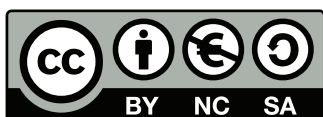


Métodos Numéricos



Este libro se distribuye bajo una licencia CC BY-NC-SA 4.0.

Eres libre de distribuir y adaptar el material siempre que reconozcas a los autores originales del documento, no lo utilices para fines comerciales y lo distribuyas bajo la misma licencia.

creativecommons.org/licenses/by-nc-sa/4.0/

Métodos Numéricos

Daniel Zufri Quesada

Doble grado de ingeniería informática y matemáticas

Universidad de Granada

https://gitlab.com/danizufrique/danielzq_public/tree/master/LP0

Índice

I Tema 1. Introducción a los problemas del análisis Numérico	1
1. Introducción a los métodos numéricos: algoritmo	1
1.1. Espacios normados	1
1.2. Problemas bien planteados. Estabilidad	13
1.3. Algoritmos. Algoritmo PageRank de Google	19
2. Errores de redondeo. Iteradores	21
2.1. Sistema posicional y números máquina	21
2.2. Redondeo en sistemas de punto flotante y su aritmética	24
II Tema 2. Resolución numérica de sistemas de ecuaciones lineales	1
1. Métodos directos: Gauss y versiones, factorización de matrices	1
1.1. Sistemas triangulares	1
1.2. Métodos de Gauss y Gauss-Jordan. Pivotaje	1
1.3. Métodos de factorización	4
2. Métodos iterativos: métodos de Jacobi y Gauss-Seidel	11
2.1. Métodos iterativos: convergencia y consistencia	11
2.2. Generación de métodos iterativos	13
3. Análisis de error	21
III Tema 3. Interpolación	24
1. Interpolación polinómica: Lagrange y Newton. Error de interpolación	24
1.1. Polinomio de interpolación tipo Lagrange	24
1.2. Forma de Newton del polinomio de interpolación	25
1.3. Error de interpolación. Convergencia y estabilidad. Polinomios de Chebyshev	26

1.4. Otros problemas de interpolación: Hermite y caso general	33
2. Interpolación mediante funciones splines	35
2.1. Funciones splines lineales	36
2.2. Funciones splines cúbicas	37
 IV Tema 4. Aproximación	 40
1. Aproximación por mínimos cuadrados discreta y continua	40
1.1. Principio del mínimo	40
1.2. Aproximación por mínimos cuadrados discreta	41
1.3. Aproximación por mínimos cuadrados general: caso continuo	47

I | Tema 1. Introducción a los problemas del análisis Numérico

1 Introducción a los métodos numéricos: algoritmo

Comenzaremos con un ejemplo de recurrencia en el que observaremos que al redondear el primer valor, se acumula el error y los siguientes valores se desbordan.

Sea la recurrencia $x_n := \int_0^1 x^n e^x dx$ con $n \geq 1$. Si resolvemos la integral para

(i) $n = 0$, tenemos que $x_0 = e - 1$.

(ii) $n \in \mathbb{N}$, tenemos que $x_n = e - nx_{n-1}$

Por lo que esta sucesión, $\{x_n\}_{n \geq 1} \subset \mathbb{R}_+$, la cual es monótona y acotada, es decreciente y tiende a 0, es decir, $\lim_{n \rightarrow \infty} x_n = 0$.

Veamos que si redondeamos x_0 y calculamos a partir de él, los siguientes términos de la sucesión, se acumula el error.

Si $n = 12$, tenemos que $x_{12} = 0.1951$. Si ahora redondeamos $x_0 = 1.7183$ e iterando según este valor hasta $n = 12$, obtenemos que $x_{12} = 8704.39$. Luego, este valor es, con diferencia, mayor que el que habíamos calculado sin redondeo y x_n no tiende a 0.

Concluimos que el redondeo, a veces, conlleva errores muy grandes.

1.1 Espacios normados

Si usamos las normas en los problemas numéricos, sabremos si los problemas están bien planteados, los errores cometidos, la convergencia...

Definición 1.1 (Norma). Sea E un espacio vectorial real, diremos que una aplicación $\|\cdot\| : E \rightarrow \mathbb{R}$ es una **norma** en E si verifica las siguientes propiedades:

(i) $\|x\| \geq 0$, con $x \in E$.

Además, $\|x\| = 0 \Leftrightarrow x = 0$.

(ii) $\|x + y\| \leq \|x\| + \|y\|$, con $x, y \in E$ (desigualdad triangular)

(iii) $\|\lambda x\| = |\lambda| \|x\|$, con $x \in E, \lambda \in \mathbb{R}$

Demostración 1.1. $E = \mathbb{R}^N, \|x\|_\infty = \max |x_j|$

(i) $x \in \mathbb{R}^N$:

- $\|x\|_\infty = \max_{j=1,\dots,N} |x_j| \geq 0$
- $\|x\|_\infty \iff 0 \leq \max_{j=1,\dots,N} |x_j| = 0 \iff |x_1| = \dots = |x_N| = 0$
 $\iff x_1 = \dots = x_n = 0 \iff x = 0$

(ii) $x, y \in \mathbb{R}^N \Rightarrow \|x+y\|_\infty \leq \|x\|_\infty + \|y\|_\infty$

$$\|x+y\|_\infty = \max_{j=1,\dots,N} |x_j+y_j| \leq \max_{j=1,\dots,N} (|x_j| + |y_j|) \leq \max_{j=1,\dots,N} |x_j| + \max_{j=1,\dots,N} |y_j| = \|x\|_\infty + \|y\|_\infty$$

(iii) $x \in \mathbb{R}^N, \lambda \in \mathbb{R} \Rightarrow \|\lambda x\|_\infty = |\lambda| \|x\|_\infty$

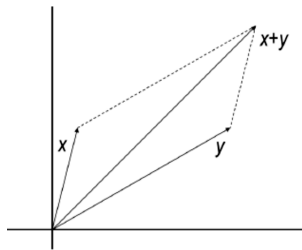
$$\|\lambda x\|_\infty = \max_{j=1,\dots,N} |\lambda x_j| = \max_{j=1,\dots,N} |\lambda| |x_j| = |\lambda| \max_{j=1,\dots,N} |x_j| = |\lambda| \|x\|_\infty$$

□

Definición 1.2 (Espacio normado). Sea E un espacio vectorial real. Si este espacio admite una norma, entonces E se llama **espacio normado**.

Trabajaremos en \mathbb{R} , aunque en \mathbb{C} es lo mismo.

Veamos la interpretación geométrica de la desigualdad triangular, usando la norma euclídea en \mathbb{R}^2 o \mathbb{R}^3 .



Las siguientes normas son las que vamos a utilizar.

Definición 1.3 (Norma p). Sean $E = \mathbb{R}^N, p \in \mathbb{R}_+, p \geq 1$ y $x \in \mathbb{R}^N$, entonces:

$$\|x\|_p := \left(\sum_{j=1}^N |x_j|^p \right)^{1/p}$$

Si $p = 2$, entonces la norma es **euclídea**.

Definición 1.4 (Norma del máximo). Sean $E = \mathbb{R}^N$ y $x \in \mathbb{R}^N$, entonces:

$$\|x\|_{\infty} := \max\{|x_j| : j = 1, \dots, N\}$$

Definición 1.5 (Norma de Frobenius). Sean $E = \mathbb{R}^{M \times N}$ y $A \in \mathbb{R}^{M \times N}$, entonces:

$$\|A\|_F := \sqrt{\sum_{i=1}^M \sum_{j=1}^N a_{ij}^2}$$

Nota. Esta última norma es bastante inútil.

Nota. Si estamos en un espacio vectorial real $C([a, b])$, esto significa que este espacio está compuesto por todas las funciones continuas en el intervalo cerrado $[a, b]$. Si el espacio es $C^k([a, b])$, significa que está formado por las funciones de clase k , es decir, funciones derivables hasta orden k y esas derivadas son continuas.

Definición 1.6 (Norma del máximo). Sean $E = C([a, b])$ y $f \in C([a, b])$, entonces:

$$\|f\|_{\infty} := \max \{ |f(x)| : a \leq x \leq b \}$$

Definición 1.7. Sean $E = C^k([a, b])$, $k \in \mathbb{N}$ y $f \in C^k([a, b])$ entonces:

$$\|f\|_k := \max \{ \|f^{(j)}\| : j = 0, \dots, k \}$$

Nota. No entra en el examen

Nota. Si una función es derivable siempre es continua.

Ahora que ya tenemos definidas las normas, podemos calcular el error cometido al aproximar los vectores.

Definición 1.8 (Error absoluto). Sean E un espacio normado, $x \in E$ y $x^* \in E$ una aproximación de x , entonces la siguiente operación calcula el error absoluto:

$$\|x^* - x\|_\infty$$

Definición 1.9 (Error relativo). Sean E un espacio normado, $x \in E$ y $x^* \in E$ una aproximación de x , entonces la siguiente operación calcula el error relativo:

$$\frac{\|x^* - x\|_\infty}{\|x\|_\infty}$$

Veamos una aplicación de estos errores.

Ejercicio 1.1. Calcula los errores absolutos y relativos de:

(i) $E = \mathbb{R}$, $x = 1/4$, $x^* = 0.23$.

(ii) $E = \mathbb{R}^3$, $x = (1/5, 2, 1)$, $x^* = (0.19, 2.2, 0.9)$.

(iii) $E = C([0, \pi/2])$, $f(t) = \sin(t)$, $f^*(t) = t$

Solución.

(i) *error absoluto:*

$$|x^* - x| = |0.23 - 1/4| = 0.02$$

error relativo:

$$\frac{|x^* - x|}{|x|} = \frac{|0.23 - 1/4|}{|1/4|} = 0.08$$

(ii) *error absoluto:*

$$\|x^* - x\|_\infty = \|(0.19, 2.2, 0.9) - (1/5, 2, 1)\|_\infty = \|(-0.01, 0.2, -0.1)\|_\infty = \max\{0.01, 0.2, 0.1\} = 0.2$$

error relativo:

$$\frac{\|x^* - x\|_\infty}{\|x\|_\infty} = \frac{\|(-0.01, 0.2, -0.1)\|_\infty}{\|(1/5, 2, 1)\|_\infty} = \frac{0.2}{2} = 0.1$$

(iii) *error absoluto*:

$$\|f^* - f\|_\infty = \|t - \sin(t)\|_\infty = \frac{\pi}{2} - 1$$

error relativo:

$$\frac{\|f^* - f\|_\infty}{\|f\|_\infty} = \frac{\pi}{2} - 1$$

Definición 1.10 (Distancia). Se define la **distancia** entre dos vectores $x, y \in E$ como

$$\text{dist}(x, y) := \|x - y\|$$

Definición 1.11 (Converge). Se dice que $\{x_n\}_{n \geq 1}$ en E **converge** a $x_0 \in E$ si

$$\forall \varepsilon > 0 \Rightarrow [\exists n_0 \in \mathbb{N} : n \geq n_0 \Rightarrow \|x_n - x_0\| < \varepsilon]$$

es decir,

$$\lim_{n \rightarrow \infty} x_n = x_0 \Leftrightarrow \lim_{n \rightarrow \infty} \|x_n - x_0\| = 0$$

Definición 1.12 (Continua). Sean X, Y subconjuntos no vacíos de sendos espacios normados y sea $f : X \rightarrow Y$, diremos que f es **continua** en $x_0 \in X$ si

$$\forall \varepsilon > 0 \Rightarrow [\exists \delta > 0 : x \in X \wedge \|x - x_0\| < \delta \Rightarrow \|f(x) - f(x_0)\| < \varepsilon]$$

Proposición 1.1. Sea $x \in \mathbb{R}^N \Rightarrow \|x\|_\infty \leq \|x\|_1 \leq N\|x\|_\infty$

Definición 1.13 (Normas equivalentes). Sean $\|\cdot\|$ y $\|\cdot\|_*$ dos normas, se dice que son **equivalentes** si $\exists c_1, c_2 > 0$ tales que

$$\forall x \in E \Rightarrow c_1\|x\| \leq \|x\|_* \leq c_2\|x\|$$

Proposición 1.2. Sean $\|\cdot\|$ y $\|\cdot\|_*$ dos normas, entonces la convergencia de sucesiones y la continuidad son equivalentes para ambas normas.

Teorema 1.1. Todas las normas en un espacio normado finito dimensional son equivalentes.

Observemos que para calcular el límite de la norma del máximo tenemos que calcular el límite de cada coordenada.

Proposición 1.3. Sea \mathbb{R}^N un espacio normado finito dimensional y consideremos la norma $\|\cdot\|_\infty$ en este espacio, entonces:

$$\lim_{n \rightarrow \infty} x_n = x_0 \Leftrightarrow \lim_{n \geq 1} (x_n)_j = (x_0)_j \quad \forall j \in \{1, \dots, N\}$$

Demostración 1.2.

$$\lim_{n \rightarrow \infty} x_n = x_0 \Leftrightarrow \lim_{n \rightarrow \infty} \|x_n - x_0\|_\infty = 0 \Leftrightarrow \max \{ |(x_n - x_0)_j| : j = 1, \dots, N \} = 0 \Leftrightarrow \lim_{n \geq 1} (x_n)_j = (x_0)_j$$

□

Ejemplo 1.1. Aplicación de la anterior proposición:

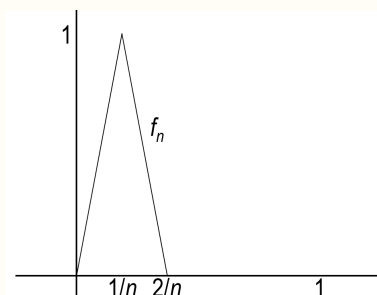
$$\lim_{n \rightarrow \infty} \left(\left(1 + \frac{1}{n} \right)^n, \frac{(-1)^n}{n^2} \right) = (e, 0)$$

La anterior proposición también se puede aplicar para cualquier norma de \mathbb{R}^N y de $\mathbb{R}^{M \times N}$.

Ejercicio 1.2. Comprueba que la norma del máximo en $C([0, 1])$ no es equivalente a la norma $\|\cdot\|_1$ definida para cada $f \in C([0, 1])$ como

$$\|f\|_1 := \int_0^1 |f(x)| dx$$

(Indicación: para cada $n \geq 2$, considera la función f_n cuya gráfica es la poligonal que une los puntos $(0, 0)$, $(1/n, 1)$, $(2/n, 0)$, $(1, 0)$).



Solución. Tenemos que

$$E = C([0, 1]), \quad \|f\|_\infty = \max \{ |f(x)| : 0 \leq x \leq 1 \}, \quad \|f\|_1 := \int_0^1 |f(x)| dx$$

Luego

$$\|f_n\|_\infty = 1 \quad y \quad \|f_n\|_1 = 1/n$$

(Lo cual coincide con el área).

Si fueran equivalentes, entonces

$$\exists \alpha, \beta > 0 \, f \in E \Rightarrow \alpha \|f_n\|_1 \leq \|f_n\|_\infty \leq \beta \|f_n\|_1$$

Lo cual es una contradicción, pues si $n \geq 1 \Rightarrow \|f_n\|_\infty \leq \beta \|f_n\|_1 \Leftrightarrow 1 \leq \frac{\beta}{n} \Leftrightarrow n \leq \beta$ y n no está acotada.

Por lo que no son equivalentes.

Proposición 1.4. Sean $M, N \in \mathbb{N}$ y consideremos sendas normas en \mathbb{R}^N y \mathbb{R}^M , que sin lugar a ambigüedad notaremos indiferentemente como $\|\cdot\|$. Entonces la aplicación que notaremos igualmente como $\|\cdot\|$ define una norma en $\mathbb{R}^{M \times N}$:

$$\|A\| := \sup \left\{ \|Ax\| : x \in \mathbb{R}^N \wedge \|x\| = 1 \right\} \quad \forall A \in \mathbb{R}^{M \times N}$$

Definición 1.14 (Norma Inducida). Se define la norma **inducida** en $\mathbb{R}^{M \times N}$ como:

$$\|A\| := \sup \left\{ \|Ax\| : x \in \mathbb{R}^N \wedge \|x\| = 1 \right\} \quad \forall A \in \mathbb{R}^{M \times N}$$

Proposición 1.5. Con la notación de la proposición anterior, si $A \in \mathbb{R}^{M \times N}$ entonces

$$\|A\| := \sup \left\{ \frac{\|Ax\|}{\|x\|} : x \in \mathbb{R}^N \wedge x \neq 0 \right\}$$

En particular,

$$\|Ax\| \leq \|A\| \|x\|$$

Proposición 1.6. Consideremos la norma $\|\cdot\|_1$ en \mathbb{R}^N y en \mathbb{R}^M , entonces la norma $\|\cdot\|_1$ inducida

en $\mathbb{R}^{M \times N}$ es

$$\|A\|_1 = \max \left\{ \sum_{i=1}^M |a_{ij}| : j = 1, \dots, N \right\} \quad \forall A \in \mathbb{R}^{M \times N}$$

Es decir, es el máximo de las sumas de los valores absolutos de cada **columna**.

Proposición 1.7. Consideremos la norma $\|\cdot\|_\infty$ en \mathbb{R}^N y en \mathbb{R}^M , entonces la norma $\|\cdot\|_\infty$ inducida en $\mathbb{R}^{M \times N}$ es

$$\|A\|_\infty = \max \left\{ \sum_{j=1}^N |a_{ij}| : i = 1, \dots, M \right\} \quad \forall A \in \mathbb{R}^{M \times N}$$

Es decir, es el máximo de las sumas de los valores absolutos de cada **fila**.

Demostración 1.3. Vamos a demostrar que se da la doble desigualdad, luego se dará la igualdad.

Probaremos primero que

$$\|A\|_\infty \geq \max \left\{ \sum_{j=1}^N |a_{ij}| : i = 1, \dots, M \right\}$$

$$\text{Sea } \text{sign}(a) := \begin{cases} -1 & \text{si } a < 0 \\ 1 & \text{si } a \geq 0 \end{cases} \quad \forall a \in \mathbb{R}$$

Tenemos que

$$\left\| [\text{sign}(a_{11}), \dots, \text{sign}(a_{1N})]^T \right\|_\infty = 1 \quad \Rightarrow \quad \|A\|_\infty \geq \left\| A [\text{sign}(a_{11}), \dots, \text{sign}(a_{1N})]^T \right\|_\infty \geq \sum_{j=1}^N |a_{1j}|$$

Hacemos lo mismo con $[\text{sign}(a_{i1}), \dots, \text{sign}(a_{iN})]^T \quad \forall i = 2, \dots, M$ y obtenemos que

$$\|A\|_\infty \geq \max \left\{ \sum_{j=1}^N |a_{ij}| : i = 1, \dots, M \right\}$$

Ahora probaremos que

$$\|A\|_\infty \leq \max \left\{ \sum_{j=1}^N |a_{ij}| : i = 1, \dots, M \right\}$$

Sea $x \in \mathbb{R}^N$ tal que $\|x\|_\infty = 1$, entonces:

$$\begin{aligned} \|Ax\|_\infty &= \left\| \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \vdots \\ a_{M1} & \cdots & a_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \right\|_\infty = \left\| \left[\sum_{j=1}^N a_{1j}x_j, \dots, \sum_{j=1}^N a_{Mj}x_j \right]^T \right\|_\infty = \\ &= \max \left\{ \left| \sum_{j=1}^N a_{ij}x_j \right| : i = 1, \dots, M \right\} \leq \max \left\{ \sum_{j=1}^N |a_{ij}| |x_j| : i = 1, \dots, M \right\} \leq \\ &\leq \max \left\{ \sum_{j=1}^N |a_{ij}| : i = 1, \dots, M \right\} \end{aligned}$$

□

Por lo que si ya hemos calculado alguna de estas dos últimas normas, podemos saber la otra sin tener que volver a calcular el máximo, es decir, la relación entre ambas viene en la siguiente proposición.

Proposición 1.8. $\|A\|_1 = \|A^T\|_\infty \quad \forall A \in \mathbb{R}^{M \times N}$

Hay que tener en cuenta que $\|\cdot\|_2$ no induce en $\mathbb{R}^{M \times N}$ Frobenius.

Además, para las matrices **cuadradas** tenemos la siguiente definición.

Definición 1.15 (Radio espectral). Sea $A \in \mathbb{R}^{N \times N}$, denotaremos como **radio espectral de A** a:

$$\rho(A) := \max \{ |\lambda| : \lambda \in \mathbb{C} \wedge \det(A - \lambda I) = 0 \}$$

La siguiente proposición muestra una manera más fácil de calcular la norma euclídea de un vector de \mathbb{R}^N , que es calculando la suma de las coordenadas al cuadrado.

Proposición 1.9. $\|x\|_2 = \sqrt{x^T x} \quad \forall x \in \mathbb{R}^N$

Definición 1.16 (Matriz semidefinida positiva). Diremos que $A \in \mathbb{R}^{N \times N}$ es **semidefinida positiva** $\Leftrightarrow x^T A x \geq 0 \quad \forall x \in \mathbb{R}^N$

Las matrices semidefinidas positivas tienen la siguiente propiedad.

Ejercicio 1.3. Demuestra las siguientes proposiciones:

(i) Sea $A \in \mathbb{R}^{N \times N}$ semidefinida positiva. Si λ es un valor propio de $A \Rightarrow \lambda \geq 0$.

(ii) Sea $P \in \mathbb{R}^{N \times N}$ una matriz ortogonal, entonces

$$\{x \in \mathbb{R}^N : \|x\|_2 = 1\} = \{P^T x : x \in \mathbb{R}^N \wedge \|x\|_2 = 1\}$$

(iii) Si $\lambda_1, \dots, \lambda_N \geq 0 \Rightarrow \sup \left\{ \sqrt{\sum_{i=1}^N \lambda_i y_i^2} : y \in \mathbb{R}^N \wedge \|y\|_2 = 1 \right\} = \sqrt{\max \{\lambda_i : i = 1, \dots, N\}}$

Lema 1.1. $(AB)^t = B^t A^t$.

Solución.

(i) Como λ es valor propio de A , entonces $\exists x \in \mathbb{R}^N : x \neq 0 \wedge Ax = \lambda x$

Luego

$$0 \leq x^T Ax = x^T \lambda x = \lambda x^T x = \lambda \|x\|_2^2$$

Como $x \neq 0$, entonces $0 \leq \lambda$

(ii) Vamos a demostrar la doble inclusión, lo que dará la igualdad.

▪ $\mathcal{I} \supseteq ?$

Sea $x \in \mathbb{R}^N : \|x\|_2 = 1 \Rightarrow \mathcal{I} \|P^T x\|_2 = 1 ?$

$$\|P^T x\|_2 = \sqrt{x^T P P^T x} = \sqrt{x^T x} = \|x\|_2 = 1$$

▪ $\mathcal{I} \subseteq ?$

$$1 = \|x\|_2 = \sqrt{x^T x} = \sqrt{x^T I x} = \sqrt{x^T P P^T x} = \|P^T x\|_2$$

(iii) Queda como ejercicio.

