j=1}^{n-1} f(x_j) + f(b) \right\}$	$-\frac{(b-a)}{12} h^2 f''(\mu)$
Simpson	$\frac{h}{3} \left\{ f(x_0) + 2 \sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4 \sum_{j=1}^{n/2} f(x_{2j-1}) + f(x_n) \right\}$	$-\frac{(b-a)}{180} h^4 f^{(4)}(\mu)$

## Clase 15 - Integración numérica (3)

### Reglas gaussianas

Las fórmulas de cuadratura consideradas hasta ahora, simples y compuestas, se construyen usando valores funcionales en puntos conocidos. Es decir, todas tienen la forma

$$\int_a^b f(x) dx \approx \sum_{i=0}^n a_i f(x_i), \quad (1)$$

y son exactas para polinomios de grado  $\leq n$ . En esta fórmula **los nodos**  $x_0, x_1, \dots, x_n$  **son conocidos a priori** y los coeficientes  $a_0, a_1, \dots, a_n$  se determinan únicamente de manera que (1) sea una igualdad, para ciertos polinomios.

Por ejemplo, si usáramos la regla simple del Trapecio, cuyos nodos son los extremos del intervalo:

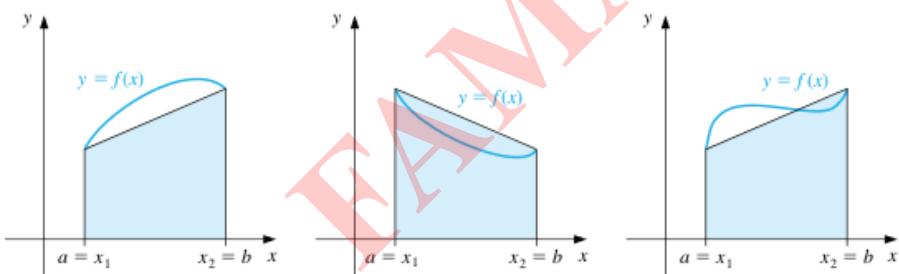


Figura 1: Regla del Trapecio (simple)

En cambio, si se pudieran elegir adecuadamente los dos nodos se podría mejorar la precisión:

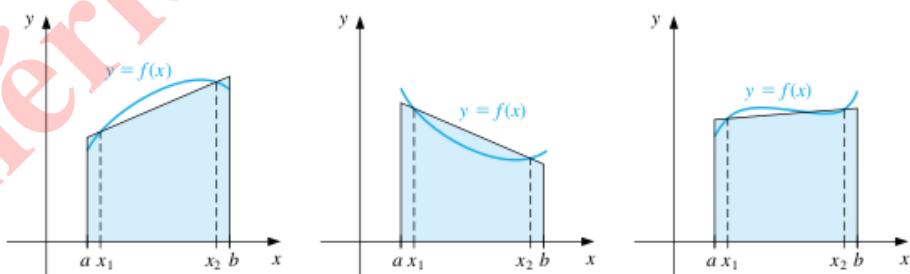


Figura 2: Regla con dos nodos

Recordemos que:

Regla	$n$	Número de puntos ( $n + 1$ )	Precisión
Rectángulo	0	1	0
Punto medio	0	1	1
Trapecio	1	2	1
Simpson	2	3	3

Es decir que para  $(n + 1)$  puntos, la precisión de cada regla de cuadratura es  $(n + 1)$  o  $n$ .

Las reglas gaussianas permiten seleccionar convenientemente los nodos, además de los coeficientes, de manera óptima en el sentido que la regla de integración sea exacta para polinomios del grado más alto posible (precisión). Es decir que se deberán determinar los  $(n + 1)$

nodos  $x_0, \dots, x_n$  y los  $(n+1)$  coeficientes  $a_0, \dots, a_n$  de manera que la regla de cuadratura con funciones de peso

$$\int_a^b w(x)f(x)dx \approx \sum_{i=0}^n a_i f(x_i), \quad (2)$$

sea exacta para polinomios de grado  $\leq 2n+1$ .

Antes de estudiar el caso general, veamos el siguiente ejemplo.

**Ejemplo:** tomando la función de peso  $w(x) = 1$ , determinar los nodos  $x_0$  y  $x_1$  y los coeficientes  $a_0$  y  $a_1$  tal que la regla de cuadratura

$$\int_{-1}^1 f(x)dx \approx a_0 f(x_0) + a_1 f(x_1), \quad (3)$$

sea exacta para polinomios  $p$  de grado  $\leq 2 \cdot 1 + 1 = 3$ .

- Si  $grado(p) = 0$ , basta considerar  $p(x) \equiv 1$ , entonces

$$2 = \int_{-1}^1 1 dx = a_0 + a_1. \quad (4)$$

- Si  $grado(p) = 1$ , basta considerar  $p(x) = x$ , entonces

$$0 = \int_{-1}^1 x dx = a_0 x_0 + a_1 x_1. \quad (5)$$

- Si  $grado(p) = 2$ , basta considerar  $p(x) = x^2$ , entonces

$$\frac{2}{3} = \int_{-1}^1 x^2 dx = a_0 x_0^2 + a_1 x_1^2. \quad (6)$$

- Si  $grado(p) = 3$ , basta considerar  $p(x) = x^3$ , entonces

$$0 = \int_{-1}^1 x^3 dx = a_0 x_0^3 + a_1 x_1^3. \quad (7)$$

Si  $a_0 = a_1 = 0$ , por (4) se tiene un absurdo, por lo tanto no pueden ser simultáneamente cero.

Supongamos ahora uno de los dos coeficientes es cero y el otro no, por ejemplo que  $a_0 = 0$  y  $a_1 \neq 0$ . Luego por (4) se deduce que  $a_1 = 2$ , y por (5) se tiene que  $x_1 = 0$ , y por (6) se llega a un absurdo. Análogamente si suponemos que  $a_0 \neq 0$  y  $a_1 = 0$ . Por lo tanto ambos coeficientes deben ser distintos de cero.

De la ecuación (5), si  $x_0 = 0$ , y como  $a_1 \neq 0$ , resulta que  $x_1 = 0$  y por (6) se llega a un absurdo. Por lo tanto  $x_0$  no puede ser 0 y por un razonamiento análogo  $x_1$  tampoco es 0.

Luego  $a_0, a_1, x_0$  y  $x_1$  son distintos de 0.

De (5) y (7) se tiene que

$$\frac{a_0 x_0^3}{a_0 x_0} = \frac{-a_1 x_1^3}{-a_1 x_1},$$

de donde se deduce que  $x_0^2 = x_1^2$  y por lo tanto  $|x_0| = |x_1|$ . Ahora se tienen dos casos.

Si  $x_0 = x_1$ , por (4) y (5) se obtendría que  $x_0 = 0$  y por lo tanto es absurdo. Por lo tanto,  $x_0 = -x_1$ .

Usando (6) se sabe que  $(a_0 + a_1)x_0^2 = 2/3$ , y entonces  $x_0^2 = 1/3$  y de aquí se tiene que  $x_0 = -\frac{\sqrt{3}}{3}$  y  $x_1 = \frac{\sqrt{3}}{3}$ .

Por último, usando (4) y (5) se obtiene que  $a_0 = a_1 = 1$ .

El teorema siguiente, debido a Gauss, indica como deben elegirse los  $(n+1)$  nodos de manera de obtener una fórmula de cuadratura que sea exacta para polinomios de hasta grado  $2n+1$ . El resultado será enunciado y probado para una función de peso  $w$  general, pero puede ser aplicado al caso simple donde  $w(x) \equiv 1$ .

**Teorema 1.** *Sea  $w$  una función de peso positiva definida en  $[a, b]$  y  $q$  un polinomio no nulo de grado exactamente  $(n+1)$  que es ortogonal a todo polinomio  $p$  de grado  $\leq n$  (con respecto a  $w$ ), es decir,*

$$\int_a^b q(x)p(x)w(x) dx = 0. \quad (8)$$

*Si  $x_0, x_1, \dots, x_n$  son las  $(n+1)$  raíces de  $q$ , entonces la fórmula*

$$\int_a^b f(x)w(x) dx \approx \sum_{i=0}^n a_i f(x_i) \quad (9)$$

*con  $a_i = \int_a^b w(x) \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x-x_j}{x_i-x_j} dx$ , será exacta para todo polinomio  $f$  de grado  $\leq 2n+1$ .*

*Demostración.* Veamos primero que la fórmula de cuadratura es exacta para polinomios de grado  $\leq n$ .

Sea  $f$  un polinomio de grado  $\leq n$ , entonces  $f$  es el único polinomio que interpola los  $n+1$  nodos  $x_0, x_1, \dots, x_n$ , y por lo tanto  $f$  se puede reescribir como  $f(x) = \sum_{i=1}^n f(x_i) \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x-x_j}{x_i-x_j}$ .

Luego

$$\int_a^b f(x)w(x) dx = \sum_{i=1}^n f(x_i) \int_a^b w(x) \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x-x_j}{x_i-x_j} dx = \sum_{i=0}^n a_i f(x_i).$$

Por lo tanto la regla de cuadratura es exacta para polinomios de grado  $\leq n$ .

Ahora supongamos que  $f$  es un polinomio tal que  $n+1 \leq \text{grado}(f) \leq 2n+1$ . Dividimos  $f$  por el polinomio  $q$  y, por el algoritmo de la división de polinomios, se sabe que existen polinomios  $p$  (cociente) y  $r$  (resto) tal que

$$f(x) = p(x)q(x) + r(x), \quad (10)$$

con  $\text{grado}(p) \leq n$  (pues  $\text{grado}(f) \leq 2n+1$ ) y  $\text{grado}(r) \leq n$  o  $r(x) \equiv 0$ .

Luego,

$$\int_a^b q(x)p(x)w(x) dx = 0, \quad (11)$$

por hipótesis, pues el grado de  $q$  es  $n+1$ .

Además, como  $x_i$  es una raíz de  $q(x)$  para  $i = 0, \dots, n$ , se tiene que

$$f(x_i) = p(x_i)q(x_i) + r(x_i) = r(x_i), \quad \text{para } i = 0, \dots, n. \quad (12)$$

Luego como  $\text{grado}(r) \leq n$ , entonces

$$\int_a^b r(x)w(x)dx = \sum_{i=0}^n a_i r(x_i). \quad (13)$$

Finalmente, por (10), (11), (12) y (13), se tiene que

$$\int_a^b f(x)w(x)dx = \int_a^b (p(x)q(x) + r(x))w(x)dx = \int_a^b r(x)w(x)dx = \sum_{i=0}^n a_i r(x_i) = \sum_{i=0}^n a_i f(x_i),$$

que es lo que se quería probar.  $\square$

El teorema anterior afirma que las raíces de los polinomios ortogonales son los nodos más convenientes para tener más precisión en las reglas gaussianas. El siguiente resultado resume algunas propiedades importantes de estos puntos. La demostración puede consultarse en los libros de la Bibliografía.

**Teorema 2.** *Sean  $w$  una función de peso en  $[a, b]$  y  $\{\phi_0, \dots, \phi_n\}$  un conjunto polinomios tales que  $\text{grado}(\phi_k) = k$  y son  $w$ -ortogonales en  $[a, b]$  en el sentido que*

$$\int_a^b \phi_i(x)\phi_j(x)w(x)dx = 0, \quad \text{si } i \neq j.$$

*Si  $x_0, \dots, x_{k-1}$  son las  $k$  raíces de  $\phi_k$  entonces tales raíces son reales, simples y pertenecen al intervalo abierto  $(a, b)$ .*

**Observación:** si la función de peso que se utiliza en el intervalo  $[-1, 1]$  es  $w(x) \equiv 1$  se obtienen los polinomios de Legendre:

$$\begin{aligned} P_0(x) &\equiv 1, & P_1(x) &= x, & P_2(x) &= x^2 - \frac{1}{3}, \\ P_3(x) &= x^3 - \frac{3}{5}x, & P_4(x) &= x^4 - \frac{6}{7}x^2 + \frac{3}{35}. \end{aligned}$$

Para estos polinomios se conocen sus raíces así como los coeficientes  $a_i$  de las fórmulas de cuadratura:

grado del polinomio	raíces ( $x_i$ )	coeficientes ( $a_i$ )
2	-0.5773502692	1.0000000000
	0.5773502692	1.0000000000
3	-0.7745966692	0.5555555556
	0.0000000000	0.8888888889
4	0.7745966692	0.5555555556
	-0.8611363116	0.3478548451
	-0.3399810436	0.6521451549
	0.3399810436	0.6521451549
	0.8611363116	0.3478548451

La siguiente muestra como es posible utilizar las reglas gaussianas en diferentes intervalos.

**Observación:** supongamos que se conoce una fórmula de integración numérica

$$\int_{-1}^1 f(x) dx \approx \sum_{i=0}^n a_i f(x_i),$$

para valores determinados de los coeficientes  $a_i$ , para  $i = 0, \dots, n$ , que dependen de los nodos  $x_0, \dots, x_n \in [-1, 1]$  y se desea obtener una fórmula de integración numérica para

$$\int_a^b f(x) dx.$$

Esto se puede hacer muy fácilmente haciendo el cambio de variables  $t : [a, b] \rightarrow [-1, 1]$ , dado por  $t = \alpha x + \beta$ , donde los coeficientes  $\alpha$  y  $\beta$  se determinan fácilmente resolviendo el sistema lineal

$$\begin{cases} \alpha a + \beta = -1 \\ \alpha b + \beta = 1 \end{cases}$$

y por lo tanto  $t = \frac{2}{b-a}x - \frac{a+b}{b-a}$ . Entonces  $x = \frac{b-a}{2}t + \frac{a+b}{2}$  y  $\frac{dx}{dt} = \frac{b-a}{2}$ , es decir,  $dx = \frac{b-a}{2}dt$ . Luego,

$$\int_a^b f(x) dx = \int_{-1}^1 f\left(\frac{b-a}{2}t + \frac{a+b}{2}\right) \frac{b-a}{2} dt = \frac{b-a}{2} \int_{-1}^1 f\left(\frac{b-a}{2}t + \frac{a+b}{2}\right) dt.$$

**Ejemplo:** se desea estimar el valor de la integral

$$\int_1^{1.5} e^{-x^2} dx$$

usando las reglas gaussianas con la función de peso  $w(x) \equiv 1$  y usando polinomios de grado 2 y 3. Se sabe que esa integral es aproximadamente  $\approx 0.1093643$ . Notar que si el grado es  $m$ , entonces el índice de los nodos de la regla de cuadratura es  $n = m - 1$ , pues los subíndices comienzan desde 0. Por lo tanto para grados  $m = 2$  y  $3$  se tienen  $n = 1$  y  $2$  y la precisión  $(2n+1)$  será 3 y 5, respectivamente.

