Métodos Numéricos I

José Luis Bravo Trinidad, heredados por Javier Cabello Sánchez

Índice general

Tema 1 Métodos directos de resolución de sistemas lineales

El objetivo de este tema es obtener *la* solución del sistema de ecuaciones lineales

$$Ax = b, \quad A \in \mathcal{M}_{n\times n}, \quad b \in \mathbb{R}^n,$$

donde A es una matriz no singular y estudiar la dependencia de la solución de los errores en A y b.

En asignaturas anteriores (Álgebra Lineal II y Métodos Computacionales II), se ha estudiado cómo discutir y resolver el sistema por eliminación gaussiana. En este tema nos centraremos en resolver mediante factorización de matrices y en analizar los errores producidos al resolver un sistema.

Recordemos que la solución de un sistema lineal se puede obtener por la regla de Cramer. Sin embargo, este método es muy poco eficiente. En particular, si calculamos los determinantes por desarrollo de adjuntos (es decir, sin usar la eliminación gaussiana) se necesitarían más de $100!$ operaciones elementales para una matriz A de 100×100, lo que lo hace imposible de utilizar en la práctica.

Comenzaremos recordando algunos elementos del análisis matricial, después estudiaremos cómo resolver los sistemas mediante factorización de matrices y finalmente estudiaremos las normas matriciales y los errores en las soluciones de los sistemas lineales.

1.1 Elementos básicos del análisis matricial

En general, trabajaremos sobre los números complejos, pues los métodos iterativos (que estudiaremos en el siguiente tema), son más simples en este contexto. Además, el primer gran resultado que veremos (Teorema 1.2) solamente se cumple en $\mathbb{C}$, en $\mathbb{R}$ es falso.

Dado un vector columna v de $\mathbb{C}^n$, definimos su traspuesto conjugado como el vector fila tal que cada coordenada es la conjugada de la coordenada correspondiente de v y lo denotamos v^*.

Definimos el producto escalar de dos vectores $u, v \in \mathbb{C}^n$ como v^*u. Es sencillo probar que es un producto interior. Además

$$v^*v = \|v\|_2 := \sqrt{\sum_{k=1}^{n} |v_k|^2}.$$

1.1.1 Matriz traspuesta conjugada

Dada una matriz $A \in \mathcal{M}_{m\times n}$, $A = (a_{ij})_{i=1\ldots m}^{j=1\ldots n}$ definiremos su traspuesta conjugada[1] como

$$A^* := (\overline{a}_{ji}) \in \mathcal{M}_{n\times m}$$

Proposición 1.1

Sean $A, B \in \mathcal{M}_{m\times n}$, $C \in \mathcal{M}_{n\times p}$. Se verifica:

1. $A^{**} = A$.
2. $(A + B)^* = A^* + B^*$.
3. $(\lambda A)^* = \overline{\lambda} A^*$.

[1] La matriz traspuesta conjugada también se denota A^H. En algunos textos se denomina matriz adjunta (no confundir con la matriz de los adjuntos) o hermitiano.

4. $(AC)^* = C^*A^*$.

Demostración Ejercicio. □

Proposición 1.2

Si A es invertible, entonces A^* es invertible y se verifica

$$(A^*)^{-1} = (A^{-1})^*.$$

Además, $\det A^* = \overline{\det A}$.

Demostración Ejercicio. □

Sean

$$A = (a_{ij})_{1\le i,j\le n} \in \mathcal{M}_n, \quad I = (\delta_{ij})_{1\le i,j\le n}.$$

Decimos que A es

- simétrica si es real y $A = A^t$.
- hermítica si $A = A^*$.
- ortogonal si es real y $AA^t = A^tA = I$.
- unitaria si $AA^* = A^*A = I$.
- normal si $AA^* = A^*A$.

1.1.2 Matrices triangulares y diagonal dominantes

Atendiendo a la naturaleza de los elementos, definimos los siguientes tipos de matrices:

- A es diagonal si $a_{ij} = 0, i \neq j, 1 \le i, j \le n$.
- A es triangular superior si $a_{ij} = 0, i > j, 1 \le i, j \le n$.
- A es triangular inferior si $a_{ij} = 0, i < j, 1 \le i, j \le n$.
- A es (estrictamente) diagonal dominante si

$$|a_{ii}| \underset{(>)}{\ge} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \quad 1 \le i \le n.$$

Proposición 1.3

Toda matriz estrictamente diagonal dominante es invertible.

Demostración

Sea A estrictamente diagonal dominante. Para probar la existencia de inversa, vamos a aplicar Rouché-Frobenius. En particular, mostraremos que $Ax = 0$ tiene solución única (lo que implica rango máximo). Suponemos que no por reducción al absurdo (existe z tal que $Az = 0$ y $z \neq 0$). Sea i_0 el índice donde se alcanza el máximo de los módulos de los componentes de z (estrictamente positivo)

Como $Az = 0$, $\sum a_{ij}z_j = 0$. En particular para i_0, pero

$$|a_{i_0i_0}||z_{i_0}| = |\sum_{j=1,j\neq i_0}^{n} a_{i_0j}z_j| \le \sum_{j=1,j\neq i_0}^{n} |a_{i_0j}z_j| \le \sum_{j=1,j\neq i_0}^{n} |a_{i_0j}||z_{i_0}|$$

$$= |z_{i_0}| \sum_{j=1,j\neq i_0}^{n} |a_{i_0j}| < |z_{i_0}||a_{i_0i_0}|.$$

De esta contradicción, A es invertible.

□

✍ **Ejercicio 1.1** Obtener un ejemplo de una matriz diagonal dominante que no sea invertible.

1.1.3 Partición de matrices

Dada una matriz $D \in \mathcal{M}_n$, diremos que

$$D = \begin{pmatrix} D_{11} & D_{12} & \cdots & D_{1k_D} \\ D_{21} & D_{22} & \cdots & D_{2k_D} \\ \vdots & \vdots & \ddots & \vdots \\ D_{m_D1} & D_{m_D2} & \cdots & D_{m_Dk_D} \end{pmatrix},$$

es una partición de D si para cada $1 \leq i, j \leq n$, D_{ij} es una matriz con el mismo número de filas que $D_{i\bar{j}}$ para todo $\bar{j}$ y con el mismo número de columnas que $D_{\bar{i}j}$ para todo $\bar{i}$.

Supongamos que $A = (A_{ij})$, $B = (B_{ij})$, $C = (C_{ij})$ son particiones de $A, B, C \in \mathcal{M}_n$.

Teorema 1.1

Si cada producto $A_{is}B_{sj}$ se puede formar y

$$C_{ij} = \sum_s A_{is}B_{sj},$$

entonces $C = AB$.

♡

Demostración Ejercicio. □

✍ **Ejercicio 1.2** Elegir dos matrices y comprobar el resultado. Se recomienda practicar este producto.

1.1.4 Autovalores y autovectores

Recordemos algunas definiciones y propiedades de los autovalores.

Sea $A \in \mathcal{M}_n$. Se denomina

- Polinomio característico de A a $p(\lambda) = \det(A - \lambda I)$.
- Autovalores de A a las raíces del polinomio característico.
- Espectro de A, $\mathrm{Sp}(A)$, al conjunto de autovalores de A.
- Radio espectral de A, $\rho(A) := \max\{|\lambda| : \lambda \in \mathrm{Sp}(A)\}$.
- Autovector asociado a un autovalor $\lambda \in \mathrm{Sp}(A)$ a todo vector v que satisfaga $Av = \lambda v$.

Proposición 1.4

1. *A es invertible si y sólo si $0 \notin \mathrm{Sp}(A)$.*
2. *Para todo autovalor λ, $\dim(\mathrm{Ker}(A - \lambda I)) > 0$ (todo autovalor tiene un autovector no nulo asociado).*
3. *$\mathrm{Sp}(AB) = \mathrm{Sp}(BA)$. En particular, el espectro es invariante por cambio de base.*
4. *Los autovectores asociados a un autovalor λ constituyen el subespacio vectorial $\mathrm{Ker}(A - \lambda I)$, de dimensión menor o igual que la multiplicidad de λ como raíz del polinomio característico.*
5. *Si λ_1, λ_2 son dos autovalores distintos de A, entonces*

$$\mathrm{Ker}(A - \lambda_1 I) \cap \mathrm{Ker}(A - \lambda_2 I) = \{0\}.$$

♠

Demostración

1. Trivial
2. Teorema de Rouché-Frobenius (el rango no puede ser máximo pues el determinante es nulo).
3. Sea v un autovector de AB con autovalor asociado λ. Entonces $ABv = \lambda v$. Luego
$$BA(Bv) = B\lambda v = \lambda Bv,$$
de donde Bv autovector de BA con autovalor asociado λ.
4. Si expresamos la matriz en una base tal que los primeros vectores son los autovectores asociados, entonces la aplicación tiene en la diagonal el autovalor, tantas veces como autovectores, luego la multiplicidad algebraica es mayor.
5. Sencilla (ejercicio). De hecho, los subespacios vectoriales asociados a los distintos autovectores están en suma directa. Si la suma de sus dimensiones es n, la matriz es diagonalizable.

□

Proposición 1.5

1. *Los autovalores de una matriz hermítica son siempre reales.*
2. *Los autovalores de una matriz unitaria tienen módulo* 1. *De hecho, las matrices unitarias de orden n se corresponden con las isometrías lineales de $(\mathbb{C}^n, \|\cdot\|_2)$.*

♠

Demostración Sea A, λ autovalor, v autovector.

1) Si A es hermítica:
$$\lambda v^* v = v^* \lambda v = v^* A v = v^* A^* v = (Av)^* v = (\lambda v)^* v = \bar{\lambda} v^* v,$$
luego $\lambda = \bar{\lambda} \in \mathbb{R}$.

2) Si A es unitaria:
$$|\lambda|^2 v^* v = \bar{\lambda}\lambda v^* v = \bar{\lambda} v^* \lambda v = (\lambda v)^* \lambda v = (Av)^* Av = v^* A^* Av = v^* v.$$
Entonces $|\lambda|^2 = 1$.

Además se tiene que, dados $u, v \in \mathbb{C}^n$,
$$\langle Au, Av\rangle = (Au)^*(Av) = u^* A^* Av = u^* v = \langle u, v\rangle,$$
por lo que la aplicación "multiplicar por A" es una isometría para $\|\cdot\|_2$. □

Corolario 1.1

Los autovalores de una matriz simétrica son reales.

♡

1.1.5 Factorización de Schur

Se dice que las matrices $A, B \in \mathcal{M}_n$ son semejantes si existe $P \in \mathbb{M}_n$ invertible, tal que
$$B = P^{-1} A P.$$
A y B son semejantes si y sólo si representan la misma aplicación lineal en dos bases distintas.

Proposición 1.6

Sean $A, B \in \mathbb{M}_n$ matrices semejantes. Se verifica:

- $\det A = \det B$.
- $\operatorname{tr} A = \operatorname{tr} B$.
- $\operatorname{Sp}(A) = \operatorname{Sp}(B)$.

♠

Demostración Se ha visto en asignaturas anteriores. □

La descomposición de Schur muestra que toda matriz es semejante a una triangular superior mediante un cambio de variables unitario.

Teorema 1.2 (Descomposición de Schur)

Sea $A \in \mathcal{M}_n(\mathbb{C})$.

1. *Existe una matriz unitaria U tal que la matriz U^*AU es triangular superior. Además los elementos de la diagonal son los autovalores de A.*
2. *(Teorema de Descomposición espectral) A es normal si y sólo si existe U (unitaria) tal que U^*AU es diagonal.*

♡

Demostración

1) Por inducción.

El resultado es trivial si $n = 1$, así que supondremos $n \geq 2$.

Suponemos cierto hasta $n-1$. Sea λ_1 un autovalor de A y x_1 un autovector no nulo asociado a λ_1 normalizado, es decir: $\|x_1\|_2^2 = x_1^* x_1 = 1$, $Ax_1 = \lambda_1 x_1$. Entonces, podemos extender x_1 hasta una base ortonormal (por ejemplo, con la ortonormalización de Gram-Schmidt), $x_2, \ldots, x_n$ de modo que la matriz X cuyas columnas son $x_1, \ldots x_n$ es unitaria. Es más, existe un vector u y una matriz $A_1 \in \mathcal{M}_{n-1}(\mathbb{C})$ tal que

$$X^*AX = \begin{pmatrix} \lambda_1 & u \\ 0 & A_1 \end{pmatrix}.$$

Por inducción existe U_1 tal que $U_1^* A_1 U_1$ es triangular superior y los elementos de la diagonal son los autovalores de A_1. Tomamos

$$U = \begin{pmatrix} 1 & 0 & \\ 0 & & U_1 \end{pmatrix}.$$

U es unitaria (ejercicio). Entonces

$$\begin{aligned} U^*AU &= \begin{pmatrix} 1 & 0 \\ 0 & U_1^* \end{pmatrix} X^*AX \begin{pmatrix} 1 & 0 \\ 0 & U_1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & U_1^* \end{pmatrix} \begin{pmatrix} \lambda_1 & u \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & U_1 \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & uU_1 \\ 0 & U_1^* A U_1 \end{pmatrix}, \end{aligned}$$

que es triangular superior. Por ser semejantes, A y U^*AU tienen los mismos autovalores, que son los elementos de la diagonal de U^*AU.

2) Si A es normal, sea $R = U^*AU$, donde U es la construida en el apartado anterior y R es triangular superior. Entonces

$$R^*R = (U^*AU)^*(U^*AU) = U^*A^*AU = U^*AA^*U = (U^*AU)(U^*A^*U) = RR^*$$

Luego R es normal. Pero el elemento (1,1) de R^*R es $|\lambda_1|^2$ y el de RR^* es

$$|\lambda_1|^2 + \sum_{k=2}^{n} |r_{1k}|^2.$$

Luego $r_{1k} = 0$ para todo $k \neq 1$. Repitiendo el proceso, se obtiene que es diagonal.

Reciprocamente, si U^*AU es diagonal, D, entonces $A = UDU^*$ y $A^*A = AA^*$.

□

✍ **Ejercicio 1.3** Usando la factorización de Schur, demostrar que la traza A es igual a la suma de los autovalores y que el determinante de A es el producto de los autovalores.

1.1.6 Matriz definida positiva

Una matriz hermítica $A \in \mathcal{M}_n$ es definida positiva (resp. semidefinida positiva) si

$$v^*Av > 0,\ v \in \mathbb{K}^n \backslash \{0\} \quad (\text{resp. } v^*Av \geq 0,\ v \in \mathbb{K}^n \backslash \{0\}).$$

Proposición 1.7

Sea $A \in \mathcal{M}_n$ una matriz hermítica. Se verifica:

1. *A es definida positiva si y sólo si $\mathrm{Sp}(A) \subset \mathbb{R}^+$.*
2. *A es semidefinida positiva si y sólo si $\mathrm{Sp}(A) \subset \mathbb{R}^+ \cup \{0\}$.*

♠

Demostración 1) Sea A def. positiva. Sea $\lambda \in \mathrm{Sp}(A)$ y v un autovector asociado no nulo.

$$0 < v^*Av = v^*\lambda v = \lambda v^*v$$

Como $v^*v > 0$, tenemos que $\lambda > 0$.

Supongamos que A es hermítica y los autovalores son reales positivos. Por el Teorema de Schur existe U unitaria tal que

$$U^*AU = D = \mathrm{diag}([\lambda_1, \ldots, \lambda_n)]$$

Entonces para todo vector v no nulo, se verifica

$$v^*Av = v^*UDU^*v = (U^*v)^*D(U^*v) = w^*Dw = \sum \lambda_i |w_i|^2 > 0.$$

2) Ejercicio. □

Proposición 1.8

Sea $A \in \mathcal{M}_n$. Entonces A^*A es una matriz hermítica y semidefinida positiva.
Si además A es invertible, entonces A^*A es definida positiva.

♠

Demostración Se deja como ejercicio.

□

1.2 Normas matriciales

En esta sección veremos cómo extender el concepto de norma a las aplicaciones lineales. Para complementar esta sección, se puede consultar el libro de Infante y Rey.

1.2.1 Norma matricial

Una norma (vectorial) sobre $\mathcal{M}_n$ es una aplicación,

$$\|\cdot\| \colon \mathcal{M}_n \to \mathbb{R}^+ \cup \{0\}, \quad A \to \|A\|,$$

que verifica:

1. $\|A\| = 0$ si y sólo si $A = 0$.
2. $\|A + B\| \leq \|A\| + \|B\|$, para todo $A, B \in \mathcal{M}_n$.

3. $\|\lambda A\| = |\lambda|\|A\|$, para todo $\lambda \in \mathbb{C}$, $A \in \mathcal{M}_n$.

Decimos que es una norma matricial si además verifica

4. $\|AB\| \leq \|A\|\|B\|$, para todo $A, B \in \mathcal{M}_n$.

Ejemplo 1.1 Sea $A \in \mathcal{M}_n$, $A = (a_{ij})_{1\leq i,j\leq n}$. Podemos definir las siguientes normas (demostraremos más tarde que son normas matriciales):

- $\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|$.
- $\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$.
- $\|A\|_2 = \rho(A^*A)^{1/2}$.

  Sin embargo

$$\|A\| = \max_{1\leq i,j\leq n} |a_{ij}|$$

no es una norma matricial (ejercicio).

1.2.2 Normas matriciales compatibles e inducidas

Sea $\|\cdot\|_v$ una norma vectorial en $\mathbb{K}^n$ y $\|\cdot\|_M$ una norma matricial en $\mathcal{M}_n$.

Decimos que $\|\cdot\|_M$ es una norma matricial compatible si

$$\|Av\|_v \leq \|A\|_M \|v\|_v, \quad \forall v \in \mathbb{K}^n,\ A \in \mathcal{M}_n.$$

Decimos que $\|\cdot\|_M$ es una norma matricial inducida si

$$\|A\|_M = \max_{v\neq 0} \frac{\|Av\|_v}{\|v\|_v} = \max_{\|v\|_v=1} \|Av\|_v, \quad \forall A \in \mathcal{M}_n.$$

Si $\|\cdot\|_M$ es una norma matricial inducida entonces $\|I\|_M = 1$.

Ejemplo 1.2 Las tres normas que más utilizaremos, $\|\cdot\|_1, \|\cdot\|_\infty, \|\cdot\|_2$, son inducidas (lo demostraremos después):

- $\|A\|_1$ inducida por $\|v\|_1 = \sum_{i=1}^n |v_i|$, $v \in \mathbb{K}^n$.
- $\|A\|_\infty$ inducida por $\|v\|_\infty = \max_{i=1}^n |v_i|$, $v \in \mathbb{K}^n$.
- $\|A\|_2$ inducida por $\|v\|_2 = \sqrt{\sum_{i=1}^n |v_i|^2}$, $v \in \mathbb{K}^n$.
- La norma de Frobenius:

$$\|A\| = \left(\sum_{1\leq i,j\leq n} |a_{ij}|^2 \right)^{1/2} = \operatorname{tr}(A^*A)^{1/2}$$

  no es una norma inducida.

  Es una norma matricial, pues es una norma vectorial (la norma euclídea, considerando las matrices como vectores) y cumple:

$$\|AB\|^2 = \sum_{1\leq i,j\leq n} |\sum_{k=1}^n a_{ik}b_{kj}|^2 \overset{(Des.Cauchy-Schwarz)}{\leq} \sum_{1\leq i,j\leq n} \sum_{k=1}^n |a_{ik}|^2 \sum_{k=1}^n |b_{kj}|^2$$

$$= \left(\sum_{1\leq i,k\leq n} |a_{ik}|^2 \right) \left(\sum_{1\leq k,j\leq n} |b_{kj}|^2 \right) = \|A\|^2\|B\|^2.$$

  No es inducida pues $\|I\| = \sqrt{n} \neq 1$, y para toda norma inducida, $\|I\|_M = 1$.

  Es compatible con la norma euclídea (ejercicio).

Proposición 1.9

Si $\|\cdot\|_v$ es una norma vectorial sobre $\mathbb{K}^n$, entonces la norma inducida

$$\|A\|_M := \max_{\|v\|_v=1} \|Av\|_v$$

es una norma matricial, compatible con $\|\cdot\|_v$.

Demostración En primer lugar, está bien definida pues $x \to \|Ax\|_v$ es una función continua, luego alcanza su máximo en el compacto $\{\|v\|_v = 1\}$. Además, ambas normas son compatibles

$$\|A\|_M = \max_{v\neq 0} \frac{\|Av\|_v}{\|v\|_v} \geq \frac{\|Av\|_v}{\|v\|_v}, \forall\, v \neq 0,$$

luego $\|Av\|_v \leq \|A\|_M \|v\|_v$ para todo $v \neq 0$.

Veamos que es una norma matricial. Sean $A, B \in \mathcal{M}_n$.

1. $\|A\|_M = 0$ si y sólo si $\max \|Av\|_v/\|v\|_v = 0$ si y sólo si $\|Av\|_v = 0$ si y sólo si $Av = 0$ sii $A = 0$.
2. Sea u de norma 1 donde se alcanza el supremo de $\|(A+B)v\|_v$

$$\|A+B\|_M = \|(A+B)u\|_v \leq \|Au\|_v + \|Bu\|_v \leq \|A\|_M\|u\|_v + \|B\|_M\|u\|_v = \|A\|_M + \|B\|_M.$$

3. Sea $\lambda \in \mathbb{C}$. Entonces

$$\|\lambda A\|_M = \max \|\lambda Av\|_v = |\lambda| \max \|Av\|_v = |\lambda|\|A\|_M$$

4. Sea u de norma 1 donde se alcanza el supremo de $\|(AB)v\|_v$

$$\|AB\|_M = \|(AB)u\|_v \leq \|A\|_M\|Bu\|_v \leq \|A\|_M\|B\|_M\|u\|_v = \|A\|_M\|B\|_M.$$

□

1.2.3 Normas matriciales inducidas por las vectoriales

Proposición 1.10

1. *La norma inducida por $\|v\|_1 = \sum_{i=1}^n |v_i|$, $v \in \mathbb{K}^n$, es*

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|.$$

2. *La norma inducida por $\|v\|_\infty = \max_{1\leq i\leq n} |v_i|$, $v \in \mathbb{K}^n$, es*

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|.$$

3. *La norma inducida por $\|v\|_2 = \sqrt{\sum_{i=1}^n |v_i|^2}$, $v \in \mathbb{K}^n$, es*

$$\|A\|_2 = \rho(A^*A)^{1/2}.$$

Demostración

1) Sea A una matriz. Sea y de norma 1 donde se alcance $\|A\|_1$,

$$\|A\|_1 = \|Ay\|_1 = \sum_i |\sum_{j=1}^n a_{ij}y_j| \leq \sum_i \sum_j |a_{ij}y_j| = \sum_j |y_j| \sum_i |a_{ij}|$$

$$\leq \sum_j |y_j| (\max_k \sum_i |a_{ik}|) = (\max_k \sum_i |a_{ik}|)\|y\|_1 \leq \max_k \sum_i |a_{ik}|.$$

Por otra parte, supongamos que el $\max_k \sum_i |a_{ik}|$ se alcanza en el índice k_0 y consideremos el vector e_{k_0}

$$\|A\|_1 \geq \|Ae_{k_0}\|_1 = \sum_i |a_{ik_0}| = \max_{1\leq k\leq n} \sum_i |a_{ik}|$$

2) Sea A una matriz. Sea y de norma 1 donde se alcance $\|A\|_\infty$

$$\|A\|_\infty = \|Ay\|_\infty = \max_i |\sum_{j=1}^{n} a_{ij}y_j| \leq \max_i \sum_j |a_{ij}||y_j|$$

$$\leq \max_j |y_j| \max_i \sum_j |a_{ij}| = \|y\|_\infty \max_i \sum_j |a_{ij}| = \max_i \sum_j |a_{ij}|$$

Por otra parte, supongamos que el $\max_i \sum_j |a_{ij}|$ se alcanza en el índice I. Consideremos el vector v definido por $v_j = \bar{a}_{Ij}/|a_{Ij}|$ si $a_{Ij} \neq 0$, $v_j = 0$ en caso contrario. Entonces $\|v\|_\infty = 1$, $\|Av\|_\infty \geq \sum_j |a_{Ij}|$

$$\|Av\|_\infty = \max_i |\sum_j a_{ij}v_j| \leq \max_i \sum_j |a_{ij}||v_j| \leq \max_i \sum_j |a_{ij}|.$$

Luego $\|Av\|_\infty =^* \sum_j |a_{Ij}| \leq \|A\|_\infty$.

