Análisis Numérico Avanzado

Ahmed Ould Khaoua, Ph. D.

1 de agosto de 2010

Índice general

Introducción

En la literatura Matemática, hay dos tipos de textos de Análisis Numérico: los que tienen la fama de ser teóricos y generalmenente destinados a estudiantes avanzados de Matemáticas por el alto nivel de conocimientos, particularmente en Análisis Funcional, requerido para poder leerlos. Estos textos generalmente se usan con estudiantes que hayan visto un curso elemental en el área. El segundo tipo es de los textos que hacen un énfasis importante sobre la parte algorítmica y computacional. El público potencial de estos libros son los estudiantes de Ingeniería que no buscan más que resolver numéricamente un problema dado con algoritmos ya existentes. Así que muchos resultados matemáticos en estos libros son enunciados brevemente o simplemente omitidos sobretodo los resultados de convergencia.

Este libro se puede considerar como entre los dos tipos de textos citados. Así que los estudiantes de Matemáticas e Ingeniería y profesores en el área de Análisis Numérico pueden perfectamente seguir los distintos capítulos del texto sin perderse en los resultados teóricos complejos ni aburrise con largas tablas de resultados numéricos de un algoritmo cuya convergencia no ha sido probada. El perfil general de lector no require sino un nivel básico en Matemáticas equivalente a Cálculo Vectorial o funciones de varias variables. Pero dado que el texto es autosuficiente, hasta los resultados de dificultad relativa son demostrados rigurosamente.

La idea del libro nació cuando dicté por la primera vez el curso de Análisis Numérico de Honores en el departamento de Matemáticas de la Universidad de los Andes. La necesidad de redactar las notas del curso persistió en los semestres posteriores a medida que el esquema del curso se aclaraba cada vez que dictaba el curso.

Aunque los capítulos tienen cierta independencia entre sí, hay un hilo conductor a lo largo de este documento. El objetivo se trazó de manera que el lector puede resolver al final un problema sencillo de Ecuaciones con Derivadas Parciales con el método de Elementos Finitos en todos sus detalles empezando por la discretización del dominio, pasando por la construción del sistema lineal que resulta de la discretización y terminando con la resolución numércia de éste. Esta última parte que es la componente más numérica de la resolución del problema está desarollada en tres capítulos, a saber, Métodos Directos, Métodos Iterativos y Métodos basados en la optimización.

Los dos capítulos anteriores a estos forman la imprescindible base teórica de Álgebra Lineal para estudiar rigurosamente los algoritmos de dichos métodos.

Los dos primeros capíulos tienen como objetivo la integración numérica dado que en la solución de las Ecuaciones con Derivadas Parciales aparecen muchas integrales en la construción del sistema y su cálculo exacto es demasiado costoso o imposible.

Ahmed Ould Khaoua
Enero 2004

Capítulo 1

Interpolación

Ejemplo 1.1 (Ejemplo ilustrativo del problema)**.** Supongamos que tenemos una barra metálica recta de longitud l. Nos interesa conocer la temperatura a lo largo de la barra pero el dispositivo de medición disponible sólo puede dar la temperatura en los extremos de ésta.

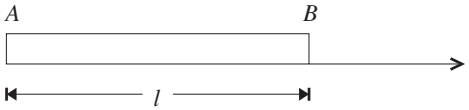

Figura 1.1.

Denotemos T_x a la temperatura en un punto M a distancia x de A en la barra. Conociendo T_0 y T_l queremos aproximar la temperatura $T_{l/2}$ en la mitad de la barra. El método intuitivo es tomar el promedio de T_0 y T_l, es decir

$$T_{l/2} \simeq \frac{1}{2}(T_0 + T_l).$$

Por ejemplo, si $T_0 = 10^\circ$ y $T_l = 50^\circ$,

$$T_{l/2} \simeq 30^\circ.$$

De manera similar $T_{l/4} \simeq 20^\circ$. De ese modo

$$T_x \simeq mx + b$$

y remplazando los datos en los extremos tenemos

$$T_0 = b, \qquad T_l = ml + T_0.$$

Así

$$m = \frac{T_l - T_0}{l}$$

y la ecuación de la aproximación de la temperatura en la barra es

$$T_x = \frac{T_l - T_0}{l}\, x + T_0.$$

Ahora, supongamos que la situación es diferente. El dispositivo de medición da la temperatura en los extremos y en la mitad de la barra, es decir tenemos $T_0, T_{l/2}, T_l$ y se requiere aproximar la temperatura en un punto M a distancia x de A en la barra.

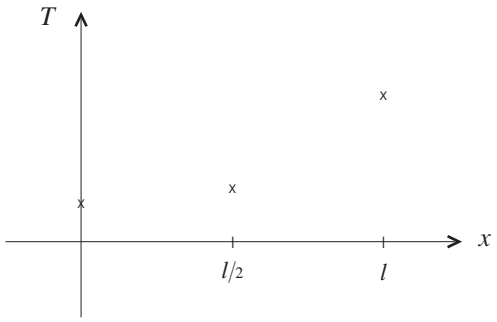

Figura 1.2.

Con los datos que se muestren en la figura (1.2), no se puede usar la anterior técnica de aproximación porque hay 3 datos y no son colineales. Una manera de hacerlo es tratar de encontrar una función cuya gráfica pase por los tres puntos. El problema de interpolación es encontrar tal función.

El problema general de la interpolacion se platea de la siguiente manera. Sea $I = [a, b]$ un intervalo en $\mathbb{R}$ y sean $x_0 < x_1 < \ldots < x_n$, $n + 1$ puntos de I. Supongamos que f es una función definida en I tal sus valores son conocidos en los puntos x_i, $i = 0, \ldots, n$.

Definición 1.1. Una función g se dice interpolante de f en I con respecto a los puntos x_i, $i = 0, \ldots, n$, si g se conoce en todo I y $g(x_i) = f(x_i)$, $i = 0, \ldots, n$. Si g es un polinomio, se le dice interpolante polinomial.

Nota 1.1.
En la práctica se usa más la interpolación polinomial porque es más fácil manejar polinomios en los cálculos.

1.1. Interpolación de Lagrange

Sean $I = [a, b]$, $x_0 < x_1 < \ldots < x_n$, puntos de I y f una función definida sobre I.

Problemática 1.1. Encontrar un polinomio P de grado menor o igual a n tal que

$$f(x_i) = P(x_i), \qquad i = 0, \ldots, n.$$

Definición 1.2. Los polinomios de Lagrange asociados a los puntos x_i, $i = 0, \ldots, n$, son los polinomios de grado n definidos por

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \left(\frac{x - x_j}{x_i - x_j} \,. \right)$$

Nota 1.2.

1. Para todo $i = 0, \ldots, n$, L_i es un polinomio de grado n.
2. L_i tiene n raíces reales que son $x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$.
3. Para todo $i, j = 0, \ldots, n$
$$L_i(x_j) = \delta_{ij},$$
donde δ_{ij} es el símbolo de Kronecker definido por
$$\delta_{ij} = \begin{cases} 0, & \text{si } i \neq j, \\ 1, & \text{si } i = j. \end{cases}$$

1.1.1. Existencia y unicidad del polinomio de interpolación

Teorema 1.1. *Dados $n + 1$ puntos x_i, $i = 0, \ldots, n$, distintos de un intervalo $I = [a, b]$ y $n + 1$ números reales y_i, $i = 0, \ldots, n$, entonces existe un único polinomio P_n de grado menor o igual a n tal que*

$$P_n(x_i) = y_i, \qquad i = 0, \ldots, n.$$

Nota 1.3.

En este teorema los números reales y_i, $i = 0, \ldots, n$, representan los valores $f(x_i)$, $i = 0, \ldots, n$, cuando se trata de la interpolación de la función f y en este caso P_n se llama el polinomio de interpolacion de lagrange de f en los puntos x_i, $i = 0, \ldots, n$.

Demostración. Sea

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

P_n es un polinomio de grado menor o igual a n siendo suma de los polinomios de Lagrange L_i, $i = 0, \ldots, n$. Además, para cualquier x_j, $j = 0, \ldots, n$,

$$\begin{aligned} P_n(x_j) &= \sum_{i=0}^{n} y_i \, L_i(x_j), \\ &= \sum_{i=0}^{n} y_i \, \delta_{ij}, \\ &= y_j. \end{aligned}$$

Es decir P_n interpola f en los puntos $x_0, x_1, \ldots, x_n$.

Mostremos la unicidad. Designamos $d^\circ P$ el grado de P. Supongamos que P_n y Q_n son dos polinomios tales que $d^\circ P_n \leq n$, $d^\circ Q_n \leq n$ y $P_n(x_i) = Q_n(x_i)$, $i = 0, \ldots, n$. Sea el polinomio

$$W_n(x) = P_n(x) - Q_n(x).$$

Notamos que $d^\circ W_n \leq n$ y $W(x_i) = 0$ para todo $i = 0, \ldots, n$. Así, W_n es un polinomio de grado menor o igual a n con $(n+1)$ raíces, lo que es imposible a menos que $W_n(x) \equiv 0$ entonces, $Q_n(x) = P_n(x)$ y P_n es único. □

Nota 1.4.
En el teorema anterior el polinomio de interpolación de Lagrange no se presenta en la forma

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0.$$

Eso representa una dificultad numérica mayor por necesitarse muchas operaciones para calcular $P_n(\alpha)$ para un α real dado. Existen algoritmos que permiten la reducción de los el cálculos para evaluar $P_n(\alpha)$ (Ver problemas).

Ejercicio 1.1. Calcular el número de operaciones (suma, multiplicación) necesario para calcular $P_n(\alpha)$.

Nota 1.5.
En el ejemplo de la aproximación de la temperatura se construyó un polinomio de grado 1 (caso lineal) cuando se usaron las temperaturas medidas en los extremos.

$$P_1(x) = \frac{T_l - T_0}{l}\, x + T_0$$

En el caso de mediciones en tres puntos $T_0, T_{l/2}, T_l$ se construye un polinomio de grado 2,

$$P_2(x) = T_0\, L_0(x) + T_{l/2}\, L_1(x) + T_l\, L_2(x),$$

donde

$$\begin{aligned} L_0 &= \frac{(x - l/2)(x - l)}{(-l/2)(-l)} = \frac{1}{l^2}\,(2x - l)(x - l), \\ L_1 &= \frac{x(x - l)}{l/2(-l/2)} = -\frac{4}{l^2}\, x(x - l), \\ L_2 &= \frac{x(x - l/2)}{l(l/2)} = \frac{1}{l^2}\, x(2x - l). \end{aligned}$$

El polinomio de interpolación de Lagrange para el problema es

$$P_2(x) = \frac{T_0}{l^2}\,(2x - l)(x - l) - \frac{4T_{l/2}}{l^2}\, x(x - l) + \frac{T_l}{l^2}\, x(2x - l).$$

1.2. Interpolación de Hermite

Vimos que con el conocimiento de las parejas $(x_i, f(x_i))$, $x_i \in I = [a, b]$, $i = 0, \ldots, n$, se puede hallar el polinomio de interpolación de grado menor o igual a n. Supongamos ahora que se quiere encontrar un polinomio "más parecido" a f usando información adicional sobre f en los puntos x_i, $i = 0, \ldots, n$.

Lo más natural en este caso es pensar en las derivadas de f en los puntos x_i, $i = 0, \ldots, n$.

El problema se plantea de la siguiente manera. Sean $I = [a, b]$ y $x_0 < x_1 < \ldots < x_n$ $(n+1)$ puntos en I. Sean $k_0, k_1, k_2, \ldots, k_n$, $n+1$ enteros no negativos y f una función definida en I tal que los números

$$f^{(j)}(x_i), \qquad i = 0, \ldots, n, \quad j = 0, \ldots, k_i,$$

son definidos y conocidos. Denotemos igualmente $m = n + k_0 + \ldots + k_n$.

Nota 1.6.
$f^{(j)}(x_i)$ denota la $j-$ésima derivada de f en el punto x_i y $f^{(0)} = f$.

Problemática 1.2. Encontrar un polinomio P_m de grado menor o igual a m tal que

$$P_m{}^{(j)}(x_i) = f^{(j)}(x_i), \qquad i = 0, \ldots, n, \quad j = 0, \ldots, k_i. \tag{1.1}$$

Definición 1.3. Un polinomio como (1.1) se llama polinomio de interpolación de Hermite de la función f en los puntos x_i, $i = 0, \ldots, n$, con respecto a los enteros k_i, $i = 0, \ldots, n$.

Teorema 1.2. *El polinomio de interpolación de Hermite existe y es único.*

Antes de demostrar este teorema introducimos la noción de raíz múltiple y miramos su relación con las derivadas de polinomios.

Definición 1.4. Sea $P(x)$ un polinomio con coeficientes en $\mathbb{K}$ ($\mathbb{K}$=$\mathbb{R}$ o $\mathbb{C}$) y sean $\alpha \in \mathbb{K}$ un escalar y $m \in \mathbb{N}$. Se dice que α es una raíz de P de multiplicidad m si $(x-\alpha)^m$ divide a $P(x)$ pero $(x-\alpha)^{m+1}$ no lo divide.

En otras palabras, α es raíz de multiplicidad m de P si $P(x)$ se escribe en la forma

$$P(x) = (x-\alpha)^m \, Q(x),$$

con $Q(\alpha) \neq 0$.

Lema 1.1. *α es una raíz de multiplicidad m del polinomio P si y sólo si para todo k, $0 \leq k \leq m-1$,*

$$P^{(k)}(\alpha) = 0 \quad y \quad P^{(m)}(\alpha) \neq 0. \tag{1.2}$$

Demostración. Sea P un polinomio de grado n. Aplicando la formula de Taylor de orden $n+1$ al polinomio P en un punto α se tiene

$$P(x) = \sum_{i=0}^{n} \frac{(x-\alpha)^i}{i!} \, P^{(i)}(\alpha).$$

Es evidente que si $P^{(k)}(\alpha) = 0, k = 0, \ldots, m-1$, y $P^{(m)}(\alpha) \neq 0$ entonces α es raíz de multiplicidad m. Recíprocamente, derivar la ecuación

$$P(x) = (x - \alpha)^m \, Q(x),$$

permite deducir el resultado. □

Ahora demostremos el teorema (1.2).

Demostración. Si escribimos P_m el polinomio de interpolación de Hermite en la forma

$$P_m(x) = a_0 + a_1 x + \cdots + a_m x^m,$$

las ecuaciones (1.1) forman un sistema lineal de $m+1$ ecuaciones y $m+1$ incongnitas que son los coeficientes $a_i, i = 0, \ldots m$. De este modo el teorema quedará demostrado si probamos que este sistema lineal tiene una solución única.

Pero esto es equivalente a probar que el sistema lineal homogéneo asociado tiene la solución trivial como solución única. Tal sistema se escribe en la forma

$$P_m^{(j)}(x_i) = 0, \qquad i = 0, \ldots, n, \quad j = 0, \ldots, k_i. \tag{1.3}$$

Estas ecuaciones significan que para todo $i = 0, \ldots, n$, x_i es raíz de multiplicidad mayor o igual a $k_i + 1$ del polinomio P_m. En otras palabras P_m, que es un polinomio de grado m, tiene al menos $m+1$ raíces lo que significa que P_m es el polinomio nulo, es decir,la única solución posible del sistema (1.3) es la solución trivial. □

Nota 1.7.
Es evidente que la interpolación de Lagrange es un caso particular de la de Hermite, con $k_i = 0$, $i = 0, \ldots, n$.

1.3. Cálculo del error de interpolación

Sea P_m el polinomio de interpolación de una función f en los puntos x_i, $i = 0, \ldots, n$ del intervalo $I = [a, b]$ con respecto a los enteros k_i, $i = 0, \ldots, n$.

Problemática 1.3. Es natural preguntarse qué tan grande es el error que uno está cometiendo cuando aproxima $f(x)$ con $P_m(x)$ para un $x \in I$.

Notemos que sin condiciones o datos adicionales sobre la función f, no hay ninguna manera de controlar el error $e(x)$ definido por:

$$e(x) = |f(x) - P_m(x)| \, .$$

Miremos por qué, para el caso simple $k_i = 0$, $i = 0, \ldots, n$. Sean P_i, $i = 0, \ldots, n$, $n+1$ puntos del plano de coordenadas (x_i, y_i), $i = 0, \ldots, n$, donde $x_i \in I$,

$i = 0, \ldots, n$ y sea $\overline{x} \in I$. El polinomio de interpolación de Lagrange asociado a los puntos P_i $i = 0, \ldots, n$, queda fijo (por su unicidad) al fijar estos puntos. $i = 0, \ldots, n$. Es evidente que para M, un real positivo arbitrario dado, existe una infinidad de funciones f tales que

$$f(\overline{x}) = M \quad \text{y} \quad f(x_i) = y_i, \quad i = 0, \ldots, n.$$

Además, el polinomio de interpolación asociado a los puntos (x_i, y_i), $i = 0, \ldots, n$, tiene un valor fijo $P_m(\overline{x})$. De ese modo

$$\begin{aligned} e(\overline{x}) &= |f(\overline{x}) - P_m(\overline{x})|, \\ &\geq |f(\overline{x})| - |P_m(\overline{x})|, \\ &\geq M - P_m(\overline{x}). \end{aligned}$$

Así, $e(\overline{x})$ toma valores tan grandes como uno quiera.

Teorema 1.3. *Sean $I = [a, b]$, x_i, $i = 0, \ldots, n$, $n+1$ puntos de I, k_i, $i = 0, \ldots, n$, enteros no negativos, $m = n + \sum_{i=0}^{n} k_i$ y P_m el polinomio de interpolación de Hermite de la función f definida en I con respecto a x_i, $i = 0, \ldots, n$, y k_i, $i = 0, \ldots, n$. Si $f \in \mathcal{C}^{m+1}(I)$ entonces para todo $x \in I$, existe $\theta_x \in I$ tal que*

$$f(x) - P_m(x) = \frac{p_m(x)}{(m+1)!} f^{(m+1)}(\theta_x), \tag{1.4}$$

donde

$$p_m(x) = \prod_{i=0}^{n} (x - x_i)^{k_i+1}.$$

Demostración. Si x es uno de los x_i, $i = 0, \ldots, n$, los dos lados de la igualdad (1.4) valen cero. Ahora, si $x \neq x_i$, $i = 0, \ldots, n$, definamos la función polinomial Q_m en la variable t por

$$Q_m(t) = P_m(t) + \frac{f(x) - P_m(x)}{p_m(x)} p_m(t).$$

Notemos que $d^\circ Q_m \leq m + 1$ por la presencia de $p_m(t)$. Además, para un j entero positivo

$$Q_m{}^{(j)}(t) = P_m{}^{(j)}(t) + \frac{f(x) - P_m(x)}{p_m(x)} p_m^{(j)}(t).$$

Tenemos para todo $i = 0, \ldots, n$, para todo $j \leq k_i$

$$p_m^{(j)}(x_i) = 0.$$

Así

$$Q_m{}^{(j)}(x_i) = P_m{}^{(j)}(x_i) = f^{(j)}(x_i) \forall j \leq k_i$$

y también tenemos

$$Q_m(x) = f(x).$$

La función F, definida por

$$F(t) = Q_m(t) - f(t),$$

se anula $m+2$ veces y su derivada F' se anula al menos $m+1$ veces. Siguiendo el proceso deducimos que existe un θ_x en I tal que $F^{m+1}(\theta_x) = 0$ es decir

$$Q_m{}^{(m+1)}(\theta_x) = f^{(m+1)}(\theta_x).$$

Pero

$$Q_m{}^{(m+1)}(t) = 0 + \frac{f(x) - P_m(x)}{p_m(x)}\, p_m^{(m+1)}(t).$$

Como $p_m(t)$ es un polinomio mónico de grado $m+1$ deducimos que

$$p_m^{(m+1)} = (m+1)!.$$

Entonces

$$Q_m{}^{(m+1)}(\theta_x) = \frac{f(x) - P_m(x)}{p_m(x)}\,(m+1)!$$

lo que implica que

$$f(x) - P_m(x) = \frac{p_m(x)}{(m+1)!}\, f^{(m+1)}(\theta_x).$$

□

Corolario 1.1. *Si f satisface las condiciones del teorema (1.3) tenemos*

$$e(x) \leq \frac{M_{m+1}}{(m+1)!}\,|p_m(x)|$$

donde

$$M_{m+1} = \sup_{x\in I}\left|f^{(m+1)}(x)\right|.$$

Varias preguntas surgen examinando la desigualdad del corolario (1.1)

Nota 1.8.

1. Primera pregunta:
   ¿Bajo qué condiciones la cantidad

   $$\frac{M_{m+1}}{(m+1)!}\,|p_m(x)|$$

   tiende a cero independientemente de x cuando m tiende a infinito? En otras palabras, ¿la sucesión de polinomios de interpolación converge uniformemente a f?

2. Segunda pregunta:
   Si uno quiere interpolar una función f dada optimizando el error, el único margen de maniobra es la cantidad $|p_m(x)|$ que, en síntesis, depende sólo de los puntos x_i, $i = 0, \ldots, n$, y de los enteros k_i, $i = 0, \ldots, n$. ¿Cuál es entonces la mejor selección de estos parámetros para optimizar $e(x)$?

Ilustremos esta discusión con los siguientes ejemplos.

Ejemplo 1.2. Sean $I = [a, b]$ y $f \in \mathcal{C}^\infty(I)$. Supongamos que existe un real $M > 0$ tal que

$$|f^{(k)}(x)| \leq M, \qquad \forall x \in I, \quad \forall k \in \mathbb{N}.$$

Interpolando f en el intervalo I con respecto a los puntos x_i, $i = 0, \ldots, n$, y los enteros $k_0, \ldots, k_n$, el error $e(x)$ es acotado de la siguiente manera

$$e(x) \leq \frac{M}{(m+1)!} |p_m(x)|, \tag{1.5}$$

con $m = n + k_0 + \ldots + k_n$. Recordemos que

$$p_m(x) = \prod_{i=0}^{n} (x - x_i)^{k_i+1},$$

dado que los x_i están en $[a, b]$. Entonces para todo x en $[a, b]$

$$|x - x_i| \leq b - a.$$

Así

$$\prod_{i=0}^{n} |x - x_i|^{k_i+1} \leq (b - a)^{m+1}$$

y la fórmula (1.5) da

$$e(x) \leq \frac{M}{(m+1)!} (b - a)^{m+1}.$$

Sabemos que para todo real x

$$\lim_{n \to 0} \frac{x^n}{n!} = 0,$$

entonces

$$\lim_{m \to \infty} e(x) = 0, \qquad \forall x \in I.$$

Es decir que si se requiere una precisión muy alta en la aproximación con el polinomio de interpolación es suficiente aumentar el número de puntos de interpolación o los datos sobre las derivadas en estos puntos.

Ejemplo 1.3. En este ejemplo, llamado de Runge, veremos que la convergencia no es siempre posible si la función interpolada presenta algunas "patologías". Sean $I = [-5,\ 5]$ y

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

Para n entero positivo se realiza la interpolación de f en los puntos equidistantes

$$x_i = -5 + ih, \quad i = 0, \ldots, n, \quad h = 10/n$$

y se calcula el error en el punto $\hat{x} = 5 - h/2$. Los resultado obtenidos, para distintos valores de n, están resumidos en la tabla siguiente donde la divergencia de los polinomios de interpolación es evidente.

n	$f(\hat{x}) - P_n(\hat{x})$
2	-.08977122020
6	-.1864413536
10	-.6868592004
14	-2.822954659
18	-12.15212603
22	-53.74463799
26	-242.0061508
30	-1103.835361
34	-5083.699217
38	-23589.72229

Un analisis, no muy sencillo, de la función f muestra que sus derivadas tienen la propiedad

$$M_n \approx n!.$$

Recordemos que $M_n = \sup_{x \in I} |f^{(n)}(x)|$

1.4. Interpolación en los puntos de Chebyshev

Vimos en la nota (1.8.2) que la manera de reducir la cota superior del error de interpolación para una función dada es elegir los puntos x_i, $i = 0, \ldots, n$, y los enteros k_i, $i = 0, \ldots, n$, de manera que

$$\max_{a \le x \le b} |p_m(x)|$$

sea mínimo.

En esta sección vamos a buscar estos puntos para el caso de interpolación de Lagrange, es decir $k_i = 0$, $i = 0, \ldots, n$. Inicialmente damos los puntos para la interpolación en el intervalo $[-1, 1]$ y después, con un cambio de variable sencillo, generalizamos esta búsqueda para intervalos arbitrarios de tipo $[a, b]$, $-\infty < a < b < \infty$.

1.4.1. Polinomios de Chebyshev

Para $n \in \mathbb{N}$, definamos las funciones

$$T_n(x) = \cos(n \arc\cos x)$$

sobre el intervalo $[-1, 1]$.

Proposición 1.1. *Para todo $n \geq 1$,*

1. *La función $T_n(x)$ satisface la relación inductiva $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$;*

2. *las soluciones de la ecuación $T_n(x) = 0$ son*

$$x_k = \cos\left(\frac{2k+1}{2n}\,\pi\right), \qquad k = 0, \ldots, n-1;$$

3. *La función $T_n(x)$ toma valores entre -1 y 1 $T_n(x)$ y alcanza el máximo*

$$\max_{-1\leq x\leq 1} |T_n(x)| = 1$$

en los puntos

$$y_k = \cos\left(\frac{k\pi}{n}\right), \qquad k = 0, \ldots, n.$$

Nota 1.9.
Dado que $T_0(x) = 1, T_1(x) = x$ y por la parte (1) de la proposición deducimos que para todo $n \geq 0$, $T_n(x)$ es un polinomio de grado n en x sobre el intervalo $[-1, 1]$.

Definición 1.5. Los polinomios $T_n(x)$ se llaman polinomios de Chebyshev y sus raíces x_i $k = 0, \ldots, n-1$ se llaman puntos de Chebyshev.

Demostración.

1.

$$\begin{aligned} T_{n+1}(x) &= \cos\big((n+1)\arccos x\big), \\ &= \cos(n\arccos x)\cos(\arccos x) - \operatorname{sen}(n\arccos x)\operatorname{sen}(\arccos x), \\ &= xT_n(x) - \tfrac{1}{2}\cos\big((n-1)\arccos x\big) + \tfrac{1}{2}\cos\big((n+1)\arccos x\big), \\ &= xT_n(x) - \tfrac{1}{2}T_{n-1}(x) + \tfrac{1}{2}T_{n+1}(x). \end{aligned}$$

Así

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x);$$

2. La ecuación $T_n(x) = 0$ equivale a

$$\begin{aligned} n\arccos x &= (2k+1)\frac{\pi}{2}, \\ \text{es decir } \arccos x &= (2k+1)\frac{\pi}{2n}, \qquad k = 0, \ldots, n-1. \end{aligned}$$

Así

$$x = \cos\left(\frac{2k+1}{2n}\,\pi\right), \qquad k = 0, \ldots, n-1.$$

Queda mostrado el punto 2.

3. Sabemos que $|T_n(x)| \leq 1$, $\forall x \in [-1,1]$ y que $|T_n(-1)| = |T_n(1)| = 1$. Ahora analicemos los extremos $T_n(x)$ en el intervalo $]-1,1[$.

$$T_n'(x) = \frac{n}{\sqrt{1-x^2}} \operatorname{sen}(n \operatorname{arc\,cos} x)$$

es nula si y sólo si

$$\operatorname{arc\,cos} x = \frac{k\pi}{n}, \qquad k = 1, \ldots, n-1,$$

así que $|T_n(x)| = 1$ en los puntos

$$y_k = \cos\left(\frac{k\pi}{n}\right), \qquad k = 1, \ldots, n-1$$

y en estos puntos el polinomio T_n vale

$$T_n(y_k) = (-1)^k, \qquad k = 1, \ldots, n-1.$$

Agregamos los puntos $y_0 = -1,\ y_n = 1$ y (3) queda demostrado.

□

El teorema siguiente muestra que el error es minimo si los puntos de interpolación coinciden con las raices del polinomio de Chebyshev.

Teorema 1.4. *Para todo $n \geq 1$,*

1. *el coeficiente de x^n en T_n es 2^{n-1};*
2. *para todo polinomio Q_n mónico de grado n tenemos*

$$\frac{1}{2^{n-1}} = \max_{-1<x<1} \left|\frac{T_n(x)}{2^{n-1}}\right| \leq \max_{x\in[-1,1]} Q_n(x).$$

Demostración.

1. Es evidente a partir de la proposición;
2. Como
$$\max_{-1<x<1} |T_n(x)| = 1$$
deducimos que
$$\max_{-1<x<1} \left|\frac{T_n(x)}{2^{n-1}}\right| = \frac{1}{2^{n-1}}.$$
Supongamos que existe un polinomio Q_n mónico de grado n tal que
$$\max_{-1<x<1} |Q_n(x)| < \frac{1}{2^{n-1}}.$$

Ahora consideremos el polinomio

$$W_{n-1}(x) = Q_n(x) - \frac{T_n(x)}{2^{n-1}}.$$

W_{n-1} es un polinomio de grado $\leq n-1$.

En los puntos

$$y_k = \cos\frac{k\pi}{n}\,, \qquad k = 0, \ldots, n,$$
$$W_{n-1}(y_k) = Q_n(y_k) - \frac{(-1)^k}{2^{n-1}}\,, \qquad k = 0, \ldots, n.$$

Como

$$\max_{-1<x<1} |Q_n(x)| < \frac{1}{2^{n-1}},$$

deducimos que W_{n-1} cambia de signo al menos n veces, es decir, W_{n-1} tiene al menos n raíces lo que contradice el hecho que el grado de W_{n-1} sea menor o igual a $(n-1)$.

□

Nota 1.10.

1. La parte (2) de este teorema nos dice que T_n es la solución del problema de optimización dado por

$$\min_{Q_n \in P_n} \left[\max_{-1<x<1} |Q_n(x)| \right],$$

donde P_n es el conjunto de lo polinomios mónicos de grado $\leq n$;

2. Dado que los polinomios mónicos están totalmente determinados por sus raíces, si se escogen

$$x_k = \cos\left(\frac{(2k+1)\pi}{2n}\right), \qquad k = 0, \ldots, n-1,$$

(raíces del polinomio T_n) como puntos de interpolación de una función f en el intervalo $[-1, 1]$ aseguramos que el error de interpolación satiface

$$e(x) = \frac{M_n}{n!}\,\frac{1}{2^{n-1}}$$

pues

$$\max_{-1<x<1} |(x - x_0)\ldots(x - x_{n-1})| \leq \frac{1}{2^{n-1}}.$$

En el caso interpolación en un intervalo $[a, b]$ se puede hacer el cambio de variable

$$t = \frac{1}{2}\big[(b-a)x + a + b\big]$$

que envía el intervalo $[-1, 1]$ al intervalo $[a, b]$. Ilustremos esto con el siguiente ejemplo.

Ejemplo 1.4. Sea la función $f(x) = \cos(x^2 + 10)$ definida sobre el intervalo $[1,\ 2]$. Para un entero natural n, se construyen los polinomios de interpolación de Lagrange P_n en los puntos equidistantes $x_i = 1 + i/n$, $i = 0, \ldots, n$, y Q_n en los puntos de Chebyshev $t_i = \cos\left(\frac{2k+1}{2(n+1)}\,\pi\right), \qquad k = 0, \ldots, n$; respectivamente. Se calcula el error en los puntos $a_i = 1 + i/n^2$, $i = 0, \ldots, n^2$ para comparar la eficiencia de aproximación de las dos interpolaciones. Los resultados de la tabla siguiente corresponden a $n = 4$.

x	$f(x) - Q_n(x)$	$f(x) - P_n(x)$
1	.006151161452	-.004425697988
1.062500000	-.0052213045	-.1329372639
1.125000000	-.0059099095	-.2667803115
1.187500000	-.0015960724	-.4028073422
1.250000000	.0035636880	-.5370412570
1.312500000	.0069207574	-.6646923520
1.375000000	.0072519428	-.7802408017
1.437500000	.0045964842	-.8775996460
1.500000000	$1{,}10^{-9}$	-.9503708471
1.562500000	-.0048423344	-.9922021095
1.625000000	-.0080433166	-.9972444096
1.687500000	-.0080710720	-.9606993515
1.750000000	-.0043615418	-.8794316548
1.812500000	.0020447625	-.7526059079
1.875000000	.0078999083	-.5822894220
1.937500000	.0072541870	-.3739465808
2.000000000	-.0088422590	-.1367372182

Como se ve en la tabla, el error en el caso de los puntos de Chebyshev es de lejos menor que el error cometido en el caso de los puntos equiditantes

Capítulo 2

Integración numérica

Introducción

Sean $I = [a, b]$ y f una función definida sobre I tales que

$$J = \int_a^b f(x)\,dx \quad \text{existe},$$

es decir f es integrable en el sentido de Riemann.

Si f no se conoce en todos los puntos de I, o si f tiene una forma analítica compleja de manera que la integral J es difícil o imposible de calcular usando una función primitiva, es necesario buscar métodos numéricos para aproximar la integral.

Ejemplo 2.1.

1. No es posible calcular la integral

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

   utilizado el teorema fundamental de clalculo por que la la función $x \to e^{-x^2}$ no le conocemos una antiderivada.

2. Sabemos que

   $$\ln 2 = \int_0^1 \frac{1}{1+x}\,dx.$$

   Si queremos realizar una aproximación de $\ln 2$, la integración numérica será una alternativa evaluando la integral de manera aproximada.

3. Las funciones c, s definidas por las integrales

   $$c(t) = \int_0^t \cos dx(\frac{\pi}{2}w^2)dt, \qquad s(t) = \int_0^t \operatorname{sen} dx(\frac{\pi}{2}w^2)dt.$$

aparecen en los probleams de difracción de la luz. Los valores de estas funciones en un t dado, no se pueden evaluar sin aproximaciones por no conocer a funciones antiderivadas de $w \to \cos(w^2)$ y $w \to \operatorname{sen}(w^2)$.

La integración numérica consiste en aproximar este tipo de integrales en la forma siguiente

$$\int_a^b f(x)\,dx \simeq \sum_{i=0}^{n} c_i\, f(\alpha_i) \tag{2.1}$$

donde $c_i,\ \alpha_i$, $i = 0, \ldots, n$, son reales.

Definición 2.1. La fórmula (2.1) se llama fórmula de cuadratura numérica.

Problemática 2.1. Para construir un método de cuadratura hay que escoger un entero n y $2(n+1)$ reales (los α_i y c_i, $i = 0, \ldots, n$).

Como en toda construcción numérica, la mayor preocupación es que la aproximación de la fórmula (2.1) sea lo más precisa posible. Así que la búsqueda de los parámetros n, α_i y c_i, $i = 0, \ldots, n$, se hace con este objetivo.

Nota 2.1.

La formula de cuadratura numérica (2.1) no es sino una suma de Riemann de la función lo que es natural dado que la integral de Riemann es un limite de sumas de Riemann.