Una manera más sencilla de calcular la norma de una matriz es la siguiente.

Proposición 1.10. $\|A\|_2 = \sqrt{\rho(A^T A)}$, con $A \in \mathbb{R}^{M \times N}$.

Lema 1.2. Si $A \in \mathbb{R}^{M \times N} \Rightarrow A^T A \in S_N$

Definición 1.17. Una matriz ortogonal es una matriz cuadrada cuya matriz inversa coincide con su matriz traspuesta. El conjunto de matrices ortogonales constituyen una representación lineal del grupo ortogonal $O(n, \mathbb{R})$.

Demostración 1.4. Como $A^T A$ es simétrica, entonces $\exists P \in \mathbb{R}^{N \times N}$ ortogonal tal que

$$P^T A^T A P = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$$

donde $\lambda_1, \dots, \lambda_N$ son los valores propios de $A^T A$, no negativos (ya que $A^T A$ es semidefinida positiva).

Como P es ortogonal, entonces $\{x \in \mathbb{R}^N : \|x\|_2 = 1\} = \{P^T x : x \in \mathbb{R}^N, \|x\|_2 = 1\}$.

$$y = P^T x \Leftrightarrow Py = x$$

$$\begin{aligned} \|A\|_2 &= \sup \{\|Ax\|_2 : x \in \mathbb{R}^N, \|x\|_2 = 1\} = \sup \{\sqrt{x^T A^T A x} : x \in \mathbb{R}^N, \|x\|_2 = 1\} = \\ &= \sup \{\sqrt{y^T P^T A^T A P y} : y \in \mathbb{R}^N, \|y\|_2 = 1\} = \sup \left\{ \sqrt{\sum_{i=1}^N \lambda_i y_i^2} : y \in \mathbb{R}^N, \|y\|_2 = 1 \right\} = \\ &= \sqrt{\max \{\lambda_i : i = 1, \dots, N\}} = \sqrt{\rho(A^T A)} \end{aligned}$$

□

Definición 1.18 (Norma matricial). Una norma en $\mathbb{R}^{N \times N}$ se dice **matricial** cuando

$$\|AB\| \leq \|A\| \|B\| \quad \forall A, B \in \mathbb{R}^{N \times N}$$

Hay que tener en cuenta de que no toda norma en $\mathbb{R}^{N \times N}$ es matricial, por ejemplo:

Ejemplo 1.2. Vamos a utilizar la siguiente norma

$$\|A\| := \max \{|a_{ij}| : i, j = 1, \dots, N\} \quad \forall A \in \mathbb{R}^{N \times N}$$

Sean

$$A = B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Luego

$$2 = \|AB\| > \|A\|\|B\| = 1$$

Por lo que esta norma no es matricial.

Proposición 1.11. Toda norma en $\mathbb{R}^{N \times N}$ inducida por una norma en \mathbb{R}^N es matricial.

Demostración 1.5. Sea $\|\cdot\|$ una norma en $\mathbb{R}^{N \times N}$ inducida por una norma en \mathbb{R}^N , entonces

$$\|A\| := \sup \left\{ \|Ax\| : x \in \mathbb{R}^N \wedge \|x\| = 1 \right\}$$

Sabemos que la norma es inducida, luego se cumple que

$$\|Ax\| \leq \|A\|\|x\| \quad \forall x \in \mathbb{R}^N$$

Tenemos que probar que $\|AB\| \leq \|A\|\|B\|$

$$\begin{aligned} \|AB\| &= \sup \left\{ \|ABx\| : x \in \mathbb{R}^N \wedge \|x\| = 1 \right\} \leq \sup \left\{ \|A\|\|Bx\| : x \in \mathbb{R}^N \wedge \|x\| = 1 \right\} \leq \\ &\leq \sup \left\{ \|A\|\|B\|\|x\| : x \in \mathbb{R}^N \wedge \|x\| = 1 \right\} = \|A\|\|B\| \end{aligned}$$

□

Teorema 1.2. $\lim_{n \rightarrow \infty} A^n = 0 \Leftrightarrow \rho(A) < 1 \quad \forall A \in \mathbb{R}^{N \times N}$

Este teorema nos deja dos importantes consecuencias.

Corolario 1.1. Sea $A \in \mathbb{R}^{N \times N}$ una matriz triangular, entonces

$$\lim_{n \rightarrow \infty} A^n = 0 \Leftrightarrow \max \{ |a_{ii}| < 1 : i = 1, \dots, N \}$$

Corolario 1.2. Sean $N \geq 1$, $A \in \mathbb{R}^{N \times N}$ y $\|\cdot\|$ una norma matricial en $\mathbb{R}^{N \times N}$ tal que $\|A\| < 1$, entonces $\rho(A) < 1$.

Hay que tener en cuenta que no se cumple la implicación contraria, por ejemplo:

Ejemplo 1.3. Sea

$$A = \begin{bmatrix} 0.5 & 500 \\ 0 & 0.5 \end{bmatrix} \Rightarrow \rho(A) = 0.5 < 1$$

pero

$$\|A\|_{\infty} = 500.5 \geq 1$$

1.2 Problemas bien planteados. Estabilidad

Nos planteamos el siguiente problema:

Sean X e Y subconjuntos no vacíos de sendos espacios normados reales, $f : X \rightarrow Y$ una aplicación, $y_0 \in Y$. Entonces tenemos que encontrar $x_0 \in X : f(x_0) = y_0$.

Denotaremos a x_0 como la solución que resuelve el problema determinado por f y a y_0 los datos, es decir, son números. Si tenemos un conjunto finito de números, usaremos el vector de \mathbb{R}^N o matriz, y si tenemos infinitos datos, usaremos una función.

Ejemplo 1.4. Sean $A \in \mathbb{R}^{M \times N}$ e $y \in \mathbb{R}^M$. Determinar una solución del sistema de ecuaciones lineales cuya matriz de coeficientes sea A y su vector de términos independientes sea y

$$X = \mathbb{R}^N, \quad Y = \mathbb{R}^M, \quad f(x) = Ax = y$$

Definición 1.19 (Problema bien planteado). Un problema está **bien planteado** cuando es **unisolvente** y **estable**:

- (i) $\exists! x_0 \in X : f(x_0) = y_0$.
- (ii) x_0 depende continuamente de los datos y_0 .

Ejemplo 1.5. Veamos un problema mal planteado.

Sean $X := \mathbb{R}$, $Y := \mathbb{R}_+$ y $f(x) := |x|$, $\forall x \in X$

Observamos que este problema no es unisolvente, ya que si $y_0 = 1$ (lo mismo vale $\forall y_0 > 0$) tenemos que $f(-1) = 1 = f(1)$.

Definición 1.20 (Resolvente). Denotaremos a la función g como la **resolvente** de f si g es la

inversa de f , para todo $y \in Y$ unisolvente.

Ejemplo 1.6. Sean $X := \mathbb{R}$, $Y := \mathbb{R}_+$ y $f(x) := e^x$, $\forall x \in X$

Entonces este problema es unisolvente, luego tiene resolvente (inversa): $g(y) = \log(y)$, $\forall y_0 \in Y$.

Podemos ver la **estabilidad** de un problema intuitivamente, es decir, a pequeñas perturbaciones de los datos y_0 corresponden pequeñas perturbaciones de la solución x_0 .

Ejemplo 1.7. Consideremos el problema siendo $X := [-1, 1] =: Y$ y la función

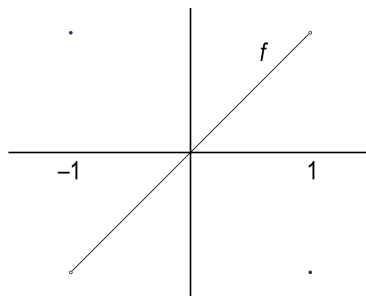
$$f(x) := \begin{cases} x & \text{si } -1 < x < 1 \\ -x & \text{si } x = \pm 1 \end{cases}$$

Entonces este problema es unisolvente, ya que la función $f : [-1, 1] \rightarrow [-1, 1]$ es biyectiva y $g = f$.

Pero para que esté bien planteado, tiene que ser, además, estable. Veamos que no es estable, ya que a pequeñas perturbaciones del dato $y_0 = -1$ (lo mismo con $y_0 = 1$) no corresponden pequeñas perturbaciones de $x_0 = 1$.

Sea $y_n := -1 + \frac{1}{n} \rightarrow y_0 = -1$, entonces $g(y_n) = -1 + \frac{1}{n} \rightarrow -1$, ya que $y_n = g(y_n)$.

Luego $|-1 - x_0| = 2$, es decir, las perturbaciones de x_0 son muy grandes.



Ahora nos planteamos la siguiente cuestión: ¿La estabilidad del problema y la continuidad de la resolvente g tienen algo que ver? Veamos en el siguiente ejemplo que no siempre es así.

Ejemplo 1.8. Consideremos el problema siendo la función

$$f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$$

$$f(x) := \left(\frac{x}{10}\right)^{10} \quad \forall x \geq 0$$

Tenemos que este problema es unisolviente y su resolvente es

$$g : \mathbb{R}_+ \rightarrow \mathbb{R}_+$$

$$g(y) = 10y^{\frac{1}{10}} \quad \forall y \geq 0$$

Tenemos que g es continua, luego pequeños cambios de los datos $y \in \mathbb{R}_+$ nos llevan a cambios cercanos de las correspondientes soluciones $x \in \mathbb{R}_+$ pero no controlados, se aproximan a una velocidad diferente. Vamos a comprobarlo.

Tenemos que $y_0 = 0 = x_0$. Sea y un dato próximo a y_0 , por ejemplo $y = 10^{-10}$. Entonces la solución correspondiente es

$$x = g(y) = 10y^{\frac{1}{10}} = 10 \cdot \left(10^{-10}\right)^{\frac{1}{10}} = \frac{10}{10} = 1$$

Por lo que tenemos lo siguiente:

$$|y - y_0| = 10^{-10}$$

$$|x - x_0| = 1$$

Es decir, el número 10^{-10} indica la velocidad con la que se mueven los datos y , en comparación con el número 1, que indica que los datos x se mueven mucho más rápido, lo que nos lleva a que, aunque g sea continua, la velocidad de aproximación es diferente, luego no hay estabilidad.

Luego podríamos decir que la estabilidad es una condición que fuerce un control de los valores de las soluciones en función de los datos, de forma que pequeñas perturbaciones de y_0 generen perturbaciones pequeñas y controladas de x_0 .

Formalización del concepto de estabilidad –en un contexto métrico se conoce como *lipschitzianidad local*–.

Definición 1.21 (Aplicación estable). Sean X e Y subconjuntos no vacíos de sendos espacios normados, $g : Y \rightarrow X$ una aplicación e $y_0 \in Y$. Diremos que g es **estable** en y_0 cuando

$$\exists \delta, M > 0 : \sup \left\{ \frac{\|g(y) - g(y_0)\|}{\|y - y_0\|} : y \in Y \wedge 0 < \|y - y_0\| < \delta \right\} < M$$

y que g es **estable** si lo es en todos y cada uno de los elementos de Y . Otra forma de definirlo es: g estable en $y_0 \iff \exists \delta > 0, M > 0 : \|y - y_0\| < \delta \Rightarrow \|g(y) - g(y_0)\| \leq M\|y - y_0\|$

Teorema 1.3. Teorema del Valor Medio. Sea $f : [a, b] \rightarrow \mathbb{R}$, con f continua en $[a, b]$ y derivable en (a, b) . Tomando el segmento \overline{ab} , $\exists c$ tal que $f'(c) = \frac{f(b) - f(a)}{b - a}$.

La estabilidad es más fuerte que la continuidad pero más débil que la clase 1.

Definición 1.22 (Problema estable). Un problema es **estable** en $y_0 \in Y$ si su resolvente $g : Y \rightarrow X$ lo es en dicho punto, y es **estable** si lo es en cualquier dato de Y .

Habrá mejor comportamiento cuanto más pequeño sea M .

Proposición 1.12. g es estable en $y_0 \Rightarrow g$ es continua en y_0

La implicación \Leftarrow se cumple al tomar la resolvente del ejemplo anterior en 0, o si se quiere, la función raíz cuadrada en 0.

Proposición 1.13. Toda función real de variable real de clase C^1 es estable.

Para demostrar esta proposición hay que usar el Teorema del Valor Medio. Además, el recíproco no es cierto.

Ejemplo 1.9. Sea f una función de variable real de clase C^1 , vamos a medir su estabilidad en $x_0 \in \mathbb{R}$.

Si $x_0 f(x_0) \neq 0$, entonces el cociente entre el error relativo cometido cerca de $f(x_0)$ y el error relativo de x_0 es:

$$\left| \frac{\frac{f(x)-f(x_0)}{f(x_0)}}{\frac{x-x_0}{x_0}} \right| = \left| \frac{f(x)-f(x_0)}{x-x_0} \right| \left| \frac{x_0}{f(x_0)} \right|$$

Si $x_0 f(x_0) = 0$, entonces los errores absolutos cerca de $f(x_0)$ e x_0 es:

$$\left| \frac{f(x)-f(x_0)}{x-x_0} \right|$$

De forma precisa:

Definición 1.23 (Condicionamiento de una aplicación). Dada una función $f \in C^1(\mathbb{R})$ y un punto $x_0 \in \mathbb{R}$, entonces el **condicionamiento relativo** de f en x_0 viene dado por

$$c(f, x_0) := \left| \frac{f'(x_0) \cdot x_0}{f(x_0)} \right|$$

siempre que $x_0 \cdot f(x_0) \neq 0$.

Además, el **condicionamiento absoluto** de f en x_0 es

$$C(f, x_0) := |f'(x_0)|$$

Ídem en funciones reales de variable real definidas en intervalos de \mathbb{R} y de clase C^1 .

Para calcular el condicionamiento en funciones de varias variables, usamos lo siguiente:

$$f \in C^1(\mathbb{R}^N, \mathbb{R}^M), f = [f_1 \cdots f_M],$$

$$\frac{\partial f}{\partial x}(x_0) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x_0) & \frac{\partial f_1}{\partial x_2}(x_0) & \cdots & \frac{\partial f_1}{\partial x_N}(x_0) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1}(x_0) & \frac{\partial f_M}{\partial x_2}(x_0) & \cdots & \frac{\partial f_M}{\partial x_N}(x_0) \end{bmatrix} \in \mathbb{R}^{M \times N}$$

$$c(f, x_0) := \frac{\left\| \frac{\partial f}{\partial x}(x_0) \right\| \|x_0\|}{\|f(x_0)\|}$$

$$C(f, x_0) := \left\| \frac{\partial f}{\partial x}(x_0) \right\|$$

Nota. No entra en el examen.

Definición 1.24 (Condicionamiento de un problema). Si en un problema (P) la resolvente es de clase C^1 e $y_0 \in Y$, el **condicionamiento relativo** o **absoluto** de (P) son los de su resolvente en dicho punto.

Aunque se trata de un concepto bastante ambiguo, suele decirse que una aplicación (o un problema) está **bien condicionado** en un punto si su condicionamiento es pequeño y en caso contrario, **mal condicionado**.

Ejemplo 1.10. Sea

$$f : (0, 1) \rightarrow (0, \pi/2)$$

$$f(x) := \arcsen(x)$$

Veamos si está bien planteado.

La resolvente es

$$g : (0, \pi/2) \rightarrow (0, 1)$$

$$g(y) := \sen(y)$$

Luego está bien planteado.

Veamos si está bien condicionado.

Para ello calculamos el condicionamiento relativo en $y_0 \in (0, \pi/2)$.

$$c(g, y_0) = \frac{|g'(y_0)y_0|}{|g(y_0)|} = y_0 \frac{\cos(y_0)}{\sen(y_0)}$$

Como $c(g, y_0) \in (0, 1)$, entonces está bien condicionado.

Ejemplo 1.11. Sea $A \in \mathbb{R}^{N \times N}$ regular e $y \in \mathbb{R}^N$, vamos a encontrar $x \in \mathbb{R}^N$ tal que $f(x) = y$.

Sea $f(x) = Ax$. Como A es regular, el problema es unisolvante, luego tiene resolvente

$$g : \mathbb{R}^N \rightarrow \mathbb{R}^N$$

$$g(y) := A^{-1}y$$

Además, el problema es estable en todo $y_0 \in \mathbb{R}^N$, ya que:

$$g(y) - g(y_0) = A^{-1}y - A^{-1}y_0 = A^{-1}(y - y_0) \quad y \quad \|AB\| \leq \|A\|\|B\|$$

$$\|g(y) - g(y_0)\|_\infty \leq \|A^{-1}\|_\infty \|y - y_0\|_\infty$$

Ahora calcularemos el condicionamiento relativo.

$$c(g, y_0) = \frac{\left\| \frac{\partial g}{\partial y}(y_0) \right\| \|y_0\|}{\|g(y_0)\|} = \frac{\|A^{-1}\| \|y_0\|}{\|A^{-1}y_0\|} = \frac{\|A^{-1}\| \|y_0\|}{\|x_0\|}$$

Luego $Ax_0 = y_0$.

Si $\|A^{-1}\|$ es grande, x_0 e y_0 son del mismo orden, entonces el condicionamiento será grande. Observémoslo en el siguiente ejemplo.

Ejemplo 1.12. Sea $A = \begin{bmatrix} 1 & 1 \\ 1 & 0.999 \end{bmatrix}$, un dato $y_0 = \begin{bmatrix} 2 \\ 1.999 \end{bmatrix}$ y su solución $x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Veamos que hay un mal condicionamiento, ya que el condicionamiento es demasiado grande en comparación con el resto. Para ello, calculamos antes la inversa de la matriz A y la norma del máximo de esta, de la solución x_0 y del dato y_0 .

$$A^{-1} = \begin{bmatrix} -999 & 1000 \\ 1000 & -1000 \end{bmatrix}$$

$$\|A^{-1}\|_\infty = 2000, \quad \|x_0\|_\infty = 1, \quad \|y_0\|_\infty = 2$$

Luego el condicionamiento se calculará con la expresión obtenida anteriormente usando la norma infinito.

$$c(g, y_0) = \frac{\|A^{-1}\|_\infty \|y_0\|_\infty}{\|x_0\|_\infty} = 4000$$

Como hemos dicho, es un condicionamiento demasiado grande y nos preguntamos ¿cuál es el problema? Pues que al calcular las soluciones de otros datos próximos al inicial, las soluciones obtenidas no son próximas a la solución inicial. Comprobémoslo.

Sea y un dato próximo a y_0 , por ejemplo $y = \begin{bmatrix} 2 \\ 1.998 \end{bmatrix}$. Es próximo, ya que $\|y - y_0\|_\infty = 0.001$.

Su solución es $x = A^{-1}y = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$. No es próxima a la solución inicial, ya que $\|x - x_0\|_\infty = 1$.

La interpretación geométrica de este resultado es: las columnas de A forman base de $\mathbb{R}^{2 \times 2}$, calculamos las coordenadas de y en y_0 en dicha base. Un pequeño cambio entre y e y_0 genera un cambio grande de coordenadas por ser los vectores de la base casi paralelos.

Ahora nos planteamos la siguiente pregunta: ¿está globalmente acotado el condicionamiento relativo de un sistema de ecuaciones lineales? La respuesta es afirmativa y nos da una idea de cómo de estable es.

La siguiente definición será útil para calcular el condicionamiento de una matriz cuando la norma es inducida.

Definición 1.25 (Condicionamiento de una matriz regular). Si $\|\cdot\|$ denota la norma matricial en $\mathbb{R}^{N \times N}$ inducida por una norma en \mathbb{R}^N , que notaremos igualmente como $\|\cdot\|$, entonces definimos el **condicionamiento** $c(A)$ de una matriz regular $A \in \mathbb{R}^{N \times N}$ como

$$c(A) := \|A^{-1}\| \|A\|$$

Además,

$$\sup \{c(g, y_0) : y_0 \neq 0\} = \sup \left\{ \frac{\|A^{-1}\| \|y_0\|}{\|A^{-1}y_0\|} : y_0 \neq 0 \right\} = \sup \left\{ \frac{\|A^{-1}\| \|Ax_0\|}{\|x_0\|} : x_0 \neq 0 \right\} = \|A^{-1}\| \|A\|$$

Lo que nos lleva a que si tenemos un mal condicionamiento del sistema, el condicionamiento de la matriz de coeficientes es grande, y viceversa.

También tenemos que $c(A) \geq 1$.

Si retomamos el ejemplo anterior, usando la norma del máximo en \mathbb{R}^2 y la correspondiente norma matricial inducida, obtenemos que $c(A) = 4000$.

1.3 Algoritmos. Algoritmo PageRank de Google

Definición 1.26 (Algoritmo). Procedimiento que describe de forma precisa, y siempre mediante un número finito de operaciones aritméticas y lógicas elementales, la resolución de un problema.

El algoritmo recoge las instrucciones que permiten al ejecutor del mismo resolver completamente el problema. El ejecutor suele ser un ordenador y, de hecho, en la mayoría de los casos, no puede ser una persona.

Definición 1.27 (Análisis Numérico). Se ocupa de diseñar algoritmos que permitan la resolución efectiva de problemas bien planteados y que involucran números reales.

Definición 1.28 (Complejidad de un algoritmo). Medida del tiempo de ejecución y que suele expresarse en términos de un parámetro asociado al problema.

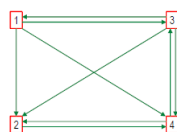
Al desarrollar y analizar un algoritmo obtenemos precisión, estabilidad y efectos de la representación finita de los números reales.

El Algoritmo PageRank de Google mide la relevancia de las páginas web (dominios...) con enlaces en común. Por lo que aquí nuestro problema es calcular la relevancia de una página (P), teniendo en cuenta el número de enlaces de otras páginas a (P) y la importancia de las páginas que establecen enlace con (P).

Nuestro modelo es:

- Denotaremos al conjunto (finito) de páginas web como $1, 2, \dots, x_N$.
- Denotaremos a las páginas como $1, 2, \dots, i$.
- La relevancia de la página i se representa por $x_i \geq 0$.
- La página i será más importante que la página j si $x_i > x_j$.

Ejemplo 1.13. Consideremos la estructura de enlaces entre cuatro páginas web.



La relevancia de cada página será la suma de los cocientes entre la relevancia de la página que entra y los enlaces que salen de esta.

$$x_1 = \frac{x_3}{3}, \quad x_2 = \frac{x_1}{3} + \frac{x_3}{3} + \frac{x_4}{2}, \quad x_3 = \frac{x_1}{3} + \frac{x_4}{2}, \quad x_4 = \frac{x_1}{3} + x_2 + \frac{x_3}{3}$$

Por lo que siempre obtendremos un sistema compatible indeterminado con un parámetro (rango = 3).