3) Sabemos que los autovalores de A^*A son todos reales positivos o nulos. Además, por ser hermítica, el Teorema de Schur implica que existe una base ortonormal $\{u_1, \ldots, u_n\}$ formada por autovectores de A^*A (U). Sea y de norma 1 donde se alcanza $\max_{\|v\|=1} \|Av\|_2^2$. Si $y = \sum \alpha_i u_i$,

$$1 = \|y\|_2^2 = y^*y = (\sum \bar{\alpha}_i u_i^*)(\sum \alpha_j u_j) = \sum_{ij} \bar{\alpha}_i \alpha_j u_i^* u_j$$

$$= \sum_i \bar{\alpha}_i \alpha_i = \sum |\alpha_i|^2$$

Además

$$\|Ay\|_2^2 = (Ay)^*(Ay) = y^*A^*Ay = \sum_i \bar{\alpha}_i u_i^*(A^*A) \sum_j \alpha_j u_j = \sum_{ij} \bar{\alpha}_i \alpha_j u_i^*(A^*A)u_j =$$

$$= \sum_{ij} \bar{\alpha}_i \alpha_j u_i^* \lambda_j u_j = \sum \lambda_i |\alpha_i|^2 \leq \rho(A^*A) \sum |\alpha_i|^2 = \rho(A^*A).$$

Por otra parte, Sea λ_I el autovalor que da el radio espectral y u_I un autovector de norma 1 asociado.

$$\max_{\|v\|=1} \|Av\|_2^2 \geq \|Au_I\|_2^2 = (Au_I)^*(Au_I) = u_I^*(A^*A)u_I = u_I^*\lambda_I u_I$$

$$= \lambda_I |u_I|^2 = \lambda_I = \rho(A^*A).$$

□

1.2.4 Error y condicionamiento

Sea $\tilde{x}$ una solución aproximada de $Ax = b$. Definimos el *vector de error* como

$$x_\delta = \tilde{x} - x,$$

donde x es la solución exacta del problema.

Definimos el *vector de error residual* como

$$b_\delta = A\tilde{x} - b.$$

Nótese que $Ax_\delta = b_\delta$.

Fijada una norma, definimos los errores relativos asociados a los errores anteriores como

$$E = \frac{\|x_\delta\|}{\|x\|}, \quad R = \frac{\|b_\delta\|}{\|b\|}.$$

Consideremos una norma matricial inducida por una vectorial (denotaremos ambas como $\|\cdot\|$).

Sea $A \in \mathcal{M}_n$. Denominamos *condicionamiento* de A respecto a la norma $\|\cdot\|$ a

$$\text{cond}(A) := \|A\|\|A^{-1}\|.$$

Proposición 1.11

Sea $\|\cdot\|$ una norma matricial inducida y $A \in \mathcal{M}_n$ una matriz invertible. Se verifican las siguientes propiedades:

1. $\text{cond}(A) \geq 1$.
2. $\text{cond}(A) = \text{cond}(A^{-1})$.
3. $\text{cond}(\lambda A) = \text{cond}(A)$ *para todo* $\lambda \in \mathbb{R}\backslash\{0\}$.

♠

Demostración 1. $1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\|$.

2 y 3 triviales (ejercicio). □

Teorema 1.3

Sean $b, b_\delta \in \mathbb{R}^n$ no idénticamente nulos. Denotemos x y $x + x_\delta$ a las soluciones respectivas de los sistemas lineales

$$Ax = b \quad A(x + x_\delta) = b + b_\delta.$$

Entonces se verifica

$$\frac{1}{\text{cond}(A)} \frac{\|b_\delta\|}{\|b\|} \leq \frac{\|x_\delta\|}{\|x\|} \leq \text{cond}(A) \frac{\|b_\delta\|}{\|b\|}$$

Además, para toda matriz A invertible, existen $b, b_\delta \in \mathbb{R}^n$ no idénticamente nulos tal que las desigualdades se alcanzan.

♡

Demostración De $A(x + x_\delta) = b + b_\delta$, tenemos que $x_\delta = A^{-1}b_\delta$.

Entonces $\|x_\delta\| \leq \|A^{-1}\|\|b_\delta\|$

Por otra parte, $\|b\| \leq \|A\|\|x\|$, entonces $1/\|x\| \leq \|A\|/\|b\|$.

Juntando ambas desigualdades tenemos la que queremos.

Por otra parte, por definición de norma inducida, existe x tal que $\|Ax\| = \|A\|\|x\|$. Definimos $b = Ax$ para este x.

De nuevo por definición de norma inducida, existe b_δ tal que $\|A^{-1}b_\delta\| = \|A^{-1}\|\|b_\delta\|$.

Para los sistemas lineales $Ax = b$ y $A(x + x_\delta) = b + b_\delta$, tenemos que $x_\delta = A^{-1}b_\delta$, luego $\|x_\delta\| = \|A^{-1}\|\|b_\delta\|$ y $\|A\|\|x\| = \|b\|$ y de ahí se sigue la igualdad.

□

Teorema 1.4 (ver [1, p. 62])

Sean $b, b_\delta \in \mathbb{R}^n$ no identicamente nulos, $A, A_\delta \in \mathcal{M}_n$ tales que A y $A + A_\delta$ son matrices invertibles y denotemos x y $x + x_\delta$ a las soluciones respectivas de

$$Ax = b, \quad (A + A_\delta)(x + x_\delta) = b + b_\delta.$$

Entonces se verifica

$$\frac{\|x_\delta\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A)\|A_\delta\|/\|A\|} \left(\frac{\|b_\delta\|}{\|b\|} + \frac{\|A_\delta\|}{\|A\|} \right)$$

♡

1.3 Factorización de matrices

El objetivo es resolver el sistema de ecuaciones lineales

$$\mathbf{Ax} = \mathbf{b},$$

donde $\mathbf{A} \in \mathcal{M}_n(\mathbb{K})$, $\mathbf{b} \in \mathbb{K}^n$ son conocidos.

1.3.1 Orden de una sucesión

Decimos que una sucesión $\{x_n\}$ es del orden menor o igual que otra sucesión $\{y_n\}$ si existen constantes C, N tales que $|x_n| < C|y_n|$ para $n > N$ y lo denotamos

$$\{x_n\} = \mathcal{O}(\{y_n\}) \quad (\{x_n\} \in \mathcal{O}(\{y_n\})).$$

Decimos que dos sucesiones $\{x_n\}, \{y_n\}$ son del mismo orden, y lo denotamos $\{x_n\} = \Theta(\{y_n\})$ si

$$\{x_n\} = \mathcal{O}(\{y_n\}), \quad \{y_n\} = \mathcal{O}(\{x_n\}).$$

Decimos que el orden de $\{x_n\}$ es estrictamente menor que el de $\{y_n\}$ si $\lim_{n\to\infty} x_n/y_n = 0$ y lo denotamos $\{x_n\} = o(\{y_n\})$.

Ejemplo 1.3 Es fácil comprobar que

$$\{n\}, \{100n\}, \{n+1\}, \{n^2+n\} = \mathcal{O}(\{n^2\}).$$

Sin embargo $\{n^2\} \neq \mathcal{O}(\{n\})$. Es más, si p es un polinomio de grado k, entonces

$$p(n) = \Theta(n^k).$$

✍ **Ejercicio 1.4** Probar que si $x_n = 2x_{n-1} + 1$, $x_0 = 1$, entonces

$$\{x_n\} = \mathcal{O}(2^n).$$

✍ **Ejercicio 1.5** Probar que si x_n es el número de cifras en base 10 de n, entonces

$$\{x_n\} = \mathcal{O}(\ln n).$$

1.3.2 Sistemas fáciles de resolver

Consideremos que tenemos el sistema de ecuaciones lineales

$$\mathbf{Ax} = \mathbf{b},$$

donde $\mathbf{A} \in \mathcal{M}_n(\mathbb{K})$, $\mathbf{b} \in \mathbb{K}^n$ son conocidos.

Para ciertas matrices $\mathbf{A}$ el sistema es fácil de resolver:

1. $\mathbf{A}$ diagonal. El número de operaciones está en $\mathcal{O}(n)$.
2. $\mathbf{A}$ triangular superior. Por sustitución regresiva. El número de operaciones está en $\mathcal{O}(n^2)$.
3. $\mathbf{A}$ triangular inferior. Por sustitución progresiva. El número de operaciones está en $\mathcal{O}(n^2)$.
4. $\mathbf{A}$ ortogonal o unitaria. Multiplicando por su traspuesta o traspuesta conjugada. El número de operaciones está en $\mathcal{O}(n^2)$.

1.3.3 Factorización de matrices

Supongamos que la matriz $\mathbf{A}$ factoriza como producto de varias matrices

$$\mathbf{A} = \mathbf{M_1 M_2} \dots \mathbf{M_k},$$

de modo que las matrices $\mathbf{M_i}$ se correspondan con matrices de sistemas "fáciles de resolver".

Entonces, podemos resolver el sistema recursivamente:

$$\mathbf{M_1 y_1 = b}, \mathbf{M_2 y_2 = y_1}, \dots \mathbf{M_k y_k = y_{k-1}},$$

y tendremos que la solución es $\mathbf{x = y_k}$.

Por ejemplo, si $\mathbf{A = M_1 M_2}$, basta resolver $\mathbf{M_1 y = b}$ y $\mathbf{M_2 x = y}$.

✍ **Ejercicio 1.6**

Resolver el sistema $Ax = b$, con $b = (1, 1, 0)$ y

$$A = \begin{pmatrix} 2 & -1 & \sqrt{2} \\ -1 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

sabiendo que su factorización de Schur es $A = URU^*$ con

$$U = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ -\sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R = \begin{pmatrix} 3 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

1.3.4 Factorización LU

Decimos que una matriz invertible $\mathbf{A}$ admite una factorización LU si se puede escribir en la forma:

$$\mathbf{A = LU}$$

donde

- $\mathbf{L}$ es una matriz triangular inferior ($l_{ij} = 0$ si $i < j$).
- $\mathbf{U}$ es una matriz triangular superior ($u_{ij} = 0$ si $i > j$).

Si conocemos una factorización LU de una matriz $\mathbf{A}$, $\mathbf{A = LU}$, podemos resolver el sistema $\mathbf{Ax = b}$ con el siguiente procedimiento:

- Resolvermos mediante sustitución progresiva el sistema
$$\mathbf{Ly = b}.$$
- Obtenemos la solución $\mathbf{x}$ resolviendo por sustitución regresiva el sistema
$$\mathbf{Ux = y}.$$

También es útil para calcular el determinante, inversas, etc.

Obtener una factorización LU de una matriz $\mathbf{A}$ es equivalente a resolver el sistema

$$\mathbf{A = LU}$$

donde los elementos de $\mathbf{L}$ y $\mathbf{U}$ son las incógnitas.

Tenemos n^2 ecuaciones y $n^2 + n$ incógnitas.

Debemos fijar n condiciones. Algunas posibilidades:

- Factorización de Doolittle, si la diagonal de L está formada por unos.
- Factorización de Crout, si la diagonal de U está formada por unos.

Por omisión, se considerará que la factorización LU es la de Doolittle.

Teorema 1.5

Si los n menores principales de la matriz $A \in \mathcal{M}_n$ son no singulares, entonces la matriz A admite una factorización LU. Recíprocamente, todos los menores principales de una matriz de la forma LU son no singulares siempre que L y U sean (triangulares e) invertibles. ♡

Demostración Supongamos que $A = LU$. Particionando adecuadamente las tres matrices, tenemos

$$\begin{pmatrix} A_{n-1} & a^{n-1} \\ a_{n-1} & a_{nn} \end{pmatrix} = \begin{pmatrix} L_{n-1} & 0^{n-1} \\ l_{n-1} & l_{nn} \end{pmatrix} \cdot \begin{pmatrix} U_{n-1} & u^{n-1} \\ 0_{n-1} & u_{nn} \end{pmatrix} = \begin{pmatrix} L_{n-1}U_{n-1} & L_{n-1}u^{n-1} \\ l_{n-1}U_{n-1} & l_{n-1}u^{n-1} + l_{nn}u_{nn} \end{pmatrix},$$

por lo que A_{n-1} es no singular.

Veamos que si todos los menores son no singulares entonces A admite una factorización de Doolittle.

Por inducción, el menor principal de orden $n-1$ admite una factorización $A_{n-1} = L_{n-1}U_{n-1}$, donde L_{n-1} y U_{n-1} son no singulares.

Ahora planteamos el sistema de ecuaciones $L_{n-1}(u_{1,n} \dots u_{n-1,n}) = (a_{1,n} \dots a_{n-1,n})$, que es resoluble porque L_{n-1} es no singular.

Lo mismo para $(l_{n,1} \dots l_{n,n-1})U_{n-1} = (a_{n,1} \dots a_{n,n-1})$.

Por último $a_{n,n} = \sum_{s=1}^{n-1} l_{n,s}u_{s,n} + l_{n,n}u_{n,n}$ y de ahí despejamos $u_{n,n}$ (tomamos $l_{n,n} = 1$). □

1.3.5 Factorización LU con pivoteo

Dada $\sigma \in S_n$, denominamos **matriz de permutaciones asociada a** σ a la matriz

$$P = \begin{pmatrix} e_{\sigma(1)} \\ \dots \\ e_{\sigma(n)} \end{pmatrix}.$$

Para toda matriz $A \in \mathcal{M}_n$, $\mathbf{PA}$ permuta por σ las filas de $\mathbf{A}$.

Lema 1.1

Sea P una matriz de permutaciones. Entonces en cada fila y en cada columna de P hay un único 1 y el resto de posiciones contienen ceros. Además, $\det(P) = \operatorname{sig}(\sigma)$ y

$$PP^t = P^tP = I.$$

♡

Demostración Basta escribir P como producto de matrices correspondientes a trasposiciones de filas. Su inversa es ella misma y al trasponer intercambiamos los productos. □

Definición 1.1

Decimos que $A \in \mathcal{M}_n$ admite una factorización LU con pivoteo si existen

1. *una matriz de permutación P,*
2. *una matriz triangular inferior $L = (l_{ij})_{1 \le i,j \le n}$, con $l_{ii} = 1$, $1 \le i \le n$ y $|l_{ij}| \le 1$ $1 \le i, j \le n$,*
3. *una matriz triangular superior U invertible*

tales que $PA = LU$.

♣

Si A admite una factorización LU con pivoteo entonces A factoriza como $A = P^tLU$.

Vamos a probar que siempre que la matriz sea invertible, existe dicha factorización.

Teorema 1.6

Toda matriz $A \in \mathcal{M}_n$ invertible admite una factorización LU con pivoteo. ♡

Demostración

Decimos que P_k, L_k, U_k son una factorización LU con pivoteo de A hasta el paso k si $P_kA = L_kU_k$ y

1. P_k es una matriz de permutaciones.
2. L_k es una matriz triangular inferior tal que $L_k = (l_{ij}^{(k)})$, $l_{ii}^{(k)} = 1$, $1 \leq i \leq n$, $|l_{ij}^{(k)}| \leq 1$, $1 \leq i, j \leq n$, y $l_{ij}^{(k)} = 0$ para $i > k$, $i \neq j$.
3. U_k es "triangular superior hasta la fila k" tal que $U_k = (u_{ij}^{(k)})_{1 \leq i,j \leq n}$ y $u_{ij}^{(k)} = 0$ si $j < i \leq k$.

Sea k el mayor valor para el que tenemos la factorización anterior. Veamos que $k = n$. Supongamos que $k < n$.

En primer lugar, si $u_{jk}^{k} = 0$ para todo $j \geq k$, es fácil probar que $det(U_k) = 0$ y llegamos a contradicción.

Sea l tal que $|u_{lk}^{(k)}| = \text{máx}\{|u_{jk}^{(k)}| : j \geq k\}$. Entonces, si P_{kl} es la matriz que permuta las filas k y l, tenemos que

$$P_{kl}P_kA = (P_{kl}L_kP_{kl})(P_{kl}U_k).$$

Si denotamos

$$P_{k+1} = P_{kl}P_k, \quad \tilde{L}_k = (P_{kl}L_kP_{kl}), \quad \tilde{U}_k = (P_{kl}U_k)$$

entonces $P_{k+1}A = \tilde{L}_k\tilde{U}_k$ y $P_{k+1}, \tilde{L}_k, \tilde{U}_k$ es una factorización LU de $P^{k+1}A$ hasta el paso k, pero ahora $|\tilde{u}_{kk}^{k}| = \text{máx}\{|\tilde{u}_{jk}^{k}| : j \geq k\}$. Entonces, por la factorización LU usual, existen matrices L_{k+1} y U_{k+1}, tales que si $P_{k+1} = P_{kl}P_k$, se verifica que $P_{k+1}A = L_{k+1}U_{k+1}$ con las propiedades buscadas, luego $k = n$.

□

1.3.5.1 Algoritmo para obtener la factorización LU con pivoteo

Sea $A \in \mathcal{M}_n$. Definimos las matrices $P^{(k)} = (p_{ij}^{(k)}), L^{(k)} = (l_{ij}^{(k)}), U^{(k)} = (u_{ij}^{(k)})$, $k = 0, 1, \ldots, n-1$, determinadas de la siguiente manera:

$$P^{(0)} = L^{(0)} = Id, \quad U^{(0)} = A.$$

Para cada $k = 1, \ldots, n-1$

1. Sea $\tau = (k, s)$, $s \geq k$ de modo que $|u_{i,k}^{k-1}| \leq |u_{s,k}^{k-1}|$.
2. Definimos $P^{(k)}$ como la matriz obtenida de $P^{(k-1)}$ permutando las filas k y s.
3. Definimos $\tilde{L}^{(k)} = (\tilde{l}_{ij}^{(k)})$ como la matriz obtenida de $L^{(k-1)}$ permutando las filas k y s, excepto los elementos de la diagonal.
4. Definimos $l_{ik}^{(k)} = \tilde{u}_{ik}^{(k)}/\tilde{u}_{kk}^{(k)}$, $j = k+1, \ldots, n$ y $l_{ij}^{(k)} = \tilde{l}_{ij}^{(k)}$ para el resto.
5. Definimos

$$u_{ij}^{(k)} = \tilde{u}_{ij}^{(k)} - \frac{\tilde{u}_{ik}^{(k)}}{\tilde{u}_{kk}^{(k)}}\tilde{u}_{kj}^{(k)}, \quad k+1 \leq i \leq n, \ k \leq j \leq n.$$

y $u_{ij}^{(k)} = \tilde{u}_{ij}^{(k)}$ para el resto.

La factorización buscada es $P = P^{(n-1)}$, $L = L^{(n-1)}$, $U = U^{(n-1)}$.

Ejemplo 1.4

Tomemos la matriz:

$$\mathbf{A} = \begin{pmatrix} 5 & -1 & -1 & 1 \\ -2 & 1 & 1 & 2 \\ 18 & 1 & 2 & 3 \\ 5 & -2 & 1 & 4 \end{pmatrix}$$

Tomamos para el primer paso las matrices:

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathbf{U} = \begin{pmatrix} 5 & -1 & -1 & 1 \\ -2 & 1 & 1 & 2 \\ 18 & 1 & 2 & 3 \\ 5 & -2 & 1 & 4 \end{pmatrix}$$

Permutamos la fila 1 y 3 de $\mathbf{U}$ y de $\mathbf{P}$

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathbf{U} = \begin{pmatrix} 18 & 1 & 2 & 3 \\ -2 & 1 & 1 & 2 \\ 5 & -1 & -1 & 1 \\ 5 & -2 & 1 & 4 \end{pmatrix}$$

Ahora hacemos cero debajo de la diagonal y guardamos los multiplicadores en $\mathbf{L}$

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{9} & 1 & 0 & 0 \\ \frac{5}{18} & 0 & 1 & 0 \\ \frac{5}{18} & 0 & 0 & 1 \end{pmatrix}, \mathbf{U} = \begin{pmatrix} 18 & 1 & 2 & 3 \\ 0 & \frac{10}{9} & \frac{11}{9} & \frac{7}{3} \\ 0 & -\frac{23}{18} & -\frac{14}{9} & \frac{1}{6} \\ 0 & -\frac{41}{18} & \frac{4}{9} & \frac{19}{6} \end{pmatrix}$$

Permutamos las filas 2 y 4 de $\mathbf{U}$ y de $\mathbf{P}$ y las de $\mathbf{L}$ pero omitiendo la diagonal.

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{5}{18} & 1 & 0 & 0 \\ \frac{5}{18} & 0 & 1 & 0 \\ -\frac{1}{9} & 0 & 0 & 1 \end{pmatrix}, \mathbf{U} = \begin{pmatrix} 18 & 1 & 2 & 3 \\ 0 & -\frac{41}{18} & \frac{4}{9} & \frac{19}{6} \\ 0 & -\frac{23}{18} & -\frac{14}{9} & \frac{1}{6} \\ 0 & \frac{10}{9} & \frac{11}{9} & \frac{7}{3} \end{pmatrix}$$

Finalmente, hacemos cero debajo de la diagonal ($\mathbf{P}$ permanece inalterado)

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{5}{18} & 1 & 0 & 0 \\ \frac{5}{18} & \frac{23}{41} & 1 & 0 \\ -\frac{1}{9} & -\frac{20}{41} & -\frac{59}{74} & 1 \end{pmatrix}, \mathbf{U} = \begin{pmatrix} 18 & 1 & 2 & 3 \\ 0 & -\frac{41}{18} & \frac{4}{9} & \frac{19}{6} \\ 0 & 0 & -\frac{74}{41} & -\frac{66}{41} \\ 0 & 0 & 0 & \frac{96}{37} \end{pmatrix}$$

✍ **Ejercicio 1.7**

Obtener la factorización LU con pivote de

$$\mathbf{A} = \begin{pmatrix} 1/2 & 5 & 0 \\ 1 & 6 & -6 \\ 1/3 & 5 & 0 \end{pmatrix}$$

1.3.6 Factorización de Cholesky

Sea $A \in \mathcal{M}_n(\mathbb{R})$. Decimos que A admite una *factorización LU de Cholesky* si existe una matriz **real** triangular inferior $\mathbf{L}$ tal que los elementos de su diagonal son positivos y

$$\mathbf{A} = \mathbf{L}\mathbf{L}^\top.$$

Teorema 1.7

Una matriz $\mathbf{A}$ *real es simétrica y definida positiva si y sólo si admite una factorización LU de Choleski.* ♡

Demostración Supongamos que A es simétrica y definida positiva. En particular, los menores principales han de ser no nulos. Por lo que admite una factorización LU.

Por ser simétrica, veamos que admite una factorización LDL^t:

$$LU = A = A^t = (LU)^t = U^t L^t$$

Como L es invertible, $U = L^{-1}U^t L^t$. Luego $U(L^t)^{-1} = L^{-1}U^t$, pero entonces ha de ser triangular inferior y superior, luego es diagonal. Sea

$$D = U(L^t)^{-1} = L^{-1}U^t.$$

Es más, $U = DL^t$, luego $A = LDL^t$.

Como A es definida positiva, tenemos que D también ha de serlo (escribir $D = L^{-1}A(L^{-1})^t$ y aplicar def. pos de A a e_k) Entonces podemos definir $D^{1/2}$ y $\tilde{L} = LD^{1/2}$. Tenemos $A = \tilde{L}(\tilde{L})^t$

Recíprocamente, si $A = LL^t$, entonces,

$$v^t Av = v^t LL^t v = \|v^t L\|_2^2.$$

□

Ejemplo 1.5

Consideremos la matriz:

$$\mathbf{A} = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 2 & -4 \\ -2 & -4 & 11 \end{pmatrix}.$$

Como es simétrica, calculamos su factorización:

$$\mathbf{L} = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -3 & 1 \end{pmatrix}, \ \mathbf{L}^t = \begin{pmatrix} 2 & 1 & -1 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{pmatrix}.$$

(Luego es definida positiva.)