2.1. Métodos compuestos

Consideramos la integral

$$I = \int_a^b f(x)\,dx$$

que queremos aproximar. En estos métodos se divide el intervalo $I = [a, b]$ en k subintervalos

$$[t_i, t_{i+1}], \qquad i = 0, \ldots, k-1,$$

donde $t_0 = a$ y $t_k = b$. Así

$$\int_a^b f(x)\,dx = \sum_{i=0}^{k-1} \int_{t_i}^{t_{i+1}} f(x)\,dx.$$

Ahora en lugar de aproximar la integral global I, se aproximan cada una de las k integrales

$$I_i = \int_{t_i}^{t_{i+1}} f(x)\,dx, \qquad i = 0, \ldots, k-1.$$

La idea es utilizar un polinomio de interpolación que aproxime f en el intervalo $[t_i, t_{i+1}]$ para aproximar I_i. Dado que hay k integrales para aproximar es

mejor estandarizar el proceso de aproximación, es decir tratar de encontrar una forma de aproximación algorítmica. Para eso, se considera el cambio de variable

$$x = \frac{(t_{i+1} - t_i)t + (t_i + t_{i+1})}{2},$$
$$dx = \left(\frac{t_{i+1} - t_i}{2}\right) dt.$$

Entonces

$$\int_{t_i}^{t_{i+1}} f(x)\, dx = \left(\frac{t_{i+1} - t_i}{2}\right) \int_{-1}^{1} g_i(t)\, dt,$$

donde

$$g_i(t) = f\left(\frac{(t_{i+1} - t_i)t + (t_i + t_{i+1})}{2}\right).$$

Si denotamos

$$h_i = t_{i+1} - t_i,$$
$$\int_{t_i}^{t_{i+1}} f(x)\, dx = \frac{h_i}{2} \int_{-1}^{1} g_i(t)\, dt.$$

El problema se reduce entonces a aproximar integrales de tipo

$$\int_{-1}^{1} g(t)\, dt.$$

Sean x_i, $i = 0, \ldots, n$, puntos del intervalo $[-1, 1]$ y k_i, $i = 0, \ldots, n$, enteros no negativos. Sea P_m el polinomio de interpolación de Hermite de g con respecto a los x_i y k_i, $i = 0, \ldots, n$. Sabemos que existe un real $\theta_x \in [-1, 1]$ tal que

$$P_m(x) - g(x) = \frac{g^{(m+1)}(\theta_x)}{(m+1)!}\, p_m(x),$$

donde

$$m = n + \sum_{i=0}^{n} k_i \qquad \text{y} \qquad \theta_x \in [-1, 1].$$

Así

$$\int_{-1}^{1} g(t)\, dt = \int_{-1}^{1} P_m(x)\, dx - \frac{1}{(m+1)!} \int_{-1}^{1} g^{(m+1)}(\theta_x) p_m(x)\, dx.$$

En resto de este capitulo, consideramos el caso donde $k_i = 0$, $i = 0, \ldots, n$, es decir la interpolación de Lagrange,

$$P_m(x) = \sum_{i=0}^{n} g(x_i)\, L_i(x),$$

donde L_i son los polinomios de Lagrange

$$L_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \left(\frac{x - x_j}{x_i - x_j} \right).$$

Entonces

$$\int_{-1}^{1} g(t)\, dt = \sum_{i=0}^{n} c_i\, g(x_i) - e,$$

donde

$$c_i = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq i}}^{n} \left(\frac{x - x_j}{x_i - x_j} \right) dx$$

y

$$e = \frac{1}{(m+1)!} \int_{-1}^{1} g^{(m+1)}(\theta_t)\, p_m(t)\, dt.$$

Es evidente que la aproximación se hace despreciando el término e. Así

$$\int_{-1}^{1} g(t)\, dt \simeq \sum_{i=0}^{n} c_i\, g(x_i).$$

Casos particulares

1. Aproximación por rectángulos.
   En este caso la interpolación es de grado 0. Es decir $n = 0$, $L_0(x) = 1$ y

$$c_0 = \int_{-1}^{1} L_0(x)\, dx = 2.$$

$$\int_{t_i}^{t_{i+1}} f(x)\, dx = \frac{h_i}{2} \int_{-1}^{1} g(t)\, dt \simeq \frac{h_i}{2} 2g(x_i) = h_i f(z_i),$$

   donde $z_i \in [t_i, t_{i+1}]$.

$$\int_a^b f(x)\, dx = \sum_{i=0}^{k-1} \int_{t_i}^{t_{i+1}} f(x)\, dx,$$

$$\int_a^b f(x)\, dx \simeq \sum_{i=0}^{k-1} h_i\, f(z_i).$$

   Si $z_i = t_i$, el método es de los rectángulos del lado izquierdo (figura 2.1).

   Si $z_i = t_{i+1}$, el método corresponde a los rectángulos del lado derecho (figura 2.2).

   Si $z_i = \frac{t_i + t_{i+1}}{2}$, el método se llama de los puntos medios (figura 2.3).

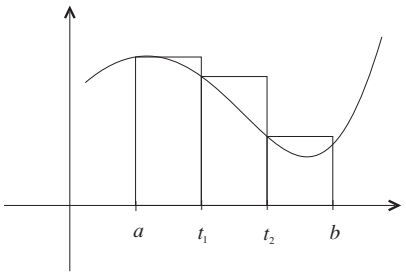

Figura 2.1.

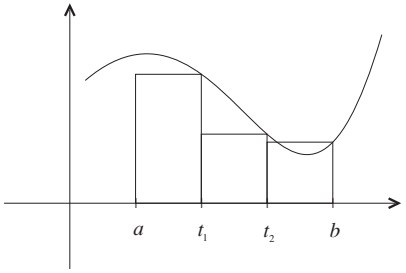

Figura 2.2.

2. Aproximación por trapezoides (figura 2.4).

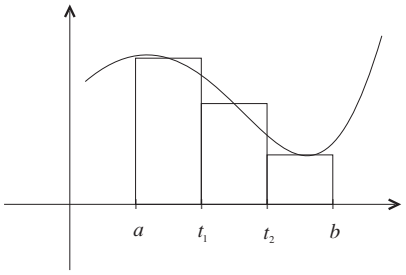

Figura 2.3.

Se usa en este caso la interpolación lineal, es decir de grado 1.

$$\int_{-1}^{1} g(t)\,dt \simeq c_0\, g(-1) + c_1\, g(1),$$
$$c_0 = \int_{-1}^{1} \frac{x - x_1}{x_0 - x_1}\,dx = \int_{-1}^{1} \frac{x-1}{-2}\,dx = 1,$$
$$c_1 = \int_{-1}^{1} \frac{x+1}{2}\,dx = 1,$$
$$g(-1) = f(t_i), \qquad g(1) = f(t_{i+1}).$$

Así

$$\int_{t_i}^{t_{i+1}} f(x)\,dx \simeq \frac{h_i}{2}\left(f(t_i) + f(t_{i+1})\right)$$

y

$$\int_a^b f(x)\,dx \simeq \sum_{i=0}^{k-1} \frac{h_i}{2}\left(f(t_i) + f(t_{i+1})\right). \tag{2.2}$$

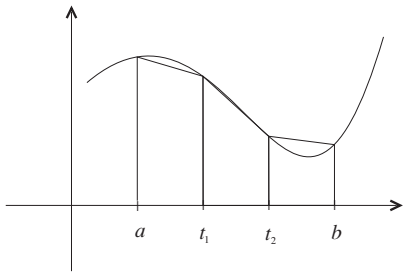

Figura 2.4.

3. Interpolación cuadrática.
   Para este caso la interpolación es de grado 2, es decir $k = 2$ y los puntos de interpolación son los extremos del intervalo y su punto medio. Así

$$\int_{-1}^{1} g(t)\,dt = c_0\, g(-1) + c_1 g(0) + c_2\, g(1),$$
$$c_0 = \int_{-1}^{1} \frac{x(x-1)}{(-1)(-2)}\,dx = \frac{1}{2}\left[\frac{1}{3}x^3 - \frac{1}{2}x^2\right]_{-1}^{1} = \frac{1}{3},$$
$$c_1 = \int_{-1}^{1} \frac{x^2-1}{-1}\,dx = -\left[\frac{x^3}{3} - x\right]_{-1}^{1} = \frac{4}{3},$$
$$c_2 = \int_{-1}^{1} \frac{x(x+1)}{2}\,dx = \frac{1}{2}\left[\frac{x^3}{3}\right]_{-1}^{1} = \frac{1}{3}.$$

Así tenemos la fórmula llamada de Simpson

$$\int_a^b f(t)\,dt = \sum_{i=0}^{k-1} \frac{h_i}{2}\left[\frac{1}{3}f(t_i) + \frac{4}{3}f(\frac{t_i+t_{i+1}}{2}) + \frac{1}{3}f(t_{i+1})\right]. \qquad (2.3)$$

2.2. Métodos de cuadratura gaussiana

Sea la fórmula de integración numérica dada por

$$\int_a^b f(x)\,dx \simeq \sum_{i=0}^{n} c_i\, f(x_i). \qquad (2.4)$$

Definición 2.2. Se dice de la fórmula (2.4) que es de grado m ($m \in \mathbb{N}$) si es exacta para todos los polinomios de grado $\leq m$ y existe al menos un polinomio de grado $m+1$ para el que no lo es.

Nota 2.2.

Dadosa los puntos x_i, $i = 0, \ldots, n$, en $[a, b]$ puntos distintos, sabemos que la fórmula (2.4) donde

$$c_i = \int_a^b L_i(x)\,dx, \qquad i = 0, \ldots, n$$

y L_i son los polinomios de Lagrange asociados a los puntos x_i, $i = 0, \ldots, n$, es de grado n. Dado que el error es

$$e = \int_a^b \frac{f^{(n+1)}(\theta_x)}{(n+1)!}\, p_n(x)\,dx$$

y si f es un polinomio de grado n, su derivada de orden $(n+1)$ es nula.

Problemática 2.2. Queremos encontrar una fórmula de integración numérica que sea de orden lo más grande posible.

Hasta ahora los puntos x_i, $i = 0, \ldots, n$, han sido arbitrarios en $[a, b]$ y proporcionan una fórmula de grado n. Una manera de proceder entonces es escoger estos puntos de manera que la fórmula (2.4) sea de grado mayor.

Definición 2.3. Sean $I = [a, b]$ y w una función definida e integrable en I. Supongamos además que $w(x) \geq 0$ para todo $x \in I$ y que $\int_a^b w(x) > 0$. Para f, g funciones definidas sobre I denotamos

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

si esta integral está bien definida. Llamamos a $\langle f, g\rangle$ producto escalar de f y g, y w se llama función peso.

Propiedades

Para todas f, g, h funciones tales que sus productos escalares esten definidos. Tenemos

1. $\langle \alpha f + \beta g, h\rangle = \alpha \langle f, h\rangle + \beta \langle g, h\rangle , \quad \forall \alpha, \beta \in \mathbb{R}.$
2. $\langle f, g\rangle = \langle g, f\rangle.$
3. $\langle f, f\rangle \geq 0.$
4. $\langle f, f\rangle = 0$ si y solo si $f \equiv 0.$

Notación

Se denota $\|f\| = \sqrt{\langle f, f\rangle}$.

Definición 2.4. Dos funciones f y g son ortogonales con respecto al producto escalar $\langle \cdot, \cdot \rangle$ si $\langle f, g\rangle = 0$.

Teorema 2.1. *Sean $I = [a, b]$ y w función peso. Existe una sucesión de polinomios $(P_n)_{n\in\mathbb{N}}$ tal que para todo $n \in \mathbb{N}$*

i) P_n *es mónico de grado* n *para* $n \geq 1$

ii) $\langle P_n, Q_{n-1}\rangle = 0$ *para todo polinomio* Q_{n-1} *de grado* $\leq n - 1$.

iv) $P_n = (x - \lambda_n)P_{n-1} - \mu_n P_{n-2}$, *donde*

$$\mu_n = \frac{\|P_{n-1}\|^2}{\|P_{n-2}\|^2} \qquad y \qquad \lambda_n = \frac{\langle xP_{n-1}, P_{n-1}\rangle}{\|P_{n-1}\|^2}.$$

Demostración. Consideramos $B = \{1, x, x^2, \ldots, x^n, \ldots\}$ la base canónica de los polinomios con coeficientes reales. Denotemos

$$\begin{aligned} P_0 &= 1, \\ P_{m+1} &= x^{m+1} - \sum_{k=1}^{m} \frac{\langle x^{m+1}, P_k\rangle}{\|P_k\|^2} P_k. \end{aligned}$$

Los P_m son los polinomios obtenidos por el proceso de ortogonalización de Gram–Schmidt aplicado a la base B. Es fácil ver (por inducción) que

$$\langle P_m, P_k\rangle = 0, \qquad \forall k < m.$$

Como $(P_n)_{n\in\mathbb{N}}$ forman una base del espacio de los polinomios, deducimos que $\langle P_m, Q_k\rangle = 0$ para todo polinomio de grado $k < m$. □

Ahora vamos a construir un método de cuadratura gaussiana usando los polinomios ortogonales. Para eso en lugar de aproximar $\int_a^b f(x)\,dx$ consideramos las integrales de tipo

$$\int_a^b w(x)\,f(x)\,dx,$$

donde $w(x)$ es una función peso. Se considera la fórmula de tipo

$$\int_a^b w(x)\,f(x)\,dx = \sum_{i=0}^{n} c_i\,f(\alpha_i). \tag{2.5}$$

El objetivo, recordémoslo, es encontrar los α_i y los c_i de manera que la fórmula (2.5) sea de orden máximo.

Sean P_n los polinomios ortogonales asociados al intervalo $I = [a, b]$ y la función peso $w(x)$. Sean x_i, $i = 0, \ldots, n-1$, las raíces del polinomio P_n.

Teorema 2.2. *Para todo $n \geq 1$, las raíces x_i de P_n son distintas, reales y pertenecen al intervalo $I = [a, b]$.*

Demostración. Sea $E = \{y_i,\ i = 1, \ldots, k\}$ el conjunto de raíces reales del polinomio P_n que están en I y que tienen una multipicidad impar. A priori, es posible que E sea vacío. Ponemos

$$h(x) = \begin{cases} \prod_{i=1}^{k} (x - y_i), & \text{si } k > 0, \\ 1, & \text{si } k = 0. \end{cases}$$

Ahora las raíces del polinomio $Q(x) = P_n(x) \cdot h(x)$ que están en el intervalo I son todas de multiplicidad par además $Q(x)$ es un polinomio mónico no nulo, entonces

$$\int_a^b w(x)\,Q(x)\,dx > 0.$$

Pero si $k < n$ por ortogonalidad tendremos

$$\int_a^b w(x)\,Q(x)\,dx = \int_a^b w(x)\,P_n(x)\,h(x)\,dx = 0.$$

Así que $k = n$ y el teorema queda probado. □

Ahora veremos que elegir las raíces de los polinomios P_n, $n \geq 1$, como los puntos de integración numérica mejora de manera muy sensible el orden de la cuadratura gaussiana como lo muestra el teorema siguiente.

Teorema 2.3. *Si se escogen los α_i, $i = 0, \ldots, n-1$, como raíces de P_n y*

$$c_i = \int_a^b w(x)\,L_i(x)\,dx,$$

entonces el método

$$\int_a^b w(x)\, f(x)\, dx = \sum_{i=0}^{n-1} c_i\, f(\alpha_i) \tag{2.6}$$

es de orden $(2n-1)$.

Definición 2.5. La formula dada por (2.6) se llama cuagratura gaussiana.

Demostración. Sea T un polinomio de grado menor o igual que $(2n-1)$. Por la división Euclidiana sabemos que existen dos polinomios q, r tales que

$$T = P_n\, q + r \qquad \text{y} \qquad d^\circ r < n \text{ ó } r = 0.$$

Ahora

$$T(\alpha_i) = P_n(\alpha_i)\, q(\alpha_i) + r(\alpha_i) = r(\alpha_i), \qquad i = 0, \ldots, n-1$$

porque los α_i son las raíces de P_n. Así

$$\int_a^b w(x)\, T(x)\, dx = \int_a^b w(x)\, P(x)\, q(x)\, dx + \int_a^b w(x)\, r(x)\, dx.$$

Pero como $d^\circ q < n$, por ortogonalidad

$$\int_a^b w(x)\, T(x)\, dx = \int_a^b w(x)\, r(x)\, dx.$$

Siendo $d^\circ r < n$ la fórmula (2.5) es exacta, es decir que

$$\int_a^b w(x)\, r(x)\, dx = \sum_{i=0}^{n-1} c_i\, r(\alpha_i).$$

Pero $r(\alpha_i) = T(\alpha_i)$ entonces

$$\int_a^b w(x)\, T(x)\, dx = \sum_{i=0}^{n-1} c_i\, T(\alpha_i),$$

es decir la fórmula es exacta para T. □

Ejemplo 2.2. En este ejemplo se comparan los resultados numéricos del cálculo de la integral

$$\int_{-1}^{1} e^{(4x^2)} \cos(4\pi x) dx.$$

El aspecto osilatorio de la función f y su fuerte gradiente en los extremos del intervalo de integración generan dificultades obvias de integración numérica como lo muestra la figura (2.5). El valor de la integral es 3.68584253940776 En la tabla siguiente se presenta el error cometido en los métodos ultilizatos que son él de trapezoides (2.2) , de Simpson (2.3), de cuadratura con interpolación en $n+1$ puntos equidistantes (columna Equidistantes.) y de cuadratura gaussiana (columna Legendre.) utilizando los poliniomios de Legendre que corresponden a la función peso $w = 1$ y el intérvalo $[-1,\ 1]$.

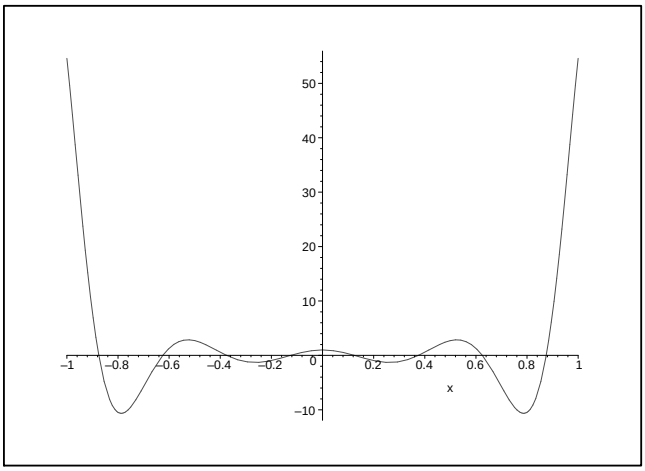

Figura 2.5.

n	Trapezoides	Simpson	Equidistantesqui.	Legendre.
4	26.831514	0.694563	17.432917	4.417462
12	2.373763	0.087200	0.232951	$0{,}34 \times 10^{-7}$
20	0.774535	0.011623	0.000852	$0{,}15 \times 10^{-7}$
28	0.383574	0.003038	0.000016	0.

Capítulo 3

Preliminares de álgebra lineal

3.1. Nomenclatura

Denotamos $\mathbb{K}$ al cuerpo de los reales $\mathbb{R}$ o los complejos $\mathbb{C}$, $\mathbb{K}^* = \mathbb{K} - \{0\}$, $\mathbb{N}$ es el conjunto de los enetros naturales, $\mathbb{N}^* = \mathbb{N} - \{0\}$. Sean m, n dos enteros positivos. $\mathbb{K}^{m\times n}$ es el espacio vectorial de las matrices de m filas y n columnas. Si $A \in \mathbb{K}^{m\times n}$, denotamos A_{ij} o a_{ij} al elemento de la fila i y la columna j de la matriz A.

La suma de dos matrices y la multiplicación de una matriz por un escalar, que dan a $\mathbb{K}^{m\times n}$ estructura de espacio vectorial, se definen por:

$$\begin{aligned} (A+B)_{ij} &= A_{ij} + B_{ij} && 1 \le i \le m, \quad 1 \le j \le n, \\ (\lambda A)_{ij} &= \lambda\, A_{ij} && \text{para } \lambda \in \mathbb{K}. \end{aligned}$$

El espacio $\mathbb{K}^{m\times n}$ es de dimensión $m \times n$. La base canónica de $\mathbb{K}^{m\times n}$ es el conjunto

$$B = \{E^{ij} \in \mathbb{K}^{m\times n};\ 1 \le i \le m,\ 1 \le j \le n\},$$

definido de la siguiente manera

$$E^{ij}_{kl} = \begin{cases} 1, & \text{si } k = i \text{ y } l = j, \\ 0, & \text{si no.} \end{cases}$$

Para $n \in \mathbb{N}$ denotemos también

$$\mathbb{K}^n = \mathbb{K}^{n\times 1},$$

es decir que los elementos de $\mathbb{K}^n$ son los vectores columna.

3.2. Multiplicación de matrices

Sean $A \in \mathbb{K}^{m\times n}$, $B \in \mathbb{K}^{n\times p}$. El producto de las matrices A y B, en este orden, es la matriz AB definida en $\mathbb{K}^{m\times p}$

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj}, \qquad 1 \le i \le m,\ 1 \le j \le p.$$

La matriz nula es la matriz cuyas entradas son nulas y se nota O_M o O si no hay lugar a confusión

Una matriz A es cuadrada si su número de filas es igual a su número de columnas, es decir si $A \in \mathbb{K}^{n\times n}$ para un n en $\mathbb{N}^*$.

La matriz identidad de $\mathbb{K}^{n\times n}$ es la matriz I definida por

$$I_{ij} = \delta_{ij},$$

donde δ_{ij} es el símbolo de Kronecker.

Para una matriz cuadrada, y $k \in \mathbb{N}$ se define la potencia $k$isima de A por $A^{k+1} = AA^k$ poniendo $A^0 = I$

Ejercicio 3.1. Sean A, B, C matrices. Demostrar que

1. $AI = IA = A$.
2. $A(B + C) = AB + AC$.
3. $(B + C)A = BA + CA$.
4. $A(BC) = (AB)C$.
5. $O_M \cdot A = O_M$.
6. $0 \cdot A = O_M$.

Las dimensiones de las matrices son tales que las expresiones que aparecen en el ejercicio esten bien definidas.

Definición 3.1. Sea una matriz A en $\mathbb{K}^{n\times n}$. Se dice que A es invertible si existe una matriz B en $\mathbb{K}^{n\times n}$ tal que

$$AB = BA = I.$$

La matriz B se llama la inversa de A y se nota A^{-1}.

Ejercicio 3.2. Demostrar que

1. La matriz nula no es invertible.
2. La matriz identidad es invertible.
3. Si A, B son invertibles en $\mathbb{K}^{n\times n}$, entonces AB lo es también.
4. Si $\lambda \in \mathbb{K}^*$ y $A \in \mathbb{K}^{n\times n}$ es invertible, entonces λA también es invertible.
5. Si $A \in \mathbb{K}^{n\times n}$ es invertible y $k \in \mathbb{N}$, entonces la potencia A^k de A lo es también.

3.3. Notación por bloques

Sean m, n enteros no negativos y sean

$$\begin{cases} m = m_1 + m_2 + \cdots + m_l & \text{y} \\ n = n_1 + n_2 + \cdots + n_p. \end{cases} \tag{3.1}$$

una descomposición de m y n y $A \in \mathbb{K}^{m\times n}$. Denotamos para $1 \leq r \leq l$, $1 \leq s \leq p$, A^{rs} la matriz de $\mathbb{K}^{m_r \times n_s}$ definida por la fórmula

$$A^{rs}_{ij} = A_{m_1+m_2+\cdots+m_{r-1}+i, n_1+n_2+\cdots+n_{s-1}+j}$$

para $1 \leq i \leq m_r$, $1 \leq j \leq n_s$. Es decir

$$A = \begin{pmatrix} A^{11} & A^{12} & \cdots & A^{1p} \\ A^{21} & A^{22} & \cdots & A^{2p} \\ \vdots & \vdots & & \vdots \\ A^{l1} & A^{l2} & \cdots & A^{lp} \end{pmatrix}. \tag{3.2}$$

En (3.2) A se presenta según la descomposición (3.1) de m y n.

Nota 3.1.
Si las matrices tienen la misma descomposición en bloques, es evidente que la suma preserva la estructura por bloques.

Ejercicio 3.3. Sea

$$\begin{cases} m = m_1 + \cdots + m_l, \\ n = n_1 + \cdots + n_p, \\ k = k_1 + \cdots + k_q. \end{cases} \tag{3.3}$$

Sean $A \in \mathbb{K}^{m\times n}$, $B \in \mathbb{K}^{n\times k}$ dadas en bloques según la descomposición (3.3). Demostrar que $AB \in \mathbb{K}^{m\times k}$ tiene estructura por bloques adecuada y que

$$(AB)^{ij} = \sum_{t=1}^{p} A^{it}\, B^{tj}, \tag{3.4}$$

para $1 \leq i \leq l$, $1 \leq j \leq q$. Los productos del lado izquierdo de la ecuación (3.4) son matriciales.

3.4. Matrices particulares

Sea $A \in \mathbb{K}^{n\times n}$. Se dice que A es

a) diagonal si $A_{ij} = 0$, para todo $i \neq j$.

b) triangular superior si $A_{ij} = 0$, para $i > j$.

c) triangular inferior si $A_{ij} = 0$, para $i < j$.

Proposición 3.1.

1. *El producto de dos matrices diagonales es una matriz diagonal.*
2. *El producto de dos matrices triangulares superiores (respectivamente inferiores) es una matriz triangular superior (respectivamente inferior).*

Demostración. Mostremos la parte 2. El caso diagonal va a ser un caso particular del caso triangular.
Sean T y S dos matrices triangulares superiores en $\mathbb{K}^{n\times n}$. Sea $i > j$

$$(TS)_{ij} = \sum_{k=1}^{n} T_{ik}\, S_{kj},$$

$$(TS)_{ij} = \underbrace{\sum_{k=1}^{j} T_{ik}\, S_{kj}}_{(I)} \; + \underbrace{\sum_{k=j+1}^{n} T_{ik}\, S_{kj}}_{(II)}.$$

La parte (I) es nula porque

$$k \leq j < i, \qquad \text{así} \quad T_{ik} = 0.$$

La parte (II) es nula porque

$$k > j, \qquad \text{así} \quad S_{kj} = 0.$$

Así

$$(TS)_{ij} = 0, \qquad \text{si} \quad i > j.$$

□

Ejercicio 3.4.

1. Mostrar que si T y S son triangulares superiores entonces
$$(TS)_{ii} = T_{ii}\, S_{ii}, \qquad \text{para} \quad i = 1, \ldots, n.$$
2. Mostrar que si una matriz triangular es invertible entonces sus elementos diagonales son no nulos.
3. Sean $\{e_k,\ k = 1, \ldots, n\}$ la base canónica de $\mathbb{K}^n$, T una matriz triangular superior invertible y x un vector de $\mathbb{K}^n$ tal que
$$Tx = e_k.$$
Mostrar que $x_j = 0$ para $j > k$.
4. Deducir del punto (3) que la inversa de una matriz triangular superior (respectivamente inferior) invertible es una matriz triangular superior (respectivamente inferior).

3.5. Transpuesta y adjunta de una matriz

Definición 3.2. Sea $A \in \mathbb{K}^{m\times n}$. Se define la transpuesta de A por la matriz $A^t \in \mathbb{K}^{m\times n}$ dada por

$$A^t_{ij} = A_{ji}, \qquad 1 \leq i \leq n, 1 \leq j \leq m.$$

La adjunta A^* de la Matriz A se define por

$$A^*_{ij} = \overline{A_{ji}}, \qquad 1 \leq i \leq n, 1 \leq j \leq m.$$

donde $\overline{A_{ij}}$ indica el complejo conjugado de A_{ij}.

Nota 3.2.

No hay mucho interés de considerar la transpuesta de una matriz con coeficientes complejos, entonces cuando la matriz es real se habla de su transpuesta, y cuando es compleja se habla de su adjunta. Además, si $A \in \mathbb{R}^{n\times n}$ entonces $A^t = A^*$.

3.5.1. Propiedades de la transpuesta y la adjunta de una matriz

Sean A, B matrices en $\mathbb{K}^{m\times n}$, C en $\mathbb{K}^{n\times p}$ y $\lambda \in \mathbb{K}$.

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

Definición 3.3.

1. Una matriz A en $\mathbb{R}^{n\times n}$ es simétrica si $A^t = A$.
2. Una matriz A en $\mathbb{C}^{n\times n}$ es hermitiana si $A^* = A$.

Ejercicio 3.5.

1. Demostrar que la suma de dos matrices simétricas (respectivamente hermitianas) es simétrica (respectivamente hermitiana).
2. ¿Es el producto de dos matrices simétricas (respectivamente hermitianas) simétrico (respectivamente hermitiano)?
3. Demostrar que si A es invertible entonces
$$\left(A^{-1}\right)^* = \left(A^*\right)^{-1}.$$
4. Deducir que si A es hermitiana entonces su inversa A^{-1} lo es también.

Definición 3.4.

1. Una matriz U en $\mathbb{C}^{n\times n}$ se dice unitaria si $UU^* = I$.
2. Una matriz S en $\mathbb{R}^{n\times n}$ se dice ortogonal si $SS^t = I$.

Nota 3.3.
El producto de matrices unitarias (respectivamente ortogonales) es unitario (respectivamente ortogonal)

3.6. Matrices con diagonal estrictamente dominante

Definición 3.5. Sea $A \in \mathbb{K}^{n\times n}$. Se dice que A tiene diagonal estrictamente dominante si $\forall i = 1, \ldots, n$

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \tag{3.5}$$

Proposición 3.2. *Si $A \in \mathbb{K}^{n\times n}$ es una matriz con diagonal estrictamente dominante entonces*

1. *Para todo $i = 1, \ldots, n$, $a_{ii} \neq 0$.*
2. *A es invertible.*

Demostración.

1. Por la desigualdad (3.5) se tiene tenemos $a_{ii} \neq 0$, $i = 1, \ldots, n$.
2. Supongamos que A no es invertible entonces existe $x \in \mathbb{K}^n$ tal que $x \neq 0$ y $Ax = 0$. Sea k el índice donde

$$|x_k| = \max_{i=1,\ldots,n} |x_i| .$$

Ahora $Ax = 0$ implica $(Ax)_k = 0$, es decir

$$a_{kk}x_k = -\sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj}x_j,$$

$$|a_{kk}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}| \frac{|x_j|}{|x_k|} \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}| .$$

Contradicción con la desigualdad (3.5).

□

Ejercicio 3.6. Sea $A = D + L + U \in \mathbb{K}^{n\times n}$ donde D,L y U son respectivamente la parte diagonal, la parte triangular inferior estricta y la parte triangular superior estricta de la matriz A. Mostrar que si A tiene diagonal estrictamente dominante entonces

1. $D^{-1}A$ tiene diagonal estrictamente dominante.
2. $(D + aL + bU)$ tiene diagonal estrictamente dominante donde a, b son escalares de $\mathbb{K}$ tales que $|a| \leq 1$ y $|b| \leq 1$.

3.7. Matrices de permutaciones

Para $n \in \mathbb{N}^*$ denotemos $J_n = \{1, 2, \ldots, n\}$.

Definición 3.6. Una permutación de J_n es una biyección de J_n en sí mismo. S_n designa el conjunto de las permutaciones de J_n. Una transposición de J_n es una permutación que intercambia dos elementos de J_n entre sí y deja el resto de los elementos sin cambio.Usualmente una permutación σ se nota de la manera siguiente

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n) \end{pmatrix}$$

Ejemplo 3.1. $n = 4$.

$i : J_4 \to J_4$,
$i(1) = 1, \quad i(2) = 2, \quad i(3) = 3, \quad i(4) = 4,$
$i \in S_4$ y se llama identidad.

$\sigma : J_4 \to J_4$,
$\sigma(1) = 3, \quad \sigma(2) = 4, \quad \sigma(3) = 4, \quad \sigma(4) = 1,$
$\sigma \notin S_4$.

$\alpha : J_4 \to J_4$,
$\alpha(1) = 1, \quad \alpha(2) = 3, \quad \alpha(3) = 4, \quad \alpha(4) = 2,$
$\alpha \in S_4$.

$\beta : J_4 \to J_4$,
$\beta(1) = 1, \quad \beta(2) = 4, \quad \beta(3) = 3, \quad \beta(4) = 2,$
β es una transposición.

Definición 3.7. Sean $n \in \mathbb{N}^*$ y $\sigma \in S_n$. Se asocia a la permutación σ la matriz Σ^σ definida en $\mathbb{K}^{n\times n}$ por

$$\Sigma^\sigma_{ij} = \delta_{i\sigma(j)}, \qquad 1 \leq i, j \leq n,$$

donde δ_{ij} es el símbolo de Kronecker. Σ^σ se llama matriz de permutación asociada a σ.

Ejemplo 3.2. En el ejemplo (3.1), la matriz de permutación asociada a α es

$$\Sigma^\alpha = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

La matriz de permutación asociada a β es

$$\Sigma^\beta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Ejercicio 3.7. Calcular $(\Sigma^\alpha)(\Sigma^\alpha)^t$ y $\left(\Sigma^\beta\right)\left(\Sigma^\beta\right)^t$.

Proposición 3.3. *Sean n un entero positivo y α, β dos permutaciones de S_n. Tenemos*

1. $\Sigma^i = I$
2. $\Sigma^\alpha \Sigma^\beta = \Sigma^{\alpha \circ \beta}$
3. $(\Sigma^\alpha)\left(\Sigma^{\alpha^{-1}}\right) = I$
4. $(\Sigma^\alpha)(\Sigma^\alpha)^t = I$

Demostración.

1. Evidente.
2. Sean $1 \leq i, j \leq n$.

$$\left(\Sigma^\alpha \Sigma^\beta\right)_{ij} = \sum_{k=1}^{n} \left(\Sigma^\alpha_{ik} \cdot \Sigma^\beta_{kj}\right),$$

$$\left(\Sigma^\alpha \Sigma^\beta\right)_{ij} = \sum_{k=1}^{n} \left(\delta_{i\,\alpha(k)} \cdot \delta_{k\,\beta(j)}\right),$$

   pero $\delta_{k\,\beta(j)} = 0$ excepto cuando $k = \beta(j)$. Así

$$\left(\Sigma^\alpha \Sigma^\beta\right)_{ij} = \delta_{i\,\alpha(\beta(j))} = \Sigma^{\alpha\circ\beta}_{ij}.$$

3. El resultado se obtiene usando (1) y (2).

4. Sean $1 \leq i, j \leq n$.

$$\left(\Sigma^\alpha \left(\Sigma^\alpha\right)^t\right)_{ij} = \sum_{k=1}^{n} \left(\Sigma^\alpha_{ik} \cdot \Sigma^\alpha_{jk}\right),$$

$$\left(\Sigma^\alpha \left(\Sigma^\alpha\right)^t\right)_{ij} = \sum_{k=1}^{n} \left(\delta_{i\,\alpha(k)} \cdot \delta_{j\,\alpha(k)}\right),$$

pero $\delta_{j\,\alpha(k)} = 0$ excepto cuando $\alpha(k) = j$, es decir $k = \alpha^{-1}(j)$. Así

$$\left(\Sigma^\alpha \left(\Sigma^\alpha\right)^t\right)_{ij} = \delta_{i\,\alpha(\alpha^{-1}(j))} = \delta_{ij}.$$

Lo que significa que la transpuesta de cualquier matriz de permutación es la matriz de permutación asociada a la inversa de esta permutación.

□

3.8. Determinantes

Adoptemos en esta sección la siguiente notación. Escribimos una matriz A en $\mathbb{K}^{n\times n}$ en la forma

$$A = (c_1|\ldots|c_n)$$

donde c_i, $i = 1, \ldots, n$, son las columnas de A.

Definición 3.8.

1. Se le dice a una función

$$L : \mathbb{K}^{n\times n} \to \mathbb{K} \qquad \text{multilineal o } n\text{-lineal}$$

si es lineal para cada columna, es decir si $\forall i$, $i = 1, \ldots, n$, la función

$$x \to L(c_1|\ldots|c_{i-1}|x|c_{i+1}|\ldots|c_n)$$

de $\mathbb{K}^n$ en $\mathbb{K}$ es lineal. Así

$$L(c_1|\ldots|c_{i-1}|\lambda x + \mu y|c_{i+1}|\ldots|c_n) = \\ \lambda L(c_1|\ldots|c_{i-1}|x|c_{i+1}|\ldots|c_n) + \mu L(c_1|\ldots|c_{i-1}|y|c_{i+1}|\ldots|c_n).$$

2. Una función n-lineal es alternada si $L(A) = 0$ para toda matriz A que tenga dos columnas iguales.

Lema 3.1. *Si L es una función n-lineal alternada entonces*

$$L(A) = -L(A') \quad \textit{donde}$$

A' es la matriz obtenida permutando dos columnas de A entre sí.

Demostración. Sean $A \in \mathbb{K}^{n\times n}$ e i, j dos índices fijos en $\{1, \ldots, n\}$ tales que $i < j$. Escribamos $A = (c_1|\ldots|c_n)$.

$$L(c_1|\ldots|c_i + c_j|\ldots|c_j + c_i|\ldots|c_n) = 0$$

por tener dos columnas iguales, pero

$$\begin{aligned} L(c_1|\ldots|c_i + c_j|\ldots|c_j + c_i|\ldots|c_n) = \\ L(c_1|\ldots|c_i|\ldots|c_j|\ldots|c_n) + L(c_1|\ldots|c_i|\ldots|c_i|\ldots|c_n) + \\ L(c_1|\ldots|c_j|\ldots|c_i|\ldots|c_n) + L(c_1|\ldots|c_j|\ldots|c_j|\ldots|c_n) = 0. \end{aligned}$$

Pero $(c_1|\ldots|c_j|\ldots|c_i|\ldots|c_n) = A'$ entonces $L(A) = -L(A')$. □

Definición 3.9. Se dice que una función $D : \mathbb{K}^{n\times n} \to \mathbb{K}$ es una función determinante si

a) D es n-lineal.

b) D es alternada.

c) $D(I) = 1$.

Ejemplo 3.3. Sea D una función determinante para $n = 2$ y sea $I = (e_1|\, e_2)$, donde $\{e_1, e_2\}$ es la base canónica de $\mathbb{K}^2$. Una Matriz $A \in \mathbb{K}^{2\times 2}$ de la forma

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

se puede escribir como

$$A = (ae_1 + ce_2|\, be_1 + de_2).$$

La multilinealidad de L implica

$$\begin{aligned} D(A) &= D(ae_1 + ce_2, be_1 + de_2), \\ &= ab\, D(e_1, e_1) + ad\, D(e_1, e_2) + cb\, D(e_2, e_1) + cd\, D(e_2, e_2), \\ &= ad\, D(e_1, e_2) + cb\, D(e_2, e_1), \\ &= (ad - bc)\, D(e_1, e_2), \\ &= (ad - bc)\, D(I), \\ &= ad - bc. \end{aligned}$$

Nota 3.4.

Deducimos del ejemplo (3.3) que no hay sino una única función determinante para $n = 2$.