Haciendo el cambio de variables, resulta

$$y = 4x - 5, \quad x = \frac{1}{4}y + \frac{5}{4}.$$

Luego,

$$\int_1^{1.5} e^{-x^2} dx = \frac{1}{4} \int_{-1}^1 e^{-(y+5)^2/16} dy.$$

Si el grado es  $m = 2$ , es decir ( $n = 1$ ) con los puntos  $x_0$  y  $x_1$ , entonces

$$\int_1^{1.5} e^{-x^2} dx \approx 0.1094003$$

Si el grado es  $m = 3$ , es decir ( $n = 2$ ) con los puntos  $x_0$ ,  $x_1$  y  $x_2$ , entonces

$$\int_1^{1.5} e^{-x^2} dx \approx 0.1093642$$

# Clase 16 - Resolución de sistemas lineales

## El problema

Se desea encontrar  $x_1, \dots, x_n$  solución del siguiente sistema de ecuaciones lineales:

$$\left\{ \begin{array}{l} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \dots + a_{2n}x_n = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \dots + a_{3n}x_n = b_3 \\ \vdots \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \dots + a_{nn}x_n = b_n \end{array} \right.$$

Es conveniente escribir este sistema en forma matricial:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix},$$

es decir  $Ax = b$ , con  $A \in \mathbb{R}^{n \times n}$  y  $b \in \mathbb{R}^n$ . Así, el problema de esta unidad puede formularse de la siguiente manera:

Dada  $A \in \mathbb{R}^{n \times n}$  y  $b \in \mathbb{R}^n$  hallar  $x_*$  solución de  $Ax = b$ .

Este es uno de los problemas más importantes en Análisis Numérico en general porque la resolución de muchos problemas de diferentes disciplinas o aplicaciones termina en la resolución de sistemas lineales. Para lo cual es muy importante disponer de métodos numéricos y computacionales que sean rápidos y eficientes.

A continuación enunciamos un teorema básico de Álgebra lineal que será de utilidad en lo que sigue.

**Teorema 1.** para toda matriz  $A \in \mathbb{R}^{n \times n}$ , las siguientes propiedades son equivalentes:

1. existe  $A^{-1}$ , la inversa de  $A$ . ( $A^{-1}A = AA^{-1} = I$ );
2. la matriz  $A$  es no singular, es decir, si  $Ax = 0$  entonces  $x = 0$ ;
3.  $\det(A) \neq 0$ ;
4. las columnas de  $A$  forman una base de  $\mathbb{R}^n$ ;
5. las filas de  $A$  forman una base de  $\mathbb{R}^n$ ;
6. para cada  $b \in \mathbb{R}^n$  existe un único  $x \in \mathbb{R}^n$  solución de  $Ax = b$ .

Si bien se conocen métodos matemáticos de Álgebra lineal para resolver ese sistema lineal, nuestro objetivo será estudiar métodos y algoritmos numéricos para resolver este problema de una forma sistemática y eficiente. Esta eficiencia estará asociada al costo computacional de un algoritmo. Este costo se mide en términos de cantidad de operaciones y en espacio de almacenamiento. Dado que los problemas que estudiaremos son simples y pequeños nos

ocuparemos sólo de contabilizar la cantidad de operaciones que realiza cada algoritmo para obtener la solución buscada.

Antes de estudiar el caso de una matriz general, consideraremos la resolución de sistemas lineales de dos casos muy simples.

## Sistemas diagonales

**Definición 1.** Una matriz  $A$  es diagonal si y sólo si  $a_{ij} = 0$  para  $i \neq j$ .

Sea  $A \in \mathbb{R}^{n \times n}$  una matriz diagonal

$$A = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}.$$

Luego  $A$  es no singular si y sólo sí  $a_{ii} \neq 0$ ,  $i = 1, \dots, n$  si y sólo si  $\det(A) = a_{11}a_{22} \dots a_{nn} \neq 0$ .

Entonces  $Ax = b$  si y sólo si  $a_{ii}x_i = b_i$ , para  $i = 1, \dots, n$ , si y sólo si  $x_i = \frac{b_i}{a_{ii}}$ , para  $i = 1, \dots, n$ .

**Algoritmo 1:** resolución de sistema lineal con matriz diagonal.

```

input  $n, A, b$ 
for  $i = 1, \dots, n$  do
     $x_i \leftarrow b_i / a_{ii}$ 
output  $i, x_i$ 
end for
end
```

**Costo computacional:**  $n$  operaciones (productos) en punto flotante, y así se tiene  $\mathcal{O}(n)$  flops.

**Observación:** notar que como la matriz  $A$  es diagonal y no singular, los únicos elementos no nulos están precisamente en la diagonal. Entonces, sería suficiente almacenar los elementos diagonales de  $A$  en un vector  $d = \text{diag}(A)$ , y así calcular  $x(i) = b(i)/d(i)$  para  $i = 1, \dots, n$ .

## Sistemas triangulares

**Definición 2.** Una matriz  $A$  es triangular inferior (superior) si y sólo si  $a_{ij} = 0$  para  $i < j$  ( $a_{ij} = 0$  para  $i > j$ ).

$A \in \mathbb{R}^{n \times n}$  es triangular inferior y no singular si y sólo si  $\det(A) = a_{11}a_{22} \dots a_{nn} \neq 0$ , con

$$A = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

Entonces resolver  $Ax = b$  es equivalente a

$$\begin{cases} a_{11}x_1 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \\ \vdots &\vdots \\ a_{i1}x_1 + \dots + a_{ii}x_i &= b_i \\ \vdots &\vdots \\ a_{n1}x_1 + \dots + a_{nn}x_n &= b_n \end{cases} \implies \begin{cases} x_1 &= b_1/a_{11} \\ x_2 &= \frac{1}{a_{22}}(b_2 - a_{21}x_1) \\ \vdots &\vdots \\ x_i &= \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j) \\ \vdots &\vdots \\ x_n &= \frac{1}{a_{nn}}(b_n - \sum_{j=1}^{n-1} a_{nj}x_j) \end{cases}$$

**Algoritmo 2:** resolución de sistema lineal con matriz triangular inferior.

**input**  $n, A, b$

**for**  $i = 1, \dots, n$  **do**

$$x_i \leftarrow \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j \right) / a_{ii} \quad (\text{no sumar si } i = 1)$$

**output**  $i, x_i$

**end for**

**end**

**Costo computacional:** vamos a contar las sumas/restas y productos/divisiones que se realizan para cada  $i$  y luego sumamos:

$i$	sumas	productos
1	0	1
2	1	2
3	2	3
$\vdots$	$\vdots$	$\vdots$
n	n-1	n
Total	$\frac{(n-1)n}{2}$	$\frac{n(n+1)}{2}$

Así, la cantidad total de operaciones es:  $\frac{(n-1)n}{2} + \frac{n(n+1)}{2} = n^2$ , es decir  $\mathcal{O}(n^2)$  flops.

(Recordar que  $\sum_{i=1}^n i = \frac{n(n+1)}{2}$ ).

Análogamente, Si  $A \in \mathbb{R}^{n \times n}$  es triangular superior y no singular, es decir,

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix},$$

con  $\det(A) = a_{11}a_{22}\dots a_{nn} \neq 0$ .

Entonces resolver  $Ax = b$  es equivalente a

$$\left\{ \begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & b_1 \\ a_{22}x_2 + \cdots + a_{2n}x_n & = & b_2 \\ \vdots & & \vdots \\ a_{ii}x_i + \cdots + a_{in}x_n & = & b_i \\ \vdots & & \vdots \\ a_{nn}x_n & = & b_n \end{array} \right. \implies \left\{ \begin{array}{lcl} x_n & = & b_n/a_{nn} \\ x_{n-1} & = & \frac{1}{a_{n-1,n-1}}(b_{n-1} - a_{n-1,n}x_n) \\ \vdots & & \vdots \\ x_i & = & \frac{1}{a_{ii}}(b_i - \sum_{j=i+1}^n a_{ij}x_j) \\ \vdots & & \vdots \\ x_1 & = & \frac{1}{a_{11}}(b_1 - \sum_{j=2}^n a_{1j}x_j) \end{array} \right.$$

**Algoritmo 3:** resolución de sistema lineal con matriz triangular superior.

```

input  $n, A, b$ 
for  $i = n, \dots, 1$  do
     $x_i \leftarrow \left( b_i - \sum_{j=i+1}^n a_{ij}x_j \right) / a_{ii}$  (no sumar si  $i = n$ )
output  $i, x_i$ 
end for
end
```

**Costo computacional:** de manera análoga al caso anterior se puede ver que la cantidad total de operaciones es:  $n^2$ , es decir  $\mathcal{O}(n^2)$  flops.

## Eliminación gaussiana

Dado un sistema lineal  $Ax = b$ , la idea consiste en transformarlo en otro sistema equivalente  $Ux = y$ , que sea más fácil de resolver. Recordar que sistemas equivalentes tienen la misma solución  $x$ . La matriz  $U$  del nuevo sistema será triangular superior, cuya resolución ya vimos que es muy sencilla. Recordemos también que existen 3 tipos de operaciones elementales por filas que permiten transformar un sistema lineal ampliado en otro sistema lineal ampliado equivalente:

- reemplazar una fila por un múltiplo escalar (no nulo) de ella misma ( $f_i \leftarrow \lambda f_i, \lambda \neq 0$ ).
- restar a una fila un múltiplo escalar (no nulo) de otra fila ( $f_i \leftarrow f_i - \lambda f_j, \lambda \neq 0$ ).
- intercambiar dos filas ( $f_i \leftrightarrow f_j$ ).

El método de eliminación gaussiana se basa fundamentalmente en las dos primeras operaciones elementales por filas. Es decir, se realizan  $(n - 1)$  pasos para obtener la matriz triangular  $U$  partiendo de  $A$ :

$$A \xrightarrow{\text{Paso 1}} A^{(1)} \xrightarrow{\text{Paso 2}} A^{(2)} \xrightarrow{\text{Paso 3}} \dots \xrightarrow{\text{Paso } n-1} A^{(n-1)} = U$$

y las mismas operaciones que se realizan en  $A$  se deben realizar en el vector  $b$ .

Para fijar ideas, veamos con detalles cómo se hacen estos pasos en un ejemplo numérico. Consideremos el sistema lineal:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix}$$

**Paso 1:** para pasar de la matriz  $A$  a la matriz  $A^{(1)}$  se deben realizar operaciones elementales por filas para obtener ceros en las posiciones de la primera columna a partir de la segunda fila hasta la última.

Para esto, denotemos  $f_i$  a la  $i$ -ésima fila de  $A$ . Llamamos a  $f_1$  como la fila pivote y a  $a_{11} = 6$  será llamado el elemento pivote. Entonces haremos las siguientes operaciones:

$$\begin{aligned} f_2 &\leftarrow f_2 - 2f_1 \\ f_3 &\leftarrow f_3 - (1/2)f_1 \\ f_4 &\leftarrow f_4 - (-1)f_1 \end{aligned}$$

Así obtenemos,

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ 21 \\ -26 \end{bmatrix}$$

**Paso 2:** para pasar de  $A^{(1)}$  a  $A^{(2)}$  se deben realizar operaciones elementales por filas para obtener ceros en las posiciones de la segunda columna a partir de la tercera fila hasta la última. La fila pivote será la nueva  $f_2$  y el elemento pivote será  $a_{22}^{(1)} = -4$ . Se realizan las siguientes operaciones:

$$\begin{aligned} f_3 &\leftarrow f_3 - 3f_2 \\ f_4 &\leftarrow f_4 - (-1/2)f_2 \end{aligned}$$

Así obtenemos,

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -21 \end{bmatrix}$$

**Paso 3:** por último, para pasar de  $A^{(2)}$  a  $A^{(3)} = U$  se deben realizar operaciones elementales por filas para modificar la última fila:

$$f_4 \leftarrow f_4 - 2f_3$$

Así obtenemos,

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -3 \end{bmatrix}$$

cuya solución es:  $x = (1, -3, -2, 1)$ .

**Observación 1:** notar que, en general, al pasar de  $A$  a  $A^{(1)}$  en el Paso 1, se obtiene una matriz de la forma

$$\left[ \begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{array} \right] \xrightarrow{\text{Paso 1}} \left[ \begin{array}{cccc} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & & \dots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{array} \right]$$

Más específicamente, se calcula

$$a_{ij}^{(1)} = \begin{cases} a_{ij} & \text{si } i = 1, j = 1, \dots, n \\ 0 & \text{si } i = 2, \dots, n, j = 1 \\ a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j} & \text{si } i = 2, \dots, n, j = 2, \dots, n \end{cases} \quad \text{y} \quad b_i^{(1)} = \begin{cases} b_i & \text{si } i = 1 \\ b_i - \frac{a_{i1}}{a_{11}} b_1 & \text{si } i = 2, \dots, n. \end{cases}$$

Los coeficientes  $m_{i,1} = a_{i1}/a_{11}$  para  $i = 2, \dots, n$  se denominan multiplicadores.

**Costo computacional:** calcularemos por separado la cantidad de sumas/restas y de productos/divisiones.

Para el Paso 1:

- **Sumas/restas:**  $(n-1)^2$  (submatriz  $(n-1) \times (n-1)$ ) y  $(n-1)$  (vector  $b$ ).  
En total son:  $(n-1)^2 + (n-1) = (n-1)(n-1+1) = n(n-1) = n^2 - n$ .
- **Productos/divisiones:**  $(n-1)$  (multiplicadores),  $(n-1)^2$  (submatriz  $(n-1) \times (n-1)$ ) y  $(n-1)$  (vector  $b$ ).  
En total son:  $(n-1)^2 + 2(n-1) = (n-1)(n-1+2) = (n-1)(n+1) = n^2 - 1$ .

**Observación 2:** para el Paso 2, es análogo a lo anterior, pero se trabaja con una matriz de un orden menor, es decir  $(n-1) \times (n-1)$ , y por lo tanto se requieren:  $(n-1)^2 - (n-1)$  sumas/restas y  $(n-1)^2 - 1$  productos/divisiones.

**Observación 3:** en general, para pasar de la matriz  $A^{(k-1)}$  a la matriz  $A^{(k)}$  en el Paso k, se calcula

$$a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)} & \text{si } i = 1, \dots, k, j = 1, \dots, n \\ 0 & \text{si } i = k+1, \dots, n, j = k \\ a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)} & \text{si } i = k+1, \dots, n, j = k+1, \dots, n \end{cases}$$

y

$$b_i^{(k)} = \begin{cases} b_i^{(k-1)} & \text{si } i = 1, \dots, k \\ b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)} & \text{si } i = k+1, \dots, n. \end{cases}$$

**Ejercicio:** mostrar que el costo total de operaciones del método de eliminación gaussiana para resolver un sistema  $Ax = b$ , con  $A \in \mathbb{R}^{n \times n}$  es  $\mathcal{O}(\frac{2}{3}n^3)$  flops.

Ayuda: recordar que  $\sum_{k=1}^n k = \frac{n(n+1)}{2}$  y  $\sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6}$ .

---

**Algoritmo 4:** eliminación gaussiana para obtener un sistema lineal con matriz triangular superior.

```
input n,A,b
for k = 1,...,n - 1 do
    for i = k + 1,...,n do
        if (akk = 0) STOP!
        m ← aik/akk
        for j = k + 1,...,n do
            aij ← aij - makj
        endfor (j)
        bi ← bi - mbk
    endfor (i)
endfor (k)
output A,b
end
```

**Observación 1:** notar que estamos almacenando la matriz triangular superior  $U$  en la misma matriz  $A$  y el vector modificado y se sobreescribe en el mismo vector  $b$ . No es necesario anular los elementos que están debajo de la diagonal, sólo se tiene en cuenta que se utilizarán los elementos de la parte triangular superior.

**Observación 2:** una vez obtenida la matriz triangular superior  $U$  y el vector  $y$ , se puede aplicar el algoritmo 3 para resolver sistemas con matrices triangulares superiores.