Tema 1 Ejercicios

1. Encontrar ejemplos de matrices A tales que
   (a). A no sea normal.
   (b). A sea normal pero no unitaria.
   (c). A sea unitaria pero no hermítica.
2. 👥 Demostrar que el producto de dos matrices triangulares superiores es una matriz triangular superior. Demostrar que si A es una matriz triangular superior e invertible, entonces su inversa es triangular superior.
3. 👥 Sea A una matriz subdividida en bloques de la forma
$$A = \begin{pmatrix} B & C \\ 0 & I \end{pmatrix},$$
donde los bloques son de $n \times n$. Demuestre que si $B - I$ es no singular, entonces para $k \geq 1$,
$$A^k = \begin{pmatrix} B^k & (B^k - I)(B - I)^{-1}C \\ 0 & I \end{pmatrix},$$
4. Calcular la norma uno, infinito y euclídea de las siguientes matrices:
$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ -1 & 0 & -1 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix}.$$
5. Calcular la norma dos de la siguiente matriz. Calcular los valores $\alpha \in [0, 2\pi]$ para los que la norma uno e infinito sea máxima.
$$A = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}.$$
6. 💻 Consideremos la matriz
$$A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}.$$
   (a). Representar en una gráfica la imagen de la bola unidad (con la norma euclídea) por la aplicación lineal dada por la matriz. En la misma gráfica, representar la imagen de los vectores $(1, 0)$ y $(0, 1)$.
   (b). En la gráfica anterior, representar la imagen de los autovectores de A^*A. Utilizar la orden arrow.
   (c). Añadir a la gráfica la circunferencia de centro el origen y radio $\sqrt{\rho(A^*A)}$.
   (d). Considerar las matrices
$$B = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad C = \begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix}.$$
   Calcular sus autovalores y tratar de deducir cómo será la imagen de la bola unidad. Repetir el proceso anterior con estas matrices.
7. 👥 Demostrar que la norma matricial (norma de Fröbenius)
$$\|A\| = \left(\sum_{1 \leq i,j \leq n} |a_{ij}|^2 \right)^{1/2}, \quad (A = \{a_{ij}\} \in \mathcal{M}_n),$$
no está inducida por ninguna norma vectorial, pero es compatible con la norma vectorial euclídea (utilizar la desigualdad de Cauchy-Schwarz).

8. 👥 Demostrar que la norma matricial
$$\|A\| = \sum_{1\le i,j\le n} |a_{ij}|, \quad (A = \{a_{ij}\} \in \mathcal{M}_n),$$
no está inducida por ninguna norma vectorial, pero es compatible con la norma vectorial $\|\cdot\|_1$.
9. Demostrar que $\|A\| = \max_{1\le i,j\le n} |a_{ij}|$ no es una norma matricial.
10. 👥 Sea $A \in \mathcal{M}_n$. Se verifica
   (a). Si $\|\cdot\|$ es una norma matricial inducida tal que $\|A\| < 1$, entonces la matriz $I + A$ es invertible y se tiene que
$$\frac{1}{1+\|A\|} \le \|(I+A)^{-1}\| \le \frac{1}{1-\|A\|}.$$
   (b). Si una matriz de la forma $I + A$ es singular, entonces necesariamente $\|A\| \ge 1$ para cualquier norma matricial (inducida o no).
11. 👥 Sea A una matriz hermítica tal que $\Delta_k > 0$, $1 \le k \le n$. Entonces A es definida positiva.
12. 👥 Sea $A \in \mathcal{M}_n$ una matriz real, simétrica y definida positiva. Entonces
   (a). $a_{ii} > 0$ para todo $1 \le i \le n$.
   (b). $\max_{1\le i,j\le n} a_{ij} = \max_{1\le i\le n} a_{ii}$.
13. 💻 Consideremos las matrices
$$A = \begin{pmatrix} 1 & 0 \\ -1 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ -2 & 2 \end{pmatrix},$$
   (a). Calcular la imagen de la bola unidad con norma euclídea por las aplicaciones lineales definidas por dichas matrices.
   (b). Calcular la norma 2 de ambas matrices.
   (c). Obtener las direcciones en las que la imagen de la bola unidad alcanza el máximo y el mínimo.
14. 👥 Demostrar que para cualquier matriz no singular A y cualquier norma matricial $\|\cdot\|$,
$$\|I\| \ge 1, \quad \|A^{-1}\| \ge \frac{1}{\|A\|}.$$
15. 👥 Demostrar que el producto de matrices ortogonales es una matriz ortogonal y que el determinante de una matriz ortogonal vale ± 1.
16. 👥 Sea $\{v_1, \ldots, v_n\}$ una base de vectores ortonormales de $\mathbb{R}^n$. Dado $v \in \mathbb{C}$, obtener en términos de $v_1, \ldots, v_n$ y v, los coeficientes $c_k \in \mathbb{R}$ tales que
$$v = \sum_{i=1}^{n} c_i v_i.$$
17. 👥 Demostrar (usando la norma euclídea matricial)
   (a). $\rho^{1/2}\left((A^* + B^*)(A+B)\right) \le \rho^{1/2}(A^*A) + \rho^{1/2}(B^*B)$.
   (b). $\rho\left((AB)^*(AB)\right) \le \rho(A^*A)\rho(B^*B)$.
18. 👥 Sea $A \in \mathcal{M}_n$ invertible y $\|\cdot\|$ una norma matricial inducida. Probar
   (a). $\mathrm{cond}(A) \ge 1$
   (b). $\mathrm{cond}(A) = \mathrm{cond}(A^{-1})$
   (c). $\mathrm{cond}(\lambda A) = \mathrm{cond}(A)$, para todo $\lambda \in \mathbb{K}$, $\lambda \ne 0$.
19. 💻 Consideramos el sistema
$$\begin{pmatrix} 1 & 1/30 \\ 1 & 1/31 \end{pmatrix} x = \left(\frac{31}{30}, \frac{32}{31}\right).$$

Supongamos que el término independiente se ha obtenido midiendo con un error en cada coordenada de $\pm 0{,}01$. Acotar el error en las soluciones.

20. 💻 Sea
$$A = \begin{pmatrix} 4 & -1 & 0 & 1 \\ -1 & 4 & 0 & 0 \\ 1 & 0 & 5 & -1 \\ -1 & 1 & 1 & 4 \end{pmatrix}.$$
   (a). Calcula las normas 1,2 e infinito de la matriz A.
   (b). Calcula un vector unitario con la norma 1, u, tal que $\|A\|_1 = \|A\|_1\|u\|_1$.
   (c). Idem para las normas 2 e ∞.
   (d). Calcular los autovectores unitarios (usar la orden de sage *eigenvectors_right*) con la norma euclídea de A^tA. Calcular su imagen por A. Comprobar que tanto los autovectores como sus imágenes son ortogonales, es decir, forman una base. Calcular $\min_{\|u\|_2=1} \|Au\|_2$. Repetir el proceso cambiando A por A^{-1}. ¿Qué relación hay entre la $1/\|A^{-1}\|_2$ y $\min_{\|u\|_2=1} \|Au\|_2$?
   (e). Calcular $\min_{\|u\|_1=1} \|Au\|_2$.
   (f). Idem para la norma infinito.
21. 💻 La matriz de Hilbert de $n \times n$ está definida como
$$H_n = \left(\frac{1}{i+j-1}\right)_{i,j=1,\dots,n}.$$
Son matrices típicamente mal condicionadas.
   (a). Calcula los números de condición de la matriz de Hilbert de dimensión 5 con la norma 1 y con la norma infinito.
   (b). Crea varias matrices aleatorias de dimensión 4. Calcula su determinante y su número de condición (con tu norma preferida). ¿Qué relación hay?
   (c). Crea muchas matrices aleatorias de dimensión 4 y representa en una gráfica el determinante y el número de condición.
22. 💻 Consideramos el sistema
$$\begin{pmatrix} 1 & 1/30 \\ 1 & 1/31 \end{pmatrix} x = b,$$
con b un vector unitario con la norma 2. Suponemos que tenemos un pequeño error en el término independiente, es decir, en lugar de b, tenemos $b + \tilde{b}$ y sabemos que $\|\tilde{b}\|_2/\|b\|_2 \leq 1$.
   (a). Calcular el número de condición de la matriz del sistema con la norma euclídea.
   (b). Acotar el error relativo de las soluciones.
   (c). Encontrar b de norma 1 y $\tilde{b}$ en las condiciones anteriores para que el error sea máximo.
   (d). Idem para mínimo.
23. Demostrar que si p es un polinomio de grado d, entonces $p(n) \in \mathcal{O}(n^d)$.
24. Probar que $n \ln n \notin \mathcal{O}(n)$ y que $n^2 \notin \mathcal{O}(n \ln n)$.
25. Sea
$$f(n) = \sum_{i=1}^{n} i(i+3).$$
Demostrar (sin calcular el sumatorio) que $f(n) \in \mathcal{O}(n^3)$.

26. Sea
$$f(n) = \sum_{i=1}^{n} \sum_{j=i}^{n-1} (i+1)j.$$
Demostrar que $f(n) \in \mathcal{O}(n^4)$.

27. 👥 Sea $f(n)$ el n-ésimo término de la sucesión de Fibonacci. Demostrar que $f(n) \notin \mathcal{O}(n)$ y $f(n) \in \mathcal{O}(2^n)$.

28. 👥 Al aplicar la factorización LU con pivote a una matriz diagonalmente dominante, ¿siempre se elige como pivote el elemento de la diagonal?

29. A partir de la eliminación gaussiana, obtener la factorización LU de las siguientes matrices:
$$A = \begin{pmatrix} 3 & 0 & 3 \\ 0 & -1 & 3 \\ 1 & 3 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 & \frac{1}{3} & 0 \\ 0 & 1 & 3 & -1 \\ 3 & -3 & 0 & 6 \\ 0 & 2 & 4 & -6 \end{pmatrix}$$

30. A partir de la eliminación gaussiana con pivote, obtener la factorización LU de las siguientes matrices:
$$A = \begin{pmatrix} 3 & 0 & 3 \\ 0 & -1 & 3 \\ 1 & 3 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 & \frac{1}{3} & 0 \\ 0 & 1 & 3 & -1 \\ 3 & -3 & 0 & 6 \\ 0 & 2 & 4 & -6 \end{pmatrix}$$

31. 💻 Mediante factorización LU, obtener la inversa de las siguientes matrices
$$A = \begin{pmatrix} 3 & 0 & 3 \\ 0 & -1 & 3 \\ 1 & 3 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 & \frac{1}{3} & 0 \\ 0 & 1 & 3 & -1 \\ 3 & -3 & 0 & 6 \\ 0 & 2 & 4 & -6 \end{pmatrix}$$

32. Considerar la matriz
$$A = \begin{pmatrix} 1 & 0 & 0 & 2 \\ 0 & 3 & 0 & 0 \\ 0 & 9 & 4 & 0 \\ 5 & 0 & 8 & 10 \end{pmatrix}$$
   (a). Determinar una matriz triangular inferior M con diagonal unitaria y una matriz triangular superior U tal que $MA = U$.
   (b). Determinar una matriz triangular inferior L con diagonal unitaria y una matriz triangular superior U tal que $A = LU$. Mostrar que $ML = I$ (es decir, $L = M^{-1}$).

33. 💻 Consideremos la matriz
$$A = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ 3 & -4 & 4 \end{pmatrix}.$$
   (a). Obtener la factorización LU de la matriz A.
   (b). Utilizar dicha factorización para obtener la inversa de A.
   (c). Obtener la factorización LU de A^tA. A partir de dicha factorización, obtener la factorización de la forma LDL^t y a partir de ella la factorización de Choleski.

34. Probar que la siguiente matriz no puede factorizarse como producto LU, donde L es una matriz triangular

inferior y U es una matriz triangular superior.

$$A = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix}$$

35. Factorizar la matriz

$$A = \begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix}$$

(a). Como LU donde L es triangular inferior con diagonal unitaria y U es triangular superior.

(b). Usando la factorización anterior, factorizarla como LDU donde L es triangular inferior con diagonal unitaria, D es diagonal y U es triangular superior con diagonal unitaria.

(c). Usando la factorización anterior, factorizarla como LU donde L es triangular inferior y U es triangular superior con diagonal unitaria.

(d). Usando la factorización anterior, factorizarla como LL^t, donde L es triangular inferior.

36. 💻 Consideremos la matriz:

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ -2 & -4 & 4 \end{pmatrix}$$

(a). Obtener la factorización de Schur. ¿Qué ocurre? Calcular los autovalores de la matriz A y discutir porqué no factoriza en los reales como una matriz unitaria por una matriz triangular. Obtener la factorización en los números complejos en doble precisión (CDF)

(b). Resolver el sistem $Ax = b$ con $b = (1, 2, 3)$ usando la factorización anterior.

(c). Consideremos ahora la aplicación lineal dada por

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

Calcular su factorización de Schur. Representar las columnas de la matriz U y también las de AU.

(d). Obtener la factorización de Schur de cualquiera de las matrices anteriores "paso a paso", siguiendo la demostración vista en clase. Para calcular un autovector se puede usar la función *eigenvectors_right* y para extender a una base ortonormal, se puede utilizar la función *gram_schmidt* de Sage.

Tema 2 Métodos iterativos para sistemas lineales

Los métodos directos que hemos estudiado tienen inconvenientes si se aplican a sistemas de ecuaciones de grandes dimensiones, porque requieren muchas operaciones y son sensibles a errores de redondeo. Además, son especialmente poco prácticos cuando las matrices son dispersas, es decir, tienen muchos ceros. Veremos que estos problemas se pueden resolver con los métodos iterativos.

Por otra parte, veremos métodos iterativos para el cálculo de los autovalores y autovectores de una matriz.

2.1 Sucesiones de matrices

Una sucesión de matrices $\{A_r\}_{r=1,2,\ldots}$, $A_r \in \mathcal{M}_n$ se dice que converge a la matriz $A \in \mathcal{M}_n$ si para una (cualquier) norma matricial

$$\lim_{r\to\infty} \|A_r - A\| = 0.$$

Proposición 2.1

Sea $A \in \mathcal{M}_n$. Entonces se verifica

1. *$\rho(A) \leq \|A\|$ para toda norma matricial.*
2. *Para todo $\epsilon > 0$, existe una norma matricial inducida $\|\cdot\|$ tal que*

$$\|A\| \leq \rho(A) + \epsilon.$$

Demostración

1) Sea λ_0 el autovalor que da el radio espectral y sea v un autovalor asociado no nulo.

Sea u un vector tal que $vu^t \neq 0$. Entonces

$$|\lambda_0|\|vu^t\| = \|\lambda_0 vu^t\| = \|Avu^t\| \leq \|A\|\|vu^t\|,$$

luego $\rho(A) = |\lambda_0| \leq \|A\|$

2) Dada A, por el Teorema de Schur, existe U (invertible) tal que $U^{-1}AU$ es triangular superior.

Nótese que los elementos de la diagonal de $U^{-1}AU$ son los autovalores de A. Es decir

$$B = U^{-1}AU = \begin{pmatrix} \lambda_1 & b_{12} & \ldots & b_{1n} \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & b_{n-1,n} \\ 0 & \ldots & 0 & \lambda_n \end{pmatrix},$$

Dado $\epsilon > 0$ tomamos $\delta > 0$ tal que para $1 \leq i \leq n-1$,

$$\sum_{j=i+1}^{n} |\delta^{j-i} b_{ij}| = |\delta| \sum_{j=i+1}^{n} |\delta^{j-i-1} b_{ij}| \leq \epsilon$$

Construimos la matriz $D_\delta = \mathrm{diag}(1, \delta, \delta^2, \ldots, \delta^{n-1})$

Entonces

$$C = (UD_\delta)^{-1} A (UD_\delta) = \begin{pmatrix} \lambda_1 & \delta b_{12} & \ldots & \delta^{n-1} b_{1n} \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \delta b_{n-1,n} \\ 0 & \ldots & 0 & \lambda_n \end{pmatrix}$$

Luego la aplicación

$$\|\cdot\|_R\colon \mathcal{M}_n \to \mathbb{R}, \quad \|B\|_R := \|(UD_\delta)^{-1}B(UD_\delta)\|_\infty$$

es una norma matricial (inducida por $\|v\| := \|(UD_\delta)^{-1}v\|_\infty$) que verifica

$$\|A\|_R = \|(UD_\delta)^{-1}\cdot A\cdot (UD_\delta)\|_\infty = \|C\|_\infty = \max_{1\le i\le n}\sum_{j=1}^{n}|c_{ij}|.$$

Es decir,

$$\|A\| = \max_{1\le i\le n}\left(|\lambda_i| + \sum_{j=i+1}^{n}|\delta^{j-i}u_{ij}|\right) \le \epsilon + \max\lambda_i = \epsilon + \rho(A).$$

□

Teorema 2.1

Sea $A\in\mathcal{M}_n$. Son equivalentes:

1. $\lim_{k\to\infty} A^k = 0$
2. $\lim_{k\to\infty} A^k v = 0$, *para todo* $v\in\mathbb{C}^n$.
3. $\rho(A) < 1$.
4. *Existe una norma matricial (inducida) tal que* $\|A\| < 1$.

♡

Demostración

$1\Rightarrow 2$ Sea $\|\cdot\|_v$ una norma vectorial y $\|\cdot\|_M$ la norma matricial inducida. Entonces $\|A^k v\|_v \le \|A^k\|_M\|v\|_v$.
Luego $\lim_{k\to\infty}\|A^k v\|_v \le \lim_{k\to\infty}\|A^k\|_M\|v\|_v = 0$.

$2\Rightarrow 3$ Supongamos que $\rho(A)\ge 1$ entonces existe un autovalor λ tal que $|\lambda|\ge 1$
Sea v un autovector no nulo asociado $A^k v = \lambda^k v$
Luego $\|A^k v\|_v = |\lambda|^k\|v\| \ge \|v\|$, en contradicción con que la sucesión converge a cero

$3\Rightarrow 4$ Supongamos $\|A\|\ge 1$ para toda norma matricial
Existe una norma matricial inducida tal que $\|A\|\le\rho(A)+\epsilon$.
Luego $1\le\|A\|\le\rho(A)+\epsilon$. Tomando límite, $\rho(A)\ge 1$.

$4\Rightarrow 1$ Supongamos que existe una norma tal que $\|A\|<1$. Para dicha norma $\|A^k\|\le\|A\|^k\to 0$.

□

Teorema 2.2

Sea $A\in\mathcal{M}_n$ y $\|\cdot\|$ una norma matricial. Entonces

$$\lim_{k\to\infty}\|A^k\|^{1/k} = \rho(A).$$

♡

Demostración

Por una parte, $\rho(A)\le\|A\|$. Además, si λ es autovalor de A, entonces λ^k autovalor de A^k.
Luego, para todo k, $\rho(A)^k = \rho(A^k)\le\|A^k\|$. Por tanto

$$\rho(A)\le\|A^k\|^{1/k}$$

Por otra parte, para cada $\epsilon>0$ consideramos la matriz

$$A_\epsilon = \frac{1}{\rho(A)+\epsilon}A.$$

Entonces $\rho(A_\epsilon)<1$. Por el Teorema anterior, $A_\epsilon^k\to 0$. Por definición de convergencia, existe $K>0$ tal que $\|A_\epsilon^k\|<1$ para todo $k>K_\epsilon$.

Luego para todo $k > K_\epsilon$

$$\|A_\epsilon^k\| = \frac{1}{(\rho(A)+\epsilon)^k}\|A^k\| < 1.$$

De donde $\|A^k\| < (\rho(A)+\epsilon)^k$. Luego para todo $k > K_\epsilon$,

$$\|A^k\|^{1/k} < \rho(A) + \epsilon.$$

Tomando límite en k, tenemos que

$$\lim_{k\to\infty} \|A^k\|^{1/k} \leq \rho(A) + \epsilon, \quad \text{para todo } \epsilon > 0.$$

Luego

$$\lim_{k\to\infty} \|A^k\|^{1/k} = \rho(A).$$

□

Finalmente, recordemos el Teorema del punto fijo de Banach que necesitaremos posteriormente.

Teorema 2.3 (Teorema del punto fijo de Banach)

Sea $T\colon \mathbb{K}^n \to \mathbb{K}^n$ una aplicación contractiva (es decir, existe $K < 1$ tal que $\|T(y)-T(x)\| \leq K\|y-x\|$ para todo $x, y \in \mathbb{K}^n$). Entonces existe un único punto fijo de T, $z \in \mathbb{K}^n$.

Es más, para cualquier $x_0 \in \mathbb{K}^n$, sea $\{x_k\}$ la sucesión definida por

$$x_{k+1} = T(x_k), \quad n > 0.$$

Entonces se verifica

$$\lim_{k\to\infty} x_k = z.$$

♡

2.2 Métodos iterativos para resolución de sistemas

Los métodos directos que hemos estudiado tienen inconvenientes si se aplican a sistemas de ecuaciones de grandes dimensiones, porque requieren muchas operaciones y son sensibles a errores de redondeo.

Los métodos iterativos están especialmente indicados en la resolución de este tipo de sistemas, o en aquellos en las que las matrices son dispersas (poseen muchos ceros).

Consisten en definir una sucesión de puntos $\mathbf{x_0}$, $\mathbf{x_1}$, ... que converjan a la sucesión del sistema. Para aplicarlo, escribiremos $\mathbf{A} = \mathbf{M} - \mathbf{N}$. Entonces

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

Eligiendo una matriz M que sea fácil de resolver, podemos aplicar el método del punto fijo partiendo de un vector x_0 y calculando los siguientes vectores de modo recursivo resolviendo el sistema

$$Mx_{k+1} = Nx_k + b.$$

Esto es equivalente a $x_{k+1} = G(x_k)$, donde

$$G(x) = M^{-1}(Nx + b).$$

Como

$$\|G(y) - G(x)\| \leq \|M^{-1}N\|\|y - x\| \quad \text{para todo } x, y \in \mathbb{K}^n,$$

si $\|M^{-1}N\| < 1$, entonces la función G es contractiva.

Teorema 2.4

Sean $A, M, N \in \mathcal{M}_n$ tales que M es invertible y $A = M - N$. Dado $x_0 \in \mathbb{K}^n$, definimos la sucesión $x_{k+1} = G(x_k)$, donde $G(x) = M^{-1}(Nx + b)$.
Una condición necesaria y suficiente para que la sucesión anterior converja para todo valor inicial (a un punto fijo de G) es

$$\rho(M^{-1}N) < 1.$$

♡

Demostración En primer lugar, si $\rho(M^{-1}N) < 1$, entonces existe una norma matricial inducida tal que $\|M^{-1}N\| < 1$. Así que G es contractiva y la sucesión converge al único punto fijo de G.

Recíprocamente, si la sucesión del punto fijo converge a x para cualquier punto inicial x_0, como

$$x_{k+1} - x = M^{-1}N(x_k + b) - M^{-1}N(x + b) = M^{-1}N(x_k - x),$$

entonces

$$0 = \lim_{n\to\infty} x_k - x = \lim_{n\to\infty} \left(M^{-1}N\right)^k (x_0 - x).$$

Nótese que $x_0 - x$ es cualquier punto de $\mathbb{K}^n$. Luego $\rho(M^{-1}N) < 1$.