Ahora sea $A \in \mathbb{K}^{n\times n}$,

$$A = (c_1|\dots|c_n).$$

Para cada columna c_j podemos escribir

$$c_j = \sum_{i=1}^{n} a_{ij}\, e_j.$$

Por multilinealidad, si D es una función determinante

$$\begin{aligned} D(e_1|\dots|c_i|\dots|c_j|\dots|e_n) &= \sum_{\substack{k=1\\h=1}}^{n} a_{ik}\, a_{jh}\, D(e_1|\dots|e_k|\dots|e_h|\dots|e_n), \\ &= a_{ii}\, a_{jj} - a_{ij}\, a_{ji}. \end{aligned}$$

El caso anterior era sólo para dos columnas de A y el resto son los vectores de la base canónica. Se puede mostrar fácilmente que

$$D(A) = \sum_{\sigma\in S_n} (-1)^{sg(\sigma)} \prod_{i=1}^{n} a_{i\,\sigma(i)} \tag{3.6}$$

donde S_n es el conjunto de las permutaciones de J_n y $sg(\sigma)$ es la signatura de σ, es decir el número de transposiciones que la compone módulo 2.

Del resultado (3.6) podemos deducir que en $\mathbb{K}^{n\times n}$ no hay sino una sola función determinante que denotaremos det en el futuro.

3.8.1. Propiedades del determinante

Si $A, B \in \mathbb{K}^{n\times n}$ y $\lambda \in \mathbb{K}$, entonces

1. $\det(\lambda A) = \lambda^n \det(A)$
2. $\det(AB) = \det(A)\, \det(B)$
3. $\det(I) = 1$
4. $\det(A)\, I = (A)\, [\mathrm{cof}(A)]^t$

La matriz de cofactores $\mathrm{cof}(A)$ está dada por

$$[\mathrm{cof}(A)]_{ij} = (-1)^{(i+j)}\, a_{ij}\, \det(A(i,j)),$$

donde $A(i,j)$ es la matriz $(n-1)\times(n-1)$ obtenida borrando la fila i y la columna j de A.

Proposición 3.4. *Una matriz cuadrada A es invertible si y sólo si su determinante es no nulo.*

Demostración. Si A es invertible, tenemos

$$A\,A^{-1} = I \qquad \text{lo que implica que}$$
$$\det(A)\ \det(A^{-1}) = \det(I) = 1.$$

Así

$$\det(A) \neq 0.$$

Recíprocamente, si $\det(A) \neq 0$, sabemos que

$$A\ [\mathrm{cof}(A)]^t = (\det A)\, I$$

lo que significa que A es invertible y

$$A^{-1} = \frac{1}{\det(A)}\ [\mathrm{cof}(A)]^t\,.$$

□

Ejercicio 3.8. Mostrar que si $A \in \mathbb{K}^{n\times n}$ entonces

1. $\det(A^*) = \overline{\det(A)}$.
2. $\det(A^n) = (\det(A))^n$.
3. $\det(A) = \prod_{i=1}^{n}(A_{ii})$ si A es triangular.

Ejercicio 3.9.

1. Sean A, B dos matrices cuadradas, siendo A invertible, y c un escalar. Demostrar que
$$\det(AB - cI) = \det(BA - cI).$$
2. Sea A una matriz singular, es decir no invertible. Demostrar que existe un $\varepsilon_0 > 0$, tal que $\forall\varepsilon$, $0 < \varepsilon \leq \varepsilon_0$
$$\det(A + \varepsilon I) \neq 0.$$
3. Deducir de los numerales (1) y (2) que para todas las matrices A, B en $\mathbb{K}^{n\times n}$
$$\det(AB - cI) = \det(BA - cI).$$

Ejercicio 3.10.

1. Sea A una matriz de $\mathbb{K}^{n\times n}$ con la estructura por bloques siguiente
$$A = \left(\begin{array}{c|c} K & H \\ \hline O & L \end{array}\right)$$
donde $K \in \mathbb{K}^{p\times p}$, $L \in \mathbb{K}^{q\times q}$, O es la matriz nula y $n = p + q$. Demostrar que
$$\det(A) = \det(K)\ \det(L).$$

2. Deducir que si A es una matriz triangular por bloques entonces

$$\det(A) = \prod_{i=1}^{k} \det(A^{ii})$$

donde A^{ii}, $i = 1, \ldots, k$, son los bloques diagonales (los A^{ii} son cuadrados).

3.9. Elementos propios de una matriz

Sean $A \in \mathbb{K}^{n\times n}$ y $\lambda \in \mathbb{K}$.

Definición 3.10. Se dice que λ es un valor propio de A si existe un vector x en $\mathbb{K}^n$ no nulo tal que

$$Ax = \lambda x.$$

En este caso x se llama vector propio asociado a λ.

Nota 3.5.

1. Si x es vector propio de A asociado a λ entonces

$$Ax = \lambda x$$

o también

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

lo que significa que la matriz $(\lambda I - A)$ no es invertible es decir

$$\det(\lambda I - A) = 0. \tag{3.7}$$

De la misma manera, si λ verifica la ecuación (3.7) se deduce que λ es valor propio de A.

2. De las propiedades del determinante, podemos decir que $\det(\lambda I - A)$ es un polinomio de grado exactamente n en λ con coeficiente dominante 1.

3. De la nota (2) y del Teorema Fundamental del Álgebra concluimos que una matriz A en $\mathbb{K}^{n\times n}$ tiene exactamente n valores propios en $\mathbb{C}$ (eventualmente repetidos)

Definición 3.11. Sea $A \in \mathbb{K}^{n\times n}$. El espectro de A es

$$\text{esp}(A) = \{\lambda \in \mathbb{C};\ \lambda \text{ es valor propio de } A\}.$$

Ejercicio 3.11. Sean $A, B \in \mathbb{C}^{n\times n}$. ¿Verdadero o falso?

1. A es invertible si y sólo si $0 \in \text{esp}(A)$.

2. $\text{esp}(\alpha I + A) = \{\alpha + \lambda;\ \lambda \in \text{esp}(A)\}$.

3. Si $\lambda \in \text{esp}(A)$ y $\mu \in \text{esp}(B)$ entonces $\lambda + \mu \in \text{esp}(A + B)$.

4. Si $\lambda \in \text{esp}(A)$ y $\mu \in \text{esp}(B)$ entonces $\lambda\mu \in \text{esp}(AB)$.

5. Si $Q(x)$ es un polinomio con coeficientes en $\mathbb{C}$ entonces $\lambda \in \text{esp}(A)$ implica $Q(\lambda) \in \text{esp}(Q(A))$.

Ejercicio 3.12. ¿Es la recíproca del punto (5) del ejercicio (3.11) verdadera?

Proposición 3.5. *Sean T una matriz invertible en $\mathbb{C}^{n\times n}$ y A, B dos matrices tales que*

$$A = T^{-1}BT$$

entonces $\text{esp}(A) = \text{esp}(B)$.

Demostración. Tenemos

$$\lambda \in \text{esp}(A)$$

si y sólo si

$$\det(A - \lambda I) = 0. \tag{3.8}$$

Multiplicando la ecuación (3.8) por $\det T \det T^{-1}$ sale

$$\det T \cdot \det(A - \lambda I) \cdot \det(T)^{-1} = 0$$

es decir,

$$\det(TAT^{-1} - \lambda I) = 0,$$

lo que significa

$$\lambda \in \text{esp}(B).$$

□

Definición 3.12. El radio espectral $\rho(A)$ de una matriz cuadrada A es

$$\rho(A) = \text{máx}\left\{|\lambda|\,,\ \lambda \in \text{esp}(A)\right\}$$

Ejercicio 3.13. ¿Verdadero o falso? Si $\rho(A) = 0$ entonces $A = 0$.

Ejercicio 3.14. Sea $a \in \mathbb{C}$. Estudiar $\rho(A)$ donde

$$A = \begin{pmatrix} a & 2a^2 \\ 1 & a \end{pmatrix}.$$

Definición 3.13. Una matriz A en $\mathbb{C}^{n\times n}$ es diagonalizable si existen una matriz diagonal Δ y una matriz invertible P tales que

$$A = P\Delta P^{-1}. \tag{3.9}$$

Nota 3.6.

1. Si la ecuación (3.9) tiene lugar entonces

$$\mathrm{esp}(A) = \{\Delta_{ii},\ i = 0, \ldots, n\}.$$

2. Si la ecuación (3.9) es verdadera entonces

$$Q(A) = P^{-1}\, Q(\Delta)\, P, \qquad \text{para todo polinomio } Q.$$

   Lo que implica que $Q(A)$ es también diagonalizable.

3. Supongamos que una matriz A en $\mathbb{K}^{n\times n}$ es diagonalizable. Sean Δ la matriz diagonal y P una matriz invertible tales que

$$A = P\Delta P^{-1},$$

   o también

$$AP = P\Delta. \tag{3.10}$$

   Si escribimos la matriz P en la forma

$$P = (c_1|c_2|\ldots|c_n)$$

   donde c_i, $i = 1, \ldots, n$, son las columnas de P, la ecuación (3.10) será equivalente a

$$Ac_i = \lambda_i c_i, \qquad i = 1, \ldots, n, \tag{3.11}$$

   donde λ_i, $i = 1, \ldots, n$, son los coeficientes Δ_{ii}.
   La ecuación (3.11) significa que las columnas de P son justamente los vectores propios de A.

Ejercicio 3.15. Sea

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

Mostrar que A no es diagonalizable.

Ejercicio 3.16. ¿Verdadero falso?

1. Dada una matriz diagonal D. Todo vector en $\mathbb{K}^n - \{0\}$ es un vector propio de la matriz D.

2. La suma de dos matices diagonalizables es diagonalizable.

Proposición 3.6. *Los vectores propios asociados a valores propios distintos de una matriz son linealmente independientes.*

Demostración. Por inducción sobre k el número de valores propios. Denotemos v_i, $i = 1, \ldots, k$, los vectores propios asociados a los valores propios λ_i, $i = 1, \ldots, k$. Si $k = 1$, dado que los vectores propios son no nulos

$$\alpha_1 \, v_1 = 0 \quad \text{implica} \quad \alpha_1 = 0.$$

Supongamos que todo subconjunto de k vectores propios asociados a valores propios distintos son linealmente independientes. Sean ahora v_i, $i = 1, \ldots, k+1$, $k+1$ vectores propios asociados a valores propios distintos.
Si

$$\sum_{i=1}^{k+1} \alpha_i \, v_i = 0, \tag{3.12}$$

aplicando A a la ecuación (3.12) obtenemos

$$\sum_{i=1}^{k+1} \alpha_i \, A v_i = 0.$$

Como v_i, $i = 1, \ldots, k+1$, son vectores propios

$$\sum_{i=1}^{k+1} \alpha_i \, \lambda_i \, v_i = 0. \tag{3.13}$$

Multiplicando la ecuación (3.12) por λ_{k+1} y restándola de la ecuación (3.13)

$$\sum_{i=1}^{k} \alpha_i (\lambda_i - \lambda_{k+1}) v_i = 0.$$

La hipótesis de inducción implica

$$\alpha_i (\lambda_i - \lambda_{k+1}) = 0, \qquad \text{para todo} \quad i = 1, \ldots, k.$$

Pero los λ_i, $i = 1, \ldots, k+1$, son distintos es decir $\alpha_i = 0$, $i = 1, \ldots, k$. Reemplazando en la ecuación (3.12) resulta que

$$\alpha_{k+1} = 0,$$

lo que significa que $v_1, \ldots, v_{k+1}$ son linealmente independientes. □

Corolario 3.1. *Si $A \in \mathbb{K}^{n \times n}$ tiene n valores propios distintos entonces A es diagonalizable.*

Demostración. Sean $\lambda_1, \ldots, \lambda_n$ los valores propios de A y sean $v_1, \ldots, v_n$ vectores propios asociados a los λ_i, $i = 1, \ldots, n$. Tenemos

$$A \, v_i = \lambda_i \, v_i, \qquad i = 1, \ldots, n. \tag{3.14}$$

Sea $P = (v_1|v_2|\dots|v_n)$. La ecuación (3.14) significa

$$AP = P\Lambda,$$

es decir

$$A = P\Lambda P^{-1},$$

donde

$$\Lambda = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}.$$

□

3.10. Producto escalar y normas en $\mathbb{C}^n$

Sean a, b dos vectores en $\mathbb{C}^n$. Se define el producto escalar $\langle a, b\rangle$ por

$$\boxed{\langle a, b\rangle = b^* a}$$

Nota 3.7.
Recordemos que:

a) Si $b \in \mathbb{C}^n$, $b^* \in \mathbb{C}^{1\times n}$ y el producto $b^* \cdot a$ es producto matricial cuyo resultado está en $\mathbb{C}^{1\times 1}$ es decir es un escalar.

b) $b_i^* = \overline{b_i}$, $i = 1, \dots, n$.

3.10.1. Propiedades del producto escalar

Para todo a, b, c en $\mathbb{C}^n$, y α, β en $\mathbb{C}$. Tenemos

1. $\langle \alpha a + \beta b, c\rangle = \alpha \langle a, c\rangle + \beta \langle b, c\rangle$.
2. $\langle a, b\rangle = \overline{\langle b, a\rangle}$.
3. $\langle a, a\rangle$ es real no negativo.
4. Si $a \neq 0$, $\langle a, a\rangle > 0$.

Nota 3.8.
Notemos para un $a \in \mathbb{C}^n$, $\|a\|_2 = \sqrt{\langle a, a\rangle}$.

Definición 3.14. 1. Dos vectores a, b en $\mathbb{C}^n$ se dicen ortogonales si

$$\boxed{\langle a, b\rangle = 0}$$

2. Un conjunto E de vectores de $\mathbb{C}^n$ es ortogonal si

$$\forall (a,b) \in E^2, \qquad \langle a,b\rangle = 0.$$

Nota 3.9.

1. Si $\langle a,b\rangle = 0$ se denota $a \perp b$.
2. Si $A \subset \mathbb{C}^n$, el ortogonal de A es el conjunto definido por

$$A^{\perp} = \{x \in \mathbb{C}^n;\ \langle x,a\rangle = 0\ \forall a \in A\}.$$

Propiedades

a) $\forall A \subset \mathbb{C}$, $0 \in A^{\perp}$.

b) $\{0\}^{\perp} = \mathbb{C}^n$.

c) $\forall A \subset \mathbb{C}$, $A^{\perp}$ es un subespacio vectorial $\mathbb{C}^n$.

Proposición 3.7. *Si E es un subconjunto ortogonal de $\mathbb{C}^n$ y $0 \notin E$ entonces E es un conjunto de vectores linealmente independientes.*

Demostración. Sea E un conjunto ortogonal y sean $a_1,\ldots,a_k$ k elementos de E. Supongamos que existen escalares $\alpha_1,\ldots,\alpha_k$ tales que

$$\sum_{i=1}^{k} \alpha_i\, a_i = 0.$$

Para un j fijo entre 1 y k

$$\left\langle a_j, \sum_{i=1}^{k} \alpha_i\, a_i \right\rangle = 0.$$

Por ortogonalidad

$$\alpha_j \langle a_j, a_j\rangle = 0$$

lo que implica que

$$\alpha_j = 0, \qquad \text{dado que} \quad \langle a_j, a_j\rangle > 0.$$

□

Corolario 3.2. *Una familia ortogonal de $\mathbb{C}^n$ no puede tener más que n elementos.*

Teorema 3.1. *sean $\nu_1,\ldots,\nu_k$ k vectores de $\mathbb{K}^n$ linealmente independientes. Existen $u_1,\ldots,u_k$ k vectores ortogonales tales que las dos familias $\{\nu_1,\ldots,\nu_k\}$ y $\{u_1,\ldots,u_k\}$ generan el mismo espacio.*

Demostración. El resultado es debido al conocido proceso de ortogonalización de Gram–Schmidt.
La familia u_i, $i = 1, \ldots, k$, se obtiene de la siguiente forma:

$$\begin{aligned} u_1 &= \nu_1, \\ u_j &= \nu_j - \sum_{i=1}^{j-1} \frac{\langle u_i, \nu_i \rangle}{\langle u_i, u_i \rangle} u_i. \end{aligned}$$

□

Ejercicio 3.17. Verificar que la familia $\{u_i,\ i = 1, \ldots, k\}$ definida en la demostración anterior es ortogonal y genera el mismo subespacio vectorial que la familia $\{\nu_i,\ i = 1, \ldots, k\}$.

3.10.2. Producto escalar y matrices

Teorema 3.2. *Sean $A \in \mathbb{K}^{m\times n}$, $x \in \mathbb{C}^n$ y $y \in \mathbb{C}^m$. Tenemos*

$$\langle Ax, y \rangle = \langle x, A^* y \rangle .$$

Demostración.

$$\begin{aligned} \langle Ax, y \rangle &= y^* Ax, \\ &= (y^* A)x, \\ &= (A^* y)^* x, \\ &= \langle x, A^* y \rangle . \end{aligned}$$

□

Corolario 3.3. *Si A es una matriz hermitiana, es decir que $A = A^*$, en $\mathbb{C}^{n\times n}$ entonces $\langle Ax, x \rangle$ es real para todo $x \in \mathbb{C}^n$.*

Demostración. Sea $x \in \mathbb{C}^n$.

$$\begin{aligned} \langle Ax, x \rangle &= \langle x, A^* x \rangle , \\ &= \langle x, Ax \rangle , \\ &= \overline{\langle Ax, x \rangle}, \end{aligned}$$

así que $\langle Ax, x \rangle$ es igual a su conjugado es decir es real. □

Ejercicio 3.18.

1. Mostrar que si $A \in \mathbb{C}^{n\times n}$ y $\langle Ax, x \rangle$ es real para todo $x \in \mathbb{C}^n$, entonces A es hermitiana.

2. Hallar una matriz real no simétrica A tal que $\langle Ax, x\rangle \geq 0$, $\forall x \in \mathbb{R}^n$.

Definición 3.15. Se dice de una matriz hermitiana $A \in \mathbb{C}^{n\times n}$ que es definida positiva si

$$\langle Ax, x\rangle > 0, \qquad \forall x \in \mathbb{C}^n - \{0\}.$$

Se dice de una matriz hermitiana $A \in \mathbb{C}^{n\times n}$ que es semi–definida positiva si

$$\langle Ax, x\rangle \geq 0, \qquad \forall x \in \mathbb{C}^n.$$

Proposición 3.8. *Para toda matriz cuadrada A en $\mathbb{C}^{n\times n}$, las matrices A^*A y AA^* son semi–definidas positivas.*
Además, son definidas positivas si y sólo si A es invertible.

Demostración. Sea $A \in \mathbb{C}^{n\times n}$.

$$(AA^*)^* = (A^*)^*A^* = AA^*,$$

entonces AA^* es hermitiana.
Sea $x \in \mathbb{C}^n$.

$$\langle AA^*x, x\rangle = \langle A^*x, A^*x\rangle = \|A^*x\|^2 \geq 0.$$

Si A es invertible, A^* también lo es y si $x \neq 0$ entonces $\|A^*x\| > 0$. □

Proposición 3.9. *Si A es definida positiva entonces A es invertible.*

Demostración. Si A no fuera invertible existiría $x \neq 0$ tal que $Ax = 0$. Así $\langle Ax, x\rangle = 0$.
Contradicción con el carácter definido positivo de A. □

Definición 3.16. Sea $A \in \mathbb{K}^{n\times n}$. Las submatrices principales de A son las n matrices cuadradas $A(k)$ en $\mathbb{K}^{k\times k}$, $k = 1, \ldots, n$, definidas por

$$(A(k))_{ij} = A_{ij}, \quad 1 \leq i \leq k, 1 \leq j \leq k.$$

Teorema 3.3. *Una matriz $A \in \mathbb{K}^{n\times n}$ es hermitiana definida positiva si y sólo si todas sus submatrices principales lo son.*

Demostración. Evidentemente la condición es suficiente porque $A(n) = A$. La condición es necesaria porque en primer lugar las matrices $A(k)$, $i = 1, \ldots, n$, son hermitianas y en segundo lugar para un vector $x \in \mathbb{K}^k$ no nulo se define el vector y $\hat{x} \in \mathbb{K}^n$ tales que

$$\hat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

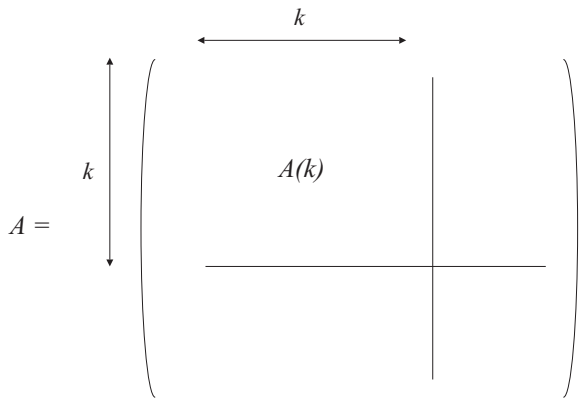

Figura 3.1.

$$\langle A\hat{x}, \hat{x}\rangle = \langle A(k)x, x\rangle > 0,$$

así $A(k)$ es hermitiana definida positiva. □

3.11. Valores propios de matrices particulares

Teorema 3.4. *Sea* $A \in \mathbb{C}^{n\times n}$.

1. *Si* A *es hermitiana entonces* $\text{esp}(A) \subset \mathbb{R}$.
2. *Si* A *es semi–definida positiva entonces* $\text{esp}(A) \subset \mathbb{R}_+$.
3. *Si* A *es definida positiva entonces* $\text{esp}(A) \subset \mathbb{R}_+^*$.
4. *Si* A *es unitaria entonces* $\text{esp}(A) \subset S^1$ *donde* $S^1 = \{\lambda \in \mathbb{C};\ |\lambda| = 1\}$.

Demostración. Sea $A \in \mathbb{C}^{n\times n}$.

1. Supongamos A hermitiana. Sea $\lambda \in \text{esp}(A)$ y sea u vector propio asociado a λ. Tenemos $Au = \lambda u$ y $\langle Au, u\rangle = \langle u, Au\rangle$ porque A es hermitiana. Reemplazando en la última igualdad

   $$\langle \lambda u, u\rangle = \langle u, \lambda u\rangle,$$

   es decir

   $$\lambda \langle u, u\rangle = \overline{\lambda} \langle u, u\rangle.$$

   Como $u \neq 0$,

   $$\lambda = \overline{\lambda}.$$

   Así $\lambda \in \mathbb{R}$.

2. Sean A semi–definida positiva y $\lambda \in \text{esp}(A)$. Sea $u \in \mathbb{C}^n$ tal que $Au = \lambda u$, $u \neq 0$. Tenemos
$$0 \leq \langle Au, u\rangle = \langle \lambda u, u\rangle = \lambda \langle u, u\rangle .$$
Como $\langle u, u\rangle > 0$ deducimos que $\lambda \geq 0$.

3. Si A es definida positiva entonces A es semi–definida positiva e invertible. Así $\text{esp}(A) \subset]0 \quad +\infty[$.

4. Sea A unitaria. Sean λ valor propio de A y $u \neq 0$ tales que $Au = \lambda u$. Tenemos
$$\langle Au, Au\rangle = \langle A^* Au, u\rangle = \langle u, u\rangle .$$
Pero
$$\langle Au, Au\rangle = \langle \lambda u, \lambda u\rangle = \lambda\overline{\lambda} \langle u, u\rangle = |\lambda|^2 \langle u, u\rangle .$$
Así
$$|\lambda|^2 \langle u, u\rangle = \langle u, u\rangle .$$
Como $u \neq 0$, $|\lambda|^2 = 1$, es decir $|\lambda| = 1$.

□

Teorema de Schur

Para toda matriz $A \in \mathbb{C}^{n\times n}$ existen una matriz U unitaria (es decir $U^*U = I$) y una matriz triangular superior T tales que

$$A = UTU^*. \tag{3.15}$$

Nota 3.10.

1. Veamos lo qué significa una matriz unitaria.
Sea $U = (c_1|\ldots|c_n)$ unitaria.
$$U^* = \begin{pmatrix} {c_1}^* \\ {c_2}^* \\ \vdots \\ {c_n}^* \end{pmatrix}.$$
$$(U^*U)_{ij} = \langle c_i, c_j\rangle = \delta_{ij}$$
lo que significa que las columnas de U forman una base ortonormal de $\mathbb{C}^n$.

2. Se dice que A es ortonormalmente triangularizable si satisface la ecuación (3.15).

Demostración. Por inducción sobre n.
Para $n = 1$

$$\forall a \in \mathbb{C}, \qquad a = (1)^* a \cdot (1).$$

Supongamos la propiedad válida para $n-1$.
Sean λ_1 un valor propio de $A \in \mathbb{C}^{n\times n}$ y x_1 un vector propio asociado a λ_1 tal que $\|x\|_2 = 1$.
Construyamos una base ortonormal $\{x_1, \ldots, x_n\}$ (Notar que la base contiene x_1).
Consideremos la matriz

$$V = (x_1|\ldots|x_n), \qquad V^* = \left(\begin{array}{c} {x_1}^* \\ \hline {x_2}^* \\ \hline \vdots \\ \hline {x_n}^* \end{array}\right).$$

V es unitaria por construcción. Tenemos

$$AV = (\lambda x_1, Ax_2|\ldots|Ax_n),$$
$$V^*AV = \left(\begin{array}{c|c} \lambda_1 & b_{n-1}^* \\ \hline 0 & A_{n-1} \end{array}\right),$$

donde

$$b_{n-1} \in \mathbb{C}^{n-1} \qquad \text{y} \qquad A_{n-1} \in \mathbb{C}^{n-1\times n-1}.$$

La hipótesis de inducción implica que existen

$$W_{n-1} \quad \text{unitaria en} \quad \mathbb{C}^{n-1\times n-1} \quad \text{y}$$
$$T_{n-1} \quad \text{triangular superior en} \quad \mathbb{C}^{n-1\times n-1}$$

tales que

$$A_{n-1} = W_{n-1}T_{n-1}W_{n-1}^*.$$

Ponemos ahora

$$W_n = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W_{n-1} \end{array}\right),$$

$$W_n^*W_n = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W_{n-1}^* \end{array}\right)\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W_{n-1} \end{array}\right) = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W_{n-1}^*W_{n-1} \end{array}\right) = I.$$

Así W_n es unitaria.
Ahora el producto

$$W_n^* V^* A V W_n = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W_{n-1}^* \end{array}\right) \left(\begin{array}{c|c} \lambda_1 & b_{n-1}^* \\ \hline 0 & A_{n-1} \end{array}\right) \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W_{n-1} \end{array}\right),$$

$$W_n^* V^* A V W_n = \left(\begin{array}{c|c} \lambda_1 & b_{n-1}^* \\ \hline 0 & W_{n-1}^* A_{n-1} \end{array}\right) \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W_{n-1} \end{array}\right),$$

$$W_n^* V^* A V W_n = \left(\begin{array}{c|c} \lambda_1 & b_{n-1}^* W_{n-1} \\ \hline 0 & W_{n-1}^* A_{n-1} W_{n-1} \end{array}\right),$$

$$W_n^* V^* A V W_n = \left(\begin{array}{c|c} \lambda_1 & b_{n-1}^* W_{n-1} \\ \hline 0 & T_{n-1} \end{array}\right).$$

Así, el producto $W_n^* V^* A V W_n$ es una matriz triangular. Recordando que el producto de dos matrices uniatarias es unitario y tomando

$$V W_n = U,$$

tenemos

$$A = U T U^*,$$

donde

$$T = \left(\begin{array}{c|c} \lambda_1 & b_{n-1}^* U_{n-1} \\ \hline 0 & T_{n-1} \end{array}\right).$$

□

Definición 3.17. Se dice que una matriz $A \in \mathbb{C}^{n\times n}$ es normal si A conmuta con su adjunta. Es decir

$$AA^* = A^* A.$$

Ejemplo 3.4. Las matrices simétricas, hermitianas, ortogonales, unitarias son normales.

Definición 3.18. Se dice que una matriz A es unitariamente (respectivamente ortogonalmente) semejante a una matriz B si existe una matriz unitaria U (respectivamente ortogonal) tal que

$$A = UBU^*$$

Si B es diagonal se dice que A es unitariamente (respectivamente ortogonalmente) diagonalizable.

Teorema 3.5. *Toda matriz normal es unitariamente diagonalizable. Es decir si*

$$AA^* = A^*A,$$

existen U unitaria y Δ diagonal tales que

$$A = U\Delta U^*.$$

Demostración. Sabemos del Teorema de Schur (página 50) que existen U unitaria y T triangular superior tales que

$$A = UTU^*.$$

La normalidad de A implica

$$TT^* = T^*T. \tag{3.16}$$

Para los elementos diagonales de la igualdad de la ecuación (3.16) tenemos para $i = 1, \ldots, n$

$$(TT^*)_{ii} = (T^*T)_{ii},$$

es decir

$$\sum_{k=1}^{n} T_{ik}\, T^*_{ki} = \sum_{k=1}^{n} T^*_{ik}\, T_{ki},$$

pero

$$T^*_{ik} = \overline{T_{ki}}.$$

Así

$$\sum_{k=1}^{n} |T_{ik}|^2 = \sum_{k=1}^{n} |T_{ki}|^2 \quad \forall i = 1, \ldots, n \tag{3.17}$$

La estructura triangular superior de T reduce la igualdad de la ecuación (3.17) a

$$\sum_{k=i}^{n} |T_{ik}|^2 = \sum_{k=1}^{i} |T_{ki}|^2 . \tag{3.18}$$

Eliminando $|T_{ii}|^2$ de la ecuación (3.18), obtenemos

$$\sum_{k=i+1}^{n} |T_{ik}|^2 = \sum_{k=1}^{i-1} |T_{ki}|^2 , \qquad i = 1, \ldots, n.$$

Para $i = 1$,

$$\sum_{k=2}^{n} |T_{1k}|^2 = 0.$$

Así

$$T_{1k} = 0, \qquad n \geq k \geq 2.$$

Para $i = 2$,

$$\sum_{k=3}^{n} |T_{2k}|^2 = |T_{12}|^2 = 0.$$

Así

$$T_{2k} = 0, \qquad n \geq k \geq 3.$$

Continuando así sucesivamente deducimos que la parte estrictamente triangular superior de T es nula, lo que implica que T es diagonal. □

Nota 3.11.

Con este resultado, podemos deducir que las matrices hermitianas, simétricas, unitarias y ortogonales son todas diagonalizables de manera unitaria u ortogonal.

3.12. Raíz cuadrada de una matriz hermitiana definida positiva

Teorema 3.6. *Si $A \in \mathbb{K}^{n\times n}$ es una matriz hermitiana definida positiva (hermitiana semi–definida positiva), existe una matriz $A^{1/2}$ hermitiana definida positiva (respectivamente hermitiana semi–definida positiva) tal que*

$$\left(A^{1/2}\right)\left(A^{1/2}\right) = \left(A^{1/2}\right)^2 = A.$$

Definición 3.19. La matriz $A^{1/2}$ se llama raíz cuadrada de A.

Demostración. A es normal entonces existe U unitaria tal que

$$A = U^* D U$$

donde D es diagonal con $D_{ii} > 0$ (respectivamente $D_{ii} \geq 0$), $i = 1, \ldots, n$, por el carácter definido positivo (respectivamente semi–definido positivo) de la matriz.

Sea $D^{1/2}$ la matriz diagonal definida por

$$D_{ii}^{1/2} = \sqrt{D_{ii}}, \qquad i = 1, \ldots, n.$$

Tenemos

$$\left(D^{1/2}\right)^2 = D.$$

Ahora

$$A = U^* D U = U^* D^{1/2} D^{1/2} U = \left(U^* D^{1/2} U\right)\left(U^* D^{1/2} U\right).$$

Con $A^{1/2} = \left(U^* D^{1/2} U\right)$, tenemos

$$A = \left(A^{1/2}\right)^2.$$

Ahora

$$\left(A^{1/2}\right)^* = \left(U^* D^{1/2} U\right),$$

es decir $A^{1/2}$ es hermitiana. Además, $\operatorname{esp} A^{1/2} = \left\{\sqrt{D_{ii}},\ i = 1, \ldots, n\right\}$, lo que significa que $A^{1/2}$ es hermitiana definida positiva (respectivamente semi–definida positiva). □

3.13. El cociente de Rayleigh de una matriz hermitiana

Definición 3.20. Sea $A \in \mathbb{C}^{n\times n}$ hermitiana. El cociente de Rayleigh asociado a A es la función

$$\begin{aligned} q_A : (\mathbb{C}^n)^* &\to \mathbb{R} \\ x &\mapsto q_A(x) = \frac{\langle x, Ax\rangle}{\langle x, x\rangle}. \end{aligned}$$

Nota 3.12.

1. Recordemos que $(\mathbb{C}^n)^*$ es el conjunto $\mathbb{C}^n - \{0\}$.
2. La definición de q_A tiene sentido porque A es hermitiana.
3. Para todo $x \in (\mathbb{C}^n)^*$, $\forall \alpha \in \mathbb{C}$
$$q_A(\alpha x) = q_A(x).$$
4. Si x es vector propio de A asociado a $\lambda \in \text{esp}(A)$ entonces
$$q_A(x) = \lambda.$$

Teorema 3.7. *Sea $A \in \mathbb{C}^{n\times n}$ una matriz hermitiana y $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ sus valores propios. Entonces*

$$\lambda_1 = \min_{x\in(\mathbb{C}^n)^*} q_A(x), \qquad \lambda_n = \max_{x\in(\mathbb{C}^n)^*} q_A(x).$$

Demostración. Siendo hermitiana, A es unitariamente diagonalizable, es decir existe una base ortonormal de vectores propios $u_1, \ldots, u_n$ de $\mathbb{C}^n$, asociados respectivamente a $\lambda_1, \ldots, \lambda_n$.
Ahora sea $x \in \mathbb{C}^n$. Existen escalares $x_1, \ldots, x_n$ tales que

$$x = \sum_{i=1}^{n} x_i\, u_i.$$

Por ortonormalidad tenemos

$$\langle x, x\rangle = \sum_{i=1}^{n} |x_i|^2 .$$

Así

$$q_A(x) = \frac{\left\langle A\left(\sum_{i=1}^{n} x_i\, u_i\right), \sum_{i=1}^{n} x_i\, u_i\right\rangle}{\langle x, x\rangle} = \frac{\sum_{i=1}^{n} \lambda_i\ |x_i|^2}{\sum_{i=1}^{n} |x_i|^2},$$

de manera que

$$q_A(x) \geq \min_{i=1,\ldots,n} \lambda_i \frac{\sum |x_i|^2}{\sum |x_i|^2} = \lambda_1$$

y

$$q_A(x) \leq \max_{i=1,\ldots,n} \lambda_i \frac{\sum |x_i|^2}{\sum |x_i|^2} = \lambda_N,$$

para todo $x \in (\mathbb{C}^n)^*$.
Para terminar la demostración es suficiente notar que

$$q_A(u_1) = \lambda_1 \quad \text{y} \quad q_A(u_n) = \lambda_n.$$

□

Capítulo 4

Normas vectoriales y matriciales

4.1. Introducción

En este capitulo prensentamos de manera sencilla las normas en los espacios vectoriales en dimension finita. Estas herramientas son necesarias para evaluar la qualidad de la aproxiamción cuando buscamos una solución aproximada.

4.2. Normas vectoriales

Definición 4.1. Una norma vectorial en $\mathbb{K}^n$ es una función $N : \mathbb{K}^n \to \mathbb{R}_+$ tal que para todo x, y en $\mathbb{K}^n$ y todo λ en $\mathbb{K}$

i) $N(x) = 0$ si y sólo si $x = 0$.

ii) $N(\lambda x) = |\lambda|\ N(x)$.

iii) $N(x + y) \leq N(x) + N(y)$.

Nota 4.1.
Usualmente las normas se denotan $\|\cdot\|$.

Ejemplo 4.1 (Ejemplos importantes)**.** $\mathbb{K}$=$\mathbb{C}$ ó $\mathbb{R}$.

1. Para $p \in [1 \quad +\infty[$ y $x \in \mathbb{K}^n$ se define

$$\|x\|_p = \left(\sum |x_i|^p\right)^{1/p}$$

$x^t = (x_1, x_2, \ldots, x_n)$.

2. Para $x \in \mathbb{K}^n$

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i| .$$

Mostremos que $\|\cdot\|_p$ es una norma en $\mathbb{K}^n$ para $p \in [1 \quad \infty[$. La dificultad está sólo en la desigualdad triangular iii).

Empecemos por el caso $p = 1$.
Para x, y en $\mathbb{K}^n$, se tiene

$$\|x+y\|_1 = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i|\,,$$
$$\|x+y\|_1 \leq \|x\|_1 + \|y\|_1\,.$$

Para el caso $p > 1$ la demostración es menos sencilla.