**Observación 3:** si para algún  $k$  se tiene que  $a_{kk} = 0$ , se podría intercambiar la fila  $k$  por otra fila  $j$ , con  $j = k + 1, \dots, n$  tal que  $a_{jk} \neq 0$ . Esto se llama estrategia de pivoteo y no se estudia en este curso.

**Observación 4:** notar que el ciclo ( $j$ ) del algoritmo, donde se van recorriendo las columnas de la matriz, comienza desde  $k + 1$  y no desde  $k$ . Esto se debe a que, en cada fila, el coeficiente que multiplica a la fila pivote (multiplicador) se define de manera de anular el coeficiente de la columna  $k$ . Entonces no tiene sentido hacer las operaciones en la columna  $k$  pues ya se sabe que el resultado va a dar cero.

Por último enunciamos un resultado que da condiciones suficientes para poder aplicar eliminación gaussiana para resolver un sistema lineal.

**Teorema 2.** Sean  $A_k = A(1 : k, 1 : k)$  es la submatriz de  $A$  formada por las primeras  $k$  filas y  $k$  columnas de  $A$  para todo  $k = 1, \dots, n$ . Si  $\det(A_k) \neq 0$  para  $k = 1, \dots, n$ , entonces es posible realizar el proceso completo de eliminación gaussiana y, por lo tanto, el sistema  $Ax = b$  tiene única solución.

# Clase 17 - Resolución de sistemas lineales (2)

## El problema

Dada  $A \in \mathbb{R}^{n \times n}$  y  $b \in \mathbb{R}^n$  hallar  $x_*$  solución de  $Ax = b$ .

## Factorización LU

Vimos que, usando eliminación gaussiana, esto se puede resolver con un costo computacional de  $\mathcal{O}(\frac{2}{3}n^3)$  flops. Este costo computacional se debe principalmente a las operaciones que se deben realizar para transformar la matriz  $A$  en una matriz triangular superior  $U$ , que es mucho mayor que la cantidad de operaciones que se realizan para transformar el vector  $b$  en otro vector  $y$ .

Ahora supongamos que se deben resolver varios sistema lineales  $Ax = b$ , con la misma matriz  $A$  y diferentes vectores  $b$ . Sería poco eficiente repetir todas las operaciones elementales por fila del proceso de eliminación gaussiana siendo que sólo el vector  $b$  es modificado y la matriz  $A$  se mantiene igual.

La idea de la factorización LU consiste en factorizar la matriz como un producto de dos matrices más simples:  $A = LU$ , donde  $L$  es una matriz triangular inferior con elementos diagonales iguales a uno ( $l_{ii} = 1, i = 1, \dots, n$ ) y  $U$  triangular superior. Así:

$$Ax = b \iff LUx = b \iff \begin{cases} Ly = b \\ Ux = y \end{cases}$$

Es decir que, conocidas las matrices  $L$  y  $U$  de la factorización de  $A$ , en lugar resolver  $Ax = b$ , se deben resolver dos sistemas triangulares para cada nuevo vector  $b$ .

Estas matrices  $L$  y  $U$  se obtienen a partir de  $A$  de una forma constructiva, como veremos a continuación. La idea es escribir  $A = LU$ , o equivalentemente  $LU = A$ , y tratar de despejar los coeficientes de  $L$  y de  $U$ , usando inducción:

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ l_{21} & 1 & 0 & \dots & 0 \\ l_{31} & l_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ 0 & 0 & u_{33} & \dots & u_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & u_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix}.$$

Vamos a determinar la primera fila de  $U$  y la primera columna de  $L$ .

Si multiplicamos la primera fila de  $L$  por la columna  $j$  de  $U$  para  $j = 1, \dots, n$ , obtenemos:

$$a_{1j} = 1u_{1j} + 0u_{2j} + 0u_{3j} + \dots + 0u_{nj},$$

por lo tanto,

$$u_{1j} = a_{1j}, \quad \text{para } j = 1, \dots, n.$$

Si multiplicamos cada fila de  $L$  por la primera columna de  $U$ , obtenemos:

$$a_{i1} = l_{i1}u_{11}, \quad \text{para } i = 2, \dots, n,$$

por lo tanto,

$$l_{i1} = \frac{a_{i1}}{u_{11}}, \quad \text{para } i = 2, \dots, n.$$

Ahora, supongamos conocidas las  $(k - 1)$  filas de  $U$  y  $(k - 1)$  columnas de  $L$  y vamos a determinar la fila  $k$  de  $U$  y la columna  $k$  de  $L$ .

Si multiplicamos la fila  $k$  de  $L$  por la columna  $j$  de  $U$  para  $j = k, \dots, n$ , obtenemos:

$$a_{kj} = \sum_{m=1}^{k-1} l_{km} u_{mj} + l_{kk} u_{kj},$$

por lo tanto,

$$u_{kj} = a_{kj} - \sum_{m=1}^{k-1} l_{km} u_{mj}, \quad \text{para } j = k, \dots, n.$$

Si multiplicamos cada fila de  $L$ , desde la  $k + 1$  hasta la última, por la columna  $k$  de  $U$ , obtenemos:

$$a_{ik} = \sum_{m=1}^{k-1} l_{im} u_{mk} + l_{ik} u_{kk},$$

por lo tanto,

$$l_{ik} = \frac{1}{u_{kk}} (a_{ik} - \sum_{m=1}^{k-1} l_{im} u_{mk}) \quad \text{para } i = k + 1, \dots, n.$$

A continuación veremos el algoritmo para realizar la factorización LU.

**Algoritmo:** Factorización LU.

```

input  $n, A = (a_{ij})$ 
for  $k = 1, \dots, n$  do
    for  $j = k, \dots, n$  do
         $u_{kj} \leftarrow a_{kj} - \sum_{m=1}^{k-1} l_{km} u_{mj}$ 
    end for ( $j$ )
    for  $i = k + 1, \dots, n$  do (no ejecutar si  $k = n$ )
         $l_{ik} \leftarrow \frac{1}{u_{kk}} (a_{ik} - \sum_{m=1}^{k-1} l_{im} u_{mk})$ 
    end for ( $i$ )
end for ( $k$ )
output  $L = (l_{ij}), U = (u_{ij})$ 
end
```

El siguiente resultado da una condición suficiente para poder realizar la factorización LU.

**Teorema 1.** Sea  $A \in \mathbb{R}^{n \times n}$  tal que las submatrices  $A_k = A(1:k, 1:k)$  para  $k = 1, \dots, n-1$  son no singulares. Entonces existen únicas matrices  $L, U \in \mathbb{R}^{n \times n}$  tal que  $L$  es triangular inferior con 1 en la diagonal ( $l_{ii} = 1, i = 1, \dots, n$ ) y  $U$  es triangular superior. Además,  $\det(A) = u_{11}u_{22}\dots u_{nn}$ .

*Demostración.* La prueba se hará por inducción en la dimensión de la matriz.

Si  $n = 1$ , entonces si  $A = LU$ , basta tomar  $L = [1]$  y  $U = [a_{11}]$ . También se cumple que  $\det(A) = u_{11}$ .

Supongamos que la factorización es válida para dimensión  $k - 1$ , es decir, dada  $A_{k-1} \in \mathbb{R}^{(k-1) \times (k-1)}$ , existen  $L_{k-1}, U_{k-1} \in \mathbb{R}^{(k-1) \times (k-1)}$  tal que  $A_{k-1} = L_{k-1}U_{k-1}$ . Veamos para  $k$ , o sea, veremos que para una matriz  $A_k \in \mathbb{R}^{k \times k}$  existen  $L_k, U_k \in \mathbb{R}^{k \times k}$  tal que  $A_k = L_kU_k$ . Para determinar esto vamos a particionar convenientemente:

$$\left[ \begin{array}{c|c} A_{k-1} & a \\ \hline b^T & c \end{array} \right] = \left[ \begin{array}{c|c} L_{k-1} & 0 \\ \hline d^T & 1 \end{array} \right] \left[ \begin{array}{c|c} U_{k-1} & e \\ \hline 0^T & u_{kk} \end{array} \right],$$

donde  $a, b, d, e \in \mathbb{R}^{k-1}$  y  $c, u_{kk} \in \mathbb{R}$ .

Realizando los productos en bloques del lado derecho e igualando al lado izquierdo, obtenemos:

$$A_{k-1} = L_{k-1}U_{k-1} \quad (1)$$

$$L_{k-1}e = a \quad (2)$$

$$d^T U_{k-1} = b^T \quad (3)$$

$$d^T e + u_{kk} = c \quad (4)$$

De (2), y como  $L_{k-1}$  es triangular inferior con unos en la diagonal, es claro que existe un único  $e \in \mathbb{R}^{k-1}$ .

De (3), aplicando transpuesta a ambos lados,  $(d^T U_{k-1})^T = b$ , es decir,  $U_{k-1}^T d = b$ , y como  $U_{k-1}$  tiene inversa por hipótesis inductiva y por que las matrices  $A_k$  son no singulares, entonces existe un único  $d \in \mathbb{R}^{k-1}$ .

Luego, de (4),  $u_{kk} = c - d^T e$  está bien definido y está únicamente determinado.

De esta manera se ha probado que existen  $L_k$  y  $U_k$  tales que  $A_k = L_kU_k$ .

Por último, como  $A = LU$ , entonces

$$\det(A) = \det(LU) = \det(L) \det(U) = 1 \cdot (u_{11}u_{22} \dots u_{nn}) = u_{11}u_{22} \dots u_{nn}.$$

□

**Observación 1:** El costo de realizar la factorización LU es el mismo que para realizar la eliminación gaussiana, es decir,  $\mathcal{O}(\frac{2}{3}n^3)$ .

**Observación 2:** Si la matriz  $A$  no será usada posteriormente, una implementación eficiente de la factorización LU, podría sobreescribir las matrices  $L$  y  $U$  sobre la misma  $A$ , teniendo en cuenta que los coeficientes de la diagonal de  $L$  son todos iguales a 1. Más aún, se puede ver que los elementos de la matriz  $L$  debajo de la diagonal son los multiplicadores que aparecen en la eliminación gaussiana.

**Observación 3:** este es el método más eficiente para calcular el determinante de una matriz. El costo computacional será aproximadamente el mismo que para realizar la factorización LU.

---

Recordemos la clásica regla para calcular el determinante para una matriz  $2 \times 2$ :

$$\det(A) = \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

Si usáramos esta regla se requieren 2 productos y una resta para calcular el determinante. Lamentablemente, esta cantidad aumenta increíblemente cuando crece el orden  $n$  de la matriz. Por ejemplo, si  $A$  fuera una matriz  $20 \times 20$ . Entonces se requieren del orden de  $20! \approx 2.4 \cdot 10^{18}$  operaciones. Si se usara una computadora que realiza  $10^9$  operaciones en punto flotante por segundo, entonces se necesitarían 76 años para calcular ese determinante.

**Observación 4:** nunca se debe calcular la inversa de una matriz. Siempre es mucho más eficiente reformular el problema para resolver sistemas lineales. Sin embargo, si el único objetivo del problema es obtener la matriz de una inversa de una matriz, la forma más eficiente consiste en realizar una factorización LU de la matriz  $A$  y luego resolver  $n$  sistemas lineales de la forma  $LUX = e_i$ , donde  $e_i$  es el  $i$ -ésimo vector de la base canónica que tiene un 1 en la posición  $i$ , y 0 en las otras componentes, para  $i = 1, \dots, n$ .

A continuación veremos algunas definiciones y resultados que serán de utilidad para la clase siguiente.

**Definición 1.** Una norma vectorial en  $\mathbb{R}^n$  es una función que asigna a cada  $x \in \mathbb{R}^n$  un número real no negativo denotado por  $\|x\|$  y llamado norma de  $x$ , tal que se satisfacen las siguientes tres propiedades para todo  $x, y \in \mathbb{R}^n$  y para todo  $\alpha \in \mathbb{R}$ :

1.  $\|x\| > 0$  si  $x \neq 0$ , y  $\|0\| = 0$ ;
2.  $\|\alpha x\| = |\alpha| \|x\|$ ;
3.  $\|x + y\| \leq \|x\| + \|y\|$ .

**Ejemplos:**

1. norma euclídea (o norma 2):  $\|x\|_2 = \left( \sum_{i=1}^n |x_i|^2 \right)^{1/2}$ ;

2. norma 1:  $\|x\|_1 = \sum_{i=1}^n |x_i|$ ;

3. para  $p \geq 1$ , norma  $p$ :  $\|x\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}$ ;

4. norma infinito:  $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$ ;

**Definición 2.** Dados dos vectores  $x, y \in \mathbb{R}^n$  y una norma vectorial  $\|\cdot\|$  se define la distancia entre  $x$  e  $y$  por:

$$d(x, y) = \|x - y\|.$$

El concepto de normas puede extenderse al conjunto de matrices.

**Definición 3.** Una **norma matricial** en  $\mathbb{R}^{n \times n}$  es una función que asigna a cada  $A \in \mathbb{R}^{n \times n}$  un número real no negativo denotado por  $\|A\|$  y llamado norma de  $A$ , tal que se satisfacen las siguientes cuatro propiedades para todo  $A, B \in \mathbb{R}^{n \times n}$  y para todo  $\alpha \in \mathbb{R}$ :

1.  $\|A\| > 0$  si  $A \neq 0$ , y  $\|0\| = 0$ ;
2.  $\|\alpha A\| = |\alpha| \|A\|$ ;
3.  $\|A + B\| \leq \|A\| + \|B\|$ ;
4.  $\|AB\| \leq \|A\| \|B\|$ .

**Ejemplo:** norma de Frobenius

$$\|A\|_F = \left( \sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2 \right)^{1/2}.$$

**Definición 4.** Dada una norma vectorial en  $\mathbb{R}^n$  y una matriz  $A \in \mathbb{R}^{n \times n}$  se define la norma matricial inducida por

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

**Proposición 1.** Sean  $A \in \mathbb{R}^{n \times n}$  y  $\|\cdot\|$  una norma vectorial en  $\mathbb{R}^n$  que induce una norma matricial, entonces:

1.  $\|Ax\| \leq \|A\| \|x\|$  para todo  $x \in \mathbb{R}^n$ ;
2. existe  $\bar{x}$  con  $\|\bar{x}\| = 1$  tal que  $\|A\bar{x}\| = \|A\|$ .

*Demostración.* - Probemos la primera parte.

Si  $x = 0$ , entonces  $\|Ax\| = 0 = \|A\| \|x\|$ .

Si  $x \neq 0$ , entonces  $\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \geq \frac{\|Ax\|}{\|x\|}$  para todo  $x$ , y por lo tanto  $\|Ax\| \leq \|A\| \|x\|$  para todo  $x \in \mathbb{R}^n$ .

- Ahora probemos la segunda parte.