□

2.2.1 Métodos de Jacobi y Gauss-Seidel

El **método de iteración de Jacobi** consiste en definir M como una matriz diagonal con la misma diagonal que A y

$$\mathbf{N} = M - A$$

Así, la sucesión se construye partiendo de un valor inicial $\mathbf{x}_0$ y definiendo

$$\mathbf{M}\mathbf{x}_{k+1} = N\mathbf{x}_k + \mathbf{b},\ k \geq 0$$

Nótese que en este caso, M^{-1} es una matriz diagonal tal que el i-ésimo elemento de su diagonal es a_{ii}^{-1}, $1 \leq i \leq n$.

El **método de iteración de Gauss-Seidel** se obtiene al tomar M como una matriz triangular inferior cuyos elementos no nulos coinciden con los de A, es decir, si $A = (a_{ij})$ y $M = (m_{ij})$, entonces $m_{ij} = a_{ij}$ si $i \geq j$ y $m_{ij} = 0$ si $i < j$.

La matriz $N = (n_{ij}) = M - A$ verifica $n_{ij} = -a_{ij}$ si $i < j$ y $n_{ij} = 0$ si $i \geq j$.

La sucesión se construye partiendo de un valor inicial $\mathbf{x}_0$ y definiendo

$$\mathbf{M}\mathbf{x}_{k+1} = N\mathbf{x}_k + \mathbf{b},\ k \geq 0$$

Ejemplo 2.1

Ejemplo 4: Dado el sistema de ecuaciones

$$\begin{array}{rcl} 2x - y & = & 9 \\ x + 6y - 2z & = & 15 \\ 4x - 3y + 8z & = & 1 \end{array}, \text{ con } \mathbf{A} = \begin{pmatrix} 2 & -1 & 0 \\ 1 & 6 & -2 \\ 4 & -3 & 8 \end{pmatrix}$$

tenemos:

$$\mathbf{D} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 8 \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 4 & -3 & 0 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix}$$

Partimos de un valor inicial

$$\mathbf{x}_0 = (0,0,0)$$

El método de Jacobi consiste en definir

$$\mathbf{D}\mathbf{x}_{k+1} = -(\mathbf{L}+\mathbf{U})\mathbf{x}_k + \mathbf{b},\ k \geq 0$$

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 8 \end{pmatrix} \begin{pmatrix} x_{k+1} \\ y_{k+1} \\ z_{k+1} \end{pmatrix} = - \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & -2 \\ 4 & -3 & 0 \end{pmatrix} \begin{pmatrix} x_k \\ y_k \\ z_k \end{pmatrix} + \begin{pmatrix} 9 \\ 15 \\ 1 \end{pmatrix}$$

Por tanto,

$$\begin{aligned} 2x_{k+1} &= y_k + 9 \\ 6y_{k+1} &= -x_k + 2z_k + 15 \\ 8z_{k+1} &= -4x_k + 3y_k + 1 \end{aligned}$$

Partiendo de $\mathbf{x}_0 = (0,0,0)$,
$$\begin{aligned} 2x_1 &= y_0 + 9 = 9 \\ 6y_1 &= -x_0 + 2z_0 + 15 = 15 \\ 8z_1 &= -4x_0 + 3y_0 + 1 = 1 \end{aligned}$$

Ejemplo 2.2 Para el método de Jacobi, tenemos:

- $\mathbf{x}_0 = (0,0,0)$
- $\mathbf{x}_1 = (4{,}5, 2{,}5, 0{,}125)$
- $\mathbf{x}_2 = (5{,}75, 1{,}7916667, -1{,}1875)$
- $\mathbf{x}_3 = (5{,}3958333, 1{,}14583333, -2{,}078125)$
- $\mathbf{x}_4 = (5{,}07291667, 0{,}90798611, -2{,}14322917)$

Nótese que, en este caso, el sistema es resoluble, y la solución real es $(5, 1, -2)$, hacia la cual converge la sucesión creada por el método.

Ejemplo 2.3

El método de Gauss-Seidel consiste en definir

$$(\mathbf{D}+\mathbf{L})\mathbf{x}_{k+1} = -\mathbf{U}\mathbf{x}_k + \mathbf{b},\ k \geq 0$$

$$\begin{pmatrix} 2 & 0 & 0 \\ 1 & 6 & 0 \\ 4 & -3 & 8 \end{pmatrix} \begin{pmatrix} x_{k+1} \\ y_{k+1} \\ z_{k+1} \end{pmatrix} = - \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_k \\ y_k \\ z_k \end{pmatrix} + \begin{pmatrix} 9 \\ 15 \\ 1 \end{pmatrix}$$

Por tanto,

$$\begin{aligned} 2x_{k+1} &= y_k + 9 \\ x_{k+1} + 6y_{k+1} &= 2z_k + 15 \\ 4x_{k+1} - 3y_{k+1} + 8z_{k+1} &= 1 \end{aligned}$$

Partiendo de $\mathbf{x}_0 = (0,0,0)$,
$$\begin{aligned} 2x_1 &= y_0 + 9 \\ 6y_1 &= -x_1 + 2z_0 + 15 \\ 8z_1 &= -4x_1 + 3y_1 + 1 \end{aligned}$$

Para el método de Gauss-Seidel, tenemos:

- $\mathbf{x}_0 = (0,0,0)$

- $\mathbf{x}_1 = (4{,}5, 1{,}75, -1{,}46875)$
- $\mathbf{x}_2 = (5{,}375, 1{,}11458333, -2{,}14453125)$
- $\mathbf{x}_3 = (5{,}05729167, 0{,}94227430, -2{,}05029297)$
- $\mathbf{x}_4 = (4{,}97113715, 0{,}98804615, -1{,}99005127)$

2.2.1.1 Convergencia de los métodos iterativos

Teorema 2.5

Dado el sistema $\mathbf{Ax} = \mathbf{b}$, si $\mathbf{A}$ es estrictamente diagonal dominante, entonces existe una única solución del sistema y los métodos iterativos de Jacobi y Gauss-Seidel producen una sucesión de vectores que converge a dicha solución. ♡

Demostración Para el método de Jacobi, basta calcular la norma infinito de $M^{-1}N$. Como A es estrictamente diagonal dominante, se obtiene que $\|M^{-1}N\|_\infty < 1$. Luego es contractiva.

Para el método de Gauss-Seidel, vamos a probar que todos los autovalores de $M^{-1}N$ tienen módulo menor que 1.

Sea $v = (v_1, \ldots, v_n)$ un autovector de $M^{-1}N$ de norma infinito 1 y autovalor λ.

$$M^{-1}Nv = \lambda v, \quad \text{i.e.} \quad Nv = \lambda Mv.$$

Sea i tal que $v_i = 1$ (si es necesario, cambiamos de signo el autovector). Entonces

$$-\sum_{j=i+1}^{n} a_{ij}v_j = \lambda \sum_{j=1}^{i} a_{ij}v_j.$$

Luego

$$|\lambda| = \left| \frac{-\sum_{j=i+1}^{n} a_{ij}v_j}{\sum_{j=1}^{i} a_{ij}v_j} \right| < \frac{\sum_{j=i+1}^{n} |a_{ij}|}{|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|} < 1.$$

□

2.2.2 Criterio de parada

Consideremos un método iterativo de la forma

$$Mx^{(n+1)} = Nx^{(n)} + b,$$

donde $x^{(0)}$ es el punto inicial escogido, $B = M^{-1}N$ la matriz del método y b un vector constante.

Denotemos $\delta_k = \|x^{(k)} - x^{(k-1)}\|$. Si $\varepsilon_{k+1} = \|x - x^{(k+1)}\|$, entonces una estimación para dicho valor es

$$\varepsilon_{k+1} \approx \frac{\delta_{k+1}^2}{\delta_k - \delta_{k+1}}.$$

Vamos a justificar dicho critero. Si x es la solución, entonces $x = M^{-1}(Nx + b)$ y tenemos

$$\varepsilon_{k+1} = \|x - x^{(k+1)}\| = \|M^{-1}(Nx + b) - M^{-1}(Nx^{(k)} + b)\| = \|Bx - Bx^{(k)}\| \leq \|B\|\|x - x^{(k)}\| = \|B\|\varepsilon_k.$$

Por la desigualdad triangular,

$$\varepsilon_{k+1} = \|x - x^{(k+1)}\| \leq \|B\|\|x - x^{(k)}\| \leq \|B\| \left(\|x - x^{(k+1)}\| + \|x^{(k+1)} - x^{(k)}\| \right) = \|B\|(\varepsilon_{k+1} + \delta_{k+1}).$$

Despejando (más nos vale que $\|B\| < 1$), obtenemos

$$\varepsilon_{k+1} \leq \frac{\|B\|}{1 - \|B\|}\delta_{k+1}.$$

Por otra parte, podemos estimar $\|B\|$ del siguiente modo:

$$\delta_{k+1} = \|x^{(k+1)} - x^{(k)}\| = \|Bx^{(k)} - Bx^{(k-1)}\| \leq \|B\|\delta_k$$

De aquí,

$$\|B\| \geq \delta_{k+1}/\delta_k.$$

Tomando la estimación $\|B\| \approx \delta_{k+1}/\delta_k$ y sustituyendo en la desigualdad anterior, tenemos el resultado buscado.

2.3 Métodos iterativos para el cálculo de autovalores y autovectores

Sea $A \in \mathcal{M}_n$ una matriz. Recordemos que sus autovalores pueden ser calculados como las raíces del polinomio

$$p(\lambda) = \det(A - \lambda I),$$

y, dado un autovalor λ, podemos calcular su autovector asociado como la solución del sistema lineal homogéneo (singular)

$$(A - \lambda I)x = 0.$$

Sin embargo estos problemas (en especial el primero) están mal condicionados. Por ello, estudiaremos métodos para determinar directamente los autovalores.

Teorema 2.6 (Teorema de Gershgorin)

Sea $A \in \mathcal{M}_n(\mathbb{C})$. La unión de todos los discos

$$K_i = \{\mu \in \mathbb{C} \colon |\mu - a_{ii}| \leq \sum_{\substack{k=1 \\ k \neq i}}^{n} |a_{ik}|\}$$

contiene todos los autovalores de la matriz $A = (a_{ij})_{1 \leq i,j \leq n}$. ♡

Demostración Sea λ un autovalor y v un autovector asociado. Entonces

$$Av = \lambda v.$$

En particular $\lambda v_i = \sum_{j=1}^{n} a_{ij} v_j$.

Sea i la coordenada donde se alcanza $\|v\|_\infty$. Podemos suponer que $v_i = 1$ y $|v_j| \leq 1$ para todo $j \neq i$. Entonces

$$|\lambda - a_{ii}| = |(\lambda - a_{ii})v_i| = |\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} v_j| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|.$$

Es decir, $\lambda \in K_i$. □

Corolario 2.1

Si la unión $M_1 = \cup_{j=1}^{m} K_{i_j}$ de m discos K_{i_j} $j = 1, \ldots, m$ y la unión M_2 de los discos restantes son disjuntas, entonces M_1 contiene exactamente m autovalores de A. ♡

Demostración (Necesita variable compleja)

Mediante permutaciones de filas y columnas, podemos asumir que los índices de M_1 son $1, \ldots, m$.

Sea D la matriz diagonal con la misma diagonal que A y $B = A - D$. Consideremos $A_t = D + tB$. Los autovalores de A_t son funciones continuas respecto de t, ya que son los ceros del polinomio característico (válido únicamente si son simples, si no, hay que hacer una construcción un poco más delicada). Aplicando

el Teorema de Gershgorin a la matriz A_t, se obtiene que, para $t = 0$, los primeros m autovalores pertenecen a M_1. Como para todo t en un camino que una 0 y 1 y evite los autovalores múltiples, los autovalores $\lambda_i(t)$, $1 \leq i \leq m$, pertenecen a la unión de M_1 y M_2 y $\lambda_i(t)$, $1 \leq i \leq m$, es un conexo, tenemos que $\lambda_i(t) \in M_1$.

□

2.3.1 Métodos para calcular un autovalor. Método de la potencia.

Sea $A \in \mathcal{M}_n$ y supongamos que sus autovalores, $\lambda_1, \ldots, \lambda_n$ verifican

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$$

y que existen n autovectores linealmente independientes, $u_1, u_2, \ldots, u_n$ (tal que $Au_i = \lambda_i u_i$).

Sea $x_0 = \alpha_1 u_1 + \ldots + \alpha_n u_n$, con α_1 y consideremos la sucesión $\{x_k\}$ definida por

$$x_k = Ax_{k-1} = \ldots = A^k x_0.$$

Entonces

$$x_k = \lambda_1^k \left(\alpha_1 u_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^k \alpha_2 u_2 + \ldots + \left(\frac{\lambda_n}{\lambda_1}\right)^k \alpha_n u_n \right).$$

Es decir, para k suficientemente grande, se verifica

$$x_{k+1} \approx \lambda_1 x_k.$$

El **método de la potencia** se basa en aproximar el valor de λ_1, aplicando a ambos términos una función lineal $\phi\colon \mathbb{K}^n \to \mathbb{K}$, por ejemplo:

1. Método de Rayleigh
$$\lambda_1 \approx \frac{x_k^* x_{k+1}}{x_k^* x_k}.$$
2. Método del valor máximo. Sea j la posición de la mayor coordenada en módulo de x_k, entonces
$$\lambda_1 \approx \frac{e_j^* x_{k+1}}{e_j^* x_k}.$$

El método de la potencia también permite calcular un vector propio asociado a λ_1. Conocido λ_1, se verifica que

$$\lim_{k\to\infty} \frac{x_k}{\lambda_1^k} = \alpha_1 u_1$$

Si el autovalor de módulo máximo no es único, existen modificaciones del método que permiten estimarlo. En particular, si λ_1 es un autovalor múltiple (y verifica que su módulo es estrictamente mayor que el del resto de autovalores), entonces el método de la potencia también estima λ_1.

El **método de la potencia inversa** permite calcular el autovalor de módulo mínimo. Sea $A \in \mathcal{M}_n$ invertible y diagonalizable, con autovalores $\lambda_1, \ldots, \lambda_n$ tal que

$$|\lambda_1| \geq |\lambda_2| \geq \ldots > |\lambda_n| \ (> 0).$$

Entonces los autovalores de A^{-1} son $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$ y verifican

$$|\lambda_1^{-1}| \leq |\lambda_2^{-1}| \leq \ldots < |\lambda_n^{-1}|.$$

Por tanto, podemos aproximar λ_n^{-1} mediante el método de la potencia, aplicado a la matriz A^{-1}

El **método de la potencia desplazada** permite mejorar la convergencia del método de la potencia y poder aplicarlo en el caso de autovalores distintos con el mismo módulo. Consideremos la matriz $A - \mu I$, donde μ se denomina desplazamiento. Si λ es un autovalor de A, entonces $\lambda - \mu$ es un autovalor de $A - \mu I$. De este modo,

si dos autovalores tienen el mismo módulo, podemos tomar un desplazamiento para que en la matriz desplazada no tengan el mismo módulo y poder aplicar el método de la potencia.

Una variante del método anterior permite obtener el autovalor más próximo a un número complejo μ dado. Sean $\bar{\lambda}_1, \ldots, \bar{\lambda}_n$ los autovalores de $A - \mu I$ y supongamos que

$$|\bar{\lambda}_1| \geq |\bar{\lambda}_2| \geq \ldots > |\bar{\lambda}_n| > 0.$$

Entonces los autovalores de $(A - \mu I)^{-1}$ son $\bar{\lambda}_1^{-1}, \ldots, \bar{\lambda}_n^{-1}$ y verifican

$$|\bar{\lambda}_1^{-1}| \leq |\bar{\lambda}_2^{-1}| \leq \ldots < |\bar{\lambda}_n^{-1}|.$$

Por tanto, podemos aproximar $(\lambda_n - \mu)^{-1}$ mediante el método de la potencia, aplicado a la matriz $(A - \mu I)^{-1}$.

2.3.2 Métodos basados en transformaciones matriciales

Los métodos anteriores permiten estimar los autovalores de uno en uno. Existen métodos basados en transformaciones matriciales que permiten aproximar todos los autovalores a la vez. En este apartado estudiaremos uno de ellos.

2.3.2.1 Método QR de Francis-Kublanovskaya

El Método QR de Francis-Kublanovskaya es un método iterativo que, bajo ciertas condiciones permite obtener todos los autovalores de una matriz.

Dada una matriz $A \in \mathcal{M}_n(\mathbb{C})$, supondremos que podemos factorizarla como

$$A = QR,$$

donde Q es una matriz unitaria ($Q^*Q = I$) y R es una matriz triangular superior. Esta factorización se denomina **factorización QR**.

Sea $A_1 = A$ y Q_1, R_1 una factorización QR de A_1. Definimos la matriz

$$A_2 = R_1 Q_1 = Q_1^* A_1 Q_1 = Q_1^{-1} A_1 Q_1.$$

En particular A_1 y A_2 tienen los mismos autovalores. El método QR de Francis-Kublanovskaya consiste en ir calculando recursivamente

$$A_k = R_{k-1} Q_{k-1},$$

donde $Q_{k-1} R_{k-1} = A_{k-1}$, R_{k-1} es triangular superior y Q_{k-1} es unitaria.

Teorema 2.7 (Ver [1])

Sea $A \in \mathcal{M}_n(\mathbb{C})$ tal que sus autovalores λ_i, $1 \leq i \leq n$ verifican

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|.$$

Entonces

$$\lim_{k \to \infty} A_k = \begin{pmatrix} \lambda_1 & c_{12} & \ldots & c_{1n} \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & c_{n-1,n} \\ 0 & \ldots & 0 & \lambda_n \end{pmatrix}.$$

Es más,

$$(A_k)_{i,i-1} = \mathcal{O}\left(\left|\frac{\lambda_i}{\lambda_{i-1}}\right|^k\right), \quad i = 2, \ldots, n.$$

Si además A es simétrica, entonces $c_{ij} = 0$, $1 \leq i \neq j \leq n$. ♡

2.3.2.2 Factorización QR Householder

Dado un vector v unitario (con norma euclídea) de $\mathbb{R}^n$, la **reflexión de Householder** es la matriz

$$P = I - 2vv^t.$$

Proposición 2.2

La reflexión de Householder es una matriz simétrica y ortogonal. Además

1. $PP = I_d$.
2. $\|Pc\|_2 = \|c\|_2$ *para todo* $c \in \mathbb{R}^n$.
3. *Geométricamente, se corresponde con una reflexión respecto al hiperplano ortogonal a* v.

♠

Demostración Es fácil comprobar que P es simétrica y ortogonal. En particular $PP = Id$, $\|P\|_2 = 1$ y la imagen por P de un vector c es otro vector con la misma norma (euclídea), ya que

$$\|Pc\|_2^2 = (Pc)^t Pc = c^t P^t Pc = c^t c = \|c\|_2^2.$$

Por otra parte, si $w = v(v^t c)$, w es la proyección de c en v. Entonces $c - w$ es un vector ortogonal a v, es decir, un vector del hiperplano ortogonal y

$$Pc = c - 2vv^t c = c - 2w$$

es el simétrico de c respecto a ese hiperplano. □

La reflexión de Householder permite transformar un vector en otro en la dirección de uno de los ejes.

Proposición 2.3

Dado $c \in \mathbb{R}^n$, existe una reflexión tal que $Pc = \pm\|c\|e_1$.

♠

Demostración Basta tomar

$$v = \frac{u}{\|u\|}, \quad u = c \pm \|c\|e_1.$$

Entonces

$$Pc = c - 2vv^t c = c - u\frac{2u^t c}{\|u\|^2}$$

Ahora bien, como

$$\frac{2u^t c}{\|u\|^2} = \frac{2\left(\|c\|^2 \pm \|c\|c_1\right)}{2\|c\|^2 \pm 2c_1\|c\|} = 1,$$

tenemos que

$$Pc = c - 2vv^t c = c - u\frac{2u^t c}{\|u\|^2} = c - (c \pm \|c\|e_1) = \mp\|c\|e_1$$

□

Algoritmo de factorización.

Sea $A \in \mathcal{M}_n$ y sean c la primera columna. Definimos la matriz

$$Q^{(1)} = I - 2vv^t, \quad v = \frac{u}{\|u\|}, \quad u = c \pm \|c\| e_1.$$

Entonces $Q^{(1)}$ es una matriz unitaria tal que la primera columna de $R^{(1)} = Q^{(1)}A$ tiene ceros debajo de la diagonal.

Definimos $Q^{(k)}$, $R^{(k)}$ por recursivamente. Sea $\tilde{R}$ la submatriz de $R^{(k-1)}$ formada por las últimas $n-k+1$ filas y columnas. Procediendo como antes, existe una reflexión de Householder $\tilde{Q}$ tal que $\tilde{Q}\tilde{R}$ tiene ceros debajo de la diagonal. Entonces definimos

$$Q^{(k)} = \begin{pmatrix} Id & 0 \\ 0 & \tilde{Q} \end{pmatrix}, \quad R^{(k)} = Q^{(k)}R^{(k-1)}.$$

Finalmente $Q = Q^{(1)}Q^{(2)} \dots Q^{(n-1)}$ y $R = Q^{(n-1)} \dots Q^{(2)}Q^{(1)}A$ producen la factorización QR buscada.

Ejemplo 2.4 Obtener una factorización QR de

$$A = \begin{pmatrix} \frac{16}{25} & -\frac{14}{25} & -2 \\ -\frac{12}{25} & -\frac{52}{25} & -1 \\ -\frac{3}{5} & -\frac{3}{5} & -3 \end{pmatrix}.$$

Tema 2 Ejercicios

1. 👥 Demostrar que para matrices de orden 2, el método de Jacobi converge si y sólo si el de Gauss-Seidel converge.
2. Aplicar tres pasos del método de Jacobi para aproximar la solución de los siguientes sistemas (partiendo de $(0,0,0,0)$). Calcular el error residual.

$$a)\begin{cases} 4x_1+2x_2=0 \\ 2x_1-5x_2+2x_3=3 \\ 4x_3-2x_4=0 \\ x_3-3x_4=0 \end{cases} \quad b)\begin{cases} 10x_1+x_2=1 \\ 10x_2+x_3=0 \\ 10x_3+x_4=1 \\ x_1+10x_4=0 \end{cases} \quad c)\begin{cases} x_1+x_3=1 \\ -2x_2-4x_3=2 \\ 2x_3-2x_4=3 \\ 5x_1+4x_4=4 \end{cases}$$

3. Aplicar tres pasos del método de Gauss-Seidel para aproximar la solución de los siguientes sistemas (partiendo de (0,0,0,0)). Calcular el error residual.

$$a)\begin{cases} 4x_1+2x_2=0 \\ 2x_1-5x_2+2x_3=3 \\ 4x_3-2x_4=0 \\ x_3-3x_4=0 \end{cases} \quad b)\begin{cases} 10x_1+x_2=1 \\ 10x_2+x_3=0 \\ 10x_3+x_4=1 \\ x_1+10x_4=0 \end{cases} \quad c)\begin{cases} x_1+x_3=1 \\ -2x_2-4x_3=2 \\ 2x_3-2x_4=3 \\ 5x_1+4x_4=4 \end{cases}$$

4. En general no se puede comparar la convergencia de los métodos de Jacobi y Gauss-Seidel. Probar que
   (a). Si $A=\begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$, entonces el método de Jacobi converge, pero el de Gauss-Seidel no.
   (b). Si $A=\begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$, entonces el método de Gauss-Seidel converge, pero el de Jacobi no.
5. Dada la matriz $A=\begin{pmatrix} 1 & 0 & \alpha \\ 1 & 3 & \alpha \\ 1 & 0 & 2 \end{pmatrix}$, probar para qué valores de α el método de Jacobi es convergente.
6. Dada la matriz $A=\begin{pmatrix} 1 & \alpha & 0 \\ \beta & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix}$, probar para qué valores de α y β los métodos de Jacobi y Gauss-Seidel son convergentes.
7. Consideremos la matriz

$$A=\begin{pmatrix} 1 & 1 & 0 \\ 2 & 3 & 2 \\ 1 & 1 & 2 \end{pmatrix}.$$

   y el vector $b=(1,2,3)$.
   (a). 👥 ¿Existe x_0 tal que el método de Jacobi es convergente?
   (b). 💻 Encontrar un vector inicial de modo que el método de Jacobi para el sistema $Ax=b$ sea convergente. Aplicar 1000 iteraciones del método y representar los errores residuales.
8. 💻 Se considera un disipador en forma de barra. Dividimos barra en 10 secciones de la misma longitud. Denotamos la temperatura de la sección i como x_i. Se sabe que en estado estacionario (es decir, cuando

alcanza el equilibrio), se verifica el siguiente sistema de ecuaciones:

$$\begin{cases} 4x_1 - x_2 = 10, \\ -x_1 + 4x_2 - x_3 = 0, \\ -x_2 + 4x_3 - x_4 = 0, \\ \dots, \\ -x_8 + 4x_9 - x_{10} = 0, \\ -x_9 + 4x_{10} = 0. \end{cases}$$

(a). Calcular la solución mediante factorización LU con pivote.