Primero:
Para todo α, β en $\mathbb{R}_+$ tenemos

$$\alpha\beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q} \tag{4.1}$$

donde p y q son números reales positivos que cumplen la igualdad

$$\frac{1}{q} + \frac{1}{p} = 1$$

y se llama conjugado real de p.
La demostración de la ecuación (4.1) se basa en la convexidad de la función $x \mapsto e^x$. Sabemos que

$$e^{\theta x + (1-\theta)y} \leq \theta e^x + (1-\theta)e^y. \tag{4.2}$$

Eso es debido a que la segunda derivada del exponencial es siempre positiva. Ahora, en la desigualdad (4.2), tomamos $\theta = \frac{1}{p}$, $x = p\,\ln\alpha$, $y = q\,\ln\beta$. Así

$$e^{\theta x + (1-\theta)y} = e^{\ln\alpha + \ln\beta} = \alpha\beta \leq \theta e^x + (1-\theta)e^y = \frac{1}{p}\,\alpha^p + \frac{1}{q}\,\beta^q.$$

Segundo:
Ahora si x, y son dos vectores en $\mathbb{K}^n$ tenemos para todo $i = 1, \ldots, n$

$$\frac{|x_i y_i|}{\|x\|_p\ \|y\|_q} \leq \frac{1}{p}\,\frac{|x_i|^p}{\|x\|_p^p} + \frac{1}{q}\,\frac{|y_i|^q}{\|y\|_q^q}.$$

Sumando esta desigualdad para $i = 1, \ldots, n$ resulta

$$\sum_{i=1}^{n} \frac{|x_i y_i|}{\|x\|_p\ \|y\|_q} \leq \frac{1}{p}\,\frac{\sum\limits_{i=1}^{n} |x_i|^p}{\sum\limits_{i=1}^{n} |x_i|^p} + \frac{1}{q}\,\frac{\sum\limits_{i=1}^{n} |y_i|^q}{\sum\limits_{i=1}^{n} |y_i|^q} = \frac{1}{p} + \frac{1}{q} = 1.$$

Así

$$\sum_{i=1}^{n} |x_i y_i| \leq \|x\|_p \, \|y\|_q \,. \tag{4.3}$$

Esta desigualdad es conocida como la desigualdad de Hölder. Su caso particular para $p = 2$ es la famosa desigualdad de Cauchy–Schwarz.

Volvamos a la demostración.
Tenemos para todo $i = 1, \ldots, n$,

$$(|x_i| + |y_i|)^p = |x_i| \, (|x_i| + |y_i|)^{p-1} + |y_i| \, (|x_i| + |y_i|)^{p-1} \,. \tag{4.4}$$

Sumando las igualdades (4.4)

$$\sum_{i=1}^{n} (|x_i| + |y_i|)^p = \sum_{i=1}^{n} |x_i| \, (|x_i| + |y_i|)^{p-1} + \sum_{i=1}^{n} |y_i| \, (|x_i| + |y_i|)^{p-1} \,.$$

Usando la desigualdad de Hölder viene

$$\sum_{i=1}^{n} (|x_i| + |y_i|)^p \leq \|x\|_p \left(\sum_{i=1}^{n} (|x_i| + |y_i|)^{(p-1)q} \right)^{1/q} + \|y\|_p \left(\sum_{i=1}^{n} (|x_i| + |y_i|)^{(p-1)q} \right)^{1/q} .$$

Pero $(p-1)q = p$. Así

$$\sum_{i=1}^{n} (|x_i| + |y_i|)^p \leq \left(\|x\|_p + \|y\|_q \right) \left(\sum_{i=1}^{n} (|x_i| + |y_i|)^p \right)^{1/q} .$$

Simplificando y usando $\frac{1}{p} + \frac{1}{q} = 1$ tenemos

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p \,.$$

Ejercicio 4.1.

1. Verificar que $\|\cdot\|_\infty$ es una norma en $\mathbb{K}^n$.

2. Mostrar que para $x \in \mathbb{K}^n$ fijo

$$\lim_{p \to \infty} \|x\|_p = \|x\|_\infty \,.$$

   Ayuda: empezar con $n = 2$.

Nota 4.2.

La norma $\|\cdot\|_2$ se llama norma euclidiana y es la única norma que viene de un producto escalar en $\mathbb{K}^n$.

4.3. Normas matriciales relativas

Sean $A \in \mathbb{K}^{m\times n}$ y $\|\cdot\|$, $\|\cdot\|'$ dos normas vectoriales definidas respectivamente en $\mathbb{K}^m$ y $\mathbb{K}^n$. Dado que para todo $x \in \mathbb{K}^n$ las componentes de Ax son funciones lineales en los x_i, podemos encontrar una constante $c > 0$ tal que

$$\|Ax\| \leq c \|x\|', \quad \forall x \in \mathbb{C}^n,$$

Con esta observación la siguiente definición tiene sentido.

Definición 4.2. La norma matricial de A relativa a las normas vectoriales $\|\cdot\|$ y $\|\cdot\|'$ se define por

$$\|A\| = \sup_{x\in(\mathbb{K}^n)^*} \frac{\|Ax\|}{\|x\|'}. \tag{4.5}$$

Nota 4.3.

1. En la práctica, las normas usadas en $\mathbb{K}^n$ y $\mathbb{K}^m$ son las normas $\|\cdot\|_p$ para $p \in [1 \quad +\infty[$. Tomando el mismo p en $\mathbb{K}^n$ y $\mathbb{K}^m$se denota para una matriz $A \in \mathbb{K}^{m\times n}$
$$\|A\|_p = \sup_{x\in(\mathbb{K}^n)^*} \frac{\|Ax\|_p}{\|x\|_p}.$$

2. Las normas más usadas son $\|\cdot\|_1$, $\|\cdot\|_2$ y $\|\cdot\|_\infty$.

3. $$\|A\| = \sup_{x\in(\mathbb{K}^n)^*} \frac{\|Ax\|}{\|x\|} = \max_{x\in(\mathbb{K}^n)^*} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|\leq 1} \|Ax\| = \max_{\|x\|=1} \|Ax\|$$
porque de un lado el conjunto $\{x;\ \|x\| = 1\}$ es compacto, entonces la función $x \to \frac{\|Ax\|}{\|x\|}$ alcanza sus extremos y del otro .
$$\|Ax\| / \|x\| = \|A(x/\|x\|)\| \qquad \text{y} \qquad \|(x/\|x\|)\| = 1$$

4. Para toda norma matricial relativa a una norma vectorial tenemos
$$\|Ax\| \leq \|A\| \cdot \|x\|.$$

Proposición 4.1. *Sea $\|\cdot\|$ una norma matricial relativa a una norma vectorial $\|\cdot\|$. Tenemos*

i) $\|A\| = 0$ *si y sólo si* $A = 0,\ \forall A \in \mathbb{K}^{m\times n}$.

ii) $\|\lambda A\| = |\lambda| \cdot \|A\|, \qquad \forall \lambda \in \mathbb{K},\ \forall A \in \mathbb{K}^{m\times n}$.

iii) $\|A + B\| \leq \|A\| + \|B\|, \qquad \forall A, B \in \mathbb{K}^{m\times n}$.

iv) $\|AB\| \leq \|A\| \cdot \|B\|$.

Demostración.

i) Si $A = 0$, $\|A\| = 0$ es evidente.
Si $\|A\| = 0 \implies \|Ax\| = 0,\ \forall x \in \mathbb{K}^n \implies Ax = 0,\ \forall x \in \mathbb{K}^n \implies A = 0.$

ii) $\|\lambda A\| = \max\limits_{\|x\|=1} \|\lambda Ax\| = \max\limits_{\|x\|=1} (|\lambda| \cdot \|Ax\|) = |\lambda| \cdot \max\limits_{\|x\|=1} (\|Ax\|) = |\lambda| \cdot \|A\|.$

iii) $\|A + B\| = \max\limits_{\|x\|=1} \|(A + B)x\| = \max\limits_{\|x\|=1} \|Ax + Bx\| \leq \max\limits_{\|x\|=1} \|Ax\| + \max\limits_{\|x\|=1} \|Bx\| =$
$\|A\| + \|B\|.$

iv) $\|A\,Bx\| \leq \|A\| \cdot \|Bx\| \leq \|A\| \cdot \|B\| \cdot \|x\|.$
Así
$$\frac{\|A\,Bx\|}{\|x\|} \leq \|A\| \cdot \|B\|\,.$$

□

Nota 4.4.

Es importante precisar que existen normas sobre el espacio de las matrices que verifican las cuatro propiedades pero no son relativas a ninguna norma vectorial como lo muestra el ejercicio siguiente.

Ejercicio 4.2. Para $A \in \mathbb{K}^{n\times n}$ se define

$$\|A\|_s = \left(\sum_{i,j=1}^{n} |A_{ij}|^2 \right)^{1/2}.$$

$\|\cdot\|_s$ se llama norma de Schur.

1. Verificar que $\|\cdot\|_s$ satisface las cuatro propiedades de la proposición (4.1).
2. Calcular $\|I\|_s$.
3. Deducir que $\|\cdot\|_s$ no es relativa a ninguna norma vectorial.

Teorema 4.1. *Sea* $A \in \mathbb{K}^{m\times n}$.

1. $\|A\|_1 = \max\limits_{j=1,\ldots,n} \sum\limits_{i=1}^{m} |a_{ij}|.$
2. $\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)}.$
3. $\|A\|_\infty = \max\limits_{i=1,\ldots,m} \sum\limits_{j=1}^{n} |a_{ij}|.$

Demostración.

1. Sea $x \in \mathbb{K}^n$ tal que $\|x\|_1 \leq 1$ es decir

$$\sum_{i=1}^{n} |x_i| \leq 1.$$

$$\begin{aligned}
\|Ax\|_1 &= \sum_{i=1}^{m} |Ax|_i = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} |a_{ij}| \ |x_j| \right), \\
\|Ax\|_1 &\leq \sum_{j=1}^{n} \sum_{i=1}^{m} |a_{ij}| \ |x_j| \leq \left(\max_j \sum_{i=1}^{m} |a_{ij}| \right) \sum_{j=1}^{n} |x_j|, \\
\|Ax\|_1 &\leq \max_j \left(\sum_{i=1}^{m} |a_{ij}| \right) \|x\|_1, \\
\|Ax\|_1 &\leq \max_j \left(\sum_{i=1}^{m} |a_{ij}| \right).
\end{aligned}$$

Mostremos que esta desigualdad se vuelve igualdad para un x particular. Sea k el índice que satisface

$$\max_j \sum_{i=1}^{n} |a_{ij}| = \sum_{i=1}^{n} |a_{ik}|.$$

y definamos el vector $x \in \mathbb{K}^n$ de la siguiente manera

$$x_i = 0 \qquad \text{si} \qquad i \neq k, \qquad x_k = 1,$$

Con esta elección de x se tiene

$$\|Ax\|_1 = \sum_{i=1}^{n} |a_{ik}| = \max_j \sum_{i=1}^{n} |a_{ij}|.$$

Así

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

2.

$$\|A\|_2^2 = \max_{x \neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \max_{x \neq 0} \frac{\langle Ax, Ax \rangle}{\langle x, x \rangle} = \max_{x \neq 0} \frac{\langle A^*Ax, x \rangle}{\langle x, x \rangle}.$$

Según el cociente de Rayleigh y dado que A^*A es hermitiana, $\text{esp}(A^*A) \subset \mathbb{R}$ y

$$\max_{x \neq 0} \frac{\langle A^*Ax, x \rangle}{\langle x, x \rangle} = \lambda_n,$$

donde $\lambda_n = \max\{\lambda_i, \quad \lambda_i \in \text{esp}\, A^*A\}$.

Como A^*A es semi definida positiva $\text{esp}(A) \subset \mathbb{R}_+$ y $\lambda_n = \rho(A^*A)$. Así $\|A\|_2 = \sqrt{\rho(A^*A)}$.

Como $\text{esp}(A^*A) = (AA^*)$ deducimos que $\|A\|_2 = \sqrt{\rho(AA^*)}$.

3. Sea $x \in \mathbb{K}^n$ tal que $\|x\|_\infty \leq 1$.

$$\|Ax\|_\infty = \max_{1\leq i\leq m} \left| \sum_{j=1}^{n} a_{ij}\, x_j \right| \leq \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|\ |x_j| \leq \max_i \sum_{j=1}^{n} |a_{ij}|$$

porque

$$\max_i |x_j| \leq 1.$$

Ahora sea k el índice donde

$$\max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{kj}|$$

y sea x el vector definido por

$$x_j = \frac{\overline{a_{kj}}}{|a_{kj}|} \qquad \text{si } a_{kj} \neq 0$$

y

$$x_j = 1 \qquad \text{si } a_{kj} = 0.$$

Tenemos $\|x\|_\infty = 1$ y

$$\|Ax\|_\infty = \sum_{j=1}^{n} \left| a_{kj}\, \frac{\overline{a_{kj}}}{|a_{kj}|} \right| = \sum_{j=1}^{n} |a_{kj}| = \max_i \sum_{j=1}^{n} |a_{ij}|\,.$$

□

Proposición 4.2.

1. *Si $U \in \mathbb{K}^{n\times n}$ es unitaria y $A \in \mathbb{K}^{n\times n}$ entonces*

$$\|UA\|_2 = \|AU\|_2 = \|A\|_2\,.$$

2. *Si A es normal*

$$\|A\|_2 = \rho(A).$$

Demostración.

1. Por el teorema (4.1) tenemos

$$\|UA\|_2 = \rho(A^*U^*UA)^{1/2} = \rho(A^*A)^{1/2} = \|A\|_2\,,$$
$$\|AU\|_2 = \rho(U^*A^*AU)^{1/2} = \rho(AUU^*A^*)^{1/2} = \rho(AA^*)^{1/2} = \|A\|_2\,.$$

2. Si A es normal, es unitariamente diagonalizable, entonces existen U unitaria y D diagonal tales que

$$A = UDU^*,$$
$$\|A\|_2 = \|D\|_2 = \rho(D) = \rho(A).$$

□

Observamos de las notas anteriores que existe una estrecha relación entre el radio espectral de una matriz y su norma euclidiana. En algunos casos particulares son iguales.

Veamos ahora si tal relación se extiende a otras normas matriciales.

Proposición 4.3. *Sea $A \in \mathbb{K}^{n\times n}$. Para toda norma matricial relativa a una norma vectorial tenemos*

$$\rho(A) \leq \|A\|. \tag{4.6}$$

Demostración. Sean $\lambda \in \text{esp}(A)$ y $u \in \mathbb{K}^n$ tales que $Au = \lambda u$ y $\|u\| = 1$.

$$\|A\| = \max_{\|x\|=1} \|Ax\| \geq \|Au\| = |\lambda|\ \|u\| = |\lambda|.$$

Así

$$\|A\| \geq \rho(A).$$

□

Nota 4.5.

En realidad, este resultado se puede extender a cualquier norma matricial que satisfaga la desigualdad

$$\|AB\| \leq \|A\|\ \|B\|. \tag{4.7}$$

Definición 4.3. Una norma matricial que satisface la desigualdad (4.7) se le dice norma submultiplicativa.

Nota 4.6.

1. Existen normas matriciales (normas en el sentido que cumplen con las tres primeras propiedades de la proposición (4.1) de la página 60) y que no son submultiplicativas como lo muestra el ejemplo siguiente.

   En $\mathbb{K}^{2\times 2}$;

$$|||A||| = \max_{i,j} |a_{ij}|.$$

   Es evidente que $|||\cdot|||$ cumple con *i) ii) iii)* de dicha proposición pero si tomamos

$$A = \begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad C = \begin{pmatrix} 3 & 4 \\ 0 & 1 \end{pmatrix},$$

$$|||A||| = 3, \qquad |||B||| = 1 \qquad \text{y} \qquad |||AB||| = 4 > |||A||| \cdot |||B|||.$$

2. Por la proposción anterior se tiene

$$\rho(A) \leq \|A\|_s ,$$

donde $\|\cdot\|_s$ es la norma de Schur.

La desigualdad de la proposición (4.3) es tan estrecha que se tiene e siguiente resultado.

Teorema 4.2. *Sean $A \in \mathbb{K}^{n\times n}$ y $\varepsilon > 0$. Existe una norma matricial relativa tal que*

$$\|A\| < \rho(A) + \varepsilon.$$

Nota 4.7.
Este resultado se puede leer de la siguiente forma

$$\rho(A) = \text{ínf}\{\|A\| , \quad \|\cdot\| \text{ es una norma multiplicativa}\}.$$

Antes de demostrar el teorema hagamos el ejercicio siguiente.

Ejercicio 4.3. Sean $\|\cdot\|$ una norma matricial relativa y V una matriz invertible. Mostrar que la función

$$\begin{aligned} \|\cdot\|_V : \mathbb{K}^{n\times n} &\to \mathbb{R}_+ \\ A &\mapsto \|A\|_V = \left\|VAV^{-1}\right\| \end{aligned}$$

es una norma matricial relativa a una norma vectorial.

Demostración. Según el teorema de Schur existen una matriz triangular superior T y una matriz unitaria U tales que

$$A = U^*TU.$$

Sea $\alpha > 0$ y definimos

$$\Delta = \begin{pmatrix} 1 & & & & 0 \\ & \alpha & & & \\ & & \alpha^2 & & \\ & & & \ddots & \\ 0 & & & & \alpha^{n-1} \end{pmatrix},$$

$$\Delta_{ij} = \alpha^{i-1}\,\delta_{ij}, \qquad i,j = 1,\ldots,n.$$

Tenemos

$$\Delta^{-1}T\Delta = \begin{pmatrix} t_{11} & \alpha t_{12} & \alpha^2 t_{13} & \cdots & \alpha^{n-1}t_{1n} \\ & t_{22} & \alpha t_{23} & \ddots & \\ & & \ddots & \ddots & \alpha^2 t_{n-2,n} \\ & & & t_{n-1,n-1} & \alpha t_{n-1,n} \\ 0 & & & & t_{nn} \end{pmatrix}.$$

Notamos que las potencias de α en las diagonales de $\Delta^{-1}T\Delta$ son constantes. Ahora calculemos la norma $\left\|\Delta^{-1}T\Delta\right\|_\infty$. Tenemos

$$\left\|\Delta^{-1}T\Delta\right\|_\infty = \max_i\left[|t_{ii}| + \alpha\, c_i\right]$$

donde

$$c_i = \sum_{j=i+1}^{j=n} |t_{ij}|\, \alpha^{j-i-1}.$$

Podemos también escribir

$$\left\|\Delta^{-1}T\Delta\right\|_\infty = \max_i |t_{ii}| + \max_i \alpha\, c_i.$$

Para $\alpha = 0$, $\alpha\, c_i = 0$ y $\alpha\, c_i$ es un polinomio en α. Entonces por continuidad del polinomio, para todo $\varepsilon > 0$, existe $\delta_i > 0$ tal que si $\alpha < \delta_i$ tenemos $|\alpha\, c_i| < \varepsilon$. Entonces

$$\begin{aligned} &\text{si} \qquad \alpha < \min_i \delta_i \\ &\max_i \alpha\, c_i < \varepsilon. \end{aligned}$$

Para un tal α se tiene

$$\left\|\Delta^{-1}T\Delta\right\|_\infty \leq \max_i |t_{ii}| + \varepsilon.$$

Pero $\text{esp}(A) = \{t_{ii}, \quad i = 1, \ldots, n\}$. Así

$$\left\|\Delta^{-1}T\Delta\right\|_\infty \leq \rho(A) + \varepsilon.$$

Ahora

$$\begin{aligned} T &= UAU^{-1}, \\ \Delta^{-1}T\Delta &= \Delta^{-1}UAU^{-1}\Delta = (U^{-1}\Delta)^{-1}AU^{-1}\Delta. \end{aligned}$$

Llamamos $V = U^{-1}\Delta$, así

$$\Delta^{-1}T\Delta = V^{-1}AV$$

y

$$\left\|\Delta^{-1}T\Delta\right\|_\infty = \left\|V^{-1}AV\right\|_\infty = \|A\|_V \qquad \text{(según el ejercicio 4.3).}$$

Es decir

$$\|A\|_V \leq \rho(A) + \varepsilon.$$

Así construimos una norma $\|\cdot\|_V$ que satisface $\|A\| \leq \rho(A) + \varepsilon$. □

4.4. Convergencia de vectores y matrices

4.4.1. Sucesiones de vectores

Consideremos el sistema de ecuaciones

$$\begin{cases} 3x^2 - 10 \ln y = 1 \\ \cos y - \ln(xy) = 0 \end{cases}$$

donde (x, y) es la pareja de incógnitas.

Es evidente que es imposible resolver este sistema de manera analítica (o al menos para mi :-)). Para problemas de este tipo la única manera que queda es encontrar una aproximación de la solución. Para eso tenemos que estar seguros *a priori* de la existencia de la solución (lo que es el caso de este ejemplo).

Sabiendo que la solución existe, viene el proceso de buscar un aproximación a la misma. Usualmente se construye una sucesión de aproximaciones de manera iterativa de modo que la aproximación sea "mejor" a medida que el número de iteraciones va creciendo.

Para el caso de nuestro ejemplo, se construye una sucesión $(x_0, y_0), (x_1, y_1), \ldots$ que "converge" a la solución exacta. Así estamos usanando vectores $(x_i, y_i)_{i\in\mathbb{N}}$. Definamos ahora la convergencia de vectores.

Notación

Para una sucesión de $\mathbb{K}^k$, adoptemos la convención de notar sus términos de la siguiente manera

$$(x^n)_{n\in\mathbb{N}} \qquad x^n = (x_1^n, \ldots, x_k^n)^t$$

es decir el índice de abajo es de la componente en $\mathbb{K}^n$y el de arriba es el de la suceción.

Definición 4.4. Sean $(x^n)_{n\in\mathbb{N}}$ una sucesión en $\mathbb{K}^k$ y $\|\cdot\|$ una norma vectorial en $\mathbb{K}^k$. Se dice que $(x^n)_{n\in\mathbb{N}}$ es convergente si existe un vector x en $\mathbb{K}^k$ tal que

$$\begin{gathered} \forall \varepsilon > 0, \ \exists N \in \mathbb{N} \quad \text{tal que} \\ n \geq N \quad \text{implica} \quad \|x^n - x\| < \varepsilon. \end{gathered}$$

Nota 4.8.

1. En práctica, se trabaja con las normas $\|\cdot\|_1$, $\|\cdot\|_2$ y $\|\cdot\|_\infty$.
2. A priori la convergencia de una sucesión de vectores depende de la norma, pero como veremos en el teorema siguiente la convergencia para una norma implica la convergencia para cualquier otra norma en $\mathbb{K}^k$.

Definición 4.5. Decimos que dos normas $\|\cdot\|$ y $\|\cdot\|'$ en $\mathbb{K}^n$ son equivalentes si existen $\alpha > 0$, $\beta > 0$ tales que

$$\forall x \in \mathbb{K}^n, \qquad \alpha \|x\| \leq \|x\|' \leq \beta \|x\| .$$

Nota 4.9.
Es evidente que si dos normas $\|\cdot\|$ y $\|\cdot\|'$ satisfacen

$$\|\cdot\|' < \beta \|\cdot\|$$

entonces la convergencia en $\|\cdot\|$ implica la convergencia en $\|\cdot\|'$. (Verificarlo a título de ejercicio)

Así si dos normas en $\mathbb{K}^n$ son equivalentes tienen las mismas sucesiones convergentes.

Teorema 4.3. *1. Sea $E = \{1, 2, \infty\}$. Para todo p, q en E, existe $\alpha_{pq} > 0$ tal que*

$$\forall x \in \mathbb{K}^n, \qquad \|x\|_p \leq \alpha_{pq} \|x\|_q .$$

2. Todas las normas en $\mathbb{K}^n$ son equivalentes.

Demostración. Sea $x \in \mathbb{K}^n$. La norma $\|x\|_\infty$ es igual a uno de los números $|x_1|, |x_2|, \ldots, |x_n|$ entonces

$$\|x\|_\infty \leq \sum_{i=1}^{n} |x_i| = \|x\|_1 .$$

Así $\alpha_{\infty 1} = 1$.

De la misma manera

$$\forall i = 1, \ldots, n, \qquad \|x\|_\infty \geq |x_i| .$$

Sumando

$$n \|x\|_\infty \geq \|x\|_1 .$$

Así $\alpha_{1\infty} = n$.

Igualmente, entre $\|\cdot\|_1$ y $\|\cdot\|_2$ notemos que

$$\|x\|_2 \geq |x_i| , \qquad \forall i = 1, \ldots, n.$$

Sumando

$$n \|x\|_2 \geq \|x\|_1 ,$$

de manera que $\alpha_{12} = n$ y

$$\|x\|_1 \geq |x_i| , \qquad i = 1, \ldots, n.$$

Elevando al cuadrado y sumando

$$n \|x\|_1^2 \geq \sum_{i=1}^{n} |x_i|^2 .$$

Tomando la raíz

$$\|x\|_2 \leq \sqrt{n} \|x\|_1 .$$

Así $\alpha_{21} = \sqrt{n}$.

La equivalencia entre $\|\cdot\|_\infty$ y $\|\cdot\|_2$ se trata de la misma manera. Hacerlo como ejercicio.

La demostración de la segunda parte del teorema necesita argumentos de continuidad fuera del ámbito de este curso. □

Nota 4.10.

La convergencia de una sucesión en $\mathbb{K}^n$es independiente de la elección de la norma.

Ejercicio 4.4. Mostrar que si $(x^n)_{n\in\mathbb{N}}$ es una sucesión en $\mathbb{K}^k$ entonces tenemos equivalencia entre:

i) $(x^n)_{n\in\mathbb{N}}$ es convergente en $\mathbb{K}^k$.

ii) $\forall i,\ i = 1, \ldots, k,\ (x_i^n)_{n\in\mathbb{N}}$ es convergente en $\mathbb{K}$.

4.4.2. Convergencia de sucesiones de matrices

En esta parte, en realidad no nos interesamos en las sucesiones de matrices de manera general sino sólo en las sucesiones de potencias de matrices, es decir en las que son de la forma

$$(A^k)_{k\in\mathbb{N}},$$

para una matriz $A \in \mathbb{K}^{n\times n}$, y donde $A^k = k$-ésima potencia de A.

Nota 4.11.

Dado que una matriz $A \in \mathbb{K}^{n\times n}$ puede ser considerada como vector en $\mathbb{K}^{mn}$ y como todas las normas en $\mathbb{K}^{mn}$ son equivalentes, estudiar las sucesiones de matrices se vuelve lo mismo que estudiar sucesiones de vectores.

Definición 4.6. Sea $A \in \mathbb{K}^{n\times n}$. Se dice que la sucesión $(A^k)_{k\geq 0}$ es convergente si existe una norma matricial tal que

$$\lim_{k\to\infty} \left\|A^k\right\| = 0.$$

Nota 4.12.

Sabemos que si $(A^k)_{k\geq 0}$ converge para una norma, converge para cualquier otra norma matricial (relativa o no).

Teorema 4.4. *Sea $A \in \mathbb{K}^{n\times n}$. Las siguientes 4 aserciones son equivalentes.*

i) $(A^k)_{k\geq 0}$ *es convergente.*

ii) $\forall x \in \mathbb{K}^n, \quad A^k x \to 0.$

iii) $\rho(A) < 1.$

iv) Existe una norma multiplicativa tal que $\|A\| < 1$.

Demostración. Hacemos la demostración en el sentido *i)* $\Longrightarrow$ *ii)* $\Longrightarrow$ *iii)* $\Longrightarrow$ *iv)* $\Longrightarrow$ *i)*.

i) $\Longrightarrow$ *ii)*. Sea $x \in \mathbb{K}^n$.
Tenemos

$$\left\|A^k x\right\| \leq \left\|A^k\right\| \cdot \|x\| \tag{4.8}$$

para toda norma inducida. El segundo miembro de la desigualdad de la ecuación (4.8) tiende a cero por *i)*. Así la sucesión vectorial $A^k x$ converge a 0.

ii) $\Longrightarrow$ *iii)*. Sean $\lambda \in \text{esp}(A)$ y u un vector propio asociado a λ.
Tenemos $Au = \lambda u$. Aplicando A k veces

$$A^k u = \lambda^k u. \tag{4.9}$$

ii) implica que $\lambda^k u \to 0$. Es decir $\left\|\lambda^k u\right\| \to 0$ para toda norma vectorial. Es decir $|\lambda|^k \|u\| \to 0$ lo que sólo es posible si $|\lambda| < 1$. Así $\rho(A) < 1$.

iii) $\Longrightarrow$ *iv)*. Sabemos del teorema (4.2) de la pagina 65 que $\forall \varepsilon > 0$, existe una norma matricial inducida $\|\cdot\|$, tal que

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

Si escogemos

$$\varepsilon = \frac{1 - \rho(A)}{2}$$

deducimos que $\|A\| < 1$.

iv) $\Longrightarrow$ *i)*. Tenemos

$$\left\|A^k\right\| \leq \|A\|^k, \qquad \forall k \geq 0.$$

Así $\lim \left\|A^k\right\| = 0$ para la norma proporcionada por *iv*. □

4.5. Sensibilidad a perturbaciones y Condicionamiento de una matriz

4.5.1. Ejemplo de introducción

Consideramos el siguiente sistema lineal

$$Ax = b, \tag{4.10}$$

donde

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 1 \end{pmatrix} \text{ y } b = \begin{pmatrix} 6 \\ 11 \\ 9 \end{pmatrix}.$$

La matriz A es invertible y su inversa es dada por

$$A^{-1} = \begin{pmatrix} 21 & -13 & 2 \\ -13 & 8 & -1 \\ 2 & -1 & 0 \end{pmatrix}$$

La solución de este sistema es el vector $u^t = (1, 1, 1)$. Ahora perturbemos levelente la matriz A agregándole una matriz de perturbación que notaremos δA

$$\delta A = \begin{pmatrix} 0{,}056 & 0{,}072 & 0{,}018 \\ 0{,}087 & -0{,}026 & 0{,}013 \\ 0{,}039 & -0{,}04 & 0{,}072 \end{pmatrix}$$

La solución del sistema perturbado $(A + \delta A)x = b$ es el vector

$$u + \delta u = (2786.59, -1740.92, 222.18)^t$$

Por comodidad medimos los errores con la normas infinitas (matricial y vectorial). Así, el error relativo cometido en la matriz es

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} \approx \frac{0{,}146}{11} \approx 0{,}0012$$

y el error relativo que resultó en la solución es

$$\frac{\|\delta u\|_\infty}{\|u\|_\infty} = \frac{2785{,}59}{1} \approx 2785{,}59$$

La "patología" de este sistema lineal se presenta en que un error de orden 0.0012 en los datos (la matriz en este caso) generó un error de orden de 2785. Es decir que el error relativo se multiplicó por $2785/0{,}012 \approx 232133$ veces. Intolerable!

analicemos ahora como afectan los cambios pequeños en el vector b del sistema (4.10) los resultados del sistema perturbado. La solución de este sistema lineal con

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 1 \end{pmatrix} \text{ y } b' = \begin{pmatrix} 6 + 0.094 \\ 11 - 0.098 \\ 9 + 0.093 \end{pmatrix}.$$

es el vector

$$u + \delta u = (\ 4.433999999, \text{-}1.098999999, 1.286000000\ ).$$

Los errores relativos son

$$\frac{\|\delta b\|_\infty}{\|b\|_\infty} \approx \frac{0.098}{11} \approx 0.009 \quad \frac{\|\delta u\|_\infty}{\|u\|_\infty} \approx \frac{3.433}{1} \approx 3.433\ ,$$

es decir que el error relativo se multiplicó por $3.433/0.009 \approx 382$ veces.

4.5.2. El análisis teórico

Analicemos como afectan las perturbaciones en los datos del sistema lineal la calidad de los resultados. Para eso consideremos el sistema lineal $Ax = b$ cuya matriz A es invertible y sea u su solución. Sea $\|.\|$ una norma matricial relativa a una norma vectorial que denotaremos de la misma manera.

Pertubaciones en la matriz

Proposición 4.4. *Sea $\delta A \in \mathbb{K}^{n\times n}$. Si $u + \delta u$ es una solución del sistema lineal*

$$(A + \delta A)x = b,$$

entonces

$$\frac{\|\delta u\|}{\|u + \delta u\|} \leq \|A\| \, \|A^{-1}\| \frac{\|\delta A\|}{\|A\|} \tag{4.11}$$

Demostración. Tenemos

$$(A + \delta A)(u + \delta u) = b.$$

Escrita de otra manera

$$A(u + \delta u) + \delta A(u + \delta u) = b$$

Multiplicando esta última ecuación por A^{-1} y teniendo en cuenta que $Au = b$ se obtiene

$$\delta u = -A^{-1}\delta A(u + \delta u)$$

de donde se tiene

$$\|\delta u\| \leq \|A^{-1}\| \, \|\delta A\| \|u + \delta u\|$$

lo que implica

$$\frac{\|\delta u\|}{\|u + \delta u\|} \leq \|A^{-1}\| \, \|\delta A\| = \|A\| \, \|A^{-1}\| \frac{\|\delta A\|}{\|A\|}$$

□

Pertubaciones en el vector b

Proposición 4.5. *Si δb es una perturbación en el vector b entonces la solución $u + \delta u$ del sistema lineal*

$$Ax = b + \delta b$$

satisface

$$\frac{\|\delta u\|}{\|u\|} \leq \|A\| \, \|A^{-1}\| \frac{\|\delta b\|}{\|b\|} \tag{4.12}$$

Demostración. Tenemos $b = Au$ así

$$\|b\| \leq \|A\| \, \|u\|. \tag{4.13}$$

También $A(u + \delta u) = b + \delta b$. Desarrollando esta ecuación y remplazando b, se tiene $A\delta u = \delta b$, es decir $\delta u = A^{-1}\delta b$ lo que implica

$$\|\delta u\| \leq \|A^{-1}\| \, \|\delta b\| \tag{4.14}$$

Combinando (4.13) y (4.14) se obtiene la desigualdad buscada (4.12). □

Notemos que en las dos proposiciones anteriores el error relativo del vector solución del sistema lineal inicial es controlado por el factor $\|A\| \, \|A^{-1}\|$ y que aparece en las dos acotaciones (4.11) y (4.12).

Definición 4.7. Sea $A \in \mathbb{K}^{n\times n}$ una matriz invertible y $\|.\|$ una norma matricial. El número de condición de la matriz A asociado a la norma $\|.\|$ es el número dado por

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

Nota 4.13.

1. El número de condición se define siempre con una norma relativa, pero dado que todas las normas son equivalentes, este hecho no influye mucho en la estimación del error.

2. Cuando no hay ambigüedad, y es lo que sucede a menudo, omitimos la dependencia de la norma en la notación del número de condición.

3. El caso particular de la norma euclidiana tiene su propia notación.

$$\kappa(A) = \text{cond}_{\|.\|_2}(A)$$

4. A la luz de los resultados anteriores, notemos que mientras más pequeño es el número de condición de la matriz, menos es la amplificación del error en la solución del sistema lineal.

Propiedades del número de condición

1. Para toda matriz A se tiene

$$\text{cond}(A) \geq 1$$

pues

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = \text{cond}(A)$$

2. Si $A \in \mathbb{K}^{n\times n}$ es normal entonces

$$\kappa(A) = \frac{\lambda_n}{\lambda_1},$$

donde

$$\lambda_1 = \text{mín}\{|\lambda|,\ \lambda \in \text{esp}(A)\} \text{ y } \lambda_n = \text{máx}\{|\lambda|,\ \lambda \in \text{esp}(A)\},$$

en efecto, por el teorema de Schur, la normalidad de A implica la existencia de $U \in \mathbb{K}^{n\times n}$ unitaria y $D \in \mathbb{K}^{n\times n}$ diagonal tal que $A = U^*DU$. Así por la invariancia de la norma euclidiana por la multiplicación por matrices unitarias,

$$\|A\|_2 = \|U^*DU\|_2 = \|D\|_2 = |\lambda_n|$$

y

$$\|A^{-1}\|_2 = \|U^*D^{-1}U\|_2 = \|D^{-1}\|_2 = |\lambda_1|^{-1}$$

3. Si $\kappa(A) = 1$ necesariamente A es unitaria pues en este caso todos los valores singulares de A (i.e los valores propios de AA^*) son de módulo 1.

4. Para todo $\alpha \in \mathbb{K}, \alpha \neq 0$

$$\text{cond}(\alpha A) = \text{cond}(A) \tag{4.15}$$

Decimos que una matriz invertible está **bien acondicionada** si su número de condición está relativamente cerca de 1 y se le dice **mal acondicionada** si su número de condición está muy grande comparado con 1. Con este concepto la matriz del ejemplo introductorio debe estar muy mal acondicionada puesto que los errores "admisibles y realistas "de los datos generaron resultados muy lejos de la solución exacta. Un simple cálculo muestra que el numero de condición de la matriz A al respecto de la norma $\|.\|_\infty$ es

$$\text{cond}(A) = 11 \times 36 = 396$$

Lo que es mucho para una matriz 3×3. Otra observación que podemos hacer es que debido a (4.15) no sirve de nada multiplicar un sistema lineal por un escalar para mejorarle el número de condición.