Si tomamos x_4 como parámetro y le asignamos el valor 10, tenemos lo siguiente:

$$x_1 = 1.875, \quad x_2 = 7.5, \quad x_3 = 5.625, \quad x_4 = 10$$

2 Errores de redondeo. Iteradores

Las principales fuentes de error suelen ser equivocarse al elegir el modelo de un problema, la medida de los datos experimentales, error al truncar o al redondear y consecuentemente, al operar, se produce una propagación del error.

2.1 Sistema posicional y números máquina

Los ordenadores trabajan con un subconjunto finito de números reales, los **números máquina**, subconjunto que depende de las especificaciones del ordenador.

SISTEMAS POSICIONALES DE NUMERACIÓN

Sea $b \in \mathbb{N}$ la base binaria o decimal, según el estándar ISO/IEC/IEEE60559:2011 del IEEE (Institute of Electrical and Electronic Engineers, www.ieee.org). Sea $s \in \{0, 1\}$ el signo, $N, M \in \mathbb{N} \cup 0$ y sea x_k la cifra en la posición k tal que $0 \leq x_k < b$, $\forall k = -M, \dots, N$.

Definición 2.1 (Representación Posicional). La **representación posicional** de un número real es:

$$x = (-1)^s \sum_{n=-M}^N x_n b^n$$

$$x_b := (-1)^s \cdot (x_N \dots x_1 x_0 . x_{-1} x_{-2} \dots x_{-M})_b$$

Además, denominaremos al punto entre x_0 y x_{-1} como **punto binario** o **punto decimal**.

Ejemplo 2.1. Sea $x_{10} = x = 101.11$, su representación posicional es $x = 1 \cdot 10^2 + 1 \cdot 10^0 + 1 \cdot 10^{-1} + 1 \cdot 10^{-2}$.

Sea $y_2 = (101.11)_2$, su representación posicional es $y = 1 \cdot 2^2 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2}$.

Variando $M, N \in \mathbb{N} \cup \{0\}$ tenemos un subconjunto denso de \mathbb{R} . Además, la serie geométrica de razón menor que 1 converge. Se puede extender a infinito:

$$x = (-1)^s \sum_{n=-\infty}^N x_n b^n$$

$$x_b := (-1)^s \cdot (x_N \dots x_1 x_0 . x_{-1} x_{-2} \dots)_b$$

El siguiente ejemplo aprenderemos representar en forma posicional infinita a partir de otra infinita.

Ejemplo 2.2.

$$(0.\widehat{001})_2 = \sum_{n=1}^{\infty} (2^{-3})^n = \frac{1}{7} = 0.1\widehat{42857}$$

Para calcular la convergencia de esta serie, como la razón es menor que 1, hemos usado lo siguiente:

$$\alpha \neq 1, n > k \geq 0 \Rightarrow \sum_{j=k}^n \alpha^j = \frac{\alpha^k - \alpha^{n+1}}{1 - \alpha}$$

En particular, si

$$|\alpha| < 1 \Rightarrow \sum_{j=k}^{\infty} \alpha^j = \frac{\alpha^k}{1 - \alpha}$$

NÚMEROS MÁQUINA

Estos números los utiliza el ordenador. Denotamos a la base como b y a las posiciones de memoria como N .

Definición 2.2 (Representación con punto fijo). Sea $k \in \mathbb{N}$ fijo, $N - k - 1$ dígitos enteros, k dígitos tras el punto a_n tal que $0 \leq a_n \leq b - 1$. Entonces las N posiciones de memoria son: $N = \text{signo} + \text{cifras significativas} = 1 + N - 1$. La representación con punto fijo es la siguiente:

$$(-1)^s b^{-k} \cdot \sum_{n=0}^{N-2} a_n b^n \sim (-1)^s \cdot (a_{N-2} \dots a_k . a_{k-1} \dots a_0)_b$$

Definición 2.3 (Representación con número flotante). Sea $t \in \mathbb{N}$ el número máximo de dígitos o cifras significativas a_n tales que $0 \leq a_n \leq b - 1$, sea $m = a_1 \dots a_t$ la mantisa tal que $0 \leq m \leq b^t - 1$, sea $e \in \mathbb{Z}$ el exponente, tal que $L \leq e \leq U$ donde $L, U \in \mathbb{Z}$, $L \leq U$. Entonces las N posiciones de memoria son $N = \text{signo} + \text{cifras significativas} + \text{dígitos del exponente} = 1 + t + N - t - 1$. La representación con punto flotante es:

$$(-1)^s b^e \cdot \sum_{n=1}^t a_n b^{-n} \sim (-1)^s \cdot (0.a_1 \dots a_t) \cdot b^e = (-1)^s \cdot m \cdot b^{e-t}$$

Ejemplo 2.3. Sea $x = -3.4567$ y la base $b = 10$, vamos a representarlo de las dos maneras.

- (i) Representación con punto fijo con $k = 4$ y $N = 6$. Luego $x = (-1) \cdot 10^{-4} \cdot (3 \cdot 10^4 + 4 \cdot 10^3 + 5 \cdot 10^2 + 6 \cdot 10 + 7 \cdot 10^0) = (-1)(3.4567)$

(ii) Representación con punto flotante con $t = 6$ y $e = 2$. Luego $x = (-1) \cdot 10^2 \cdot (0 \cdot 10^{-1} + 3 \cdot 10^{-2} + 4 \cdot 10^{-3} + 5 \cdot 10^{-4} + 6 \cdot 10^{-5} + 7 \cdot 10^{-6}) = (-1)(0.034567) \cdot 10^2$

Usualmente hay dos representaciones de punto flotante: la precisión simple y la doble.

Si un número no está normalizado, es decir, la cifra significativa principal a_1 no es 0, entonces ese número tendrá varias representaciones. Veámoslo en el siguiente ejemplo.

Ejemplo 2.4. Sean $b = 2$, $t = 3$, $L = 1$, $U = 3$ y punto flotante para $x = 1$, entonces:

$$(0.100) \cdot 2^1 = (0.010) \cdot 2^2 = (0.001) \cdot 2^3$$

Para evitar este problema, usaremos la representación normalizada que viene definida a continuación.

Definición 2.4 (Notación del sistema normalizado de punto flotante).

$$\mathbb{F}(b, t, L, U) := \{0\} \cup \left\{ (-1)^s b^e \sum_{n=1}^t a_n b^{-n} : s = 0, 1, a_1 \neq 0, 0 \leq a_1, \dots, a_t \leq b-1, L \leq e \leq U \right\}$$

Proposición 2.1. Sean $t \in \mathbb{N}$, $L, U \in \mathbb{Z}$ con $L \leq U$ y $x \in \mathbb{F}(b, t, L, U)$. Entonces:

- (i) $-x \in \mathbb{F}(b, t, L, U)$.
- (ii) $b^{L-1} \leq |x| \leq b^U(1 - b^{-t})$.
- (iii) $\text{card}(\mathbb{F}(b, t, L, U)) = 2(b-1)b^{t-1}(U-L+1) + 1$.

Demostración como ejercicio

Demostración 2.1.

Veamos cuál es el número no nulo más pequeño.

$$x = (0.100 \dots 0 \cdot 0)b^L = \frac{1}{b} \cdot b^L = b^{L-1}$$

El mayor será:

$$y = (0.b-1 \dots b-1) \cdot b^U = b^U(b-1)\left(\frac{1}{b} + \frac{1}{b^2} + \dots + \frac{1}{b^t}\right) = b^U(b-1)\frac{\frac{1}{b} - \frac{1}{b^{t+1}}}{1 - \frac{1}{b}} = b^U(b-1)\frac{b^t-1}{b^{t+1}} \frac{b}{b-1} = b^U(1-b^{-t})$$

□

Veamos un importante ejemplo en el que se representan todos los números positivos de un sistema de punto flotante normalizado.

Ejemplo 2.5. Sea $\mathbb{F}(2, 4, -1, 1)$, entonces los números estrictamente positivos de ese sistema de punto flotante son:

$$\begin{aligned}
 (0.1111) \cdot 2 &= \frac{15}{8} & (0.1110) \cdot 2 &= \frac{7}{4} & (0.1101) \cdot 2 &= \frac{13}{8} & (0.1100) \cdot 2 &= \frac{3}{2} \\
 (0.1011) \cdot 2 &= \frac{11}{8} & (0.1010) \cdot 2 &= \frac{5}{4} & (0.1001) \cdot 2 &= \frac{9}{8} & (0.1000) \cdot 2 &= 1 \\
 (0.1111) \cdot 2^0 &= \frac{15}{16} & (0.1110) \cdot 2^0 &= \frac{7}{8} & (0.1101) \cdot 2^0 &= \frac{13}{16} & (0.1100) \cdot 2^0 &= \frac{3}{4} \\
 (0.1011) \cdot 2^0 &= \frac{11}{16} & (0.1010) \cdot 2^0 &= \frac{5}{8} & (0.1001) \cdot 2^0 &= \frac{9}{16} & (0.1000) \cdot 2^0 &= \frac{1}{2} \\
 (0.1111) \cdot 2^{-1} &= \frac{15}{32} & (0.1110) \cdot 2^{-1} &= \frac{7}{16} & (0.1101) \cdot 2^{-1} &= \frac{13}{32} & (0.1100) \cdot 2^{-1} &= \frac{3}{8} \\
 (0.1011) \cdot 2^{-1} &= \frac{11}{32} & (0.1010) \cdot 2^{-1} &= \frac{5}{16} & (0.1001) \cdot 2^{-1} &= \frac{9}{32} & (0.1000) \cdot 2^{-1} &= \frac{1}{4}
 \end{aligned}$$

(El primer decimal no puede ser 0) Además, podemos calcular lo siguiente:

Número de elementos de $\mathbb{F}(2, 4, -1, 1)$ (positivos, negativos y cero):

$$2(b-1)b^{t-1}(U-L+1)+1=49$$

Valores mínimo y máximo (positivos):

$$b^{L-1} = \frac{1}{4}, \quad b^U(1-b^{-t}) = \frac{15}{8}$$

El sistema $\mathbb{F}(b, t, L, U)$ no se distribuye uniformemente, aunque sí por bloques.

Definición 2.5 (Épsilon máquina). Para un sistema de punto flotante $\mathbb{F}(b, t, L, U)$ con $L \leq 1 \leq U$, el **épsilon máquina**, que se escribe como ε_M , es la distancia entre el menor número de $\mathbb{F}(b, t, L, U)$ mayor que 1 y la propia unidad, es decir,

$$\varepsilon_M := b^{1-t}$$

Como $L \leq 1 \leq U \Rightarrow 1 \in \mathbb{F}(b, t, L, U)$

2.2 Redondeo en sistemas de punto flotante y su aritmética

En este apartado vamos a trabajar sobre el contexto de los números máquina representados en el sistema de punto flotante $\mathbb{F}(b, t, L, U)$. Primero, tengamos en cuenta lo siguiente:

(i) $\mathbb{F}(b, t, L, U) \neq \mathbb{R}$.

(ii) El resultado de operar con dos números de $\mathbb{F}(b, t, L, U)$ no queda dentro de dicho sistema

necesariamente, por ejemplo:

$$\frac{1}{4}, \frac{9}{32} \in \mathbb{F}(2, 4, -1, 1), \text{ pero } \frac{1}{4} + \frac{9}{32} = \frac{17}{32} \notin \mathbb{F}(2, 4, -1, 1)$$

Definición 2.6 (Truncatura). Fijado un sistema de punto flotante concreto $\mathbb{F}(b, t, L, U)$ (al que no se hará referencia si no hay lugar a ambigüedad) si $x \in \mathbb{R}$ es el número real

$$x = (-1)^s b^e \sum_{n=1}^{\infty} a_n b^{-n}$$

entonces su **truncatura** (en dicho sistema) es un número de $\mathbb{F}(b, t, L, U)$

$$tr(x) := (-1)^s \cdot (0.a_1 \dots a_t) \cdot b^e$$

Ejemplo 2.6. Sea el sistema de punto flotante $\mathbb{F}(2, 4, -1, 1)$ y el número real $x = 1.6875$, calcularemos su truncatura. Primero lo pasaremos a la base indicada:

$$1.6875 = (0.11011) \cdot 2$$

Ahora calculamos la truncatura:

$$tr(1.6875) = (0.1101) \cdot 2 = \frac{13}{8} = 1.625$$

Es decir, hemos quitado la última cifra, ya que en el número había 5 cifras y según el sistema, el número de cifras es 4. Por último lo hemos pasado a base decimal.

Definición 2.7. Para un sistema de punto flotante $\mathbb{F}(b, t, L, U)$, el **redondeo** del número real

$$x = (-1)^s b^e \sum_{n=1}^{\infty} a_n b^{-n}$$

es el número de $\mathbb{F}(b, t, L, U)$

$$rd(x) = tr \left(x + (-1)^s \frac{b}{2} \frac{b^e}{b^{t+1}} \right)$$

El **redondeo** se puede calcular de otra manera.

Proposición 2.2. Para un sistema de punto flotante $\mathbb{F}(b, t, L, U)$, el **redondeo** del número real

$$x = (-1)^s b^e \sum_{n=1}^{\infty} a_n b^{-n}$$

es el número de $\mathbb{F}(b, t, L, U)$

$$rd(x) := (-1)^s \cdot (0.a_1 \dots a_{t-1} r_t) \cdot b^e$$

donde

$$r_t := \begin{cases} a_t & \text{si } a_{t+1} < \frac{b}{2} \\ a_t + 1 & \text{si } a_{t+1} \geq \frac{b}{2} \end{cases}$$

Ejemplo 2.7. Sea el sistema de punto flotante $\mathbb{F}(2, 4, -1, 1)$ y el número real 1.6875, vamos a calcular su redondeo. Como la base es 2, vamos a pasarlo a esa base: $1.6875 = (0.11011) \cdot 2$ Para redondearlo, lo podemos hacer de dos formas.

(i) Directamente con 5ª cifra tras el punto ($t = 4$):

$$rd(1.6875) = (0.1110) \cdot 2 = \frac{7}{4} = 1.75$$

(ii) Con la definición:

$$rd(1.6875) = tr \left(x + (-1)^0 \frac{2}{2} \frac{2}{b^5} \right) =$$

$$tr((0.11011) \cdot 2 + (0.00001) \cdot 2) = tr((0.11100) \cdot 2) = (0.1110) \cdot 2 = 1.75$$

Cuando tenemos $x \in \mathbb{F}(b, t, L, M) \Rightarrow b^{L-1} \leq |x| \leq b^U(1 - b^{-t})$, truncar o redondear x :

$$b^{L-1} \leq |x| \leq b^U(1 - b^{-t})$$

- Operación en $\mathbb{F}(b, t, L, U)$: valor absoluto excede la cota superior **overflow**, se interrumpe el proceso computacional en curso.
- Operación en $\mathbb{F}(b, t, L, U)$: valor absoluto menor que la cota inferior **underflow**, menos drástico, suele reemplazarse por 0.

Los errores de redondeo o truncatura están acotados, se muestran en la siguiente propiedad.

Proposición 2.3. Consideremos el sistema de punto flotante $\mathbb{F}(b, t, L, U)$, con $L \leq e \leq U$ y sea $x = (-1)^s b^e \sum_{n=1}^{\infty} a_n b^{-n} \in \mathbb{R}$. Entonces:

$$(i) \quad |x - tr(x)| \leq b^{e-t}$$

$$(ii) \quad \frac{|x - tr(x)|}{|x|} \leq \varepsilon_M$$

$$(iii) \quad |x - rd(x)| \leq \frac{1}{2} b^{e-t}$$

$$(iv) \frac{|x - rd(x)|}{|x|} \leq \frac{\varepsilon_M}{2}$$

Demostración 2.2.

(i)

$$|x - tr(x)| = b^e \left| \sum_{n=t+1}^{\infty} a_n b^{-n} \right| \leq b^e (b-1) \sum_{n=t+1}^{\infty} b^{-n} = b^{e-t}$$

La primera igualdad se da, ya que al restar x y $tr(x)$ el factor común es b^e y se eliminan los términos del 1 hasta la t . En la segunda desigualdad hemos acotado por la cifra más grande. Finalmente, hemos calculado la suma.

(ii) Usando (i) y que $a_1 \geq 1$ tenemos que

$$\frac{|x - tr(x)|}{|x|} \leq \frac{b^{e-t}}{b^e \sum_{n=1}^{\infty} a_n b^{-n}} \leq \frac{b^{e-t}}{b^e \frac{1}{b}} = b^{1-t} = \varepsilon_M$$

(iii)

$$|x - rd(x)| = b^e |(0.a_1 \dots a_t a_{t+1} \dots) - (0.a_1 \dots r_t)|$$

$$\leq | (0.a_1 \dots a_t a_{t+1} \dots) - (0.a_1 \dots r_t) | \leq \frac{b^{-t}}{2} ?$$

- Si $a_{t+1} < \frac{b}{2}$ y $a_t = r_t$ [Relación ejercicios]:

$$|(0.a_1 \dots a_t a_{t+1} \dots) - (0.a_1 \dots r_t)| = (0.0 \dots 0 a_{t+1} \dots) \leq (0.0 \dots 0 \frac{b}{2}) = \frac{b^{-t}}{2}$$

- Si $a_{t+1} \geq \frac{b}{2}$ y $a_t + 1 = r_t$:

$$|(0.a_1 \dots a_t a_{t+1} \dots) - (0.a_1 \dots r_t)| = \frac{1}{b^t} - \left(\frac{a_{t+1}}{b^{t+1}} + \dots \right) \leq \frac{1}{b^t} - \frac{b}{2b^{t+1}} = \frac{b^{-t}}{2}$$

(iv)

$$a_1 \geq 1 \Rightarrow |x| \geq b^e b^{-1} \Rightarrow_{(iii)} \frac{|x - rd(x)|}{|x|} \leq \frac{\frac{1}{2} b^{e-t}}{b^e b^{-1}} = \frac{1}{2} b^{1-t} = \frac{\varepsilon_M}{2}$$

□

Definición 2.8 (Precisión máquina). La cota que aparece en el error relativo del redondeo recibe el nombre de **precisión máquina** (o **unidad de redondeo**) y se denota por u , es decir,

$$u := \frac{1}{2} b^{1-t} = \frac{1}{2} \varepsilon_M$$

Corolario 2.1. Dados $\mathbb{F}(b, t, L, U)$ y $x \in \mathbb{R}$, con $b^{L-1} \leq |x| \leq b^U(1 - b^{-t})$, se tiene que

$$rd(x) = (1 + \mu)x$$

para cierto $\mu \in \mathbb{R}$ tal que $|\mu| \leq u$

Demostración 2.3.

$$|x - rd(x)| \leq u|x| \Leftrightarrow x - |x|u \leq rd(x) \leq x + |x|u$$

esto es,

$$rd(x) = x + \kappa|x|u$$

con $|\kappa| \leq 1$, es decir,

$$rd(x) = (1 + \mu)x$$

para cierto $\mu \in \mathbb{R}$ tal que $|\mu| \leq u$

□

Las operaciones con números máquina en $\mathbb{F}(b, t, L, U)$ no son necesariamente internas.

Definición 2.9. Sea $\bullet : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ una operación, definimos la **operación máquina** como

$$\bullet_M : \mathbb{F}(b, t, L, U) \times \mathbb{F}(b, t, L, U) \rightarrow \mathbb{F}(b, t, L, U)$$

$$\bullet_M(x, y) := rd(x \bullet y), \quad (x, y \in \mathbb{F}(b, t, L, U))$$

Corolario 2.2.

$$x, y \in \mathbb{F}(b, t, L, U) \Rightarrow \bullet_M(x, y) = (1 + \mu)x \bullet y$$

para cierto $\mu \in \mathbb{R}$ tal que $|\mu| \leq u$

Aunque una única operación genere un error pequeño, una sucesión finita de operaciones es susceptible de producir la llamada **propagación del error**, que puede ser considerable. Veámoslo en el siguiente ejemplo.

Ejemplo 2.8. Sea la sucesión $x_n := \int_0^1 x^n e^x dx$, vamos a integrar por partes para obtener una

expresión más sencilla.

$$\int_0^1 x^n e^x dx = x^n e^x \Big|_0^1 - n \int_0^1 x^{n-1} e^x dx = e - n \int_0^1 x^{n-1} e^x dx$$

Por lo que $x_n = e - nx_{n-1}$.

Si $x \in [0, 1] \Rightarrow x^n e^x \leq x^{n-1} e^x$.

Tenemos que la sucesión $\{x_n\}_{n \geq 0}$ es decreciente y positiva, luego converge a un número ℓ .

Calculamos ese número:

$$x_{n-1} = \frac{e - x_n}{n} \Rightarrow \ell = 0$$

Redondeando en el sistema de punto flotante $b = 10$ y $t = 7$ obtenemos que:

$$x_{15} = (0.1604004)$$

$$x_0 = (0.1718281828) \cdot 10$$

$$x_{15} = (0.6004419078) \cdot 10^3$$

Lo que produce una propagación del error.

Ahora vamos a calcular el condicionamiento de una sucesión de funciones directamente relacionadas con la sucesión $\{x_n\}_{n \geq 0}$. Primero tenemos que calcular usando la recurrencia x_n , siendo $x_n = f_n(x_0)$ donde $f_n : \mathbb{R} \rightarrow \mathbb{R}$

$$x_1 = e - x_0$$

$$x_2 = e - 2(e - x_0) = -e + 2x_0$$

$$x_3 = e - 3(-e + 2x_0) = 4e - 6x_0$$

Por inducción, para $n \geq 1$:

$$(-1)^n (n!x_0 - \alpha_n e)$$

con $\alpha_n \in \mathbb{N}$ independiente de x_0 (es un valor irrelevante para calcular el condicionamiento).

Tenemos que:

$$x_n = f_n(x_0)$$

$$f_n(x) = (-1)^n (n!x - \alpha_n e), \quad (x \in \mathbb{R})$$

Ahora ya podemos calcular el condicionamiento de f_n en x_0 :

$$c(f_n, x_0) = \frac{n!x_0}{x_n} \geq \frac{n!x_0}{x_0} = n!$$

A continuación vamos a estudiar las operaciones aritméticas elementales y los errores debidos que se producen en estas debidos a truncaturas, redondeos o a cualquier otra circunstancia.

Proposición 2.4. Sea x el dato, μ_x el error relativo para x y $(1 + \mu_x)x$ su valor aproximado.

(i) Suma (igual para la resta).