(b). Calcular la solución mediante dos pasos de Gauss-Seidel. ¿Cuál es el error cometido?

9. Determinar los círculos de Geršgorin de la matriz

$$A = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{pmatrix}.$$

Utilizando dichos círculos, acotar el radio espectral de A.

10. 💻 Consideremos la matriz

$$A = \begin{pmatrix} 2 & -1 & 1 & 2 \\ 1 & 3 & 2 & -2 \\ 1 & -2 & 1 & 1 \\ 1 & -2 & 1 & 2 \end{pmatrix}$$

(a). Calcular sus autovectores con Sage.

(b). Para cada autovalor, obtener un autovector con norma infinito 1 y que tenga una coordenada igual a 1.

(c). Usando el resultado obtenido en el apartado anterior, ¿estarán los autovalores contenidos en los discos de Gershgorin correspondientes a las filas 1 y 3? Comprobar.

11. 👥 Usando el Teorema de Geršgorin, probar que toda matriz A estrictamente diagonal dominante es no singular. Idem si A^t es estrictamente diagonal dominante.

12. 👥 Encontrar una matriz $\mathcal{M}_3(\mathbb{R})$ tal que los discos de Gershgorin tengan todos radio estrictamente positivo y que tenga un autovalor que esté en el borde de uno de los discos.

13. 👥 Encontrar una matriz $\mathcal{M}_3(\mathbb{R})$ tal que los discos de Gershgorin tengan todos radio estrictamente positivo y que tenga un disco que no contenga autovalores.

14. Aplicar tres iteraciones del método de la potencia para aproximar el autovalor dominante de la siguiente matriz

$$A = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{pmatrix}.$$

15. Aplicar tres iteraciones del método de la potencia para aproximar el autovalor dominante de la siguiente matriz

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix},$$

tomando como vector inicial $x_0 = (1, -1, 2)$.

16. Aplicar tres iteraciones del método de la potencia inversa para aproximar el autovalor de menor valor absoluto de la siguiente matriz

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

tomando como vector inicial $x_0 = (-1, 0, 1)$.

17. Aplicar tres iteraciones del método de la potencia inversa desplazada para aproximar cada uno de los autovalores de la siguiente matriz

$$A = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 5 & 1 \\ 0 & 1 & 7 \end{pmatrix},$$

tomando como vectores iniciales $(1, 1, 0), (1, 2, 1), (0, 1, 4)$.

18. Encontrar una factorización QR de la siguiente matriz

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

19. 💻 Aplicar 10 pasos del método QR para estimar los autovalores de la matriz

$$\begin{pmatrix} 5{,}0 & -4{,}0 & 1{,}0 \\ -4{,}0 & 6{,}0 & -4{,}0 \\ 1{,}0 & -4{,}0 & 5{,}0 \end{pmatrix}.$$

Tema 3 Resolución aproximada de ecuaciones no lineales

Los métodos numéricos para obtener un cero de f producen una sucesión $\{x_n\}$ tal que

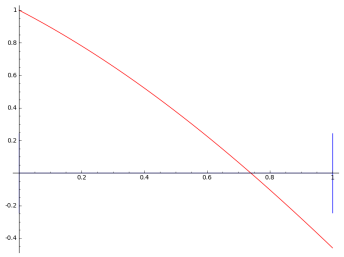
$$\lim_{n\to\infty} x_n = \xi, \quad \text{donde } f(\xi) = 0.$$

Sea $\{x_n\}$ una sucesión producida por un método decimos que *converge a ξ con orden $p \geq 1$ si existen $C > 0$ y $n_0 \in \mathbb{N}$ tal que*

$$\frac{|x_{n+1} - \xi|}{|x_n - \xi|^p} \leq C, \quad \textit{para todo } n \geq n_0.$$

✍ ***Ejercicio 3.1*** *Obtener el orden de las siguientes sucesiones:* $1/n$, $1/n^2$, 2^{-n}, $x_{n+1} = x_n^2$ *con* $x_0 < 1$.

3.1 Métodos iterativos de dos puntos



Teorema 3.1 (Teorema de Bolzano)

Sea f una función continua en un intervalo $[a, b]$ tal que $f(a)f(b) < 0$. Entonces, existe $\xi \in (a, b)$ tal que $f(\xi) = 0$.

♡

Los métodos de dos puntos tratan de reducir la anchura del intervalo manteniendo el cambio de signo.

3.1.1 Método de la bisección

Sea $f \in C[a, b]$ tal que $f(a)f(b) < 0$. El método de la bisección consiste en definir dos sucesiones $\{a_n\}$, $\{b_n\}$, determinadas por $a_0 = a$, $b_0 = b$, y para $n > 0$

$$a_n = \begin{cases} a_{n-1} & \text{si } f(c_{n-1})f(a_{n-1}) \leq 0, \\ c_{n-1} & \text{en otro caso.} \end{cases}$$

$$b_n = \begin{cases} b_{n-1} & \text{si } f(c_{n-1})f(a_{n-1}) > 0, \\ c_{n-1} & \text{en otro caso.} \end{cases}$$

donde

$$c_{n-1} = \frac{a_{n-1} + b_{n-1}}{2}.$$

Método de la bisección

Teorema 3.2

Sean $a < b \in \mathbb{R}$, $f \in \mathcal{C}([a,b])$ tal que $f(a)f(b) < 0$ y sean $\{a_n\}$, $\{b_n\}$ los extremos izquierdo y derecho, respectivamente, de los intervalos obtenidos por el método de la bisección.

Entonces existe $\xi \in (a,b)$ tal que $f(\xi) = 0$ y

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \xi.$$

Además,

$$b_n - \xi \leq \frac{b-a}{2^n}, \quad \xi - a_n \leq \frac{b-a}{2^n}.$$

♡

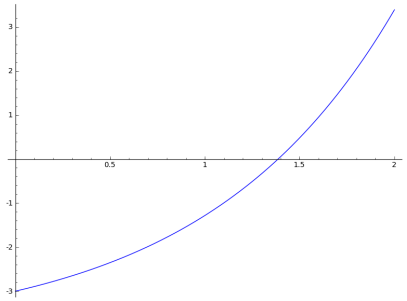
Sea $f \in \mathcal{C}[a,b]$ tal que $f(a)f(b) < 0$. El *método de la regula falsi se define análogamente al método de bisección tomando*

$$c_{n-1} = \frac{f(b_{n-1})a_{n-1} - f(a_{n-1})b_{n-1}}{f(b_{n-1}) - f(a_{n-1})}.$$

En este método, no tenemos asegurado que la anchura de los intervalos converja a cero. Puede ocurrir que a partir de un término sólo se actualice a_n ó b_n. En todo caso, se puede demostrar que la sucesión c_n converge a un cero de f.

3.2 Métodos de un punto

Métodos de un punto



Tenemos una función $f(x)$ **continua** *y un punto inicial x_0.*

Queremos calcular un cero de $f(x)$ próximo a x_0.

Los métodos de un punto tratan de obtener el cero creando una sucesión $\{x_n\}$ que converja a dicho valor.

3.2.1 Método del punto fijo

El objetivo es obtener un **punto fijo** *de una función $F(x)$ dada, es decir, un valor c tal que $F(c) = c$.*

A partir de la función $F(x)$ y de un punto inicial x_0, definimos la sucesión del método del punto fijo como

$$\boxed{x_n = F(x_{n-1})}.$$

Proposición 3.1

Sea F continua, $x_0 \in \mathbb{R}$ y $\{x_n\}$ la sucesión definida por $x_n = F(x_{n-1})$. Si $\{x_n\}$ converge a $c \in \mathbb{R}$, entonces $F(c) = c$.

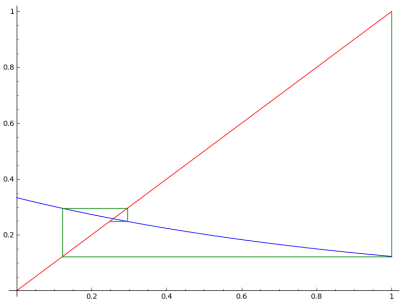
♠

Demostración Si $x_n \to c$ cuando $n \to \infty$,

$$c = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} F(x_n) = F\left(\lim_{n\to\infty} x_n\right) = F(c).$$

□

Ejemplo 3.1 *Consideremos el método del punto fijo para $F(x) = e^{-x}/3$ (gráfica azul), con $x_0 = 1$. Podemos representar las primeras iteraciones en una gráfica mediante un diagrama de Verhulst (o diagrama de telaraña - cobweb diagram). La gráfica roja es la identidad, el punto buscado es aquel en el que se cortan.*



Los valores que obtenemos al aplicar el método son los siguientes:

- $x_0 = 1$
- $x_1 = F(x_0) = 0{,}122626$
- $x_2 = F(x_1) = 0{,}294865$
- $x_3 = F(x_2) = 0{,}248211$

Podemos aplicar el Teorema del punto fijo de Banach para asegurar la existencia y unicidad del punto fijo. En este contexto, el teorema tiene el siguiente enunciado.

Teorema 3.3

Supongamos que $F \in C^1([a,b])$.

1. *Si $F(x) \in [a,b]$ para todo $x \in [a,b]$, entonces F tiene un punto fijo en $[a,b]$.*
2. *Si además $F'(x)$ está definida y es continua en (a,b) y $|F'(x)| < 1$ para todo $x \in (a,b)$, entonces F tiene un único punto fijo c en $[a,b]$.*

♡

Demostración El primer punto se sigue del Teorema de Bolzano aplicado a la función $F(x) - x$.

El segundo, del Teorema de Valor Medio, considerando que haya dos puntos fijos $c_1 \neq c_2$ en a, b, entonces

$$|c_1 - c_2| = |F(c_1) - F(c_2)| = |F'(\xi)||c_1 - c_2| < |c_1 - c_2|.$$

De esta contradicción, $c_1 = c_2$. □

Teorema 3.4

Supongamos que $F \in C^1([a,b])$ verifica:

1. *$F(x) \in [a,b]$ para todo $x \in [a,b]$.*
2. *$F'(x)$ está definida y es continua en (a,b) y $|F'(x)| \leq K < 1$ para todo $x \in (a,b)$.*

Sea $x_0 \in [a,b]$, sea c el único punto fijo de F y sea $\{x_n\}$ definida recursivamente por $x_n = F(x_{n-1})$. Entonces

$$\lim_{n\to\infty} x_n = c, \quad E_n := |c - x_n| \leq K^n \frac{|x_1 - x_0|}{1 - K}.$$

♡

Demostración

Tenemos que

$$|c - x_{n+1}| = |F(c) - F(x_n)| \leq K|c - x_n|.$$

Por inducción,

$$|c - x_n| \leq K^n|c - x_0|.$$

En consecuencia $x_n \to c$ cuando $n \to \infty$. Además,

$$|c - x_0| \leq |c - x_1| + |x_1 - x_0| \leq K|c - x_0| + |x_1 - x_0|.$$

Luego

$$|c - x_0| \leq |x_1 - x_0|/(1 - K).$$

□

✍ ***Ejercicio 3.2***

Consideremos la función $f(x) = ax$, para $a \in \mathbb{R}$. Calcular para qué valores de a el método del punto fijo es convergente para cualquier condición inicial en el intervalo $[-1, 1]$ y qué valores de a no es convergente para toda condición inicial en $[-1, 1]$.

Pensar qué ocurre si $a \in \mathbb{C}$.

3.2.2 Extrapolación de Aitken

Supongamos que $F \in \mathcal{C}([a, b])$ y que c un punto fijo de F. Si denotamos por $\lambda = F'(c)$ y $\{x_n\}$ la sucesión del método del punto fijo, entonces podemos estimar el error como:

$$c - x_n = F(c) - F(x_{n-1}) = F'(\xi)(c - x_{n-1}) \approx \lambda(c - x_{n-1}).$$

Despejando c, tenemos

$$c \approx \frac{x_n - \lambda x_{n-1}}{1 - \lambda} = x_n + \frac{\lambda}{1 - \lambda}(x_n - x_{n-1}).$$

Proposición 3.2

Supongamos que $x_n \to c$, $n \to \infty$. La sucesión

$$\lambda_n := \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}}$$

converge a λ cuando $n \to \infty$. ♠

Demostración Aplicando el Teorema de Valor Medio

$$\frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}} = \frac{F(x_{n-1}) - F(x_{n-2})}{x_{n-1} - x_{n-2}} = F'(\xi_n) \to F(c) = \lambda.$$

□

Una estimación del error, llamada **Fórmula de extrapolación de Aitken** *es*

$$c - x_n \approx \frac{\lambda_n}{1 - \lambda_n}(x_n - x_{n-1}).$$

Esta fórmula puede utilizarse para acelerar la convergencia.

3.2.3 Método de Newton-Raphson

Partimos de una función $f(x)$ y un punto inicial x_0.

Definimos x_1 como el punto donde la tangente a la gráfica de $f(x)$ en el punto $(x_0, f(x_0))$ corta el eje x.

La ecuación de la tangente es:

$$y - f(x_0) = f'(x_0)(x - x_0)$$

Despejando el punto de corte con el eje x,

$$x_1 = x_0 - f(x_0)/f'(x_0).$$

Repitiendo el proceso,

$$\boxed{x_{n+1} = x_n - f(x_n)/f'(x_n)}$$

Convergencia

Teorema 3.5

Supongamos que f, f', f'' son continuas en un entorno de una raíz, c, de f y que $f'(c) \neq 0$.

Existe un entorno U de c tal que si $x_0 \in U$, entonces la sucesión $\{x_n\}$ del método de Newton-Raphson está contenida en U y

$$\lim_{n\to\infty} x_n = c.$$

Además,

$$E_{n+1} := c - x_{n+1} = -\frac{1}{2}\frac{f''(\xi_n)}{f'(x_n)}E_n^2,$$

para cierto ξ_n entre x_n y c.

♡

Demostración

Desarrollando en serie de Taylor,

$$f(c) = f(x_n) + f'(x_n)(c - x_n) + f''(\xi)(c - x_n)^2/2.$$

Luego

$$c - x_{n+1} = c - x_n + f(x_n)/f'(x_n) = \frac{f(x_n) + f'(x_n)(c - x_n)}{f'(x_n)} = -\frac{f''(\xi)}{2f'(x_n)}(c - x_n)^2.$$

Definimos

$$c(\delta) = \frac{\text{máx}_{|x-c|\leq\delta}\,|f''(x)|}{2\,\text{mín}_{|x-c|\leq\delta}\,|f'(x)|}.$$

Elegimos δ tal que $\rho = \delta c(\delta) < 1$ (existe porque $\delta c(\delta) \to 0$ cuando $\delta \to 0$). Entonces, si $e_0 = |c - x_0| \leq \delta$,

$$e_1 = |c - x_1| \leq c(\delta)e_0^2 \leq \delta c(\delta)e_0 = \rho e_0 < e_0 \leq \delta.$$

Luego

$$e_2 \leq \rho e_1 \leq \rho^2 e_0,$$

Repitiendo el proceso

$$e_n \leq \rho e_{n-1} \leq \rho^n e_0.$$

Como $\rho < 1$, $e_n \to 0$ cuando $n \to \infty$.

□

Proposición 3.3

Sea $f \in \mathcal{C}^2([a, b])$, con una raíz c en $[a, b]$.

Si $f'(c) > 0$ y $f''(x) > 0$ para todo $x \in [c, b]$, entonces la sucesión $\{x_n\}$ converge a c para todo valor inicial $x_0 \in [c, b]$.

Se verifica un resultado análogo si $f'(c) < 0$ y $f''(x) < 0$ para todo $x \in [a, b]$. ♠

Demostración Se deja como ejercicio. □

3.2.4 Método de la secante

Partimos de una función $f(x)$ y de dos puntos iniciales x_0, x_1. Calculamos el siguiente punto x_2 como el punto en el que la secante determinada por los puntos $(x_0, f(x_0))$ y $(x_1, f(x_1))$ corta el eje x.

La ecuación de la secante es:

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

Imponemos que corte el eje x:

$$x_2 = x_0 - f(x_0)\frac{x_1 - x_0}{f(x_1) - f(x_0)} = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)}.$$

Repitiendo el proceso,

$$\boxed{x_{n+2} = \frac{x_n f(x_{n+1}) - x_{n+1} f(x_n)}{f(x_{n+1}) - f(x_n)}}$$

Teorema 3.6

Sea c una solución de $f(x) = 0$ y supongamos que $f'(r) \neq 0$. Entonces el método de la secante converge en un entorno de c y es de orden $(1 + \sqrt{5})/2$. ♡

Demostración Ver Kincaid-Cheney. □

3.2.5 Sistemas no lineales

Queremos resolver un sistema de ecuaciones no lineales

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0, \\ \ldots \\ f_n(x_1, \ldots, x_n) = 0. \end{cases}$$

3.2.5.1 Punto fijo

El método del punto fijo consiste en escribir el sistema no lineal en la forma $F(x) = x$, donde $F\colon \mathbb{R}^n \to \mathbb{R}^n$ es una función continua. Fijado $x_0 \in \mathbb{R}^n$, definimos la sucesión

$$x_{n+1} = F(x_n).$$

Por continuidad, si la sucesión $\{x_n\}$ definida por el método del punto fijo converge, es decir, existe $z \in \mathbb{R}^n$ tal que

$$\lim_{n\to\infty} x_n = z,$$

entonces z es un punto fijo de F y por tanto una solución del sistema original.

Por el Teorema del punto fijo de Banach, una condición suficiente para que la sucesión converja es que la aplicación F sea contractiva.

Teorema 3.7

Supongamos que $F\colon D \subset \mathbb{R}^n \to \mathbb{R}^n$, $F \in \mathcal{C}^1(D)$, existe $z \in D$ tal que $F(z) = z$ y $\rho(JF(z)) < 1$. Entonces existe un entorno U de z tal que para todo $x_0 \in U$, la sucesión definida por el método del punto fijo converge a z. ♡

Demostración Existe una norma matricial inducida tal que $\|JF(z)\| < 1$. Tomemos una bola centrada en z tal que $\|JF(x)\| \leq K < 1$ para todo x en la bola. Entonces

$$F(y) - F(x) = \int_0^1 JF(x + t(y-x))(y-x)dt.$$

Luego

$$\|F(y) - F(x)\| \leq \int_0^1 \|JF(x + t(y-x))(y-x)\|dt \leq \int_0^1 \|JF(x + t(y-x))\|\|y-x\|dt < \| \leq K\|y-x\|$$

De donde obtenemos que es contractiva. □

3.2.5.2 Newton-Raphson

Consideremos de nuevo el sistema

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0, \\ \ldots \\ f_n(x_1, \ldots, x_n) = 0. \end{cases}$$

Podemos escribir el sistema como $\mathbf{f}(x) = 0$, para $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{f} = (f_1, \ldots, f_n)$. Sean $x, z \in \mathbb{R}^n$ tal que $f(z) = 0$ y supongamos $\mathbf{f}$ de clase 2. Entonces, desarrollando en serie de Taylor,

$$0 = f(z) \approx \mathbf{f}(x) + \mathrm{J}\,\mathbf{f}(x) \cdot (z - x).$$

El método de Newton-Raphson consiste en, partiendo de x_0, generar la sucesión definida por

$$\mathrm{J}\,\mathbf{f}(x_n) \cdot (x_{n+1} - x_n) = -\mathbf{f}(x_n).$$

Teorema 3.8

Supongamos que $f\colon D \subset \mathbb{R}^n \to \mathbb{R}^n$, $f \in \mathcal{C}^2(D)$, existe $z \in D$ tal que $f(0) = 0$.
Entonces existe un entorno U de z tal que para todo $x_0 \in U$, la sucesión definida por el método de Newton-Raphson converge a z. ♡

Demostración El método de Newton-Raphson es equivalente a aplicar el método del punto fijo a la función

$$F(x) = x - (\mathrm{J}\,\mathbf{f}(x))^{-1}\mathbf{f}(x).$$

Es fácil ver que $\mathrm{J}\,F(z)$ es la matriz cero, por lo que aplicando el Teorema del método del punto fijo concluimos. Concretamente,

$$\mathrm{J}\,F(x) = \mathrm{J}(\mathrm{Id}(x)) - \mathrm{J}\left([\mathrm{J}\,\mathbf{f}(x)]^{-1} \cdot \mathbf{f}(x)\right) = \mathrm{Id}_n - \mathrm{J}\left([\mathrm{J}\,\mathbf{f}(x)]^{-1}\right) \cdot \mathbf{f}(x) - [\mathrm{J}\,\mathbf{f}(x)]^{-1} \cdot \mathrm{J}(\mathbf{f}(x)).$$

Hay que tener en cuenta que todas las cosas que estamos sumando son matrices cuadradas de orden n, aunque no estamos *multiplicando* matrices cuadradas de orden n: $\mathrm{J}\left([\mathrm{J}\,\mathbf{f}(x)]^{-1}\right)$ es un tensor de orden 3 (matriz de $n \times n \times n$ o aplicación lineal de $(\mathbb{R}^n)^3$ en $(\mathbb{R}^n)^2 \cong \mathcal{M}_n$, como cada cual lo quiera ver) y $\mathbf{f}(x) \in \mathbb{R}^n$. En cualquier caso, el último sumando es $-\mathrm{Id}$ y, al evaluar en z, $\mathrm{J}\left([\mathrm{J}\,\mathbf{f}(x)]^{-1}\right) \cdot \mathbf{f}(x) = \mathrm{J}\left([\mathrm{J}\,\mathbf{f}(z)]^{-1}\right) \cdot \mathbf{f}(z) = 0$.
□