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x \neq 0} \left\| \frac{Ax}{\|x\|} \right\| = \sup_{x \neq 0} \left\| A \left( \frac{x}{\|x\|} \right) \right\| = \sup_{y: \|y\|=1} \|Ay\|.$$

Como  $S = \{x \in \mathbb{R}^n | \|x\| = 1\}$  es cerrado y acotado (compacto), la función continua  $x \rightarrow \|Ax\|$  alcanza sus valores extremos, en particular existe un  $\bar{x}$  tal que  $\|\bar{x}\| = 1$  y

$$\|A\| = \sup_{y: \|y\|=1} \|Ay\| = \max_{y: \|y\|=1} \|Ay\| = \|A\bar{x}\|.$$

□

**Ejemplos:**

1.  $\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|$ ;

---

2.  $\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$ ;

3.  $\|A\|_2 = \sqrt{\rho(A^T A)}$ ,  
donde  $\rho(B)$  es el radio espectral de  $B$  y se define por  $\max\{|\lambda| : \lambda \text{ es autovalor de } B\}$ .

A. Numérico - FAMAF 2021

# Clase 18 - Resolución de sistemas lineales (3)

## El problema

Dada  $A \in \mathbb{R}^{n \times n}$  y  $b \in \mathbb{R}^n$  hallar  $x_*$  solución de  $Ax = b$ .

Hasta ahora vimos dos métodos numéricos para resolver este problema: eliminación gaussiana y factorización LU. Ambos métodos son considerados métodos directos y se sabe que luego de un número finito de pasos se obtiene una solución, salvo errores de redondeo. Por otro lado, existe otra familia de métodos para resolver este problema llamados iterativos o indirectos.

## Métodos iterativos

Los métodos iterativos para sistemas lineales generan una sucesión de vectores  $\{x^{(k)}\}$ , a partir de un vector inicial  $x^{(0)}$ , que convergen a la solución de  $Ax = b$ , bajo adecuadas hipótesis. La convergencia significa que esta sucesión se detiene cuando se alcanza una precisión determinada luego de un cierto número de iteraciones. El éxito computacional de estos métodos requiere que procedimiento sea simple y de bajo costo computacional. En general estos métodos son adecuados para matrices grandes y que tienen muchos coeficientes iguales a cero (matrices ralas).

Estudiaremos dos métodos iterativos muy conocidos: método de Jacobi y método de Gauss-Seidel. Presentaremos brevemente estos métodos y algunos resultados simples de convergencia. Para fijar ideas veremos inicialmente un ejemplo muy sencillo.

**Ejemplo:** consideremos el sistema lineal

$$\begin{bmatrix} 7 & -6 \\ -8 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \end{bmatrix} \iff \begin{cases} 7x_1 - 6x_2 = 3 \\ -8x_1 + 9x_2 = -4 \end{cases}, \quad (1)$$

cuya solución es  $x_* = (0.2, -0.2666\dots)$ . A partir del sistema lineal (1), podemos despejar las variables:

$$\begin{cases} x_1 = \frac{6}{7}x_2 + \frac{3}{7} \\ x_2 = \frac{8}{9}x_1 - \frac{4}{9} \end{cases} \quad (2)$$

Dada una aproximación  $x^{(k-1)} = (x_1^{(k-1)}, x_2^{(k-1)})$ , y de (2) se puede generar el siguiente método iterativo:

$$\begin{cases} x_1^{(k)} = \frac{6}{7}x_2^{(k-1)} + \frac{3}{7} \\ x_2^{(k)} = \frac{8}{9}x_1^{(k-1)} - \frac{4}{9} \end{cases},$$

el cual es conocido como método de Jacobi.

También se puede generar otro método iterativo de la siguiente manera:

$$\begin{cases} x_1^{(k)} = \frac{6}{7}x_2^{(k-1)} + \frac{3}{7} \\ x_2^{(k)} = \frac{8}{9}x_1^{(k)} - \frac{4}{9} \end{cases},$$

el cual es conocido como método de Gauss-Seidel.

La tabla siguiente muestra algunas iteraciones de ambos métodos, comenzando en ambos casos con  $x^{(0)} = (x_1^{(0)}, x_2^{(0)}) = (0, 0)$ . Recordemos que la solución es:  $x_* = (0.2, -0.2666\dots)$ .

It.	M. de Jacobi	M. de Gauss-Seidel
$k$	$x_1^{(k)}$	$x_2^{(k)}$
0	0	0
10	0.14865	-0.19820
20	0.18682	-0.24909
30	0.19662	-0.26215
40	0.19913	-0.26551
50	0.19978	-0.26637

Los métodos de Jacobi y Gauss-Seidel son conocidos también como métodos de separación (splitting) y parten de una misma idea básica, que consiste en escribir la matriz como  $A = M - N$ , donde  $M$  es una matriz no singular. Así tenemos que

$$\begin{aligned} Ax &= b \\ (M - N)x &= b \\ Mx &= Nx + b \end{aligned} \tag{3}$$

$$x = M^{-1}(Nx + b) \tag{4}$$

$$x = (M^{-1}N)x + M^{-1}b. \tag{5}$$

Dado  $x^{(0)} = (x_1^{(0)}, x_2^{(0)})$ , la última ecuación sugiere definir un método iterativo de la forma:

$$x^{(k+1)} = (M^{-1}N)x^{(k)} + M^{-1}b, \quad \text{para } k \geq 0. \tag{6}$$

A continuación veremos algunos resultados de convergencia de estos métodos iterativos.

**Teorema 1.** *Sea  $b \in \mathbb{R}^n$  y  $A = M - N \in \mathbb{R}^{n \times n}$  donde  $A$  y  $M$  son matrices no singulares. Si  $\|M^{-1}N\| < 1$  para alguna norma matricial inducida entonces la sucesión generada por la ecuación (6) converge a la solución de  $Ax = b$  para cualquier vector inicial  $x^{(0)}$ .*

*Demostración.* Restando la ecuación (6) de (5), donde en (5) ponemos la solución  $x^*$  se obtiene:

$$x^{(k+1)} - x^* = (M^{-1}N)(x^{(k)} - x^*), \quad \text{para } k \geq 0. \tag{7}$$

Ahora, utilizando alguna norma matricial inducida, se obtiene:

$$\|x^{(k+1)} - x^*\| \leq \|(M^{-1}N)\| \|x^{(k)} - x^*\|, \quad \text{para } k \geq 0. \tag{8}$$

Luego, restando este último paso se tiene:

$$\|x^{(k+1)} - x^*\| \leq \|(M^{-1}N)\|^{k+1} \|x^{(0)} - x^*\|, \quad \text{para } k \geq 0. \tag{9}$$

Usando que  $\|(M^{-1}N)\| < 1$ , se tiene que

$$\lim_{k \rightarrow \infty} \|x^{(k)} - x^*\| = 0,$$

para cualquier vector inicial  $x^{(0)}$ . □

La principal diferencial entre el método de Jacobi y el de Gauss-Seidel consiste en cómo es elegida la matriz  $M$ . Por simplicidad vamos a descomponer la matriz  $A$  de la siguiente forma:  $A = L + D + U$ , donde  $L$  es la parte triangular inferior de  $A$  (sin la diagonal),  $D$  es la diagonal de  $A$  y  $U$  es la parte triangular superior de  $A$  (sin la diagonal). Es importante no confundir con las matrices  $L$  y  $U$  de la factorización LU.

### Método de Jacobi

En el método de Jacobi se toma  $M = D$ , y por lo tanto,

$$N = M - A = D - (L + D + U) = -(L + U),$$

esto es,

$$M = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix} \quad \text{y} \quad N = -\begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ a_{21} & 0 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix}.$$

Si miramos la  $i$ -ésima componente en (3) tenemos que

$$\begin{aligned} [Dx^{(k+1)}]_i &= [b - (L + U)x^{(k)}]_i \\ a_{ii}x_i^{(k+1)} &= b_i - [a_{i1} \dots a_{i,i-1} 0 a_{i,i+1} \dots a_{in}] \begin{bmatrix} x_1^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix} \\ a_{ii}x_i^{(k+1)} &= b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \\ x_i^{(k+1)} &= \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)}), \end{aligned}$$

para  $i = 1, \dots, n$ .

Ahora veremos un pseudocódigo del método de Jacobi. Notar que en el algoritmo usaremos  $x$  para representar el vector inicial y sobreescribiremos las sucesivas aproximaciones en el mismo vector. La variable  $MaxIt$  representa el número máximo de iteraciones permitidas y  $Xtol$  la tolerancia para la norma de la diferencia de dos aproximaciones vectoriales sucesivas.

**Algoritmo:** método de Jacobi.

**input**  $n, A, b, x, MaxIt, Xtol$

**for**  $k = 1, \dots, MaxIt$  **do**

**for**  $i = 1, \dots, n$  **do**

$$u_i \leftarrow \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^n a_{ij}x_j)$$

**end for** ( $i$ )

**if**  $\|u - x\| < Xtol$  **STOP, output:**  $u$  es la solución,  $k$  iteración.

**for**  $i = 1, \dots, n$  **do**

$x_i \leftarrow u_i$

**end for** ( $i$ )

**end for** ( $k$ )

**output**  $x$

**end**

Antes de enunciar un resultado de convergencia, daremos una definición.

**Definición 1.** Una matriz  $A$ , de orden  $n \times n$ , es **diagonalmente dominante** si

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|, \quad \text{para } i = 1, \dots, n.$$

**Teorema 2.** Si  $A$  es diagonalmente dominante, entonces la sucesión generada por el método de Jacobi converge a la solución de  $Ax = b$ , para cualquier vector inicial  $x^{(0)} \in \mathbb{R}^n$ .

*Demostración.* En el método de Jacobi, la matriz  $M$  es la diagonal de  $A$  y debe ser inversible para que el método esté bien definido, por lo que  $a_{ii} \neq 0, i = 1, \dots, n$ . La matriz de iteración está dada por

$$\begin{aligned} M^{-1}N &= - \begin{bmatrix} 1/a_{11} & 0 & \dots & 0 \\ 0 & 1/a_{22} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1/a_{nn} \end{bmatrix} \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ a_{21} & 0 & \dots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix} \\ &= - \begin{bmatrix} 0 & a_{12}/a_{11} & \dots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & \dots & a_{2n}/a_{22} \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & \dots & 0 \end{bmatrix} \end{aligned} \quad (10)$$

Luego,

$$\|M^{-1}N\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n \frac{|a_{ij}|}{|a_{ii}|} < 1,$$

pues  $A$  es diagonalmente dominante. Finalmente, la convergencia es una consecuencia directa del teorema anterior.

□

## Método de Gauss-Seidel

En el método de Gauss-Seidel se toma  $M = L + D$ , y por lo tanto,

$$N = M - A = L + D - (L + D + U) = -U,$$

esto es,

$$M = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{bmatrix} \quad \text{y} \quad N = - \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & 0 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

Luego por (3) tenemos que

$$\begin{aligned}(L+D)x^{(k+1)} &= b - Ux^{(k)} \\ Dx^{(k+1)} &= b - Lx^{(k+1)} - Ux^{(k)}.\end{aligned}$$

Así, analizando la  $i$ -ésima componente, obtenemos

$$\begin{aligned}[Dx^{(k+1)}]_i &= [b - Lx^{(k+1)} - Ux^{(k)}]_i \\ a_{ii}x_i^{(k+1)} &= b_i - [a_{i1} \dots a_{i,i-1} 0 \dots 0] \begin{bmatrix} x_1^{(k+1)} \\ \vdots \\ x_n^{(k+1)} \end{bmatrix} - [0 \dots 0 a_{i,i+1} \dots a_{in}] \begin{bmatrix} x_1^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix} \\ a_{ii}x_i^{(k+1)} &= b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \\ x_i^{(k+1)} &= \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)}),\end{aligned}$$

para  $i = 1, \dots, n$ .

Ahora veremos un pseudocódigo del método de Gauss-Seidel. Notar que en el algoritmo usaremos  $x$  para representar el vector inicial y sobreescribiremos las sucesivas aproximaciones en el mismo vector. La variable  $MaxIt$  representa el número máximo de iteraciones permitidas y  $Xtol$  la tolerancia para la norma de la diferencia de dos aproximaciones vectoriales sucesivas.

**Algoritmo:** método de Gauss-Seidel.

```

input  $n, A, b, x, MaxIt, Xtol$ 
for  $k = 1, \dots, MaxIt$  do
    for  $i = 1, \dots, n$  do
         $u_i \leftarrow 0$ 
    end for ( $i$ )
    for  $i = 1, \dots, n$  do
         $u_i \leftarrow \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}u_j - \sum_{j=i+1}^n a_{ij}x_j)$ 
    end for ( $i$ )
    if  $\|u - x\| < Xtol$  STOP, output:  $u$  es la solución,  $k$  iteración.
    for  $i = 1, \dots, n$  do
         $x_i \leftarrow u_i$ 
    end for ( $i$ )
end for ( $k$ )
output  $x$ 
end
```

El resultado de convergencia que enunciamos para este método es similar al del método de Jacobi, aunque su demostración no es tan directa.

---

**Teorema 3.** Si  $A$  es diagonalmente dominante, entonces la sucesión generada por el método de Gauss-Seidel converge a la solución de  $Ax = b$ , para cualquier vector inicial  $x^{(0)} \in \mathbb{R}^n$ .

**Observación:** existen resultados que muestran que si ambos métodos convergen, el método de Gauss-Seidel lo hace más rápido que el método de Jacobi. Esto es razonable, pues al calcular la  $i$ -ésima componente, en el método de Gauss-Seidel, se utiliza información recientemente calculada en la misma iteración. Por otro lado, si ambos métodos divergen, el método de Gauss-Seidel lo hace más rápido que el método de Jacobi. Por esta razón el método era más atractivo computacionalmente que el de Jacobi. Sin embargo, el método de Jacobi ha vuelto ha cobrar popularidad en las últimas décadas, pues es paralelizable mientras que el método de Gauss-Seidel no lo es.

Por último vamos a enunciar un resultado auxiliar y un teorema que, bajo ciertas hipótesis, asegura la convergencia de los métodos como Jacobi o Gauss-Seidel.

**Teorema 4.** Para cada matriz  $A \in \mathbb{R}^{n \times n}$ , se cumple que  $\rho(A)$  es igual al  $\inf\{\|A\|\}$  sobre todas las normas matriciales inducidas.

**Teorema 5.** Una condición necesaria y suficiente para que la sucesión generada por el método iterativo  $x^{(k+1)} = (M^{-1}N)x^{(k)} + M^{-1}b$ , para  $k \geq 0$ , converja a la única solución de  $Ax = b$  para todo  $x^{(0)}$  inicial, es que  $\rho(M^{-1}N) < 1$ .

# Clase 19 - Programación lineal

## Introducción

La programación lineal (PL) es uno de los mecanismos más naturales para tratar una gran cantidad de problemas del mundo real de un modo sencillo. Es una subárea de Optimización donde, como es de esperar, todas las funciones involucradas para formular el problema son lineales. Aunque las funciones lineales son de las más simples, existen muchos problemas de economía, producción, planning, logística, redes, scheduling, transporte y otras áreas que pueden formularse como un problema de programación lineal. Además, los aspectos matemáticos y computacionales de PL son muy interesantes. Por un lado, la matemática de PL es sencilla, poderosa y elegante y utiliza enfoques geométricos y de Álgebra lineal, que están conectados entre sí. Por otro lado, los aspectos computacionales son muy importantes por las potenciales aplicaciones y variantes de los algoritmos para resolver el problema. El método más conocido, es el Simplex, y fue propuesto por Dantzig en 1947. En estas clases, por una cuestión de tiempo, presentaremos algunas nociones básicas de PL y el método Simplex en una versión resumida. Para entender en qué consiste un problema de PL consideraremos inicialmente un ejemplo.