Capítulo 5

Métodos directos de solución de sistemas lineales

Introducción

Nos interesamos en este capítulo en los métodos directos de solución de sistemas lineales de tipo

$$Ax = b, \tag{5.1}$$

donde $A \in \mathbb{K}^{n\times n}$, $b \in \mathbb{K}^n$ con

$$\det(A) \neq 0.$$

En estos métodos el sistema (5.1) se resuelve en un número finito de pasos lo cual significa que se halla la solución exacta en un número finito de operaciones.

Nota 5.1.
Cuando se dice solución exacta no se toman en cuenta los errores de cálculo ni las representaciones de los números en las máquinas de cálculo.

5.1. Sistemas triangulares

Los sistemas lineales con matrices triangulares (superiores o inferiores) no necesitan ningún cálculo sofisticado.

1. Si el sistema lineal es de tipo

   $$Ux = b,$$

   donde U es una matriz triangular superior invertible de tamaño n y $b \in \mathbb{K}^n$, entonces la solución se da por substitución hacia arriba con la fórmula

   $$x_{n-i} = \frac{1}{a_{n-i,n-i}} \left(b_{n-i} - \sum_{n-i+1}^{n} a_{n-i,j}\, x_j \right), \qquad i = 0, \ldots, n-1,$$

que también se puede escribir

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j=i+1}^{n} a_{ij}\, x_j \right), \qquad i = n, \ldots, 1.$$

2. Si la matriz L del sistema triangular es inferior invertible entonces la solución del sistema lineal

$$Lx = b$$

es dada por substitución hacia abajo con la fórmula

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j \right), \qquad i = 1, \ldots, n.$$

Nota 5.2.

Calculemos el número de operaciones necesarias para resolver un sistema triangular. Para calcular x_i se necesitan $(n - i + 1)$ multiplicaciones y $(n - i)$ adiciones, $i = 1, \ldots, n$. Así el número total de operaciones es

$$\sum_{i=1}^{n} (2n - 2i + 1) = n^2.$$

5.2. El método de eliminación de Gauss

Sea el sistema lineal

$$Ax = b, \tag{5.2}$$

donde $A \in \mathbb{K}^{n \times n}$ es una matriz invertible y $b \in \mathbb{K}^n$.

5.2.1. Operaciones elementales de filas

El método de Gauss consiste en realizar operaciones de ecuaciones en el sistema lineal (5.2) para llegar a un sistema lineal equivalente

$$Ux = b',$$

donde $U \in \mathbb{K}^{n \times n}$ es una matriz triangular superior y $b' \in \mathbb{K}^n$.

Nota 5.3.

1. Las operaciones elementales mencionadas son las siguientes:

   a) multiplicar una ecuación por un escalar no nulo;

   b) permutar dos ecuaciones;

   c) reemplazar una ecuación por la suma de ésta ecuación y otra.

2. Con sistemas lineales equivalentes se quiere decir que se obtiene uno del otro a partir de una sucesión finita de operaciones elementales de ecuaciones.

Para facilitar las notaciones consideremos la matriz $M \in \mathbb{K}^{n\times(n+1)}$ definida por

$$M_{ij} = A_{ij}, \qquad 1 \le i, j \le n$$

y

$$M_{i,n+1} = b_i, \qquad i = 1, \ldots, n,$$

$$M = \overset{n+1}{\left(\begin{array}{c|c} A & b \end{array}\right)}\ {\scriptstyle n}\ .$$

Definición 5.1. La matriz M así definida se llama matriz aumentada del sistema (5.2).

5.2.2. Los pasos del método de Gauss

Paso 1

En este paso se hacen las operaciones en las filas de la matriz aumentada M en tal forma que todos los elementos de la primera columna que están debajo del elemento a_{11} sean nulos. Más explícitamente, si $a_{11} \neq 0$ se multiplica la primera fila por

$$c_{i1} = -\frac{a_{i1}}{a_{11}}$$

y se suma el resultado a la fila i, $i = 2, \ldots, n$. Con estas operaciones, la matriz M se transforma en la matriz $M^{(1)}$ dada por

$$M^{(1)} = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1,n} & m_{1,n+1} \\ 0 & m_{22}^{(1)} & \cdots & m_{2,n}^{(1)} & m_{2,n+1}^{(1)} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & m_{n,2}^{(1)} & \cdots & m_{n,n}^{(1)} & m_{n,n+1}^{(1)} \end{pmatrix}.$$

Si $a_{11} = 0$, se realiza un intercambio de filas de manera que la nueva matriz tenga un elemento no nulo en la posición $(1, 1)$.

Paso 2

En este paso, las operaciones de filas se hacen sobre la matriz $M^{(1)}$ de manera que la matriz obtenida, $M^{(2)}$, tenga todos los elementos de la segunda columna y que estén abajo de la diagonal nulos. Obviamente estas operaciones se hacen sin perder los ceros de la primera columna obtenidos en el paso 1.

Si $m_{22}^{(1)} \neq 0$, la matriz $M^{(2)}$ es dada por

$$M^{(2)} = \begin{pmatrix} m_{11} & m_{12} & \cdots & \cdots & m_{1,n} & m_{1,n+1} \\ 0 & m_{22}^{(1)} & \cdots & \cdots & m_{2,n}^{(1)} & m_{2,n+1}^{(1)} \\ \vdots & 0 & m_{33}^{(2)} & \cdots & m_{3,n}^{(2)} & m_{3,n+1}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & m_{n,3}^{(2)} & \cdots & m_{n,n}^{(2)} & m_{n,n+1}^{(2)} \end{pmatrix},$$

donde

$$m_{ij}^{(2)} = -\frac{m_{i2}^{(1)}}{m_{22}^{(1)}} m_{2j}^{(1)} + m_{ij}^{(1)}, \qquad i \geq 3, j \geq 3.$$

Si $m_{22}^{(1)} = 0$, se realiza intercambio de la segunda fila con una fila inferior en la matriz $M^{(1)}$ de manera que la nueva matriz tenga un elemento no nulo en la posición $(2, 2)$.

El paso general k

Supongamos que ya tenemos la matriz $M^{(k-1)}$ donde los elementos debajo de la diagonal de las columnas $1, 2, \ldots, (k-1)$ son nulos y queremos calcular $M^{(k)}$. Así

$$M^{(k-1)} = \begin{pmatrix} m_{11} & & & & & \cdots & m_{1,n+1} \\ 0 & m_{22}^{(1)} & & & & \cdots & m_{2,n+1}^{(1)} \\ & 0 & \ddots & & & & \vdots \\ & & & m_{k-1,k-1}^{(k-2)} & \cdots & \cdots & m_{k-1,n+1}^{(k-2)} \\ & & & 0 & m_{kk}^{(k-1)} & \cdots & m_{k,n+1}^{(k-1)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & m_{n,k}^{(k-1)} & \cdots & m_{n,n+1}^{(k-1)} \end{pmatrix}.$$

Si $m_{kk}^{(k-1)} = 0$ efectuamos un intercambio de la fila k y otra abajo de esta (veremos en un ejerciocio más adelante que si $\det(A) \neq 0$ este intercambio es siempre posible) de manera que la nueva matriz tenga en la posición (k, k) un elemento no nulo y sigamos las operaciones.

Si $m_{kk}^{(k-1)} \neq 0$ (eventualmente después de un intercambio) multiplicamos la fila k por

$$c_{ik} = -\frac{m_{ik}^{(k-1)}}{m_{kk}^{(k-1)}}$$

y sumamos el resultado a la fila i, $i = k+1, \ldots, n$. El resultado es la matriz $M_{(k)}$ dada por

$$M^{(k)} = \begin{pmatrix} m_{11} & & & & & & m_{1,n+1} \\ 0 & \ddots & & & & & m^{(1)}_{2,n+1} \\ \vdots & & \ddots & & & & \vdots \\ & & & m^{(k-1)}_{k,k} & & & \\ & & & 0 & m^{(k)}_{k+1,k+1} & \cdots & m^{(k)}_{k+1,n+1} \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \cdots & & 0 & m^{(k)}_{n,k+1} & \cdots & m^{(k)}_{n,n+1} \end{pmatrix},$$

donde

$$m^{(k)}_{ij} = -\frac{m^{(k-1)}_{ik}}{m^{(k-1)}_{kk}}\, m^{(k-1)}_{kj} + m^{(k-1)}_{ij},$$

para

$$k+1 \leq i \leq n, k+1 \leq j \leq n+1.$$

Al paso $n-1$ la matriz $M^{(n-1)}$ tiene la estructura de la figura (5.2), es decir, la submatriz principal U de orden n de $M^{(n-1)}$ es triangular superior.

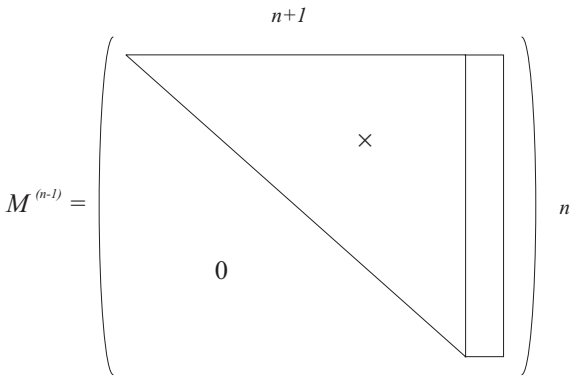

Figura 5.1.

Si llamamos $b^{(n-1)}$ la última columna de $M^{(n-1)}$, el sistema lineal que hay que resolver al fin de esta eliminación es

$$Ux = b^{(n-1)}.$$

5.3. Estudio general del método de Gauss

Definición 5.2. Sea A una matriz invertible de orden n. Decimos que A es directamente Gauss–reducible si en la eliminación de Gauss no se necesita en ningún paso intercambiar filas.

Nota 5.4.

Esta definición significa que A es directamente Gauss–reducible si y sólo si para todo $k = 1, \ldots, n-1$

$$m_{kk}^{k-1} \neq 0,$$

con

$$m_{11}^{0} = a_{11}.$$

Ahora supongamos que A es directamente Gauss–reducible. Pasar de la matriz M (recordar que M es la matriz A aumentada por b) a la matriz $M^{(1)}$ es equivalente a multiplicar la matriz M por la matriz cuadrada

$$L^{(1)} = \begin{pmatrix} 1 & & & & 0 \\ c_{21} & 1 & & & \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ c_{n1} & 0 & & & 1 \end{pmatrix},$$

es decir,

$$L^{(1)} M = M^{(1)}.$$

De la misma manera pasar de $M^{(1)}$ a $M^{(2)}$ equivale a multiplicar $M^{(1)}$ por la matriz

$$L^{(2)} = \begin{pmatrix} 1 & & & & 0 \\ 0 & 1 & & & \\ \vdots & c_{32} & 1 & & \\ \vdots & \vdots & 0 & \ddots & \\ 0 & c_{n2} & & & 1 \end{pmatrix},$$

es decir,

$$L^{(2)} M^{(1)} = M^{(2)}$$

y de manera general

$$M^{(k)} = L^{(k)} M^{(k-1)},$$

donde

$$L^{(k)} = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & c_{k+1,k} & \ddots & \\ & & \vdots & & \ddots \\ 0 & & c_{n,k} & & 1 \end{pmatrix}.$$

Recordemos que

$$c_{ik} = -\frac{m_{ik}^{(k-1)}}{m_{kk}^{(k-1)}}.$$

Agrupando estos pasos, resulta

$$L^{(n-1)} \dots L^{(2)} L^{(1)} M = M^{(n-1)}. \tag{5.3}$$

Si escribimos la ecuación (5.3) en términos del sistema inicial

$$Ax = b,$$

tenemos

$$L^{(n-1)} \dots L^{(2)} L^{(1)} Ax = L^{(n-1)} \dots L^{(2)} L^{(1)} b,$$

pero

$$L^{(n-1)} \dots L^{(2)} L^{(1)} A = U. \tag{5.4}$$

Así

$$Ux = \Lambda b, \tag{5.5}$$

donde U es una matriz triangular y Λb es el vector dado por la última columna de la matriz $M^{(n-1)}$.

La solución del sistema $Ax = b$ se reduce a resolver el sistema triangular (5.5).

5.4. Descomposición LU

Supongamos que tenemos el sistema

$$Ax = b,$$

con $A \in \mathbb{K}^{n\times n}$ invertible y directamente Gauss–reducible. La eliminación de Gauss nos lleva al sistema

$$Ux = \Lambda b, \tag{5.6}$$

con

$$\Lambda = L^{(n-1)} \dots L^{(2)} L^{(1)}.$$

Miremos más de cerca la estructura de la matriz Λ.

Definición 5.3. Sea $1 \le k \le n-1$. Una matriz $V \in \mathbb{K}^{n\times n}$ es de tipo $L^{(k)}$ si existen escalares $\alpha_{k+1}, \dots, \alpha_n$ tales que

$$V = I + \sum_{i=k+1}^{n} \alpha_i \, e_i e_k^*,$$

donde $\{e_i,\ i = 1, \dots, n\}$ es la base canónica de $\mathbb{K}^n$.

Nota 5.5.
Las matrices de tipo $L^{(k)}$ tienen la estructura de la figura siguiente

$$\begin{array}{c} \\ k \\ \downarrow \\ \end{array}$$
$$\begin{pmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & \alpha_{k+1} & \ddots & & \\ & & \vdots & & \ddots & \\ 0 & & \alpha_n & & & 1 \end{pmatrix}$$

Proposición 5.1.

a) *La inversa de una matriz de tipo L^k es de tipo L^k. Además tenemos*

$$\left(I + \sum_{i=k+1}^{n} \alpha_i \, e_i e_k^*\right)^{-1} = I - \sum_{i=k+1}^{n} \alpha_i \, e_i e_k^*.$$

b) *Si M y N son matrices de tipo L^k y L^l respectivamente con $k < l$ entonces la matriz MN tiene la forma*

$$\begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & \times & \ddots & & & \\ & & \times & & 1 & & \\ & & \times & & \times & \ddots & \\ 0 & & \times & 0 & \times & 0 & 1 \end{pmatrix},$$
$$\begin{array}{ccc} & \uparrow & \uparrow \\ & k & l \end{array}$$

es decir, si

$$M = I + \sum_{i=k+1}^{n} \alpha_i \, e_i e_k^*, \qquad N = I + \sum_{j=l+1}^{n} \beta_j \, e_j e_l^*,$$

entonces

$$MN = M + N - I.$$

Demostración.

a)

$$\left(I+\sum_{i=k+1}^{n}\alpha_i\, e_i e_k^*\right)\left(I-\sum_{j=k+1}^{n}\alpha_j\, e_j e_k^*\right)=I+\sum_{i=k+1}^{n}\alpha_i\, e_i e_k^*-\sum_{j=k+1}^{n}\alpha_j\, e_j e_k^*-\sum_{i,j=k+1}^{n}\alpha_i\alpha_j e_i e_k^* e_j e_k^*,$$
$$=I-\sum_{i,j=k+1}^{n}\alpha_i\alpha_j e_i\delta_{jk}e_k^*. \tag{5.7}$$

Como j empieza en $k+1$, entonces $\delta_{jk}=0$. Así

$$\left(I+\sum_{i=k+1}^{n}\alpha_i\, e_i e_k^*\right)^{-1}=I-\sum_{i=k+1}^{n}\alpha_i\, e_i e_k^*.$$

b)

$$MN=\left(I+\sum_{i=k+1}^{n}\alpha_i\, e_i e_k^*\right)\left(I+\sum_{j=l+1}^{n}\beta_j\, e_j e_l^*\right),$$
$$=I+\sum_{i=k+1}^{n}\alpha_i\, e_i e_k^*+\sum_{j=l+1}^{n}\beta_j\, e_j e_l^*+\sum_{\substack{i=k+1\\ j=l+1}}^{n}\alpha_i\beta_j e_i e_k^* e_j e_l^*,$$
$$=M+N-I+\sum_{\substack{i=k+1\\ j=l+1}}^{n}\alpha_i\beta_j e_i\delta_{jk}e_l^*.$$

Tenemos que j empieza por $l+1$. Así $j>l$ y $l\geq k$. Así $j>k$ y $\delta_{jk}=0$ lo que significa que

$$MN=M+N-I.$$

□

Cuidado:

El orden es importante como lo muestra el ejemplo siguiente.

$$A=\begin{pmatrix}1&0&0\\0&1&0\\0&1&1\end{pmatrix},\qquad B=\begin{pmatrix}1&0&0\\1&1&0\\1&0&1\end{pmatrix},\qquad AB=\begin{pmatrix}1&0&0\\1&1&0\\2&1&1\end{pmatrix},$$

$$AB\neq A+B-I$$

porque A es de tipo $L^{(2)}$ y B de tipo $L^{(1)}$.

Nota 5.6.

De manera general si $1 \leq k_1 < k_2 < \ldots < k_p \leq n$ con $p \leq n-1$ y $M_1, \ldots, M_p$ son matrices de tipo $L^{(k_1)}, L^{(k_2)}, \ldots, L^{(k_p)}$ entonces

$$M_1 M_2 \ldots M_p = M_1 + M_2 + \cdots M_p - (p-1)I$$

Volvamos ahora a la ecuación (5.6) de la página 81. La eliminación de Gauss da

$$Ux = \Lambda b,$$

con U triangular superior y

$$\Lambda = L^{(n-1)} \ldots L^{(2)} L^{(1)}.$$

Λ es invertible entonces

$$\Lambda^{-1} = {L^{(1)}}^{-1} {L^{(2)}}^{-1} \ldots {L^{(n-1)}}^{-1}.$$

Las matrices ${L^{(k)}}^{-1}$, $k = 1, \ldots, n-1$ son de tipo $L^{(k)}$ (de hecho son el origen de la definición). Así

$$\Lambda^{-1} = \sum_{i=1}^{n-1} \left(L^{(i)}\right)^{-1} - (n-2)I.$$

Denotemos $\Lambda^{-1} = L$. Según la nota (5.5) L es triangular inferior con unos en la diagonal. Además como

$$L^{(k)} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & c_{k+1,k} & \ddots & & \\ & & \vdots & & \ddots & \\ & & c_{n,k} & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ k \\ \downarrow \end{matrix},$$

deducimos que

$$L = \begin{pmatrix} 1 & & & & 0 \\ -c_{21} & 1 & & & \\ \vdots & -c_{32} & \ddots & & \\ \vdots & & \ddots & 1 & \\ -c_{n1} & \cdots & & -c_{n,n-1} & 1 \end{pmatrix}.$$

De la ecuación (5.4) de la página 81 tenemos

$$\Lambda A = U.$$

Siendo $\Lambda^{-1} = L$, tenemos

$$A = LU.$$

Acabamos de demostrar el siguiente teorema.

Teorema 5.1. *Si $A \in \mathbb{K}^{n\times n}$ es una matriz directamente Gauss–reducible existen una matriz U triangular superior y una matriz L triangular inferior con unos en la diagonal tales que*

$$A = LU.$$

Nota 5.7.

No es necesario que la matriz A sea invertible para que tenga la descomposición LU.

Ejemplo 5.1.

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

Pero tenemos el resultado siguiente:

Proposición 5.2. *Si A es invertible y directamente Gauss–reducible entonces la descomposición $A = LU$ es única.*

Demostración. Supongamos que $A = LU = L^{'}U^{'}$.

$$L^{'-1}L = U^{'}U^{-1}. \tag{5.8}$$

El lado izquierdo de (5.8) es una matriz triangular inferior con unos en la diagonal y el lado derecho es una matriz triangular superior. Así $U^{'}U^{-1} = I$ es decir $U^{'} = U$ y $L^{'} = L$. □

Teorema 5.2. *Una matriz invertible es directamente Gauss–reducible si y sólo si todas sus submatrices principales son invertibles.*

Demostración. Sea $A \in \mathbb{K}^{n\times n}$ invertible. Supongamos que para todo $k = 1, \ldots, n$

$$\det(A(k)) \neq 0,$$

donde $A(k)$, $k = 1, \ldots, n$, son las submatrices principales de A.

En el proceso de la descomposición LU de la matriz después de la etapa $k-1$ tenemos

$$\begin{matrix} k\text{-}1 & n\text{-}k+1 \end{matrix} \qquad \begin{matrix} k\text{-}1 & n\text{-}k+1 \end{matrix} \qquad \begin{matrix} k\text{-}1 & n\text{-}k+1 \end{matrix}$$

$$\begin{array}{r} k\text{-}1 \\ n\text{-}k+1 \end{array} \left(\begin{array}{c|c} A^{(1)} & A^{(2)} \\ \hline A^{(3)} & A^{(4)} \end{array}\right) = \left(\begin{array}{c|c} L^{(1)} & 0 \\ \hline L^{(3)} & L^{(4)} \end{array}\right) \left(\begin{array}{c|c} U^{(1)} & U^{(2)} \\ \hline 0 & U^{(4)} \end{array}\right) \begin{array}{l} k\text{-}1 \\ n\text{-}k+1 \end{array} \tag{5.9}$$

donde

$$A^{(1)} = A(k-1)$$

y L_1, U_1 son matrices triangulares inferior y superior respectivamente.

Ahora tenemos que mostrar que para pasar a la siguiente etapa en la eliminación de Gauss no hay que intercambiar filas. En otras palabras, mostrar que $U_{11}^{(4)}$ es no nulo.

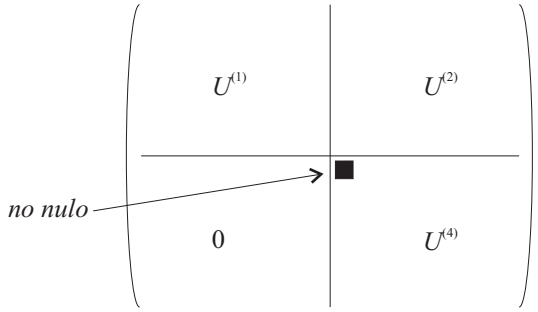

Figura 5.2.

Para eso hacemos una partición de tipo $k, n-k$ en el sistema (5.9), es decir

$$\begin{array}{r} \\ k \\ n\text{-}k \end{array} \begin{array}{c} \begin{array}{cc} k & n\text{-}k \end{array} \\ \left(\begin{array}{c|c} \tilde{A}^{(1)} & \tilde{A}^{(2)} \\ \hline \tilde{A}^{(3)} & \tilde{A}^{(4)} \end{array}\right) \end{array} = \begin{array}{c} \begin{array}{cc} k & n\text{-}k \end{array} \\ \left(\begin{array}{c|c} \tilde{L}^{(1)} & 0 \\ \hline \tilde{L}^{(3)} & \tilde{L}^{(4)} \end{array}\right) \end{array} \begin{array}{c} \begin{array}{cc} k & n\text{-}k \end{array} \\ \left(\begin{array}{c|c} \tilde{U}^{(1)} & \tilde{U}^{(2)} \\ \hline \tilde{U}^{(3)} & \tilde{U}^{(4)} \end{array}\right) \end{array} \tag{5.10}$$

La matriz $\tilde{U}^{(1)}$ sigue siendo triangular superior. Así

$$\det \tilde{U}^{(1)} = \prod_{i=1}^{k} \left(\tilde{U}_{ii}^{(1)}\right)$$

y el producto matricial por bloques de (5.10) da

$$\tilde{L}^{(1)} = \tilde{A}^{(1)} \tilde{U}^{(1)}.$$

Como $\det \tilde{L}^{(1)} = 1$ y $A(k) = \tilde{A}^{(1)}$, decimos que

$$\prod_{i=1}^{n} \tilde{U}_{ii}^{(1)} = \det(A(k)).$$

Como por hipótesis $\det(A(k)) \neq 0$, se tiene

$$\tilde{U}_{ii}^{(1)} \neq 0,$$

para todo $i = 1, \ldots, k$ y en particular $\tilde{U}_{kk}^{(1)} \neq 0$ que es justamente $\tilde{U}_{11}^{(4)}$ de la ecuación (5.9). Este proceso inductivo garantiza la descomposición LU sin intercambiar las filas. El paso de $A(1)$ a $A(2)$ es evidente dado que $\det(A(1)) = A_{11}$.

Inversamente, si existen L triangular inferior con unos en la diagonal y U triangular superior tales que

$$A = LU, \tag{5.11}$$

para todo $k = 1, \ldots, n$,

$$\left(\begin{array}{c|c} A^{(1)} & A^{(2)} \\ \hline A^{(3)} & A^{(4)} \end{array}\right) = \left(\begin{array}{c|c} L^{(1)} & 0 \\ \hline L^{(3)} & L^{(4)} \end{array}\right) \left(\begin{array}{c|c} U^{(1)} & U^{(2)} \\ \hline 0 & U^{(4)} \end{array}\right)$$

(k y $n\text{-}k$)

y

$$\det A^{(1)} = \det L^{(1)} \det U^{(1)} = \prod_{i=1}^{k} U_{ii}^{(1)}.$$

Así

$$\det(A(k)) = \prod_{i=1}^{k} U_{ii}^{(1)} = \prod_{i=1}^{k} U_{ii}.$$

Como U es invertible (porque A lo es) deducimos que $\det(A(k)) \neq 0$. □

Como aplicaciones de este teorema tenemos los resultados siguientes sobre matrices con diagonal estrictamente dominante y matrices hermitianas definitads positivas:

Teorema 5.3. *Toda matriz con diagonal estrictamente dominate es directamente Gauss-reducible*

Demostración. las submatrices principales de una matriz con diagonal estrictamente dominante lo son también entonces son invertibles. □

Ejercicio 5.1. Mostrar que si $A \in \mathbb{K}^{n\times n}$ es con diagonal estrictamente dominante entonces todas la matrices intermedias $A^{(k)}$ de la eliminación de Gauss son también con diagonal estrictamente dominante.

Teorema 5.4. *Toda matriz hermitiana definida positiva A se descompone en la forma $A = LU$ donde L es una matriz triangular inferior con unos en su diagonal y U es una matriz triangular superior con elementos en la diagonal estrictamente positivos. En particular A es directamente Gauss–reducible.*

Demostración. Sea $A \in \mathbb{K}^{n\times n}$ hermitiana definida positiva. Tenemos del teorema (3.3)

$$detA(k) > 0,\ k = 1 \dots n$$

(recordamos que $A(k)$ es la submatriz principal de A de tamaño k).Es decir A es directamente Gauss-reducible. De la ecuación $A = LU$, tenemos para $k = 1 \dots n$

$$\det A(k) = \det L(k) \det U(k) = \det U(k) > 0$$

pusto que L tiene unos en la diagonal.
Dado que para todo k,

$$\det U(k) = \prod_{i=1}^{i=k} U_{ii}$$

deducimos que $U_{ii} > 0$ para todo $k = 1 \dots n$ □

5.5. Descomposición de Cholesky

Sea $A \in Knn$ una matriz hermitiana definida positiva. Escribamos

$$A = LU$$

Sea D la matriz diagonal definida por $D_{ii} = U_{ii},\ i = 1 \dots n$. Sabemos de la sección anterior que los elementos de $D_{ii} > 0$ para todo $i = 1 \dots n$. Así

$$\begin{aligned} A &= LU \\ &= LDD^{-1}U \\ &= LDV \end{aligned}$$

donde $V = D^{-1}U$. Por la construción de D, los elementos diagonales de V son unos y dado que A es hermitiana tenemos

$$A = V^*D^*L^* = V^*DL^*$$

De la unicidad de la descomposición mostrada en la proposición (5.2) deducimos que $V^* = L$ y por supuesto que $L^* = V$, así

$$A = LDL^*.$$

Ahora sea Δ la matriz diagonal definida positiva dada por

$$\Delta^2 = D$$

Podemos ahora escribir A en la formas

$$A = LDL^* = L\Delta^2L^* = L\Delta\Delta^*L^* = (L\Delta)(L\Delta)^*$$

Teorema 5.5. Descomposición de Cholesky
Una matriz $A \in \mathbb{K}^{n\times n}$ es hermitiana definida positiva si y solo si existe una matriz triangular inferior invertible tal que

$$A = LL^* \tag{5.12}$$

Demostración. Por lo anterior la condición es necesaria tomando como matriz triangular inferior $L\Delta$. La condición es suficiente por que si A se descompone en la forma (5.12) entonces A es evidentemente hermitiana además si x es un vector no nulo se tiene

$$\langle Ax, x\rangle = \langle LL^*x, x\rangle = \langle L^*x, L^*x\rangle = \|L^*x\|_2^2 > 0$$

□

5.5.1. Algoritmo de Cholesky

Este algoritmo se usa para resolver sistemas lineales $Ax = b$ cuando la matriz A es hermitiana definida positiva. La idea es encontrar la matriz triagular inferior L de la descomposición de Cholesky.

$$\begin{cases} L_{11} = \sqrt{A_{11}} \\ \text{para } i = 2\ldots n,\ L_{i1} = A_{i1}/L_{11} \\ \text{para } j = 2\ldots n-1,\ L_{i1} = A_{i1}/L_{11} \\ \begin{cases} L_{jj} = \sqrt{A_{jj} - \sum_{k=1}^{j-1} L_{jk}^2} \\ \text{para } i = j+1,\ldots,n\ L_{ij} = (A_{ij} - \sum_{k=1}^{j-1} L_{ik}L_{jk})/L_{jj} \end{cases} \\ L_{nn} = \sqrt{A_{nn} - \sum_{k=1}^{n-1} L_{nk}^2} \end{cases}$$

Capítulo 6

Métodos iterativos de solución de sistemas lineales

Consideremos el sistema lineal

$$Ax = b, \tag{6.1}$$

donde $A \in \mathbb{K}^{n\times n}$, $\det(A) \neq 0$ y $b \in \mathbb{K}^n$.

Los métodos iterativos para resolver sistemas de este tipo consisten en construir una sucesión de vectores $(x^n)_{n\in\mathbb{N}}$ en $\mathbb{K}^n$ que converge a la solución del sistema $A^{-1}b$. Lo que significa que no se busca la solución exacta del sistema sino una aproximación con una precisión deseada.

Nos interesa construir sucesiones $(x^k)_{k\in\mathbb{N}}$ de la forma

$$\begin{aligned} &x^0 \quad \text{dado y} \\ &x^{k+1} = Tx^k + c, \end{aligned} \tag{6.2}$$

donde T es una matriz en $\mathbb{K}^{n\times n}$ y c un vector. Escoger un método iterativo es escoger la matriz T y el vector c.

Definición 6.1.

1. Sea $\overline{x}$ la solución exacta del sistema lineal de la ecuación (6.1). El error e^k cometido en la iteración k en el algoritmo de la ecuación (6.2) se define por
$$e^k = x^k - \overline{x}.$$
2. Llamamos residuo al vector r^k definido por
$$r^k = b - Ax^k.$$

Nota 6.1.

El error e^k mide qué tan cerca está x^k de $\overline{x}$ mientras r^k mide qué tan bien satisface x^k el sistema lineal $Ax = b$.

Consideremos el algoritmo

$$\begin{aligned} &x^0 \quad \text{dado y} \\ &x^{k+1} = Tx^k + c. \end{aligned} \tag{6.3}$$

Si la sucesión x^k converge a la solución del sistema lineal $Ax = b$, es decir si

$$\lim_{k\to\infty} x^k = A^{-1}b,$$

tenemos entonces en la ecuación (6.3)

$$A^{-1}b = TA^{-1}b + c.$$

Así

$$c = (I - T)A^{-1}b. \tag{6.4}$$

La ecuación (6.4) es una condición necesaria para que el algoritmo converja a la solución $\overline{x} = A^{-1}b$.

Definición 6.2. Decimos que el algoritmo de la ecuación (6.3) es consistente si c y T verifican la ecuación (6.4).

Teorema 6.1. *Si el algoritmo de la ecuación (6.3) es consistente entonces $(x^k)_{k\geq 0}$ converge a la solución $\overline{x}$ del sistema $Ax = b$ si y sólo si $\rho(T) < 1$.*

Demostración.

$$x^{k+1} = Tx^k + c.$$

Restando $\overline{x}$ de los dos lados

$$e^{k+1} = Tx^k + c - \overline{x}$$

y por la condición de consistencia

$$\begin{aligned} e^{k+1} &= Tx^k + (I - T)\overline{x} - \overline{x}, \\ e^{k+1} &= T(x^k - \overline{x}), \\ e^{k+1} &= Te^k \end{aligned}$$

para todo $k \geq 0$. Así

$$e^k = T^k e^0,$$

es decir que x^k converge a $\overline{x}$ cuando $k \to \infty$ lo que equivale a decir que $e^k \to 0$ que a su vez es equivalente, según el teorema (4.4) de la página 69, a $\rho(T) < 1$. □

6.1. Construcción de los métodos iterativos

Definición 6.3. Sea $A \in \mathbb{K}^{n\times n}$. Una descomposción regular de A (en Inglés *regular splitting*) es escribir A en la forma $A = M - N$ donde M es una matriz invertible.

Ahora consideremos el sistema lineal en cuestión

$$Ax = b. \tag{6.5}$$

Sea $A = M - N$ una descomposición regular de A. De la ecuación (6.5) podemos escribir

$$Mx = Nx + b,$$

o también

$$x = M^{-1}Nx + M^{-1}b. \tag{6.6}$$

La ecuación (6.6) sugiere el algoritmo

$$\begin{aligned} &x^{k+1} = M^{-1}Nx^k + M^{-1}b, \\ &x^0 \quad \text{dado.} \end{aligned}$$

Por construcción, este algoritmo es consistente y la convergencia a la solución del sistema de la ecuación (6.5) es equivalente a $\rho\left(M^{-1}N\right) < 1$.

6.1.1. Casos de descomposiciones particulares

Consideremos el sistema lineal de la ecuación (6.5) y escribamos A en la forma

$$A = D - E - F,$$

donde D es diagonal, E es triangular inferior estricta y F es triangular superior estricta.

$$A = \begin{pmatrix} \diagdown & & -F \\ & D & \\ -E & & \diagdown \end{pmatrix}.$$

Asumamos que D es invertible.

Método de Jacobi

$$M = D, \qquad N = E + F$$

y el algoritmo es dado en la forma matricial con

$$\begin{aligned} &x^0 \quad \text{dado y} \\ &x^{k+1} = D^{-1}(E + F)x^k + D^{-1}b, \quad k \geq 0, \end{aligned}$$

que podemos escribir componente por componente en la forma

$$\begin{aligned} &x^0 \quad \text{dado,} \\ &x_i^{k+1} = \frac{1}{a_{ii}} \left(-\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}\, x_j^k - b_i \right), \qquad i = 1, \ldots, n. \end{aligned}$$

Método de Gauss–Seidel

Para este método

$$M = D - E \qquad N = F$$

Así el algoritmo en la forma matricial está dado por

$$\begin{aligned} &x^0 \quad \text{dado,} \\ &x^{k+1} = (D-E)^{-1} F x^k + (D-E)^{-1} b, \quad k \geq 0. \end{aligned}$$

Podemos también escribirlo en la forma

$$(D-E)x^{k+1} = Fx^k + b$$

o también

$$Dx^{k+1} = Ex^{k+1} + Fx^k + b.$$

Así

$$x^{k+1} = D^{-1}\left(Ex^{k+1} + Fx^k + b\right), \quad k \geq 0.$$

La forma puntual del algoritmo de Gauss–Seidel es

$$\begin{aligned} &x_i \quad \text{dado,} \\ &x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}\, x_j^k \right), \qquad i = 1, \ldots, n. \end{aligned}$$

Método de relajación (SOR)

Sea ω una parámetro real no nulo. Se definen M y N como

$$\begin{aligned} M &= \frac{1}{\omega}(D - \omega E), \\ N &= \frac{1}{\omega}\big[(1-\omega)D + \omega F\big]. \end{aligned}$$

Denotemos la matriz $M^{-1}N$ por $\mathcal{L}_\omega$. Miremos como es $\mathcal{L}_\omega$.

$$\begin{aligned}
\mathcal{L}_\omega &= \left[\frac{1}{\omega}(D-\omega E)\right]^{-1}\left[\left(\frac{1}{\omega}-1\right)D+F\right],\\
\mathcal{L}_\omega &= \omega(D-\omega E)^{-1}\left[\left(\frac{1}{\omega}-1\right)D+F\right],\\
\mathcal{L}_\omega &= (D-\omega E)^{-1}\big[(1-\omega)D+\omega F\big],\\
\mathcal{L}_\omega &= \big[D\left(I-\omega D^{-1}E\right)\big]^{-1}\big[D\left((1-\omega)I+\omega D^{-1}F\right)\big],\\
\mathcal{L}_\omega &= (I-\omega L)^{-1}D^{-1}D\left[(1-\omega)I+\omega U\right],
\end{aligned}$$

donde $L = D^{-1}E$ y $U = D^{-1}F$. Así

$$\mathcal{L}_\omega = (I-\omega L)^{-1}\big[(1-\omega)I-\omega U\big]. \tag{6.7}$$

La iteración en la forma matricial es

$$\begin{aligned}
&x^0 \quad \text{dado y}\\
&x^{k+1} = \mathcal{L}_\omega x^k + c_\omega, \quad k \geq 0,
\end{aligned}$$

donde $\mathcal{L}_\omega$ está dada por la ecuación (6.7) y

$$c_\omega = \left(\frac{D}{\omega}-E\right)^{-1} b,$$

que podemos también escribir en la forma

$$\begin{aligned}
&x^0 \quad \text{dado y}\\
&(D-\omega E)x^{k+1} = [(1-\omega)D+\omega F]x^k + \omega b.
\end{aligned}$$

Lo que nos da la forma puntual del algoritmo de relajación

$$x^0 \quad \text{dado},$$

$$x_i^{k+1} = \frac{1}{a_{ii}}\left[\omega b_i - \omega\sum_{j=1}^{i-1} a_{ij}\,x_j^{k+1} + a_{ii}\,x_i^k - \omega\sum_{j=i}^{n} a_{ij}\,x_j^k\right], \qquad i = 1,\ldots,n.$$

Nota 6.2.