Tema 3 Ejercicios

1. *Probar que la ecuación $e^x + 2x = 0$ tiene una única raíz. Acotar dicha raíz mediante el método de la bisección con un error menor de 10^{-2}.*
2. *Aplicar el Método de bisección a $f(x) = x^3 - 16 = 0$, a fin de determinar la raíz cúbica de 16 con un error menor que $0{,}125$.*
3. *Estudiar si el método del punto fijo para la función $F(x) = -\operatorname{sin}(x)+x+1/2$ converge para condiciones iniciales en el intervalo $[0, 1]$.*
4. *Acotar el error cometido al aplicar tres pasos del método del punto fijo para calcular un punto fijo de $F(x) = -\operatorname{sin}(x) + x + 1/2$ partiendo de $x_0 = 0{,}5$. Comprobar las condiciones en el intervalo $[0, 1]$.*
5. *Encontrar un intervalo donde la función $F(x) = \exp(-x)/3$ tenga un punto fijo y tal que el método del punto fijo converja para cualquier valor inicial en dicho intervalo.*
6. 👥 *Probar que si F es una función de clase $\mathcal{C}^1$ y $[a, b]$ es un intervalo tal que $F(x) \in [a, b]$ para todo $x \in [a, b]$ y $0 < F'(x) < 1$, entonces la sucesión obtenida al aplicar el método del punto fijo converge monótonamente a un punto fijo de F.*
7. 👥 *Probar que si F es una función de clase $\mathcal{C}^1$ y $[a, b]$ es un intervalo tal que $F(x) \in [a, b]$ para todo $x \in [a, b]$ y $-1 < F'(x) < 0$, entonces F tiene un único punto fijo c en dicho intervalo, y si x_n es la sucesión obtenida al aplicar el método del punto fijo, entonces $(x_n - c)(x_{n+1} - c) < 0$ y $|x_n - c|$ converge monotonamente (decreciente) a cero.*
8. 👥 *Demostrar que si F es una función de clase $\mathcal{C}^1$ y tiene un punto fijo x_0 tal que $|F'(x_0)| < 1$, entonces existe un intervalo tal que el método del punto fijo converge para toda condición inicial en dicho intervalo.*
9. *Encontrar una solución de $e^x - 4x - \operatorname{sin}(x) = 0$ en el intervalo $[0, 1]$ mediante el método del punto fijo. Probar al menos tres transformaciones de la ecuación y estudiar la convergencia del método.*
10. 👥 *Discutir la convergencia del método del punto fijo para la función $F(x) = ae^x$, $a \in \mathbb{R}$, para condiciones iniciales en el intervalo $[0, 1]$.*
11. *Aplicar cuatro pasos del Método de Newton-Raphson para encontrar una raíz cúbica de 50 partiendo de $x_0 = 4$. Acotar el error.*
12. *Aproximar mediante 4 iteraciones del método de Newton-Raphson partiendo de $x_0 = -0{,}5$ la posición del mínimo de $f(x) = e^x + x^2/2$. Acotar el error.*
13. 👥 *Obtener una función diferenciable, con un único cero en $[0, 1]$ tal que al aplicar el método de Newton-Raphson partiendo de $x_0 = 0$, obtengamos $x_1 = 1$, $x_2 = 0$, ...*
14. 👥 *Obtener un ejemplo de función diferenciable tal que al aplicar el método de Newton-Raphson, obtengamos un 3-ciclo, es decir, la sucesión repita siempre los mismos tres valores.*
15. 💻 *Supongamos que las ecuaciones del movimiento de un proyectil son*

$$y = f(t) = 4605(1 - e^{-t/15}) - 147t, \quad x = r(t) = 22400(1 - e^{-t/15}).$$

Determine el tiempo transcurrido hasta el impacto con error menor que 10^{-10}.
16. 💻 *Halle el punto de la parábola $y = x^2$ que está más cerca del punto $(3, 1)$ con error menor que 10^{-10}. Utilizar el método de Aitken para estimar el error.*
17. 💻 *Halle el punto de la curva $y = \operatorname{sen}(x - \operatorname{sen}(x))$ que está más cerca del punto $(2{,}1, 0{,}5)$ con error menor que 10^{-10}. Utilizar el método de Aitken para estimar el error.*

18. 💻 *Halle con error menor que* 10^{-10}*, el valor de* x *para el que es mínima la distancia vertical entre las gráficas de las funciones* $f(x) = x^2 + 2$ *y* $g(x) = x/5 - \operatorname{sen}(x)$*. Utilizar el método de Aitken para estimar el error.*
19. 💻 *La curva formada por un cable colgante se llama catenaria. Supongamos que el punto más bajo de una catenaria es el origen* $(0, 0)$*, entonces la ecuación de la catenaria es* $y = C\cosh(x/C) - C$*. Si queremos determinar la catenaria que pasa por los puntos* $(\pm a, b)$*, entonces debemos resolver la ecuación* $b = C\cosh(a/C) - C$*, donde la incógnita es* C*.*
    (a). *Pruebe que la catenaria que pasa por los puntos* $(\pm 10, 6)$ *es*
    $$y = 9{,}1889\cosh(x/9{,}1889) - 9{,}1889.$$
    (b). *Halle la catenaria que pasa por los puntos* $(\pm 12, 5)$*.*
20. 👥 *Si* p *es una raíz de multiplicidad* n *de una función* f *de clase* $n + 1$*, entonces* $f(x) = (x - p)^n q(x)$*, donde* $q(p) \neq 0$*.*
    (a). *Pruebe que, en este caso,* $h(x) = f(x)/f'(x)$ *tiene una raíz simple en* p*.*
    (b). *Pruebe que si aplicamos el método de Newton-Raphson para hallar la raíz simple* p *de* $h(x)$*, entonces el método queda:*
    $$x_{k+1} = x_k - \frac{f(x_k)f'(x_k)}{(f'(x_k))^2 - f(x_k)f''(x_k)}$$
    (c). *Aplicar el método anterior para obtener una raíz (el 0) de* $f(x) = \operatorname{sen}(x^3)$ *partiendo de* $x_0 = 1$*. Compararlo con el método usual de Newton-Raphson.*
21. *Aproximar una solución de*
    $$\begin{cases} x - y^3/10 - z^2/20 = 1, \\ x^2/10 + y - z^3/20 = 1, \\ x/20 - y^3/20 + z = 1, \end{cases}$$
    aplicando tres pasos del método del punto fijo, partiendo de $(1, 1, 1)$
22. *Aproximar una solución de*
    $$\begin{cases} x - y^3/10 - z^2/20 = 1, \\ x^2/10 + y - z^3/20 = 1, \\ x/20 - y^3/20 + z = 1, \end{cases}$$
    aplicando tres pasos del método de Newton-Raphson, partiendo de $(1, 1, 1)$*.*
23. 💻 *Aproximar la distancia entre las curvas paramétricas definidas mediante las ecuaciones*
    $$r(t) = (\cos t, 2\sin t) \quad t \in [0, 2\pi]$$
    y
    $$s(t) = (4 - \sin t\cos t, 5 + \cos 2t) \quad t \in [0, 2\pi].$$
    Elegir el método y aplicar cinco iteraciones.
24. 💻 *Aproximar la ecuación de una recta tangente común a las curvas paramétricas*
    $$r(t) = (\cos t, \sin t) \quad t \in [0, 2\pi]$$
    y
    $$s(t) = (3 + 2\sin t, \cos t) \quad t \in [0, 2\pi].$$
    Elegir el método y aplicar cinco iteraciones.

Tema 4 Interpolación y aproximación polinomial

La **interpolación** *consiste en obtener una función que pase por una serie de puntos prefijados.*

- **Interpolación polinomial.** *Elegimos los polinomios de grado menor o igual que el número de puntos menos 1.*
- **Interpolación a trozos.** *Elegimos funciones definidas a trozos por polinomios.*
- **Otras interpolaciones.** *Aunque no lo estudiaremos en este tema, también existe la interpolación por funciones racionales (aproximantes de Padé), por funciones trigonométricas, etc.*

4.1 Polinomio interpolador

Denotaremos $\mathcal{P}_n$ el conjunto de polinomios (reales) de grado menor o igual que n.

Teorema 4.1

Dados $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$, $x_i \neq x_j$ si $i \neq j$, existe un único polinomio en $\mathcal{P}_n$, $P_n(x)$, tal que

$$P_n(x_k) = y_k,\ k = 0, 1, \ldots, n.$$

♡

Demostración $\mathcal{P}_n$ es un espacio vectorial de dimensión $n+1$.

Consideremos la aplicación lineal $\phi\colon P_n \to \mathbb{R}^{n+1}$ definida como $\phi(P) = (P(x_0), \ldots, P(x_n))$.

Bastaría ver es un automorfismo, ya que dados $(y_0, \ldots, y_n)$ el polinomio interpolador es la preimagen por ϕ.

Si tomamos la base $1, x, x^2, \ldots$, entonces su matriz es la matriz de Vandermonde y el determinante es no nulo.

Se puede demostrar, de manera alternativa, comprobando que el núcleo de ϕ es cero. Esto es inmediato, ya que el núcleo está formado por los polinomios de grado menor o igual que n que tienen las $n+1$ raíces $x_0, x_1, \ldots, x_n$.

□

El único polinomio de grado menor o igual que n, $P_n(x)$ que verifica

$$P_n(x_k) = y_k,\ k = 0, 1, \ldots, n$$

se denomina **polinomio interpolador de los puntos**.

Ejemplo 4.1 *Consideremos los puntos:*

$$(0, 4), (1, 3), (2, 1), (3, 4)$$

Tenemos que resolver el sistema:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 1 \\ 4 \end{pmatrix}$$

Resolvemos y obtenemos

$$P(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 = 4 + \frac{3}{2} x - \frac{7}{2} x^2 + x^3.$$

Corolario 4.1

Consideremos el espacio vectorial generado por funciones $\{f_0(x), \ldots, f_n(x)\}$. *Entonces existe una única función de dicho espacio que interpola* $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ *si y sólo si*

$$\begin{vmatrix} f_0(x_0) & \ldots & f_n(x_0) \\ \vdots & \ddots & \vdots \\ f_0(x_n) & \ldots & f_n(x_n) \end{vmatrix} \neq 0$$

♡

Si $f(x) = \sum_{i=0}^{n} a_i f_i(x)$, *es la función de interpolación del corolario anterior, entonces* $a_0, \ldots, a_n$ *es la solución del sistema lineal*

$$\begin{pmatrix} f_0(x_0) & \ldots & f_n(x_0) \\ \vdots & \ddots & \vdots \\ f_0(x_n) & \ldots & f_n(x_n) \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \ldots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} y_0 \\ \ldots \\ y_n \end{pmatrix}$$

Supongamos que los puntos (x_k, y_k) *pertenecen a la gráfica de una función* f *(es decir,* $y_k = f(x_k)$*). Entonces el siguiente resultado da una idea del error que se comete al sustituir la función por su polinomio interpolador.*

Teorema 4.2

Sean $x_0 < \ldots < x_n \in \mathbb{R}$.

Dada $f \in \mathcal{C}^{n+1}([x_0, x_n])$, *si* $P(x)$ *es el polinomio interpolador de* $(x_k, f(x_k))$, $k = 0, \ldots, n$, $x_i \neq x_j$ *si* $i \neq j$, *entonces para todo* $x \in [x_0, x_n]$, *existe* $\xi \in [x_0, x_n]$ *tal que*

$$f(x) - P_n(x) = \frac{f^{n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\ldots(x - x_n)$$

♡

Demostración Consideramos un punto $\bar{x} \in (x_0, x_n), \bar{x} \neq x_i$; las funciones dadas por $\omega(x) = \prod_{i=0}^{n}(x - x_i)$, $g(x) = f(x) - P_n(x)$ y el valor $c = g(\bar{x})/\omega(\bar{x})$.

Entonces la función $\phi(x) = g(x) - c\omega(x)$ tiene al menos $n+1$ ceros en $[x_0, x_n]$, que son $x_0, \ldots, x_n, \bar{x}$. Por el Teorema de Rolle, su derivada $(n+1)$-ésima tiene un cero en (x_0, x_n), digamos ξ. La derivada $(n+1)$-ésima de P_n es 0 y la de ω es $(n+1)!$, luego

$$0 = f^{n+1)}(\xi) - c(n+1)!$$

Despejando

$$f(\bar{x}) - P_x(\bar{x}) = \frac{f^{n+1)}(\xi)}{(n+1)!}\omega(\bar{x}).$$

□

4.1.1 Polinomios de Lagrange

Dados $x_0, x_1, \ldots, x_n \in \mathbb{R}$, $x_i \neq x_j$ *si* $i \neq j$, *se definen los* **polinomios (base) de Lagrange** *como*

$$L_i(x) = \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)}$$

$L_i(x)$ *es el polinomio interpolador de* $(x_i, 1)$ *y* $(x_j, 0)$, *para* $j \neq i$.

Proposición 4.1

El polinomio interpolador de (x_0, y_0), (x_1, y_1), ..., (x_n, y_n) *es*

$$P_n(x) = \sum_{i=0}^{n} y_i L_i(x)$$

♠

4.1.2 Método de Newton

Dados $x_0, x_1, \ldots, x_n \in \mathbb{R}$, $x_i \neq x_j$ *si* $i \neq j$, *se definen los* **polinomios (base) de Newton** *como*

$$1, \quad x - x_0, \quad (x - x_0)(x - x_1), \quad \ldots, (x - x_0)(x - x_1)\cdots(x - x_{n-1}).$$

La matriz de la aplicación lineal $\phi\colon \mathcal{P}_n \to \mathbb{R}^{n+1}$ *definida como*

$$\phi(P) = (P(x_0), \ldots, P(x_n)),$$

es triangular inferior (con diagonal no nula).

Dada una función $f(x)$ *y* $n+1$ *valores* $x_0, x_1, \ldots, x_n \in \mathbb{R}$ *tales que* $x_i \neq x_j$ *si* $i \neq j$, *se definen las* **diferencias divididas** *de* $f(x)$ *recursivamente como*

$$f[x_k] = f(x_k),$$
$$f[x_r, \ldots, x_k] = \frac{f[x_{r+1}, \ldots, x_k] - f[x_r, \ldots, x_{k-1}]}{x_k - x_r}, \quad 0 \leq r < k \leq n.$$

Obviamente esto mismo se puede definir sin la función f, *simplemente teniendo unos ciertos valores* $y_0, \ldots, y_n \in \mathbb{R}$.

Teorema 4.3

Sean $x_0, x_1, \ldots, x_n \in \mathbb{R}$ *tales que* $x_i \neq x_j$ *si* $i \neq j$ *y sea* f *una función definida (al menos) sobre esos puntos. Entonces el polinomio interpolador de los puntos* $(x_0, f(x_0)), \ldots, (x_n, f(x_n))$ *es*

$$P_n(x) = \sum_{j=0}^{n} f[x_0, x_1, \ldots, x_j] \prod_{i=0}^{j-1} (x - x_i).$$

♡

Demostración Lo demostraremos por inducción sobre el grado del polinomio.

Si $n = 0$ entonces es trivial, para $n = 1$ tenemos

$$P_1(x) = f[x_0] + \tfrac{f[x_1] - f[x_0]}{x_1 - x_0}(x - x_0).$$

Veamos que si es cierto para el polinomio de grado $n-1$ entonces también lo es para el de grado n. Por hipótesis de inducción en los puntos $x_0, \ldots, x_{n-1}$ y $x_1, \ldots, x_n$, tenemos que

$$P_{n-1}(x) = f[x_0] + f[x_0, x_1](x - x_0) + \ldots + f[x_0, x_1, \ldots, x_{n-1}](x - x_0)\ldots(x - x_{n-2}).$$
$$\widetilde{P}_{n-1}(x) = f[x_1] + f[x_1, x_2](x - x_1) + \ldots + f[x_1, x_2, \ldots, x_n](x - x_1)\ldots(x - x_{n-1}).$$

Buscaremos a_n para que

$$P_n(x) = P_{n-1}(x) + a_n(x - x_0)\ldots(x - x_{n-1}).$$

Nótese que basta con despejar $a_n = [P_n(x_n) - P_{n-1}(x_n)]/[(x_n - x_0)\ldots(x_n - x_{n-1})]$.

Es claro que

$$P_n(x) = \widetilde{P}_{n-1}(x) + \frac{x - x_n}{x_n - x_0}(\widetilde{P}_{n-1}(x) - P_{n-1}(x)).$$

Igualando los coeficientes de x^n en la anterior igualdad tenemos

$$a_n = \frac{1}{x_n - x_0}(f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]) = f[x_0, x_1, \ldots, x_n].$$

□

Corolario 4.2

La diferencia dividida $f[x_0, \ldots, x_n]$ es invariante frente a permutaciones de $x_0, \ldots, x_n$, es decir, si $\sigma \in S_{n+1}$, entonces

$$f[x_0, \ldots, x_n] = f[x_{\sigma(0)}, \ldots, x_{\sigma(n)}].$$

♡

Demostración En la demostración anterior no hemos elegido ningún orden para los puntos $x_0, \ldots, x_n$ y $f[x_0, \ldots, x_n]$ es el coeficiente de x^n, luego la unicidad del polinomio de interpolación nos da el resultado. □

Teorema 4.4

Sean P_n el polinomio interpolador de una función f en $n+1$ puntos distintos, $x_0, x_1, \ldots, x_n \in \mathbb{R}$. Para todo $\bar{x} \neq x_0, \ldots, x_n$,

$$f(\bar{x}) - P_n(\bar{x}) = f[x_0, x_1, \ldots, x_n, \bar{x}] \prod_{i=0}^{n} (\bar{x} - x_i).$$

♡

Demostración Basta considerar el polinomio interpolador de f en $x_0, \ldots, x_n, \bar{x}$. Tenemos

$$P_{n+1}(x) = P_n(x) + f[x_0, x_1, \ldots, x_n, \bar{x}] \prod_{i=0}^{n} (x - x_i).$$

Ahora basta tener en cuenta que $P_{n+1}(\bar{x}) = f(\bar{x})$. □

Corolario 4.3

Si $f \in \mathcal{C}^n([a, b])$ y $x_0, \ldots, x_n \in [a, b]$, entonces existe $\xi \in (a, b)$ tal que

$$f[x_0, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}$$

♡

4.1.3 Interpolación de Hermite

Teorema 4.5

Sean $x_0 < x_1 < \ldots < x_n$, y

$$y_i^{(j)}, \quad 0 \leq i \leq n, \quad 0 \leq j \leq m_i.$$

Existe un único polinomio de grado menor o igual que $m := n + \sum_{i=0}^{n} m_i$ tal que

$$P^{(j)}(x_i) = y_i^{(j)}, \quad 0 \leq i \leq n, \quad 0 \leq j \leq m_i.$$

♡

El polinomio definido por el Teorema anterior se denomina **polinomio interpolador de Hermite**.

Definimos las diferencias divididas como

$$f[x_k, x_k] = f'(x_k),$$

$$f[x_k, x_k, x_k] = \frac{f''(x_k)}{2},$$

$$\ldots,$$

$$f[x_k, \overset{m}{\ldots}, x_k] = \frac{f^{m-1)}(x_k)}{(m-1)!}.$$

Denotemos

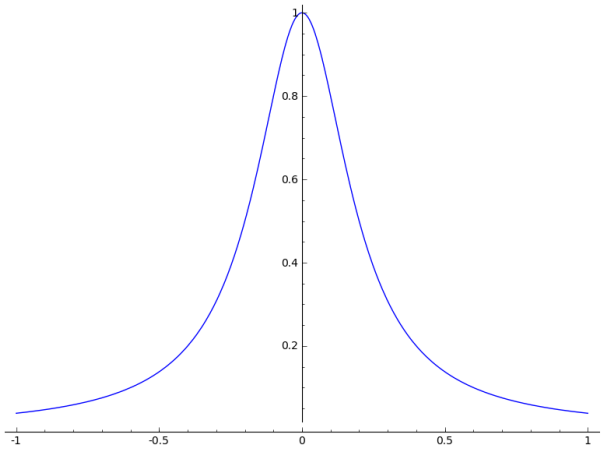
$$\{\tilde{x}_i\}_{i=0}^{m} = \{x_0, \overset{m_0+1}{\ldots}, x_0, x_1, \overset{m_1+1}{\ldots}, x_1, \ldots, x_n, \overset{m_n+1}{\ldots}, x_n\}$$

Entonces el polinomio interpolador de Hermite es

$$P(x) = \sum_{j=0}^{m} f[\tilde{x}_0, \ldots, \tilde{x}_j] \prod_{i=0}^{j-1} (x - \tilde{x}_i).$$

4.1.4 Fenómeno de Runge

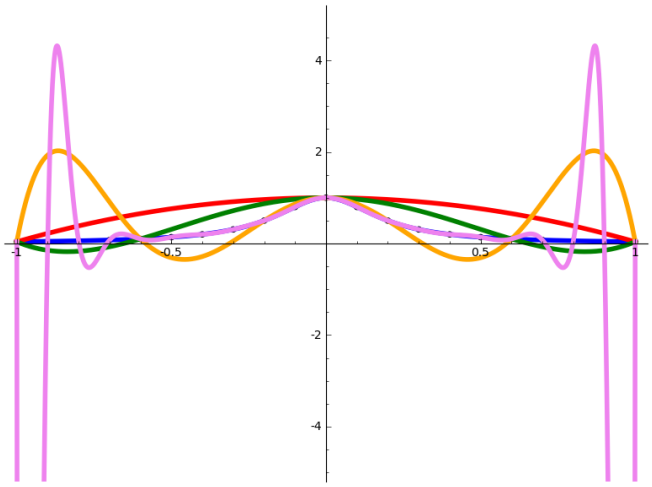
El ajuste de una curva mediante polinomios de interpolación de grado alto, esto es, para un conjunto numeroso de datos, suele resultar poco satisfactoria, pues produce oscilaciones en los extremos.



Runge observó que si interpolamos la función

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

en puntos equiespaciados, la distancia entre la función y el polinomio de interpolación, con la norma infinito, tiende a infinito cuando aumentamos el número de puntos.



4.2 Interpolación polinomial a trozos: splines

La interpolación polinomial a trozos consiste en construir un polinomio de grado 1, 2 o 3 para cada par de nodos consecutivos (x_k, y_k) y (x_{k+1}, y_{k+1}). La curva definida mediante estos "trozos" se denomina **spline**. *Estudiaremos:*

1. *Interpolación lineal a trozos: splines lineales*
2. *Interpolación cúbica a trozos: splines de Hermite*
3. *Interpolación cúbica a trozos: splines cúbicos*

4.2.1 Splines lineales

Consisten simplemente en unir los nodos o puntos mediante segmentos. Así, dado el par de nodos $(x_k, y_k), (x_{k+1}, y_{k+1})$, definimos el segmento que los une:

$$S_k(x) = y_k + \frac{y_{k+1} - y_k}{x_{k+1} - x_k}(x - x_k),\ x \in [x_k, x_{k+1}]$$

con lo que el spline lineal de los puntos $(x_0, y_0), \ldots, (x_n, y_n)$ es la función definida a trozos:

$$S(x) = S_k(x), x \in [x_k, x_{k+1}],\ k = 0, \ldots, n-1$$

4.2.2 Interpolación cúbica a trozos de Hermite

Nos dan

- $x_0 < x_1 < \ldots < x_n$,
- $y_0, y_1, \ldots, y_n$,
- $y'_0, y'_1, \ldots, y'_n$.

Queremos una función f tal que

$$f(x_k) = y_k, \quad f'(x_k) = y'_k, \quad 0 \leq k \leq n.$$

Vamos a obtener una función de la forma

$$f(x) = \Big\{ a_k x^3 + b_k x^2 + c_k x + d_k, \quad x \in (x_{k-1}, x_k).$$

Es decir, en cada intervalo (x_{k-1}, x_k) tomamos el polinomio cúbico de Hermite.

4.2.3 Splines cúbicos

Se define la curva en $[x_0, x_n]$ a trozos

$$S(x) = S_k(x), x \in [x_{k-1}, x_k],\ k = 1, \ldots, n$$

donde $S_k(x)$ es un polinomio cúbico y se impone que

1. *S interpole: $S_k(x_{k-1}) = y_{k-1}$, $S_k(x_k) = y_k$, $k = 1, \ldots, n$.*
2. *S sea derivable en cada x_k: $S'_{k-1}(x_k) = S'_k(x_k)$, $k = 1, \ldots, n-1$.*
3. *S sea dos veces derivable en cada x_k: $S''_{k-1}(x_k) = S''_k(x_k)$, $k = 1, \ldots, n-1$.*

Para tener una única solución del sistema, formado por $4n-2$ ecuaciones y $4n$ incógnitas, debemos imponer dos restricciones más. Según la forma de imponerlas, podemos definir distintos tipos de splines.