**Ejemplo:** un agricultor debe comprar fertilizantes (abono) para sus campos. El ingeniero agrónomo le dijo que cada kilogramo de fertilizante le alcanza para  $10m^2$  de su campo, y debido a las características propias de esas tierras, el fertilizante debe contener (al menos): 3 g de fósforo (P), 1.5 g de nitrógeno (N) y 4 g de potasio (K) por cada  $10m^2$ . En el mercado existen 2 tipos de fertilizantes: T1 y T2. El fertilizante T1 contiene 3 g de P, 1 g de N y 8 g de K y cuesta \$ 10 por kilogramo. En cambio, el fertilizante T2 contiene 2 g de P, 3 g de N y 2 g de K y cuesta \$ 8 por kilogramo. El agricultor desea saber cuántos kilogramos de cada fertilizante debe comprar, por cada  $10m^2$  de campo, de modo de minimizar el costo total cubriendo los requerimientos de su suelo.

Vamos a resumir toda esta información en la siguiente tabla y a continuación definiremos el problema:

	Tipo 1 ( $x_1$ )	Tipo 2 ( $x_2$ )	Necesidades mínimas
P (fósforo)	3	2	3
N (nitrógeno)	1	3	1.5
K (potasio)	8	2	4
Costo	10	8	

### Incógnitas:

$x$  : cantidad de fertilizante de tipo 1 (T1), en kg.

$y$  : cantidad de fertilizante de tipo 2 (T2), en kg.

### Restricciones (requerimientos):

$$\begin{aligned}3x + 2y &\geq 3 \\x + 3y &\geq 1.5 \\8x + 2y &\geq 4 \\x \geq 0, y \geq 0 & \quad (\text{no negatividad})\end{aligned}$$

Las últimas restricciones de negatividad son razonables pues la cantidad de fertilizantes no puede ser negativa.

**Función objetivo (costo):**  $f(x, y) = 10x + 8y$ .

**Objetivo:** Minimizar  $f(x, y) = 10x + 8y$ .

Así, este problema puede ser formulado en la siguiente forma:

$$\begin{array}{ll} \text{Minimizar} & f(x, y) = 10x + 8y \\ \text{sujeto a} & 3x + 2y \geq 3 \\ & x + 3y \geq 1.5 \\ & 8x + 2y \geq 4 \\ & x \geq 0, y \geq 0 \end{array}$$

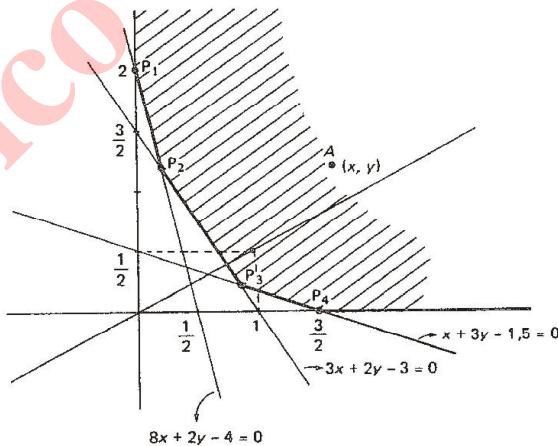
o de una forma más compacta:

$$\begin{array}{ll} \text{Minimizar} & c^T x \\ \text{sujeto a} & Ax \geq b \\ & x \geq 0 \end{array}$$

donde  $c = (10, 8)$ ,  $x = (x_1, x_2)$  y

$$A = \begin{bmatrix} 3 & 2 \\ 1 & 3 \\ 8 & 2 \end{bmatrix} \quad y \quad b = \begin{bmatrix} 3 \\ 1.5 \\ 4 \end{bmatrix}$$

Este problema se puede representar gráficamente como se ve en la Figura 1. La región sombreada se llama **región factible** y establece el conjunto de posibles soluciones, es decir los posibles valores de  $x$  e  $y$  que satisfacen las restricciones.



**Figura 1:** Representación gráfica del ejemplo de los fertilizantes.

El objetivo será encontrar la **solución óptima** en la región factible, es decir  $(x_*, y_*)$  en la región factible que minimiza la función objetivo (costo)  $f$ .

En general, los problemas de PL están definidos por:

- un vector de variables  $x \in \mathbb{R}^n$  que son no negativas, es decir,  $x_i \geq 0$  para  $i = 1, \dots, n$ .
- una función objetivo lineal  $f$ , que deberá ser minimizada o maximizada. Como esta función debe ser lineal será de la forma  $f(x) = c^T x = c_1 x_1 + \dots + c_n x_n$ .
- un conjunto de restricciones que deben satisfacer las variables. Estas restricciones estarán dadas por ecuaciones (igualdades) o inecuaciones (desigualdades) lineales.

- 
- una región (o conjunto) factible definida por la **intersección** de todas las restricciones lineales que deben satisfacer las variables.

Por lo tanto, los problemas de PL se escriben de la siguiente forma:

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeto a} & Cx = d \\ & Rx \geq s \\ & x \geq 0 \end{array} \quad (1)$$

donde  $f$  es la función objetivo a minimizar (o maximizar); el sistema lineal  $Cx = d$  corresponde al conjunto de ecuaciones lineales de igualdad que deben satisfacer las variables;  $Rx \geq s$  corresponde al conjunto de restricciones lineales de desigualdad que deben satisfacer las variables (podrían ser con  $\leq$  en vez de  $\geq$ ); y por último, las restricciones de no negatividad  $x \geq 0$ . Observar que en el ejemplo anterior no había restricciones de igualdad.

Se denomina **forma estándar** cuando el problema de PL es formulado como

$$\begin{array}{ll} \text{Minimizar} & c^T x \\ \text{sujeto a} & Ax = b \\ & x \geq 0 \end{array} \quad (2)$$

Veamos que cualquier problema de PL en formato general puede ser llevado a la forma estándar:

- Conversión de maximización en minimización: si el problema original fuera:

$$\text{maximizar } z = 4x_1 - 3x_2 + 6x_3 = c^T x,$$

multiplicando sólo la función objetivo por  $(-1)$  se convierte en

$$\text{minimizar } \hat{z} = -4x_1 + 3x_2 - 6x_3 = -c^T x.$$

Una vez que el problema de minimización es resuelto, el valor de la función objetivo se debe multiplicar por  $(-1)$ , de manera que  $z_* = -\hat{z}_*$ , aunque las variables óptimas son las mismas.

- Si alguna componente  $b_i < 0$  para algún  $i$ , se debe multiplicar por  $(-1)$  a toda la restricción, cambiando el signo de la desigualdad si la hubiera.
- Si la cota inferior de una variable no fuera cero, por ejemplo  $x_1 \geq 5$ , se puede definir  $\hat{x}_1 = x_1 - 5 \geq 0$ .
- Si una variable tuviera una cota superior, por ejemplo  $x_1 \leq 7$ , se la tratará como cualquiera de las otras restricciones.
- Si una variable fuera irrestricta (libre), por ejemplo la variable  $x_2$ , entonces debe reemplazarse por:  $x_2 = \hat{x}_2 - x'_2$ , donde  $\hat{x}_2, x'_2 \geq 0$ .
- Una restricción de desigualdad con  $\leq$  se convierte en una restricción de igualdad agregando una **variable de holgura** (slack), por ejemplo:  $2x_1 + 7x_2 - 3x_3 \leq 10$  es equivalente a  $2x_1 + 7x_2 - 3x_3 + s_1 = 10$  con  $s_1 \geq 0$ .

- Una restricción de desigualdad con  $\geq$  se convierte en una restricción de igualdad agregando una **variable de exceso**, por ejemplo:  $6x_1 - 2x_2 + 4x_3 \geq 15$  es equivalente a  $6x_1 - 2x_2 + 4x_3 - s_1 = 15$  con  $s_1 \geq 0$ .
- Una restricción de igualdad puede convertirse en 2 restricciones de desigualdad, por ejemplo:  $2x_1 + 3x_2 = 1$  es equivalente a  $2x_1 + 3x_2 \leq 1$  y  $2x_1 + 3x_2 \geq 1$ .

Entonces, el problema del ejemplo puede ser escrito en la forma estándar:

$$\begin{array}{ll} \text{Minimizar} & z = 10x_1 + 8x_2 \\ \text{sujeto a} & 3x_1 + 2x_2 - s_1 = 3 \\ & x_1 + 3x_2 - s_2 = 1.5 \\ & 8x_1 + 2x_2 - s_3 = 4 \\ & x_1, x_2, s_1, s_2, s_3 \geq 0 \end{array}$$

Ahora veremos algunas definiciones y resultados sobre la geometría del problema de PL.

**Definición 1.** Dados  $a = (a_1, a_2, \dots, a_n) \in \mathbb{R}^n$ ,  $a \neq 0$ , y  $b \in \mathbb{R}$ , se definen:

1.  $\{x \in \mathbb{R}^n | a_1x_1 + a_2x_2 + \dots + a_nx_n = b\}$  es llamado **hiperplano** (afín si  $b \neq 0$ ) del espacio  $n$ -dimensional;
2.  $\{x \in \mathbb{R}^n | a_1x_1 + a_2x_2 + \dots + a_nx_n \leq b\}$  es llamado **semiespacio** (cerrado) del espacio  $n$ -dimensional.

**Observación:** un hiperplano divide al espacio en dos semiespacios.

**Ejemplos:**

1.  $x_1 + x_2 = 2$  es un hiperplano en  $\mathbb{R}^2$ ;
2.  $x_1 + x_2 + x_3 = 1$  y  $x_3 = 2$  son 2 hiperplanos afines en  $\mathbb{R}^3$ .

**Definición 2.** Un conjunto  $S$  de  $\mathbb{R}^n$  es **convexo** si para todo par de puntos distintos  $x, y \in S$ , el segmento que los une también está contenido en  $S$ , es decir,  $(\alpha x + (1 - \alpha)y) \in S$  para todo  $\alpha \in [0, 1]$ . (Ver Figura 2).

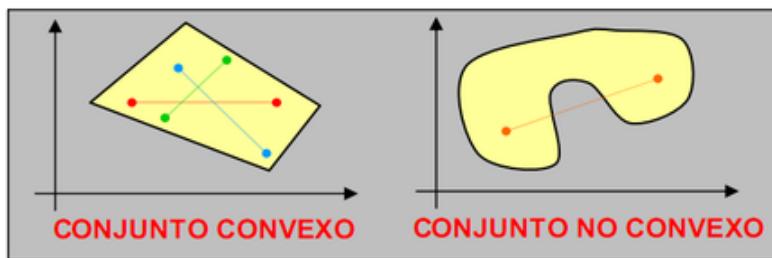


Figura 2: Conjuntos convexo y no convexo.

Los siguientes dos lemas son muy fáciles de demostrar y sus demostraciones se dejan como ejercicio.

**Lema 1.** la intersección finita de conjuntos convexos es un conjunto convexo.

**Lema 2.** todo semiespacio cerrado es un convexo.

**Definición 3.** La intersección de una cantidad finita de semiespacios cerrados de  $\mathbb{R}^n$  se denomina **región poliedral cerrada** de  $\mathbb{R}^n$ .

**Observación 1:** por los dos lemas anteriores, toda región poliedral cerrada es un conjunto convexo.

**Observación 2:** el conjunto de restricciones de un problema de programación lineal dado por  $\Omega = \{x \in \mathbb{R}^n | Ax = b, Rx \leq s\}$  es una región poliedral cerrada. Usualmente, el conjunto factible de un problema de PL suele llamarse **politopo**, siempre que sea no vacío. Además, si esta región es acotada suele llamarse simplemente **poliedro**. Que sea acotada, significa que si  $(x_1, \dots, x_n) \in \Omega$  entonces existen constantes positivas  $k_i$  tales  $|x_i| \leq k_i$  para  $i = 1, \dots, n$ .

Para caracterizar las regiones poliedrales como  $\Omega$ , estamos interesados en los **vértices**, los cuales se determinan por la intersección de las ecuaciones que definen los semiespacios de  $\Omega$ .

En el Ejemplo 1, cuya región factible  $\Omega$  está dada por las restricciones

$$\begin{aligned} 3x + 2y &\geq 3 & (r1) \\ x + 3y &\geq 1.5 & (r2) \\ 8x + 2y &\geq 4 & (r3) \\ x &\geq 0 & (r4) \\ y &\geq 0 & (r5) \end{aligned}$$

y está graficada en la Figura 1, los vértices son:

$$P_1 = (0, 2), \quad P_2 = \left(\frac{1}{5}, \frac{6}{5}\right), \quad P_3 = \left(\frac{6}{7}, \frac{3}{14}\right), \quad P_4 = \left(\frac{3}{2}, 0\right).$$

Por ejemplo, el punto  $P_2 = \left(\frac{1}{5}, \frac{6}{5}\right)$  se obtiene de resolver el siguiente sistema lineal:

$$\begin{cases} 3x + 2y = 3 \\ 8x + 2y = 4 \end{cases}$$

que corresponden a los hiperplanos (rectas) que definen las restricciones (r1) y (r3).

**Observación:** el punto  $(0, \frac{3}{2})$  es solución de

$$\begin{cases} 3x + 2y = 3 \\ x = 0 \end{cases}$$

correspondiente a las restricciones (r1) y (r4) pero **no es un vértice de la región factible  $\Omega$** .

**Definición 4.** Sea  $\Omega$  una región poliedral cerrada de  $\mathbb{R}^n$  determinada por un sistema de  $m$  inecuaciones lineales. Se llaman **vértices** de  $\Omega$  a los puntos pertenecientes a  $\Omega$  que satisfacen uno de los posibles sistemas de  $n$  ecuaciones lineales independientes (obtenido de las  $m$  inecuaciones).

**Ejercicio:** determinar todos los vértices de la región poliedral cerrada  $\Omega$  en  $\mathbb{R}^3$  definida por

$$\begin{cases} x + y + z \leq 3 \\ y - z \leq 2 \\ x - 2y \leq 1 \\ x \geq 0 \end{cases}$$

Geométricamente, los vértices de una región poliedral cerrada  $\Omega$  de  $\mathbb{R}^n$  son los puntos extremos, esto es, puntos de  $\Omega$  que no están contenidos en el interior de algún segmento contenido en la región. (Formalizaremos la definición de punto extremo en la próxima clase).

## Clase 20 - Programación lineal (2)

### El problema

Recordemos la formulación del problema de programación lineal

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeto a} & Cx = d \\ & Rx \geq s \\ & x \geq 0 \end{array}$$

o en su forma estándar

$$\begin{array}{ll} \text{Minimizar} & c^T x \\ \text{sujeto a} & Ax = b \\ & x \geq 0 \end{array}$$

con  $b \geq 0$ .

En la clase anterior dijimos que la región factible es una región poliedral cerrada y que los vértices de esa región serán importantes al momento de buscar las soluciones del problema de PL.

### Método gráfico para el caso bidimensional

Cuando el problema tiene 2 variables la forma más simple de buscar la solución de un problema de PL es el **método gráfico**. Para fijar ideas, consideraremos el siguiente ejemplo en dimensión 2:

$$\begin{array}{ll} \text{Minimizar} & z = -x_1 - 2x_2 \\ \text{sujeto a} & -2x_1 + x_2 \leq 2 \\ & -x_1 + x_2 \leq 3 \\ & x_1 \leq 3 \\ & x_1, x_2 \geq 0. \end{array}$$

Se puede pensar en la función objetivo como  $z = f(x_1, x_2) = c^T(x_1, x_2)$ , con  $c = (-1, -2)$ . La región factible de este problema puede verse en la Figura 1.