Sean los datos $x, y \in \mathbb{R}$ tales que $x + y \neq 0$ y sus valores aproximados $(1 + \mu_x)x, (1 + \mu_y)y$, entonces el error cometido al realizar la operación suma es:

$$(1 + \mu_x)x + (1 + \mu_y)y = x + y + \mu_x x + \mu_y y = (x + y) \left(1 + \frac{\mu_x x + \mu_y y}{x + y} \right) = (x + y) \left(1 + \frac{x}{x + y} \mu_x + \frac{y}{x + y} \mu_y \right)$$

Luego

$$\mu_{x+y} = \frac{x}{x+y} \mu_x + \frac{y}{x+y} \mu_y$$

Además, si x e y tienen el mismo signo se puede controlar el error relativo:

$$|\mu_{x+y}| \leq |\mu_x| + |\mu_y|$$

Aunque si x e y tienen signos opuestos, μ_{x+y} puede dispararse si $x + y \approx 0$, generándose un error relativo enorme, conocido como **error de cancelación**.

(ii) Multiplicación. [No la va a pedir porque usa la fórmula de Taylor]

Los errores relativos de x e y son pequeños.

$$(1 + \mu_x)x \cdot (1 + \mu_y)y = (1 + \mu_x + \mu_y + \mu_x \mu_y)xy \approx (1 + \mu_x + \mu_y)xy$$

Luego

$$\mu_{xy} \approx \mu_x + \mu_y$$

(iii) División.

Los errores relativos de x e y son pequeños, siendo $y \neq 0$.

$$\frac{(1 + \mu_x)x}{(1 + \mu_y)y} = \frac{x}{y} (1 + \mu_x)(1 - \mu_y + \mu_y^2 - \mu_y^3 + \dots) \approx \frac{x}{y} (1 + \mu_x - \mu_y)$$

Por lo que

$$\mu_{x/y} \approx \mu_x - \mu_y$$

Ejemplo 2.9. (i) Sistema de punto flotante $b = 10, t = 7$.

$$\sqrt{30 + 10^{-5}} - \sqrt{30} = (0.5477226) - (0.5477226) = 0$$

Cociente (buen comportamiento, siguiente ejemplo) y una suma

$$\frac{10^{-5}}{\sqrt{30 + 10^{-5}} + \sqrt{30}} = (0.9128710) \cdot 10^{-6}$$

Redondeo del valor exacto $(0.9128709) \cdot 10^{-6}$

II | Tema 2. Resolución numérica de sistemas de ecuaciones lineales

En este tema vamos a hacer un tratamiento numérico de los sistemas de ecuaciones lineales. Vamos a usar dos métodos:

- (i) Métodos directos: Gauss, versiones y factorizaciones.
- (ii) Métodos iterativos: Jacobi y Gauss-Seidel.

1 Métodos directos: Gauss y versiones, factorización de matrices

Si tenemos un sistema $N \times N$ unisolvente, con N grande, la regla de Cramer es ineficiente, ya que tenemos un gran número de operaciones elementales. La regla de Cramer hace $N + 1$ determinantes + N divisiones, lo cual lleva a

$$(N + 1) \cdot (N! \cdot N - 1) + N = (N + 1) \cdot N! - 1 \text{ operaciones}$$

1.1 Sistemas triangulares

Sean la matriz $U \in \mathbb{R}^{N \times N}$ **triangular superior** con elementos diagonales no nulos, el vector de incógnitas $x \in \mathbb{R}^N$ y el vector de términos independientes b , tenemos el siguiente sistema triangular: $Ux = b$. Este sistema se resuelve por **sustitución hacia atrás** y el algoritmo para resolverlo es:

$$x_i = \frac{1}{u_{ii}} \left(b_i - \sum_{j=i+1}^N u_{ij} x_j \right), \text{ con } i = N, \dots, 1$$

Sea la matriz $L \in \mathbb{R}^{N \times N}$ **triangular inferior** con elementos diagonales no nulos, el vector de incógnitas $x \in \mathbb{R}^N$ y el vector de términos independientes b , tenemos el siguiente sistema triangular: $Lx = b$. Este sistema se resuelve por **sustitución hacia adelante** y el algoritmo para resolverlo es:

$$x_i = \frac{1}{l_{ii}} \left(b_i - \sum_{j=1}^{i-1} l_{ij} x_j \right), \text{ con } i = 1, \dots, N$$

1.2 Métodos de Gauss y Gauss-Jordan. Pivotaje

(i) Método de Gauss

Si tenemos un sistema $Ax = b$ unisolvente, con el método de Gauss obtenemos otro sistema

$Ux = c$ equivalente con U triangular superior, que ya hemos visto como se resuelve.

Los datos son $N \geq 1, A \in \mathbb{R}^{N \times N}, b \in \mathbb{R}^N$.

Sea $A^{(1)} := A$. Suponemos que $a_{kk}^{(k)} \neq 0$ con $k = 1, \dots, N$ (en caso contrario hemos terminado y no es posible llegar a un sistema triangular equivalente). Definimos recursivamente los multiplicadores:

$$m_{ik} := \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \text{ con } i = k+1, \dots, N$$

$$a_{ij}^{(k+1)} := a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}, \text{ con } i = k+1, \dots, N; j = k, \dots, N$$

$$b_i^{(k+1)} := b_i^{(k)} - m_{ik}a_{kj}^{(k)}, \text{ con } i = k+1, \dots, N$$

Por lo que obtenemos un sistema triangular superior equivalente que se resuelve por sustitución hacia atrás:

$$Ux = c$$

donde $U := A^{(N)}, c := b^{(N)}$.

$$A^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & \dots & \dots & \dots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & \dots & \dots & \dots & a_{2N}^{(2)} \\ \vdots & & \ddots & & & & \vdots \\ 0 & \dots & \dots & 0 & a_{kk}^{(k)} & \dots & a_{kN}^{(k)} \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \dots & \dots & 0 & a_{Nk}^{(k)} & \dots & a_{NN}^{(k)} \end{bmatrix}$$

Ejemplo 1.1. Sean

$$A := \begin{bmatrix} 1 & 1 & 3 \\ 0.1 & 1 & 1 \\ 1 & 2 & 0 \end{bmatrix}, \quad b := \begin{bmatrix} 5 \\ 2.1 \\ 3 \end{bmatrix}$$

Entonces vamos a calcular el sistema triangular superior equivalente. Tenemos que

$$A^{(1)} = A, \quad b^{(1)} = b$$

$$A^{(2)} = \begin{bmatrix} 1 & 1 & 3 \\ 0 & 0.9 & 0.7 \\ 0 & 1 & -3 \end{bmatrix}, \quad b^{(2)} = \begin{bmatrix} 5 \\ 1.6 \\ -2 \end{bmatrix}$$

$$A^{(3)} = \begin{bmatrix} 1 & 1 & 3 \\ 0 & 0.9 & 0.7 \\ 0 & 0 & -3.7 \end{bmatrix}, \quad b^{(3)} = \begin{bmatrix} 5 \\ 1.6 \\ -3.7 \end{bmatrix}$$

Ahora resolvemos por sustitución hacia atrás y tenemos que

$$x_1 = x_2 = x_3 = 1$$

Proposición 1.1. Sean $A \in \mathbb{R}^{N \times N}$ una matriz cuadrada y $b \in \mathbb{R}^N$, entonces son equivalentes:

- El correspondiente método de Gauss puede completarse hasta el paso N -ésimo.
- Para cada $k = 1, \dots, N$ la k -ésima submatriz principal de A

$$A_k = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix}$$

es regular.

Demostración 1.1. ¿(i) \Rightarrow (ii) ?

$$\det(A_k) = \det \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{kk}^{(k)} \end{bmatrix} = a_{11}^{(1)} \cdots a_{kk}^{(k)} \Rightarrow A_k \text{ REGULAR, con } k = 1, \dots, N$$

¿(ii) \Rightarrow (i) ?

Supongamos que no podemos completar el método de Gauss, sea $k := \min \{l \in \{1, \dots, N\} : a_{ll}^{(l)} = 0\}$

Como A_1 es regular, entonces $k \geq 2$. Hacemos el método de Gauss hasta $A^{(k)}$:

$$A^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} & a_{1k+1}^{(1)} & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} & a_{2k+1}^{(2)} & \cdots & a_{2N}^{(2)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & a_{k+1k}^{(k)} & \cdots & a_{kN}^{(k)} \\ 0 & 0 & \cdots & a_{k+1k}^{(k)} & a_{k+1k+1}^{(k)} & \cdots & a_{k+1N}^{(k)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{Nk}^{(k)} & a_{Nk+1}^{(k)} & \cdots & a_{NN}^{(k)} \end{bmatrix}$$

$$\det(A_k) = \det(A^{(k)})_k = 0$$

□

Los errores de redondeo afectan al método de Gauss. Además, evitaremos dividir por coeficientes relativamente pequeños y también evitaremos que algún $a_{kk}^{(k)}$ del método de Gauss sea nulo.

Nota. Este metodillo de Gauss es una mierda.

(ii) **Método de Gauss con pivotaje** (o pivotaje parcial), variante adaptativa de Gauss.

En este método se modifica el paso k , por lo que antes de definir los multiplicadores m_{ik} , la matriz $A^{(k+1)}$ y el vector $b^{(k+1)}$, intercambiaremos de posición si es necesario dos de las filas k, \dots, N de la matriz $A^{(k)}$ de forma que el elemento del vector que tiene mayor valor absoluto sea $a_{kk}^{(k)}$.

$$\begin{bmatrix} a_{kk}^{(k)} \\ \vdots \\ a_{Nk}^{(k)} \end{bmatrix} \rightsquigarrow \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & \dots & \dots & \dots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & \dots & \dots & \dots & a_{2N}^{(2)} \\ \vdots & & \ddots & & & & \vdots \\ 0 & \dots & \dots & 0 & a_{kk}^{(k)} & \dots & a_{kN}^{(k)} \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \dots & \dots & 0 & a_{Nk}^{(k)} & \dots & a_{NN}^{(k)} \end{bmatrix}$$

Finalmente, el sistema equivalente al de partida se resuelve por sustitución hacia atrás. El método de Gauss con pivotaje tiene sentido hasta el último paso N -ésimo \Leftrightarrow la matriz de coeficientes A es regular (el sistema es compatible determinado).

Nota. Otra variante adaptativa del método de Gauss, **pivotaje total o completo**, no solo reordena las filas k, \dots, N de $A^{(k)}$, sino también las columnas k, \dots, N de forma que el elemento de la submatriz de dicha matriz correspondiente a las mencionadas filas y columnas tenga a $a_{kk}^{(k)}$ como el elemento de mayor valor absoluto. Sin embargo, requiere un mayor número de operaciones.

(iii) **Método de Gauss-Jordan.**

Para terminar con las variantes del método de Gauss, mencionemos el llamado **método de Gauss-Jordan**, que consiste en hacer ceros no solo debajo de $a_{kk}^{(k)}$ sino también por encima, con el mismo tipo de fórmula. Sin embargo, su coste en operaciones aritméticas es superior.

1.3 Métodos de factorización

Los métodos de factorización se usan cuando tenemos sistemas con la misma matriz de coeficientes, para resolverlos más rápido.

Definición 1.1 (Factorización LU). Sea $Ax = b$ un sistema compatible determinado, L una matriz triangular inferior y U una matriz triangular superior, entonces:

$$A = LU$$

Como $A = LU$, entonces $LUx = b$. Por lo que para resolver este sistema, se siguen los siguientes pasos:

- (i) Calculamos las matrices L y U .
- (ii) Sea $y := Ux$, entonces tenemos que resolver y de $Ly = b$ por sustitución hacia delante.
- (iii) Por último, resolvemos x de $Ux = y$ por sustitución hacia atrás.

Ejemplo 1.2. Sean

$$A = \begin{bmatrix} 1 & 3 & -1 \\ 2 & 8 & 4 \\ -1 & 3 & 4 \end{bmatrix}, \quad b = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix}$$

Primero calculamos las matrices L y U . Como $A = LU$, entonces:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 3 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 3 & -1 \\ 0 & 2 & 6 \\ 0 & 0 & -15 \end{bmatrix}$$

Ahora tenemos que resolver el sistema $Ly = b$.

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix}$$

Lo resolvemos por sustitución hacia adelante y obtenemos que $y_1 = -1$, $y_2 = 4$, $y_3 = -13$.

Ahora resolvemos el sistema $Ux = y$.

$$\begin{bmatrix} 1 & 3 & -1 \\ 0 & 2 & 6 \\ 0 & 0 & -15 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \\ -13 \end{bmatrix}$$

Lo resolvemos por sustitución hacia atrás y obtenemos que $x_1 = \frac{5}{3}$, $x_2 = \frac{-3}{5}$, $x_3 = \frac{13}{15}$. Que es la solución del sistema inicial $Ax = b$.

Nota. Tenemos que tener en cuenta que no siempre podemos obtener una factorización tipo LU para una matriz regular. Solo se podrá si $a_{11} = l_{11}u_{11} \neq 0$ y L, U son regulares.

Proposición 1.2. Sean $N \geq 1$, $A \in \mathbb{R}^{N \times N}$, $b \in \mathbb{R}^N$ y supongamos que aplicando el método de Gauss al sistema $Ax = b$ se obtiene una matriz triangular superior $A^{(N)}$ y un vector $b^{(N)}$ de forma que el sistema $A^{(N)}x = b^{(N)}$ es equivalente al de partida. Entonces

$$A = LU,$$

siendo

$$U = A^{(N)}$$

$$L = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ m_{21} & 1 & 0 & \cdots & 0 \\ m_{31} & m_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{N1} & m_{N2} & m_{N3} & \cdots & 1 \end{bmatrix},$$

donde los coeficientes de la parte inferior de L son los multiplicadores del método de Gauss definidos recursivamente.

Demostración 1.2. Sea $k = 1, \dots, N-1$, usando la notación del método de Gauss:

$$A^{(k+1)} = E_k A^{(k)}$$

$$E_k := \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & -m_{k+1\ k} & 1 & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & \cdots & -m_{Nk} & 0 & \cdots & 1 \end{bmatrix}$$

$$e_k^T := [0, 0, \dots, 0, 1, 0, \dots, 0] \text{ donde el } 1 \text{ está en la posición } k$$

$$m_k^T := [0, \dots, 0, m_{k+1\ k}, \dots, m_{Nk}]$$

Sea I_N la matriz identidad de orden N , entonces

$$E_k = I_N - m_k e_k^T$$

Entonces E_k es regular y, además,

$$E_k^{-1} = I_N + m_k e_k^T$$

$A^{(k+1)} = E_k A^{(k)}$ y usando el método de Gauss hasta el paso N , tenemos que

$$E_{N-1} E_{N-2} \cdots E_2 E_1 A = U$$

Luego

$$\begin{aligned}
 A &= E_1^{-1} E_2^{-1} \cdots E_{N-2}^{-1} E_{N-1}^{-1} U = (I_N + m_1 e_1^T)(I_N + m_2 e_2^T) \cdots (I_N + m_{N-1} e_{N-2}^T)(I_N + m_{N-1} e_{N-1}^T) U = \\
 &= \left(I_N + \sum_{k=1}^{N-1} m_k e_k^T \right) U = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ m_{21} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ m_{N1} & m_{N2} & m_{N3} & \cdots & 1 \end{bmatrix} U
 \end{aligned}$$

□

Teorema 1.1. Sea $A \in \mathbb{R}^{N \times N}$ una matriz regular. Entonces equivalen

- (i) Para cualquier $b \in \mathbb{R}^N$, el método de Gauss para el correspondiente sistema de ecuaciones lineales puede completarse hasta el paso N -ésimo.
- (ii) A admite una factorización LU .
- (iii) Las N submatrices principales de A son regulares.

Teorema 1.2. Sea A una matriz regular, entonces:

- (i) El método de Gauss con pivotaje es factible, para cualquier sistema de ecuaciones lineales que tenga a A por matriz de coeficientes.
- (ii) Salvo la eventual permutación de algunas de sus filas, A admite una factorización LU .

Demostración 1.3. P matriz que se obtiene al aplicar a la identidad de orden N las mismas permutaciones de filas que a A .

$$Ax = b \iff PAx = Pb$$

Dado que $PA = LU$:

$$Ly = Pb \quad Ux = y$$

Definición 1.2 (Factorización de Doolittle). En el sistema $LUx = b$, los coeficientes de la diagonal de L son 1, es decir

$$l_{11} = \cdots = l_{NN} = 1$$

A la hora de programar la factorización de Doolittle, el algoritmo es:

Sea $i = 1, \dots, N$, entonces

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}, \quad \text{con } j = i, \dots, N$$

Y supuesto que $u_{ii} \neq 0$

$$l_{ji} = \frac{1}{u_{ii}} \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk} u_{ki} \right), \quad \text{con } j = i+1, \dots, N$$

Definición 1.3 (Factorización de Crout). En el sistema $LUx = b$, los coeficientes de la diagonal de U son 1, es decir

$$u_{11} = \dots = u_{NN} = 1$$

Ejemplo 1.3. Sea

$$A = \begin{bmatrix} 1 & -2 & 0 & 3 \\ -2 & 3 & 1 & -6 \\ -1 & 4 & -4 & 3 \\ 5 & -8 & 4 & 0 \end{bmatrix}$$

Vamos a usar la factorización de Crout.

$$\begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 0 & 3 \\ -2 & 3 & 1 & -6 \\ -1 & 4 & -4 & 3 \\ 5 & -8 & 4 & 0 \end{bmatrix}$$

En la fila 1 de A obtenemos que

$$l_{11} = 1$$

$$u_{12} = -2$$

$$u_{13} = 0$$

$$u_{14} = 3$$

En la fila 2 de A obtenemos que

$$l_{21} = -2$$

$$l_{21} u_{12} + l_{22} = 3 \Leftrightarrow l_{22} = -1$$

$$l_{21} u_{13} + l_{22} u_{23} = 1 \Leftrightarrow u_{23} = -1$$

$$l_{21} u_{14} + l_{22} u_{24} = -6 \Leftrightarrow u_{24} = 0$$

En la fila 3 de A obtenemos que

$$l_{31} = -1$$

$$l_{31}u_{12} + l_{32} = 4 \Leftrightarrow l_{32} = 2$$

$$l_{31}u_{13} + l_{32}u_{23} + l_{33} = -4 \Leftrightarrow l_{33} = -2$$

$$l_{31}u_{14} + l_{32}u_{24} + l_{33}u_{34} = 3 \Leftrightarrow u_{34} = -3$$

En la fila 4 de A obtenemos que

$$l_{41} = 5$$

$$l_{41}u_{12} + l_{42} = -8 \Leftrightarrow l_{42} = 2$$

$$l_{41}u_{13} + l_{42}u_{23} + l_{43} = 4 \Leftrightarrow l_{43} = 6$$

$$l_{41}u_{14} + l_{42}u_{24} + l_{43}u_{34} + l_{44} = 0 \Leftrightarrow l_{44} = 3$$

Por lo que

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & -1 & 0 & 0 \\ -1 & 2 & -2 & 0 \\ 5 & 2 & 6 & 3 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & -2 & 0 & 3 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Ejercicio 1.1. Comprueba aplicando el algoritmo anterior que la matriz regular

$$\begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ 0.2 & 0.9 & 1.2 & 1.5 \\ 0.3 & 1.6 & 2.9 & 3.5 \\ 0.4 & 2.3 & 4.6 & 6.5 \end{bmatrix}$$

admite una factorización tipo Doolittle.

Ejercicio 1.2. Diseña un algoritmo para determinar la factorización tipo Crout de una matriz regular que la admita. (Indicación: se puede aprovechar el algoritmo de Doolittle: si $A^T = LU$, entonces $A = U^T L^T$). Unicidad fijando unos en una de las diagonales

Ejercicio 1.3. Considera los sistemas de ecuaciones lineales con matriz de coeficientes

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 2 & 1 & 3 & 1 \\ 0 & 1 & 3 & 3 \\ 1 & 1 & 4 & 2 \end{bmatrix}$$

y términos independientes: $[1333]^t, [1465]^t, [1445]^t$.

Resuélvelos por el método más eficiente.

Definición 1.4 (Matriz definida positiva). Sea $A \in \mathbb{R}^{N \times N}$, se dice que es **definida positiva** si

$$x^T A x > 0, \quad \forall x \in \mathbb{R}^N \setminus \{0\}$$

La factorización LU en matrices simétricas definidas positivas es válida siempre. Además, $L = U^T$.

Definición 1.5 (Factorización tipo Cholesky). Sea $A \in \mathbb{R}^{N \times N}$ una matriz simétrica y definida positiva. Entonces existe una única matriz triangular superior $U \in \mathbb{R}^{N \times N}$ con coeficientes positivos en su diagonal principal y de forma que

$$A = U^T U$$

A la hora de programar la factorización tipo Cholesky, el algoritmo es:

Para todo $j = 1, \dots, N$

$$i = 1, \dots, j-1 \Rightarrow u_{ij} = \frac{1}{u_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} u_{ki} u_{kj} \right)$$

$$u_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} u_{kj}^2}$$

Además, se eliminan los elementos de debajo de la diagonal principal, ya que hay simetría.

Proposición 1.3. Una matriz cuadrada A es simétrica y definida positiva si, y solo si, admite una factorización tipo Cholesky.

Ejercicio 1.4. Demuestra que una matriz cuadrada A es simétrica y definida positiva si, y solo si, admite una factorización tipo Cholesky.

Ejercicio 1.5. Comprueba que la matriz

$$A = \begin{bmatrix} 4 & 2 & -2 \\ 2 & 2 & -3 \\ -2 & -3 & 14 \end{bmatrix}$$

es definida positiva.

Ejercicio 1.6. Resuelve el sistema lineal $Ax = \begin{bmatrix} 4 \\ 0 \\ 2 \end{bmatrix}$ por el método de Cholesky.

Ejercicio 1.7. Encuentra una matriz cuadrada de orden 3×3 que sea simétrica pero no definida positiva.

2 Métodos iterativos: métodos de Jacobi y Gauss-Seidel

Los métodos iterativos nos dan la solución de un sistema de ecuaciones lineales cuadrado y compatible determinado como límite de una sucesión. Cada término de la sucesión se genera de forma recursiva a partir del anterior, los cuales reciben el nombre de iteradores. Además, cuando tenemos sistemas de grandes dimensiones y la matriz de coeficientes es dispersa (número de coeficientes no nulos relativamente pequeño), tenemos problemas prácticos: análisis matricial de estructuras, método de elementos finitos...

2.1 Métodos iterativos: convergencia y consistencia

Sea la matriz $A \in \mathbb{R}^{N \times N}$ regular y un vector $b \in \mathbb{R}^N$, entonces tenemos un sistema de ecuaciones lineales cuadrado y unisolviente $Ax = b$.