1. El método de Gauss–Seidel es un caso particular del método de relajación con $\omega = 1$.
2. Las multiplicaciones de las matrices D, E y F por ω o por $(1-\omega)$, que se hacen sólo una vez, son los únicos cálculos adicionales con respecto a lo que se hace en el método de Gauss–Seidel.
3. Cuando $\omega < 1$ se habla de método de subrelajación y cuando $\omega > 1$ de método de sobrerrelajación.

Proposición 6.1. *Si el método SOR converge entonces* $\omega \in]0\ 2[$.

Demostración. Para que el método SOR converja es necesario que $\rho(\mathcal{L}_\omega) < 1$. Calculemos el polinomio característico de $\mathcal{L}_\omega$.

$$P_\lambda(\mathcal{L}_\omega) = \det(\lambda I - \mathcal{L}_\omega) = \det\left[\lambda I - (I - \omega L)^{-1}\big((1-\omega)I - \omega U\big)\right].$$

Como $\det(I - \omega L) = 1$ entonces

$$P_\lambda(\mathcal{L}_\omega) = \det\left[\lambda(I - \omega L) - \big((1-\omega)I - \omega U\big)\right].$$

$$|\det(\mathcal{L}_\omega)| = |P_0(\mathcal{L}_\omega)| = |(1-\omega)^n| = |1-w|^n$$

pero al mismo tiempo

$$|\det(\mathcal{L}_\omega)| = \left|\prod_{i=1}^{n} \lambda_i\right| = \prod_{i=1}^{n} |\lambda_i|$$

donde $\lambda_1, \ldots, \lambda_n$ son los valores propios de $\mathcal{L}_\omega$. Así

$$|1-\omega|^n = \prod_{i=1}^{n} |\lambda_i| < \rho(\mathcal{L}_\omega)^n < 1.$$

Esto significa que si SOR converge entonces $|1-\omega|^n < 1$ es decir $0 < \omega < 2$. □

6.2. Convergencia de los métodos iterativos

Dado que los métodos iterativos de tipo

$$\begin{aligned} &x^0 \quad \text{dado y} \\ &x^{k+1} = (M^{-1}N)x^k + M^{-1}b, \qquad k \geq 0, \end{aligned}$$

no convergen a menos que $\rho(M^{-1}N) < 1$, la convergencia de los métodos de Jacobi y de Relajación (SOR) puede fallar para algunas matrices A. Estudiaremos en las secciones siguientes la convergencia de estos algoritmos para matrices particulares.

6.2.1. Matrices con diagonal estrictamente dominante

Teorema 6.2. *Si* $A \in \mathbb{K}^{n\times n}$ *tiene diagonal estrictamente dominante*

$$\rho\left(D^{-1}(D-A)\right) < 1.$$

Nota 6.3.

Deducimos de este teorema que el algoritmo de Jacobi converge dado que su matriz de iteración es justamente $D^{-1}(D-A) = I - D^{-1}A$.

Demostración. Sabemos por la dominancia estricta de la diagonal de A que $\forall i = 1, \ldots, n$

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}} |a_{ij}|\,,$$

es decir

$$\sum_{\substack{j=1 \\ j \neq i}} \left| \frac{a_{ij}}{a_{ii}} \right| < 1, \qquad i = 1, \ldots, n.$$

Pero

$$\frac{a_{ij}}{a_{ii}} = (D^{-1}A)_{ij}, \qquad 1 \leq i, j \leq n$$

y

$$(D^{-1}A)_{ii} = 1, \qquad i = 1, \ldots, n.$$

Así

$$\left\| I - D^{-1}A \right\|_\infty = \max_i \sum_{\substack{j=1 \\ j \neq i}} \left| \frac{a_{ij}}{a_{ii}} \right| < 1.$$

Como

$$\rho\left(I - D^{-1}A\right) \leq \left\| I - D^{-1}A \right\|_\infty,$$

deducimos que el método de Jacobi converge. □

Teorema 6.3. *Sea $0 < \omega \leq 1$. Si A es una matriz con diagonal estrictamente dominante entonces el método de relajación (SOR) converge.*

Demostración. Mostremos que $\rho(\mathcal{L}_\omega) < 1$ donde $\mathcal{L}_\omega = (I - \omega L)^{-1}\big[(1-\omega)I - \omega U\big]$. El polinomio característico de $\mathcal{L}_\omega$ es

$$P_\lambda(\mathcal{L}_\omega) = \det(\lambda I - \mathcal{L}_\omega) = \det\big(\lambda(I - \omega L) - (1-\omega)I - \omega U\big),$$

porque $\det(I - \omega L) = 1$. Así

$$P_\lambda(\mathcal{L}_\omega) = \det\big((\lambda + \omega - 1)I - \omega\lambda L - \omega U\big) = (\lambda + \omega - 1)^n \det(I - aL - bU),$$

donde

$$a = \frac{\lambda\omega}{\lambda + \omega - 1} \qquad b = \frac{\omega}{\lambda + \omega - 1}.$$

Procedamos por contradicción. Sea $\lambda \in \text{esp}(\mathcal{L}_\omega)$ y supongamos que $|\lambda| \geq 1$. en primero no se puede tener $\lambda + \omega - 1 = 0$ porque tendríamos $|\lambda| = |\omega - 1| < 1$.

Mostremos que $\det(I - aL - bU) = 0$ también daría contradicción si $|\lambda| \geq 1$. Sea $\lambda = re^{i\theta}$,

$$\begin{aligned} a &= \frac{\lambda\omega}{\lambda + \omega - 1}, \\ &= \frac{\lambda\omega(\overline{\lambda} + \omega - 1)}{(r\cos\theta + \omega - 1)^2 + r^2 \operatorname{sen}^2\theta}, \\ &= r\omega \frac{r + \cos\theta(\omega - 1) + i(\omega - 1)\operatorname{sen}\theta}{r^2 - 2r\cos\theta(1 - \omega) + (\omega - 1)^2}, \end{aligned}$$

y

$$\begin{aligned} |a|^2 &= r^2\omega^2 \frac{r^2 + \cos^2\theta(\omega - 1)^2 + (\omega - 1)^2 \operatorname{sen}^2\theta + 2r\cos\theta(\omega - 1)}{\left[r^2 - 2r\cos\theta(1 - \omega) + (\omega - 1)^2\right]^2}, \\ &= \frac{r^2\omega^2}{[r^2 - 2r\cos\theta(1 - \omega) + (\omega - 1)^2]}. \end{aligned}$$

Pero

$$r^2 - 2r(1 - \omega)\cos\theta + (\omega - 1)^2 \geq r^2 - 2r(1 - \omega) + (\omega - 1)^2 = (r - 1 + \omega)^2,$$

y dado que

$$r - 1 + \omega > r\omega$$

pues $r \geq 1$ y $0 < \omega \leq 1$, se tiene entonces

$$|a| \leq \frac{r\omega}{(r - 1 + \omega)} \leq 1.$$

Como se supone que $|\lambda| \geq 1$ y que $|a| = |\lambda|\ |b|$, entonces $|b| \leq |a| \leq 1$.

Deducimos del ejercicio 3.6 que la matriz $(I - aL - bU)$ también tiene diagonal estrictamente dominante, es decir que es invertible, lo que contradice $\det(I - aL - bU) = 0$. Así $\rho(\mathcal{L}_\omega) < 1$. □

6.2.2. Matrices hermitianas definidas positivas

Teorema 6.4. *Sean $A \in \mathbb{K}^{n\times n}$ hermitiana definida positiva y $A = M - N$ un splitting regular. Entonces el algoritmo converge si la matriz $Q = M + M^* - A$ es definida positiva.*

Demostración. Supongamos que Q es definida positiva (que es hermitiana es evidente). Mostremos que $\rho\left(M^{-1}N\right) < 1$. Denotemos $J = M^{-1}N$, es decir $J = I - M^{-1}A$. Sea $A^{1/2}$ la matriz raíz cuadrada de la matriz A y sea

$$K = A^{1/2}JA^{-1/2} = A^{1/2}\left(I - M^{-1}A\right)A^{-1/2} = I - A^{1/2}M^{-1}A^{1/2}.$$

Ahora

$$KK^* = \left(I - A^{1/2}M^{-1}A^{1/2}\right)\left(I - A^{1/2}M^{*-1}A^{1/2}\right),$$
$$KK^* = I - A^{1/2}M^{-1}A^{1/2} - A^{1/2}M^{*-1}A^{1/2} + A^{1/2}M^{-1}AM^{*-1}A^{1/2},$$
$$KK^* = I - A^{1/2}\left(M^{-1} + M^{*-1} - M^{-1}AM^{*-1}\right)A^{1/2},$$
$$KK^* = I - A^{1/2}M^{-1}\left(M^* + M - A\right)M^{*-1}A^{1/2},$$
$$KK^* = I - A^{1/2}M^{-1}QM^{*-1}A^{1/2}.$$

Pero $Q^{1/2}$ existe. Así

$$KK^* = I - A^{1/2}M^{-1}Q^{1/2}Q^{1/2}M^{*-1}A^{1/2},$$
$$KK^* = I - \left(A^{1/2}M^{-1}Q^{1/2}\right)\left(A^{1/2}M^{-1}Q^{1/2}\right)^*.$$

Si denotamos $H = A^{1/2}M^{-1}Q^{1/2}$ tenemos

$$KK^* = I - HH^*.$$

Las matrices KK^* y HH^* son hermitianas definidas positivas entonces sus valores propios son reales positivos. Pero

$$\operatorname{esp}(KK^*) = \left\{1 - \lambda;\ \lambda \in \operatorname{esp}(HH^*)\right\}.$$

Así $1-\lambda > 0 \implies 1 > \lambda$ y $\lambda > 0$ lo que significa que $\rho(KK^*) < 1$. Así $\|K\|_2 < 1$ y de hecho $\rho(K) < 1$. □

Corolario 6.1. *Sea $A \in \mathbb{K}^{n\times n}$ hermitiana definida positiva. Entonces*

1. *Si $2D - A$ es hermitiana definida positiva el método de Jacobi converge.*
2. *Para todo $\omega \in]0, 2[$ el método de relajación converge.*

Demostración. Miremos bajo qué condiciones la matriz $M + M^* - A$ es hermitiana definida positiva para ambos casos.

Para el caso de Jacobi $M = D$. Como A es hermitiana definida positiva deducimos que D también lo es. Así

$$M + M^* - A = 2D - A.$$

Para el caso de relajación

$$M = \left(\frac{D}{\omega} - E\right), \qquad D = D^*.$$

Así

$$M + M^* - A = \left(\frac{D}{\omega} - E\right) + \left(\frac{D}{\omega} - E^*\right) - D + E + E^* = \frac{2D}{\omega} - D = \frac{2-\omega}{\omega}D$$

que es una matriz hermitiana definida positiva. □

6.2.3. Búsqueda del parámetro óptimo del SOR para matrices tridiagonales por bloques

Notación

Sea $A \in \mathbb{K}^{n\times n}$ con la forma por bloques siguiente

$$A = \begin{pmatrix} A_1 & B_1 & & & \\ C_1 & A_2 & B_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & B_{p-1} \\ & & & C_{p-1} & A_p \end{pmatrix}, \tag{6.8}$$

donde A_i, $i = 1, \ldots, p$, son matrices cuadradas.

En esta sección vamos a estudiar la convergencia de los métodos de Jacobi y SOR para este tipo de matrices. La elección de estas matrices no es fortuita pues el método de diferencias finitas las genera.

Proposición 6.2. *Sea $A \in \mathbb{K}^{n\times n}$ con la forma de la ecuación (6.8). Denotemos para un $t \in \mathbb{K}$, $t \neq 0$*

$$A_t = \begin{pmatrix} A_1 & t^{-1}B_1 & & & \\ tC_1 & A_2 & t^{-1}B_2 & & \\ & tC_2 & \ddots & \ddots & \\ & & \ddots & \ddots & t^{-1}B_{p-1} \\ & & & tC_{p-1} & A_p \end{pmatrix}.$$

Tenemos $\det(A_t) = \det(A)$.

Demostración. Es sencillo verificar que

$$A_t = Q_t A Q_t^{-1},$$

donde

$$Q_t = \begin{pmatrix} tI_1 & & & \\ & t^2 I_2 & & \\ & & \ddots & \\ & & & t^p I_p \end{pmatrix}$$

con I_j la matriz identidad del mismo tamaño que A_j, $j = 1, \ldots, p$. □

Teorema 6.5. *Sea $A \in \mathbb{K}^{n\times n}$ tridiagonal por bloques. Tenemos*

$$\rho(\mathcal{L}_1) = \rho(J)^2, \tag{6.9}$$

donde $\mathcal{L}_1$ y J son las matrices de Gauss–Seidel y de Jacobi respectivamente asociadas a A.

Nota 6.4.

La ecuación (6.9) significa que los métodos de Jacobi y de Gauss–Seidel convergen o divergen simultáneamente.

Demostración. Descompongamos $A = D-E-F$. Los polinomios característicos de $\mathcal{L}_1$ y J son respectivamente

$$P_1(\lambda) = \det\left(\lambda I - (D-E)^{-1}F\right),$$
$$P_J(\lambda) = \det\left(\lambda I - D^{-1}(E+F)\right).$$

Tenemos

$$P_J(\lambda) = \det D^{-1}\ \det(\lambda D - E - F),$$
$$P_J(\lambda) = \det D^{-1}\ \det(\lambda D + E + F), \qquad t=-1 \text{ en la proposición (6.2)},$$
$$P_J(\lambda) = \left(\det D^{-1}\right)(-1)^n\ \det(-\lambda D - E - F),$$
$$P_J(\lambda) = \left(\det D^{-1}\right)(-1)^n\ (\det D)\det\left(-\lambda I - D^{-1}(E+F)\right)$$
$$P_J(\lambda) = (-1)^n\ P_J(-\lambda).$$

Es decir, si λ es valor propio de J, $-\lambda$ también lo es.

$$P_1(\lambda) = \det(D-E)^{-1}\ \det(\lambda D - \lambda E - F).$$

Sea $\lambda^{1/2}$ una raíz cuadrada de λ.

$$P_1(\lambda) = \det(D-E)^{-1}\ \det\left(\lambda^{1/2}\left[\lambda^{1/2}D - \lambda^{1/2}E - \lambda^{-1/2}F\right]\right),$$
$$P_1(\lambda) = \det(D-E)^{-1}\ \lambda^{n/2}\det\left(\lambda^{1/2}D - \lambda^{1/2}E - \lambda^{-1/2}F\right),$$
$$P_1(\lambda) = \det(D-E)^{-1}\ \lambda^{n/2}\det\left(\lambda^{1/2}D - E - F\right), \qquad t=\lambda^{1/2} \text{ en la proposición (6.2)},$$
$$P_1(\lambda) = \det(D-E)^{-1}\ \lambda^{n/2}\det(D)\det\left(\lambda^{1/2}I - D^{-1}(E+F)\right),$$
$$P_1(\lambda) = \lambda^{n/2}\det\left(I - D^{-1}E\right)^{-1}\ P_J\left(\lambda^{1/2}\right),$$
$$P_1(\lambda) = \lambda^{n/2}\ P_J\left(\lambda^{1/2}\right).$$

Así si

$$\lambda \neq 0 \quad \text{y} \quad \lambda \in \text{esp}(\mathcal{L}_1) \implies \lambda \in \text{esp}(J) \quad \text{y} \quad -\lambda \in \text{esp}(J).$$

Si

$$\lambda \in \text{esp}(J) \implies -\lambda \in \text{esp}(J) \implies \lambda^2 \in \text{esp}(\mathcal{L}_1).$$

□

Capítulo 7

Métodos basados en la optimización para sistemas lineales: Métodos del gradiente

Introducción

Sean K un subconjunto de $\mathbb{R}^n$ y $J : K \to \mathbb{R}$ una función. El problema de optimización asociado a la función J y el conjunto K es encontrar $\overline{x} \in K$ tal que

$$J(\overline{x}) = \inf_{x \in K} J(x) = \min_{x \in K} J(x).$$

Si tal $\overline{x}$ existe, se le dice solución del problema de optimización

$$\min_{x \in K} J(x).$$

Nota 7.1.
Si $J : K \to \mathbb{R}$ es una función y $\alpha > 0, \beta$ son escalares entonces los problemas

$$\min_{x \in K} J(x) \qquad \text{y} \qquad \min_{x \in K} (\alpha J(x) + \beta)$$

tienen las mismas soluciones.

En este capítulo nos interesamos en resolver sistemas lineales usando técnicas de optimización. Para eso, dado un sistema lineal

$$Ax = b, \tag{7.1}$$

se le asocia un problema de optimización

$$\min_{x \in K} J(x) \tag{7.2}$$

de manera que $\overline{x}$ es solución de (7.1) si y sólo si $\overline{x}$ es solución de (7.2).

Generalmente, se busca una aproximación de la solución $\overline{x}$ (supuestamente única) del problema (7.2). Este método se resume en construir una sucesión

$$(x^n)_{n\in\mathbb{N}} \subset \mathbb{R}^n$$

tal que

$$J(x^n) \quad \text{converja a} \quad J(\overline{x})$$

esperando que x^n converja a $\overline{x}$ cuando $n \to \infty$.

Definición 7.1. Se dice que $(x^n)_{n\in\mathbb{N}}$ es una sucesión minimizante del problema (7.2) si $J(x^n) \to \text{ínf}_{x\in K}\, J(x)$.

7.1. Construcción de la Función J

Sean $A \in \mathbb{R}^{n\times n}$ una matriz simétrica definida positiva y $b \in \mathbb{R}^n$. Se define la función

$$\begin{aligned} J : \mathbb{R}^n &\to \mathbb{R}, \\ x &\mapsto \frac{1}{2}\,\langle Ax, x\rangle - \langle b, x\rangle\,. \end{aligned}$$

Teorema 7.1. *El problema*

$$\min_{x\in\mathbb{R}^n} J(x)$$

tiene una solución $\overline{x}$ única y es justamente la solución del sistema $Ax = b$.

Demostración. Para $x \in \mathbb{R}^n$, sea

$$e(x) = x - \overline{x},$$

donde

$$\overline{x} = A^{-1}b.$$

Definimos

$$J_1(x) = \big\langle A\, e(x), e(x)\big\rangle.$$

Dado que A es simétrica definida positiva tenemos

i) $J_1(x) \geq 0$.

ii) $J(x) = 0$ si y sólo si $e(x) = 0$, es decir $x = \overline{x}$, lo que significa que $\overline{x}$ es la solución única del problema

$$\min_{x\in\mathbb{R}^n} J_1(x).$$

Para terminar la demostración falta mostrar que los problemas

$$\min_{x\in\mathbb{R}^n} J_1(x) \qquad \text{y} \qquad \min_{x\in\mathbb{R}^n} J(x)$$

tienen las mismas soluciones. Desarrollemos $J_1(x)$:

$$\begin{aligned}
J_1(x) &= \big\langle A\,e(x), e(x)\big\rangle,\\
&= \big\langle Ax - A\overline{x}\,,\ x-\overline{x}\,\big\rangle,\\
&= \big\langle Ax - b\,,\ x-\overline{x}\,\big\rangle,\\
&= \big\langle Ax, x\big\rangle - \big\langle Ax, \overline{x}\big\rangle - \big\langle b, x\big\rangle + \big\langle b, \overline{x}\big\rangle,\\
&= \big\langle Ax, x\big\rangle - \big\langle x, A\overline{x}\big\rangle - \big\langle b, x\big\rangle + \big\langle b, \overline{x}\big\rangle,\\
&= \big\langle Ax, x\big\rangle - 2\big\langle b, x\big\rangle + \big\langle b, \overline{x}\big\rangle\\
&= 2J(x) + \big\langle b, \overline{x}\big\rangle.
\end{aligned}$$

Dado que $\big\langle b, \overline{x}\big\rangle$ es una constante y teniendo en cuenta la nota (7.1) de la página 103, deducimos el resultado. □

7.2. Planteamiento general del método

La sucesión $(x^n)_{n\in\mathbb{N}}$ que minimiza el problema

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\,\langle Ax, x\rangle - \langle b, x\rangle$$

se construye en la forma siguiente:

$$x^{n+1} = x^n + \alpha_n p^n,$$

donde α_n y p^n son respectivamente escalar y vector que van a ser elegidos de forma que x^n converja a $\overline{x}$.

Definición 7.2. p^n se llama dirección de descenso.

7.2.1. Elección de α_n

Supongamos que p^n ha sido elegido. Ahora vamos a hallar un α_n adecuado de manera que

$$J(x^{n+1}) < J(x^n).$$

Para eso:

$$\begin{aligned}
J(x^{n+1}) &= \frac{1}{2}\,\big\langle A(x^n + \alpha_n p^n)\,,\ x^n + \alpha_n p^n\big\rangle - \big\langle b\,,\ x^n + \alpha_n p^n\big\rangle,\\
&= \frac{1}{2}\,\big\langle Ax^n, x^n\big\rangle - \big\langle b, x^n\big\rangle + \frac{1}{2}\,\alpha_n\big\langle Ap^n, x^n\big\rangle\\
&\qquad + \frac{1}{2}\,\alpha_n\big\langle Ax^n, p^n\big\rangle + \frac{1}{2}\,\alpha_n^2\big\langle Ap^n, p^n\big\rangle - \alpha_n\big\langle b, p^n\big\rangle,\\
&= J(x^n) + \frac{1}{2}\,\alpha_n^2\big\langle Ap^n, p^n\big\rangle + \alpha_n\big\langle Ax^n - b, p^n\big\rangle. \qquad (7.3)
\end{aligned}$$

La expresión de $J(x^{n+1})$ en la ecuación (7.3) es un polinomio de grado 2 en α_n así que alcanza su mínimo cuando

$$\alpha_n = \frac{\langle b - Ax^n,\, p^n \rangle}{\langle Ap^n, p^n \rangle}. \tag{7.4}$$

Con este valor de α_n,

$$J(x^{n+1}) = J(x^n) - \frac{1}{2}\frac{\langle b - Ax^n,\, p^n \rangle^2}{\langle Ap^n, p^n \rangle}. \tag{7.5}$$

Definición 7.3. El valor dado en (7.4) se llama óptimo local de α.

Notamos que:

1. El vector de descenso no puede ser nulo o la sucesión será estacionaria.
2. A menos que $\langle p^n,\, b - Ax^n \rangle = 0$, $J(x^{n+1}) < J(x^n)$.
3. En la elección de p^n, evitaremos que éste sea ortogonal a $b - Ax^n$.

Recordando que el residuo r^k está dado por $r^k = b - Ax^k$, la ecuación (7.5) se vuelve

$$J(x^{n+1}) = J(x^n) - \frac{1}{2}\frac{\langle r^n, p^n \rangle^2}{\langle Ap^n, p^n \rangle}. \tag{7.6}$$

Averiguamos ahora bajo que condicion (sobre p^n) la sucesion $(x^n)_{n\in\mathbb{N}}$ converge a $\overline{x}$ Multiplicando (7.6) por 2 y sumando $\langle b, \overline{x} \rangle$ se obtiene

$$J_1(x^{n+1}) = J_1(x^n) - \frac{\langle r^n, p^n \rangle^2}{\langle Ap^n, p^n \rangle}. \tag{7.7}$$

Sabiendo que $J(x^n)$ tiende a cero si y sólo si x^n converge a $\overline{x}$, la solución del sistema $Ax = b$, transformamos la ecuación (7.7) en la forma

$$J_1(x^{n+1}) = k_n J_1(x^n), \tag{7.8}$$

donde k_n es un escalar que nos ayudará a elegir la dirección p^n.
De la ecuación (7.8) tenemos

$$k_n = 1 - \frac{\langle r^n, p^n \rangle^2}{\langle Ap^n, p^n \rangle}\frac{1}{J_1(x^n)}.$$

Si podemos mostrar que existe un $0 < k < 1$ tal que

$$\forall n, \qquad 0 < k_n < k,$$

tendremos

$$J_1(x^n) \leq k^n J_1(x^0),$$

lo que garantiza que

$$x^n \to \overline{x}.$$

Miremos de cerca cómo es k_n.

$$k_n = 1 - \frac{1}{J_1(x^n)} \frac{\langle r^n, p^n \rangle^2}{\langle Ap^n, p^n \rangle},$$

pero $J_1(x^n)$ se puede tambien escribir en la forma

$$J_1(x^n) = \left\langle r^n, A^{-1} r^n \right\rangle.$$

Recordando que los valores propios de A^{-1} son los inversos de los de A y usando el cociente de Rayleigh obtenemos

$$J_1(x) \leq \frac{1}{\lambda_{min}} \left\| r^n \right\|^2,$$

o también

$$\frac{1}{J_1(x^n)} \geq \frac{1}{\left\| r^n \right\|^2} \lambda_{min}. \tag{7.9}$$

De la misma manera

$$\langle Ap^n, p^n \rangle \leq \lambda_{max} \left\| p^n \right\|^2,$$

o también

$$\frac{1}{\langle Ap^n, p^n \rangle} \geq \frac{1}{\lambda_{max}} \frac{1}{\left\| p^n \right\|^2}. \tag{7.10}$$

Multiplicando (7.9) y (7.10) tenemos

$$\frac{1}{J_1(x^n)} \frac{1}{\langle Ap^n, p^n \rangle} \geq \frac{\lambda_{min}}{\lambda_{max}} \frac{1}{\left\| p^n \right\|^2 \left\| r^n \right\|^2}.$$

Así

$$1 - \frac{1}{J_1(x^n)} \frac{1}{\langle Ap^n, p^n \rangle} \langle r^n, p^n \rangle^2 \leq 1 - \frac{\lambda_{min}}{\lambda_{max}} \frac{\langle r^n, p^n \rangle^2}{\left\| r^n \right\|^2 \left\| p^n \right\|^2}$$

y

$$k_n \leq 1 - \frac{\lambda_{min}}{\lambda_{max}} \frac{\langle r^n, p^n \rangle^2}{\left\| r^n \right\|^2 \left\| p^n \right\|^2}. \tag{7.11}$$

Denotamos

$$\delta_n = \frac{\langle r^n, p^n \rangle^2}{\left\| r^n \right\|^2 \left\| p^n \right\|^2}.$$

Observemos que por la desigualdad de Cauchy-Schwartz tenemos

$$\delta_n \leq 1$$

Teniendo en cuenta que

$$\frac{\lambda_{min}}{\lambda_{max}} = \frac{1}{K(A)} \leq 1,$$

la ecuación (7.11) se vuelve entonces

$$k_n \leq 1 - \frac{\delta_n}{K(A)}.$$

Recordemos que queremos encontrar un real $0 < k < 1$ tal que

$$k_n < k < 1.$$

es decir

$$1 - \frac{\delta_n}{K(A)} < k,$$
$$\delta_n > (-k+1)K(A).$$

Entonces la única condición para que $J_1(x^n)$ tienda a cero es que exista una constante $\beta > 0$ tal que

$$\delta_n > \beta, \qquad \forall n \in \mathbb{N}.$$

En este caso se toma

$$k = 1 - \frac{\beta}{K(A)} < 1$$

Recopilemos estos resultados en el siguiente teorema:

Teorema 7.2. *Si la dirección de descenso p^n es elegida de manera que*

$$\forall n, \qquad \frac{\langle r^n, p^n \rangle^2}{\|r^n\|^2 \|p^n\|^2} > \beta > 0, \tag{7.12}$$

donde β es una real arbitrario fijo, entonces el método

$$\begin{cases} x^{k+1} = x^k + \alpha_k p^k \text{ con } \alpha_k = \frac{\langle r^k, p^k \rangle}{\langle Ap^k, p^k \rangle}, \\ x^0 \text{ arbitrario.} \end{cases}$$

converge a la solución del sistema $Ax = b$.

7.2.2. Elección de direcciones de descenso

Método del gradiente

Es evidente en la ecuación (7.12) que la elección más natural de p^n es

$$p^n = r^n.$$

Así

$$\frac{\langle r^n, p^n \rangle^2}{\|r^n\|^2 \|p^n\|^2} = 1.$$

Este método se llama método del gradiente y su iteración está dada por

$$x^0 \text{ arbitrario y}$$
$$\text{para } k \geq 0, \quad x^{k+1} = x^k + \frac{\left\|r^k\right\|^2}{\langle Ar^k, r^k\rangle} r^k.$$

El método del gradiente es muy fácil de implementar pero no es muy práctico por requerir, a veces, muchas iteraciones para alcanzar la precisión deseada. Por eso vamos a ver otro método en que se asegura que el número de iteraciones no puede exceder el tamaño del sistema lineal. Lo que significa de hecho, y es la parte sorprendente de este algoritmo, que en la construcción del método, esperando un método iterativo clásico, se llega a la solución exacta en un número finito de iteraciones. Este método se llama método del gradiente conjugado.

Métodos de los descensos A-conjugados

Definición 7.4. Sea A una matriz simétrica definida positiva. Dos vectores x, y se dicen A-conjugados sí

$$\langle Ax, y\rangle = 0.$$

Nota 7.2.

En otras palabras, x, y son A-conjugados si son ortogonales con respecto al producto punto

$$\langle x, y\rangle_A = \langle Ax, y\rangle .$$

Veremos que si los vectores de descenso son elegidos de manera que sean todos A-conjugados entre si, es decir

$$\forall n \in \mathbb{N}, \qquad \left\langle Ap^n, p^k\right\rangle = 0, \qquad k = 0, \ldots, n-1,$$

entonces el algoritmo

$$x^{k+1} = x^k + \alpha_k p^k$$

converge en a lo sumo n iteraciones.

Nota 7.3.

De todas maneras en $\mathbb{R}^n$ no se pueden tener más que n vectores A-conjugados porque la A-conjugación implica independencia lineal.

Teorema 7.3. *Si las direcciones de descenso p^i, $i = 1, \ldots, n$, son A-conjugadas entonces*

$$r^n = 0,$$

es decir $Ax^n = b$ lo que significa que la sucesión converge en a lo sumo n iteraciones.

Demostración. Tenemos

$$x^{n+1} = x^n + \alpha_n p^n, \qquad n \geq 0 \tag{7.13}$$

y

$$\alpha_n = \frac{\langle r^n, p^n \rangle}{\langle Ap^n, p^n \rangle}.$$

Dado que los p^i forman una base de $\mathbb{R}^n$, es suficiente mostrar que

$$\left\langle r^n, p^k \right\rangle = 0, \qquad k = 0, \ldots, n-1.$$

Tenemos

$$x^n = x^{n-1} + \alpha_{n-1} p^{n-1} = x^{n-2} + \alpha_{n-2} p^{n-2} + \alpha_{n-1} p^{n-1} = \cdots = x^0 + \alpha_0 p^0 + \cdots + \alpha_{n-1} p^{n-1}.$$

Así

$$Ax^n = Ax^0 + \sum_{i=0}^{n-1} \alpha_i A p^i$$

$$b - Ax^n = b - Ax^0 - \sum_{i=0}^{n-1} \alpha_i A p^i,$$

es decir

$$r^n = r^0 - \sum_{i=0}^{n-1} \alpha_i A p^i.$$

Sea k, $0 \leq k \leq n-1$.

$$\left\langle r^n, p^k \right\rangle = \left\langle r^0, p^k \right\rangle - \sum_{i=0}^{n-1} \alpha_i \left\langle Ap^i, p^k \right\rangle$$

y por la A-conjugación

$$\left\langle r^n, p^k \right\rangle = \left\langle r^0, p^k \right\rangle - \alpha_k \left\langle Ap^k, p^k \right\rangle.$$

Pero

$$\alpha_k = \frac{\left\langle r^k, p^k \right\rangle}{\left\langle Ap^k, p^k \right\rangle}.$$

Así

$$\begin{aligned}
\left\langle r^n, p^k \right\rangle &= \left\langle r^0, p^k \right\rangle - \left\langle r^k, p^k \right\rangle, \\
&= \left\langle r^0 - r^k, p^k \right\rangle, \\
&= \left\langle b - Ax^0 - b + Ax^k, p^k \right\rangle, \\
&= \left\langle Ax^k - Ax^0, p^k \right\rangle, \\
&= \left\langle A\left(x^k - x^0\right), p^k \right\rangle.
\end{aligned}$$

Ahora

$$x^k - x^0 = x^k - x^{k-1} + x^{k-1} - x^{k-2} + x^{k-2} + \cdots + x^1 - x^0 = \sum_{i=0}^{k-1} x^{i+1} - x^i.$$

Del algoritmo (7.13)

$$x^{i+1} - x^i = \alpha_i p^i.$$

Así

$$\left\langle r^n, p^k \right\rangle = \left\langle A\left(x^k - x^0\right), p^k \right\rangle = \sum_{i=0}^{k-1} \alpha_i \left\langle Ap^i, p^k \right\rangle = 0 \qquad \text{por } A\text{-conjugación.}$$

□

Ahora nuestra tarea es construir una sucesión de direcciones de descenso $p^0, \ldots, p^n$, A-conjugadas.

Un método intuitivo es el siguiente: tomar $B = \{e_1, \ldots, e_n\}$ una base arbitraria de $\mathbb{R}^n$ y, utilizando el proceso de ortogonalización de Gram–Schmidt con respecto al producto escalar $\langle \cdot, \cdot \rangle_A = x^t A y$, construir una base de $\mathbb{R}^n$

$$p^i, \qquad i = 1, \ldots, n-1,$$

tal que

$$\left\langle p^i, Ap^j \right\rangle = 0, \qquad \forall i \neq j.$$

Con esta elección de p^i el algoritmo es:

$$\begin{cases} x^0 \text{ arbitrario y} \\ \text{para } k \leq n, \text{ hacer el ciclo:} \\ \quad r^k = b - Ax^k, \\ \quad \alpha_k = \dfrac{\left\langle p^k, r^k \right\rangle}{\left\langle Ap^k, p^k \right\rangle}, \\ \quad x^{k+1} = x^k + \alpha_k p^k. \end{cases} \tag{7.14}$$

Ejemplo 7.1.

$$\begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{pmatrix},$$

$$b^t = (25/12, 77/60, 19/20, 319/420).$$

La solución exacta del sistema $Ax = b$ es

$$x^t = (1, 1, 1, 1).$$

El proceso de ortogonalización de Gram-Schmidt de la base canónica de $\mathbb{R}^4$ respecto al producto escalar $\langle x, Ay\rangle$ da la base

$$\begin{aligned} p_0^t &= (1,0,0,0),\\ p_1^t &= (-1/2,1,0,0),\\ p_2^t &= (1/6,-1,1,0),\\ p_3^t &= (-1/20,3/5,-3/2,1). \end{aligned}$$

Del algoritmo (7.14), tomando $x^0 = (0,0,0,0)$, los resultados de las iteraciones son

$$\begin{aligned} x^1 &= (25/12,0,0,0)^t,\\ x^2 &= (19/30,29/10,0,0)^t,\\ x^3 &= (21/20,2/5,5/2,0)^t \end{aligned}$$

y la cuarta iteración da la solución exacta

$$x^4 = (1,1,1,1)^t$$

como se esperaba.

Este método es supremamente costoso en número de operaciones debido al proceso de ortogonalización de Gram-Schmidt y en espacio de almacenamiento dado que en el proceso de calcular un vector en un paso dado se necesitan todos los vectores anteriores.

Método del gradiente conjugado

En este método los p^i se eligen de manera inductiva en cada paso. Explícitamente, se escoge

$$p^0 = r^0$$

y

$$p^k = r^k + \beta_k p^{k-1}. \tag{7.15}$$

Los escalares β_k se eligen de manera que las direcciones de descenso p^i, $i \geq 0$, sean A-conjugadas.