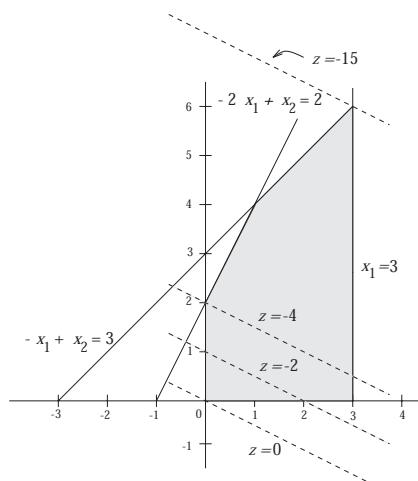
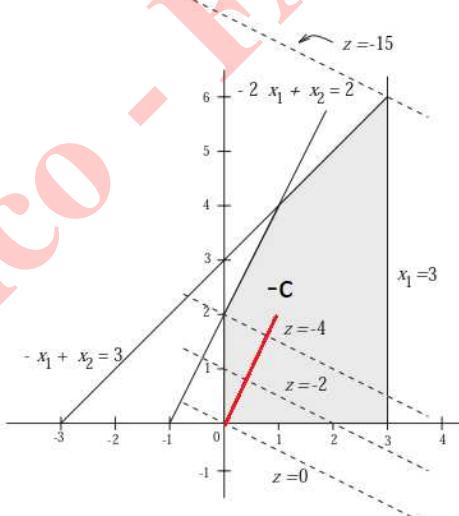


Figura 1: Solución gráfica de un problema lineal.

Notar que la figura incluye algunas líneas punteadas correspondientes a las curvas de nivel para diferentes valores de la función objetivo. Por ejemplo, el conjunto de nivel  $\{(x_1, x_2) | z = -x_1 - 2x_2 = -2\}$  es la recta  $z = -2$  que pasa por los puntos  $(2, 0)$  y  $(0, 1)$ . El conjunto de nivel  $\{(x_1, x_2) | z = -x_1 - 2x_2 = 0\}$  es la recta paralela  $z = 0$  que pasa por el origen. Recordemos que el objetivo del problema es minimizar  $z$ . Como se ve en la Figura 1, el valor de  $z$  decrece a medida que se avanza hacia la derecha, sin embargo no puede decrecer indefinidamente porque la región factible es acotada. Al continuar trazando rectas, asociadas a conjuntos de nivel, hacia la derecha llegará un momento donde esa recta intersecará por última vez al conjunto factible. En este problema el mínimo ocurre cuando  $z = -15$  en el punto  $(3, 6)$ , que corresponde a un vértice del conjunto factible. No es coincidencia que sea un vértice, como veremos más adelante en algunos resultados.

Por otro lado se sabe, por un resultado de Análisis de varias variables, que la dirección del gradiente de una función es la dirección de máximo crecimiento y, análogamente, la dirección de menos gradiente es la dirección de máximo descenso. Esto será muy útil cuando el objetivo sea maximizar o minimizar una función lineal. Ahora bien, como la función es lineal, el gradiente no es más que el vector de costos  $c$ . En este caso,  $c = (-1, -2)$  y por lo tanto, si queremos minimizar la función  $z = -x_1 - 2x_2$ , basta con trazar rectas paralelas en la dirección de  $-c$ , es decir,  $(1, 2)$  hasta intersecar por última vez a la región factible. Ver Figura 2:



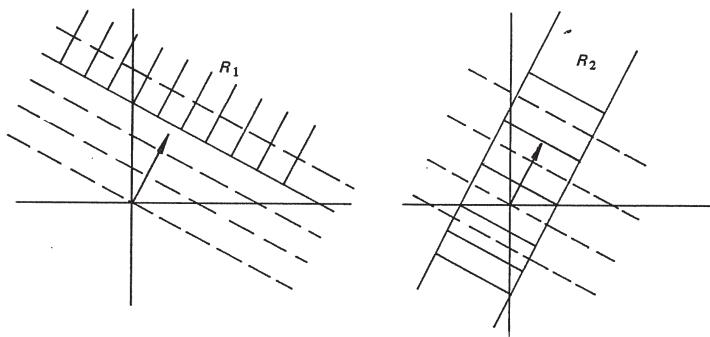
**Figura 2:** Solución gráfica de un problema lineal y vector  $c$ .

En resumen, el método gráfico para problemas de PL consiste en trazar rectas perpendiculares al vector gradiente de la función objetivo, es decir al vector de costos  $c$ , y trasladarse en una dirección u otra dependiendo si se desea minimizar o maximizar.

### Tipos de solución

Es fácil imaginar que pueden surgir diferentes tipos de problemas y soluciones para el caso de un problema de PL en dos dimensiones, debido a la geometría del conjunto factible, a la dirección del vector de costos y si se está minimizando o maximizando. En todas las figuras siguientes consideraremos  $f(x_1, x_2) = x_1 + 2x_2$ .

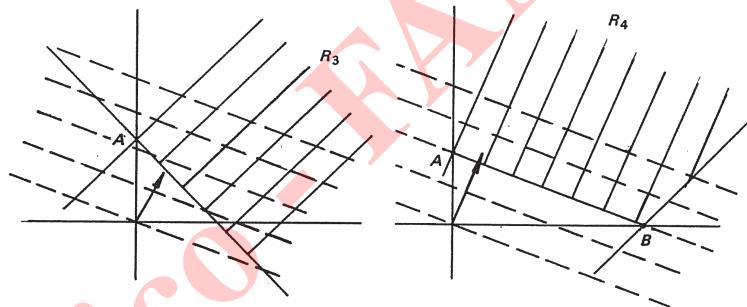
- Regiones no acotadas sin vértices:



**Figura 3:** Regiones no acotadas sin vértices.

En el caso de  $R_1$  se alcanza el valor mínimo en toda la recta (frontera de  $R_1$ ) y no hay valor máximo. En el caso de  $R_2$  no hay mínimo ni máximo. Ver Figura 3.

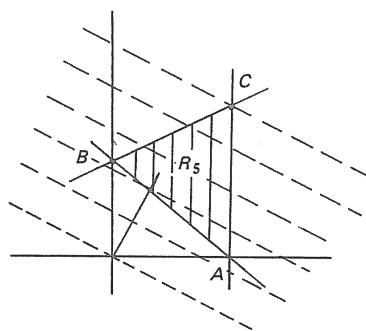
- Regiones no acotadas con vértices:



**Figura 4:** Regiones no acotadas con vértices.

En el caso de  $R_3$  no hay máximo ni mínimo. En el caso de  $R_4$  se alcanza el mínimo en los vértices  $A$  y  $B$ , por lo tanto en todo el segmento que une a estos puntos, y no hay máximo. Notar que en el caso de la región  $R_4$  hay infinitas soluciones al problema de minimización. Ver Figura 4.

- Región acotada (con al menos tres vértices):



**Figura 5:** Región acotada (con al menos tres vértices).

Asume el mínimo en el punto  $A$  y el máximo en el punto  $C$  de la región  $R_5$ . Ver Figura 5.

- Casos degenerados:

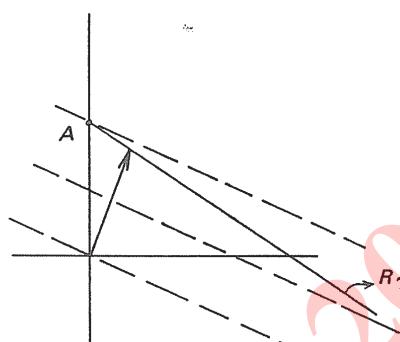
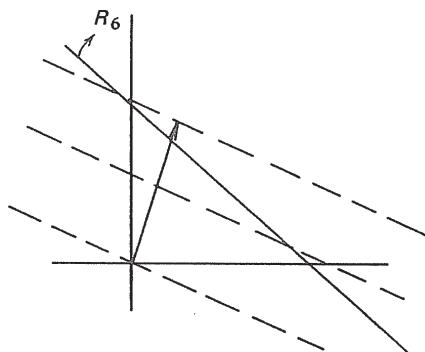


Figura 6: Casos degenerados

En el caso de la región R6 no hay mínimo ni máximo. En cambio, la región R7 tiene un máximo en el punto A y no hay mínimo. Ver Figura 6.

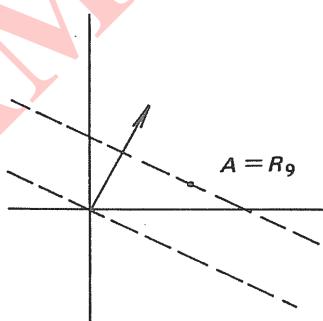
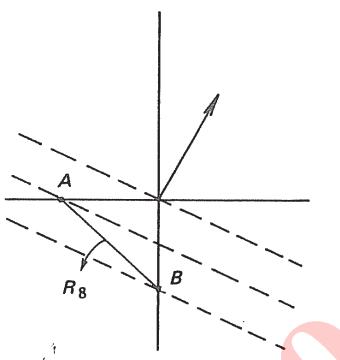


Figura 7: Casos degenerados

La región R8 tiene un mínimo en el punto B y un máximo en el punto A. La región R9 tiene un mínimo y un máximo en el punto A. Ver Figura 7.

## El caso n-dimensional

El método gráfico presentado en la sección anterior es útil sólo en casos bidimensionales y en algunos casos tridimensionales. Para dimensiones mayores que 2 el método gráfico se torna imprácticable y se requiere un método eficiente para resolver tales problemas. Para esto vamos a introducir el Método Simplex, el cual fue propuesto en 1947 por George Dantzig, para resolver problemas de PL que modelizaban situaciones de planificación económica y militar (planning). Este método consiste en un algoritmo eficiente de búsqueda, que comenzando en algún vértice de la región factible avanza hacia otro vértice vecino hasta encontrar la solución óptima de una manera inteligente y eficiente. La Programación Lineal se había desarrollado muy poco hasta 1947 debido a la dificultad computacional de resolver problemas lineales por el hecho de tener una gran número de combinaciones posibles de restricciones a ser consideradas. Con la aparición de este método, estos cálculos fueron optimizados y desde entonces el método Simplex es uno de los algoritmos más estudiados y utilizados a nivel mundial. Aunque se han propuesto otros métodos más recientes, el método Simplex sigue siendo muy competitivo y eficiente y se puede aplicar en una gran cantidad de problemas. En la actualidad, matemáticos aplicados, programadores y especialistas en computación continúan investigando en mejores implementaciones y variantes del método.

## Clase 21 - Programación lineal (3)

### Fundamentos matemáticos del método Simplex

Por razones de tiempo veremos algunos fundamentos matemáticos del método Simplex y posteriormente presentaremos la versión tabla del método, sin entrar en casos particulares ni variantes especiales que lo hacen más robusto. Comenzaremos dando algunas definiciones.

Consideremos el sistema lineal de ecuaciones de la forma

$$\left\{ \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m \end{array} \right. \quad (1)$$

Matricialmente se puede expresar como

$$Ax = b,$$

con  $A \in \mathbb{R}^{m \times n}$ ,  $x \in \mathbb{R}^n$  y  $b \in \mathbb{R}^m$ . Supongamos que  $n > m$  y que la matriz  $A$  tiene rango  $m$ .

**Definición 1. Solución básica:** dado un conjunto de  $m$  ecuaciones lineales en  $n$  incógnitas como en (1), se llamará **base** a cualquier submatriz  $B$ ,  $m \times m$  de  $A$ , no singular, formada por  $m$  columnas independientes de  $A$ . Es posible reordenar columnas de  $A$  para definir  $B$ . Como  $\text{rango}(A) = m$ , es claro que existe tal matriz  $B$ . Sea  $N$  la submatriz  $m \times (n-m)$  de  $A$ , formada por las restantes columnas. Así podemos escribir  $A$  y  $x$  particionados:

$$A = [B|N] \quad y \quad x = \begin{bmatrix} x_B \\ x_N \end{bmatrix} \quad y \text{ por lo tanto, } Bx_B + Nx_N = b.$$

Luego  $x_B = B^{-1}b - B^{-1}Nx_N$ .

Si  $x = (x_B, x_N) = (B^{-1}b - B^{-1}Nx_N, x_N)$  es tal que  $x_N = 0$ , es decir cada una de las  $(n-m)$  componentes son iguales a 0, entonces  $x_B = B^{-1}b$  y así  $x = (x_B, x_N) = (x_B, 0)$  es llamada **solución básica** de (1), con respecto a la base  $B$ . Las componentes de  $x_B$  son llamadas **variables básicas** de  $x$  y las componentes de  $x_N$  son llamadas **variables no-básicas** de  $x$ .

**Definición 2. Solución básica degenerada:** si una o más variables básicas en una solución básica toma el valor 0, se dice que esta solución es una **solución básica degenerada**. Desde el punto de vista geométrico, esto ocurre cuando se tiene un vértice determinado por la intersección de más de dos rectas o hiperplanos.

Ahora consideremos el siguiente sistema

$$\left\{ \begin{array}{l} Ax = b \\ x \geq 0 \end{array} \right. \quad (2)$$

con  $A \in \mathbb{R}^{m \times n}$ , de rango  $m$  y  $b \in \mathbb{R}^m$ , el cual representa el conjunto factible de restricciones de un problema de PL en su forma estándar.

**Definición 3.** Una solución  $x$  que satisface (2) se dice que es **factible** para este sistema. Si además esa solución  $x$  es básica se llama **solución básica factible**. Si esta solución fuera básica degenerada, se la llama **solución básica factible degenerada**.

Finalmente, consideremos el problema de PL en la forma estándar:

$$\begin{cases} \text{Minimizar} & c^T x \\ \text{sujeto a} & Ax = b \\ & x \geq 0 \end{cases} \quad (3)$$

donde  $A \in \mathbb{R}^{m \times n}$  es una matriz de rango  $m$ .

**Definición 4.** Se llama **solución básica factible óptima** a la solución básica factible que da el valor óptimo (en este caso el valor mínimo) para la función objetivo del problema (3).

El siguiente teorema, que sólo enunciaremos, establece la importancia fundamental de las soluciones básicas factibles para resolver un problema de PL. Quien esté interesado en ver la demostración puede consultarla en el libro “Linear and Nonlinear Programming” de Luenberger, Ye, 4th edition, Springer, 2015.

**Teorema 1. Teorema Fundamental de Programación Lineal.** Dado un problema de PL en la forma estándar (3), donde  $A \in \mathbb{R}^{m \times n}$ , tiene rango  $m$ . Luego,

- i) si existe una solución factible, entonces existe una solución básica factible.
- ii) si existe una solución factible óptima, entonces existe una solución básica factible óptima.

La importancia de este teorema radica en que reduce el problema de resolver un problema de PL a la búsqueda en el conjunto de soluciones básicas factibles. Como para un problema de PL de  $n$  variables y  $m$  restricciones, con  $n > m$ , se tiene (a lo sumo)

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

soluciones básicas (correspondientes al número de posibilidades de elegir  $m$  columnas de las  $n$  posibles para construir la matriz base  $B$ ). De esta manera, este teorema da una natural, pero muy ineficiente, técnica de búsqueda finita, la cual suele llamarse **método enumerativo**. Este método consiste en determinar todas las soluciones básicas factibles resolviendo todos los respectivos sistemas lineales y elegir el que produzca el menor valor objetivo, en el caso de un problema de minimización. Por otro lado, el método simplex, que veremos próximamente realiza una búsqueda inteligente sin necesidad de revisar todas las soluciones básicas factibles.