Definición 2.1 (Método iterativo). Sean $B \in \mathbb{R}^{N \times N}$, $x_0, c \in \mathbb{R}^N$, entonces se define al **método iterativo** como:

$$\begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow x_n = Bx_{n-1} + c \end{cases}$$

Definición 2.2 (Consistencia). Sea x la solución del sistema, entonces

$$\lim_{n \rightarrow \infty} x_n = x$$

$$\Downarrow (u \in \mathbb{R}^N \mapsto Bu + c \in \mathbb{R}^N \text{ continua})$$

$$x = Bx + c$$

Entonces se dice que hay **consistencia** del método con el sistema.

Proposición 2.1. La consistencia del método iterativo con el sistema equivale a

$$c = (I - B)A^{-1}b$$

Proposición 2.2. Si el método iterativo converge a la solución del sistema, entonces el método es consistente con el sistema.

El recíproco es falso. Veámoslo en el siguiente ejemplo.

Ejemplo 2.1. Sea $I \in \mathbb{R}^{N \times N}$ la matriz identidad de orden N y sea $b \in \mathbb{R}^N$. Dados el sistema de ecuaciones lineales y el método iterativo

$$2Ix = by \quad \begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow x_n = -x_{n-1} + c, \end{cases}$$

Entonces el método es consistente con el sistema si, y solo si,

$$c = b$$

y converge a la solución del sistema cuando, y solo cuando,

$$c = 2x_0$$

El pseudorrecíproco se da en la siguiente proposición.

Proposición 2.3. Supongamos que $N \geq 1$, $A, B \in \mathbb{R}^{N \times N}$ con A regular, $x_0, b, c \in \mathbb{R}^N$ y que el método iterativo

$$\begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow x_n = Bx_{n-1} + c \end{cases}$$

es consistente con el sistema unisolvente $Ax = b$. Entonces método iterativo converge a la solución del sistema cualquiera sea $x_0 \in \mathbb{R}^N \Leftrightarrow \rho(B) < 1$.

Demostración 2.1. Demostremos primero la consistencia. Sea $n \geq 1$, entonces

$$x_n - x = Bx_{n-1} + c - x = Bx_{n-1} + (I - B)x - x = B(x_{n-1} - x)$$

Resursivamente obtenemos que $x_n - x = B^n(x_0 - x)$. (2)

$\hookrightarrow \Rightarrow ?$

Basándonos en la relación de recurrencia (2) y la convergencia para todo $x_0 \in \mathbb{R}^N$, tenemos que

$$\lim_{n \rightarrow \infty} B^n(x_0 - x) = 0 \Leftrightarrow \lim_{n \rightarrow \infty} B^n = 0 \Leftrightarrow \rho(B) < 1$$

$\Leftarrow ?$

Sea $x_0 \in \mathbb{R}^N$, como $\rho(B) < 1 \Rightarrow \lim_{n \rightarrow \infty} B^n = 0$.

Usando la relación de recurrencia (2), $\|\cdot\|$ en \mathbb{R}^N y su matricial inducida en $\mathbb{R}^{N \times N}$, notada de la misma forma, y con $n \geq 1$, entonces

$$\|x_n - x\| = \|B^n(x_0 - x)\| \leq \|B^n\| \|x_0 - x\|$$

Por lo que $\lim_{n \rightarrow \infty} x_n = x$

□

Nota. Como hemos dicho, el método iterativo converge para todo $x_0 \in \mathbb{R}^N$. Aunque esto es falso si el método iterativo es consistente con el sistema y no converge para todo $x_0 \in \mathbb{R}^N$. Por ejemplo, el sistema del ejemplo anterior.

$b := 0 =: c \Rightarrow$ consistencia

Además, hay convergencia solo cuando $c = 2x_0 \Leftrightarrow x_0 = 0$ y $\rho(B) = \rho(-I) = 1$.

2.2 Generación de métodos iterativos

La proposición anterior nos orienta al procedimiento del diseño de métodos iterativos. Por lo que son automáticamente consistentes, ya que elimina el grave problema que surge al comprobar dicha condición: hay que conocer a priori la solución del sistema... ¡que es justo lo que se pretende aproximar!

Sea la matriz $A \in \mathbb{R}^{N \times N}$ regular y el vector $b \in \mathbb{R}^N$, entonces tenemos un sistema unisolviente $Ax = b$. Sea $A = M - N$, por lo que M es regular (siempre es posible por ser A regular) y con N no nula. Tenemos que

$$Ax = b \Leftrightarrow (M - N)x = b \Leftrightarrow x = M^{-1}Nx + M^{-1}b$$

Por lo que sugiere que $B = M^{-1}N$ y $c = M^{-1}b$, tenemos el siguiente método iterativo

$$\begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow x_n = M^{-1}Nx_{n-1} + M^{-1}b \end{cases}$$

que es consistente con el sistema

$$(I - B)A^{-1}b = (I - M^{-1}N)A^{-1}b = (M^{-1}M - M^{-1}N)A^{-1}b = M^{-1}(M - N)A^{-1}b = M^{-1}b = c$$

Corolario 2.1. Sean $A, M, N \in \mathbb{R}^{N \times N}$ con A y M regulares de forma que $A = M - N$ y sean $b, x_0 \in \mathbb{R}^N$. Consideremos el sistema

$$Ax = b$$

y el método iterativo

$$\begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow x_n = M^{-1}Nx_{n-1} + M^{-1}b \end{cases}$$

Entonces el método iterativo converge a la solución del sistema, cualquiera sea $x_0 \in \mathbb{R}^N \Leftrightarrow \rho(M^{-1}N) < 1$.

En resumen, todo método iterativo convergente es de esta forma:

Proposición 2.4. Sean $A, B \in \mathbb{R}^{N \times N}$, con A regular, sean $b, c \in \mathbb{R}^N$ y consideremos el sistema de ecuaciones lineales

$$Ax = b$$

y el método iterativo

$$\begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow x_n = Bx_{n-1} + c \end{cases}$$

que supondremos que converge hacia la solución del sistema para cualquier estimación inicial $x_0 \in \mathbb{R}^N$. Entonces existe una descomposición de la matriz de coeficientes

$$A = M - N$$

con $M, N \in \mathbb{R}^{N \times N}$ y M regular, tales que

$$B = M^{-1}N \quad c = M^{-1}b$$

Demostración 2.2. Convergencia del método para todo $x_0 \in \mathbb{R}^N \Rightarrow$ consistencia (bastaría para uno solo).

$$c = (I - B)A^{-1}b$$

Pretendemos que c sea $M^{-1}b = (I - B)A^{-1}b$. Esto puede conseguirse si $M^{-1} = (I - B)A^{-1}$. Como $(I - B)$ es regular por ser $\rho(B) < 1$:

$$M := A(I - B)^{-1}$$

Esta elección debe ser N con $M^{-1}N = B$.

$$N := A(I - B)^{-1}B$$

Claramente

$$A = M - N$$

□

Por lo que tenemos el siguiente método iterativo

$$\begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow x_n = M^{-1}Nx_{n-1} + M^{-1}b \end{cases}$$

con M regular y $A = M - N$.

Aunque es un poco complejo calcular la inversa de M y luego multiplicarla por otra matriz, por lo que trabajaremos sobre otro método iterativo equivalente, que consiste en multiplicar por M a ambos lados de la igualdad:

$$\begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow Mx_n = Nx_{n-1} + b \end{cases}$$

Para que converja, tenemos que calcular igualmente si $\rho(M^{-1}N) < 1$. Además, la iteración x_n es la solución del sistema resoluble sin un alto coste operativo, ya que M es triangular.

$$Mx_n = Nx_{n-1} + b$$

Ahora vamos a poner en práctica estos conceptos. Los métodos iterativos más populares son: **Jacobi** y **Gauss-Seidel**. Definamos primero algunas matrices a partir de la matriz de coeficientes A .

Sea $A \in \mathbb{R}^{N \times N}$ regular, definimos las matrices diagonales y triangulares:

$$D := \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{NN} \end{bmatrix}$$

$$E := \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ -a_{21} & 0 & 0 & \cdots & 0 \\ -a_{31} & -a_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -a_{N1} & -a_{N2} & -a_{N3} & \cdots & 0 \end{bmatrix}$$

$$F := \begin{bmatrix} 0 & \cdots & -a_{12} & -a_{1 \ N-1} & -a_{1N} \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & -a_{N-2 \ N-1} & -a_{N-2 \ N} \\ 0 & \cdots & 0 & 0 & -a_{N-1 \ N} \\ 0 & \cdots & 0 & 0 & 0 \end{bmatrix}$$

Observemos que la matriz A verifica la hipótesis adicional: $a_{11}a_{22}\dots a_{NN} \neq 0$.

Bajo esta notación definimos los siguientes métodos iterativos.

Definición 2.3 (Método de Jacobi).

$$A = M - N \text{ con } M := D \text{ y } N := E + F$$

Por lo que tenemos que

$$\begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow Dx_n = (E + F)x_{n-1} + b \end{cases}$$

Expresado en coordenadas:

$$x_0 = [x_{01}, \dots, x_{0N}]^T$$

$$i = 1, \dots, N \Rightarrow x_{ni} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1 \wedge j \neq i}^N a_{ij} x_{n-1j} \right)$$

Definición 2.4 (Método de Gauss-Seidel).

$$A = M - N \text{ con } M := D - E \text{ y } N := F$$

Por lo que tenemos que

$$\begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow (D - E)x_n = Fx_{n-1} + b \end{cases}$$

Expresado en coordenadas:

$$x_0 = [x_{01}, \dots, x_{0N}]^T$$

$$i = 1, \dots, N \Rightarrow x_{ni} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_{nj} - \sum_{j=i+1}^N a_{ij} x_{n-1j} \right)$$

- *Similitud:* Descritos a partir del esquema obtenido al despejar en $Ax = b$ x_1 de la primera ecuación, x_2 de la segunda y así hasta x_N de la N -ésima

$$\begin{cases} x_1 = \frac{1}{a_{11}}(b_1 - a_{12}x_2 - a_{13}x_3 - \dots - a_{1N}x_N) \\ x_2 = \frac{1}{a_{22}}(b_2 - a_{21}x_1 - a_{23}x_3 - \dots - a_{2N}x_N) \\ \vdots \\ x_N = \frac{1}{a_{NN}}(b_N - a_{N1}x_1 - a_{N2}x_2 - \dots - a_{NN-1}x_{N-1}) \end{cases}$$

En el método de Jacobi: a la derecha las coordenadas de una iteración para obtener a la izquierda la siguiente.

Gauss-Seidel: a la derecha las coordenadas de una iteración y las que se acaban de hallar de la siguiente más arriba para determinar las de la siguiente a la izquierda.

- *Diferencia:* En Jacobi el vector de cada iteración se calcula a partir del anterior, y en cambio, en Gauss-Seidel, el vector en cada iteración usa las coordenadas que ya se han calculado en la iteración actual.

Ahora nos preguntamos: ¿es el método de Gauss-Seidel más eficiente que el método de Jacobi? Aunque esto no se cumple siempre, veámoslo en un ejemplo.

Ejemplo 2.2. Tenemos el siguiente sistema de ecuaciones

$$\begin{bmatrix} 2 & 1 & 3 \\ -1 & 3 & 2 \\ 1 & 4 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ -1 \\ 11 \end{bmatrix}$$

Por lo que su solución es $x_1 = 1$, $x_2 = -2$, $x_3 = 3$.

Usemos los métodos de Jacobi y Gauss-Seidel para ver si convergen a esta solución. Supongamos una estimación inicial $x_0 = [0, 0, 0]^T$

Despejamos x_i de la ecuación i -ésima ($i = 1, 2, 3$) y obtenemos que

$$\begin{cases} x_1 = 4.5 - 0.5x_2 - 1.5x_3 \\ x_2 = -\frac{1}{3} + \frac{1}{3}x_1 - \frac{2}{3}x_3 \\ x_3 = \frac{11}{6} - \frac{1}{6}x_1 - \frac{3}{2}x_2 \end{cases}$$

Eso se hace para ambos métodos, pero en el siguiente paso ya distinguimos cada método.

- Jacobi:

el algoritmo parte de x_0 y para cada $n \geq 1$

$$\begin{cases} x_{n1} = 4.5 - 0.5x_{n-1,2} - 1.5x_{n-1,3} \\ x_{n2} = -\frac{1}{3} + \frac{1}{3}x_{n-1,1} - \frac{2}{3}x_{n-1,3} \\ x_{n3} = \frac{11}{6} - \frac{1}{6}x_{n-1,1} - \frac{3}{2}x_{n-1,2} \end{cases}$$

obteniendo (truncando):

$$\begin{cases} x_{01} = 0 & x_{02} = 0 & x_{03} = 0 \\ x_{11} = 4.5 & x_{12} = -0.333 & x_{13} = 1.833 \\ \dots & \dots & \dots \\ x_{20,1} = 1.308 & x_{20,2} = -1.67 & x_{20,3} = 2.702 \end{cases}$$

- Gauss-Seidel:

con la estimación inicial x_0 , para todo $n \geq 1$ tenemos que

$$\begin{cases} x_{n1} = 4.5 - 0.5x_{n-1,2} - 1.5x_{n-1,3} \\ x_{n2} = -\frac{1}{3} + \frac{1}{3}x_{n1} - \frac{2}{3}x_{n-1,3} \\ x_{n3} = \frac{11}{6} - \frac{1}{6}x_{n1} - \frac{3}{2}x_{n2} \end{cases}$$

con lo que se generan (truncando) los datos numéricos

$$\begin{cases} x_{01} = 0 & x_{02} = 0 & x_{03} = 0 \\ x_{11} = 4.5 & x_{12} = 1.166 & x_{13} = 0.305 \\ \dots & \dots & \dots \\ x_{20\ 1} = 1.035 & x_{20\ 2} = -1.961 & x_{20\ 3} = 2.968 \end{cases}$$

Por lo que en este sistema Gauss-Seidel es más eficiente, pero no en general.

Hay que tener en cuenta que la velocidad de convergencia de Jacobi es independiente de la de Gauss-Seidel, y viceversa:

Proposición 2.5. Medida de la velocidad de convergencia.

$$Ax = b, \text{ sistema unisolvente}$$

Método iterativo convergente a la solución del sistema para cualquier estimación inicial (no necesariamente Jacobi o Gauss-Seidel).

$$x_0 \text{ dado}$$

$$n \geq 1 \Rightarrow x_n = Bx_{n-1} + c$$

$$\rho(B) < 1 \text{ consistencia}$$

$$(I - B)x = c$$

compatible determinado y con la misma solución que $Ax=b$.

Demostración de la última proposición: norma y su matricial inducida, $\|B\| < 1, n \geq 1$

$$\|x_n - x\| \leq \|B\|^n \|x_0 - x\|$$

Cuanto menor sea $\|B\|$ mejor será la convergencia de la sucesión de iteradores hacia la solución del sistema.

Puede probarse que si $A \in \mathbf{R}^{N \times N}$, entonces

$$\rho(A) = \inf_{\|\cdot\|} \|A\| : \|\cdot\| \text{ es una norma matricial inducida en } \mathbf{R}^{N \times N}$$

Aunque A es real, el resultado es complejo. Es esperable que cuanto menor sea $\rho(B)$ mejor será la convergencia de la sucesión de iteradores hacia la solución del sistema. En los sistemas en los que la relación de orden entre los radios espectrales de la matriz $M^{-1}N$ para el método de Jacobi y Gauss-Seidel van en distintos sentidos (siguiente ejemplo).

Estudiemos cada una.

Proposición 2.6. Cuando menor sea $\rho(M^{-1}N)$, mejor será la convergencia de la sucesión de iteradores hacia la solución del sistema.

Ejemplo 2.3. Tenemos dos matrices

$$A_1 = \begin{bmatrix} 4 & 1 & 1 \\ 2 & -9 & 0 \\ 0 & -8 & -6 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 7 & 6 & 9 \\ 4 & 5 & -4 \\ -7 & -3 & 8 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 3 & 0 & 4 \\ 7 & 1 & 2 \\ -1 & 1 & 9 \end{bmatrix}$$

- A_1 :

$$\rho_{Jacobi} := \rho(D^{-1}(E + F)) = 0.444$$

$$\rho_{Gauss-Seidel} := \rho((D - E)^{-1}F) = 0.019$$

En este caso, Gauss-Seidel es más eficiente.

- A_2 :

$$\rho_{Jacobi} := \rho(D^{-1}(E + F)) = 0.641$$

$$\rho_{Gauss-Seidel} := \rho((D - E)^{-1}F) = 0.775$$

En este caso, Jacobi es más eficiente.

- A_3 :

$$\rho_{Jacobi} := \rho(D^{-1}(E + F)) = 1.037$$

$$\rho_{Gauss-Seidel} := \rho((D - E)^{-1}F) = 0.963$$

En este caso, Jacobi no converge, ya que $\rho_{Jacobi} \geq 1$. Solo converge Gauss-Seidel.

Ahora veamos que la convergencia para Gauss-Seidel es independiente de la de Jacobi.

Ejemplo 2.4. Tenemos dos sistemas:

$$\begin{bmatrix} 3 & 0 & 4 \\ 7 & 1 & 2 \\ -1 & 1 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 10 \\ 9 \end{bmatrix}, \quad \begin{bmatrix} -3 & 3 & -6 \\ -4 & 7 & -8 \\ 5 & 7 & -9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -6 \\ -5 \\ 3 \end{bmatrix}$$

En el primer sistema, tenemos que hay convergencia para Gauss-Seidel pero no para Jacobi:

$$\rho_{Jacobi} := \rho(D^{-1}(E + F)) = 1.037$$

$$\rho_{Gauss-Seidel} := \rho((D - E)^{-1}F) = 0.963$$

En el segundo sistema, tenemos que hay convergencia para Jacobi pero no para Gauss-Seidel:

$$\rho_{Jacobi} := \rho(D^{-1}(E + F)) = 0.813$$

$$\rho_{Gauss-Seidel} := \rho((D - E)^{-1}F) = 1.111$$

Definición 2.5 (Diagonalmente estrictamente dominante). Sea la matriz $A \in \mathbb{R}^{N \times N}$, diremos que A es diagonalmente estrictamente dominante si

$$i = 1, \dots, N \Rightarrow \sum_{j=1 \wedge j \neq i}^N |a_{ij}| < |a_{ii}|$$

Los métodos de Jacobi y de Gauss-Seidel convergen para cualquier sistema de ecuaciones lineales que tenga a A por matriz de coeficientes a su solución, cualquiera sea la elección de la estimación inicial.

Proposición 2.7. Sea $A \in \mathbb{R}^{N \times N}$ una matriz diagonalmente estrictamente dominante. Entonces los métodos de Jacobi y de Gauss-Seidel, para todo sistema de ecuaciones lineales que tenga como matriz de coeficientes A , convergen hacia su solución, independientemente de la estimación inicial que se fije.

El recíproco falso: basta considerar cualquier sistema en el que la matriz de coeficientes sea la matriz A_1 (o A_2) del penúltimo ejemplo.

Tenemos que tener en cuenta lo siguiente:

- Si se aplica al sistema de partida una transformación elemental tan simple como intercambiar de posición dos ecuaciones y se usa el mismo método iterativo con los dos sistemas, uno puede converger y otro no. La idea es que este tipo de transformación elemental no solo puede modificar claramente el hecho de que la matriz de coeficientes sea diagonalmente estrictamente dominante (que es una condición suficiente para la convergencia de Jacobi y Gauss-Seidel) sino que además puede cambiar el radio espectral.
- El número de operaciones que hay que realizar para pasar de una iteración a la siguiente en los métodos de Jacobi y Gauss-Seidel es de N^2 para un sistema de N ecuaciones y N incógnitas. Por tanto, si N es grande, requiere en principio menos operaciones que los directos. Además, y a diferencia de estos últimos, aprovecha la estructura de la matriz de coeficientes cuando es dispersa, tal y como ocurre con los sistemas que surgen en problemas de análisis de estructuras o de elementos finitos.
- La elección que hemos hecho para diseñar los métodos de Jacobi y Gauss-Seidel es la más popular, pero no la única. Una clase de métodos iterativos que los incluye y que son también de uso extendido está constituida por los llamados métodos de relajación para Jacobi y Gauss-Seidel, una especie de combinación convexa de ellos.

3 Análisis de error

Ahora vamos a estudiar el error cometido al resolver mediante un método numérico un sistema de ecuaciones lineales unisolviente y con igual número de ecuaciones e incógnitas; el error relativo cometido al resolver de forma aproximada un sistema; errores derivados del método usado, o los debidos al redondeo, propagación...; y el error relativo controlado en función únicamente del vector de términos independientes y de la solución aproximada.

Proposición 3.1. Sea $A \in \mathbb{R}^{N \times N}$ una matriz regular y $x, u, b \in \mathbb{R}^N$ con $\|x\|\|b\| \neq 0$ y de forma que $Ax = b$. Entonces para cualquier norma en \mathbb{R}^N se cumplen las desigualdades:

$$\frac{1}{c(A)} \frac{\|Au - b\|}{\|b\|} \leq \frac{\|x - u\|}{\|x\|} \leq c(A) \frac{\|Au - b\|}{\|b\|},$$

donde $c(A)$ es el condicionamiento de la matriz A relativo a la norma matricial inducida por $\|\cdot\|$.

Demostración 3.1. Primera desigualdad:

$$\|Au - b\| \leq \|A\|\|x - u\|$$

$$\|x\| \leq \|A^{-1}\|\|b\|$$

Entonces

$$\|Au - b\|\|x\| \leq c(A)\|x - u\|\|b\|$$

Segunda desigualdad (razonamiento similar, tomando $v = Au$):

$$\frac{\|x - u\|}{\|x\|} = \frac{\|A^{-1}b - A^{-1}v\|}{\|x\|} \leq \frac{\|A^{-1}\|\|b - v\|}{\|x\|} \leq \|A\|\|A^{-1}\| \frac{\|b - v\|}{\|b\|} = c(A) \frac{\|Au - b\|}{\|b\|}$$

□

La interpretación de estas desigualdades es que x es la solución del sistema $Ax = b$ y u es la solución aproximada.

La estimación del error relativo de la solución en función del condicionamiento de A y el error relativo generado al tomar Au por b es $Au - b$, que recibe el nombre de **residuo**, que en general no es nulo.

Nota. Tema 1: un residuo pequeño no garantiza un error relativo pequeño de la solución.