Proposición 7.1. *Para todo $k \geq 0$*

$$\left\langle Ap^{k+1}, p^k \right\rangle = 0$$

si y sólo si

$$\beta_{k+1} = -\frac{\left\langle r^{k+1}, Ap^k \right\rangle}{\langle Ap^k, p^k \rangle}. \tag{7.16}$$

Demostración. Multiplicando la relación

$$p^{k+1} = r^{k+1} + \beta_{k+1} p^k$$

por A y tomando el producto escalar con p^k, resulta

$$\left\langle Ap^{k+1}, p^k \right\rangle = \left\langle Ar^{k+1}, p^k \right\rangle + \beta_{k+1} \left\langle Ap^k, p^k \right\rangle .$$

Si $\left\langle Ap^{k+1}, p^k \right\rangle = 0$ tendremos

$$\beta_{k+1} = -\frac{\left\langle Ar^{k+1}, p^k \right\rangle}{\left\langle Ap^k, p^k \right\rangle}$$

y viceversa. □

Ejercicio 7.1. Mostrar que para este valor de β_k

$$\left\langle r^{k+1}, r^k \right\rangle = 0.$$

En el resto de este capítulo β_k siempre tomará el valor dado por la ecuación (7.16).

Nota 7.4.
A priori la elección de este valor para β_k no garantiza que las direcciones p^i, $i \geq 0$ sean A-conjugadas pero sí asegura que dos direcciones sucesivas son A-conjugadas.

Vamos a mostrar que esta elección de β_k es suficiente para la A-conjugación de los p^i, $i \geq 0$.

Nota 7.5.
Tenemos

$$x^{n+1} = x^n + \alpha_n p^n.$$

Multiplicando por $-A$ y sumando el vector b, resulta

$$r^{n+1} = r^n - \alpha_n A p^n. \tag{7.17}$$

Nota 7.6.
Vamos a suponer que las direcciones p^k y los residuos r^k, $k \leq n-1$, son no nulos. De hecho estas suposiciones son muy realistas dado que $r^k = 0$ implica que x^k es la solución exacta del sistema en cuestión y $p^k = 0$ implica que $x^{n+1} = x^n$, es decir que la iteración $n+1$ es inútil.

De esta manera

$$\alpha_k = \frac{\left\langle r^k, p^k \right\rangle}{\left\langle Ap^k, p^k \right\rangle}$$

está siempre bien definido.

Empecemos por unos lemas:

Lema 7.1. *Sea $k \leq n-1$.*

$$\left\langle r^k, p^i \right\rangle = 0, \qquad i = 1, \ldots, k-1.$$

Demostración. Por inducción sobre k.

$$k = 1.$$

De la ecuación (7.17)

$$\left\langle r^1, p^0 \right\rangle = \left\langle r^0 - \alpha_0 A p^0, p^0 \right\rangle.$$

Dado que $p^0 = r^0$ y que

$$\alpha_k = \frac{\left\langle r^k, p^k \right\rangle}{\left\langle A p^k, p^k \right\rangle},$$

deducimos que

$$\left\langle r^1, p^0 \right\rangle = \left\langle r^0, r^0 \right\rangle - \frac{\left\langle r^0, r^0 \right\rangle}{\left\langle A r^0, r^0 \right\rangle} \left\langle A r^0, r^0 \right\rangle = 0.$$

Supongamos que la ecuación es verdadera para k y mostremos que

$$\left\langle r^{k+1}, p^i \right\rangle = 0, \qquad i = 0, \ldots, k.$$

Si $i = k$,

$$\left\langle r^{k+1}, p^k \right\rangle = \left\langle r^k - \alpha_k A p^k, p^k \right\rangle = \left\langle r^k, p^k \right\rangle - \frac{\left\langle r^k, p^k \right\rangle}{\left\langle A p^k, p^k \right\rangle} \left\langle A p^k, p^k \right\rangle = 0.$$

Si $i = k-1$,

$$\left\langle r^{k+1}, p^{k-1} \right\rangle = \left\langle r^k, p^{k-1} \right\rangle - \alpha_k \left\langle A p^k, p^{k-1} \right\rangle.$$

Por hipótesis de inducción $\left\langle r^k, p^{k-1} \right\rangle = 0$ y $\left\langle A p^k, p^{k-1} \right\rangle = 0$ por la elección de β_k.

Si $i < k-1$,

$$\left\langle r^{k+1}, p^i \right\rangle = \left\langle r^k, p^i \right\rangle - \alpha_k \left\langle A p^k, p^i \right\rangle.$$

Por hipótesis de inducción $\left\langle r^k, p^i \right\rangle = 0$. Así

$$\left\langle r^{k+1}, p^i \right\rangle = -\alpha_k \left\langle A p^k, p^i \right\rangle = -\alpha_k \left\langle p^k, \frac{\left(r^{i+1} - r^i\right)}{-\alpha_i} \right\rangle = \frac{\alpha_k}{\alpha_i} \left\langle p^k, r^{i+1} - r^i \right\rangle = 0.$$

□

Lema 7.2. *Para todo $k \geq 0$*

$$Ap^k \in \operatorname{gen}\left\{p^{k-1}, p^k, p^{k+1}\right\}.$$

Demostración. Tenemos

$$Ap^k = -\frac{1}{\alpha_k}\left(r^{k+1} - r^k\right) = -\frac{1}{\alpha_k}\left(p^{k+1} - \beta_{k+1}p^k - p^k + \beta_k p^{k-1}\right).$$

Así

$$Ap^k \in \operatorname{gen}\left\{p^{k-1}, p^k, p^{k+1}\right\}.$$

□

Proposición 7.2. *Sea $k \leq n-1$. Para*

$$\beta_k = -\frac{\left\langle Ar^k, p^{k-1}\right\rangle}{\left\langle Ap^{k-1}, p^{k-1}\right\rangle}$$

y

$$\begin{aligned} p^0 &= r^0, \\ p^k &= r^k + \beta_k p^{k-1}, \end{aligned}$$

la familia $\left\{p^k, \quad k = 0, \ldots, n-1\right\}$ es A-conjugada.

Demostración. Por inducción sobre k.

$$k \leq n-1.$$

Mostremos que

$$\left\langle Ap^k, p^i\right\rangle = 0, \qquad i = 0, \ldots, k-1.$$

Por la proposición (7.1) de la página 112,

$$\left\langle Ap^1, p^0\right\rangle = 0.$$

Supongamos que la proposición (7.2) es verdadera hasta k y mostremos que

$$\left\langle Ap^{k+1}, p^i\right\rangle = 0, \qquad i = 0, \ldots, k.$$

Si $i = k$,

$$\left\langle Ap^{k+1}, p^k\right\rangle = 0 \qquad \text{por la proposición (7.1).}$$

Si $i \leq k-1$,

$$\begin{aligned} \left\langle Ap^{k+1}, p^i\right\rangle &= \left\langle p^{k+1}, Ap^i\right\rangle, \\ &= \left\langle r^{k+1} + \beta_{k+1}p^k, Ap^i\right\rangle, \\ &= \left\langle r^{k+1}, Ap^i\right\rangle + \beta_{k+1}\left\langle p^k, Ap^i\right\rangle. \end{aligned}$$

Por hipótesis de inducción $\langle p^k, Ap^i \rangle = 0$. Así

$$\langle Ap^{k+1}, p^i \rangle = \langle r^{k+1}, Ap^i \rangle .$$

Del lema (7.2), podemos escribir Ap^i como combinación lineal de p^{i-1}, p^i, p^{i+1} y dado que $i \leq k-1$, utilizando el lema (7.1), deducimos que

$$\langle Ap^{k+1}, p^i \rangle = \langle r^{k+1}, Ap^i \rangle = 0.$$

□

Podemos escribir β_k en una forma más *memorizable* como lo muestra la proposición siguiente.

Proposición 7.3.

$$\beta_k = -\frac{\|r^k\|^2}{\|r^{k-1}\|^2},$$

para $r^0, \ldots, r^{k-1}$ no nulos.

Demostración.

$$\beta_k = -\frac{\langle r^k, Ap^{k-1} \rangle}{\langle Ap^{k-1}, p^{k-1} \rangle}$$

$$Ap^{k-1} = -\frac{1}{\alpha_{k-1}} \left(r^k - r^{k-1} \right)$$

$$\begin{aligned} \langle r^k, Ap^{k-1} \rangle &= -\frac{1}{\alpha_{k-1}} \langle r^k, r^k - r^{k-1} \rangle \\ &= -\frac{1}{\alpha_{k-1}} \|r^k\|^2 + \frac{1}{\alpha_{k-1}} \langle r^k, r^{k-1} \rangle \\ &= -\frac{1}{\alpha_{k-1}} \|r^k\|^2 \qquad \text{teniendo en cuenta el ejercicio (7.1).} \end{aligned}$$

Del otro lado,

$$\begin{aligned} \langle Ap^{k-1}, p^{k-1} \rangle &= -\frac{1}{\alpha_{k-1}} \langle r^k - r^{k-1}, p^{k-1} \rangle \\ &= -\frac{1}{\alpha_{k-1}} \langle r^k, p^{k-1} \rangle + \frac{1}{\alpha_{k-1}} \langle r^{k-1}, p^{k-1} \rangle . \end{aligned}$$

Por el lema (7.1) $\langle r^k, p^{k-1} \rangle = 0$ y como $\langle r^{k-1}, p^{k-1} \rangle = \|r^{k-1}\|^2$, tenemos

$$\beta_k = -\frac{\|r^k\|^2}{\|r^{k-1}\|^2}.$$

□

Recopilemos estos resultados en el siguiente teorema.

Teorema 7.4. Método del gradiente conjugado
Sean $A \in \mathbb{R}^{n\times n}$ una matriz simétrica definida positiva, $b \in \mathbb{R}^n$ y $\overline{x} = A^{-1}b$. Para todo $x^0 \in \mathbb{R}^n$ la sucesión x_n definida por el algoritmo

$$\begin{cases} p^0 = r^0 \\ k \geq 0 \\ \alpha_k = \dfrac{\langle p^k, r^k \rangle}{\langle Ap^k, p^k \rangle} \\ x^{k+1} = x^k + \alpha_k p^k \\ r^{k+1} = r^k - \alpha_k A p^k \\ \beta_{k+1} = -\dfrac{\left\| r^{k+1} \right\|^2}{\left\| r^k \right\|^2} \\ p^{k+1} = r^{k+1} + \beta_{k+1} p^k \end{cases}$$

satisface $x^n = \overline{x}$.

7.3. Velocidad de convergencia

A continuación analizemos otros aspectos optimales que satisface el gradiente conjugado.Consideramos la función $h : \mathbb{R}^{k+1} \to \mathbb{R}$ definida por

$$h(u) = J(x_0 + \sum_{i=0}^{i=k} u_i p_i) \tag{7.18}$$

Estudiamemos el mí nimo de esta función. Si notemos K la matrice $n \times (k+1)$ cuyas columnas son los $k + 1p_i$'s.Así se pude escribir h en la forma

$$h(u) = J(x_0 + Ku), \tag{7.19}$$

es deiir

$$h(u) = \frac{1}{2} \langle A(x_0 + Ku), x_0 + Ku \rangle - \langle b, x_0 + Ku \rangle$$

$$h(u) = \frac{1}{2} \langle (x_0 + Ku), x_0 + Ku \rangle - \langle b, x_0 + Ku \rangle$$

Proposición 7.4.

Capítulo 8

Elementos finitos

Introducción

Muchos problemas de física, mecánica y química dan lugar a problemas de frontera que involucran ecuaciones con derivadas parciales sobre un dominio geométrico y un conjunto de condiciones sobre la frontera del dominio en cuestión.

El método de elementos finitos se usa para este tipo de ecuaciones con la idea principal de reducir la búsqueda de la función solución a un espacio vectorial de dimensión finita.

La mejor manera de presentar este capítulo es a través de casos particulares porque la teoría general es difícil.

8.1. Elementos finitos en dimensión 1

Consideremos el problema siguiente. Encontrar $u(x)$, $x \in [0,1]$,

$$\begin{cases} -u''(x) + c\,u(x) = f(x), \qquad x \in]0,1[, & \text{(a)} \\ u(0) = u(1) = 0, & \text{(b)} \end{cases} \tag{8.1}$$

donde c es una constante positiva y f una función dada.

Nota 8.1.

Las condiciones de la ecuación (8.1.b) se llaman condiciones de frontera de tipo Dirichlet homogenea.

8.1.1. Desarrollo del método

Consideremos el espacio V_0 dado por

$$V_0 = \big\{ v : [0,1] \to \mathbb{R},\ v \text{ suave},\ v(0) = v(1) = 0 \big\}.$$

Con el término suave queremos decir que v tiene la regularidad (diferenciabilidad) que requieren las operaciones en el desarrollo.

Nota 8.2.

De acuerdo con el concepto de suavidad que acabamos de dar, V_0 es un espacio vectorial de dimensión infinita.

Sea $v \in V_0$. Multipliquemos la ecuación (8.1.a) del problema por v e integremos los dos lados sobre $[0, 1]$. Resulta

$$\int_0^1 (-u''v + cuv)\, dx = \int_0^1 fv\, dx. \tag{8.2}$$

Por razones de comodidad, omitimos la dependencia de la variable x de las funciones. Integrando por partes la ecuación (8.2) se sigue

$$\Big[-u'v\Big]_0^1 + \int_0^1 u'v'\, dx + c\int_0^1 uv\, dx = \int_0^1 fv\, dx.$$

Dado que $v(0) = v(1) = 0$, concluimos que

$$\int_0^1 (u'v' + cuv)\, dx = \int_0^1 fv\, dx. \tag{8.3}$$

Denotemos

$$a(u, v) = \int_0^1 (u'v' + cuv)\, dx, \tag{8.4}$$

$$l(v) = \int_0^1 fv\, dx. \tag{8.5}$$

La ecuación (8.3) se vuelve con estas notaciones

$$a(u, v) = l(v), \qquad \forall v \in V_0. \tag{8.6}$$

La ecuación (8.6) se llama formulación variacional del problema (8.1). Esta manera de escribir el problema suele ser clave para resolverlo numéricamente como veremos ahora.

La idea del método de elementos finitos es, en lugar de buscar u en el espacio V_0, buscarlo en un subespacio vectorial V_n de dimensión finita n. Por eso fijemos V_n tal subespacio.

Sea $\varphi_1, \varphi_2, \ldots, \varphi_n$ una base de V_n. Sea u^n la aproximación de u que pertenece a V_n. $u^n \in V_0$ implica la existencia de reales $\alpha_1, \ldots, \alpha_n$ tales que

$$u^n = \sum_{i=1}^{n} \alpha_i\, \varphi_i. \tag{8.7}$$

Definición 8.1. Las funciones φ_i, $i = 1, \ldots, n$, se llaman funciones de base.

Con esta aproximación la formulación variacional (8.6) se escribe

$$a\left(\sum_{i=1}^{n} \alpha_i\, \varphi_i\ , v^n\right) = l(v^n) \qquad \forall v^n \in V_n.$$

Las propiedades de a permiten escribir

$$\sum_{i=1}^{n} a\left(\varphi_i, v^n\right) \alpha_i = l(v^n) \qquad \forall v^n \in V_n. \tag{8.8}$$

v^n es arbitrario en V_n entonces podemos elegirlo como $v^n = \varphi_i$, $i = 1, \ldots, n$. Así la ecuación (8.8) se vuelve

$$\sum_{i=1}^{n} a\left(\varphi_i, \varphi_j\right) \alpha_i = l(\varphi_j) \qquad j = 1, \ldots, n. \tag{8.9}$$

Las ecuaciones (8.9) forman un sistema lineal cuadrado

$$Ax = b, \tag{8.10}$$

donde

$$\begin{gathered} A \in \mathbb{R}^n, \qquad A_{ij} = a\left(\varphi_i, \varphi_j\right), \\ x_i = \alpha_i, \qquad b_i = l\left(\varphi_i\right), \qquad i = 1, \ldots, n. \end{gathered}$$

Así buscar u en la forma (8.7) nos lleva a resolver el sistema lineal (8.10).

Ejemplo 8.1. Consideremos el problema de frontera

$$\begin{aligned} &-u'' + 4u = 16x^3 - 16x^2 - 24x + 8 \qquad \text{en }]0, 1[, \\ &u(0) = u(1) = 0. \end{aligned}$$

Buscamos la solución en el espacio V_3 generado por las tres funciones

$$\begin{aligned} \varphi_1(x) &= x \operatorname{sen}(\pi x), \\ \varphi_2(x) &= x(1 - \cos(x - 1)), \\ \varphi_3(x) &= x(x - 1). \end{aligned}$$

Admitimos que la familia $\{\varphi_1, \varphi_2, \varphi_3\}$ es linealmente independiente (o más bien tómelo como ejercicio).

Las tres funciones satisfacen las condiciones de frontera entonces todo está listo para armar el sistema (8.10). Sabiendo que

$$A_{ij} = \int_0^1 \varphi_i' \varphi_j' + 4\, \varphi_i \varphi_j \, dx, \qquad 1 \le i, j \le 3$$

y

$$b_i = \int_0^1 \left(16x^3 - 16x^2 - 24x + 8\right) \varphi_i \, dx, \qquad i = 1, \ldots, 3,$$

llegamos al sistema lineal

$$\begin{pmatrix} 2{,}46027 & 0{,}12476 & -0{,}89463 \\ 0{,}12476 & 0{,}39215 & -0{,}11296 \\ -0{,}89463 & -0{,}11296 & 0{,}46666 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} -2{,}58331 \\ -0{,}12402 \\ 0{,}93333 \end{pmatrix}. \tag{8.11}$$

Los cálculos de estos coeficientes han sido hechos por Maple 7. La solución del sistema (8.11) es

$$\text{sol} = \begin{pmatrix} -0{,}17667 \\ 7{,}21663 \\ 3{,}40814 \end{pmatrix}.$$

La solución aproximada es entonces

$$u^3 = \sum_{i=1}^{3} \alpha_i \, \varphi_i.$$

La solución exacta de este ejemplo es

$$u(x) = 4x^2(x-1).$$

La comparación gráfica de u con u^3 está dada en la figura (8.1).

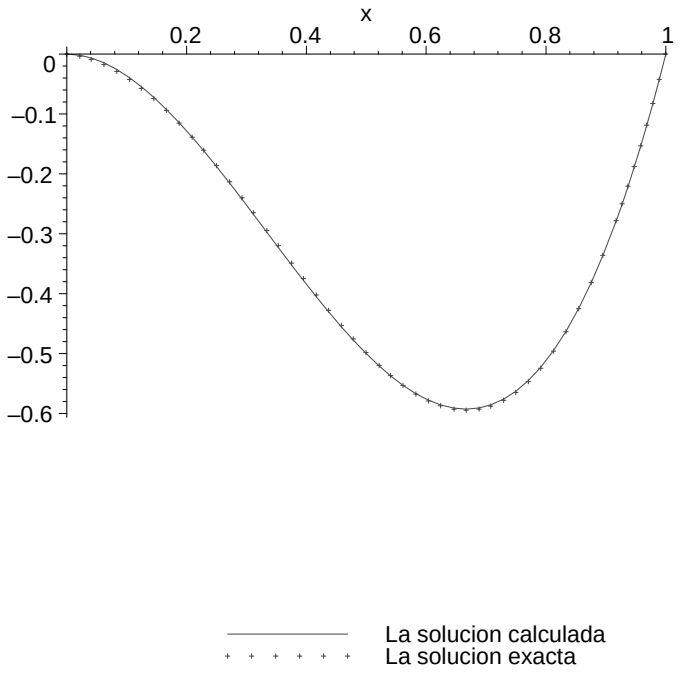

Figura 8.1.

Como lo vemos en este ejemplo la aproximación es muy aceptable pero notemos que el cálculo de los coeficientes de la matriz y del segundo miembro del sistema (8.11) es muy costoso y sería imposible si la dimensión del espacio de aproximación V_n fuera muy alta o las funciones de la base más complicadas. Con este método los coeficientes de la matriz $\big(a\,(\varphi_i, \varphi_j)\big)_{1 \le i,j \le n}$ son todos *a priori* no nulos.

Ahora vamos a trabajar en espacios $(V_n)_{n\in\mathbb{N}}$ de dimensión finita de manera que

1. Aumentar la dimensión n no cuesta mucho trabajo (fácilmente automatizable).
2. La matriz $\big(a(\varphi_i,\varphi_j)\big)_{1\le i,j\le n}$ tiene el mínimo de coeficientes no nulos.
3. No se necesitan muchos cálculos para construir el sistema lineal final.

8.1.2. Elementos finitos lineales en dimensión 1

Volvamos al problema (8.1)

$$\begin{cases} -u''(x)+c\,u(x)=f(x), & x\in[0,1],\\ u(0)=u(1)=0, \end{cases}$$

cuya formulación variacional es $a(u,v)=l(v),\ \forall v\in V_0$, donde

$$a(u,v)=\int_0^1 (u'v'+cuv)\,dx,$$
$$l(v)=\int_0^1 fv\,dx,$$

$$V_0=\big\{v:[0,1]\to\mathbb{R},\ v \text{ suave},\ v(0)=v(1)=0\big\}.$$

Ahora construimos V_n. Sean $n\in\mathbb{N}^*$ y $h=\frac{1}{n+1}$. Partimos el intervalo $[0,1]$ por la subdivisión $x_i=ih$, $i=0,\dots,n+1$.

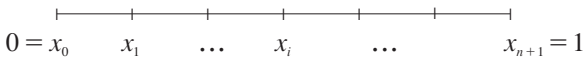

Figura 8.2.

Para cada $i=1,\dots,n$, definamos la función $\varphi_i:[0,1]\to\mathbb{R}$ como sigue

$$\varphi_i(x)=\begin{cases} 0, & x\le x_{i-1},\\ \frac{x-x_{i-1}}{h}, & x_{i-1}\le x\le x_i,\\ \frac{x_{i+1}-x}{h}, & x_i\le x\le x_{i+1},\\ 0, & x\ge x_{i+1}. \end{cases} \tag{8.12}$$

Nota 8.3.

Las funciones φ_i, $i=1,\dots,n$, son continuas, diferenciables en $[0,1]$ excepto en los puntos x_i, $i=1,\dots,n$,. Además para todo $i=1,\dots,n$,y todo $j=1,\dots,n$,

$$\varphi_i(x_j)=\delta_{ij}.$$

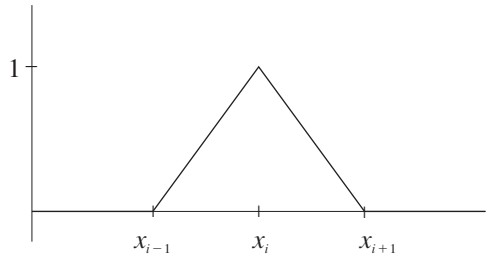

Figura 8.3.

Proposición 8.1. *La familia φ_i, $i = 1, \ldots, n$, es linealmente independiente.*

Demostración. Sea la combinación lineal

$$\sum_{i=1}^{n} \lambda_i \, \varphi_i = 0. \tag{8.13}$$

La igualdad (8.13) es funcional, es decir

$$\forall x \in [0,1], \qquad \sum_{i=1}^{n} \lambda_i \, \varphi_i(x) = 0,$$

en particular para los x_j, $j = 1, \ldots, n$. Así

$$0 = \sum_{i=1}^{n} \lambda_i \, \varphi_i(x_j) = \lambda_j, \quad j = 1, \ldots, n,$$

es decir la familia φ_i, $i = 1, \ldots, n$, es linealmente independiente. □

Proposición 8.2. *Para los φ_i, $i = 1, \ldots, n$, definidos en (8.12) tenemos $a\,(\varphi_i, \varphi_j) = 0$ si $|i - j| \geq 2$.*

Demostración. Si $|i - j| \geq 2$, no hay $x \in [0,1]$ para el cual $\varphi_i(x) \neq 0$ y $\varphi_j(x) \neq 0$. Así

$$\varphi_i(x)\varphi_j(x) = 0, \qquad \forall x \in [0,1].$$

De la misma manera

$$\varphi_i'(x)\varphi_j'(x) = 0.$$

Así

$$a\,(\varphi_i, \varphi_j) = 0.$$

□

Nota 8.4.
Este resultado significa que la matriz

$$A = \left(a\left(\varphi_i, \varphi_j\right)\right)_{1 \leq i,j \leq n}$$

es tridiagonal.

Cálculo de los coeficientes de A

$$\varphi_i'(x) = \begin{cases} 1/h, & x \in [x_{i-1}, x_i], \\ -1/h, & x \in [x_i, x_{i+1}], \\ 0, & \text{si no.} \end{cases}$$

Así para $1 \leq i \leq n$

$$\begin{aligned} a\left(\varphi_i, \varphi_i\right) &= \int_0^1 \left(\varphi_i'\right)^2 dx + \int_0^1 \varphi_i^2 \, dx, \\ &= \int_{x_{i-1}}^{x_{i+1}} \frac{1}{h^2} \, dx + c \int_{x_{i-1}}^{x_i} \left(\frac{x - x_{i-1}}{h}\right)^2 dx + c \int_{x_i}^{x_{i+1}} \left(\frac{x_{i+1} - x}{h}\right)^2 dx, \\ &= \frac{2}{h} + \frac{c}{h^2} \frac{1}{3} \left[\left(\left(x - x_{i-1}\right)^3 \right)_{x_{i-1}}^{x_i} + \left(\left(-x + x_{i+1}\right)^3 \right)_{x_i}^{x_{i+1}} \right], \\ &= \frac{2}{h} + \frac{c}{3h^2} \left(h^3 + h^3\right), \\ &= \frac{2}{h} + \frac{2ch}{3}. \end{aligned}$$

Para $1 \leq i \leq n-1$

$$a\left(\varphi_i, \varphi_{i+1}\right) = \int_0^1 \varphi_i' \varphi_{i+1}' \, dx + c \int_0^1 \varphi_i \varphi_{i+1} \, dx.$$

Pero los productos $\varphi_i' \varphi_{i+1}'$ y $\varphi_i \varphi_{i+1}$ son nulos excepto en $[x_i, x_{i+1}]$. Así

$$\begin{aligned} a\left(\varphi_i, \varphi_{i+1}\right) &= \int_{x_i}^{x_{i+1}} -\frac{1}{h^2} \, dx + c \int_{x_i}^{x_{i+1}} \frac{\left(x - x_i\right)\left(x_{i+1} - x\right)}{h^2} \, dx, \\ &= -\frac{1}{h} + \frac{ch}{6}. \end{aligned}$$

La simetría de la matriz A implica que

$$A = \begin{pmatrix} s & t & & & 0 \\ t & s & t & & \\ & t & \ddots & \ddots & \\ & & \ddots & \ddots & t \\ 0 & & & t & s \end{pmatrix},$$

donde

$$s = \frac{2}{h} + \frac{2ch}{3}, \qquad t = -\frac{1}{h} + \frac{ch}{6}.$$

Cálculo del vector b

$$b_i = \int_{x_{i-1}}^{x_{i+1}} \varphi_i(x) f(x)\, dx,$$

que se calcula de manera analítica si f es una función sencilla. Si no, se usa uno de los métodos de integración numérica vistos en el capítulo 2. El sistema tridiagonal que resulta se resuelve con un método directo o iterativo como los vistos en los capítulos 5, 6 y 7.

Ejemplo 8.2. Resolvamos el ejemplo siguiente con este método

$$\begin{aligned} &-u'' + 4u = \left[(16\pi^2 + 3)\,\mathrm{sen}(4\pi x) - 8\cos(4\pi x)\right] e^x, \\ &u(0) = u(1) = 0 \end{aligned}$$

y sabiendo que la solución exacta (figura 8.4) es

$$e^x \,\mathrm{sen}(4\pi x).$$

Para examinar la precisión del método resolvemos el problema para varios valores de n, después calculamos la norma infinita del vector U^n, solución del sistema lineal, que representa los valores de la solución aproximada u^n en los nodos x_i, $i = 1, \ldots, n$. Comparamos esta norma con la de la solución exacta U dada por

$$U_i = u(x_i), \qquad i = 1, \ldots, n.$$

Para tener una idea mejor de la aproximación, calculamos el error relativo dado por

$$e_r = \frac{\|U - U^n\|_\infty}{\|U\|_\infty}.$$

Los resultados

n	$\|U - U^n\|_\infty$	e_r
2	0,13602	0,08064
10	0,00569	0,00252
100	0,00007	0,00003

son aceptables teniendo en cuenta la precisión lograda y el trabajo que ha costado.

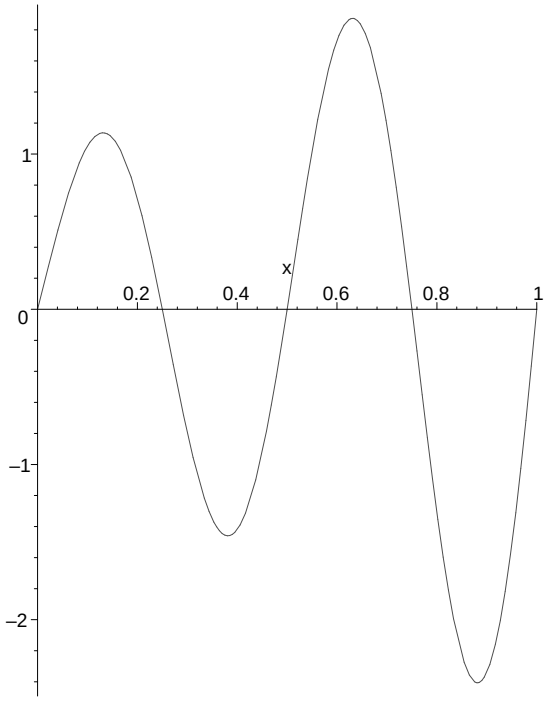

Figura 8.4. La solución exacta del ejemplo (8.2).

8.1.3. Otros tipos de condiciones de frontera

Miremos variantes del problema (8.1).

Condiciones de frontera de tipo Dirichlet general

$$\begin{cases} -u''(x) + c\,u(x) = f(x), \qquad x \in]0,1[, & \text{(a)} \\ u(0) = \alpha,\ u(1) = \beta. & \text{(b)} \end{cases} \tag{8.14}$$

Se hace, como en el caso anterior, la discretización de $[0,1]$ escogiendo en $n \in \mathbb{N}^+$

$$h = \frac{1}{n+1}, \qquad x_i = ih, \qquad i = 0, \ldots, n+1.$$

Las funciones φ_i, $i = 1, \ldots, n$, se definen como en (8.12) de la página 123 y

$$\varphi_0(x) = \begin{cases} \frac{h-x}{h}, & 0 \le x \le h, \\ 0 & h \le x \le 1. \end{cases}$$

$$\varphi_{n+1}(x) = \begin{cases} 0, & 1 \le x \le 1-h, \\ \frac{x-1+h}{h} & 1-h \le x \le 1. \end{cases}$$

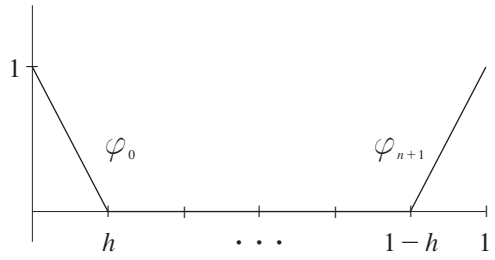

Figura 8.5.

Busquemos la formulación variacional de (8.14). Sea

$$V_{\alpha,\beta} = \big\{v,\ v \text{ suave y } v(0) = \alpha,\ v(1) = \beta\big\}.$$

Multipliquemos la ecuación (8.14.a) por un $v \in V$ e integremos por partes

$$\int_0^1 (-u''v + cuv)\,dx = \int_0^1 fv\,dx, \qquad \forall v \in V_{\alpha,\beta},$$

$$\int_0^1 u'v' + cuv\,dx - u'(1)v(1) + u'(0)v(0) = \int_0^1 fv\,dx,$$

$$\int_0^1 (u'v' + cuv)\,dx - \alpha u'(1) + \beta u'(0) = \int_0^1 fv\,dx, \qquad \forall v \in V_{\alpha,\beta}. \tag{8.15}$$

Particularmente para $v = u$, la solución del problema

$$\int_0^1 (u'u' + cuu)\,dx - \alpha u'(1) + \beta u'(0) = \int_0^1 fv\,dx, \qquad \forall v \in V. \tag{8.16}$$

Restando (8.16) de (8.15)

$$\int_0^1 u'\,(u' - v') + cu(u - v)\,dx = \int_0^1 f(v - u)\,dx.$$

Si $u, v \in V_{\alpha,\beta}$ $\quad u - v \in V_0$. Recordemos que

$$V_0 = \big\{v : [0,1] \to \mathbb{R},\ v \text{ suave},\ v(0) = v(1) = 0\big\}.$$

Así,

$$\int_0^1 u'w' + cuw\,dx = \int_0^1 fv\,dx \qquad \forall w \in V_0,$$

que se escribe también en la forma

$$a(u, v) = l(v), \qquad \forall v \in V_0.$$

Entonces la formulación variacional del problema (8.14) es la misma que cuando $\alpha = \beta = 0$, si se elige

$$u^n = \sum_{i=0}^{n+1} \lambda_i \varphi_i,$$

donde $\{\varphi_i, i = 0, \ldots, n\}$ es la base de funciones afines por trozos. De la formulación variacional resulta

$$\sum_{j=0}^{n+1} a_{ij}\lambda_i = b_i, \quad i = 0, \ldots, n+1, \tag{8.17}$$

donde

$$b_i = \int_0^1 f\varphi_i \, dx, \quad i = 0, \ldots, n+1,$$
$$a_{ij} = a(\varphi_i, \varphi_j), \quad 0 \leq i, j \leq n+1.$$

Sabemos que $\lambda_0 = \alpha,\ \lambda_{n+1} = \beta$ entonces podemos imponer directamente en (8.17)

$$\begin{array}{llll} a_{00} = 1, & b_0 = \alpha, & a_{0j} = 0, & j = 1, \ldots, n+1, \\ a_{n+1,n+1} = 1, & b_{n+1} = \beta, & a_{n+1,j} = 0 & j = 1, \ldots, n. \end{array}$$

y el sistema (8.17) se vuelve

$$\begin{pmatrix} 1 & 0 & 0 & \cdots \\ a_{10} & a_{11} & \cdots & a_{1,n+1} \\ \vdots & & & \\ a_{n0} & a_{n1} & \cdots & a_{n,n+1} \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \lambda_o \\ \lambda_1 \\ \vdots \\ \lambda_n \\ \lambda_{n+1} \end{pmatrix} = \begin{pmatrix} \alpha \\ b_1 \\ \vdots \\ b_n \\ \beta \end{pmatrix}$$

Nota 8.5.
Este sistema lineal es de dimensión $(n+2)$.

Ejemplo 8.3.

$$-u'' + 4u = e^x \left[3\cos(4\pi x) + 8\,\mathrm{sen}(4\pi x) + 16\pi^2 \cos(4\pi x)\right],$$
$$u(0) = 1, \qquad u(1) = e.$$

La solución exacta del problema es

$$u(x) = e^x \cos(4\pi x).$$

Los resultados obtenidos están resumidos en la tabla siguiente

n	$\|U - U^n\|_\infty$	e_r
2	0,10125	0,03724
12	0,00920	0,00338
100	0,00012	0,00004

Condiciones de frontera de tipo Neumann

Las condiciones de tipo Neumann son las que involucran la derivada de la función buscada en la frontera del dominio, es decir, de la forma

$$u'(a) = \alpha,$$

donde a es una extremo del intervalo del problema. Miremos cómo es la formulación variacional del problema cuando tiene este tipo de condiciones de frontera. Sea el problema

$$\begin{cases} -u''(x) + c\,u(x) = f(x), \qquad x \in]0\ 1[, & \text{(a)} \\ u(0) = 0, \qquad u'(1) = \beta. & \text{(b)} \end{cases} \tag{8.18}$$

Aquí la condición de Neumann está aplicada sólo en el punto $x = 1$.

Sea

$$V_{\text{Neu}} = \big\{ v : [0,1] \to \mathbb{R}, \ \text{ suave y } v(0) = 0 \big\}.$$

Para un $v \in V_{\text{Neu}}$ multipliquemos la ecuación $-u'' + c\,u = f$ por v e integremos sobre $[0,1]$.

$$\int_0^1 u'v' + cuv\,dx - v(1)u'(1) = \int_0^1 fv\,dx.$$

Pero $u'(1) = \beta$. Así la formulación variacional del problema (8.18) da

$$a(u,v) = l(v), \qquad \forall v \in V_{\text{Neu}},$$

donde

$$a(u,v) = \int_0^1 (u'v' + cuv)\,dx,$$
$$l(v) = \int_0^1 fv\,dx + \beta v(1).$$

Nota 8.6.

Lo que cambia en esta formulación con respecto al problema con condiciones de Dirichlet es la función l que ahora tiene un término adicional que involucra la condición de Neumann.