1. *Spline cúbico natural: este spline cúbico es el que minimiza la energía de tensión.*

$$S''(x_0) = 0, \quad S''(x_n) = 0.$$

2. *Spline periódico (suponiendo que $y_0 = y_n$):*

$$S'(x_0) = S'(x_n), \quad S''(x_0) = S''(x_n)$$

3. *Spline cúbico sujeto: Suponemos conocidos y'_0, y'_n e imponemos*

$$S'(x_0) = y'_0, \quad S'(x_n) = y'_n$$

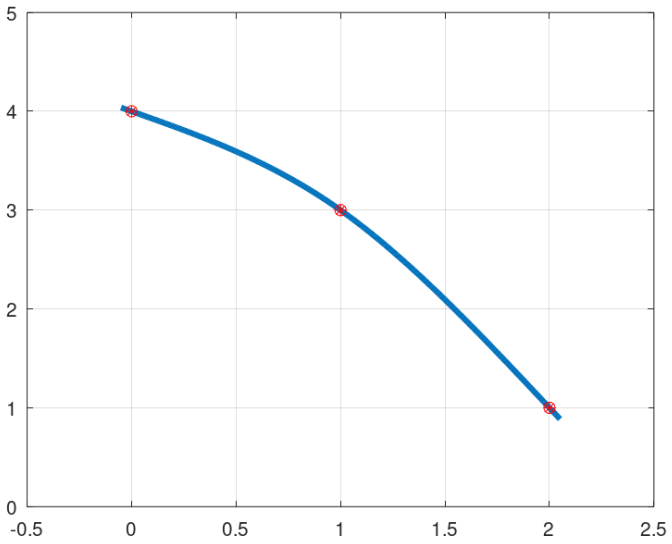
Ejemplo 4.2 *Spline cúbico natural que interpola $(0,4)$, $(1,3)$, $(2,1)$.*

Tenemos el sistema

$$\begin{array}{r} S_0(0) = 4 \to \\ S_0(1) = 3 \to \\ S_1(1) = 3 \to \\ S_1(2) = 1 \to \\ S_0'(1) = S_1'(1) \to \\ S_0''(1) = S_1''(1) \to \\ S_0''(0) = 0 \to \\ S_1''(2) = 0 \to \end{array} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 4 & 8 \\ 0 & 1 & 2 & 3 & 0 & -1 & -2 & -3 \\ 0 & 0 & 2 & 6 & 0 & 0 & -2 & -6 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 12 \end{pmatrix} \begin{pmatrix} d_0 \\ c_0 \\ b_0 \\ a_0 \\ d_1 \\ c_1 \\ b_1 \\ a_1 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 3 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Resolviendo el sistema, obtenemos

$$S_0(x) = 4 - \frac{3}{4}x - \frac{1}{4}x^3, \quad S_1(x) = \frac{7}{2} + \frac{3}{4}x - \frac{3}{2}x^2 + \frac{1}{4}x^3.$$



4.2.4 Método de los momentos

Sean $x_0 < x_1 < \ldots < x_n$. Supongamos que tenemos valores $y_0, y_1, \ldots, y_n$ y sea S el spline cúbico natural tal que $S(x_i) = y_i$. En esta sección veremos un método efectivo para calcularlo. Además, probaremos la existencia y unicidad de dicho spline.

Como vamos a imponer que $S \in \mathcal{C}^2([x_0, x_n])$, podemos definir los **momentos** *de S en $x_0, \ldots, x_n$ como*

$$z_i := S''(x_i), \quad 0 \leq i \leq n.$$

Denotemos

$$h_i := x_{i+1} - x_i, \quad b_i := \frac{y_{i+1} - y_i}{h_i},$$
$$u_i = 2(h_{i-1} + h_i), \quad v_i = 6(b_i - b_{i-1}).$$

Proposición 4.2

Los momentos z_i son las soluciones del siguiente sistema tridiagonal

$$\begin{cases} z_0 & = 0, \\ h_{i-1}z_{i-1} + u_i z_i + h_i z_{i+1} & = v_i, \quad (1 \le i \le n-1) \\ z_n & = 0. \end{cases}$$

♠

Demostración Fijado un intervalo $[x_i, x_{i+1}]$, $S''(x)$ restringido a dicho intervalo es un polinomio de grado 1, S_i, tal que $S_i''(x_i) = z_i$ y $S''(x_{i+1}) = z_{i+1}$. Entonces

$$S_i''(x) = z_{i+1}\frac{x - x_i}{h_i} + z_i\frac{x_{i+1} - x}{h_i}, \quad x \in [x_i, x_{i+1}].$$

Integrando (dos veces), tenemos

$$S_i'(x) = z_{i+1}\frac{(x - x_i)^2}{2h_i} - z_i\frac{(x_{i+1} - x)^2}{2h_i} + A,$$
$$S_i(x) = z_{i+1}\frac{(x - x_i)^3}{6h_i} + z_i\frac{(x_{i+1} - x)^3}{6h_i} + Ax + B, \quad x \in [x_i, x_{i+1}].$$

Imponiendo que $S(x_i) = y_i$ y que $S(x_{i+1}) = y_{i+1}$, tenemos

$$y_i = \frac{z_i}{6h_i}h_i^3 + Ax_i + B, \quad y_{i+1} = \frac{z_{i+1}}{6h_i}h_i^3 + Ax_{i+1} + B;$$

$$y_{i+1} - y_i = \frac{z_{i+1} - z_i}{6h_i}h_i^3 + A(x_{i+1} - x_i) = \frac{z_{i+1} - z_i}{6}h_i^2 + Ah_i.$$

Despejando, obtenemos una expresión para A:

$$A = \frac{1}{h_i}\left(y_{i+1} - y_i + \frac{z_i - z_{i+1}}{6}h_i^2\right) = \left(\frac{y_{i+1}}{h_i} - \frac{h_i z_{i+1}}{6}\right) - \left(\frac{y_i}{h_i} - \frac{h_i z_i}{6}\right). \tag{4.1}$$

Para determinar B, tenemos

$$\begin{aligned} B &= y_i - Ax_i - \frac{h_i^2 z_i}{6} = (x_{i+1} - x_i)\frac{y_i}{h_i} - \left[\frac{y_{i+1}}{h_i} - \frac{h_i z_{i+1}}{6} - \frac{y_i}{h_i}\frac{h_i z_i}{6}\right]x_i - (x_{i+1} - x_i)\frac{h_i z_i}{6} \\ &= \frac{x_{i+1}y_i}{h_i} - \frac{x_i y_i}{h_i} - \frac{x_i y_{i+1}}{h_i} + \frac{x_i h_i z_{i+1}}{6} + \frac{x_i y_i}{h_i} - \frac{x_i h_i z_i}{6} - \frac{x_{i+1} z_i h_i}{6} + \frac{x_i z_i h_i}{6} \\ &= x_{i+1}\left(\frac{y_i}{h_i} - \frac{z_i h_i}{6}\right) - x_i\left(\frac{y_{i+1}}{h_i} - \frac{z_{i+1}h_i}{6}\right). \end{aligned} \tag{4.2}$$

Así, obtenemos

$$Ax + B = \left(\frac{y_{i+1}}{h_i} - \frac{h_i}{6}z_{i+1}\right)(x - x_i) + \left(\frac{y_i}{h_i} - \frac{h_i}{6}z_i\right)(x_{i+1} - x).$$

Luego

$$\begin{aligned} S_i(x) = &\frac{z_{i+1}}{6h_i}(x - x_i)^3 + \frac{z_i}{6h_i}(x_{i+1} - x)^3 \\ &+ \left(\frac{y_{i+1}}{h_i} - \frac{h_i}{6}z_{i+1}\right)(x - x_i) + \left(\frac{y_i}{h_i} - \frac{h_i}{6}z_i\right)(x_{i+1} - x). \end{aligned}$$

Veamos ahora que $z_0, \ldots, z_n$ son soluciones del sistema lineal anterior. Es inmediato que $z_0 = z_n = 0$ por ser el spline natural. Sustituimos para obtener $S'_{i-1}(x_i)$ y $S'_i(x_i)$:

$$S'_i(x_i) = -\frac{z_i}{2h_i}(x_{i+1} - x_i)^2 + \left(\frac{y_{i+1}}{h_i} - \frac{h_i}{6}z_{i+1}\right) - \left(\frac{y_i}{h_i} - \frac{h_i}{6}z_i\right).$$

$$S'_{i-1}(x_i) = \frac{z_i}{2h_{i-1}}(x_i - x_{i-1})^2 + \left(\frac{y_i}{h_{i-1}} - \frac{h_{i-1}}{6}z_i\right) - \left(\frac{y_{i-1}}{h_{i-1}} - \frac{h_{i-1}}{6}z_{i-1}\right).$$

Ahora, teniendo en cuenta que $S'_{i-1}(x_i) = S'_i(x_i)$, tenemos

$$\frac{1}{6}h_{i-1}z_{i-1} + \frac{2}{6}(h_i + h_{i-1})z_i + \frac{1}{6}h_i z_{i+1} = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}.$$

Multiplicando por 6 obtenemos la ecuación buscada. □

De la demostración anterior obtenemos además:

Teorema 4.6

Sean $x_0 < x_1 < \ldots < x_n$. Supongamos que tenemos valores $y_0, y_1, \ldots, y_n$. El spline cúbico natural S que verifica $S(x_i) = y_i$, $0 \le i \le n$, está formado por los polinomios

$$\begin{aligned} S_i(x) =& \frac{z_{i+1}}{6h_i}(x - x_i)^3 + \frac{z_i}{6h_i}(x_{i+1} - x)^3 \\ &+ \left(\frac{y_{i+1}}{h_i} - \frac{h_i}{6}z_{i+1}\right)(x - x_i) + \left(\frac{y_i}{h_i} - \frac{h_i}{6}z_i\right)(x_{i+1} - x). \end{aligned}$$

para $x \in [x_i, x_{i+1}]$, $0 \le i < n$.

4.3 Teoría de la aproximación

El **problema de aproximación** *consiste en, dada una función continua f y una familia de funciones $\mathcal{F}$, obtener $g \in \mathcal{F}$ tal que la distancia entre f y g sea mínima.*

Por tanto, necesitaremos fijar dos elementos:

1. *La definición de distancia entre dos funciones.*
2. *La familia de funciones $\mathcal{F}$.*

Un **espacio prehilbertiano** *$\mathcal{C}$ es un espacio vectorial sobre $\mathbb{R}$ (o $\mathbb{C}$), dotado de un producto escalar, $\langle \cdot, \cdot \rangle$, es decir, una aplicación $\langle \cdot, \cdot \rangle \colon \mathcal{C}^2 \to \mathbb{R}$ bilineal, simétrica y definida positiva. A partir del producto escalar definimos una norma $\|v\| = \sqrt{\langle v, v \rangle}$.*

Ejemplos

1. *$\mathbb{R}^n$ con el producto escalar*

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i v_i.$$

2. *$\mathcal{C}_w[a, b]$ el espacio de funciones continuas sobre el intervalo $[a, b]$ con el producto escalar*

$$\langle f, g \rangle = \int_a^b f(x)g(x)w(x)\,dx,$$

donde w es una función continua y positiva en $[a, b]$.

Dadas $f, g \in \mathcal{C}$, decimos que f y g son ortogonales, $f \perp g$, si $\langle f, g \rangle = 0$.

Lema 4.1

Sea $\mathcal{C}$ un espacio prehilbertiano. Se verifica

1. $\|f+g\|^2 = \|f\|^2 + \|g\|^2 + 2\langle f, g\rangle$.
2. $f \perp g$ *si y sólo si* $\|f+g\|^2 = \|f\|^2 + \|g\|^2$.
3. $|\langle f, g\rangle| \leq \|f\|\|g\|$.
4. $\|f+g\|^2 + \|f-g\|^2 = 2(\|f\|^2 + \|g\|^2)$.

♡

Demostración

1. Aplicando la definición.
2. Pitágoras: inmediato.
3. Desigualdad de Cauchy-Schwarz.

   Supongamos que no se verifica. $|\langle f, g\rangle| > \|f\|\|g\|$ En particular, $\|g\|$ no puede ser cero, pues entonces se daría la igualdad. Podemos suponer que $\|g\| = 1$, por lo que $|\langle f, g\rangle| > \|f\|$. Reescalando, podemos suponer $\langle f, g\rangle = 1$ y tendríamos $\|f\| < 1$. Entonces llegamos a contradicción, pues

   $$0 \leq \|f-g\|^2 = \|f\|^2 - 2\langle f, g\rangle + \|g\|^2 = \|f\|^2 - 1$$

4. Igualdad del paralelogramo. Es inmediata a partir del primer punto.

□

Sea $\mathcal{C}$ un espacio prehilbertiano y $\mathcal{S}$ un subespacio de $\mathcal{C}$. Dada $f \in \mathcal{C}$, decimos que $g \in \mathcal{S}$ es la **mejor aproximación de f en $\mathcal{S}$** *si*

$$\|f-g\| \leq \|f-h\| \quad \textit{para todo } h \in \mathcal{S}.$$

Decimos que $h \in \mathcal{C}$ es **ortogonal** *a $\mathcal{S}$, $h \perp \mathcal{S}$, si $\langle h, g\rangle = 0$ para todo $g \in \mathcal{S}$.*

Teorema 4.7

Sea $\mathcal{C}$ un espacio prehilbertiano y $\mathcal{S}$ un subespacio de $\mathcal{C}$.
Dada $f \in \mathcal{C}$, $g \in \mathcal{S}$ es la mejor aproximación de f en $\mathcal{S}$ si y sólo si $f-g \perp \mathcal{S}$.

♡

Demostración Si $f-g$ es ortogonal a $\mathcal{S}$, entonces por el teorema de Pitágoras se tiene

$$\|f-h\|^2 = \|f-g+g-h\|^2 = \|f-g\|^2 + \|g-h\|^2 \geq \|f-g\|^2$$

para toda $h \in \mathcal{S}$.

Recíprocamente, si g es la mejor aproximación, consideremos $h \in \mathcal{S}$ y $\lambda > 0$. Entonces

$$0 \leq \|f-g+\lambda h\|^2 - \|f-g\|^2 = \|f-g\|^2 + 2\lambda\langle f-g, h\rangle + \lambda^2\|h\| - \|f-g\|^2 = \lambda(2\langle f-g, h\rangle + \lambda\|h\|^2).$$

Tomando límite cuando $\lambda \to 0^+$, tenemos que $\langle f-g, h\rangle \geq 0$. Análogamente, tomando $-h$ en lugar de h, $\langle f-g, h\rangle \leq 0$. En consecuencia, $\langle f-g, h\rangle = 0$, luego $f-g \perp \mathcal{S}$.

□

Ejemplo 4.3 *Aproximar* $\operatorname{sen}(x)$ *por un polinomio de la forma* $g(x) = ax + bx^3 + cx^5$ *con la norma inducida por el producto escalar*

$$\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)\,dx.$$

Dado un espacio prehilbertiano, decimos que un conjunto finito de vectores o una sucesión de vectores $f_1, f_2, \ldots$ es **ortogonal** *si*

$$\langle f_i, f_j\rangle = 0 \quad \textit{para todo } i \neq j.$$

Decimos que es **ortonormal** *si*

$$\langle f_i, f_j \rangle = \delta_{ij}.$$

Teorema 4.8

Sea $\{g_1, \ldots, g_n\}$ un conjunto ortogonal de vectores de $\mathcal{C}$. Sea $\mathcal{S}$ el subespacio generado por $\{g_1, \ldots, g_n\}$ y $f \in \mathcal{C}$. Entonces $g = \sum_{i=1}^{n} c_i g_i$ es la mejor aproximación de f en $\mathcal{S}$ si y sólo si

$$c_i = \langle f, g_i \rangle / \langle g_i, g_i \rangle, \quad i = 1, \ldots, n.$$

♡

4.3.1 Aproximación cuadrática polinomial

Dados un intervalo I y una función w, continua y positiva en el interior de I, denotaremos $\mathcal{C}_w(I)$ al espacio prehilbertiano de las funciones continuas sobre I dotado del producto escalar

$$\langle f, g \rangle = \int_I f(x) g(x) w(x)\, dx, \quad f, g \in \mathcal{C}_w(I).$$

Sea $\mathcal{P}_n$ el espacio vectorial de los polinomios de grado menor o igual que n. Dada $f \in \mathcal{C}_w(I)$, buscamos la mejor aproximación de f en $\mathcal{P}_n$.

Como hemos visto antes, necesitaremos una base ortogonal de $\mathcal{P}_n$. Para ello podemos aplicar la ortogonalización de Gram-Schmidt a $\{1, x, x^2, \ldots, x^n\}$.

Teorema 4.9

Consideremos los polinomios $p_0(x) = 1$, $p_1(x) = x - a_1$ y

$$p_i(x) = (x - a_i) p_{i-1}(x) - b_i p_{i-2}(x), \quad i = 2, \ldots, n,$$

donde

$$a_i = \frac{\langle x p_{i-1}, p_{i-1} \rangle}{\langle p_{i-1}, p_{i-1} \rangle}, \quad b_i = \frac{\langle x p_{i-1}, p_{i-2} \rangle}{\langle p_{i-2}, p_{i-2} \rangle}.$$

Entonces p_0, …, p_n, es una base ortogonal de $\mathcal{P}_n$.

♡

Demostración En primer lugar, por la definición, tenemos que p_i es un polinomio mónico de grado i, luego los polinomios $p_0, \ldots, p_n$ forman base.

Vamos a demostrar que $\langle p_j, p_i \rangle = 0$, con $i < j$ por inducción sobre j. Si $j = 1$, es fácil comprobar que $\langle p_0, p_1 \rangle = 0$.

Supongamos que es cierto para todo $j < n$ y probémoslo para n. Si $i = n - 1$ o $i = n - 1$, basta sustituir p_n por su definición y aplicar la bilinealidad.

Si $i < n - 2$, tenemos

$$\langle p_n, p_j \rangle = \langle (x - a_i) p_{n-1}(x) - b_i p_{n-2}(x), p_j \rangle.$$

Desarrollando por linealidad y usando que $\langle p_{n-1}, p_j \rangle = \langle p_{n-2}, p_j \rangle = 0$, tenemos

$$\langle p_n, p_j \rangle = \langle x p_{n-1}(x), p_j \rangle = \langle p_{n-1}(x), x p_j \rangle.$$

Por otra parte, como

$$p_{j+1}(x) = (x - a_{j+1}) p_j(x) - b_{j+1} p_{j-1}(x),$$

tenemos que

$$x p_j(x) = p_{j+1}(x) + a_{j+1} p_j(x) + b_{j+1} p_{j-1}(x).$$

Sustituyendo en la expresión de $\langle p_n, p_j \rangle$, tenemos

$$\langle p_n, p_j \rangle = \langle p_{n-1}(x), p_{j+1}(x) + a_{j+1}p_j(x) + b_{j+1}p_{j-1}(x) \rangle = 0.$$

□

Ejemplo 4.4 *Si tomamos como producto escalar*

$$\langle f, g \rangle = \int_{-1}^{1} f(x)g(x)dx.$$

Los polinomios de la base ortogonal dada por el Teorema anterior se denominan **polinomios de Legendre**.

Ejemplo 4.5 *Los polinomios de Chebyshev forman una base ortogonal de $\mathcal{P}_n$ con el producto escalar*

$$\langle f, g \rangle = \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}}dx.$$

4.3.2 Ajuste por mínimos cuadrados

La versión discreta del problema anterior es, dada una función continua f, valores $x_0 < x_1 < \ldots < x_n \in \mathbb{R}$ y el espacio de polinomios de grado menor o igual a k, $\mathcal{P}_k$, considerar la norma inducida por el producto semiescalar

$$\langle f, g \rangle = \sum_{i=0}^{n} f(x_i)g(x_i).$$

La teoría anterior es válida en este caso, sin más que cambiar el producto. El caso particular $k = 1$ se denomina regresión lineal*, con $k = 2$* regresión cuadrática*, etc. En el caso particular $k = n$ se obtiene el polinomio interpolador.*

Tema 4 Ejercicios

1. *Probar que los polinomios* $p(x) = -x^2 + 3x$ *y* $q(x) = -x^3 + 3x^2$ *interpolan los puntos* $(0,0)$, $(1,2)$, $(3,0)$. *¿Por qué no tenemos un único polinomio interpolador en este caso?*
2. 💻 *Encontrar* a_0, a_1, a_2 *tales que la función*
$$f(x) = a_0 \frac{1}{1+x^2} + a_1 \frac{1}{1+(x-1)^2} + a_2 \frac{1}{1+(x-2)^2},$$
interpole a la función e^x *en los puntos* $x = 0, 1, 2$. *Dibuja la gráfica de ambas funciones.*
3. *Consideremos la función* $f(x) = e^x$. *Acotar el error cometido a tomar* $f(0{,}5)$ *como el valor del polinomio interpolador de* f *en los puntos* $x = 0, 1, 2$.
4. *Consideremos la función* $f(x) = \cos(x)$. *Acotar el error cometido a tomar* $f(\pi/5)$ *como el valor del polinomio interpolador de* f *en los puntos* $x = 0, \pi/4, \pi/2$.
5. *Consideremos la función* $f(x) = e^x$. *Tomamos* $x_0 = 0$. *¿Cuál es el mayor valor de* x_1 *que podemos considerar para que la interpolación lineal de* f *en* x_0, x_1 *tenga un error menor de* 10^{-2}*? (es decir,* $|f(x) - p(x)| < 10^{-2}$ *para todo* $x \in (x_0, x_1)$*, donde* p *es el polinomio lineal que interpola* f *en* x_0, x_1*).*
6. *Hemos aproximado la función* $\ln x$ *por un polinomio de interpolación de grado 9 en el intervalo* $[1, 2]$ *usando puntos uniformemente distribuidos. Acotar el error cometido.*
7. *Utilizando los polinomios de Lagrange, obtener los polinomios interpoladores de las siguientes tablas:*

(a).

x	-1	0	1
y	0	3	0

(b).

x	0	1	2
y	-2	0	2

(c).

x	-1	0	1	2
y	-1	-2	0	2

8. *Obtener los polinomios interpoladores de las siguientes tablas mediante diferencias divididas:*

(a).

x	-1	0	1
y	0	3	0

(b).

x	0	1	2
y	-2	0	2

(c).

x	-1	0	1	2
y	-1	-2	0	2

9. *Sea la función* $f(x) = e^x$ *y los valores siguientes:* $f(0) = 1$, $f(0{,}5) = 1{,}64872$, $f(1) = 2{,}71828$ *y* $f(2) = 7{,}38906$*, efectuar los siguientes cálculos:*
   (a). *Aproximar* $f(0{,}25)$ *usando interpolación lineal a partir de los puntos anteriores más próximos.*
   (b). *Aproximar* $f(0{,}75)$ *usando interpolación lineal a partir de los puntos anteriores más próximos.*
   (c). *Aproximar* $f(0{,}25)$ *y* $f(0{,}75)$ *utilizando el polinomio de interpolación de los puntos* $(0, f(0))$, $(0{,}5, f(0{,}5))$, $(1, f(1))$, $(2, f(2))$.
10. *El polinomio*
$$p(x) = 1 + (x+1) + (x+1)x + (x+1)x(x-1),$$
es el polinomio interpolador de los puntos $(-1, 1)$, $(0, 2)$, $(1, 5)$, $(2, 16)$. *Calcular el polinomio interpolador de los puntos* $(-2, 8)$, $(-1, 1)$, $(0, 2)$, $(1, 5)$, $(2, 16)$, *sin recalcular todo el polinomio de interpolación*
11. 👩‍👩‍👧 *Obtener el polinomio interpolador de la siguiente tabla, usando que el polinomio interpolador de los 9 primeros puntos es* x^2*:*

x	1	2	3	4	5	6	7	8	9	10
y	1	4	9	16	25	36	49	64	81	1

12. *Dados $x_0 < x_1 \in \mathbb{R}$, $y_0, y_0', y_1, y_1' \in \mathbb{R}$, obtener el polinomio interpolador (polinomio interpolador de Hermite) determinado por $p(x_0) = y_0$, $p'(x_0) = y_0'$, $p(x_1) = y_1$, $p'(x_1) = y_1'$.*
13. *Obtener los polinomios interpoladores de las siguientes tablas:*

(a).

x	-1	0	1
y	0	3	0
y'	1	0	2

(b).

x	0	1	2
y	-2	0	2
y'	0	3	0

(c).

x	-1	0	1	2
y	2	0	0	2
y'	0	1	1	0

14. *Obtener la interpolación cúbica a trozos de Hermite de las siguientes tablas:*

(a).

x	-1	0	1
y	0	3	0
y'	1	0	2

(b).

x	0	1	2
y	-2	0	2
y'	0	3	0

(c).

x	-1	0	1	2
y	2	0	0	2
y'	0	1	1	0

15. *Obtener los splines cúbicos naturales que interpolan las siguientes tablas*

(a).

x	-1	0	1
y	0	3	0

(b).

x	0	1	2
y	-2	0	2

(c).

x	-1	0	1	2
y	-1	-2	0	2

16. *Obtener la curva polinomial que interpola los siguientes puntos, considerados en los momentos $t = 0$, $t = 1$ y $t = 2$:*

(a).

x	-1	0	-1
y	0	3	0

(b).

x	0	-1	2
y	-2	0	2

(c).

x	-1	0	1
y	-1	-2	0

(d).

x	-1	0	-1
y	1	1	1

(e).

x	0	1	-1
y	2	0	2

(f).

x	-1	0	-1
y	1	2	-3

17. *Sea $p_0, p_1, \ldots$ una sucesión de polinoios tal que, para cada n, p_n tiene exactamente grado n. Demiestre que la sucesión es linealmente independiente.*
18. 👩‍👩‍👧 *Demuestre que la matriz de Hilbert, con elementos $a_{ij} = (1 + i + j)^{-1}$ es una matriz de Gram para las funciones $1, x, x^2, \ldots, x^{n-1}$, es decir, es la matriz del sistema lineal que se obtiene al plantear el problema de aproximación para la norma inducida por cierto producto escalar.*
19. *Obtener el polinomio de grado menor o igual a 3 que mejor aproxima $f(x) = e^x$ con la norma inducida por el producto escalar*

$$\langle f, g \rangle = \int_{-1}^{1} f(x)g(x)\,dx.$$

20. *Obtener el polinomio de grado menor o igual a 3 que mejor aproxima $f(x) = e^x$ con la norma inducida por el producto escalar*

$$\langle f, g \rangle = \int_{1}^{2} f(x)g(x)\,dx.$$

21. 💻 *Obtener el polinomio de grado menor o igual a 3 que mejor aproxima $f(x) = e^x$ con la norma inducida por el producto escalar*

$$\langle f, g \rangle = \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1 - x^2}}\,dx.$$

Tema 5 Derivación e integración numéricas

Dada una función f y $a, b \in \mathbb{R}$, queremos aproximar el valor de

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

En muchas ocasiones no existe una primitiva de f, por lo que necesitaremos otro método para evaluar la integral.