Ejemplo 3.1. $0 < \alpha \leq 1$

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1-\alpha \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 2 \\ 2-\alpha \end{bmatrix}, \mathbf{u} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ solución del sistema } Ax = b, \mathbf{Au} = \begin{bmatrix} 2 \\ 2-2\alpha \end{bmatrix}$$

$$\|A\|_{\infty} = \left\| \begin{bmatrix} 1 & 1 \\ 1 & 1-\alpha \end{bmatrix} \right\|_{\infty} = 2, \quad \|A^{-1}\|_{\infty} = \left\| \begin{bmatrix} 1-\frac{1}{\alpha} & \frac{1}{\alpha} \\ \frac{1}{\alpha} & -\frac{1}{\alpha} \end{bmatrix} \right\|_{\infty} = \frac{2}{\alpha}$$

$$\frac{\alpha}{4} \frac{\|Au - b\|}{\|b\|} \leq \frac{\|x - u\|}{\|x\|} \leq \frac{4}{\alpha} \frac{\|Au - b\|}{\|b\|}$$

El control será bueno si α está próximo a 1, y será malo si α está próximo a 0. Explícitamente:

$$\frac{\alpha^2}{8} \leq 1 \leq 2$$

Proposición 3.2. Sean $N \geq 1$, $A, B \in \mathbb{R}^{N \times N}$ con A regular, $b, c \in \mathbb{R}^N$ y supongamos que el método iterativo

$$\begin{cases} x_0 \text{ dado} \\ n \geq 1 \Rightarrow x_n = Bx_{n-1} + c \end{cases}$$

converge a la solución del sistema $Ax = b$, cualquiera sea $x_0 \in \mathbb{R}^N$. Si además $\|\cdot\|$ es una norma en \mathbb{R}^N con norma matricial inducida en $\mathbb{R}^{N \times N}$ denotada de igual forma, tal que $\|B\| < 1$, entonces, para todo $n \geq 1$ se tiene:

$$(i) \quad \|x_n - x\| \leq \frac{\|B\|^n}{1-\|B\|} \|x_1 - x_0\|$$

$$(ii) \quad \|x_n - x\| \leq \|B\| \|x_{n-1} - x\|$$

$$(iii) \quad \|x_n - x\| \leq \frac{\|B\|}{1-\|B\|} \|x_n - x_{n-1}\|$$

Demostración 3.2.

¿(i)?

Supongamos $n \geq 1$, entonces

$$\|x_{n+1} - x_n\| = \|Bx_n - Bx_{n-1}\| \leq \|B\| \|x_n - x_{n-1}\|$$

inductivamente tenemos que

$$\|x_{n+1} - x_n\| \leq \|B\|^n \|x_1 - x_0\|$$

Sea $m \geq n \geq 1$, entonces

$$\|x_n - x_m\| \leq \sum_{j=0}^{m-n-1} \|x_{j+n+1} - x_{j+n}\| \leq \sum_{j=0}^{m-n-1} \|B\|^{j+n} \|x_1 - x_0\| \leq \frac{\|B\|^n}{1 - \|B\|} \|x_1 - x_0\|$$

Por último, tomamos límite en $m \rightarrow \infty$.

¿(ii)?

Supongamos $n \geq 1$, entonces

$$\|x_n - x\| = \|Bx_{n-1} + c - Bx - c\| \leq \|B\| \|x_{n-1} - x\|$$

¿(iii)?

Supongamos $n \geq 1$ y usando la desigualdad triangular tenemos que

$$\|x_{n-1} - x\| = \|x_{n-1} - x + x_n - x_n\| \leq \|x_{n-1} - x_n\| + \|x_n - x\|$$

Usando (ii)

$$\|x_n - x\| \leq \|B\| \|x_n - x_{n-1}\| + \|B\| \|x_n - x\|$$

Hemos obtenido la estimación pedida (falta reorganizar).

□

Observaciones:

- Las acotaciones (i) y (ii) dan una estimación del error absoluto, aunque hay una diferencia importante entre ambas: la primera no necesita conocer la solución exacta x .
- La estimación del error absoluto (iii) no solo se obtiene sin necesidad de conocer x sino que además constituye un criterio de parada cuando se programa el método numérico y se alcanza una tolerancia dada.

III | Tema 3. Interpolación

En este tema vamos a aprender a calcular el polinomio que pasa por unas coordenadas dadas.

1 Interpolación polinómica: Lagrange y Newton. Error de interpolación

Sean las coordenadas $(x_0, y_0), (x_1, y_1), \dots, (x_N, y_N) \in \mathbb{R}^2 : x_i \neq x_j$ con $i, j = 0, 1, \dots, N$, entonces existe una única función polinómica de grado menor o igual que N $p : \mathbb{R} \rightarrow \mathbb{R}$ tal que $p(x_i) = y_i$ con $i = 0, 1, \dots, N$.

El problema que tenemos que resolver en este tema es determinar explícitamente el polinomio de interpolación $p \in \mathbb{P}_N$.

Para calcular p debemos calcular una base de \mathbb{P}_N y sus propiedades. Veamos dos formas diferentes de hacerlo.

1.1 Polinomio de interpolación tipo Lagrange

Teorema 1.1. Sean las coordenadas $(x_0, y_0), (x_1, y_1), \dots, (x_N, y_N) \in \mathbb{R}^2$, entonces

$$\exists! p \in \mathbb{P}_N : p(x_i) = y_i : x_i \neq x_j \text{ con } i, j = 0, 1, \dots, N \Rightarrow p(x) = \sum_{i=0}^N y_i l_i(x) \text{ con } l_i(x) = \prod_{j=0 \wedge j \neq i}^N \frac{x - x_j}{x_i - x_j}$$

donde $\{l_0(x), l_1(x), \dots, l_N(x)\}$ es la base, dependiente de los x_i 's. Primero calculamos la base y luego el polinomio $p(x)$.

La fórmula anterior se conoce como **forma de Lagrange del polinomio de interpolación** y las funciones base **polinomios de Lagrange o característicos**.

Ejemplo 1.1. Sea $f : \mathbb{R} \rightarrow \mathbb{R}$ tal que $f(x) = e^x (\forall x \in \mathbb{R})$. Vamos a dar tres valores a la función para calcular el polinomio de Lagrange. Por ejemplo, en los puntos de abscisas -1, 0, 1:

$$(x_0, y_0) = (-1, e^{-1}), (x_1, y_1) = (0, 1), (x_2, y_2) = (1, e)$$

Ahora calculamos la base

$$l_0(x) = \prod_{j=0 \wedge j \neq 0}^2 \frac{x - x_j}{x_0 - x_j} = \frac{x^2}{2} - \frac{x}{2}$$

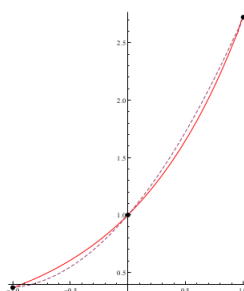
$$l_1(x) = 1 - x^2$$

$$l_2(x) = \frac{x^2}{2} + \frac{x}{2}$$

Por último, calculamos el polinomio de interpolación $p(x)$:

$$p(x) = \sum_{i=0}^2 e^{x_i} l_i(x) = \left(\frac{e}{2} + \frac{1}{2e} - 1 \right) x^2 + \left(\frac{e}{2} - \frac{1}{2e} \right) x + 1$$

Dibujándolo, en trazo discontinuo es $p(x)$ y en trazo continuo es $f(x)$, observamos que pasa por las coordenadas dadas.



1.2 Forma de Newton del polinomio de interpolación

Teorema 1.2. Sean las coordenadas $(x_0, y_0), (x_1, y_1), \dots, (x_N, y_N) \in \mathbb{R}^2$, entonces

$$\exists! p \in \mathbb{P}_N : p(x_i) = y_i : x_i \neq x_j \text{ con } i, j = 0, 1, \dots, N \Rightarrow p(x) = \sum_{i=0}^N f[x_0, \dots, x_i] \omega_i(x)$$

con

$$f[x_0, \dots, x_i] = \frac{f[x_1, \dots, x_i] - f[x_0, \dots, x_{i-1}]}{x_i - x_0} \text{ con } 1 \leq i \leq N$$

y

$$\omega_i(x) = \prod_{j=0}^{i-1} (x - x_j)$$

La expresión anterior es la conocida como **forma de Newton del polinomio de interpolación** y las funciones base **polinomios nodales**.

Otra forma de calcular $f[x_0, \dots, x_N]$ sería viéndolo como una "matriz", en la que solo nos fijaremos en la diagonal:

x_0	$f[x_0]$				
x_1	$f[x_1]$	$f[x_0, x_1]$			
x_2	$f[x_2]$	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$		
\vdots	\vdots	\vdots	\vdots	\ddots	
x_N	$f[x_N]$	$f[x_{N-1}, x_N]$	$f[x_{N-2}, x_{N-1}, x_N]$	\cdots	$f[x_0, \dots, x_N]$

donde $f[x_0] = f(x_0)$ y $\omega_0(x) = 1$.

Ejemplo 1.2. Dados los datos $(x_0, y_0) = (0.5, 1)$, $(x_1, y_1) = (1, 0.2)$, $(x_2, y_2) = (-0.25, 1)$, $(x_3, y_3) = (-0.5, 0.2)$, $(x_4, y_4) = (0.2, 1/3)$, vamos a calcular las diferencias divididas.

0.5	1				
1	0.2	-1.6			
-0.25	1	-0.64	-1.28		
-0.5	0.2	3.2	-2.56	1.28	
0.2	$\frac{1}{3}$	$\frac{4}{21}$	$-\frac{1264}{189}$	$\frac{4876}{945}$	$-\frac{36664}{2835}$

Algunas cuentas que hemos hecho para calcularlo son:

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} = \frac{0.2 - 1}{1 - 0.5} = -1.6$$

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{-0.64 - (-1.6)}{-0.25 - 0.5} = -1.28$$

1.3 Error de interpolación. Convergencia y estabilidad. Polinomios de Chebyshev

Primero, vamos a establecer la notación de algunos conceptos.

Sean $x_0, x_1, \dots, x_N \in [a, b]$, una función $f \in C([a, b])$ y su polinomio de interpolación $I_N^f \in \mathbb{P}_N$ tal que $I_N : C([a, b]) \rightarrow \mathbb{P}_N$ bien definida (unicidad polinomio interpolación). Además, tiene la propiedad de proyección, ya que $I_N^2 = I_N$.

Definición 1.1 (Error de interpolación). Definimos al error de interpolación $E_N^f(x)$ como

$$E_N^f(x) := f(x) - I_N^f(x)$$

Definición 1.2.

$$\|I_N\|_\infty := \sup \left\{ \|I_N^f\|_\infty : f \in C([a, b]), \|f\|_\infty = 1 \right\}$$

Proposición 1.1.

$$\|I_N\|_\infty = \sup \left\{ \frac{\|I_N^f\|_\infty}{\|f\|_\infty} : f \in C([a, b]), f \neq 0 \right\}$$

Definición 1.3 (Función de Lebesgue). Definimos la **función de Lebesgue** como

$$\lambda_N(x) := \sum_{i=0}^N |l_i(x)|$$

siendo l_0, \dots, l_N la base de los polinomios de Lagrange.

Definición 1.4 (Constante de Lebesgue). Definimos la **constante de Lebesgue** como

$$\Lambda_N := \|\lambda_N\|_\infty$$

Proposición 1.2.

$$\|I_N\|_\infty = \Lambda_N$$

Corolario 1.1. Sea $f \in C([a, b])$, entonces

$$\|E_N^f\|_\infty \leq (1 + \Lambda_N) \inf \{ \|f - p\|_\infty : p \in \mathbb{P}_N \}$$

Teorema 1.3. Teorema de aproximación uniforme de Weierstrass.

$$\lim_{N \rightarrow \infty} \inf \{ \|f - p\|_\infty : p \in \mathbb{P}_N \} = 0$$

Como $\lim_{N \rightarrow \infty} \Lambda_N = \infty$, no podemos asegurar convergencia uniforme, es decir, no podemos asegurar que $\lim_{N \rightarrow \infty} \|E_N^f\|_\infty = 0$.

Podemos ver si es estable según Λ_N , que solo depende de los nodos y nos sirve para medir el condicionamiento.

Ejemplo 1.3. Ejemplo de Bernstein


$$f(x) := |x| \text{ en } [-1, 1]$$

Los nodos están igualmente espaciados

$$x_i^{(N)} = -1 + \frac{2i}{N}, \text{ con } i = 0, 1, \dots, N$$

$$\lim_{N \rightarrow \infty} \|E_N f\|_\infty = \infty$$

Representamos la función f en rojo e $I_N f$ en trazo discontinuo, para $N = 4, 10, 12, 14$.

 media/Bernstein.png

Ejemplo 1.4. Ejemplo de Runge

$$f(x) := \frac{1}{1 + 25x^2}, x \in [-1, 1]$$

Los nodos están igualmente espaciados

$$x_i^{(N)} = -1 + \frac{2i}{N}, \text{ con } i = 0, 1, \dots, N$$

$$\lim_{N \rightarrow \infty} \|E_N f\|_\infty = \infty$$

Representamos la función f en rojo e $I_N f$ en trazo discontinuo, para $N = 4, 6, 10, 12$.

 media/Runge.png

Proposición 1.3. Sean los nodos x_0, x_1, \dots, x_N , los datos $(x_0, f(x_0)), (x_1, f(x_1)), \dots, (x_N, f(x_N))$ y los datos perturbados $(x_0, g(x_0)), (x_1, g(x_1)), \dots, (x_N, g(x_N))$. Entonces

$$\|I_N^f - I_N^g\|_\infty = \max \left\{ \left| \sum_{i=0}^N (f(x_i) - g(x_i)) l_i(x) \right| : x \in [a, b] \right\} \leq \Lambda_N \cdot \max \{ |f(x_i) - g(x_i)| : i = 0, 1, \dots, N \}$$

Ejemplo 1.5. Sea $f : [-1, 1] \rightarrow \mathbb{R}, f(x) := e^x$. Elegimos 21 nodos uniformemente distribuidos en $[-1, 1]$, por ejemplo

$$i = 0, 1, \dots, 20 \Rightarrow x_i := -1 + \frac{2i}{20}$$


Los datos que obtenemos son $(x_i, f(x_i))$ y los datos perturbados son $(x_i, g(x_i))$ con

$$g(x_i) = f(x_i + (-1)^i 10^{-4})$$

Tenemos que $\max \{ |f(x_i) - g(x_i)| : i = 0, \dots, 20 \} = 2.7184 \cdot 10^{-4}$

Pero $\|I_{20}f - I_{20}g\|_\infty = 1.1964$

En rojo representamos $I_{20}f$ y en verde $I_{20}g$:

 media/ej1-5_t2.png

Por lo que hay mal condicionamiento $\Lambda_{20} = 10986.7058$.

$$\|I_{20}f - I_{20}g\|_{\infty} \leq \Lambda_{20} \max \{ |f(x_i) - g(x_i)| : i = 0, \dots, 20 \}$$

$$1.1964 \leq 10986.7958 \cdot 0.2718 \cdot 10^{-3} = 2.9862$$

media/ej1-5_t2_.png

Proposición 1.4. Sean x_0, x_1, \dots, x_N números reales distintos, sea $x \in \mathbb{R}$ y sean $a := \min \{x, x_0, x_1, \dots, x_N\}$ y $b := \max \{x, x_0, x_1, \dots, x_N\}$. Supongamos además que $f \in C^{N+1}([a, b])$. Entonces existe $\xi \in]a, b[$ tal que

$$E_N^{f(x)} = \frac{f^{(N+1)}(\xi)}{(N+1)!} \omega_{N+1}(x),$$

donde $\omega_{N+1}(x)$ es el polinomio nodal de grado $N+1$.

$$f(x) = I_N f(x) + \frac{f^{(N+1)}(\xi)}{(N+1)!} \omega_{N+1}(x)$$

Es análoga a la fórmula de Taylor.

Corolario 1.2. Bajo las condiciones de la proposición anterior

$$\|E_N^f\|_{\infty} \leq \frac{\|f^{(N+1)}\|_{\infty}}{(N+1)!} (b-a)^{N+1}$$

La convergencia es uniforme: $f \in C^{\infty}([a, b])$

$$\lim_{N \rightarrow \infty} \frac{\|f^{(N+1)}\|_{\infty}}{(N+1)!} (b-a)^{N+1} = 0 \Rightarrow \lim_{N \rightarrow \infty} \|E_N^f\|_{\infty} = 0$$

Ejemplo 1.6. Sea $f(x) = e^x$, ($x \in [a, b]$), entonces

$$\frac{\|f^{(N+1)}\|_{\infty}}{(N+1)!} (b-a)^{N+1} = \frac{e^b}{(N+1)!} (b-a)^{N+1} \rightarrow 0 \quad (N \rightarrow \infty) \Rightarrow \lim_{N \rightarrow \infty} \|E_N^f\|_{\infty} = 0$$

Ejemplo 1.7. Sea $f(x) = \cos x$, ($x \in [a, b]$), entonces

$$\frac{\|f^{(N+1)}\|_{\infty}}{(N+1)!} (b-a)^{N+1} \leq \frac{1}{(N+1)!} (b-a)^{N+1} \rightarrow 0 \quad (N \rightarrow \infty) \Rightarrow \lim_{N \rightarrow \infty} \|E_N^f\|_{\infty} = 0$$

La elección de nodos influye en el error de interpolación

$$E_N^f(f) = \frac{f^{(N+1)}(\xi)}{(N+1)!} \omega_{N+1}(x)$$

Por lo que será mejor tener los nodos uniformemente distribuidos en $[a, b]$.

Ejemplo 1.8. Veamos un ejemplo con 2 puntos.

Sean $[a, b] = [x_0, x_1]$, $h := x_1 - x_0$.

Error de interpolación puntual:

$$f \in C^2([a, b]), x \in [a, b] \Rightarrow \exists \xi \in]a, b[: E_1^f(x) = \frac{f''(\xi)}{2} (x - x_0)(x - x_1)$$

Estimación uniforme para el error de interpolación:

$$\|E_1^f\|_\infty \leq \frac{\|f''\|_\infty}{2} h^2$$

Esta estimación puede ser mejorable (acotación de $\omega_2(x)$):

$$x \in [x_0, x_1] \Rightarrow \exists 0 \leq t \leq 1 : x = x_0 + th$$

$$|(x - x_0)(x - x_1)| = (x - x_0)(x_1 - x) = th(1 - t)h \leq \frac{h^2}{4} \Rightarrow \|E_1^f\|_\infty \leq \frac{\|f''\|_\infty}{8} h^2$$

Ejemplo 1.9. Veamos otro ejemplo con 3 puntos igualmente espaciados.

Sean $[a, b] = [x_0, x_2]$, $x_1 = \frac{x_2 + x_0}{2}$, $h := x_1 - x_0 = x_2 - x_1$.

Error de interpolación puntual:

$$f \in C^3([a, b]), x \in [a, b] \Rightarrow \exists \xi \in]a, b[: E_2^f(x) = \frac{f'''(\xi)}{6} (x - x_0)(x - x_1)(x - x_2)$$

Estimación uniforme para el error de interpolación:

$$\|E_2^f\|_\infty \leq \frac{\|f'''\|_\infty}{6} h^3$$

Esta estimación puede ser mejorable (acotación de $\omega_3(x)$):

$$x = x_0 + th, \quad 0 \leq t \leq 2$$

Análogo al caso de 2 puntos.

La elección de los nodos influye en el error de interpolación. En general, aun siendo mejorable, para

nodos equidistantes tenemos que

$$\|E_N^f\|_\infty \leq \frac{\|f^{(N+1)}\|_\infty}{(N+1)!} h^{N+1}$$

error de interpolación de orden $O(h^{N+1})$.

$[-1,1], [a,b] \leftrightarrow [-1,1]$ isomorfismo afín. Sea $\theta \in \mathbb{R}, N \geq 1$

$$\cos(N+1)\theta + \cos(N-1)\theta = 2\cos\theta\cos N\theta$$

$$\Downarrow$$

$$\cos(N+1)\theta = 2\cos\theta\cos N\theta - \cos(N-1)\theta$$

$$\Downarrow$$

$$\exists T_N \in \mathbb{P}_N : T_N(\cos\theta) = \cos N\theta$$

$$\cos : [0, \pi] \longleftrightarrow [-1,1] \text{ biyección}$$

$$x = \cos\theta$$

$$\cos(N+1)\theta = 2\cos\theta\cos N\theta - \cos(N-1)\theta$$

y

$$T_N(\cos\theta) = \cos N\theta$$

$$\left| \begin{array}{l} T_0(x) = 1 \\ T_1(x) = x \\ i = 0, 1, 2, \dots \Rightarrow T_{i+2}(x) = 2xT_{i+1}(x) - T_i(x) \end{array} \right.$$

donde T_i es el **polinomio de Chebyshev** de grado i . Sus propiedades importantes son:

- $T_N \in \mathbb{P}_N$ con N ceros reales, todos en $[-1,1]$ y coeficiente líder 2^{N-1} .
- $T_N \in C([-1,1]), \|T_N\|_\infty = 1$.

Los **nodos de Chebyshev** son:

$$\left| \begin{array}{l} T_N \in \mathbb{P}_N \\ T_N(\cos\theta) = \cos(N\theta) \\ \cos(N\theta) = 0 \end{array} \right. \quad x_i^{(N)} = \cos \frac{2i+1}{2N} \pi, \quad (i = 0, \dots, N-1) \text{ los } N \text{ ceros de } T_N$$

Teorema 1.4. Chebyshev. Sea $N \geq 1$ y sea $p \in \mathbb{P}_N$ un polinomio con coeficiente líder 1. Entonces

$$\max \{ |p(x)| : x \in [-1,1] \} \geq \frac{1}{2^{N-1}}$$

Demostración 1.1. Vamos a demostrando usando la reducción al absurdo. Supongamos que

$$\max \{ |p(x)| : x \in [-1, 1] \} < \frac{1}{2^{N-1}}$$

Sea

$$q(x) := \frac{1}{2^{N-1}} T_N - p(x) \in \mathbb{P}_{N-1}$$

Luego

$$q(y_0^{(N)}) > 0, q(y_1^{(N)}) < 0, q(y_2^{(N)}) > 0, \dots, (-1)^N q(y_N^{(N)}) > 0$$

Como $q \in \mathbb{P}_{N-1}$, tiene al menos N ceros reales distintos, luego $q = 0$, lo cual es una contradicción. \square

Corolario 1.3. Sean $N \geq 1, x_0, \dots, x_N \in [-1, 1]$ y sean $x_0^{(N+1)}, x_1^{(N+1)}, \dots, x_N^{(N+1)}$ los nodos de Chebyshev. Entonces en el espacio normado $C([-1, 1])$,

$$\left\| \prod_{i=0}^N (x - x_i) \right\|_{\infty} \geq \left\| \prod_{i=0}^N (x - x_i^{(N+1)}) \right\|_{\infty} = \frac{1}{2^N}$$

Con los nodos de Chebyshev tenemos que

$$\|E_N^f\|_{\infty} \leq \frac{\|f^{(N+1)}\|_{\infty}}{(N+1)!} \frac{1}{2^N}$$

1.4 Otros problemas de interpolación: Hermite y caso general

Sean los nodos distintos $x_0, x_1, \dots, x_N \in [a, b]$, los órdenes de derivación $m_0, m_1, \dots, m_N \geq 0$ y la función $f \in C^M([a, b])$, $M := \max \{m_i : i = 0, \dots, N\}$. El **problema de interpolación de Hermite** consiste en encontrar $p \in \mathbb{P}$ de grado mínimo K con

$$i = 0, 1, \dots, N \Rightarrow \left[j = 0, 1, \dots, m_i \Rightarrow p^{(j)}(x_i) = f^{(j)}(x_i) \right]$$

$$p \in \mathbb{P}_K : p^{(j)}(x_i) = f^{(j)}(x_i) \Leftrightarrow L(p) = f^{(j)}(x_i)$$

con $L : \mathbb{P}_K \longrightarrow \mathbb{R}$ forma lineal

$$p \mapsto L(p) := p^{(j)}(x_i)$$

El **problema general de interpolación** consiste en, sea E espacio vectorial real de dimensión N , $L_1, \dots, L_N : E \longrightarrow \mathbb{R}$ formas lineales, $d_1, \dots, d_N \in \mathbb{R}$, encontrar un único $p \in E : [i = 1, \dots, N \Rightarrow L_i(p) = d_i]$

Proposición 1.5. Sea E un espacio vectorial (real) de dimensión $N \geq 1$, sean $L_1, \dots, L_N : E \rightarrow \mathbb{R}$ formas lineales y sea p_1, \dots, p_N una base de E . Entonces, el problema general de interpolación admite una única solución, cualesquiera sean $d_1, \dots, d_N \in \mathbb{R}$, si, y solo si,

$$\det [L_i(p_j)]_{i,j=1}^N \neq 0$$

$[L_i(p_j)]_{i,j=1}^N$ es la matriz de coeficientes de un sistema cuadrado.