Por último veremos la conexión de las soluciones básicas factibles con los puntos extremos. Esta conexión establece la relación entre resultados algebraicos con resultados geométricos y de convexidad.

**Definición 5.** Un punto  $x$  en la región factible  $\Omega = \{x | Ax = b, x \geq 0\}$ , con  $A \in \mathbb{R}^{m \times n}$  de rango  $m$  y  $b \in \mathbb{R}^m$ , se dice que es un **punto extremo** de  $\Omega$  si y sólo si  $x$  no puede ser expresado como:

$$x = \alpha y + (1 - \alpha)z, \quad y, z \in \Omega, \quad 0 < \alpha < 1, \quad x \neq y \neq z$$

Esta definición dice básicamente que un punto extremo no puede estar en un segmento que conecte otros dos puntos distintos del mismo conjunto. Por ejemplo, cada vértice de un triángulo es un punto extremo del triángulo. Por lo tanto, todos los vértices de una región poliedral cerrada como  $\Omega$  son puntos extremos.

Ahora veremos algunos resultados que relacionan los vértices de la región factible con la solución óptima de un problema de PL.

**Teorema 2.** *Sea  $\Omega$  la región factible dada en la definición anterior. Un vector  $x$  es un punto extremo de  $\Omega$  si y sólo es una solución básica factible de  $\Omega$ .*

*Demostración.*

( $\Leftarrow$ ) Supongamos que  $x$  es una solución básica factible de  $\Omega$ , entonces  $x$  puede particionarse como  $x = (x_B, x_N)$ , con  $Bx_B = b$  y  $x_B \geq 0, x_N = 0$ .

Demostraremos que  $x$  debe ser un punto extremo, por contradicción. Si  $x$  no es un punto extremo entonces existen dos puntos distintos  $y, z \in \Omega$  tal que

$$x = \alpha y + (1 - \alpha)z, \quad 0 < \alpha < 1.$$

Como  $\alpha$  y  $(1 - \alpha)$  son positivos y  $x_N = 0$  entonces los vectores  $y, z$  también tienen se partitionan de la misma manera y tienen  $n - m$  componentes iguales a cero:

$$y = (y_B, y_N), \quad z = (z_B, z_N)$$

con  $By_B = b, y_B \geq 0, y_N = 0$ , y  $Bz_B = b, z_B \geq 0, z_N = 0$ .

Como la matriz  $B$  es inversible y  $Bx = By = Bz = b$ , entonces  $x = y = z$  lo cual contradice que  $x$  no es punto extremo. Por lo tanto hemos probado que  $x$  es un punto extremo.

( $\Rightarrow$ ) Supongamos que  $x$  es un punto extremo entonces veamos que  $x$  es una solución básica factible. También se probará por contradicción.

Como  $x$  es un punto extremo de  $\Omega$  también debe ser un punto factible, es decir,  $Ax = b$  y  $x \geq 0$ . Reordenando las variables del vector  $x$ , si fuera necesario, es posible escribir  $x$  como

$$x = \begin{bmatrix} x_B \\ x_N \end{bmatrix},$$

donde  $x_N = 0$  y  $x_B > 0$ , es decir  $x_N$  contiene las componentes de  $x$  que son iguales a cero. Entonces particionamos la matriz  $A$  de igual forma como

$$A = [B|N]$$

donde  $B$  y  $N$  contienen las columnas correspondientes a  $x_B$  y  $x_N$ , respectivamente. Notar que  $B$  podría no ser una matriz cuadrada. Si las columnas de  $B$  son linealmente independientes, entonces  $x$  es una solución básica factible y termina aquí la demostración. Ahora supongamos que las columnas de  $B$  son linealmente dependientes y construiremos dos puntos factibles distintos  $y, z$  tal que  $x = \frac{1}{2}y + \frac{1}{2}z$ , y por lo tanto,  $x$  no puede ser un punto extremo.

Sea  $B_i$  la  $i$ -ésima columna de  $B$ . Si las columnas de  $B$  son linealmente dependientes, entonces existen números reales  $r_1, \dots, r_k$ , no todos nulos, tales que

$$r_1 B_1 + r_2 B_2 + \cdots + r_k B_k = 0, \quad \text{o equivalentemente} \quad B_1 r_1 + B_2 r_2 + \cdots + B_k r_k = 0,$$

esto dice que si  $r = (r_1, r_2, \dots, r_k)$ , entonces  $Br = 0$ . Notar que

$$B(x_B \pm \alpha r) = Bx_B \pm \alpha Br = Bx_B \pm 0 = Bx_B = b,$$

para todo  $\alpha \in \mathbb{R}$ . Como  $x_B > 0$ , para valores suficientemente pequeños de  $\varepsilon > 0$  se tiene que  $x_B + \varepsilon r > 0$  y  $x_B - \varepsilon r > 0$ .

Sean

$$y = \begin{bmatrix} x_B + \varepsilon r \\ x_N \end{bmatrix} \quad y z = \begin{bmatrix} x_B - \varepsilon r \\ x_N \end{bmatrix}.$$

Luego, los vectores  $y, z$  son factibles, distintos entre sí y distintos de  $x$ . Como  $x = \frac{1}{2}y + \frac{1}{2}z$ , lo cual contradice que  $x$  es un punto extremo.  $\square$

Los corolarios que se enuncian a continuación son consecuencia de este teorema y del Teorema Fundamental de PL.

**Corolario 1.** Si el conjunto  $\Omega$  es no vacío, entonces tiene al menos un punto extremo.

**Corolario 2.** Si existe una solución óptima de un problema de programación lineal, entonces existe una solución óptima finita que es un punto extremo del conjunto de restricciones.

**Corolario 3.** El conjunto de restricciones  $\Omega$  tiene a lo sumo un número finito de puntos extremos.

**Ejemplo:** para fijar ideas vamos a determinar las soluciones básicas factibles del siguiente problema de maximización.

Maximizar	$z = 4500x_1 + 8000x_2$
sujeto a	$5x_1 + 20x_2 \leq 400 \quad (r1)$
	$10x_1 + 15x_2 \leq 450 \quad (r2)$
	$x_1 \geq 0 \quad (r3)$
	$x_2 \geq 0 \quad (r4)$

Agregando variables de holgura, las restricciones de este problema en formato estándar resultan:

$$\begin{cases} 5x_1 + 20x_2 + x_3 &= 400 \\ 10x_1 + 15x_2 + x_4 &= 450 \\ x_1, x_2, x_3, x_4 &\geq 0 \end{cases}$$

Así la matriz  $A$  y el vector  $b$  del conjunto de restricciones en el formato estándar están dados por

$$A = \begin{bmatrix} 5 & 20 & 1 & 0 \\ 10 & 15 & 0 & 1 \end{bmatrix} \quad b = \begin{bmatrix} 400 \\ 450 \end{bmatrix}.$$

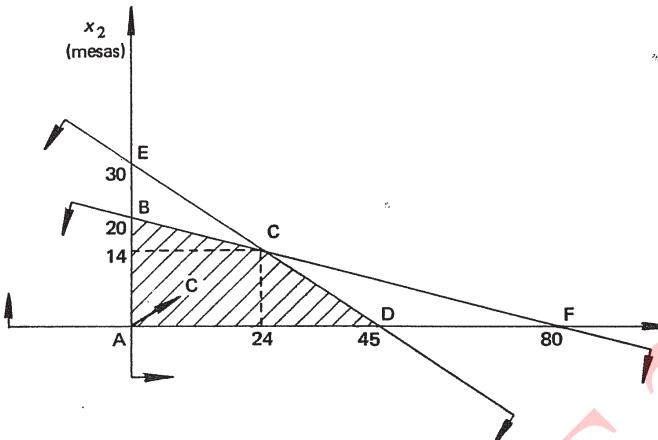
Como  $A$  tiene 2 filas y 4 columnas, se tienen a lo sumo  $\binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$  casos a considerar, como posibles vértices del conjunto factible. Ver Figura 1.

**Caso 1:** La matriz base  $B$  formada por las columnas 3 y 4 de la matriz  $A$ . Entonces

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad x_B = \begin{bmatrix} x_3 \\ x_4 \end{bmatrix}.$$

Luego

$$Bx_B = b \Rightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 400 \\ 450 \end{bmatrix}$$



**Figura 1:** Región factible del ejemplo.

Por lo tanto  $x_B = (x_3, x_4) = (400, 450)$ ,  $x_N = (x_1, x_2) = (0, 0)$ , y la solución básica asociada a esta base es  $(x_1, x_2, x_3, x_4) = (0, 0, 400, 450)$ , que corresponde al vértice A.

**Caso 2:** La matriz base  $B$  formada por las columnas 2 y 4 de la matriz  $A$ . Entonces

$$B = \begin{bmatrix} 20 & 0 \\ 15 & 1 \end{bmatrix} \quad x_B = \begin{bmatrix} x_2 \\ x_4 \end{bmatrix}.$$

Luego

$$Bx_B = b \Rightarrow \begin{bmatrix} 20 & 0 \\ 15 & 1 \end{bmatrix} \begin{bmatrix} x_2 \\ x_4 \end{bmatrix} = \begin{bmatrix} 400 \\ 450 \end{bmatrix}$$

Por lo tanto  $x_B = (x_2, x_4) = (20, 150)$ ,  $x_N = (x_1, x_3) = (0, 0)$ , y la solución básica asociada a esta base es  $(x_1, x_2, x_3, x_4) = (0, 20, 0, 150)$ , que corresponde al vértice B.

**Caso 3:** La matriz base  $B$  formada por las columnas 1 y 2 de la matriz  $A$ . Entonces

$$B = \begin{bmatrix} 5 & 20 \\ 10 & 15 \end{bmatrix} \quad x_B = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Luego

$$Bx_B = b \Rightarrow \begin{bmatrix} 5 & 20 \\ 10 & 15 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 400 \\ 450 \end{bmatrix}$$

Por lo tanto  $x_B = (x_1, x_2) = (24, 14)$ ,  $x_N = (x_3, x_4) = (0, 0)$ , y la solución básica asociada a esta base es  $(x_1, x_2, x_3, x_4) = (24, 14, 0, 0)$ , que corresponde al vértice C.

**Caso 4:** La matriz base  $B$  formada por las columnas 1 y 3 de la matriz  $A$ . Entonces

$$B = \begin{bmatrix} 5 & 1 \\ 10 & 0 \end{bmatrix} \quad x_B = \begin{bmatrix} x_1 \\ x_3 \end{bmatrix}.$$

Luego

$$Bx_B = b \Rightarrow \begin{bmatrix} 5 & 1 \\ 10 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} 400 \\ 450 \end{bmatrix}$$

Por lo tanto  $x_B = (x_1, x_3) = (45, 175)$ ,  $x_N = (x_2, x_4) = (0, 0)$ , y la solución básica asociada a esta base es  $(x_1, x_2, x_3, x_4) = (45, 0, 175, 0)$ , que corresponde al vértice D.

---

**Caso 5:** La matriz base  $B$  formada por las columnas 2 y 3 de la matriz  $A$ . Entonces

$$B = \begin{bmatrix} 20 & 1 \\ 15 & 0 \end{bmatrix} \quad x_B = \begin{bmatrix} x_2 \\ x_3 \end{bmatrix}.$$

Luego

$$Bx_B = b \Rightarrow \begin{bmatrix} 20 & 1 \\ 15 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 400 \\ 450 \end{bmatrix}$$

Por lo tanto  $x_B = (x_2, x_3) = (30, -200)$ ,  $x_N = (x_1, x_4) = (0, 0)$ , y la solución básica asociada a esta base es  $(x_1, x_2, x_3, x_4) = (0, 30, -200, 0)$ , que corresponde al punto E, que no es un vértice.

**Caso 6:** La matriz base  $B$  formada por las columnas 1 y 4 de la matriz  $A$ . Entonces

$$B = \begin{bmatrix} 5 & 0 \\ 10 & 1 \end{bmatrix} \quad x_B = \begin{bmatrix} x_1 \\ x_4 \end{bmatrix}.$$

Luego

$$Bx_B = b \Rightarrow \begin{bmatrix} 5 & 0 \\ 10 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_4 \end{bmatrix} = \begin{bmatrix} 400 \\ 450 \end{bmatrix}$$

Por lo tanto  $x_B = (x_1, x_4) = (80, -350)$ ,  $x_N = (x_2, x_3) = (0, 0)$ , y la solución básica asociada a esta base es  $(x_1, x_2, x_3, x_4) = (80, 0, 0, -350)$ , que corresponde al punto F, que no es un vértice.

## Clase 22 - Programación lineal (4)

### El método Simplex

El método Simplex es un método iterativo para resolver un problema de PL en la forma estándar. Si el problema es no degenerado, el método va recorriendo la región factible, de una manera eficiente, de una solución básica factible (vértice) a otra hasta llegar a la solución básica factible óptima. En cada iteración, se verifica si la solución actual es óptima. Si no lo es, el método elige una dirección donde se mejore el valor óptimo y se avanza en esa dirección hasta encontrar otra solución factible óptima. Luego se repite este procedimiento.

Para entender mejor el método consideraremos un ejemplo:

$$\begin{aligned} \text{minimizar } & z = -x_1 - 2x_2 \\ \text{sujeto a } & -2x_1 + x_2 \leq 2 \\ & -x_1 + 2x_2 \leq 7 \\ & x_1 \leq 3 \\ & x_1, x_2 \geq 0 \end{aligned}$$

El formato estándar de este problema está dado por

$$\begin{aligned} \text{minimizar } & z = -x_1 - 2x_2 \\ \text{sujeto a } & -2x_1 + x_2 + x_3 = 2 \\ & -x_1 + 2x_2 + x_4 = 7 \\ & x_1 + x_5 = 3 \\ & x_1, x_2, x_3, x_4, x_5 \geq 0. \end{aligned}$$

Para este problema se tienen:

$$A = \begin{bmatrix} -2 & 1 & 1 & 0 & 0 \\ -1 & 2 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad y \quad b = \begin{bmatrix} 2 \\ 7 \\ 3 \end{bmatrix}.$$

Cuando se tienen restricciones del tipo  $\leq$  y se agregan variables de holgura, se puede obtener fácilmente una solución básica factible tomando  $x_B = (x_3, x_4, x_5) = (2, 7, 3)$  y  $x_N = (x_1, x_2) = (0, 0)$ , que corresponde al origen de coordenadas, es decir, el punto  $x_a = (x_1, x_2, x_3, x_4, x_5) = (0, 0, 2, 7, 3)$  en la Figura 1.