Supondremos que conocemos los valores de la función en $n+1$ coordenadas $x_0 < x_1 < \ldots < x_n \in [a, b]$, es decir, que tenemos los puntos

$$(x_0, f(x_0)), (x_1, f(x_1)), \ldots (x_n, f(x_n)),$$

Obtendremos valores aproximados de $I(f)$ a partir de los valores de la función en los puntos anteriores.

5.1 Fórmulas de cuadratura

Denominamos **fórmula de cuadratura** *a una expresión del tipo*

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

- $x_0 < x_1 < \ldots < x_n \in [a, b]$ *son los* **nodos de cuadratura**.
- $a_0, a_1, \ldots, a_n \in \mathbb{R}$ *son los* **pesos de cuadratura**.

El **error de truncamiento** *de la fórmula es*

$$E[f] = \int_a^b f(x)dx - Q[f].$$

El **orden** *o* **grado de precisión** *de la fórmula de cuadratura es el mayor número natural m de modo que $E[P] = 0$ para cualquier polinomio P de grado $\leq m$. Se dice que el orden es exactamente m si existe un polinomio P de grado m tal que $E[P] \neq 0$*

Proposición 5.1

Una fórmula de cuadratura es de orden m si y sólo si

$$E[x^k] = 0, \quad k = 0, 1, \ldots, m.$$

El orden es exactamente m si además

$$E[x^{m+1}] \neq 0.$$

♠

Demostración Ejercicio. □

5.1.1 Fórmulas de cuadratura de tipo interpolatorio

Las **fórmulas de cuadratura de tipo interpolatorio** *consisten en, dados $a < b \in \mathbb{R}$, $x_0, \ldots, x_n \in [a, b]$ y f integrable en $[a, b]$, considerar el polinomio de interpolación de $f(x)$ en los puntos $x_0, x_1, \ldots, x_n$, P_n, y definir la fórmula de cuadratura*

$$\int_a^b f(x)dx \approx Q[f] = \int_a^b P_n(x)dx.$$

Nótese que en particular el grado de precisión es, al menos, el grado del polinomio interpolador, n.

En el caso de tomar puntos equiespaciados se denominan **Fórmulas de Newton-Cotes**. *Si los puntos equiespaciados incluyen a y b se denominan fórmulas cerradas. En ese caso, los nodos de cuadratura son*

$$x_0 = a, x_1 = a + h, \ldots, x_n = a + nh = b, \quad h = (b-a)/n.$$

Sean $l_i(x)$, $0 \leq i \leq n$ los polinomios de Lagrange de $x_0, x_1, \ldots, x_n$. Entonces

$$P_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x).$$

Sustituyendo en la fórmula de cuadratura,

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

donde

$$a_i = \int_a^b l_i(x)\, dx.$$

Recordemos que si P_n es el polinomio interpolador, el error de interpolación es

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{i=0}^{n} (x - x_i), \quad \xi_x \in [a, b].$$

Por tanto, el error de cuadratura será

$$E[f] = \int_a^b \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{i=0}^{n} (x - x_i)\, dx, \quad \xi_x \in [a, b].$$

Otra formulación equivalente es, utilizando la interpolación de Newton,

$$E[f] = \int_a^b f[x_0, x_1, \ldots, x_n, x] \prod_{i=0}^{n} (x - x_i)\, dx.$$

5.1.1.1 Trapecio simple

Tomando $x_0 = a$, $x_1 = b$, obtenemos

$$\boxed{\int_a^b f(x)dx \approx \frac{b-a}{2}(f(a) + f(b))}$$

Y el error de cuadratura es

$$\boxed{E[f] = -\frac{f''(\xi)}{12}(b-a)^3,\ \xi \in [a, b]}$$

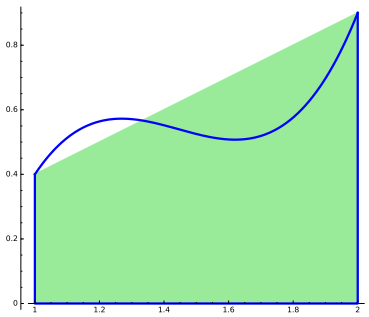


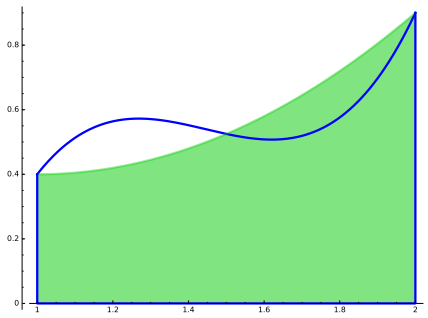
Figura 5.1: Trapecio simple

El error de curvatura se obtiene usando que $(x-a)(b-x)$ es positiva en $[a, b]$ y por el Teorema del valor medio del cálculo integral, el término de la derivada sale fuera de la integral.

5.1.1.2 Regla de Simpson

Tomando $x_0 = a$, $x_1 = \frac{a+b}{2}$, $x_2 = b$, $h = (b-a)/2$, *obtenemos la regla de Simpson:*

$$\boxed{\int_a^b f(x)dx \approx \frac{h}{3}(f(a) + 4f(x_1) + f(b))}$$



Proposición 5.2

Sea $f \in \mathcal{C}^4[a,b]$. *Entonces existe* $\xi \in (a,b)$ *tal que el error cometido con la regla de Simpson es*

$$\int_a^b f(x)\,dx - \frac{h}{3}(f(a) + 4f(x_1) + f(b)) = -\frac{f^{(4)}(\xi)}{90}h^5.$$

♠

Demostración Desarrollamos la función de h en $h = 0$ la expresión del error

$$\int_a^{a+2h} f(x)\,dx - \frac{h}{3}(f(a) + 4f(a+h) + f(a+2h)).$$

Si

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

entonces

$$F(x) = F(a)+F'(a)(x-a)+F''(a)(x-a)^2/2+F'''(a)(x-a)^3/3!+F''''(a)(x-a)^4/4!+F'''''(\xi)(x-a)^5/5!$$

Luego

$$\int_a^{a+2h} f(x)\,dx = 0 + f(a)2h + f'(a)2h^2 + f''(a)4h^3/3 + f'''(a)2h^4/3 + f''''(\xi)32h^5/5!$$

Po otro lado

$$(h/3)(f(a)+4f(a+h)+f(a+2h)) = 0+f(a)2h+f'(a)2h^2+f''(a)4h^3/3+f'''(a)2h^4/3+f''''(\xi)(4+16)h^5/(34!)$$

Es decir, el resto es

$$(32\cdot 3 - 20\cdot 5)/(3\cdot 5!) = -4/(3\cdot 5!) = -1/(3\cdot 5\cdot 3\cdot 2) = -1/90.$$

□

5.1.2 Grado de precisión de las fórmulas de tipo interpolatorio

Proposición 5.3

Una fórmula de cuadratura con $(n+1)$-nodos distintos,

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

es de interpolación si y sólo si tiene grado de precisión al menos n.

♠

Demostración

En primer lugar, si la fórmula es de interpolación, es decir,

$$I(f) = \int_a^b P_n(x)\, dx,$$

donde $P_n(x)$ es el polinomio de interpolación de f en $x_0, x_1, \ldots, x_n$, entonces, si f es un polinomio de grado $\leq n$, tenemos que $P_n = f$ y la fórmula es exacta.

Recíprocamente, supongamos que la fórmula tiene grado de precisión al menos n. Entonces, si $E_n(f)$ es el error de interpolación, $E_n(x^k) = 0$, $0 \leq k \leq n$, es decir,

$$\frac{1}{k+1}\left(b^{k+1} - a^{k+1}\right) = \sum_{i=0}^{n} a_i x_i^k, \quad 0 \leq k \leq n.$$

El sistema anterior es compatible determinado (su determinante es el de Vandermonde).

Por otra parte, si $\bar{a}_i$ son los pesos de cuadratura de la fórmula de cuadratura de tipo interpolatorio, como tiene grado de precisión al menos n, tenemos que son solución del sistema anterior. Luego la fórmula de cuadratura anterior es de tipo interpolatorio.

□

5.1.3 Fórmulas de cuadratura compuestas

Las **reglas compuestas** *consisten en descomponer el intervalo en subintervalos y aproximar la integral aproximando la integral en cada subintervalo con un método simple.*

1. *Si aplicamos la regla del trapecio simple en cada subintervalo, tenemos la regla del trapecio compuesta.*
2. *Si el número de subintervalos es par y aplicamos la regla de Simpson en cada par de subintervalos, tendremos la regla de Simpson compuesta.*

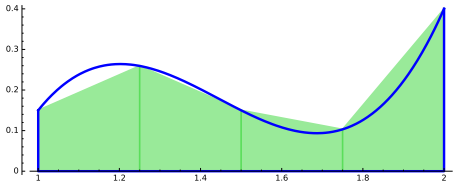
5.1.3.1 Regla del Trapecio Compuesta

Dado $h = (b-a)/n$, tomamos los nodos $x_i = a + ih$, $0 \leq i \leq n$. Entonces

$$\boxed{\int_a^b f(x)dx \approx \frac{h}{2}\left(f(x_0) + 2f(x_1) + \cdots + 2f(x_{n-1}) + f(x_n)\right).}$$

Y el error de cuadratura

$$\boxed{E[f] = -\frac{f''(\xi)}{12}(b-a)h^2,\ \xi \in [a,b].}$$



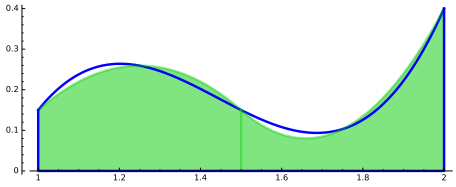
5.1.3.2 Regla de Simpson Compuesta

Dado $n = 2k$, $h = (b-a)/n$, *tomamos* $x_i = a + ih$, $0 \leq i \leq n$. *Entonces*

$$\int_a^b f(x)dx \approx \frac{h}{3}\left(f(x_0) + 2\sum_{i=1}^{k-1} f(x_{2i}) + 4\sum_{i=0}^{k-1} f(x_{2i+1}) + f(x_n)\right).$$

Y el error de cuadratura es

$$E[f] = -\frac{f^{(4)}(\xi)}{180}h^4(b-a),\ \xi \in [a,b].$$



5.2 Cuadratura adaptativa

Los métodos anteriores toman los nodos de cuadratura equiespaciados. Esto hace que se dedique el mismo esfuerzo computacional a cada subintervalo, independientemente de cómo se comporte la función en el mismo.

El **método de cuadratura adaptativa** *estima el error cometido en cada intervalo. En caso de que dicha estimación sea menor que una tolerancia fijada, se usa la aproximación obtenida. Sin embargo, si la estimación del error es mayor que la tolerancia, se divide el intervalo en dos, asignando a cada uno la mitad de la tolerancia.*

Una vez subdividido un intervalo, se aplica el mismo proceso para cada una de las dos subdivisiones.

Vamos a ilustrarlo usando el método de Simpson como fórmula de cuadratura para un intervalo. Fijada f *y* $a, b \in \mathbb{R}$, *denotaremos*

$$S(a,b) = \frac{b-a}{6}\left(f(a) + 4f((a+b)/2) + f(b)\right).$$

Entonces

$$\int_a^b f(x)\,dx - S(a,b) = -\frac{f^{(4)}(\xi_1)}{90}h^5, \quad \xi_1 \in [a,b].$$

Por otra parte, si $c = (a+b)/2$,

$$\int_a^b f(x)\,dx - S(a,c) - S(c,b) = -\frac{f^{(4)}(\xi_2)}{16 \cdot 90}h^5, \quad \xi_2 \in [a,b].$$

Supongamos que $f^{(4)}(\xi_1) \approx f^{(4)}(\xi_2)$. *Despejando de las dos ecuaciones anteriores, obtenemos*

$$S(a,b) - S(a,c) - S(c,b) = \frac{15}{16}\frac{f^{(4)}(\xi_1)}{90}h^5.$$

De donde

$$-\frac{f^{(4)}(\xi_1)}{16\cdot 90}h^5 = -\frac{S(a,b) - S(a,c) - S(c,b)}{15}.$$

Es decir, $(S(a,c) + S(c,b) - S(a,b))/15$ *es una estimación del error cometido aplicando Simpson con dos intervalos.*

Resumiendo, el procedimiento (recursivo) es el siguiente:

1. *Dada una tolerancia* T, *se comprueba si el error estimado* $(S(a,c) + S(c,b) - S(a,b))/15$ *en valor absoluto es menor que dicha tolerancia.*
2. *Si es menor, se toma como aproximación de la integral*
$$\int_a^b f(x)dx \approx S(a,c) + S(c,b) + \frac{S(a,c) + S(c,b) - S(a,b)}{15}.$$
3. *Si es mayor, se consideran las integrales en los intervalos* $[a,c]$ *y* $[c,b]$, *asignando a cada una una tolerancia* $T/2$ *y se procede análogamente. Una vez aproximadas las dos integrales, se suman los valores de las dos aproximaciones y ese es el valor utilizado como aproximación de la integral entre* a *y* b.

5.3 Fórmulas de Cuadratura Gaussiana

Consideremos una fórmula de cuadratura sobre $n+1$ *nodos,*

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

Vamos a buscar los valores de los pesos y de los nodos para que el orden de la fórmula de cuadratura sea máximo. Sabemos que para tener grado de precisión al menos n, *la fórmula de cuadratura será de tipo interpolatorio, por lo que sólo tendremos que obtener los nodos* $x_0, \ldots, x_n$.

Proposición 5.4

Dados $n \in \mathbb{N}$, $r \geq 0$, *una fórmula de cuadratura de tipo interpolatorio,* $I(f) = \sum_{i=0}^{n} a_i f(x_i)$, *tiene grado de precisión* $n+r$ *si y sólo si*

$$\int_a^b x^j \prod_{i=0}^{n}(x - x_i)\,dx = 0, \quad 0 \leq j \leq r-1.$$

♠

Demostración

Si la fórmula de cuadratura tiene grado de precisión $n+r$ entonces para todo polinomio de grado menor o igual a $n+r$ el error es cero. En particular, para los siguientes polinomios

$$P_j(x) = x^j \prod_{i=0}^{n}(x - x_i), \quad 0 \leq j \leq r-1.$$

Entonces su integral en $[a,b]$ coincide con la fórmula de cuadratura, que es cero pues se anulan en los nodos.

Supongamos que se verifica la fórmula. Denotemos

$$\Pi_n(x) = \prod_{i=0}^{n}(x - x_i)$$

La fórmula será de grado $n+r$ si el error de cuadratura de todo polinomio de grado menor o igual a $n+r$, P, es cero. Ahora bien, $P = Q\Pi_n + R$, donde Q tiene grado menor o igual a r y R tiene grado menor o igual a n.

Entonces

$$E[P] = E[Q\Pi_n] + E[R].$$

Por hipótesis, $E[Q\Pi_n] = 0$ y por ser una fórmula de cuadratura de tipo interpolatorio, $E[R] = 0$.

□

La Proposición anterior muestra que para obtener una fórmula de cuadratura de grado $n+r$ basta tomar

$$\Pi_n = \prod_{i=0}^{n}(x - x_i)$$

ortogonal a x^j, $0 \leq i \leq r-1$ con el producto escalar

$$\langle g, h\rangle = \int_a^b g(x)h(x)\,dx.$$

Consideremos una sucesión de polinomios ortogonales (con el producto escalar anterior), $\{Q_k\}_{k=0,1,\ldots}$ tal que el grado de Q_k sea k.

Proposición 5.5

Fijado $n \in \mathbb{N}$ sean $x_0, \ldots, x_n$ las raíces de Q_{n+1}. Entonces la fórmula de cuadratura de tipo interpolatorio cuyos nodos son $x_0, \ldots, x_n$ tiene orden $2n+1$.

♠

Demostración Hay que probar que las raíces han de ser reales. Si no lo fuesen, entonces tendríamos un polinomio de grado menor con el mismo número de raíces reales, que al multiplicarlo por Q_{n+1} nos daría un polinomio positivo y por tanto no ortogonal. □

Proposición 5.6

Los pesos de cuadratura son positivos.

♠

Demostración Ver ejercicios. □

Teorema 5.1

Si f es continua en $[a, b]$, entonces las fórmulas de cuadratura gaussianas convergen a la integral.

♡

Demostración Por el Teorema de Aproximación de Weierstrass, existe un polinomio tal que $|f - p| < \epsilon$. Para n suficientemente grande, las fórmulas de cuadratura son exactas para ese polinomio, luego

$$\left|\int_a^b f(x)\,dx - \sum A_i f(x_i)\right| = \left|\int_a^b f(x) - p(x)\,dx\right| + \left|\sum A_i(p(x_i) - f(x_i))\right| \leq (b-a)\epsilon + \epsilon\sum A_i = 2(b-a)\epsilon.$$

□

Vamos a calcular las fórmulas de cuadratura gaussianas para la integral $\int_{-1}^{1} f(x)\,dx$, denominadas fórmulas de Gauss-Legendre.

Para ello, comenzamos calculando los polinomios de Legendre, que forman base ortogonal. Los dos primeros serán $Q_0(x) = 1$, $Q_1(x) = x$. Los siguientes se pueden calcular mediante la fórmula recursiva

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x).$$

Ahora bastará tomar como nodos las raíces de dichos polinomios y como pesos los dados por ser una fórmula de tipo interpolatorio.

Tema 5 Ejercicios

1. *La regla del trapecio aplicada a* $\int_0^2 f(x)\,dx$ *nos da el valor 4 y la de Simpson nos da el valor 2. ¿Cuál es el valor de* $f(1)$*?*
2. *Obtener una función* $f(x)$ *tal que la regla del trapecio de el valor exacto de* $\int_{-1}^1 f(x)\,dx$*, pero la de Simpson no sea exacta.*
3. *Obtener el grado de precisión de la fórmula de cuadratura*
$$\int_{-1}^1 f(x)\,dx \approx f\left(-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right).$$
4. *Sea* $h = (b-a)/3$*,* $x_0 = a$*,* $x_1 = a+h$*,* $x_2 = b$*. Obtener el grado de precisión de la fórmula de cuadratura*
$$\int_a^b f(x)\,dx = \frac{9}{4}hf\left(x_1\right) + \frac{3}{4}hf\left(x_2\right).$$
5. *La fórmula de cuadratura*
$$\int_{-1}^1 f(x)\,dx \approx c_0 f(-1) + c_1 f(0) + c_2 f(1)$$
es exacta para todos los polinomios de grado ≤ 2*. Determinar* c_0*,* c_1*,* c_2*.*
6. *Encuentra* c_0, c_1, x_1 *tales que la fórmula de cuadratura*
$$\int_0^1 f(x)\,dx \approx c_0 f(0) + c_1 f(x_1),$$
tenga el máximo grado de precisión posible.
7. *Encuentra* x_0, c_1, x_1 *tales que la fórmula de cuadratura*
$$\int_0^1 f(x)\,dx \approx \frac{1}{2} f(x_0) + c_1 f(x_1),$$
tenga el máximo grado de precisión posible.
8. *Aproximar aplicando la regla de Simpson*
$$\int_1^3 \int_1^2 \ln(x+2y)\,dy\,dx.$$
9. *Encontrar el valor de* h *para que el error de cuadratura con la regla del trapecio compuesto para las siguientes integrales sea menor que* 10^{-4}*.*
   (a). $\int_0^2 e^{x^2}\,dx$.
   (b). $\int_0^1 \cos(\pi^2 x^2)\,dx$.
   (c). $\int_0^2 (x+4)^{-1}\,dx$.
   (d). $\int_1^2 x \ln x\,dx$.
10. *Encontrar el valor de* h *para que el error de cuadratura con la regla de Simpson para las siguientes integrales sea menor que* 10^{-4}*.*
    (a). $\int_0^2 e^{x^2}\,dx$.
    (b). $\int_0^1 \cos(\pi^2 x^2)\,dx$.
    (c). $\int_0^2 (x+4)^{-1}\,dx$.
    (d). $\int_1^2 x \ln x\,dx$.
11. *Determinar* a, b, c, d *para que la siguiente fórmula de cuadratura sea de grado tres*
$$\int_{-1}^1 f(x)\,dx \approx af(-1) + bf(1) + cf'(-1) + df'(1).$$

12. *Determinar a, b, c, d, e para que la siguiente fórmula de cuadratura sea de grado cuatro*

$$\int_{-1}^{1} f(x)\,dx \approx af(-1) + bf(0) + cf(1) + df'(-1) + ef'(1).$$

13. *Aproximar por la fórmula de Gauss-Legendre con 3 nodos las siguientes integrales:*

$$\int_{-1}^{1} \sin(\pi x)\,dx, \quad \int_{1}^{3} \ln x\,dx, \quad \int_{1}^{2} e^{x^2}\,dx.$$

14. 👥 *Demostrar que los pesos de las fórmulas de cuadratura gaussiana son siempre positivos*[1].

15. *Calcular mediante cuadratura adaptativa una aproximación con error estimado menor de 10^{-1} de las siguientes integrales:*

(a). $\int_0^1 (1+x^3)^{1/2}\,dx$
(b). $\int_1^2 (1+x^4)^{-1}\,dx$
(c). $\int_0^2 (1+x^3)^{-1/2}\,dx$
(d). $\int_0^{\pi/2} \cos(x)\,dx$
(e). $\int_0^1 xe^x\,dx$
(f). $\int_0^2 e^{x^2}\,dx$

[1] Pista: considerar los polinomios $P_j(x) = \prod_{i\neq j}(x - x_i)^2/(x_j - x_i)^2$.

Bibliografía

[1] Fausto Saleri Alfio Quarteroni Riccardo Sacco. «Numerical Mathematics». En: *Text in Applied Mathematics. Springer.* (2006).