Unisolvencia para cualesquiera $d_1, \dots, d_N \Leftrightarrow$ unisolvencia para $d_1 = \dots = d_N = 0$.

A partir de ahora, vamos a enfocar el problema con Lagrange.

Sea E espacio vectorial real de dimensión N , $L_1, \dots, L_N : E \rightarrow \mathbb{R}$ formas lineales, $d_1, \dots, d_N \in \mathbb{R}$, el problema general de interpolación consiste en encontrar un único $p \in E : [i = 1, \dots, N \Rightarrow L_i(p) = d_i]$

Sea l_1, \dots, l_N base de E de Lagrange

$$i, j = 1, \dots, N \Rightarrow L_i(l_j) = \delta_{ij} \Rightarrow p := \sum_{i=1}^N d_i l_i$$

que es la solución del problema general de interpolación. La dificultad está en determinar la base de Lagrange, que siempre existe. Veamos los distintos casos que existen.

(i) $(x_0, y_0, d_0), (x_1, y_1, d_1), \dots, (x_N, y_N, d_N) \in \mathbb{R}^3 : i, j = 0, 1, \dots, N \Rightarrow x_i \neq x_j$

$$E := \mathbb{P}_{2N+1}, \quad \{1, x, \dots, x^{2N+1}\} \text{ base}$$

$$i = 0, 1, \dots, N \Rightarrow L_{2i}(p) := p(x_i), \quad L_{2i+1}(p) = p'(x_i)$$

$$\text{encontrar un unico } p \in E : [i = 0, 1, \dots, N \Rightarrow p(x_i) = y_i, \quad p'(x_i) = d_i]$$

$$\det [L_i(p_j)]_{i,j=0}^{2N+1} \neq 0$$

$$p(x_0) = p'(x_0) = 0 \Rightarrow p(x) = (x - x_0)^2 q_0(x), \text{ para cierto } q_0 \in \mathbb{P}_{2N-1}$$

$$\exists \alpha \in \mathbb{R} : p(x) = \alpha \prod_{i=0}^{N+1} (x - x_i)^2$$

Como $p \in \mathbb{P}_{2N+1}$, entonces $p = 0$. Luego hay unisolvencia.

La base de Lagrange nos lleva a los polinomios de Hermite $h_0(x), h_1(x), \dots, h_{2N+1}(x)$. Sea $i = 0, \dots, N$, entonces

$$h_{2i}(x) = (1 - 2(x - x_i)l'_i(x_i))l_i^2(x)$$

$$h_{2i+1}(x) = (x - x_i)l_i(x)^2$$

$l_0(x), l_1(x), \dots, l_N(x)$ polinomios de Lagrange

$$i = 0, 1, \dots, N \Rightarrow l_i(x) = \prod_{j=0 \wedge j \neq i}^N \frac{x - x_j}{x_i - x_j}$$

La solución es

$$p(x) = \sum_{i=0}^N (y_i h_{2i}(x) + d_i d_{2i+1}(x))$$

Ejemplo 1.10. Sean los nodos $x_0 = -1, x_1 = -0.5, x_2 = 0, x_3 = 0.5, x_4 = 1$, sean los valores $y_0 = 1, y_1 = 2, y_2 = 0, y_3 = -2, y_4 = -1$ y sean las derivadas $d_0 = 0, d_1 = 1, d_2 = -1, d_3 = 1, d_4 = 0$. Entonces

$$p(x) = -x - \frac{833}{18}x^3 + \frac{385}{2}x^5 - \frac{740}{3}x^7 + \frac{904}{9}x^9$$

(ii) $(x_0, d_0, d_1, \dots, d_N) \in \mathbb{R}^{N+2}$

$$E := \mathbb{P}_N, \{1, x, \dots, x^N\} \text{ base}$$

$$i = 0, 1, \dots, N \Rightarrow L_i(p) := p^{(i)}(x_0)$$

El problema es encontrar un único $p \in E : [i = 0, 1, \dots, N \Rightarrow p^{(i)}(x_0) = d_i]$

$$\det [L_i(p_j)]_{i,j=0}^N = \det \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^N \\ 0 & 1 & 2x_0 & \cdots & Nx_0^{N-1} \\ 0 & 0 & 2 & \cdots & N(N-1)x_0^{N-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & N! \end{bmatrix} \neq 0$$

Hay unisolvencia.

Con la base de Lagrange obtenemos los polinomios de Taylor $t_0(x), t_1(x), \dots, t_N(x)$

$$i = 0, 1, \dots, N \Rightarrow t_i(x) = \frac{(x - x_0)^i}{i!}$$

Solución:

$$p(x) = \sum_{i=0}^N \frac{d_i}{i!} (x - x_0)^i$$

$d_i \leftrightarrow f^{(i)}(x_0)$ polinomio de Taylor.

Ejercicio 1.1. Decide razonadamente si el problema de interpolación: encontrar $p \in \mathbb{P}_3$ tal que

$$p(0) = 0, p'(0) = 0, p'(-1) = 0, p''(-0.5) = 0$$

es unisolviente.

2 Interpolación mediante funciones splines

Para mejorar la precisión tenemos que hacer una partición del intervalo $[a, b]$ y que el grado polinomial sea bajo. Sea P una **partición** de $[a, b]$ tal que $P = \{a = x_0 < x_1 < \dots < x_N = b\}$.

Definición 2.1 (Espacio de funciones splines). Dados un intervalo $[a,b]$ y una partición P del mismo, el espacio de funciones splines de clase k y grado m viene dado por

$$\mathbb{S}_m^k(P) := \left\{ s \in C^k([a,b]) : i = 0, 1, \dots, N-1 \Rightarrow s|_{[x_i, x_{i+1}]} \in \mathbb{P}_m \right\}$$

Tenemos que $\mathbb{S}_m^m = \mathbb{P}_m$ y $k < m$.

2.1 Funciones splines lineales

Empecemos en el espacio vectorial $\mathbb{S}_1^0(P)$. Tenemos que $\dim \mathbb{S}_1^0(P) = 2 \cdot N - (N-1) = N+1$ y la base usual es $B_0(x), B_1(x), \dots, B_N(x)$.

La siguiente definición es muy importante.

Definición 2.2 (Base usual). Sean $a < b$ y sea $P = \{a = x_0 < x_1 < \dots < x_N = b\}$ una partición del intervalo $[a,b]$. La **base usual** del espacio $\mathbb{S}_1^0(P)$ viene dada por

$$i = 0, 1, \dots, N \Rightarrow B_i(x) := \begin{cases} \frac{x-x_{i-1}}{x_i-x_{i-1}}, & \text{si } x \in [x_{i-1}, x_i] \\ \frac{x_{i+1}-x}{x_{i+1}-x_i}, & \text{si } x \in [x_i, x_{i+1}] \\ 0, & \text{fuera} \end{cases}$$

Ejercicio 2.1. Comprueba que, efectivamente, las $N+1$ funciones splines anteriores forman una base de $\mathbb{S}_1^0(P)$.

Indicación: para la independencia lineal basta observar que $B_i(x_j) = \delta_{ij}$, con $i, j = 0, 1, \dots, N$, mientras que para probar que forman un sistema de generadores, solo hay que demostrar que

$$s \in \mathbb{S}_1^0(P) \Rightarrow s(x) = \sum_{i=0}^N s(x_i) B_i(x)$$

Dados un intervalo $[a,b]$, una partición $P = \{a = x_0 < x_1 < \dots < x_N = b\}$ y una función $f : [a,b] \rightarrow \mathbb{R}$, se considera el **problema de interpolación en $\mathbb{S}_1^0(P)$**

$$\text{encontrar } s \in \mathbb{S}_1^0(P) : [i = 0, 1, \dots, N \Rightarrow s(x_i) = f(x_i)]$$

La estructura del problema de interpolación general es claramente unisolvante. Con la base de Lagrange tenemos $B_0(x), B_1(x), \dots, B_N(x)$. La solución, notando $s = S_N^1 f$

$$S_N^1 f(x) = \sum_{i=0}^N f(x_i) B_i(x)$$

Definición 2.3 (Error de interpolación).

$$E_N^{f(x)} = f(x) - S_N^1 f(x)$$

Sea $f \in C^2([a,b])$

$$i = 0, 1, \dots, N-1, x \in [x_i, x_{i+1}] \Rightarrow |f(x) - S_N^1 f(x)| \leq \frac{\|f''\|_\infty}{8} (x_{i+1} - x_i)^2$$

$$\Downarrow$$

$$\|f - S_N^1 f\|_\infty \leq \frac{\|f''\|_\infty}{8} (\max \{(x_{i+1} - x_i) : i = 0, \dots, N-1\})^2$$

Hemos probado:

Proposición 2.1. Con la notación anterior, si $f \in C^2([a,b])$, entonces

$$\lim_{N \rightarrow \infty} \max \{(x_{i+1} - x_i) : i = 0, \dots, N-1\} = 0 \quad \Rightarrow \quad \lim_{N \rightarrow \infty} \|f - S_N^1 f\|_\infty = 0$$

Proposición 2.2. Sea $f \in C([a,b])$, sea $P = \{a = x_0 < x_1 < \dots < x_N = b\}$ y sea $S_N^1 f$ el único elemento de $\mathbb{S}_1^0(P)$ que interpola a f en los nodos de P , i.e.,

$$i = 0, 1, \dots, N \Rightarrow S_N^1 f(x_i) = f(x_i)$$

Entonces:

(i) Si $i = 0, 1, \dots, N-1$ y $x \in [x_i, x_{i+1}]$, entonces

$$S_N^1 f(x) \leq \max \{f(x_i), f(x_{i+1})\}$$

(ii) $\|S_N^1 f\|_\infty \leq \|f\|_\infty$

(iii) $\|f - S_N^1 f\|_\infty \leq 2 \cdot \inf \{\|f - s\|_\infty : s \in \mathbb{S}_1^0(P)\}$

2.2 Funciones splines cúbicas

Ahora trabajaremos con el espacio $\mathbb{S}_3^2(P)$. Tenemos que $\dim \mathbb{S}_3^2(P) = 4N - 3(N-1) = N+3$.

Definición 2.4 (Funciones splines cúbicas naturales).

$$i = 0, 1, \dots, N \Rightarrow s(x_i) = f(x_i)$$

2 condiciones adicionales: $s''(a) = 0 = s''(b)$.

La construcción de las funciones splines cúbicas naturales es directa en cada subintervalo $[x_{i-1}, x_i]$. Por comodidad expositiva, los puntos son distribuidos uniformemente en $[a, b]$, esto es:

$$h = \frac{b-a}{N}, \quad i = 0, 1, \dots, N \quad \Rightarrow \quad x_i = a + ih$$

Vamos a establecer la notación:

$$\begin{aligned} s &:= S_N^2 f \\ s_i &:= S_N^2 f|_{[x_i, x_{i+1}]} \quad i = 0, 1, \dots, N-1 \\ y_i &:= s(x_i), \quad d_i := s'(x_i) \quad c_i := s''(x_i) \quad i = 0, 1, \dots, N \end{aligned}$$

Tenemos, para $i = 1, \dots, N$:

Primero calculamos los $c'_i s$, que son solución del sistema:

$$\begin{bmatrix} 2 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{2} & 2 & \frac{1}{2} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{2} & 2 & \frac{1}{2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{2} & 2 & \frac{1}{2} & 0 \\ 0 & \cdots & 0 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_N \end{bmatrix} = \frac{3}{h^2} \begin{bmatrix} 0 \\ y_2 - 2y_1 + y_0 \\ y_3 - 2y_2 + y_1 \\ \vdots \\ y_N - 2y_{N-1} + y_{N-2} \\ 0 \end{bmatrix}$$

Luego calculamos:

$$\begin{aligned} \alpha_{i-1} &= \frac{y_i - y_{i-1}}{h} - \frac{h}{6}(c_i - c_{i-1}) \\ \beta_{i-1} &= y_{i-1} - c_{i-1} \frac{h^2}{6} \end{aligned}$$

Por último:

$$x \in [x_{i-1}, x_i] \Rightarrow s_{i-1}(x) = c_{i-1} \frac{(x_i - x)^3}{6h} + c_i \frac{(x - x_{i-1})^3}{6h} + \alpha_{i-1}(x - x_{i-1}) + \beta_{i-1}$$

Ejemplo 2.1. Sea $[-1, 1]$, $N = 7$, $y_0 = 1, y_1 = 0.2, y_2 = 0, y_3 = 0.3, y_4 = -1, y_5 = 1, y_6 = 0.3, y_7 = -0.2$. Luego la solución del sistema es

$$c_0 = 0, c_1 = 5.3882, c_2 = 22.5474, c_3 = -58.8278, c_4 = 95.1637, c_5 = -79.277, c_6 = 23.4942, c_7 = 0$$

Ahora vamos a estudiar el **error de interpolación**.

Proposición 2.3. Con la notación anterior, si $f \in C^4([a,b])$ entonces

$$j = 0, 1, 2, 3 \Rightarrow \|f^{(j)} - S_N^2 f^{(j)}\|_\infty \leq K_j h^{4-j} \|f^{(4)}\|_\infty$$

donde $K_0 = \frac{5}{384}$, $K_1 = \frac{1}{24}$, $K_2 = \frac{3}{8}$ y $K_4 = 1$.

Por último, estudiaremos el **principio de mínima energía**.

Proposición 2.4. Sean $P=a = x_0 < x_1 < \dots < x_N = b$, $f:[a,b] \rightarrow \mathbb{R}$ y $g \in C^2([a,b])$ de forma que

$$g(x_i) = f(x_i), \quad i = 0, 1, \dots, N$$

Si además s es la función spline cúbica natural que satisface la misma condición de interpolación, entonces

$$\int_a^b s''(x)^2 dx \leq \int_a^b g''(x)^2 dx$$

dándose la igualdad si, y solo si, $g = s$.

IV | Tema 4. Aproximación

Al tener unos datos experimentales, la aproximación sirve para hacer previsiones y ajustes.

1 Aproximación por mínimos cuadrados discreta y continua

1.1 Principio del mínimo

Sea la función $f : \mathbb{R} \rightarrow \mathbb{R}$ tal que $f(x) = \frac{1}{2}ax^2 - bx$, con $a > 0$. Es una parábola, por lo que alcanza su valor mínimo en $x = \frac{b}{a}$ y tenemos que $f(b/a) = -\frac{1}{2a}b^2$.

Esto se puede extender a \mathbb{R}^N . El número a pasa a ser una matriz simétrica y definida positiva $A \in \mathbb{R}^{N \times N}$, el número b pasa a ser el vector $b \in \mathbb{R}^N$. Por lo que tenemos que:

$$f : \mathbb{R}^N \rightarrow \mathbb{R}, f(x) = \frac{1}{2}x^T Ax - b^T x$$

Teorema 1.1. Principio del mínimo. Si $A \in \mathbb{R}^{N \times N}$ es una matriz simétrica y definida positiva, $b \in \mathbb{R}^N$ y $f : \mathbb{R}^N \rightarrow \mathbb{R}$ es la función cuadrática definida en cada $x \in \mathbb{R}^N$ como

$$f(x) = \frac{1}{2}x^T Ax - b^T x$$

entonces f alcanza su mínimo en un único vector, la solución del sistema $Ax = b$, y su valor mínimo es

$$-\frac{1}{2}b^T A^{-1}b$$

Demostración 1.1. Como A es definida positiva, entonces A es regular:

$$Ax = 0 \Rightarrow x^T Ax = 0 \Rightarrow x = 0$$

$A^{-1}b$ es solución de $Ax = b$.

Sea $x := A^{-1}b$, $y \in \mathbb{R}^N$. Como A es simétrica:

$$\begin{aligned}
f(y) - f(x) &= \frac{1}{2}y^T Ay - b^T y - \frac{1}{2}x^T Ax + b^T x \\
&= \frac{1}{2}y^T Ay - \frac{1}{2}x^T Ax + x^T Ax - b^T y + b^T x \\
&= \frac{1}{2}y^T Ay - \frac{1}{2}x^T Ax + x^T Ax - x^T Ay \\
&= \frac{1}{2}(y - x)^T A(y - x)
\end{aligned}$$

Como A es definida positiva, f alcanza su mínimo en $y = x = A^{-1}b$:

$$f(A^{-1}b) = -\frac{1}{2}b^T A^{-1}b$$

□

Proposición 1.1.

• A simétrica y definida positiva $\Leftrightarrow A$ admite factorización LU Cholesky.
($A \in \mathbb{R}^{N \times N}$)

- Condición suficiente:

A regular $\Rightarrow A^T A$ simétrica y definida positiva. ($A \in \mathbb{R}^{N \times N}$)

- $A \in \mathbb{R}^{M \times N}$. Condición suficiente (más general):

$\text{rango}(A) = N \Leftrightarrow$ columnas de A son linealmente independientes $\Rightarrow A^T A$ es simétrica y definida positiva.

(y $N \leq M$).

- Condición suficiente (aún más general): $C \in \mathbb{R}^{M \times M}$ simétrica y definida positiva, $A \in \mathbb{R}^{M \times N}$:

$\text{rango}(A) = N \Leftrightarrow$ columnas de A linealmente independientes $\Rightarrow A^T C A$ simétrica y definida positiva.

(y $N \leq M$).

1.2 Aproximación por mínimos cuadrados discreta

Ahora vamos a estudiar una aplicación del principio del mínimo, vamos a ver cuál es la mejor aproximación euclídea en **dimensión finita**.

Sea S un subespacio vectorial de \mathbb{R}^M ,

$$\left\{ \begin{bmatrix} a_{11} \\ \vdots \\ a_{M1} \end{bmatrix}, \dots, \begin{bmatrix} a_{1N} \\ \vdots \\ a_{MN} \end{bmatrix} \right\} \text{ base de } S$$

Luego $\text{rango}(A) = N$, $S = \{Ax : x \in \mathbb{R}^N\}$.

Por lo que si tenemos el vector de soluciones $b \in \mathbb{R}^M$, entonces

(i) $b \in S \Leftrightarrow Ax = b$ compatible

(ii) $b \notin S \Leftrightarrow Ax = b$ incompatible

Ahora veamos si existe un vector $Ax \in S$ más próximo, usando la norma euclídea, a b .

$\exists x \in \mathbb{R}^N : y \in \mathbb{R}^N \Rightarrow \|Ax - b\|_2 \leq \|Ay - b\|_2 ?$

\Updownarrow

$\exists x \in \mathbb{R}^N : y \in \mathbb{R}^N \Rightarrow \|Ax - b\|_2^2 \leq \|Ay - b\|_2^2 ?$

\Updownarrow

$\exists x \in \mathbb{R}^N : y \in \mathbb{R}^N \Rightarrow (Ax - b)^T(Ax - b) \leq (Ay - b)^T(Ay - b) ?$

\Updownarrow

$\exists x \in \mathbb{R}^N : y \in \mathbb{R}^N \Rightarrow x^T A^T A x + b^T b - 2b^T A x \leq y^T A^T A y + b^T b - 2b^T A y ?$

\Updownarrow

$\exists x \in \mathbb{R}^N : y \in \mathbb{R}^N \Rightarrow \frac{1}{2}x^T A^T A x - b^T A x \leq \frac{1}{2}y^T A^T A y - b^T A y ?$

Por el principio del mínimo, como $A^T A$ es simétrica y definida positiva,

$$f(x) = \frac{1}{2}x^T A^T A x - b^T A x$$

Luego es cierto que existen tales vectores y la solución es

$$A^T A x = A^T b$$

Esta solución recibe el nombre de **ecuaciones normales**. Además, Ax es la mejor aproximación de b en S en el sentido de los mínimos cuadrados (discretos, $(\mathbb{R}^M, \|\cdot\|_2)$).

Ejemplo 1.1. Sean

$$S := \text{lin} \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}, \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix} \right\}, \quad b = \begin{bmatrix} 2 \\ 1 \\ 7.3 \end{bmatrix}$$

Vamos a calcular la mejor aproximación de b en S usando mínimos cuadrados.

Primero calculamos una base.