Para pasar a otra solución básica factible, alguna de las variables no básicas pasará a ser básica, y por lo tanto alguna de las variables básicas dejará de serlo y se convertirá en no básica. Para eso notemos que, a partir de la forma estándar, se obtiene que

$$\begin{aligned} x_3 &= 2 + 2x_1 - x_2 \\ x_4 &= 7 + x_1 - 2x_2 \\ x_5 &= 3 - x_1 \end{aligned} \tag{1}$$

y

$$z = -x_1 - 2x_2,$$

cuyo valor en  $x_a$  es  $z = 0$ , pues  $x_1 = x_2 = 0$ . Como el objetivo es minimizar  $z$ , si cualquiera de estas dos variables toma valores positivos  $z$  decrecerá. Es claro que conviene elegir la variable  $x_2$  porque el coeficiente de  $x_2$  hará que  $z$  decrezca más que si se eligiera  $x_1$ . Por

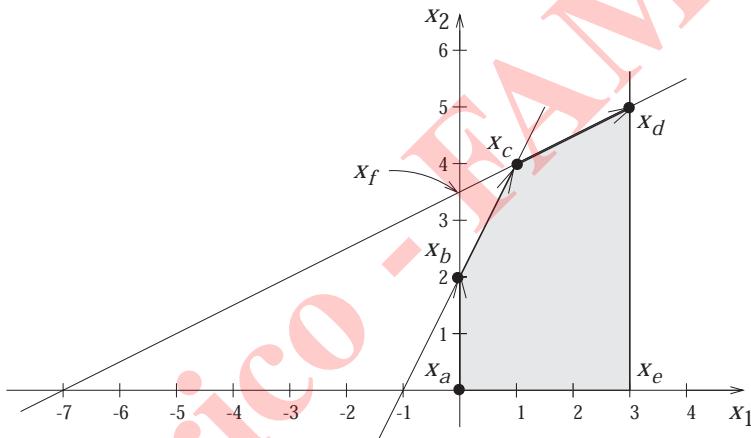
lo tanto, la variable  $x_2$  que era no básica ( $x_2 = 0$ ) pasará a ser básica, es decir tomará un valor positivo. Debido a las restricciones (1), esta variable no puede crecer indefinidamente. Manteniendo la variable no básica  $x_1$ , es decir  $x_1 = 0$ , y teniendo en cuenta (1), obtenemos

$$x_3 = 2 - x_2 \quad (2)$$

$$x_4 = 7 - 2x_2 \quad (3)$$

$$x_5 = 3. \quad (4)$$

Estas 3 ecuaciones permitirán determinar cual es la variable básica que pasará de ser básica a no básica: la variable  $x_2$  puede aumentar hasta que  $x_3$  o  $x_4$  se hagan cero sin tomar valores negativos. De (2), cuando  $x_3 = 0$ , entonces  $x_2 = 2$  y además  $x_4 = 3 > 0$ . De (3), cuando  $x_4 = 0$ , entonces  $x_2 = 7/2$  y además  $x_3 = -3/2 < 0$ . Por lo tanto,  $x_2$  puede crecer hasta el valor  $x_2 = 2$ , y  $x_3$  dejará de ser variable básica para ser no básica. Y así, la nueva solución básica factible será el punto  $x_b = (x_1, x_2, x_3, x_4, x_5) = (0, 2, 0, 3, 3)$  en la Figura 1.



**Figura 1:** El método Simplex.

Para terminar la iteración, resta actualizar el problema en términos de la nueva solución básica factible, donde las variables básicas están en  $x_B = (x_2, x_4, x_5)$  y las no básicas  $x_N = (x_1, x_3)$ . Para esto, se deben expresar las nuevas variables básicas en términos de las nuevas variables no básicas. Las variables no básicas deben aparecer en la función objetivo y en la matriz de restricciones se debe obtener una matriz identidad en las variables básicas.

- De la primera ecuación de (1), tenemos que  $x_2 = 2 + 2x_1 - x_3$ , y reemplazando en la función objetivo y en las restricciones, el problema de PL resulta

$$\begin{aligned} \text{minimizar } & z = -4 - 5x_1 + 2x_3 \\ \text{sujeto a } & x_2 = 2 + 2x_1 - x_3 \\ & x_4 = 3 - 3x_1 + 2x_3 \\ & x_5 = 3 - x_1 \\ & x_1, x_2, x_3, x_4, x_5 \geq 0. \end{aligned}$$

Como  $x_N = (x_1, x_3) = (0, 0)$ , entonces  $z = -4$  y  $x_B = (x_2, x_4, x_5) = (2, 3, 3)$ . Esto termina la primera iteración del método.

## Fórmulas generales

Consideremos el problema de PL en el formato estándar

$$\begin{aligned} & \text{minimizar} && z = c^T x \\ & \text{sujeto a} && Ax = b \\ & && x \geq 0. \end{aligned}$$

Si  $x = (x_B, x_N)$  es una solución básica factible para alguna submatriz  $B$  de  $A = [B|N]$ , entonces

$$\begin{aligned} z &= c_B^T x_B + c_N^T x_N \\ Bx_B + Nx_N &= b, \quad \Rightarrow x_B = B^{-1}b - B^{-1}Nx_N, \end{aligned}$$

entonces reemplazando en  $z$ , se obtiene  $z = c_B^T B^{-1}b + (c_N^T - c_B^T B^{-1}N)x_N$ .

Ahora, el valor actual de las variables básicas y la función objetivo se obtienen tomando  $x_N = 0$ , y denotamos  $x_B = \hat{b} = B^{-1}b$ , y  $\hat{z} = c_B^T B^{-1}b$ .

Si definimos  $c_B^T B^{-1} = y^T$ , es decir  $y = (c_B^T B^{-1})^T = B^{-T} c_B$ , entonces

$$z = y^T b + (c_N^T - y^T N)x_N.$$

Se llama **costo reducido** de  $x_j$  a  $\hat{c}_j$ , la entrada  $j$ -ésima en el vector  $\hat{c}_N^T \equiv (c_N^T - c_B^T B^{-1}N)$ . Luego,  $z = \hat{z} + \hat{c}_N^T x_N$ .

**Test de optimalidad:** se debe analizar que ocurre con la función objetivo si se aumenta cada variable no básica a partir del valor cero. Si  $\hat{c}_j > 0$  la función objetivo aumentará; si  $\hat{c}_j = 0$  la función objetivo se mantiene constante; si  $\hat{c}_j < 0$  la función objetivo disminuirá si  $x_j$  comienza a crecer a partir de cero. Si la solución no es óptima, entonces se puede elegir una variable no básica  $x_t$  con  $\hat{c}_t < 0$  para **entrar a la base**.

**¿Cuál variable dejará de ser básica?:** para esto recordemos que  $x_B = B^{-1}b - B^{-1}Nx_N$ , y dado que las variables no básicas son iguales a cero, excepto  $x_t$ , entonces

$$x_B = \hat{b} - \hat{A}_t x_t,$$

donde  $\hat{A}_t = B^{-1}A_t$ , y  $A_t$  es la columna  $t$  de  $A$ . Ahora, como hicimos en el ejemplo anterior, miremos la  $i$ -ésima componente de  $x_B$ :

$$(x_B)_i = \hat{b}_i - \hat{a}_{i,t} x_t.$$

Si  $\hat{a}_{i,t} > 0$ , entonces  $(x_B)_i$  decrecerá a medida que  $x_t$  aumente, y  $(x_B)_i$  será igual a cero cuando  $x_t = \hat{b}_i / \hat{a}_{i,t}$ . Si  $\hat{a}_{i,t} < 0$ , entonces  $(x_B)_i$  aumentará y si  $\hat{a}_{i,t} = 0$  entonces  $(x_B)_i$  se mantiene constante. Se aplicará el siguiente **test del cociente mínimo** para decidir cual variable básica pasará a ser no básica:

$$\bar{x}_s = \min_{1 \leq i \leq m} \left\{ \frac{\hat{b}_i}{\hat{a}_{i,t}} \mid \hat{a}_{i,t} > 0 \right\}.$$

Para determinar los nuevos valores de las variables básicas y la función objetivo se debe calcular:

$$x_B \leftarrow x_B - \hat{A}_t \bar{x}_s, \quad y \quad \hat{z} \leftarrow \hat{z} + \hat{c}_t \bar{x}_s.$$

Si  $\hat{a}_{i,t} \leq 0$  para todo  $i$  entonces ninguna de las variables básicas decrecerá a medida que  $x_t$  aumenta, por lo tanto  $x_t$  puede crecer tanto como se quiera. Esto significa, que la función objetivo decrecerá mientras  $x_t \rightarrow \infty$ , o sea que no tendrá un mínimo finito, y por lo tanto el problema es no acotado.

## El Algoritmo Simplex

Se comienza con una matriz base  $B$  correspondiente a una solución básica factible  $x_B = \hat{b} = B^{-1}b \geq 0$ . Los siguientes tres pasos resumen el método Simplex.

### Paso 1: test de optimalidad.

Calcular  $y^T = c_B^T B^{-1}$ ;  $\hat{c}_N^T = c_N^T - y^T N$ .

Si  $\hat{c}_N^T \geq 0$ , entonces la solución es óptima. Sino, elegir una variable  $x_t$  tal que  $\hat{c}_t < 0$ , para entrar a la base (pasar a ser básica).

### Paso 2: test del cociente mínimo.

Calcular  $\hat{A}_t = B^{-1}A_t$ . Determinar un índice  $s$  tal que

$$\frac{\hat{b}_s}{\hat{a}_{s,t}} = \min_{1 \leq i \leq m} \left\{ \frac{\hat{b}_i}{\hat{a}_{i,t}} \mid \hat{a}_{i,t} > 0 \right\}.$$

Este test determina cual es la variable que abandona la base.

Si  $\hat{a}_{i,t} \leq 0$  para todo  $i$ , el problema es no acotado.

### Paso 3: actualización.

Actualizar la matriz base  $B$  y las correspondientes variables básicas  $x_B$ .

Volver al Paso 1.

**Observación 1:** esta es una versión simplificada del método simplex. Existen variantes algorítmicas para evitar algunos problemas que pueden aparecer como ciclos o restricciones redundantes. Además, existen diferentes implementaciones del método simplex, tanto gratuitas como comerciales.

**Observación 2:** a veces no es posible disponer de una solución básica factible inicial fácilmente. Para esto existen algunos métodos que se usan previamente para obtener esta solución inicial (método de las dos fases, método de la M grande).

**Observación 3:** bajo adecuadas hipótesis el algoritmo del método simplex converge a una solución básica factible óptima o determina que el problema es no acotado.

**Observación 4:** cuando las variables involucradas son enteras, el problema es mucho más complejo, tanto matemática como computacionalmente, y esto se estudia en una subárea de Optimización llamada Programación Lineal Entera.

## Formulación tabla (tableau) del método Simplex

Para resolver problemas grandes de PL, las implementaciones computacionales eficientes del método Simplex se basan en el algoritmo descripto arriba. En cambio, para problemas pequeños es conveniente usar la formulación tabla por ser una manera compacta y sistemática de organizar las cuentas del método Simplex.

Para la formulación estándar del ejemplo:

$$\begin{aligned}
 & \text{minimizar} && z = -x_1 - 2x_2 \\
 & \text{sujeto a} && -2x_1 + x_2 + x_3 = 2 \\
 & && -x_1 + 2x_2 + x_4 = 7 \\
 & && x_1 + x_5 = 3 \\
 & && x_1, x_2, x_3, x_4, x_5 \geq 0.
 \end{aligned}$$

La tabla inicial está dada por:

base	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	LD
$-z$	-1	-2	0	0	0	0
$x_3$	-2	1	1	0	0	2
$x_4$	-1	2	0	1	0	7
$x_5$	1	0	0	0	1	3

(5)

LD, significa lado derecho en la primera fila de la tabla. Notar que en la segunda fila escribimos  $-z$ . Esto se debe a que igualamos a cero a la función objetivo:

$$z = -x_1 - 2x_2 \Rightarrow -z - x_1 - 2x_2 + 0x_3 + 0x_4 + 0x_5 = 0.$$

La primera columna indica las variables básicas y en las 3 últimas filas se indican los coeficientes de las restricciones.

La tabla del problema original y en la base actual de cada iteración están dadas por:

base	$x_B$	$x_N$	LD	base	$x_B$	$x_N$	LD
$-z$	$c_B$	$c_N$	0	$-z$	0	$c_N^T - c_B^T B^{-1} N$	$-c_B^T B^{-1} b$
$x_B$	$B$	$N$	$b$	$x_B$	$I$	$B^{-1} N$	$B^{-1} b$

Como en la Tabla 5, los coeficientes de las variables no básicas en la primera fila (costos reducidos) son negativos, significa que esa solución no es óptima. Elegimos  $x_2$  por tener el costo reducido de mayor magnitud (-2). Esa será la variable no básica que entrará a la base. Para decidir cual variable básica dejará de serlo calculamos el mínimo de los siguientes cocientes:

$$\min \left\{ \frac{2}{1}, \frac{7}{2} \right\} = 2,$$

como este valor de mínimo corresponde a la variable básica  $x_3$ , esa variable abandonará la base y será reemplazada por  $x_2$ .

El paso siguiente consiste en transformar la tabla de manera de obtener una matriz identidad en las nuevas variables básicas  $x_2, x_4, x_5$ . Antes de reemplazar  $x_3$  se debe identificar la columna de  $x_2$  en la tabla

base	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	LD
$-z$	-1	-2	0	0	0	0
$\Rightarrow x_3$	-2	1	1	0	0	2
$x_4$	-1	2	0	1	0	7
$x_5$	1	0	0	0	1	3

y transformarla, usando operaciones elementales por filas, en  $\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ . En este caso, se debe sumar 2 veces la fila de  $x_3$  a la fila de  $-z$  y restar 2 veces la fila de  $x_3$  a la fila de  $x_4$ . Así se obtiene la siguiente tabla

base	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	LD
$-z$	-5	0	2	0	0	4
$x_2$	-2	1	1	0	0	2
$x_4$	3	0	-2	1	0	3
$x_5$	1	0	0	0	1	3

Ahora se comienza la segunda iteración. En la primera fila, el costo reducido de la variable  $x_1$  es  $-5 < 0$ , por lo tanto esta solución básica factible no es óptima y  $x_1$  puede pasar a ser básica (entrar a la base). Aplicando el test del cociente mínimo, se deduce que  $x_4$  es la variable que abandonará la base:

base	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	LD
$-z$	-5	0	2	0	0	4
$x_2$	-2	1	1	0	0	2
$\Rightarrow x_4$	3	0	-2	1	0	3
$x_5$	1	0	0	0	1	3

Luego, repitiendo este procedimiento, se obtienen las siguientes tablas, de las iteraciones tres y cuatro:

base	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	LD
$-z$	0	0	-4/3	5/3	0	9
$x_2$	0	1	-1/3	2/3	0	4
$x_1$	1	0	-2/3	1/3	0	1
$\Rightarrow x_5$	0	0	2/3	-1/3	1	2

---

Finalmente,

base	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	LD
$-z$	0	0	0	1	2	13
$x_2$	0	1	0	$1/2$	$1/2$	5
$x_1$	1	0	0	0	1	3
$x_3$	0	0	1	$-1/2$	$3/2$	3

Vemos que en esta última tabla, los costos reducidos de las variables no básicas ( $x_4$  y  $x_5$ ) son ambos positivos, y por lo tanto esta solución básica factible es óptima. La solución final está claramente expresada en la última columna:  $z = -13$ ,  $x_2 = 5$ ,  $x_1 = 3$  y  $x_3 = 3$ , y las variables no básicas  $x_4 = x_5 = 0$ , lo que corresponde al punto  $x_d$  de la Figura 1.