Ahora sean $\varphi_0, \varphi_1, \ldots, \varphi_n, \varphi_{n+1}$ las funciones descritas en (8.15) y (8.16), de la página 128, y (8.12) de la página 123. Busquemos la solución del problema en la forma

$$u = \sum_{i=0}^{n+1} \alpha_i \varphi_i.$$

El sistema lineal que resulta es

$$\sum_{j=0}^{n+1} a\left(\varphi_i, \varphi_j\right) \alpha_j = l\left(\varphi_i\right), \qquad i = 0, \ldots, n+1.$$

La matriz que resulta es

$$\begin{pmatrix} \alpha & \beta & & & 0 \\ \beta & \alpha & \beta & & \\ & \beta & \ddots & \ddots & \\ & & \ddots & \ddots & \beta \\ 0 & & & \beta & \alpha \end{pmatrix}.$$

Como para todo φ_i, $i = 0, \ldots, n$,

$$\varphi_i(1) = 0$$

y

$$\varphi_{n+1}(1) = 1.$$

$$l = \begin{cases} \int_0^1 \varphi_i f \, dx, & i = 0, \ldots, n, \\ \beta + \int_0^1 \varphi_{n+1} f \, dx, & i = n+1. \end{cases}$$

Ejemplo 8.4.

$$\begin{cases} -u'' + u = [144 \operatorname{sen}(12x) - 24 \cos(12x)] \, e^x, & x \in]0\ 1[, \\ u(0) = 0, \qquad u'(1) = 26{,}06743. \end{cases}$$

Cuya solución exacta (figura 8.6) es

$$u(x) = e^x \operatorname{sen}(12x).$$

Los resultados se resumen en la tabla siguiente

n	$\|U - U^n\|_\infty$	e_r
2	0,22448	0,11649
10	0,14747	0,00596
100	0,00016	0,00006

8.2. Elementos finitos en dimensión 2

8.2.1. Preliminares de cálculo vectorial

Primera fórmula de Green

Dados un dominio acotado $\Omega \subset \mathbb{R}^n$ de frontera suave $\partial\Omega$ y un campo vectorial suave $F : \Omega \to \mathbb{R}^n$, tenemos

$$\int_{\partial\Omega} Fn \, ds = \int_\Omega \operatorname{div} F \, dx,$$

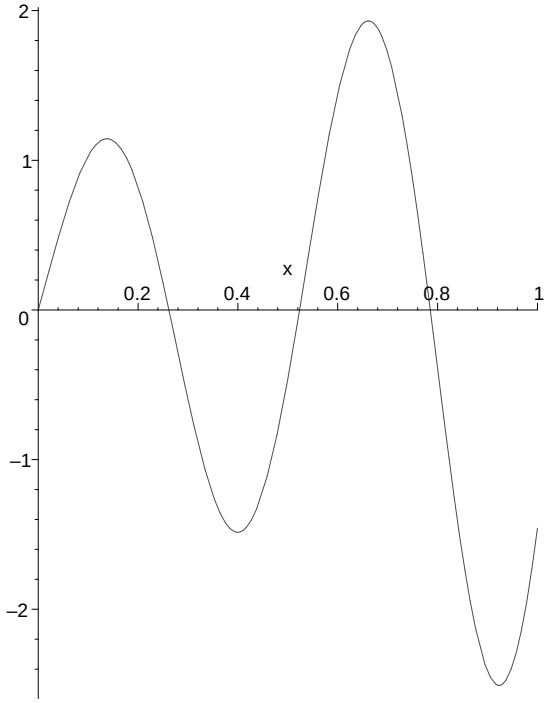

Figura 8.6. La solución exacta del ejemplo (8.4).

donde n es la normal unitaria saliente de $\partial\Omega$ que, evidentemente, depende de s y

$$\operatorname{div} F = \sum_{i=1}^{n} \frac{\partial F_i}{\partial x_i}.$$

Nota 8.7.

En esta sección, x denota una variable general de Ω y s la variable curvilínea de la frontera $\partial\Omega$.

Segunda fórmula de Green

Sean u, v dos campos escalares y suaves. Denotemos ∇u el gradiente de u definido por

$$\nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u}{\partial x_n} \right)^t.$$

El campo $v\nabla u : \Omega \to \mathbb{R}^n$ es suave de modo que se le puede aplicar la primera fórmula de Green, es decir

$$\int_{\Omega} \operatorname{div}(v\nabla u)\, dx = \int_{\partial\Omega} v\nabla u\, n\, ds.$$

Pero

$$\operatorname{div}(v\nabla u) = \nabla u \nabla v + v\Delta u$$

donde Δu es el operador laplaciano definido por

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Así

$$\int_\Omega \nabla u \nabla v \, dx + \int_\Omega v\Delta u \, dx = \int_{\partial\Omega} v\nabla u \, n \, ds$$

o también

$$-\int_\Omega v\Delta u \, dx = \int_\Omega \nabla u \nabla v \, dx - \int_{\partial\Omega} v\nabla u \, n \, ds. \tag{8.19}$$

La ecuación (8.19) se llama segunda fórmula de Green o, sencillamente, fórmula de Green.

8.2.2. Problema modelo

Como en el caso monodimensional, vamos a ver los diferentes pasos a seguir en elementos finitos a través de un ejemplo modelo.

Sean Ω un dominio del plano con frontera $\partial\Omega$ suave, c una constante positiva y $f : \Omega \to \mathbb{R}$ una función dada. Nuestro problema es buscar $u : \Omega \to \mathbb{R}$ tal que

$$\begin{cases} -\Delta u(x,y) + cu(x,y) = f(x.y) \quad \text{en } \Omega, & \text{(a)} \\ u = 0 \quad \text{en } \partial\Omega. & \text{(b)} \end{cases} \tag{8.20}$$

El problema (8.20) es un problema de frontera. La condición (8.20.b) se llama condición de frontera de tipo Dirichlet.

Formulación variacional del problema (8.20)

Consideremos el espacio vectorial

$$V_0 = \big\{v : \Omega \to \mathbb{R} \text{ suave tal que } v = 0 \text{ en } \partial\Omega\big\}.$$

Sea $v \in V_0$ arbitrario. Multipliquemos la ecuación (8.20.a) por v e integremos sobre Ω. Obtenemos

$$\int_\Omega -\Delta u \, v \, dx + c\int_\Omega uv \, dx = \int_\Omega fv \, dx, \qquad \forall v \in V_0.$$

Utilizando la fórmula de Green, obtenemos

$$\int_\Omega \nabla u \nabla v \, dx - \int_{\partial\Omega} v\nabla u \, n \, ds + \int_\Omega cuv \, dx = \int_\Omega fv \, dx.$$

Pero $v = 0$ en $\partial\Omega$. Así

$$\int_\Omega (\nabla u \nabla v + cuv)\, dx = \int_\Omega fv\, dx, \qquad \forall v \in V_0. \tag{8.21}$$

La ecuación (8.21) se llama formulación variacional del problema (8.20) y se escribe también en la forma

$$a(u, v) = l(v), \qquad \forall v \in V_0 \tag{8.22}$$

donde

$$a(u, v) = \int_\Omega (\nabla u \nabla v + cuv)\, dx$$

y

$$l(v) = \int_\Omega fv\, dx.$$

Nota 8.8.

Igual que en el caso de dimensión 1, a es una forma bilineal simétrica sobre V_0 y l es una forma lineal sobre V_0.

Sea V^n un subespacio vectorial de V_0 de dimensión finita. Vamos a buscar una solución u^n dentro de V^n que aproxime u. Para eso, sean $\varphi_1, \varphi_2, \ldots, \varphi_n$, n funciones linealmente independientes de V_0 y denotemos V^n al subespacio de V_0 generado por estas funciones. Busquemos u^n en la forma

$$u^n = \sum_{n=1}^{n} \lambda_i \varphi_i$$

donde $\lambda_1, \ldots, \lambda_n$ son escalares a ser hallados. De la formulación variacional (8.22) y variando $v = \varphi_i$, $i = 1, \ldots, n$, tenemos

$$\sum_{i=1}^{n} a(\varphi_i, \varphi_j)\lambda_j = l(\varphi_i), \qquad i = 1, \ldots, n$$

que constituye un sistema lineal

$$Ax = b$$

donde

$$A_{ij} = a(\varphi_i, \varphi_j), \qquad i, j = 1, \ldots, n$$

y

$$b_i = l(\varphi_i), \qquad i = 1, \ldots, n.$$

Construcción de los espacios V^n

La construcción de los espacios V^n se realiza con las siguientes consideraciones:

1. El proceso de hallar la base no es muy costoso.
2. La matriz que resulta tiene el mínimo de coeficientes no nulos.
3. El cálculo de los coeficientes $a(\varphi_i, \varphi_j), l(\varphi_i)$ es relativamente fácil.

Definición 8.2. El soporte de una función $h : \Omega \to \mathbb{R}$ es la clausura del conjunto $\{x \in \Omega; \quad h(x) \neq 0\}$ que denotamos $\operatorname{sop} h$.

Discretización del dominio

Una manera de conseguir que la matriz del sistema lineal que resulta sea dispersa, es decir con un número alto de coeficientes nulos, es escoger las funciones de base con soportes reducidos en el dominio Ω. Así, si dos funciones φ_i, φ_j tienen soportes disyuntos, el coeficiente A_{ij} es necesariamente nulo. Es por esta razón que partimos Ω, el dominio geométrico del problema, en subdominios de formas geométricas facilmente manejables (triángulos, rectángulos, ...).

Definición 8.3. Una discretización del dominio $\Omega \subset \mathbb{R}^2$ es la partición del dominio en subdominios $E_k, k = 1, \ldots, N_e$, tal que

$$\bigcup_{k=1}^{N_e} E_k = \Omega.$$

Para todo $i, j = 1, \ldots, N_e$, la intersección $E_i \cap E_j$ es vacía, un punto o un segmento. Los subdominios $E_k, k = 1, \ldots, N_e$, se llaman elementos y sus vértices P_i, de coordenadas (x_i, y_i), $i = 1, \ldots, N$, nodos.

8.2.3. Elemento finito lineal: triángulo con 3 nodos

Para este caso se discretiza en triángulos como lo muestra la figura (8.7).

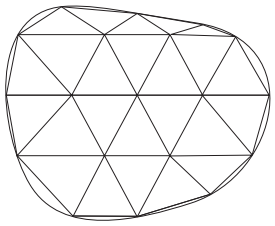

Figura 8.7.

Nota 8.9.

Si el dominio Ω no es un polígono se hace una discretización de la frontera de manera que se aproxime a un polígono como se ve en la figura (8.7). En otras palabras, se desprecian los *trozos* que exceden al polígono.

A cada nodo P_i, $i = 1, \ldots, N$, se le asocia una función φ_i tal que

$$\varphi_i(x_j, y_j) = \delta_{ij}, \qquad i, j = 1, \ldots, N$$

y φ_i es de gráfica plana sobre cada triángulo de la discretización.

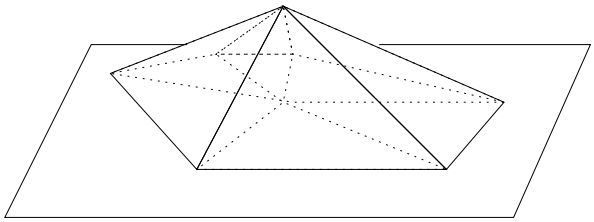

Figura 8.8.

Nota 8.10.

Para este elemento finito el soporte $\text{sop}(\varphi_i)$ de φ_i es la unión de todos los triángulos que tienen el nodo (x_i, y_i) como vértice (que podemos ver también como el piso de la carpa en la figura 8.8).

Cálculo de la matriz del sistema (matriz de rigidez)

Intuitivamente, uno calcularía el coeficiente

$$A_{ij} = \int_{\Omega} \nabla\varphi_i \nabla\varphi_j + c\,\varphi_i\varphi_j \; dx$$

de manera directa variando i y j de 1 hasta N pero esta técnica de calcular los coeficientes de A uno tras otro tiene desventajas:

a) Las funciones φ_i cambian de expresión sobre cada triángulo.

b) El número de triángulos en la intersección de los soportes no es siempre constante como se ve en el ejemplo de la figura (8.9).

$$\text{sop}(\varphi_1) \cap \text{sop}(\varphi_2) = T_1$$
$$\text{sop}(\varphi_2) \cap \text{sop}(\varphi_3) = T_2 \cup T_3$$

Así la sistematización de los cálculos no es fácil. Para evitar estos problemas se trabaja con la estrategia elemento por elemento. La idea del método se basa en

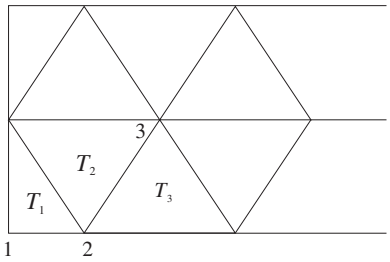

Figura 8.9.

la formulación variacional. Sabemos que

$$a(\varphi_i, \varphi_j) = \int_\Omega \nabla\varphi_i \nabla\varphi_j + c\,\varphi_i\varphi_j \; dx,$$
$$a(\varphi_i, \varphi_j) = \sum_{T_k \in \mathcal{T}} \int_{T_k} \nabla\varphi_i \nabla\varphi_j + c\,\varphi_i\varphi_j \; dx,$$

donde $\mathcal{T}$ es el conjunto de los elementos (triángulos en este caso) de la discretización. Así

$$a(\varphi_i, \varphi_j) = \sum_{T_k \in \mathcal{T}} a_k(\varphi_i, \varphi_j)$$

con

$$a_k(\varphi_i, \varphi_j) = \int_{T_k} \nabla\varphi_i \nabla\varphi_j + c\,\varphi_i\varphi_j \; dx.$$

Definición 8.4. La matriz $A^k \in \mathbb{R}^{3\times 3}$ definida por

$$A^k_{ij} = a_k(\varphi_i, \varphi_j)$$

se llama matriz elemental asociada al elemento T_k.

El número $a_k(\varphi_i, \varphi_j)$ se puede ver como la participación del elemento T_k en el coeficiente $a(\varphi_i, \varphi_j)$. La ventaja de trabajar con los $a_k(\varphi_i, \varphi_j)$ es que permite automatizar la computación de manera sencilla debido a que cada triángulo T_k tiene exactamente 3 funciones $\varphi_{k_1}, \varphi_{k_2}, \varphi_{k_3}$, no nulas en él. Esta automatización se hace de la manera siguiente:

Se recorren los triángulos uno a uno. Sea T_k un triángulo de vértices (x_{k_1}, y_{k_1}), (x_{k_2}, y_{k_2}) y (x_{k_3}, y_{k_3}). Las tres funciones $\varphi_{k_1}, \varphi_{k_2}, \varphi_{k_3}$ son las restricciones al triángulo T_k asociadas a sus vértices. Como estas restricciones son de tipo $ax + by + c$ entonces encontrar $\varphi_{k_1}, \varphi_{k_2}, \varphi_{k_3}$ se reduce a calcular los 9 coeficientes

$$a_{k_i}, b_{k_i}, c_{k_i}, \qquad i = 1, 2, 3,$$

de las funciones

$$\varphi_{k_i}(x,y) = a_{k_i}x + b_{k_i}y + c_{k_i}, \qquad i = 1,2,3.$$

Pero dado que

$$\varphi_i(x_j, y_j) = \delta_{ij},$$

el problema se reduce entonces a invertir la matriz

$$E_k = \begin{bmatrix} x_{k_1} & y_{k_1} & 1 \\ x_{k_2} & y_{k_2} & 1 \\ x_{k_3} & y_{k_3} & 1 \end{bmatrix}.$$

Así

$$E_k^{-1} = \begin{bmatrix} a_{k_1} & a_{k_2} & a_{k_3} \\ b_{k_1} & b_{k_2} & b_{k_3} \\ c_{k_1} & c_{k_2} & c_{k_3} \end{bmatrix}.$$

Nota 8.11.
Un simple cálculo muestra que

$$E_k^{-1} = \frac{1}{2\,|T_k|} \begin{bmatrix} y_{k_2} - y_{k_3} & y_{k_3} - y_{k_1} & y_{k_1} - y_{k_2} \\ x_{k_3} - x_{k_2} & x_{k_1} - x_{k_3} & x_{k_2} - x_{k_1} \\ x_{k_2}y_{k_3} - x_{k_3}y_{k_2} & x_{k_3}y_{k_1} - x_{k_1}y_{k_3} & x_{k_1}y_{k_2} - x_{k_2}y_{k_1} \end{bmatrix}.$$

Nota 8.12.
Notemos que el determinante de la matriz E_k^{-1} es $2\,|T_k|$ donde $|T_k|$ es el área de T_k.

Ahora el cálculo de los coeficientes $a_k(\varphi_i, \varphi_j)$ de la matriz elemental se vuelve muy sencillo. Tenemos

$$\nabla\varphi_{k_i}(x,y) = (a_{k_i}, b_{k_i})^t.$$

Así

$$a_k(\varphi_i,\varphi_j) = \int_{T_k} a_{k_i}a_{k_j} + b_{k_i}b_{k_j}\,dx + c\int_{T_k} \varphi_{k_i}(x,y)\,\varphi_{k_j}(x,y)\,dx,$$

$$a_k(\varphi_i,\varphi_j) = |T_k|\left(a_{k_i}a_{k_j} + b_{k_i}b_{k_j}\right) + c\int_{T_k} \varphi_{k_i}(x,y)\,\varphi_{k_j}(x,y)\,dx.$$

Enumeración de nodos

Hasta ahora nos hemos referido a los nodos P_i, de coordenadas (x_i, y_i), $i = 1, \ldots, N$, listándolos sin más detalles. Implícitamente, estamos haciendo una enumeración de los nodos. La importancia de este concepto es que en la matriz del sistema lineal un coeficiente a_{ij} es nulo si la intersección $\mathrm{sop}(\varphi_i) \cap \mathrm{sop}(\varphi_j)$ es vacía. Es decir que dos funciones asociadas a dos nodos i, j geométricamente

alejados dan lugar, generalmente, a un coeficiente nulo. Más precisamente, estamos seguros de que si i, j no son nodos del mismo triángulo, necesariamente $a_{ij} = 0$. Considerando estos aspectos, es importante hacer una enumeración de nodos de manera que la matriz tenga un ancho de banda mínimo.

Para entender este concepto, presentamos en el ejemplo a continuación dos enumeraciones diferentes para una misma discretización.

Ejemplo 8.5. En este ejemplo, los coeficientes no nulos de la matriz se denotan con una caja ■. Las casillas vacías corresponden a los coeficientes nulos de la matriz.

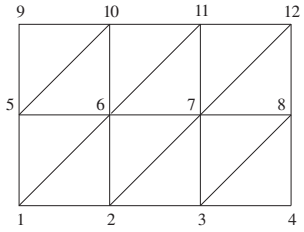

Figura 8.10. Enumeración 1 del ejemplo (8.5).

	1	2	3	4	5	6	7	8	9	10	11	12
1	■	■			■	■						
2	■	■	■			■	■					
3		■	■	■			■	■				
4			■	■				■				
5	■				■	■			■	■		
6	■	■			■	■	■			■	■	
7		■	■			■	■	■			■	■
8			■	■			■	■				■
9					■				■	■		
10					■	■			■	■	■	
11						■	■			■	■	■
12							■	■			■	■

Tabla 8.1. Estructura de la matriz de la Enumeración 1 del ejemplo (8.5).

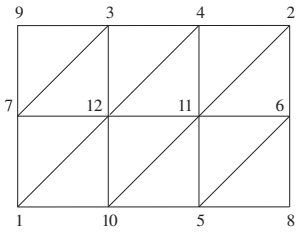

Figura 8.11. Enumeración 2 del ejemplo (8.5).

	1	2	3	4	5	6	7	8	9	10	11	12
1	■						■			■		■
2		■		■		■					■	
3			■	■			■		■			■
4		■	■	■							■	■
5					■	■		■		■	■	
6		■			■	■		■			■	
7	■		■				■		■			■
8					■	■		■				
9			■				■		■			
10	■				■					■	■	■
11		■		■	■	■				■	■	■
12	■		■	■			■			■	■	■

Tabla 8.2. Estructura de la matriz de la Enumeración 2 del ejemplo (8.5).

Buena enumeración, mala enumeración

Como se ve en el ejemplo (8.5), la estructura de la matriz en términos de la distribución de ceros está estrechamente ligada a la manera de enumerar los nodos. Tomando como criterio el ancho de la banda de la matriz, una buena enumeración es la que minimiza máx $|i - j|$, donde i y j son vértices del mismo triángulo. Es muy complejo resolver este problema de optimización para una discretización dada.

Tabla de conexiones

La técnica de trabajar elemento por elemento exige que asociemos a cada elemento los nodos que constituyen sus vértices. La tabla que agrupa esta información de llama tabla de conexiones.

Para facilitar la tarea de crear la tabla de conexiones, es importante identificar cada elemento. Es por eso que también se enumeran los elementos de 1 hasta N_e.

Nota 8.13.

Contrariamente a la enumeración de los nodos, la enumeración de los elementos no tiene incidencia alguna sobre la estructura de la matriz ni sobre el proceso numérico.

Ilustremos con un ejemplo la tabla de conexiones

Ejemplo 8.6. En la discretización de este ejemplo (figura 8.12), los números de elemento están en un círculo para diferenciarlos de los números de nodo.

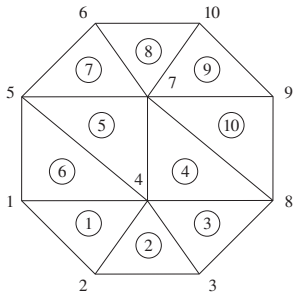

Figura 8.12. Discretización del ejemplo (8.6).

Número del elemento	Sus nodos		
1	1	2	4
2	2	3	4
3	4	3	8
4	4	8	7
5	4	7	5
6	1	4	5
7	5	6	7
8	6	7	10
9	7	9	10
10	7	8	9

Tabla 8.3. Tabla de conexiones del ejemplo (8.6).

Nota 8.14.
No es importante el orden en que se listan los nodos asociados a un elemento dado en la tabla de conexiones.

Ensamblaje de la Matriz A

Resumiendo lo que hemos hecho para calcular los coeficientes de la matriz A, usando la técnica elemento por elemento construir la matriz A es volver a *armarla* a partir de las matrices elementales dadas en la definición (8.4) de la página 137. Este proceso se llama ensamblaje de la matriz.

Ejemplo 8.7. Ejemplo sencillo de ensamblaje de la matriz
La discretización de este ejemplo está dada por la figura (8.13). La matriz $A \in \mathbb{R}^{4\times 4}$ del sistema lineal se inicializa con ceros

$$A_{ij} = 0, \qquad i, j = 1, \ldots, 4.$$

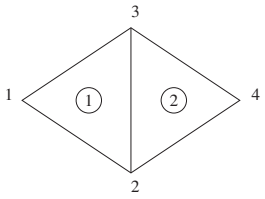

Figura 8.13. Discretización del ejemplo (8.7).

Nodos	x, y
1	x_1, y_1
2	x_2, y_2
3	x_3, y_3
4	x_4, y_4

Elemento	Sus nodos		
1	1	2	3
2	2	3	4

Tabla 8.4. Tablas de coordenadas y conexiones del ejemplo (8.7).

Para el elemento 1, tenemos

$$E_1 = \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix}$$

y

$$E_1^{-1} = \begin{bmatrix} a_1^{(1)} & a_2^{(1)} & a_3^{(1)} \\ b_1^{(1)} & b_2^{(1)} & b_3^{(1)} \\ c_1^{(1)} & c_2^{(1)} & c_3^{(1)} \end{bmatrix}.$$

Ahora calculemos los coeficientes de la matriz elemental $A^{(1)} \in \mathbb{R}^{3\times 3}$, que corresponde al elemento 1.

$$A_{ij}^{(1)} = \int_{T_1} \left(a_i^{(1)} a_j^{(1)} + b_i^{(1)} b_j^{(1)} \right) dx + c \int_{T_1} \left(a_i^{(1)} x + b_i^{(1)} y + c_i^{(1)} \right) \left(a_j^{(1)} x + b_j^{(1)} y + c_j^{(1)} \right) dx, \qquad i,j = 1,2,3.$$

Después de eso empecemos a ensamblar la matriz A con la matriz elemental calculada para el elemento 1.

$$A^{(1)} = \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} & A_{13}^{(1)} & 0 \\ A_{21}^{(1)} & A_{22}^{(1)} & A_{23}^{(1)} & 0 \\ A_{31}^{(1)} & A_{32}^{(1)} & A_{33}^{(1)} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Ahora, pasando al elemento siguiente que es el elemento 2, de la tabla de conexiones tenemos

$$E_2 = \begin{bmatrix} x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \\ x_4 & y_4 & 1 \end{bmatrix}$$

y

$$E_2^{-1} = \begin{bmatrix} a_1^{(2)} & a_2^{(2)} & a_3^{(2)} \\ b_1^{(2)} & b_2^{(2)} & b_3^{(2)} \\ c_1^{(2)} & c_2^{(2)} & c_3^{(2)} \end{bmatrix}.$$

Nota 8.15.
El índice en los coeficientes es 2 porque la matriz es la asociada al elemento 2.

De la misma manera que para el elemento 1, calculamos la matriz elemental $A^{(2)}$ correspondiente al elemento 2.

$$A_{ij}^{(2)} = \int_{T_2} \left(a_i^{(2)} a_j^{(2)} + b_i^{(2)} b_j^{(2)} \right) dx + c \int_{T_2} \left(a_i^{(2)} x + b_i^{(2)} y + c_i^{(2)} \right) \left(a_j^{(2)} x + b_j^{(2)} y + c_j^{(2)} \right) dx, \qquad i,j = 1,2,3.$$

Como

$$a_{ij} = \sum_k a_{ij}^{(k)},$$

donde la suma se hace sobre todos los elementos k comunes entre los soportes de las funciones φ_i, φ_j, entonces el ensamblaje del elemento 2 en la matriz A le da la forma

$$A = \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} & A_{13}^{(1)} & 0 \\ A_{21}^{(1)} & A_{22}^{(1)} + A_{11}^{(2)} & A_{23}^{(1)} + A_{12}^{(2)} & A_{13}^{(2)} \\ A_{31}^{(1)} & A_{32}^{(1)} + A_{21}^{(2)} & A_{33}^{(1)} + A_{22}^{(2)} & A_{23}^{(2)} \\ 0 & A_{31}^{(2)} & A_{32}^{(2)} & A_{33}^{(2)} \end{bmatrix}.$$

Cálculo del segundo miembro del sistema $Ax = b$

Sabemos que

$$b_i = \int_{\Omega} \varphi_i f \, dx,$$

o también

$$b_i = \int_{\mathrm{sop}(\varphi_i)} \varphi_i f \, dx.$$

Esta forma de escribir b_i, $i = 1, \ldots, n$, nos sugiere, como en el caso de la matriz A, trabajar elemento por elemento y hacer un ensamblaje. Así, para cada elemento k, se buscan en la tabla de conexiones los nodos k_1, k_2, k_3 que lo forman y se calcula el vector elemental $b^{(k)} \in \mathbb{R}^3$ definido por

$$b_i^{(k)} = \int_{T_k} f \varphi_i \, dx, \qquad i = 1, 2, 3.$$

Ejemplo 8.8. Para el ejemplo (8.7) de la página 142, el vector $F \in \mathbb{R}^4$ tiene la forma

$$F = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} + b_1^{(2)} \\ b_3^{(1)} + b_2^{(2)} \\ b_3^{(2)} \end{bmatrix}.$$

Integración de las condiciones de frontera de tipo Dirichlet

Normalmente, los n nodos de la discretización incluyen los nodos que están en la frontera. La combinación de la nota (8.10) de la página 136 y la condición de frontera $u(x, y) = 0$ en $\partial\Omega$ indica que $\lambda_i = 0$ para todos los nodos i de la frontera.

Así, en el sistema lineal que resulta del ensamblaje $A\Lambda = b$ hay que *forzar* las incógnitas λ_i que corresponden a los nodos de la frontera a tomar el valor 0.

Una manera de hacer eso es el proceso siguiente. Para todo nodo i de la frontera hacer

$$\begin{cases} a_{ii} = 1, \\ a_{ij} = 0, \quad i \neq j, \\ b_i = 0. \end{cases}$$

Si las condiciones de frontera son de tipo Dirichlet pero no homogéneas, es decir si son de la forma $u(x, y) = g(x, y)$, entonces se procede de la misma manera que en el caso anterior con los siguientes cambios. Si i es un nodo sometido a esta condición de frontera entonces el sistema lineal cambia de la forma

$$\begin{cases} a_{ii} = 1, \\ a_{ij} = 0, \quad i \neq j, \\ b_i = g(x_i, y_i). \end{cases}$$

Ejemplo 8.9. Ejemplo detallado

Sea $\Omega = [0\ 1] \times [0\ 1]$. Consideramos el problema $\Delta u(x, y) = -26$ en Ω con las condiciones de frontera

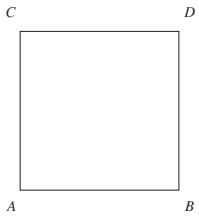

Figura 8.14.

$$u = \begin{cases} 3x^2 - x, & \text{en el segmento } [A\ B], \\ 3 + 10y^2 - 1, & \text{en el segmento } [B\ D], \\ 3x^2 - 2x + 11, & \text{en el segmento } [D\ C] \text{ y} \\ 10y^2 + y, & \text{en el segmento } [C\ A]. \end{cases}$$

La solución exacta de este problema es

$$u(x, y) = 3x^2 + 10y^2 - xy - x + y.$$

Realizamos la discretización (figura 8.15) con las siguientes características:

Tipo de elemento	Triángulo con 3 nodos
Número de nodos	16
Número de elementos	18

La tabla (8.5.1) es la tabla de coordenadas. La primera columna es el número del nodo y la segunda sus coordenadas. La tabla (8.5.2) es la tabla de conexiones.

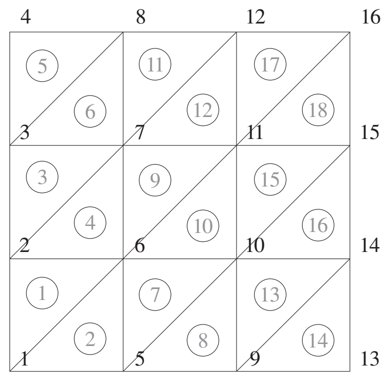

Figura 8.15. Discretización

Nota 8.16.

Es una redundancia poner la primera columna en estas tablas pero por razones de legibilidad se incluye en este ejemplo.

A continuación veremos cómo se llenan la matriz $A \in \mathbb{R}^{16\times 16}$ y el vector b del sistema lineal a medida que los elementos se recorren uno a uno.
Para el elemento 1

$$A = \begin{bmatrix}
\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{1}{2} & 1 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -\frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

y

$$b^t = \left[-\tfrac{13}{27}, -\tfrac{13}{27}, 0, 0, 0, -\tfrac{13}{27}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\right].$$

Nodos	x,y
1	0,0
2	$0,\frac{1}{3}$
3	$0,\frac{2}{3}$
4	0,1
5	$\frac{1}{3},0$
6	$\frac{1}{3},\frac{1}{3}$
7	$\frac{1}{3},\frac{2}{3}$
8	$\frac{1}{3},1$
9	$\frac{2}{3},0$
10	$\frac{2}{3},\frac{1}{3}$
11	$\frac{2}{3},\frac{2}{3}$
12	$\frac{2}{3},1$
13	1,0
14	$1,\frac{1}{3}$
15	$1,\frac{2}{3}$
16	1,1

Elemento	Sus nodos		
1	1	2	6
2	1	6	5
3	2	3	7
4	2	7	6
5	3	4	8
6	3	8	7
7	5	6	10
8	5	10	9
9	6	7	11
10	6	11	10
11	7	8	12
12	7	12	11
13	9	10	14
14	9	14	13
15	10	11	15
16	10	15	14
17	11	12	16
18	11	16	15

Tabla 8.5. Tablas de coordenadas y conexiones

Para el elemento 2

$$A = \begin{bmatrix}
1 & -\frac{1}{2} & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{1}{2} & 1 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{1}{2} & 0 & 0 & 0 & 1 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -\frac{1}{2} & 0 & 0 & -\frac{1}{2} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

y

$$b^t = \left[-\tfrac{26}{27}, -\tfrac{13}{27}, 0, 0, -\tfrac{13}{27}, -\tfrac{26}{27}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\right].$$

Finalmente, para el elemento 18

$$A = \begin{bmatrix}
1 & -\frac{1}{2} & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{1}{2} & 2 & -\frac{1}{2} & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -\frac{1}{2} & 2 & -\frac{1}{2} & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -\frac{1}{2} & 1 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{1}{2} & 0 & 0 & 0 & 2 & -1 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & -1 & 2 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 2 & -1 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & -1 & 2 & 0 & 0 & 0 & -\frac{1}{2} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 1 & -\frac{1}{2} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -\frac{1}{2} & 2 & -\frac{1}{2} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -\frac{1}{2} & 2 & -\frac{1}{2} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & -\frac{1}{2} & 1
\end{bmatrix}$$

y

$$b^t = \left[-\tfrac{26}{27}, -\tfrac{13}{9}, -\tfrac{13}{9}, -\tfrac{13}{27}, -\tfrac{13}{9}, -\tfrac{26}{9}, -\tfrac{26}{9}, -\tfrac{13}{9}, -\tfrac{13}{9}, -\tfrac{26}{9}, -\tfrac{26}{9}, -\tfrac{13}{9}, -\tfrac{13}{27}, -\tfrac{13}{9}, -\tfrac{13}{9}, -\tfrac{26}{27}\right].$$

La tabla siguiente da los nodos de la frontera (no hay un orden).

1	2	3	4	5	8	9	12	13	14	15	16

La aplicación de las condiciones de frontera a estos nodos da el sistema lineal

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

$$F^t = \left[0, \tfrac{13}{9}, \tfrac{46}{9}, 11, 0, \tfrac{-26}{9}, \tfrac{-26}{9}, \tfrac{32}{3}, \tfrac{2}{3}, \tfrac{-26}{9}, \tfrac{-26}{9}, 11, 2, \tfrac{28}{9}, \tfrac{58}{9}, 12\right],$$

cuya solución es

$$\left[0, \tfrac{13}{9}, \tfrac{46}{9}, 11, 0, \tfrac{4}{3}, \tfrac{44}{9}, \tfrac{32}{3}, \tfrac{2}{3}, \tfrac{17}{9}, \tfrac{16}{3}, 11, 2, \tfrac{28}{9}, \tfrac{58}{9}, 12\right]^t.$$

La solución exacta en los nodos es

$$\left[0, \tfrac{13}{9}, \tfrac{46}{9}, 11, 0, \tfrac{4}{3}, \tfrac{44}{9}, \tfrac{32}{3}, \tfrac{2}{3}, \tfrac{17}{9}, \tfrac{16}{3}, 11, 2, \tfrac{28}{9}, \tfrac{58}{9}, 12\right]^t,$$

lo que da un error

$$\left[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\right]^t.$$

Bibliografía

[1] Richard L. Burden and J. Douglas Faires. *Análisis numérico,* 7a. ed. Thomson Learning, 2002.

[2] Philippe G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation.* MASSON, 1990.

[3] Philippe G. Ciarlet and Jacques-Louis Lions, editors. *Handbook of numerical analysis,* Vol. 1, P.1. North-Holland, 1990.

[4] Philippe G. Ciarlet, B. Miara, and Jean-Marie Thomas. *Excercices d'analyse numérique matricielle et d'optimisation avec solutions,* 2è ed. MASSON, 1987.

[5] Michel Crouzeix and Alain L. Mignot. *Exercices d'analyse numérique des équations différentielles.* MASSON, 1993.

[6] Michel Crouziex and Alain L. Mignot. *Analyse numérique des équations différentielles,* 2è ed. rev. et augm. MASSON, 1992.

[7] Robert Dautray and Jacques-Louis Lions. *Analyse mathématique et calcul numérique,* Vol. 3. MASSON, 1987.

[8] Saber Elaydi. *An introduction to difference equations.* Springer, 1999.

[9] Daniel Euvrard. *Résolution numérique des équations aux dérivées partielles : de la physique, de la mécanique et des sciences de l'ingénieur : différences finies, éléments finis, problèmes en domaine non borné.* MASSON, 1994.

[10] Anne Greenbaum. *Iterative methods for solving linear systems.* SIAM, 1997.

[11] Kenneth Hoffman and Ray Kunze. *Álgebra lineal.* Prentice-Hall Hispanoamericana, S.A., 1973.

[12] Hayrettin Kardestuncer, Douglas H. Norrie, and F. Brezzi, editors. *Finite element handbook.* McGraw-Hill, 1987.

[13] Young W. Kwon and Hyochoong Bang. *The finite element method using MATLAB.* CRC Press, 1997.

[14] P. Lascaux and R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur,* Vol. 2, 2è ed. refondue. MASSON, 1994.

[15] M. J. D. Powell. *Approximation theory and methods.* Cambridge University Press, 1981.

[16] Pierre Arnaut Raviart and J. M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles.* MASSON, 1983.

[17] Michael Renardy and Robert C. Rogers. *An introduction to partial differential equations.* Springer-Verlag, 1993.

[18] Michelle Schatzman. *Analyse numérique.* MASSON/InterEditions, 1991.

[19] A. H. Stroud. *Approximate calculation of multiple integrals.* Prentice-Hall, 1971.

Índice alfabético