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \right\}$$

Luego

$$A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Las ecuaciones normales son

$$A^T A x = A^T b \Leftrightarrow \begin{bmatrix} 14 & 32 \\ 32 & 77 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 25.9 \\ 56.8 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3.2722 \\ -0.62 \end{bmatrix}$$

Por lo que

$$A \begin{bmatrix} 3.2722 \\ -0.62 \end{bmatrix} = \begin{bmatrix} 0.783 \\ 3.43 \\ 6.083 \end{bmatrix} \text{ mejor aproximación de } \begin{bmatrix} 2 \\ 1 \\ 7.3 \end{bmatrix} \text{ en } S$$

La mejor aproximación en el sentido de los mínimos cuadrados (discretos) tiene una interpretación geométrica. Sea S un subespacio vectorial de \mathbb{R}^M ,

$$\left\{ \begin{bmatrix} a_{11} \\ \vdots \\ a_{M1} \end{bmatrix}, \dots, \begin{bmatrix} a_{1N} \\ \vdots \\ a_{MN} \end{bmatrix} \right\} \text{ base de } S \Rightarrow S = \{Ax : x \in \mathbb{R}^N\}$$

Denotemos por $\langle \cdot, \cdot \rangle$ al producto escalar euclídeo usual en \mathbb{R}^M , entonces las ecuaciones normales se pueden expresar usando el producto escalar:

$$A^T A x = A^T b \Leftrightarrow \left\langle \begin{bmatrix} a_{1j} \\ \vdots \\ a_{Mj} \end{bmatrix}, b - Ax \right\rangle = 0, \text{ con } j = 1, \dots, N$$

Geoméricamente, el vector $b - Ax$ es perpendicular a los vectores de la base de S , equivalentemente a todos los de S ($b - Ax \in S^\perp$), luego Ax es la **proyección ortogonal** de b sobre S .

$$P_S(b) = Ax$$

Cálculo equivalente de x :

$$b - Ax = b - \sum_{i=1}^N x_i \Rightarrow \left\langle \begin{bmatrix} a_{1j} \\ \vdots \\ a_{Mj} \end{bmatrix}, b - \sum_{i=1}^N x_i \begin{bmatrix} a_{1i} \\ \vdots \\ a_{Mi} \end{bmatrix} \right\rangle = 0, \text{ con } j = 1, \dots, N$$

Ejercicio 1.1. Demuestra que si un vector es ortogonal a los de una base de un subespacio vectorial S de \mathbb{R}^M , entonces lo es a todos los vectores de S .

Ejemplo 1.2. Sean

$$S := \text{lin} \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}, \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix} \right\}, \quad b = \begin{bmatrix} 2 \\ 1 \\ 7.3 \end{bmatrix}$$

Vamos a calcular la proyección ortogonal de b sobre S , $P_S(b)$. Primero, calcularemos una base.

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \right\} \Rightarrow A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Hemos obtenido el siguiente sistema

$$\begin{cases} \left\langle \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 7.3 \end{bmatrix} - x_1 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - x_2 \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \right\rangle = 0 \\ \left\langle \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 7.3 \end{bmatrix} - x_1 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - x_2 \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \right\rangle = 0 \end{cases}$$

\Downarrow

$$\begin{bmatrix} 14 & 32 \\ 32 & 77 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 25.9 \\ 56.8 \end{bmatrix}$$

\Downarrow

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3.2722 \\ -0.62 \end{bmatrix}$$

Por lo que la proyección ortogonal de b sobre S es

$$P_S(b) = 3.2722 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - 0.62 \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = A \begin{bmatrix} 3.2722 \\ -0.62 \end{bmatrix} = \begin{bmatrix} 0.783 \\ 3.43 \\ 6.083 \end{bmatrix}$$

El siguiente teorema resume todo lo anterior.

Teorema 1.2. Sea $A \in \mathbb{R}^{M \times N}$ con $\text{rango}(A) = N$, sea S el subespacio vectorial de \mathbb{R}^M definido como

$$S := \{Ax : x \in \mathbb{R}^N\}$$

y sea $b \in \mathbb{R}^M$. Entonces existe un único vector $x \in \mathbb{R}^N$ de forma que

$$\|Ax - b\|_2 = \min \{ \|Ay - b\|_2 : y \in \mathbb{R}^N \}$$

De hecho, el vector Ax (mejor aproximación de b en S , proyección ortogonal de b sobre S) viene caracterizado por las ecuaciones normales

$$A^T Ax = A^T b,$$

equivalentemente,

$$\left\langle \begin{bmatrix} a_{1j} \\ \vdots \\ a_{Mj} \end{bmatrix}, b - \sum_{i=1}^N x_i \begin{bmatrix} a_{1i} \\ \vdots \\ a_{Mi} \end{bmatrix} \right\rangle = 0, \text{ con } j = 1, \dots, N$$

Veamos ahora una aplicación de este teorema: ajuste de datos. Sean los datos $(x_0, y_0), (x_1, y_1), \dots, (x_M, y_M) \in \mathbb{R}^2$ tales que $x_i \neq x_j$ $i, j = 0, 1, \dots, M$. Entonces existe una única función polinómica de grado menor o igual que M $p : \mathbb{R} \rightarrow \mathbb{R}$ tal que

$$p(x_i) = y_i, \quad i = 0, 1, \dots, M$$

Ejemplo 1.3. Veamos un ejemplo de ajuste de datos mediante una recta o lineal.

$$f(x) = ax + b$$

Tenemos que calcular a y b . Para ello, hay que minimizar

$$\sum_{i=0}^M (ax_i + b - y_i)^2 \Leftrightarrow \left\| a \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_M \end{bmatrix} + b \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_M \end{bmatrix} \right\|_2^2 \Leftrightarrow \left\| a \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_M \end{bmatrix} + b \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_M \end{bmatrix} \right\|_2$$

Luego

$$S := \text{lin} \left\{ \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_M \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right\}$$

Solución $P_S(y)$

$$\dim(S) = 2 \Rightarrow P_S(y) = ax + b1 \Rightarrow f(x) = ax + b$$

Ejemplo 1.4. Vamos a calcular la recta de ecuación $y = ax + b$ que mejor aproxima, en el sentido de los mínimos cuadrados, los datos $(1,1), (2,2), (0, -0.8), (-1,1), (-2,-1.1)$.

Tenemos que

$$S := \text{lin} \left\{ \begin{bmatrix} 1 \\ 2 \\ 0 \\ -1 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\}, y = \begin{bmatrix} 1 \\ 2 \\ -0.8 \\ 1 \\ -1.1 \end{bmatrix}, \dim(S) = 2$$

Luego

$$P_S(y) = a \begin{bmatrix} 1 \\ 2 \\ 0 \\ -1 \\ -2 \end{bmatrix} + b \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Por lo que

$$a = 0.62, b = 0.42 \Rightarrow f(x) = 0.62x + 0.42$$

Ejemplo 1.5. Veamos un ejemplo de ajuste de datos mediante una parábola o cuadrático.

$$f(x) = ax^2 + bx + c$$

Vamos a calcular a, b, c . Para ello, hay que minimizar

$$\sum_{i=0}^M (ax_i^2 + bx_i + c - y_i)^2 \Leftrightarrow \left\| a \begin{bmatrix} x_0^2 \\ x_1^2 \\ \vdots \\ x_M^2 \end{bmatrix} + b \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_M \end{bmatrix} + c \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_M \end{bmatrix} \right\|_2^2 \Leftrightarrow \left\| a \begin{bmatrix} x_0^2 \\ x_1^2 \\ \vdots \\ x_M^2 \end{bmatrix} + b \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_M \end{bmatrix} + c \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_M \end{bmatrix} \right\|_2$$

Luego

$$S := \text{lin} \left\{ \begin{bmatrix} x_0^2 \\ x_1^2 \\ \vdots \\ x_M^2 \end{bmatrix}, \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_M \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right\}$$

Solución $P_S(y)$ Denotemos por x^2 a $[x_0^2, x_1^2, \dots, x_M^2]^T$.

$$\dim(S) = 2 \Rightarrow P_S(y) = ax + b1 \Rightarrow f(x) = ax + b$$

Como $\dim(S) = 3$, entonces

$$P_S(y) = ax^2 + bx + c \Rightarrow f(x) = ax^2 + bx + c$$

Ejemplo 1.6. Vamos a determinar la parábola de ecuación $y = ax^2 + bx + c$ que mejor aproxima, en el sentido de los mínimos cuadrados, los datos $(1,1)$, $(2,2)$, $(0,-0.8)$, $(-1,1)$, $(-2,-1.1)$. Tenemos que

$$S := \text{lin} \left\{ \begin{bmatrix} 1 \\ 4 \\ 0 \\ 1 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 0 \\ -1 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\}, y = \begin{bmatrix} 1 \\ 2 \\ -0.8 \\ 1 \\ -1.1 \end{bmatrix}, \dim(S) = 3$$

Luego

$$P_S(y) = a \begin{bmatrix} 1 \\ 4 \\ 0 \\ 1 \\ 4 \end{bmatrix} + b \begin{bmatrix} 1 \\ 2 \\ 0 \\ -1 \\ -2 \end{bmatrix} + c \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Por lo que

$$a = 0.1, b = 0.62, c = 0.22 \Rightarrow f(x) = 0.1x^2 + 0.62x + 0.22$$

1.3 Aproximación por mínimos cuadrados general: caso continuo

Denotaremos al productor escalar usual en \mathbb{R}^N por $\langle \cdot, \cdot \rangle$.

Definición 1.1 (Producto escalar). Dado un espacio vectorial real E , diremos que una aplicación $\langle \cdot, \cdot \rangle : E \times E \longrightarrow \mathbb{R}$ es un **producto escalar** en E siempre que se cumpla lo siguiente.

- (i) $\langle x, y \rangle = \langle y, x \rangle$.
- (ii) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$.
- (iii) $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$.
- (iv) $\langle x, x \rangle \geq 0$.
- $\langle x, x \rangle = 0 \Leftrightarrow x = 0$.

donde $x, y, z \in E$ y $\lambda \in \mathbb{R}$

Definición 1.2 (Espacio euclídeo). E es un **espacio euclídeo** con producto escalar $\langle \cdot, \cdot \rangle$.

El par $(E, \langle \cdot, \cdot \rangle)$ es un **espacio euclídeo**.

Ejemplo 1.7. Veamos unos ejemplos de productos escalares usuales.

- $E = \mathbb{R}^N$

$$\langle x, y \rangle := \sum_{i=1}^N x_i y_i \quad x, y \in \mathbb{R}^N$$

- $E = \mathbb{R}^{M \times N}$

$$\langle A, B \rangle := \text{traza}(AB^T) \quad A, B \in \mathbb{R}^{M \times N}$$

- $E = C([a, b])$

$$\langle f, g \rangle := \int_a^b f(x)g(x)dx \quad f, g \in C([a, b])$$

- $E = C^k([a, b]), k \in \mathbb{N}$

$$\langle f, g \rangle := \sum_{j=0}^k \int_a^b f^{(j)}(x)g^{(j)}(x)dx \quad f, g \in C^k([a, b])$$

Proposición 1.2. Sea $(E, \langle \cdot, \cdot \rangle)$ un espacio euclídeo, $x \in E$, $N, M \geq 1$, $\lambda_1, \dots, \lambda_N, \mu_1, \dots, \mu_M \in \mathbb{R}$ y $x_1, \dots, x_N, y_1, \dots, y_M \in E$, entonces

- $\langle x, 0 \rangle = 0$.
- $\left\langle \sum_{i=1}^N \lambda_i x_i, \sum_{j=1}^M \mu_j y_j \right\rangle = \sum_{i=1}^N \sum_{j=1}^M \lambda_i \mu_j \langle x_i, y_j \rangle$

Proposición 1.3. Sea $(E, \langle \cdot, \cdot \rangle)$ un espacio euclídeo, la **norma (euclídea inducida)** es

$$\|x\| := \sqrt{\langle x, x \rangle} \quad x \in E$$

Ejemplo 1.8. Veamos unos ejmplos sobre las normas euclídeas inducidas usuales.

- $E = \mathbb{R}^N$

$$\|x\| = \sqrt{\sum_{i=1}^N x_i^2} \quad x \in \mathbb{R}^N$$

- $E = \mathbb{R}^{M \times N}$

$$\|A\| = \sqrt{\text{traza}(AA^T)} = \sqrt{\sum_{i=1}^M \sum_{j=1}^N a_{ij}^2} \quad A \in \mathbb{R}^{M \times N}$$

(norma de Frobenius).

- $E = C([a, b])$

$$\|f\| = \sqrt{\int_a^b f(x)^2 dx} \quad f \in C([a, b])$$

- $E = C^k([a, b]), k \in \mathbb{N}$

$$\|f\| = \sqrt{\sum_{j=0}^k \int_a^b f^{(j)}(x)^2 dx} \quad f \in C^k([a, b])$$

Sea $(E, \langle \cdot, \cdot \rangle)$ un espacio euclídeo, $x, y \in E$, la norma inducida por $\langle \cdot, \cdot \rangle$

$$\text{dist}(x, y) = \|x - y\|$$

$v \in E, S$ subespacio finito dimensional vectorial de E, a_1, \dots, a_N base de S , entonces ¿existe un vector $u \in S$ más próximo a v ?

$$\text{¿} \exists u \in S : w \in S \Rightarrow \|u - v\| \leq \|w - v\| \text{?}$$

⇔

$$\text{¿} \exists u \in S : w \in S \Rightarrow \|u - v\|^2 \leq \|w - v\|^2 \text{?}$$

⇔

$$\text{¿} \exists u \in S : w \in S \Rightarrow \langle u - v, u - v \rangle \leq \langle w - v, w - v \rangle \text{?}$$

Sean

$$u = \sum_{i=1}^N x_i a_i, \quad w = \sum_{i=1}^N y_i a_i$$

$$\text{¿} \exists x \in \mathbb{R}^N :$$

$$y \in \mathbb{R}^N \Rightarrow \left\langle \sum_{i=1}^N x_i a_i - v, \sum_{i=1}^N x_i a_i - v \right\rangle \leq \left\langle \sum_{i=1}^N y_i a_i - v, \sum_{i=1}^N y_i a_i - v \right\rangle ?$$

⇔

¿ $\exists x \in \mathbb{R}^N$:

$$y \in \mathbb{R}^N \Rightarrow \sum_{i=1}^N x_i \sum_{j=1}^N x_j \langle a_i, a_j \rangle - 2 \sum_{i=1}^N x_i \langle a_i, v \rangle \leq \sum_{i=1}^N \sum_{j=1}^N y_i y_j \langle a_i, a_j \rangle - 2 \sum_{i=1}^N y_i \langle a_i, v \rangle?$$

Sean

$$A := \begin{bmatrix} \langle a_1, a_1 \rangle & \langle a_1, a_2 \rangle & \cdots & \langle a_1, a_N \rangle \\ \langle a_2, a_1 \rangle & \langle a_2, a_2 \rangle & \cdots & \langle a_2, a_N \rangle \\ \vdots & \vdots & \cdots & \vdots \\ \langle a_N, a_1 \rangle & \langle a_N, a_2 \rangle & \cdots & \langle a_N, a_N \rangle \end{bmatrix} \in \mathbb{R}^{N \times N}, \quad b := \begin{bmatrix} \langle a_1, v \rangle \\ \langle a_2, v \rangle \\ \vdots \\ \langle a_N, v \rangle \end{bmatrix}$$

$$\left\langle \sum_{i=1}^N x_i a_i, \sum_{j=1}^N y_j a_j \right\rangle = x^T A y, \quad \left\langle \sum_{i=1}^N x_i a_i, v \right\rangle = b^T x \quad x, y \in \mathbb{R}^N$$

¿ $\exists x \in \mathbb{R}^N : y \in \mathbb{R}^N \Rightarrow \frac{1}{2} x^T A x - b^T x \leq \frac{1}{2} y^T A y - b^T y$?

Por el principio del mínimo y por ser A simétrica y definida positiva (hay que probarlo), entonces

$$f(x) = \frac{1}{2} x^T A x - b^T x$$

Luego la solución es

$$Ax = b$$

que recibe el nombre de **ecuaciones normales**. Además, $u = \sum_{i=1}^N x_i a_i$ es la mejor aproximación de v en S en el sentido de los mínimos cuadrados (discretos, si $\dim(E) < \infty$, continuos en caso contrario).

Ejercicio 1.2. Demuestra que la matriz A anterior es simétrica y definida positiva.

Ejercicio 1.3. Relaciona las ecuaciones normales anteriores con las obtenidas en dimensión finita ($E = \mathbb{R}^M$).

Ejemplo 1.9. Sean $E := C([-1, 1])$ con el producto escalar usual, $f(x) := |x|$, $(-1 \leq x \leq 1)$, $S := \mathbb{P}_4$. Vamos a calcular la mejor aproximación de f en S en el sentido de los mínimos cuadrados.

Primero, calculamos una base:

$$\{1, x, x^2, x^3, x^4\}$$

Las ecuaciones normales son

$$A = \left[\int_{-1}^1 x^i x^j dx \right]_{i,j=0,1,2,3,4} = \begin{bmatrix} 2 & 0 & \frac{2}{3} & 0 & \frac{2}{5} \\ 0 & \frac{2}{3} & 0 & \frac{2}{5} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} & 0 & \frac{2}{7} \\ 0 & \frac{2}{5} & 0 & \frac{2}{7} & 0 \\ \frac{2}{5} & 0 & \frac{2}{7} & 0 & \frac{2}{9} \end{bmatrix}$$

$$b = \left[\int_{-1}^1 x^i |x| dx \right]_{i=0,1,2,3,4} = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \\ 0 \\ \frac{1}{3} \end{bmatrix}$$

$$Ax = b \Leftrightarrow x = \begin{bmatrix} \frac{15}{128} \\ 0 \\ \frac{105}{64} \\ 0 \\ -\frac{105}{128} \end{bmatrix}$$

Luego

$$u(x) = \frac{15}{128} + \frac{105}{64}x^2 - \frac{105}{128}x^4 \in \mathbb{P}_4$$

es la mejor aproximación de f en S .

Ejemplo 1.10. Sean $E := C([0, 2\pi])$ con el producto escalar usual, $f(x) := \frac{x}{1.5}$ ($0 \leq x \leq 2\pi$), $S := \text{lin}\{1, \cos x, \sin x, \cos 2x, \sin 2x, \cos 3x, \sin 3x\}$. Vamos a calcular la mejor aproximación de f en S usando mínimos cuadrados.

Primero, calculamos una base

$$\{1, \cos x, \sin x, \cos 2x, \sin 2x, \cos 3x, \sin 3x\} = \{\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5, \varphi_6, \varphi_7\}$$

Las ecuaciones normales son: (ventaja: base ortogonal)

$$A = \left[\int_0^{2\pi} \varphi_i(x) \varphi_j(x) dx \right]_{i,j=1,\dots,7} = \begin{bmatrix} 2\pi & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \pi & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \pi & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \pi & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \pi & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \pi & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \pi \end{bmatrix}$$

$$b = \left[\int_0^{2\pi} \varphi_i(x) \frac{x}{1.5} dx \right]_{i=1, \dots, 7} = \begin{bmatrix} \frac{4\pi^2}{3} \\ 0 \\ -\frac{4\pi}{3} \\ 0 \\ -\frac{2\pi}{3} \\ 0 \\ -\frac{4\pi}{9} \end{bmatrix}$$

$$Ax = b \Leftrightarrow x = \begin{bmatrix} \frac{2\pi}{3} \\ 0 \\ -\frac{4}{3} \\ 0 \\ -\frac{2}{3} \\ 0 \\ -\frac{4}{9} \end{bmatrix}$$

Luego

$$u(x) = \frac{2\pi}{3} - \frac{4}{3}\sin x - \frac{2}{3}\sin 2x - \frac{4}{9}\sin 3x \in S$$

es la mejor aproximación de f en S .

Vamos a interpretar geoméricamente la mejor aproximación en el sentido de los mínimos cuadrados.

Sea $(E, \langle \cdot, \cdot \rangle)$ un espacio euclídeo, $v \in E$, S un subespacio vectorial finito dimensional de E , a_1, \dots, a_N base de S , entonces las ecuaciones normales son

$$Ax = b \Leftrightarrow \sum_{j=1}^N x_j \langle a_i, a_j \rangle = \langle a_i, v \rangle \Leftrightarrow \left\langle a_i, v - \sum_{j=1}^N x_j a_j \right\rangle = 0 \quad \text{con } i = 1, \dots, N$$

Geoméricamente, $v - \sum_{j=1}^N x_j a_j$ es perpendicular a los vectores de la base de S , equivalentemente a todos los de S ($v - \sum_{j=1}^N x_j a_j \in S^\perp$). Luego $u = \sum_{j=1}^N x_j a_j$ es la **proyección ortogonal** de v sobre S .

$$P_S(v) = u$$

Cálculo equivalente de las coordenadas $x \in \mathbb{R}^N$ de la proyección ortogonal $u = \sum_{j=1}^N x_j a_j$ en la base a_1, \dots, a_N

$$\left\langle a_i, v - \sum_{j=1}^N x_j a_j \right\rangle = 0 \quad \text{con } i = 1, \dots, N$$

Ejercicio 1.4. Demuestra que si un vector de un espacio euclídeo $(E, \langle \cdot, \cdot \rangle)$ es ortogonal a los de una base de un subespacio vectorial S de E , entonces lo es a todos los vectores de S .

Ejemplo 1.11. Sean $E := C([-1, 1])$ con el producto escalar usual, $f(x) := |x|$, $(-1 \leq x \leq 1)$, $S := \mathbb{P}_4$. Vamos a calcular la proyección ortogonal de f sobre S : $P_S(f)$.

Primero, calculamos una base

$$\{1, x, x^2, x^3, x^4\}$$

Obtenemos el siguiente sistema

$$\begin{cases} \sum_{j=0}^4 \int_{-1}^1 (x_{j+1} x^j - |x|) dx = 0 \\ \sum_{j=0}^4 \int_{-1}^1 (x_{j+1} x^j - |x|) x dx = 0 \\ \sum_{j=0}^4 \int_{-1}^1 (x_{j+1} x^j - |x|) x^2 dx = 0 \\ \sum_{j=0}^4 \int_{-1}^1 (x_{j+1} x^j - |x|) x^3 dx = 0 \\ \sum_{j=0}^4 \int_{-1}^1 (x_{j+1} x^j - |x|) x^4 dx = 0 \end{cases}$$

$$x = \left[\frac{15}{128}, 0, \frac{105}{64}, 0, -\frac{105}{128} \right]^T \Rightarrow u(x) = \frac{15}{128} + \frac{105}{64} x^2 - \frac{105}{128} x^4 \in \mathbb{P}_4$$

mejor aproximación de f en S .

Teorema 1.3. Mejor aproximación. Sean $(E, \langle \cdot, \cdot \rangle)$ un espacio euclídeo, S un subespacio vectorial finito dimensional con base $\{a_1, \dots, a_N\}$ y sea $v \in E$. Entonces existe un único vector $u \in S$ (mejor aproximación de v en S , proyección ortogonal de v sobre S) de forma que

$$\|u - v\| = \min \{\|w - v\| : w \in S\}.$$

De hecho, las coordenadas $x \in \mathbb{R}^N$ del vector u en la base a_1, \dots, a_N vienen caracterizadas por las ecuaciones normales $Ax = b$, donde

$$A := [\langle a_i, a_j \rangle]_{i,j=1}^N \in \mathbb{R}^{N \times N} \quad \text{y} \quad b := [\langle a_i, v \rangle]_{i=1}^N \in \mathbb{R}^N$$

i.e.,

$$\left\langle a_i, v - \sum_{j=1}^N x_j a_j \right\rangle = 0, \quad i = 1, \dots, N$$